TU WIEN Informatics

WiFi-Based Person-Centric Sensing

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

Doktor der Technischen Wissenschaften

by

Dipl.-Ing. Julian Strohmayer, BSc.
Registration Number 01426125

to the Faculty of Informatics

at the TU Wien

Advisor: Privatdoz. Dipl.-Ing. Dr.techn. Martin Kampel

The dissertation has been reviewed by:

Eftim Zdravevski	Peter Počta

Vienna, 20th September, 2025

Julian Strohmayer

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Dipl.-Ing. Julian Strohmayer, BSc.

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 20. September 2025

Julian Strohmayer

Danksagung

Abbildung 1: Meme-Archiv des TU Wien Computer Vision Lab.

An dieser Stelle möchte ich meinem Betreuer, Martin Kampel, meinen aufrichtigen Dank für seine kontinuierliche Unterstützung, Ermutigung und Begleitung während meines Doktorats aussprechen. Sein Vertrauen und die Freiheit, die er mir gewährte, meine eigenen Ideen zu verfolgen, waren entscheidend dafür, die Richtung dieser Arbeit maßgeblich zu prägen.

Mein herzlicher Dank gilt meinen Gutachtern, Eftim Zdravevski und Peter Počta, für ihre Zeit und Mühe, die sie in die Begutachtung dieser Arbeit investiert haben, sowie für ihr konstruktives Feedback.

Besonders bedanken möchte ich mich bei meinen Kolleginnen und Kollegen am Computer Vision Lab für die einzigartige, unterstützende und angenehme Arbeitsatmosphäre. Ganz besonders genoss ich die unterhaltsamen Veranstaltungen und Feiern (die mir eine ungesunde Dosis Leberkäse bescherten), die zahllosen Kaffeepausen mit lustigen Diskussionen sowie die erstklassige Meme-Kultur (siehe Abbildung 1).

Von Herzen danke ich meiner Familie für ihre uneingeschränkte Unterstützung und ihren Glauben an mich, nicht nur in den vergangenen Jahren, sondern während meines gesamten Lebens.

Abschließend gilt mein besonderer Dank meiner Frau Jasmin und meinem Sohn Levi. Ihre Liebe, Geduld und Bodenständigkeit waren mir täglich eine Quelle von Kraft und Perspektive. Danke, dass ihr jeden Schritt dieses Weges an meiner Seite wart.

Acknowledgements

Figure 2: Meme archive of the TU Wien Computer Vision Lab.

I would like to express my sincere gratitude to my supervisor, Martin Kampel, for his continuous guidance, encouragement, and support throughout my PhD journey. His trust and the freedom he granted me to explore my own ideas were instrumental in shaping the direction of this work.

I am deeply thankful to my reviewers, Eftim Zdravevski and Peter Počta, for their time and effort in evaluating this thesis and for their constructive feedback.

I would also like to thank my colleagues at the Computer Vision Lab for creating a uniquely supportive and enjoyable work environment. I particularly appreciated the fun events and celebrations (providing me with an unhealthy dose of Leberkäse), the countless coffee breaks filled with hilarious discussions, and the world-class meme culture (see Figure 2).

My heartfelt thanks go to my family for their unwavering support and belief in me, not only during the past years but throughout my entire life.

Finally, I am especially grateful to my wife Jasmin and my son Levi. Their love, patience, and grounding presence have been my daily source of strength and perspective. Thank you for being there, every step of the way.

Kurzfassung

WiFi-basiertes Person-Centric Sensing (PCS) bietet eine visuelle Privatsphäre wahrende Alternative zu optischen Verfahren durch passive, kontaktlose Überwachung über bestehende drahtlose Infrastruktur. Aufgrund der geringen Kosten, der diskreten Integration und der Wanddurchdringung eignet sich WiFi-basiertes PCS besonders für die großflächige Überwachung von Innenräumen. Die praktische Umsetzung wird jedoch durch fehlende öffentliche Datensätze, begrenzte Reichweite geeigneter WiFi-Hardware, ineffiziente Inferenz auf eingebetteten Geräten, mangelnde Generalisierbarkeit sowie die schwer interpretierbare Natur von WiFi-Signalen erschwert.

Zur Behebung des Datenmangels werden fünf öffentlich verfügbare Channel State Information (CSI)-Datensätze bereitgestellt: TOA, Wallhack1.8k, HALOC, 3DO und WiFiCam. Als Evaluierungsgrundlage addressieren diese spezifische Herausforderungen in Langstreckenüberwachung, Domänengeneralisierung und multimodaler Translation.

Darauf aufbauend wird gezeigt, dass gerichtete handelsübliche WiFi-Hardware Langstreckenüberwachung ermöglicht. Experimente bestätigen Präsenzdetektion, Aktivitätserkennung und Lokalisierung über bis zu 20 Meter und mehrere Räume mit einem Single-Link-Setup und belegen somit die Effektivität des vorgestellten Ansatzes.

Mit WiFlexFormer wird eine kompakte Transformer-Architektur vorgestellt, optimiert für die spektralen und temporalen Eigenschaften von CSI. Mit nur $\approx$ 50 Tsd. Parametern und $\approx$ 10 ms Inferenzzeit auf Embedded-Hardware übertrifft WiFlexFormer deutlich größere generische Vision-Architekturen als auch spezialisierte RF-Architekturen in Effizienz bei ähnlicher oder besserer Genauigkeit.

Zur Verbesserung der Generalisierbarkeit werden Datenaugmentierung und Preprocessing-Techniken untersucht, die robuste Modelle ohne Zugriff auf die Zieldomäne ermöglichen. Darauf aufbauend kombiniert das DATTA-Framework Domain-Adversarial Training, Test-Time Adaptation, zufälliges Weight Resetting und gezielte Datenaugmentierung. DATTA ermöglicht Echtzeitanpassung an dynamische Domänen und erreicht dabei State-of-the-Art-Generaliserungsfähigkeit von WiFi-basierten PCS-Modellen.

Abschließend wird mit WiFiCam ein neuer Ansatz zur Synthese von RGB-Bildern aus WiFi CSI in Through-Wall-Szenarien vorgestellt. WiFiCam rekonstruiert kohärente, visuell aussagekräftige Bilder und ermöglicht damit eine kamerafreie visuelle Überwachung. Gleichzeitig wird die Interpretierbarkeit von CSI für nachgelagerte Aufgaben verbessert.

Abstract

WiFi-based Person-Centric Sensing (PCS) offers a visual privacy-preserving alternative to optical methods by enabling passive, contactless monitoring through existing wireless infrastructure. Its low cost, unobtrusive nature, and wall-penetrating capability make it well suited for large-scale indoor monitoring applications. However, practical deployment remains constrained by data scarcity, limited sensing range of consumer off-the-shelf (COTS) hardware, computational inefficiencies, poor cross-domain generalization, and the abstract, non-intuitive nature of WiFi signals.

To address the data scarcity, five publicly available WiFi Channel State Information (CSI)-based PCS datasets are contributed: TOA, Wallhack1.8k, HALOC, 3DO, and WiFiCam. Each dataset targets distinct challenges in long-range and through-wall sensing, domain generalization, and crossmodal translation.

Building on this foundation, it is demonstrated that directional sensing with low-cost COTS WiFi systems enables long-range through-wall PCS. Experiments confirm robust presence detection, activity recognition, and localization up to 20 meters and across multiple rooms with a single-link setup, validating the effectiveness of the proposed directional sensing approach in complex indoor environments.

To support real-time inference under resource constraints, WiFlexFormer, a lightweight Transformer architecture tailored to the temporal and spectral characteristics of WiFi CSI, is introduced. With only $\approx$ 50k parameters, it achieves inference latencies of $\approx$ 10 ms on embedded hardware while matching or surpassing the performance of significantly larger generic vision and RF-specific architectures.

To improve robustness across domains, data augmentation and preprocessing strategies that enhance generalization without target-domain access are investigated. Building on these insights, the Domain-Adversarial Test-Time Adaptation (DATTA) framework is proposed. DATTA leverages domain-adversarial training, test-time adaptation, random weight resetting, and data augmentation to enable robust, real-time adaptation to domain shifts, achieving state-of-the-art cross-domain generalization performance.

Lastly, the thesis presents the first approach to synthesize RGB images from WiFi CSI in through-wall scenarios. The WiFiCam architecture, based on a multimodal variational autoencoder, reconstructs coherent, semantically meaningful images, enabling camera-free visual monitoring and improving the interpretability of CSI for downstream tasks.

These contributions address core limitations in data, hardware, efficiency, generalization, and interpretability, advancing WiFi-based PCS toward scalable real-world deployment.

Contents

CHAPTER 1

Introduction

WiFi represents a promising modality for Person-Centric Sensing (PCS), providing visual privacy-preserving, contactless, and unobtrusive monitoring suitable for large-scale indoor environments. However, practical adoption remains challenging due to a lack of publicly available datasets, limited availability of suitable sensing hardware, inefficient deep learning architectures for real-time processing, and poor cross-domain generalization of models. Beyond addressing these limitations, this thesis explores the potential of WiFi-based PCS and its capabilities in challenging long-range and through-wall scenarios, showcasing its viability as a robust alternative to traditional sensing modalities.

1.1 Motivation

PCS involves the use of wearable and contactless sensors to monitor the activities, behaviors, and interactions of individuals with their surroundings [1, 2]. Wearable sensors are physically attached to the body, capturing data such as motion and physiological signals [3]. In contrast, contactless sensors capture this information remotely by analyzing environmental interactions [3]. The fundamental challenge in PCS lies in extracting meaningful insights from sensor data to understand human behavior and context. This understanding enables a wide range of applications, including healthcare domains such as active assisted living and vital signs monitoring, as well as infrastructure and safety contexts like smart environments, security, human-robot interaction, and autonomous driving [1]. To realize these applications, several core sensing tasks must be effectively addressed, including presence detection, human activity recognition (HAR), and localization or tracking tasks that can be supported by both wearable and contactless sensing modalities. This thesis focuses on these fundamental PCS tasks due to their critical role in enabling practical applications. By capturing and analyzing motion patterns, postural changes, and activity context, these tasks support diverse use cases such as health monitoring [4, 5], rehabilitation [6], home automation [7], hazard detection [8, 9], workplace safety assessment [10], or emergency response optimization [11].

Optical modalities, particularly RGB cameras, historically dominate PCS because of their accessibility, high information density, and compatibility with deep learning frameworks [3, 12]. The development of Convolutional Neural Networks (CNNs), such as AlexNet in 2012 [13, 14], has revolutionized computer vision, enabling automated feature extraction from visual data. Cameras provide rich, fine-grained information on human appearance, posture, and motion, making them suitable for tasks like person detection [15], semantic segmentation [16], pose estimation [17], and activity recognition [18]. Additionally, their ever decreasing cost and widespread availability have solidified their position as a convenient and effective data source in PCS research and development.

Despite these advantages, optical modalities pose significant challenges in privacy-sensitive settings. Cameras inherently capture identifiable features such as color and texture [19]. In environments like homes and healthcare facilities, their presence can create a sense of surveillance, leading to discomfort and rejection [20]. This perceived privacy violation persists even when the systems aim to provide critical services such as medical emergency detection, as the mere presence of an imaging device can be perceived as invasive [21]. Attempts to mitigate these concerns with depth or thermal cameras [22, 23] have had limited success. While these alternative modalities can enhance visual privacy by reducing feature fidelity, they still facilitate person identification [24, 25, 26] and evoke psychological responses associated with surveillance due to their camera-like appearance [20]. Surveys on technology acceptance in privacy-sensitive contexts consistently highlight that camera-based systems rank among the least preferred technologies for both public and private monitoring [20, 27, 28]. This stark contrast between the capabilities of state-of-the-art PCS systems and user acceptance highlights a potential misalignment in the design of privacy-sensitive sensing technologies.

To address the privacy challenges and user rejection associated with optical modalities, a range of non-visual sensing modalities have been explored for PCS [1]. These include mechanical, electric, electromagnetic, acoustic, and environmental sensors, each varying in their privacy protection, range, system complexity, and cost. However, a viable alternative to cameras must be both privacy-preserving and contactless. Wearable sensors, while effective, are impractical due to discomfort, movement restrictions, frequent recharging, and user forgetfulness [23]. These drawbacks are particularly significant in emergency situations involving individuals suffering from cognitive impairments like dementia who might forget to wear or use the device [23].

This requirement for both contactless operation and visual privacy protection narrows the set of viable alternatives to acoustic, environmental, and electromagnetic modalities. Acoustic sensors have been explored due to their ability to infer human activity without capturing visual data. Acoustic sensors operating in the audible spectrum (20 Hz to 20 kHz) can be leveraged to passively monitor ambient sound patterns facilitating the recognition of human activities but raise privacy concerns due to their potential to capture speech, which is perceived as highly intrusive [29, 3]. Ultrasonic sensors actively emit inaudible sound waves (>20 kHz) to detect reflections, mitigating speech-related privacy issues but suffering from range limitations and environmental interference [30,

3]. Moreover, surface acoustic sensors such as accelerometers and geophones detect vibrations caused by human activity [31]. Although privacy-preserving, their effectiveness is constrained by structural dependencies and limited range.

Environmental sensors (e.g., air temperature, pressure, or humidity, etc.), have also been explored for PCS due to their inherent privacy-preserving nature. These sensors can be used to infer human activity indirectly by detecting disturbances in ambient conditions [32]. While cost-effective and non-intrusive, they provide low information density and limited range, making them unsuitable for fine-grained PCS in larger spaces without data fusion with other sensing modalities.

Finally, electromagnetic sensing approaches, such as radar and WiFi, have garnered significant attention for their ability to enable contactless, unobtrusive monitoring of human behavior without capturing visual data [33]. Radar sensors leverage the Doppler effect by emitting radio frequency (RF) signals and analyzing their reflections to detect motion, micro-gestures, and activity patterns [34]. While radar offers robust privacy protection, range and coverage in indoor environments is more limited compared to WiFi [35]. Additionally, radar systems are constrained by high hardware costs and the complexity of signal processing, which make them less suitable for large-scale or consumer-level PCS applications [35].

WiFi-based sensing represents a potential paradigm shift in PCS by leveraging the ubiquitous WiFi signals already present in most indoor environments [36]. These systems analyze variations in signal properties, such as amplitude and phase, to infer human activities [37], locations [38], and even physiological parameters [39]. Unlike radar, which requires specialized hardware and is limited by higher costs and shorter range in indoor environments [35], WiFi-based enables long-range and multi-room monitoring [40]. Standard consumer off-the-shelf (COTS) WiFi devices can monitor entire buildings [41] while maintaining low costs and minimal system complexity, making them particularly well-suited for scalable and privacy-sensitive applications.

WiFi can address the limitations of both optical and alternative non-visual modalities. Unlike cameras, WiFi signals do not capture any visual information such as color or texture, inherently preserving the visual privacy of users. In comparison to acoustic and environmental sensors, WiFi offers superior range, higher information density, and the ability to penetrate walls for seamless multi-room coverage [40]. Its accessibility, driven by the widespread deployment of WiFi infrastructure, positions WiFi-based PCS as a potential alternative to traditional optical approaches, particularly in privacy-sensitive domains like healthcare, assisted living, and smart environments. By offering a non-invasive, cost-effective, and widely accessible solution, WiFi-based sensing aligns with user preferences and technological demands, making it an ideal candidate for next-generation PCS applications.

Despite its advantages, several challenges currently limit the broader adoption of WiFi-based PCS. The availability of suitable COTS hardware remains limited, as only a small subset of devices support the capture of Channel State Information (CSI) [42],

a requirement for modern WiFi-based PCS. Additionally, there is a lack of publicly available datasets, especially for scenarios such as through-wall sensing, which hinders the development and evaluation of systems tailored for these applications [43, 44]. Another key challenge is the need for efficient deep learning architectures that can leverage the unique characteristics of WiFi signals while enabling real-time on-device inference on the edge [45, 46]. A further limitation is the poor cross-domain generalization of models trained on WiFi signals, due to the inherent sensitivity of CSI to domain variations [47], as well as the limited interpretability of WiFi signals, which hinders transparency and constrains their use in semantically meaningful or human-in-the-loop applications.

1.2 Research Questions and Contributions

This thesis is guided by four central research questions, reflecting the current limitations of WiFi-based PCS, including the restricted availability of suitable low-cost COTS WiFi systems, the need for efficient deep learning architectures that operate under resource constraints, the poor generalization of models across varying domains, and the limited interpretability of WiFi signals for downstream use. Each question targets a specific aspect of these limitations and serves as the foundation for the technical contributions presented in subsequent chapters.

RQ I: How can long-range PCS be achieved with COTS WiFi systems while minimizing complexity and cost? This research question examines the feasibility of implementing cost-effective PCS in partitioned indoor environments using minimal COTS WiFi infrastructure. While WiFi signals naturally propagate through walls and across entire buildings, conventional COTS WiFi devices are designed for communication rather than sensing. Consequently, default configurations employ omnidirectional antennas that prioritize connection stability over spatial selectivity, creating substantial challenges for through-wall PCS due to signal attenuation and multipath distortion. Moreover, a naive distributed sensing approach would involve the deployment of numerous devices throughout a building, leading to high system complexity through increased infrastructure requirements, coordination overhead, and total system cost. To address these limitations, the potential of minimal single-link configurations, consisting of just one transmitter and one receiver, is examined, focusing on whether the PCS capabilities of COTS WiFi systems can be enhanced through a directional sensing approach.

RQ II: How can deep learning architectures be designed to efficiently process WiFi CSI for real-time PCS on low-power edge devices, while accounting for the unique characteristics of WiFi signals? This research question addresses the challenge of deploying PCS on resource-constrained edge devices. Certain applications require real-time on-device inference at the edge to avoid the transmission of sensitive data and to meet the low-latency requirements of responsive PCS systems. However, deploying on low-power edge devices presents significant challenges due to strict constraints in computation, memory, and energy consumption. Generic deep learning architectures from computer vision and natural language processing domains are unoptimized for CSI processing because they either introduce excessive parameter counts or impose

inappropriate inductive biases, such as translation equivariance, that conflict with the non-shift-invariant structure of CSI. Effective architectures for CSI processing must therefore balance accuracy with compactness to enable real-time inference under edge constraints while explicitly exploiting the unique properties of WiFi signals, including multipath propagation, temporal correlations, and amplitude-phase relationships. The central challenge lies in identifying architectural principles and components that leverage these domain-specific characteristics while maintaining lightweight design suitable for practical deployment on resource-constrained devices.

RQ III: How can WiFi-based PCS models be made robust to real-world domain variations? This research question addresses the generalization challenge in WiFi-based PCS where models must maintain performance across different domains, each characterized by a unique combination of environmental, hardware, and operational conditions that collectively shape WiFi signal propagation and sensing characteristics. Domain variations encompass environmental factors such as different room layouts, furniture arrangements, and building materials; hardware variations including different WiFi chipsets and antenna configurations; temporal changes encompassing daily activity patterns and seasonal variations; and demographic diversity spanning different body types and movement patterns. These domain-specific conditions create distinct CSI signatures and signal patterns, causing substantial performance degradation when models trained in one domain encounter the unfamiliar signal characteristics of an unseen domain. To address this challenge, domain-invariant feature learning and domain adaptation techniques are explored to enable robust WiFi-based PCS across diverse deployment scenarios.

RQ IV: How can WiFi signals be made more interpretable for human understanding or downstream use? This research question explores the challenge of transforming abstract WiFi CSI into intuitive representations that humans can readily understand and interpret. While CSI contains rich information about environmental changes and human activities, its high-dimensional, complex nature makes it difficult for humans to directly comprehend or analyze, limiting both system transparency and practical applicability. The key challenge involves determining whether WiFi CSI can be translated into more familiar and interpretable modalities that preserve spatial and temporal information about human activities while making the underlying sensing processes transparent to users. Such transformation approaches not only enable intuitive interpretation of WiFi sensing results but also open new application possibilities for visual privacy-preserving monitoring and facilitate downstream tasks that benefit from more accessible data representations. The central challenge involves developing methods that can bridge the semantic gap between raw CSI patterns and human-interpretable representations while maintaining the fidelity of the original sensing information.

The contributions presented are based on eight publications that collectively advance WiFi-based PCS by addressing key limitations in data, hardware, and algorithmic processing while exploring novel applications that extend the state of the art. The contributions include the collection of specialized datasets to benchmark core PCS tasks,

Research Area	RQ	Contributions
Datasets	∀	TOA [40], Wallhack1.8k [48], HALOC [49], 3DO [50], WiFiCam [51]
WiFi Systems / PCS Feasibility	I	Development and evaluation of a series of ESP32-based WiFi systems [40, 52, 49] using passive reflectors and external directional antennas, verifying the feasibility of long-range through-wall presence detection, HAR, and localization with low-cost COTS WiFi devices.
Efficient Architectures	II	Proposal of *WiFlexFormer* [53], a highly efficient Transformer-based architecture tailored to the unique characteristics of CSI, enabling real-time WiFi-based PCS on edge devices.
Cross-Domain Generalization	III	Systematic evaluation of data augmentation [48] and preprocessing methods [50], and introduction of DATTA [54], a domain-adversarial test-time adaptation framework for robust WiFi-based PCS under domain shifts.
Interpretability / Novel Applications	IV	Proposal of the first method for synthesizing RGB images from through-wall WiFi CSI [51], enabling visual monitoring without cameras and enhancing interpretability for downstream tasks.

Table 1.1: Relation between research areas, research questions and contributions.

novel WiFi system designs for long-range and through-wall sensing, feasibility studies, the development of an efficient architecture for real-time processing of WiFi CSI, and novel frameworks for improving cross-domain generalization. The relation between research areas, research questions and the contributions made are illustrated in Table 1.1. A detailed summary of each contribution follows below.

1.2.1 WiFi System Design & Feasibility of Long-Range and Through-Wall PCS

WiFi System Designs To address the hardware-related aspects of **RQ I**, a series of novel WiFi systems are developed based on the Espressif ESP32 microcontroller platform, which enables low-cost, standalone CSI capture in a compact form factor. The novelty of these systems lies in their ability to achieve directional sensing, which is not supported by the ESP32's default omnidirectional antenna. Directionality improves spatial selectivity and signal quality by focusing transmission and reception toward the sensing target, thereby enhancing robustness and range, especially in through-wall sensing scenarios. The first system, introduced in [40], integrates the ESP32-S3 with a custom biquad antenna to enable directional sensing. In a follow-up work [52], two complementary

approaches are investigated: the use of custom passive reflectors with the built-in antenna, and design optimizations to the biquad antenna via refined reflector geometries. Finally, [49] incorporates a commercial panel antenna for improved directionality. All system designs are publicly available, providing a reproducible platform for evaluating long-range, through-wall WiFi-based PCS.

Feasibility of Long-Range and Through-Wall PCS Complementing the system development, a series of experiments is conducted to evaluate whether the proposed low-cost, single-link COTS WiFi systems are capable of supporting long-range and through-wall PCS scenarios, as targeted in **RQ I**. In [40], presence detection and HAR are successfully performed across five rooms and a distance of 20 m. Building on this setup, [52] further demonstrates reliable HAR under similar through-wall conditions. Finally, [49] shows that the same class of hardware can also enable long-range localization in a line-of-sight (LOS) scenario. These experiments confirm that low-cost, COTS WiFi systems, when equipped with directional sensing, can facilitate person-centric sensing over substantial spatial scales.

Datasets In addition to the proposed WiFi systems and feasibility studies, the following publicly available datasets are contributed, each recorded using one or more of the proposed WiFi systems. These datasets serve to verify systems functionality and to empirically evaluate the feasibility of long-range through-wall PCS under controlled and repeatable conditions.

- The *Through-Wall Office Activities (TOA)* dataset [40] is designed to evaluate presence detection and HAR performance in LOS and through-wall scenarios, serving as a foundation for assessing the sensing capabilities of the proposed WiFi systems.
- Recorded in the same environment as TOA, the *Wallhack1.8k* dataset [52] extends this work by incorporating data from two additional WiFi systems, enabling the evaluation of cross-system and cross-scenario generalization in HAR tasks.
- The *HAllway LOCalization (HALOC)* dataset [49] contains WiFi packet sequences synchronized with 3D trajectory data, supporting the benchmarking of WiFi fingerprint-based localization over large indoor areas.

Publications:

- **Julian Strohmayer** and Martin Kampel. Wifi csi-based long-range through-wall human activity recognition with the esp32. In *International Conference on Computer Vision Systems (ICVS)*, pages 41–50. Springer, 2023, doi: https://doi.org/10.1007/978-3-031-44137-0_4.
- **Julian Strohmayer** and Martin Kampel. Directional antenna systems for long-range through-wall human activity recognition. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 3594–3599, 2024, doi: https://doi.org/10.1109/ICIP51287.2024.10647666.

- **Julian Strohmayer** and Martin Kampel. Wifi CSI-based long-range person localization using directional antennas. In *The Second Tiny Papers Track at ICLR 2024*, 2024, doi: https://openreview.net/forum?id=AOJFcEh5Eb.

1.2.2 Efficient Architectures for the Processing of CSI

Real-time WiFi-based PCS on edge devices requires a careful balance between inference speed, memory footprint, and accuracy. Addressing **RQ II**, it is investigated how deep learning architectures can be designed to meet these constraints while accounting for the unique characteristics of WiFi CSI. Existing models tailored to RF/WiFi signals are overly complex, resulting in high parameter counts with limited performance gains [55], while CNN-based approaches impose translational-equivariance priors that are misaligned with the non-shift-invariant structure of CSI [46].

In this context, *WiFlexFormer* [53] is introduced as a highly efficient Transformer-based architecture that naturally captures the global spectral and temporal dependencies of WiFi CSI. Comprehensive evaluations show that *WiFlexFormer* performs competitively with both state-of-the-art RF/WiFi-specific and generic vision architectures, while requiring only $\approx$ 50k parameters (a reduction by three orders of magnitude) and achieving an inference time of $\approx$ 10 ms on a low-power single-board computer. These results make *WiFlexFormer* particularly well-suited for scalable, on-device, real-time WiFi-based PCS.

Publications:

- **Julian Strohmayer**, Matthias Wödlinger, and Martin Kampel. Wiflexformer: Efficient wifi-based person-centric sensing. *arXiv preprint arXiv:2411.04224*, 2024, doi: https://doi.org/10.48550/arXiv.2411.04224.

1.2.3 Cross-Domain Generalization

Robustness to domain shifts is a critical challenge in deploying WiFi-based PCS in real-world settings. Addressing **RQ III**, it is investigated how models can be made robust to domain variations arising from changes in the environment, sensing scenario, and hardware. To this end, a series of studies evaluate strategies for improving cross-domain generalization in the context of HAR and localization tasks.

In [48], the effectiveness of data augmentation techniques is systematically assessed using the *Wallhack1.8k* dataset to improve generalization across scenarios (LOS vs. through-wall) and WiFi systems. Complementing this, [50] explores a range of preprocessing methods, including CSI feature extraction, scaling, and dimensionality reduction, on the *3DO* dataset, which captures static, dynamic, and temporal domain variations. Finally, [54] introduces Domain-Adversarial Test-Time Adaptation (DATTA), a novel framework which integrates domain-adversarial training with test-time adaptation to facilitate rapid adaptation to unseen or changing WiFi domains while mitigating catastrophic forgetting. On public benchmarks, DATTA outperforms state-of-the-art models by up to

8.1 percentage points in F1-score, while maintaining real-time inference capabilities on edge hardware.

Datasets To support the study of cross-domain generalization, the following publicly available datasets are contributed, each capturing complementary sources of domain variability. These datasets are used to benchmark model performance under changes in environment, sensing scenario, and WiFi system configuration.

- The previously introduced *Wallhack1.8k* dataset [48] is used to assess the effectiveness of data augmentation techniques for improving cross-scenario (LOS vs. through-wall) and cross-system generalization in HAR tasks.
- The *3-Days Office (3DO)* dataset [50] provides WiFi packets with both activity and 3D trajectory labels, serving as a benchmark for evaluating cross-domain generalization in through-wall scenarios across static, dynamic, and temporal domain variations.

Publications:

- **Julian Strohmayer** and Martin Kampel. Data augmentation techniques for cross-domain wifi csi-based human activity recognition. In *IFIP International Conference on Artificial Intelligence Applications and Innovations (AIAI)*, pages 42–56. Springer, 2024, doi: https://doi.org/10.1007/978-3-031-63211-2_4.
- **Julian Strohmayer** and Martin Kampel. On the generalization of wifi-based person-centric sensing in through-wall scenarios. In *Pattern Recognition*, pages 194–211, Cham, 2025. Springer Nature Switzerland, doi: https://doi.org/10.1007/978-3-031-78354-8_13.
- **Julian Strohmayer**, Rafael Sterzinger, Matthias Wödlinger, and Martin Kampel. Datta: Domain-adversarial test-time adaptation for cross-domain wifi-based human activity recognition. *arXiv preprint arXiv:2411.13284*, 2024, doi: https://doi.org/10.48550/arXiv.2411.13284.

1.2.4 CSI-based Through-Wall Imaging

To explore how WiFi signals can be made more interpretable for human understanding and downstream use, as posed in **RQ IV**, the WiFiCam architecture, a novel approach for synthesizing RGB images directly from WiFi CSI captured in through-wall scenarios, is introduced [51]. The proposed method leverages a multimodal Variational Autoencoder (VAE) adapted for joint processing of CSI and visual data, enabling the reconstruction of images from WiFi signals during inference. This approach opens new application possibilities such as through-wall visual monitoring without cameras, offering a privacy-preserving alternative for indoor sensing. Additionally, by translating CSI into an interpretable visual representation, it facilitates downstream tasks that traditionally depend on vision-based data, such as semantic labeling of CSI time series.

Datasets To support research on CSI interpretability and enable supervised learning for image synthesis, the *WiFiCam* dataset is introduced. It is the first to link WiFi CSI with synchronized RGB images in a through-wall scenario and provides the ground truth necessary for training and evaluating generative models for visual reconstruction from WiFi signals.

- The *WiFiCam* dataset [51] links WiFi packet data with synchronized images in a through-wall setting, enabling visual reconstruction from CSI.

Publications:

- **Julian Strohmayer**, Rafael Sterzinger, Christian Stippel, and Martin Kampel. Through-wall imaging based on wifi channel state information. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 4000–4006, 2024, doi: https://doi.org/10.1109/ICIP51287.2024.10647775.

Collectively, these contributions advance WiFi-based PCS by addressing key challenges, such as robust on-device sensing in dynamic environments, and by enabling new application paradigms like through-wall visual monitoring without conventional cameras. Through reproducible COTS WiFi system designs, publicly available datasets, and tailored deep learning architectures for WiFi CSI processing, this work establishes a strong foundation for future research and real-world deployment of WiFi-based PCS technologies.

1.3 Thesis Structure

The remainder of this thesis is organized into seven chapters that build upon one another to address the challenges of WiFi-based PCS. Chapter 2 introduces the core technical concepts underlying WiFi-based PCS, including wireless communication principles, CSI representation, and sensing mechanisms. Chapter 3 surveys the evolution of WiFi-based PCS, focusing on deep learning methods for CSI processing (**RQ II**), cross-domain generalization (**RQ III**), and imaging (**RQ IV**). It highlights current limitations and motivates the need for new systems, datasets, and methods. Chapter 4 presents novel WiFi systems designed for long-range through-wall PCS (**RQ I**), reducing deployment complexity while enabling robust sensing. Chapter 5 introduces five specialized datasets (TOA, Wallhack1.8k, HALOC, 3DO, and WiFiCam) captured with the proposed systems. Each dataset targets specific challenges in long-range, cross-domain, and multimodal settings. Chapter 6 details the methodological contributions related to the four research questions: feasibility of long-range through-wall sensing (**RQ I**), efficient architectures for CSI processing (**RQ II**), techniques for cross-domain generalization (**RQ III**), and CSI-based imaging (**RQ IV**). Chapter 7 reflects on broader implications, discussing open research problems, dual-use potential, and future directions. Chapter 8 concludes with a summary of contributions and their significance for advancing WiFi-based PCS.

Note: This thesis incorporates material of the previously published works [40, 48, 49, 50, 51, 52, 53, 54]. Portions of the original text are reproduced verbatim to retain technical precision and specificity, while other parts are revised for coherence and consistency. Figures from the original publications, with, and without modifications are included (modified figures are marked by a † symbol).

CHAPTER 2

Prerequisites

This chapter provides the technical foundations for understanding WiFi-based PCS. It begins with an overview of the IEEE 802.11 protocol family, emphasizing the role of Orthogonal Frequency-Division Multiplexing (OFDM) in enabling fine-grained wireless channel measurements. It then introduces the two primary signal metrics used in PCS (Received Signal Strength Indicator (RSSI) and CSI) and presents a mathematical model of CSI in the frequency domain. This is followed by a discussion of feature extraction techniques, covering amplitude, phase, and derivative representations. The chapter concludes by outlining the physical sensing principle of WiFi-based PCS and defining the sensing scenarios considered.

2.1 WiFi

The term *WiFi* refers to a family of networking protocols for wireless local area networks (WLANs) defined by the IEEE 802.11 standards, which specify the physical and medium access control layers for wireless local area networks [56]. The 802.11 family continuously evolves, with each amendment introducing enhancements in throughput and reliability. OFDM, introduced in 802.11a, enables parallel transmission over a set of sub-channels (subcarriers) and underpins modern CSI extraction techniques [57]. Multiple-Input Multiple-Output (MIMO), standardized in 802.11n, allows the use of multiple antennas to transmit and receive simultaneous data streams, further increasing throughput and spectral efficiency [58]. As a result, recent amendments, such as 802.11ax (WiFi 6), support theoretical link rates up to 9.6 Gbit/s [59].

Although these advances primarily target high-throughput applications such as video streaming on mobile devices, WiFi-based PCS fundamentally repurposes the wireless communication channel as a sensing medium. Instead of transmitting user data, it leverages the protocol's physical-layer feedback, such as the RSSI or CSI, to capture

domain-dependent signal variations [60, 47]. These metrics, extracted through the controlled transmission of dummy packets without payload, reveal characteristic propagation patterns that encode information about human presence, movement, or activity [40]. In doing so, WiFi is effectively misused as a passive sensor, enabling contactless PCS without modifying the underlying protocol.

2.1.1 Received Signal Strength Indicator

The RSSI quantifies the power level of a received radio signal (such as WiFi) at the receiver's antenna [61]. This signal consists of multiple components arriving via different paths (multipath), resulting in time delays, varying attenuation, and phase shifts. RSSI provides a single scalar metric representing the combined power of these components. The instantaneous complex baseband voltage V at the receiver input, resulting from the superposition of these multipath components, can be represented as:

$$V = \sum_{i=1}^{N} \|V_i\| e^{-j\theta_i} \tag{2.1}$$

where $\|V_i\|$ is the amplitude and θ_i is the phase of the i-th multipath component, and N is the total number of significant multipath components. The instantaneous power is proportional to the squared magnitude of this voltage, $P_{\text{inst}} \propto \|V\|^2$. However, RSSI typically represents the average power received over a defined short period (e.g., during a packet preamble). This average power, measured in milliwatts (mW), is then converted to the logarithmic dBm scale (decibels relative to 1 milliwatt) using the standard formula:

$$\text{RSSI (in dBm)} = 10 \log_{10} \left(\frac{P_{\text{avg}}}{1 \text{ mW}} \right) \tag{2.2}$$

where P_{avg} is the average received power in milliwatts.

Although RSSI is leveraged extensively in early WiFi-based PCS research [62], its inherent limitation lies in its aggregated nature: it provides only a single measure of signal power over the entire WiFi channel, lacking the fine-grained information required for sophisticated PCS tasks [63]. This lack of detail, combined with its limited robustness even in static environments [62], has led to the adoption of CSI in modern WiFi-based PCS approaches, which offers higher information density and environmental stability [64].

2.1.2 Channel State Information

As fittingly put by Yang et al. [62]: "*In a conceptual sense, channel response is to RSSI what a rainbow (color spectrum) is to a sunbeam, where components of different wavelengths are separated.*" While RSSI aggregates the signal power across the entire WiFi channel into a single scalar value, CSI provides a fine-grained decomposition of the wireless channel, describing how individual frequency components are affected by the propagation environment. Introduced with the IEEE 802.11n standard [58], CSI captures the frequency-selective channel response by estimating the complex amplitude

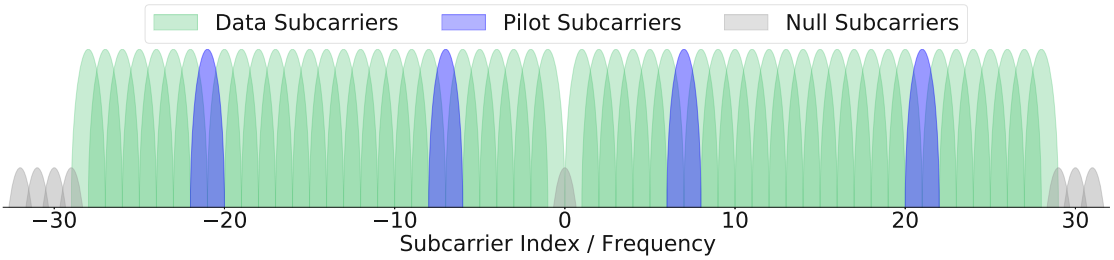


Figure 2.1: OFDM subcarrier layout in the frequency domain of a WiFi 802.11n channel.

attenuation and phase shift, experienced by each OFDM subcarrier during transmission. These estimates are derived at the receiver by comparing known reference symbols (pilots) to the received signal and are used for channel equalization. The resulting CSI matrix encodes detailed information about the multipath characteristics of the environment, capturing subtle temporal and spatial variations caused, for example, by human motion. These variations manifest as distinctive perturbations in the CSI over time, which can be linked to specific activities through appropriate signal processing or learning-based models, thereby enabling PCS [65, 60]. The resolution and structure of CSI are determined by the underlying OFDM modulation scheme used in WiFi.

Orthogonal Frequency-Division Multiplexing OFDM is a modulation scheme used in WiFi protocols, which splits the frequency band of a channel into multiple overlapping frequency sub-bands (subcarriers) over which data is transmitted in parallel for increased bandwidth. In the frequency domain, OFDM signals are modeled as $sinc(f) = \frac{sin(f)}{f}$ functions and can be transformed into the time domain through Inverse Fast Fourier Transform (IFFT). The shape properties of the *sinc* function allow the orthogonal placement of subcarriers, such that the peak of a given subcarrier aligns with the zero-crossings of all other subcarriers. Thus, when computing the sum of all subcarriers, function peaks are retained [45, 66].

A standard 2.4 GHz 802.11n WiFi channel occupies a 20 MHz-wide band, centered on a channel carrier frequency. In Europe, the 2.4 GHz band allows 13 channels, ranging from a channel carrier frequency of 2.412 to 2.472 GHz (5 MHz steps) [67]. Channel 1 for example, is centered on a carrier frequency of 2.412 GHz and occupies the 20 MHz-wide frequency band ranging from 2.402 to 2.422 GHz. As illustrated in Figure 2.1, following the OFDM scheme, the 20 MHz-wide channel is further subdivided into 64 312.5 kHz-wide subcarriers. The set of subcarriers comprises the DC subcarrier, centered on the channel carrier frequency, 52 data subcarriers used for the transmission of encoded data, 7 guard subcarriers (4 guard + 3 null) on the channel borders which mitigate interference between adjacent channels, and 4 pilot subcarriers [68]. Pilot subcarriers serve as a corrective measure for multipath effects. As a result of OFDM, each subcarrier occupies a different carrier frequency within the channel bandwidth and thus experiences different frequency selective amplitude fading effects i.e., the constructive/destructive interference caused by signals propagating over multiple paths of varying length differs between subcarriers. Pilot subcarriers allow, through the transmission of OFDM symbols known by transmitter and

receiver, for the correction of these multipath effects through subcarrier equalization [66]. While Figure 2.1 serves as a good general example to demonstrate the OFDM subcarrier layout, it should be noted that the specific numbers and types of subcarriers can vary depending on the underlying WiFi standard and transmission mode employed (e.g., 802.11n 20 MHz non-HT mode only supports 48 data subcarriers).

Mathematical Model The frequency-selective nature of multipath propagation, captured by CSI across multiple subcarriers, can be formally described using a mathematical scattering model. This model expresses how the transmitted signal is modified by the environment before reaching the receiver. Following the notation in [45], the received signal y is modeled as $y = \mathbb{H}x + \eta$, where x is the transmitted signal vector, $\eta \sim \mathcal{N}(\mu, \Sigma)$ is a Gaussian noise vector, and $\mathbb{H}$ is a complex matrix holding the Channel Frequency Response (CFR) of each subcarrier i, expressed in polar form as $h_i = A_m e^{j\phi_i}$, where A_i and ϕ_i denote amplitude and phase, respectively. These can be computed from the real $\mathcal{R}(h_i)$ and imaginary $\mathcal{I}(h_i)$ parts of the complex CFR as:

$$A_i = \sqrt{((\mathcal{I}(h_i))^2 + (\mathcal{R}(h_i))^2} \qquad (2.3) \qquad \phi_i = \text{atan2}(\mathcal{I}(h_i), \mathcal{R}(h_i)) \qquad (2.4)$$

CSI lives in the frequency domain and can thus be converted to the time domain using IFFT as given in Equation 2.5, where H_t is the Channel Impulse Response (CIR) at time t. Transformation back to the frequency domain is performed using Fast Fourier Transform (FFT), as given in Equation 2.6.

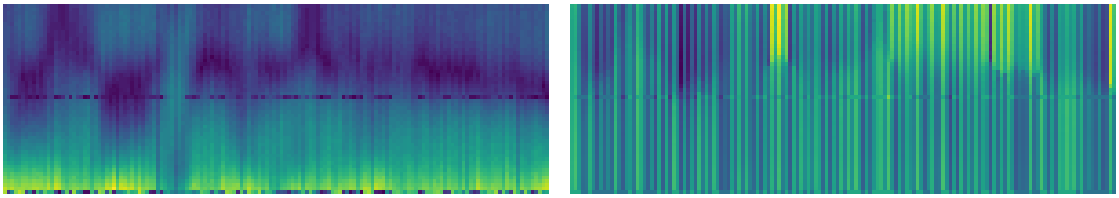
$$H_t = \sum_{m=0}^{N-1} h_m e^{-j2\pi nm/N} \qquad (2.5) \qquad h_m = \sum_{m=0}^{N-1} H_t e^{-j2\pi nm/N} \qquad (2.6)$$

2.1.3 Feature Extraction

As human activities extend over time periods exceeding the time required to transmit a single WiFi packet, WiFi-based PCS approaches consider the CSI of a set of subcarriers S over a certain number of WiFi packets w as input, resulting in a $S \times w$ CSI matrix $\mathcal{H}[t]$ of the form:

$$\mathcal{H}[t] = \begin{bmatrix} h_1[t-w+1] & h_1[t-w+2] & \cdots & h_1[t] \\ h_2[t-w+1] & h_2[t-w+2] & \cdots & h_2[t] \\ \vdots & \vdots & \ddots & \vdots \\ h_S[t-w+1] & h_S[t-w+2] & \cdots & h_S[t] \end{bmatrix}, \qquad (2.7)$$

where t is the time of extraction (or the package index). While the raw CSI $\mathcal{H}[t]$ is able to effectively capture person-centric information, it is rarely used for neural network training without prepossessing. The reason for this is that, as of August 2025, deep learning frameworks such as PyTorch or TensorFlow do not support complex-valued inputs [69]. This limitation can be circumvented by separating the real and imaginary parts of the CSI into two real-valued channels, which can then be processed. Moreover,



(a) CSI amplitude time series (b) CSI phase time series

Figure 2.2: Examples of (a) CSI amplitude $\mathcal{A}[t]$ and (b) phase matrices $\mathcal{P}[t]$ captured in a PCS scenario and visualized as spectrograms. The spectrograms show the amplitude and phase of 52 L-LTF subcarriers (802.11n 20MHz HT mode) over a time interval of approximately 1.5 seconds (150 WiFi packets). The visible temporal variations in both quantities are induced by a person walking between the transmitter and receiver. [50]

seminal approaches leveraging Complex-Valued Neural Network (CVNN) [70] for direct processing of CSI exist [46]. However, considering the added architecture complexity and the associated computational overhead, their superiority over the real-valued neural networks in WiFi-based PCS applications remain to be demonstrated. For now, real-valued features remain the de facto standard in WiFi-based PCS, as they are directly compatible with conventional deep learning frameworks. Amplitude and phase are the most fundamental and widely used representations among CSI features, as they are easy to extract without pre-processing and effectively capture person-centric information, thereby enabling a broad range of applications [45].

Amplitude The amplitude is the absolute value of the CSI and can be extracted from $\mathcal{H}[t]$, as described in Equation 2.3, resulting in the amplitude matrix $\mathcal{A}[t]$:

$$\mathcal{A}[t] = \begin{bmatrix} A_1[t-w+1] & A_1[t-w+2] & \cdots & A_1[t] \\ A_2[t-w+1] & A_2[t-w+2] & \cdots & A_2[t] \\ \vdots & \vdots & \ddots & \vdots \\ A_S[t-w+1] & A_S[t-w+2] & \cdots & A_S[t] \end{bmatrix} \tag{2.8}$$

A visual representation of $\mathcal{A}[t]$, illustrating human movement-induced amplitude variations, is shown in Figure 2.2a. CSI amplitude is the most used feature in WiFi-based PCS literature as it can be efficiently extracted and is more robust than CSI phase [71, 45].

Phase The phase (or angle) of the CSI can be extracted using Equation 2.4, resulting in the phase matrix $\mathcal{P}[t]$. A visual representation of human movement-induced phase shifts is provided in Figure 2.2b.

$$\mathcal{P}[t] = \begin{bmatrix} \phi_1[t-w+1] & \phi_1[t-w+2] & \cdots & \phi_1[t] \\ \phi_2[t-w+1] & \phi_2[t-w+2] & \cdots & \phi_2[t] \\ \vdots & \vdots & \ddots & \vdots \\ \phi_S[t-w+1] & \phi_S[t-w+2] & \cdots & \phi_S[t] \end{bmatrix}, \tag{2.9}$$

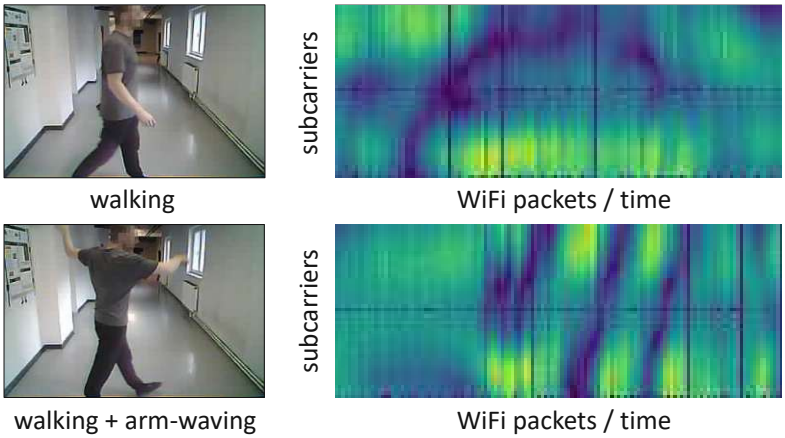


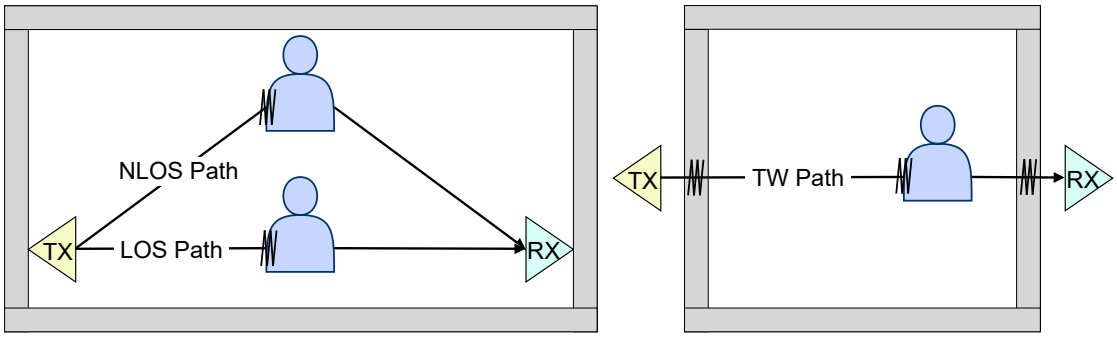
Figure 2.3: CSI amplitude spectrograms of a person walking and a person walking while simultaneously waving their arms, showing distinctive patterns caused by these activities. [48] †

While the phase can capture useful information, it often performs worse than amplitude due to noise factors, such as Carrier Frequency Offset (CFO) and Sampling Frequency Offset (SFO), which can be particularly difficult to mitigate in single-link systems [45].

Derivative Features In addition to amplitude and phase, alternative features are proposed to capture temporal and spectral dynamics in CSI data. One such feature is the first-order difference, which emphasizes short-term variations by computing the temporal derivative of the amplitude or phase across consecutive packets [39, 72]. Another feature is the Doppler Frequency Shift (DFS), which quantifies frequency changes induced by human motion and is particularly informative for assessing movement speed and direction [46]. Spectral representations are also explored. The Power Spectral Density (PSD) characterizes the distribution of signal power across frequencies, enabling activity recognition based on signal energy patterns [73]. Similarly, the Magnitude Spectrum (MS) is obtained by applying an FFT to the CSI time series and reflects the strength of each frequency component [55]. Lastly, the Body-Coordinate Velocity Profile (BVP) is proposed as a cross-domain robust feature [74]. It estimates the subject's velocity in body coordinates based on multi-link CSI measurements. While theoretically environment-independent, BVP extraction requires a multi-receiver setup and is shown to underperform DFS in practice under domain shifts [75].

2.2 PCS Working Principle

WiFi-based PCS leverages the sensitivity of WiFi signals to environmental changes, particularly those caused by human presence and motion [76]. When a person moves through the environment between a WiFi transmitter and receiver, as illustrated in Figure 2.4a, their body alters the propagation of wireless signals by introducing attenuation, scattering, and phase shifts across the multipath components. These interactions cause characteristic variations in the CSI, particularly in the amplitude and phase of the received subcarrier signals. As a result, the presence, location, or activity of a person



(a) LOS and NLOS scenarios (b) TW scenario

Figure 2.4: Overview of PCS sensing scenarios: line-of-sight (LOS), non-line-of-sight (NLOS), and through-wall (TW). TX and RX indicate transmitter and receiver locations, respectively.

manifests as distinct spatiotemporal patterns in the CSI [71, 62]. For instance, walking induces rhythmic fluctuations in amplitude and phase, while more complex activities, such as arm-waving, generate higher-frequency perturbations [48]. These patterns can be captured over time as spectrograms and used to differentiate between activities. Figure 2.3 illustrates this principle by comparing the CSI amplitude patterns of two activities: *walking* and *walking with arm-waving*. The visual distinction between these patterns underscores the discriminative power of CSI features. By learning to associate such CSI patterns with corresponding underlying physical behaviors, machine learning models can be trained to perform PCS tasks such as presence detection, activity recognition, pose-estimation or localization. The ability to infer human behavior without relying on visual information forms the foundation of WiFi-based PCS.

2.3 Sensing Scenarios

WiFi-based PCS relies on the analysis of signal perturbations induced by human presence and activity within the propagation path of a WiFi link. Depending on how the transmitted signal interacts with the environment and the human body, different sensing scenarios arise, namely line-of-sight (LOS), non-line-of-sight (NLOS), and through-wall (TW), as illustrated in Figure 2.4. These scenarios are defined by the characteristics of the dominant signal propagation paths between the transmitter and receiver.

In LOS scenarios, the direct propagation path between transmitter and receiver remains unobstructed by physical structures, but is intersected by the human body. This allows the signal to interact directly with the person, making body-induced attenuation and phase shifts the dominant source of variation in the received CSI. As a result, LOS scenarios yield highly characteristic signal patterns.

In contrast, NLOS scenarios occur when the direct path is obstructed by non-structural elements such as furniture or partial occlusions by room layout. In these cases, the WiFi signal reaches the receiver primarily through secondary reflections and diffraction paths

involving the walls, floor, ceiling, or surrounding objects. While the human body still influences the multipath components, the signal is now a mixture of person-centric and environmental reflections, complicating the extraction of person-centric features.

TW scenarios represent the most challenging condition, where one or more solid structural walls entirely block the direct path between transmitter and receiver. The signal must penetrate the wall material and undergo multiple scattering and attenuation processes before reaching the receiver. These effects drastically reduce signal strength and alter propagation characteristics, making it significantly harder to isolate human-induced variations. Despite these difficulties, TW scenarios are of particular interest, as they enable remote sensing of human behavior across room boundaries, offering an important advantage for unobtrusive, scalable indoor monitoring over optical modalities.

These scenario definitions serve as a framework for experimental design, dataset development, and evaluation of model robustness under varying propagation conditions.

Summary

RSSI provides only a coarse representation of signal strength, lacking the temporal and spectral resolution needed to capture the fine-grained variations introduced by human motion. CSI, in contrast, offers a frequency-selective view of the wireless channel that reflects multipath effects induced by human activity. Its higher information density enables learning-based models to extract subtle spatiotemporal patterns, supporting a broad range of sensing tasks. Consequently, all methods proposed herein operate on CSI, which serves as the principal signal representation throughout.

To enable learning on CSI data, real-valued features are extracted. Amplitude and phase are fundamental, with amplitude being directly derived from raw CSI without requiring preprocessing or calibration, and exhibiting lower susceptibility to noise than phase. Its simplicity and compatibility with standard deep learning frameworks make it particularly suitable for real-time PCS on embedded devices. Derivative features such as temporal differences, Doppler shifts, and spectral transforms emphasize motion-specific signatures but introduce additional computational overhead. This trade-off between efficiency and expressiveness informs the architectural considerations explored in **RQ II**.

CSI characteristics vary significantly across sensing scenarios. LOS, NLOS, and TW scenarios exhibit distinct propagation behaviors, with TW scenarios experiencing stronger attenuation and scattering due to structural obstructions. Enabling long-range PCS under such conditions requires system designs tailored to these conditions, as addressed in **RQ I**. Moreover, the substantial signal differences across scenarios can hinder model generalization, motivating investigations into cross-scenario generalization **RQ III**.

Human motion induces structured variations in CSI over time, visible in the amplitude and phase across subcarriers. These patterns encode information that learning-based models can associate with presence, activity, or location. However, their abstract and non-visual nature limits human interpretability, motivating methods that translate CSI into semantically meaningful or visual representations, as explored in **RQ IV**.

CHAPTER 3

Related Work

This chapter surveys prior work on WiFi-based PCS with a focus on approaches that utilize COTS WiFi devices, specifically commercially available consumer equipment rather than custom research hardware or software-defined radios. It provides a structured overview of the methodological foundations, particularly those relevant to the contributions presented in Chapter 6. By limiting the scope to device-free indoor sensing with COTS WiFi devices, the survey reflects the practical constraints and application goals of this thesis.

The survey is organized both chronologically and thematically. It begins by outlining early feasibility studies prior to 2015, which rely on RSSI and traditional or classical machine learning methods to demonstrate the potential of WiFi-based PCS. The subsequent focus lies on approaches from 2015 onward, exploiting CSI and deep learning methods to extract fine-grained information from wireless signals. This recent body of work is grouped into three main categories. The first addresses deep learning architectures tailored to CSI, exploring how network design choices affect performance and computational efficiency, an aspect closely tied to the processing of CSI on embedded systems (**RQ II**). The second covers strategies for achieving robustness under domain shifts, such as environmental variation, and user diversity. These works inform the development of generalizable models that maintain performance across deployment settings (**RQ III**). The third category encompasses WiFi-based imaging methods, including pose estimation and WiFi-to-image translation, which aim to derive dense, interpretable representations from CSI data (**RQ IV**). Together, these three categories reflect core methodological challenges examined, such as the efficient processing of CSI, cross-domain generalization, and interpretability

3.1 Early Work

Early work on WiFi-based PCS initially leverages RSSI measurements for indoor localization and motion detection, laying the foundation for more refined CSI-based approaches that later extend the application scope to HAR [62, 77]. Table 3.1 summarizes these seminal contributions.

Work	Year	Description
Bahl and Padmanabhan [78]	2000	First demonstration of RSSI-based localization.
Halperin et al. [79]	2011	Enabling of CSI-capturing on COTS WiFi devices.
Koshba et al. [80]	2012	First demonstration of RSSI-based motion detection.
Wu et al. [81]	2012	First demonstration of CSI-based localization.
Xiao et al. [82]	2012	First demonstration of CSI-based motion detection.
Sigg et al. [83]	2013	First demonstration of RSSI-based HAR.
Wang et al. [84]	2014	First demonstration of CSI-based HAR.

Table 3.1: Overview of early seminal works on WiFi-based PCS.

RSSI-based Sensing The *RADAR* system introduced by Bahl and Padmanabhan [78] is a cornerstone in WiFi-based PCS, demonstrating the first use of RSSI measurements for indoor localization using COTS WiFi devices. By combining empirical signal strength fingerprints with radio propagation models, RADAR enables real-time user tracking with a median localization error of 2–3 meters, thereby establishing the feasibility of location-aware applications over existing WiFi networks. Subsequent work during the first decade, exemplified by the *HORUS* system [85, 86], focuses on refining RSSI-based localization, achieving room-level accuracy in controlled environments. As a precursor to HAR, Kosba et al. [80] demonstrate RSSI-based motion detection with their *RASID* system, which employs statistical anomaly detection on RSSI measurements to identify human motion without requiring individuals to carry devices. Building on these approaches, Sigg et al. [83] explore RSSI fluctuations for HAR, extracting multiple statistical features to classify activities such as lying, standing, walking, and crawling.

CSI Capture on COTS WiFi Devices A major turning point occurs in 2011 with the release of the *Linux 802.11n CSI Tool* [79], which enables the capture of CSI from COTS WiFi devices (specifically, the *Intel WiFi Link 5300* network interface card (NIC)). This breakthrough catalyzes the shift from coarse RSSI-based techniques to more precise CSI-based approaches [62, 77], and despite the discontinuation of the *Intel WiFi Link 5300*, the tool remains widely used in WiFi-based PCS research [87].

CSI-based Sensing Leveraging the *Linux 802.11n CSI Tool*, Wu et al. [81] demonstrate the first CSI-based localization system, *FILA*, achieving sub-meter accuracy. Around the same time, Sen et al. [88] introduce *PinLoc*, a CSI fingerprint-based indoor localization system. However, both systems require users to carry WiFi-enabled target devices. The evolution continues with Xiao et al. [82], who present *FIMD*, the first device-free system to detect fine-grained human motion by capturing subtle variations in CSI data. This advancement marks a significant step toward more practical, device-free sensing. Finally, Wang et al. [84] introduce *E-eyes*, the first CSI-based system for HAR in a home environment. In an offline phase, *E-eyes* builds a CSI amplitude location-activity profile database using existing WiFi devices. At test time, it recognizes activities by matching measured profiles to those in the reference database. To overcome generalization issues inherent in domain-dependent CSI profiles, Wang et al. [89] later propose *CARM*,

a CSI-based HAR system that employs PCA-based denoising and a Discrete Wavelet Transform for velocity feature extraction, achieving higher robustness in cross-domain scenarios.

Collectively, these seminal works lay the foundation for WiFi-based PCS, marking a transition from early RSSI-based localization and motion detection to sophisticated CSI-based systems capable of precise localization and fine-grained activity recognition.

3.2 State of the Art

Modern WiFi-based PCS research builds on early feasibility studies by leveraging CSI and deep learning to extract fine-grained spatiotemporal information from wireless signals. The following works, all based on COTS WiFi hardware, address key methodological challenges related to the efficient processing of CSI, cross-domain generalization, and the derivation of interpretable representations.

3.2.1 Deep Learning Architectures for WiFi-based PCS

Early WiFi-based PCS relies on traditional signal processing and statistical machine learning, but recent advancements shift toward deep learning, enabling effective feature extraction directly from CSI [77, 90]. Deep learning architectures for WiFi-based PCS primarily fall into three categories: CNN-based, Recurrent Neural Network (RNN)-based, and Transformer-based methods, each tailored to leverage the unique spatiotemporal characteristics of CSI data. Table 3.2 summarizes the discussed WiFi-based PCS architectures, highlighting their applications and core contributions.

CNN-based Architectures CNN architectures [99, 13] are widely adopted for WiFi-based PCS due to their capability to effectively learn spatial or frequency-domain patterns from WiFi CSI spectrograms. Zhao et al. [91] introduce *RF-Pose*, the first CNN-based method mapping RF signals to 2D human skeleton keypoints, achieving through-wall pose estimation by training with paired image data. Similarly, Wang et al. [94] present *Person-in-WiFi*, a CNN-based approach for simultaneous person segmentation and 2D pose estimation from CSI using two WiFi routers, matching the performance of vision-based methods. For HAR applications, Ding et al. [55] develop *RF-Net*, employing a dual-stream CNN to extract complementary time- and frequency-domain features, combined with a meta-learning module for rapid domain adaptation in few-shot scenarios. Zhang et al.'s *CrossSense*[92] leverages a CNN-based mixture-of-experts framework to handle large-scale variability across multiple sites and activities, improving cross-domain performance. Additionally, Yang et al. [46] propose *SLNet*, introducing a novel Spectrogram Enhancement Network (SEN) and a specialized complex-valued polarized CNN architecture to improve spectrogram quality and feature extraction robustness, achieving high accuracy in HAR tasks across diverse environments.

RNN-based Architectures RNN architectures [100], particularly Long Short-Term Memory (LSTM) networks [101], are also explored due to their capability to capture

Work	Year	Appl.	Method	Summary
Yousefi et al. [65]	2017	HAR	RNN (LSTM)	Proposes an LSTM-based architecture capturing temporal dependencies in CSI for human activity recognition.
Zhao et al. [91]	2018	L,P	CNN	Introduces RF-Pose, the first CNN system for through-wall 2D human pose estimation using WiFi signals.
Zhang et al. [92]	2018	HAR	CNN	Proposes CrossSense, a CNN-based mixture-of-experts architecture enhancing cross-domain generalization in HAR.
Chen et al. [93]	2019	HAR	RNN (ABLSTM)	Introduces attention-augmented bidirectional LSTM (ABLSTM) to improve discrimination of similar human activities.
Wang et al. [94]	2019	L,P,I	CNN	Proposes Person-in-WiFi, the first CNN-based system to jointly estimate human silhouettes and poses from CSI using COTS WiFi hardware.
Jiang et al. [95]	2020	P	CNN+RNN (LSTM)	Introduces WiPose, a CNN-LSTM system for single-person 3D human pose estimation guided by skeletal constraints.
Ding et al. [55]	2020	HAR	CNN	Proposes RF-Net, a dual-stream CNN combined with meta-learning for CSI-based HAR across domains.
Zhang et al. [74]	2022	HAR	CNN+RNN (GRU)	Proposes Widar3.0, a CNN-GRU framework for robust gesture recognition using multi-link CSI data.
Yang et al. [96]	2023	HAR	Transformer	Introduces WiTransformer, applying a Transformer architecture to improve gesture recognition using CSI data.
Yang et al. [46]	2023	HAR	CVNN	Proposes SLNet, a complex-valued CNN designed specifically for direct feature extraction from raw CSI data.
Luo et al. [97]	2024	HAR	Transformer	Evaluates the performance of ViTs on CSI data, highlighting their capabilities and limitations in HAR tasks.
Yan et al. [98]	2024	P	Transformer	Introduces Person-in-WiFi 3D, a multi-person 3D pose estimation system leveraging Transformer architectures and multi-view CSI.

Table 3.2: Overview of discussed deep learning architectures in WiFi-based PCS. *Abbreviations*: HAR = human activity recognition, L = localization, P = pose estimation, I = imaging.

temporal dependencies in sequential CSI data. Yousefi et al. [65] pioneer the use of LSTMs for HAR, demonstrating improved temporal modeling and recognition performance compared to traditional classifiers. Chen et al. [93] further advance this approach by proposing an attention-augmented bi-directional LSTM (ABLSTM), improving the discrimination of activities with subtle temporal distinctions. Integrating CNN and RNN components, Jiang et al. [95] propose *WiPose*, a hybrid CNN-LSTM framework combining CSI and velocity profiles to perform accurate single-person 3D pose estimation. Similarly, the Widar3.0 system [74] employs a CNN backbone for spatial feature extraction followed by a single-layer Gated Recurrent Unit (GRU) network [102] for temporal modeling, effectively capturing short-term temporal dependencies for gesture recognition tasks.

Transformer-based Architectures Transformer [103] architectures recently demonstrate superior performance due to their ability to capture global dependencies within CSI data using self-attention mechanisms. Li et al.'s [104] two-stream Transformer model, *THAT*, captures both time-over-frequency and frequency-over-time dependencies through a multi-scale convolutional self-attention mechanism, outperforming prior CNN and RNN-

based models for HAR. Following this success, Yang et al. [96] introduce *WiTransformer*, a purely Transformer-based architecture adapted specifically for WiFi-based gesture recognition tasks. Luo et al. [97] evaluate Vision Transformers (ViTs) [105] for HAR using CSI spectrograms but highlight that naive implementations can lead to overfitting due to irrelevant visual features in CSI data [98]. Addressing multi-person 3D pose estimation, Yan et al.[98] present *Person-in-WiFi 3D*, employing a Transformer-based detection head inspired by DETR [106], achieving robust multi-person 3D pose estimation from CSI data with sub-decimeter accuracy.

3.2.2 Cross-Domain Generalization

Cross-domain generalization remains a challenge for WiFi-based PCS, as variations in environments, user populations, and hardware degrade model performance in unseen settings. Recent research addresses this through multiple complementary approaches, including domain-invariant feature extraction, virtual sample generation, transfer and few-shot learning, and big data-driven training strategies [47]. These methods aim either to reduce sensitivity to domain-specific variations, synthesize or augment data representing new conditions, or efficiently adapt existing knowledge to novel scenarios, thus enabling robust deployment across diverse and dynamic real-world domains. Tables 3.3 and 3.4 summarize the discussed works on cross-domain generalization in WiFi-based PCS, highlighting their applications, approach, and core contributions.

Domain-Invariant Feature Extraction Robust cross-domain generalization in WiFi-based PCS relies on extracting features resilient to variations caused by changing environments, hardware, and other domain-specific noise sources. Raw CSI amplitude and phase measurements are inherently sensitive to such shifts, motivating the development of alternative representations. One approach leverages temporal dynamics rather than absolute signal values: first-order difference features from consecutive amplitude [39] or phase measurements [72] emphasize motion-induced changes, effectively reducing static environmental influences. Similarly, transforming CSI signals into the frequency domain via PSD [73] exploits frequency characteristics unique to human motion, filtering out static and slowly varying noise. The BVP, introduced by Widar3.0 [107, 74], leverages physical invariances in human-induced Doppler shifts by aggregating CSI across multiple WiFi links, effectively suppressing multipath interference. Although robust, BVP's reliance on multiple spatially-arranged WiFi links limits practical deployment. Complementary approaches utilize signal decomposition and dimensionality reduction techniques to isolate activity-related information from environmental noise. FFT-based band-pass filtering, for example, isolates the human activity frequency range (0–80 Hz), effectively eliminating domain-specific interference [115, 116].

Dimensionality reduction methods like Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Uniform Manifold Approximation and Projection (UMAP) also show promise [50]. PCA reduces the dimensionality of CSI data while eliminating noisy OFDM subcarriers by focusing on axes of maximal variance [45]. ICA separates multivariate CSI signals into independent source signals, enhancing robustness

Work	Year	Appl.	Approach	Summary
Ding et al. [72]	2018	HAR, MD	DFE	Proposes using CSI phase difference between antennas to reliably capture human motion and suppress static interference.
Zhang et al. [92]	2018	HAR	VSG, TFL, BDA	Introduces a mixture-of-experts CNN architecture and CSI augmentation to enhance cross-site generalization for gesture and gait recognition.
Jiang et al. [71]	2018	HAR	TFL	Applies domain-adversarial training (DAT) to learn domain-invariant CSI features across diverse environments.
Zheng et al. [107]	2019	HAR	DFE, BDA	Introduces Body-coordinate Velocity Profile (BVP), a CSI-based representation invariant to environment, enabling robust gesture recognition.
Bai et al. [108]	2019	HAR	TFL	Proposes WiDrive, employing transfer learning (HMM-GMM) for in-car activity recognition from limited labeled data.
Wang et al. [109]	2019	HAR	BDA	Utilizes multiple spatially distributed WiFi links to robustly monitor respiration across diverse sleeping positions.
Jiang et al. [95]	2020	P	TFL	Integrates human kinematics with CNN-LSTM architecture, predicting realistic 3D poses from WiFi signals.
Zeng et al. [110]	2020	HAR	DFE	Applies ICA for blind source separation, improving CSI robustness for multi-person scenarios.
Sheng et al. [111]	2020	HAR	TFL	Proposes test-time fine-tuning of a CNN-BiLSTM architecture for efficient domain adaptation in HAR.
Hu et al. [112]	2021	HAR	TFL	Introduces DSEN (Deep Similarity Evaluation Networks) for few-shot CSI-based gesture recognition.
Kang et al. [113]	2021	HAR	TFL	Proposes a multi-source domain adversarial network with attention-driven feature disentanglement for robust gesture recognition across environments and subjects.
Zhang et al. [74]	2022	HAR	DFE, BDA	Proposes Widar3.0, leveraging body-coordinate velocity profiles from multiple WiFi links to robustly capture gestures.
Wu et al. [114]	2021	L	BDA	Proposes WiTraj, utilizing multiple WiFi links for reliable trajectory estimation regardless of posture and orientation.

Table 3.3: Overview of discussed works on cross-domain generalization in WiFi-based PCS. *Abbreviations*: DFE = Domain-invariant Feature Extraction, VSG = Virtual Sample Generation, TFL = Transfer/Few-Shot Learning, BDA = Big Data Approaches, HAR = Human Activity Recognition, MD = Motion Detection, L = Localization, P = Pose Estimation.

especially in multi-person scenarios [110]. Non-linear approaches, such as UMAP, maintain local and global data structures, further facilitating the extraction of robust CSI representations for localization and HAR tasks [117].

Complementing signal processing methods, system-level approaches like *UniFi*[118] and *AirFi*[119] leverage adversarial and multi-domain training to directly enforce learning domain-invariant features. These methods encourage stability of feature representations across domains but typically require extensive and diverse training datasets to adequately capture domain variability.

Virtual Sample Generation Virtual sample generation addresses cross-domain generalization by synthesizing diverse, physically plausible CSI patterns, helping models become robust to domain shifts. Simple yet effective approaches include controlled

noise injection and subcarrier-level dropout, mimicking channel noise and hardware variability [120]. More sophisticated techniques involve targeted augmentations; Serbetci et al. [121] introduce CSI augmentations inspired by physical propagation models, such as phase and amplitude shifts, improving cross-environment localization. Remarkably, using augmented samples equivalent to only 10% of original data matches full-dataset accuracy, demonstrating the efficiency of meaningful augmentation [121].

Generative modeling further enriches training data by capturing intricate domain variations. CrossSense [92] employs generative models to synthesize CSI samples representative of unseen environments, substantially improving cross-domain gesture recognition performance. Similarly, Variational Autoencoders (VAEs) [122, 123] and Generative Adversarial Networks (GANs) [124] effectively generate synthetic CSI data reflective of diverse environments.

Additionally, general augmentation methods, including noise injection, subcarrier dropout, and MixUp [125, 126], successfully simulate environmental and hardware variability, further enhancing model robustness [125, 120]. Despite these successes, purely synthetic augmentations might fail to fully replicate complex real-world domain variations, necessitating integration with complementary adaptation approaches for optimal performance.

Transfer and Few-Shot Learning Transfer learning enhances cross-domain generalization by leveraging knowledge from source domains, adapting efficiently to target domains with minimal or no labeled data. Such methods typically employ either parameter transfer, fine-tuning pre-trained models on limited target data, or feature representation transfer, learning shared embeddings across domains. Parameter-transfer methods like WiDrive [108], CNN-RNN, and ANN architectures [111, 92] selectively update model parameters, reducing retraining effort. Feature-representation methods, notably Domain-Adversarial Neural Networks (DANN) [132, 133, 71], enforce domain invariance via adversarial training. Kang et al. [113] further enhance this with a multi-source adversarial framework, aligning features across multiple environments and orientations, while Hao et al. [127] employ correlation alignment (CORAL) loss for unsupervised adaptation.

Few-shot learning approaches achieve even stronger adaptability from limited data by explicitly comparing similarities across domains. RFNet-based methods like Zhao et al.'s KNN-MMD [129] combine metric learning with Maximum Mean Discrepancy (MMD), yielding robust few-shot performance. CrossSense [92] integrates generative modeling with a mixture-of-experts architecture to synthesize CSI samples for unseen environments, supporting effective few-shot gesture recognition. Further contributions include Deep Similarity Evaluation Networks (DSENs) [112], attention-based Matching Networks (MatNets) [134, 135], Siamese Networks [136], and Prototypical Networks [137, 138], all leveraging metric-learning principles for robust generalization from minimal data. Collectively, transfer and few-shot learning advance rapid deployment and adaptation of WiFi-based PCS across diverse domains.

Work	Year	Appl.	Approach	Summary
Li et al. [124]	2021	HAR	VSG	Proposes GAN-based data augmentation (CrossGR) for robust gesture recognition from CSI.
Chen et al. [123]	2022	L	VSG, TFL	Proposes Fidora, employing a VAE for data augmentation and a domain-adaptive classifier for localization.
Hao et al. [127]	2022	HAR	TFL	Proposes Wi-CAL using CORAL loss for unsupervised domain adaptation, aligning CSI feature covariances across domains.
Zhang et al. [128]	2022	HAR	BDA	Introduces HandGest, leveraging multiple WiFi links to derive robust, location- and orientation-independent gesture features.
Shi et al. [116]	2022	HAR	DFE, BDA	Applies AFEE filtering to suppress environmental noise in CSI and used Matching Networks for one-shot HAR across environments.
Zhang et al. [74]	2022	HAR	DFE, BDA	Proposes Widar3.0 utilizing body-coordinate velocity profiles (BVP) for environment-agnostic gesture recognition.
Lee et al. [125]	2023	HAR	VSG	Introduces MixUp augmentation for CSI amplitude, enhancing long-term generalization of HAR models.
Serbetci et al. [121]	2023	L	VSG	Proposes physically motivated CSI augmentations, for improved cross-domain localization performance with minimal real data.
Ali et al. [39]	2023	HAR	DFE	Utilizes first-order CSI amplitude differences to robustly capture vital signs, reducing static interference.
Stahlke et al. [117]	2023	L	DFE	Applies UMAP dimensionality reduction for improved CSI-based indoor localization.
Wang et al. [119]	2024	HAR	DFE, BDA	Proposes AirFi, an adversarial training framework learning environment-invariant CSI features from diverse domains.
Liu et al. [118]	2024	HAR	DFE, BDA	Proposes UniFi, learning domain-invariant features using CSI data aggregated from multiple WiFi receivers.
Zhao et al. [129]	2024	HAR	TFL	Applies metric learning combined with Maximum Mean Discrepancy (MMD) for robust few-shot HAR.
Zhao et al. [130]	2024	HAR	TFL, BDA	Introduces CrossFi, a Siamese network using attention-based similarity for one-shot and zero-shot HAR adaptation.
Zheng et al. [131]	2025	HAR	BDA	Proposes AdaWiFi, using federated feature fusion from multiple WiFi APs for collaborative, environment-agnostic sensing without centralized data sharing.

Table 3.4: Overview of discussed works on cross-domain generalization in WiFi-based PCS. *Abbreviations*: DFE = Domain-invariant Feature Extraction, VSG = Virtual Sample Generation, TFL = Transfer/Few-Shot Learning, BDA = Big Data Approaches, HAR = Human Activity Recognition, L = Localization.

Big Data Approaches Big data approaches enhance cross-domain generalization by training on datasets covering diverse environments, user populations, and device configurations. These methods inherently expose models to greater environmental variability, reducing susceptibility to domain shifts. However, large-scale CSI data collection is labor-intensive, and publicly available datasets remain limited, often leading models to overfit specific conditions [130]. To address this, collaborative multi-link systems are explored: *Widar3.0* [74] leverages multiple WiFi links to construct a body-coordinate velocity profile (BVP), effectively capturing human-centric motion while reducing domain-dependent multipath effects. Similarly, *HandGest* [128] utilizes spatial diversity from multiple WiFi links to derive orientation-invariant gesture features, while

WiSDAR [109] employs spatially separated links to reliably monitor respiration across varying positions. Additionally, *WiTraj* [114] leverages multiple receivers positioned at different viewing angles, ensuring reliable velocity measurements regardless of user posture, orientation, or walking direction.

Recent approaches like *AirFi* [119] and *AFEE-MatNet* [116] leverage large-scale multi-environment training to directly learn domain-invariant representations. Although these big data approaches enhance robustness through extensive data diversity, they introduce practical challenges related to complex hardware configurations, high costs, and substantial deployment and maintenance overhead, limiting broader real-world applicability.

3.2.3 WiFi-based Imaging

WiFi-based imaging is a sensing paradigm that aims to reconstruct visual representations of humans and objects directly from WiFi CSI. Originally motivated by works on human pose estimation, which infer sparse representations (e.g., 2D or 3D keypoints) from CSI, recent works advance toward synthesizing dense visual representations, such as human segmentation masks, RGB images, and depth maps. Leveraging deep learning architectures, including CNNs, Transformers, GANs, and teacher-student models trained via cross-modal supervision, these works progressively narrow the fidelity gap between WiFi and optical imaging, revealing WiFi's surprising potential as a visual privacy-preserving alternative to optical cameras. Table 3.5 summarizes key works in WiFi-based imaging, highlighting their output modalities, methods, and contributions.

Human Pose Estimation WiFi-based imaging originates from works on human pose estimation aiming to infer sparse 2D/3D body keypoints from WiFi CSI. Zhao et al. [91] introduce *RF-Pose*, a pioneering system for TW 2D human skeleton estimation. RF-Pose employs cross-modal supervision from RGB images, demonstrating for the first time accurate WiFi-based pose estimation without human-labeled wireless data. Building on this concept, Wang et al. [94] propose *Person-in-WiFi*, a multi-task CNN architecture that simultaneously predicts human skeletons and binary segmentation masks from CSI. This work notably achieves camera-like accuracy using only COTS WiFi hardware. Subsequent developments continue expanding the fidelity and complexity of WiFi-based pose estimation. Yang et al. [143] introduce *MetaFi*, utilizing a CNN with custom convolutional and residual layers to infer accurate 2D keypoints suitable for avatar animation in virtual environments. The enhanced successor, *MetaFi++* [144], incorporates a Transformer-based backbone, further increasing pose estimation robustness and accuracy, reflecting the broader trend toward advanced vision-inspired architectures. Addressing the challenge of reconstructing plausible 3D human poses, Jiang et al. [95] propose *WiPose*, a hybrid CNN-LSTM architecture embedding biomechanical constraints to ensure realistic and physically consistent 3D skeleton predictions. Advancing this further, Yan et al. [98] introduce *Person-in-WiFi 3D*, a Transformer-based multi-view architecture enabling multi-person 3D pose estimation using spatially distributed WiFi transceivers. This study validates the viability of WiFi-based multi-subject 3D pose

Work	Year	Output Modalities	Method	Key Contributions
Zhao et al. [91]	2018	2D pose	CNN (teacher-student)	Demonstrates TW human pose estimation using WiFi CSI.
Wang et al. [94]	2019	2D pose, binary mask	CNN	Introduces Person-in-WiFi: simultaneous segmentation and 2D pose estimation from commodity WiFi.
Li et al. [139]	2020	semantic mask	cGAN	Proposes WiSIA, for semantic mask generation for multiple humans and objects from WiFi CSI.
Kefayati et al. [140]	2020	RGB	CNN (encoder–decoder)	Proposes WiFi-to-video synthesis (Wi2Vi), introducing novel augmentation for robustness.
Jiang et al. [95]	2020	3D pose	CNN+RNN (LSTM)	Proposes WiPose for 3D human pose estimation using skeletal constraints from CSI.
Kato et al. [141]	2021	RGB	GAN	Proposes CSI2Image, a GAN-based WiFi-to-image synthesis method, introducing an object-detection evaluation metric.
Geng et al. [142]	2022	2D pose (dense)	CNN	Proposes DensePose from WiFi, a dense pose estimation method utilizing WiFi signals, producing detailed UV maps of humans.
Yang et al. [143]	2022	2D pose	CNN	Proposes MetaFi, a CNN-based 2D pose estimation method for avatar animation.
Zhou et al. [144]	2023	2D pose	Transformer	Proposes MetaFi++ for improved WiFi-to-pose estimation utilizing a Transformer.
Chen et al. [145]	2023	binary mask	CNN (encoder–decoder)	Introduces Wi-Seg for binary human segmentation using WiFi CSI.
Yan et al. [98]	2024	3D pose	Transformer	Proposes Person-in-WiFi 3D, a WiFi-based multi-person 3D pose estimation method with Transformer-based fusion.
Wang et al. [146]	2024	3D mesh	CNN+RNN (GRU)	Proposes MultiMesh, a WiFi-based multi-person 3D mesh reconstruction method using CNN-GRU with a SMPL model.
Cao et al. [147]	2025	depth	VAE (teacher-student)	Proposes Wi-Depth, for reliable depth reconstruction from WiFi CSI via multi-task learning, utilizing a VAE-based architecture.

Table 3.5: Overview of discussed works on WiFi-based imaging.

imaging with accuracy comparable to vision and radar modalities. Recently, Wang et al. [146] expand beyond skeletons to detailed human meshes with *MultiMesh*, leveraging a CNN-GRU architecture to predict parameters of a parametric SMPL body model [148] directly from CSI-derived angle-of-arrival images. Pushing pose estimation to even finer granularity, Geng et al. [142] introduce *DensePose from WiFi*, mapping CSI signals to dense UV correspondence maps assigning every pixel on the human body to canonical 3D surface coordinates. By adapting state-of-the-art dense pose estimation methods from computer vision to WiFi signals, this work achieves detailed human pose and shape estimation entirely from WiFi signals, advancing WiFi's ability to perform complex, fine-grained human imaging tasks.

WiFi-to-Image Synthesis Inspired by advancements in pose estimation, WiFi-based imaging methods explore the direct synthesis of dense visual representations, including RGB images, segmentation masks, and depth images, from raw CSI. These approaches frame WiFi sensing as an image-to-image translation task, synthesizing meaningful visual content directly from WiFi signals. Kefayati et al. [140] pioneer this direction with

Wi2Vi, demonstrating the feasibility of synthesizing RGB video frames solely from WiFi CSI through an autoencoder architecture supervised by synchronized camera footage during training. They introduce specialized data augmentation strategies to mitigate real-world CSI variability, opening a new frontier for dynamic WiFi-to-image mapping. Extending this direction, Kato et al. [141] develop *CSI2Image*, a GAN-based framework enhancing the visual realism of WiFi-generated RGB images. This work uniquely employs an object detection model as a semantic evaluation metric, validating the information content and correctness of synthesized images. Parallel efforts explore synthesizing semantic segmentation masks. Li et al. [139] propose *WiSIA*, employing a conditional GAN architecture to simultaneously detect, segment, and identify multiple humans and objects in semantic masks generated purely from CSI data. Similarly, Chen et al. [145] introduce *Wi-Seg*, using a CNN-based encoder-decoder to effectively map CSI signals to accurate binary human segmentation masks, directly extracting person silhouettes without visual data at inference time. Moreover, previous pose estimation works, such as *Person-in-WiFi* [94], also produce binary human masks as intermediate outputs, further underscoring CSI's capacity to support semantic image representations. Extending WiFi imaging into full 3D scene understanding, Cao et al. [147] propose *Wi-Depth*, a novel teacher-student VAE architecture. By decomposing depth imaging into shape, depth, and positional estimation sub-tasks, Wi-Depth produces coherent and accurate depth maps of human subjects from CSI alone, introducing the first reliable depth reconstruction method via WiFi signals.

WiFi-based imaging steadily advances from pose estimation to dense visual reconstruction using increasingly capable deep learning models. Results indicate that semantically meaningful visual information can be inferred from CSI, enabling visual downstream tasks while preserving visual privacy. Although WiFi-to-image synthesis remains a niche area with fewer contributions than pose estimation, recent successes highlight its potential and motivate continued research.

Summary

Recent advances in WiFi-based PCS demonstrate the effectiveness of deep learning architectures tailored to the unique properties of CSI. Convolutional, recurrent, and attention-based architectures each contribute distinct capabilities for processing spatial, temporal, and spatiotemporal information, enabling efficient inference and improved performance on complex tasks, an essential step toward real-time PCS on embedded devices (**RQ II**). At the same time, generalizing across environments, hardware setups, and users remains a central challenge. Existing approaches address this through data augmentation, domain-invariant feature learning, adaptive learning strategies, and curated multi-domain datasets, with recent trends showing promise in test-time adaptation techniques (**RQ III**). Finally, WiFi-based imaging methods increasingly succeed at extracting semantically meaningful and visually interpretable representations from CSI, evolving from pose estimation to dense image synthesis. These developments reflect growing progress toward the interpretability of WiFi signals and their use in PCS applications beyond traditional recognition tasks (**RQ IV**).

CHAPTER 4

Systems

A WiFi system consists of wireless transceivers that transmit and receive signals according to standardized WiFi protocols and are primarily composed of a WiFi chipset, firmware, and antenna(s) [149]. The WiFi chipset handles the modulation, demodulation, and digital processing of signals, while the firmware controls hardware operations and communication protocols [149]. The antenna is responsible for radiating and capturing electromagnetic waves [150].

In the context of WiFi-based PCS, the choice of WiFi system dictates sensing performance and applicability. Different WiFi transceivers operate at distinct frequency bands (e.g., 2.4 GHz, 5 GHz, and recently 6 GHz), influencing signal propagation characteristics in spatially confined environments and sensing resolution. Additionally, the number of supported subcarriers directly affects the granularity of CSI measurements. Transceivers can range from single-antenna setups with low subcarrier count to MIMO configurations with high subcarrier count, enabling fine-grained spatial sensing. The option to use integrated or external antennas further affects the sensing range, coverage, and accuracy. Finally, practical considerations such as the physical size, cost, and ease of deployment often determine which WiFi transceiver suitable for a given PCS application scenario [150].

In light of these considerations, this chapter surveys existing and introduces novel WiFi systems, with a focus on enabling long-range TW PCS using low-cost COTS hardware, thereby addressing the system-level challenges posed in **RQ I** and informing the design constraints relevant to efficient CSI processing explored in **RQ II**.

4.1 Existing Solutions

Although WiFi devices are ubiquitous in everyday environments, only a limited subset of hardware-software combinations supports CSI extraction [42]. For most applications, capturing CSI is irrelevant, as WiFi primarily serves high-throughput wireless data transmission. Consequently, standard WiFi firmware and drivers do not expose CSI to end-users, limiting its accessibility for sensing applications. To overcome this restriction,

CSI Tool	Supported Devices
Linux 802.11n CSI Tool [79]	Intel Wireless Link 5300 NIC
Atheros CSI Tool [151]	Atheros AR9k NICs
PicoScenes CSI Tool [152]	802.11ac/ax NICs
Nexmon CSI Tool [153]	Raspberry Pi 3B+/4B, Google Nexus 5/6P, Asus RT-AC86U
AX-CSI Tool [154]	802.11ax routers
Espressif CSI Tool[1], ESP32 CSI Tool [155], Wi-ESP CSI Tool [156]	Espressif ESP32 (all variants)

Table 4.1: Overview of existing CSI-capturing tools and supported COTS WiFi devices.

researchers propose custom tools involving modifications to firmware and device drivers, enabling CSI extraction from selected COTS WiFi devices such as network interface cards (NICs), routers, single-board computers, smartphones, and microcontrollers. These tools vary in their supported hardware and WiFi standards, offering researchers and practitioners a range of options tailored to specific needs. An overview of existing solutions is summarized in Table 4.1.

Linux CSI Tool Introduced by Halperin et al. in 2011, the *Linux 802.11n CSI Tool* [79] lays the foundation for modern CSI-based sensing by enabling CSI extraction on COTS WiFi devices. It specifically supports the now-legendary *Intel Wireless Link 5300* NIC, which remains one of the most widely used WiFi systems in research. This tool allows for the collection of CSI data across 30 subcarriers per antenna pair. The *Intel Wireless Link 5300* supports up to three transceiving antennas (3×3 MIMO), resulting in a total of 90 subcarriers captured per device.

Atheros CSI Tool Recognizing the need for broader hardware compatibility, Xie et al. develop the *Atheros CSI Tool* [151], extending CSI extraction capabilities to Atheros AR9k series NICs, including models such as AR9580, AR9590, AR9344, and QCA9558. The tool supports capturing up to 56 (52 usable) subcarriers at a 20 MHz bandwidth and 114 (108 usable) subcarriers at a 40 MHz bandwidth, providing improved resolution over its predecessors.

PicoScenes CSI Tool Developed by Jiang et al., the *PicoScenes CSI Tool* [152] represents a significant advancement in CSI extraction, supporting modern WiFi NICs up to the IEEE 802.11ax standard, such as the *Intel WiFi 6E AX210*, *Intel WiFi 6 AX200*, and *Qualcomm Atheros AR9300*. Notably, it maintains backward compatibility with older NICs like the *Intel Wireless Link 5300*.

Nexmon CSI Tool Gringoli et al. introduce the *Nexmon CSI Tool* [153], enabling CSI extraction on modern Broadcom and Cypress WiFi chips. This tool supports up to four transceiving antennas (4×4 MIMO), accommodating bandwidths up to 80 MHz in both the 2.4 GHz and 5 GHz bands. It supports a variety of device classes such as smartphones

[1]Espressif CSI Tool, https://github.com/espressif/esp-csi, accessed: 26.03.2025

(Google Nexus 5 and Nexus 6P), single-board computers (Raspberry Pi 3B+/4B), and routers such as the Asus RT-AC86U. Prior CSI extraction tools mainly target NICs, limiting practical applicability due to dependence on external computing hardware. By enabling standalone CSI capture on mobile and embedded devices, the *Nexmon CSI Tool* expands practical WiFi-based sensing applications and facilitates broader real-world deployments.

AX-CSI Tool In 2021, Gringoli et al. present the *AX-CSI Tool* [154], marking a milestone as the first system capable of extracting CSI from IEEE 802.11ax COTS WiFi devices equipped with the Broadcom 43684 WiFi chipset. This tool supports CSI extraction with bandwidths up to 160 MHz and configurations up to 4×4 MIMO, providing high-resolution channel information. Its introduction opens new avenues for research and development in high-throughput WiFi sensing applications.

ESP32 CSI Tools The ESP32 microcontroller series (ESP32 and ESP32-S/C/P/H) has become a popular platform for compact, low-cost WiFi sensing, motivating the development of multiple external CSI extraction tools supporting IEEE 802.11 b/g/n standards. Initially, Hernandez and Bulut introduce the *ESP32 CSI Tool* in 2020 [155], providing researchers with the first accessible means to collect CSI on ESP32 devices. Around the same time, Atif et al. present the *Wi-ESP CSI Tool* [156], a lightweight framework offering similar capabilities. Building upon these foundational efforts, in 2021, *Espressif Systems*, the manufacturer of the ESP32, integrates native CSI capturing directly into their *Espressif IoT Development Framework* (ESP-IDF) through the official *Espressif CSI Tool*[1]. Given its official support and seamless integration within Espressif's development ecosystem, this tool emerges as the default choice for current ESP32-based WiFi sensing applications. In accordance with the IEEE 802.11n standard, the tool supports extracting 56 (52 usable) and 114 (108 usable) subcarriers in 20 MHz and 40 MHz WiFi channels, respectively.

4.1.1 CSI-Capturing Devices Comparison

NICs, including the *Intel Wireless Link 5300* (Linux CSI Tool), Atheros AR9k (Atheros CSI Tool), and modern 802.11ax NICs (PicoScenes CSI Tool), are widely used in research settings. They support MIMO configurations (e.g., 3×3 MIMO for *Intel Wireless Link 5300*, 4×4 for newer NICs) and allow the use of external antennas. However, they require a host device (e.g., a laptop with a mini PCIe slot), making them bulky and costly, limiting standalone deployment. The PicoScenes CSI Tool offers the broadest compatibility, supporting both legacy and WiFi 6 (802.11ax) NICs, but remains dependent on external computing hardware.

The Nexmon CSI Tool enables CSI capture on smartphones (Google Nexus 5/6P), Raspberry Pi (3B+/4B), and routers (e.g., Asus RT-AC86U). Unlike NICs, smartphones and Raspberry Pi offer standalone CSI capture, but smartphones lack MIMO and external antenna support, while current Raspberry Pi variants have become expensive (>$50). Routers, while supporting up to 4×4 MIMO and external antennas, require a separate host device for data extraction. The AX-CSI Tool (Broadcom 43684) supports 160 MHz-

wide channels and 4×4 MIMO, offering the highest-resolution CSI capture. However, like NICs, it lacks standalone functionality and requires an external host, making it unsuitable for compact, embedded applications.

Compared to these platforms, the ESP32 series stands out for its cost-efficiency, compact form factor, and native support for standalone CSI capture. Although it supports only single-antenna CSI extraction and operates solely in the 2.4 GHz band (IEEE 802.11b/g/n), these constraints are offset by its ability to integrate external antennas and deploy multiple nodes in parallel. The platform's low power requirements and official firmware support make it a highly practical option for embedded WiFi-based PCS systems.

4.1.2 The ESP32 Microcontroller Series

The ESP32 microcontroller series by *Espressif Systems* is prevalent in embedded and IoT applications, owing to its low cost (under USD 10), compact size (25.5 mm × 18 mm), and integrated wireless connectivity [42, 157]. Current versions feature a dual-core Tensilica Xtensa LX7 processor (up to 240 MHz) with 520 KiB SRAM, supporting IEEE 802.11b/g/n WiFi and Bluetooth 5 (Low Energy). Peripherals include GPIOs, ADCs, DACs, UARTs, SPI, and I²C. The series expands into four variants: ESP32-S adds security features and USB OTG; ESP32-C uses a RISC-V core and supports Bluetooth 5.0; ESP32-P offers extended memory and interfaces; and ESP32-H targets ultra-low-power applications.

Native CSI extraction is supported via the official *Espressif CSI Tool*, integrated into the ESP-IDF framework, simplifying development and enabling direct access to CSI without firmware modification. While limited to single-antenna CSI capture and 2.4 GHz operation, the platform supports flexible antenna configurations. For instance, the *ESP32-S3-WROOM-1*[2] module includes a built-in Printed Inverted-F Antenna (PIFA)[3], whereas the *ESP32-S3-WROOM-1U*[2] offers an I-PEX MHF I connector for external antennas (see Figure 4.1), enabling directional or high-gain setups advantageous in long-range or TW sensing scenarios. Its low power consumption and standalone operation make it particularly well-suited for distributed, low-cost WiFi-based PCS at the edge [158].

ESP32-S3-WROOM-1

ESP32-S3-WROOM-1U

Figure 4.1: ESP32-S3-WROOM-1/1U modules with integrated PIFA or I-PEX MHF I antenna connector.

[2]ESP32-S3-WROOM-1/1U, https://www.espressif.com/sites/default/files/documentation/esp32-s3-wroom-1_wroom-1u_datasheet_en.pdf, accessed: 26.03.2025

[3]ESP32 PIFA, https://www.ti.com/lit/an/swra117d/swra117d.pdf, accessed: 26.03.2025

4.2 Proposed Systems

Among available solutions for CSI extraction from COTS WiFi devices, the ESP32 microcontroller series stands out for its unique combination of low cost, compact form factor, and integrated wireless connectivity [42]. Consequently, the WiFi systems presented leverage the ESP32 for PCS. While prior research demonstrates its feasibility for WiFi-based PCS, existing works typically employ the ESP32 in its default configuration, relying on the built-in PIFA [45, 159, 157]. Due to its low gain ($\approx$ 2 dBi) and omnidirectional radiation pattern, the PIFA proves unsuitable for long-range or TW scenarios where high signal strength and directional stability are critical. Although a number of works explore the use of passive reflectors [160] or external antennas (specifically omnidirectional rod antennas) [161, 162, 163] to mitigate this shortcoming, these efforts remain limited to short-range scenarios.

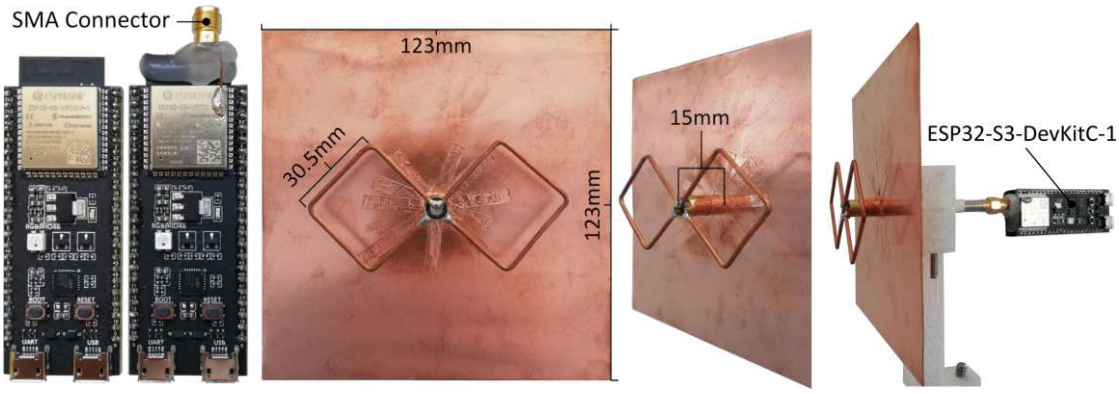
Despite being highly cost-effective, the ESP32's potential can be expanded by extending its sensing range. If a single device could monitor multiple rooms, floors, or even entire buildings, the cost per unit area would be further reduced. The majority of WiFi-based PCS systems use a point-to-point transmitter-receiver configuration with the monitored area positioned between them [164, 161, 165]. Directing energy toward this region, rather than radiating uniformly, can substantially enhance forward signal strength and stability in the presence of obstacles. This makes directional sensing a natural fit for long-range and TW scenarios. However, this approach remains largely unexplored in prior ESP32-based PCS work.

To explore the feasibility of long-range TW PCS with COTS WiFi devices (**RQ I**), two directional sensing methods are investigated: the integration of passive reflectors with the standard PIFA and the utilization of external directional antennas. The development follows an iterative design process, where insights from each system inform the next, leading to progressive improvements in signal quality, sensing performance, and hardware integration. The systems are referred to as system $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}0$, $\mathcal{C}1$, $\mathcal{C}_2$, and $\mathcal{D}$. This naming convention is used throughout this and the following chapters whenever referring to a specific system. The proposed systems are presented in chronological order, detailing their design rationale, implementation, and the lessons learned throughout their development.

4.2.1 System $\mathcal{A}$

System $\mathcal{A}$, first introduced in the seminal work on pairing the ESP32-S3 with external directional antennas for long-range TW HAR [40], marks the initial step in the development of the WiFi systems proposed. It is based on the ESP32-S3-DevKitC-1[4], featuring the ESP32-S3-WROOM-1 module[5] (see Figure 4.2a). Although the built-in PIFA of this module provides basic WiFi connectivity, preliminary short-range tests indicate insufficient gain ($\approx$ 2 dBi) and poor stability, making it an unsuitable choice for a WiFi system targeting challenging long-range and TW sensing scenarios. To overcome

[4]ESP32-S3-DevKitC-1, https://docs.espressif.com/, accessed: 26.03.2025
[5]ESP32-S3-WROOM-1, https://www.espressif.com/, accessed: 26.03.2025



(a) ESP32-S3-DevKitC-1 (b) custom directional biquad antenna (c) System $\mathcal{A}$

Figure 4.2: Overview of the initial system prototype for long-range TW PCS. (a) ESP32-S3-DevKitC-1 development board with built-in PIFA and SMA connector attached, (b) 2.4 GHz directional biquad antenna, and (c) system $\mathcal{A}$, antenna with ESP32-S3-DevKitC-1 attached. [40]

these limitations, system $\mathcal{A}$ employs an external antenna solution. Specifically, an SMA connector is soldered directly to the ESP32-S3-WROOM-1 module's PIFA trace, as shown in Figure 4.2a, allowing easy connection of external antennas. Two antennas operating at 2.4 GHz are evaluated: a commercial 8 dBi omnidirectional dipole antenna and a custom directional biquad antenna (Figure 4.2b), chosen due to its higher gain ($\approx$ 10-12 dBi) and narrower beamwidth ($\approx 70°$)) [166][6].

Antenna *gain*, measured in decibels relative to an isotropic antenna (dBi), quantifies how effectively an antenna directs electromagnetic radiation in a particular direction compared to an idealized antenna radiating equally in all directions (i.e., 0 dB gain). Higher dBi values thus correspond to more concentrated directional energy, improving range, signal quality and reducing interference, while sacrificing some omnidirectional coverage.

The custom directional biquad antenna used in system $\mathcal{A}$ is constructed from readily available materials and consists of two primary elements: a plane reflector and a radiating element. The reflector is made from a 123 mm×123 mm sheet of 0.2 mm-thick copper. Its radiating element employs a vertically polarized biquad design, constructed from 2.5 mm² solid core copper wire, shaped into a biquad loop with an edge length of 30.5 mm, corresponding to a $\frac{1}{4}$ wavelength at 2.448 GHz (IEEE 802.11b central carrier frequency). The radiating element is mounted at a spacing of 15.25 mm ($\frac{1}{8}$ wavelength) in front of the reflector using a central soldered copper tube. A short length of low-loss, 50 Ω impedance coaxial cable connects the radiating element through the copper tube to the ESP32-S3-WROOM-1 via the SMA connector (see Figure 4.2c).

To verify the feasibility of long-range, TW PCS, the two antenna configurations are

[6]Biquad antenna construction, https://martybugs.net/wireless/biquad/, accessed: 26.03.2025

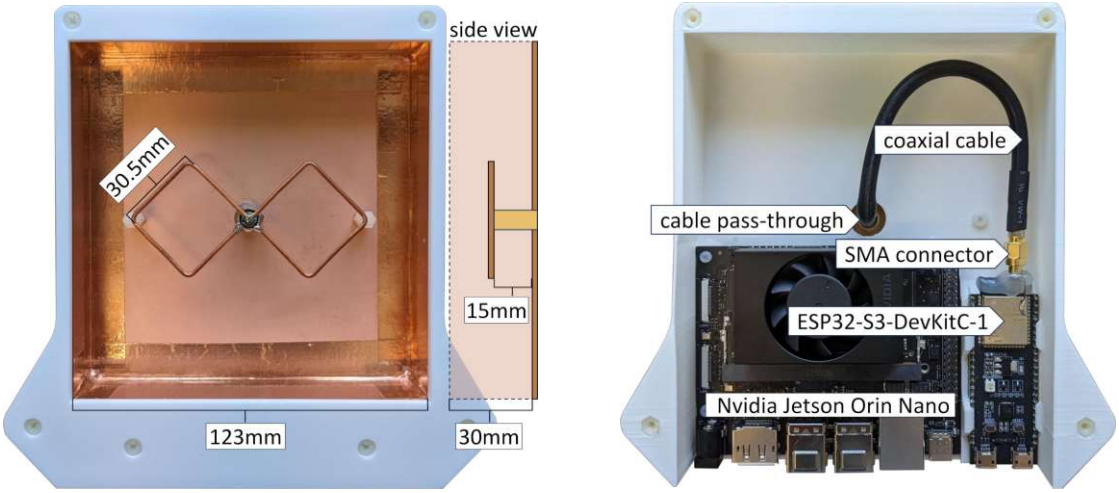
tested in a challenging HAR scenario covering a distance of 18.5 m. The test environment spans five rooms, separated by four 25 cm-thick brick walls. In a point-to-point setup, transmitter and receiver units, each equipped with identical antennas, are placed at opposite ends of the test environment. The TW signal strength is evaluated by measuring the mean RSSI over 1,000 WiFi packets. Results clearly indicate the directional biquad antenna's superior performance, achieving an RSSI of -67 dBm, significantly outperforming the omnidirectional dipole antenna with an RSSI of -79 dBm (i.e., the biquad antenna achieves 15.85× higher signal power). For reference, the built-in PIFA fails entirely to establish a stable connection in this challenging scenario. Based on these results, the directional biquad antenna is selected for system $\mathcal{A}$. Beyond its superior signal strength, this antenna also reduces external noise, a notable challenge for omnidirectional antennas [160]. Its 70° beamwidth facilitates comprehensive room coverage in most PCS scenarios while maintaining constraints on the recording environment. Moreover, it aligns closely with typical camera fields of view, allowing for easy integration and providing a sense of the antenna beam's coverage area.

4.2.2 System $\mathcal{B}$

System $\mathcal{B}$, proposed in follow-up work on long-range TW HAR [52], builds upon the foundation laid by system $\mathcal{A}$. The design of system $\mathcal{B}$ addresses specific shortcomings identified in the initial system, particularly focusing on enhanced robustness, ease of manufacturing, improved antenna performance, and integrated on-device inference capabilities. Like its predecessor, system $\mathcal{B}$ combines the ESP32-S3-DevKitC-1 development board, featuring the ESP32-S3-WROOM-1 module, with a custom 2.4 GHz directional biquad antenna. However, notable improvements are introduced to both antenna design and mechanical robustness (Figure 4.3).

First, to simplify the antenna manufacturing process and improve structural integrity, the original copper-sheet reflector used in system $\mathcal{A}$ is replaced with blank printed circuit board (PCB) material, consisting of a single-sided copper layer (35 µm thickness). This change enhances planarity and ease of handling, reducing deformation risks associated with copper sheets. Additionally, side lips, measuring 30 mm in depth, are introduced to the reflector's edges, while retaining its original 123 mm×123 mm dimensions. Although optional, these side lips reduce side-lobe radiation and external noise from lateral directions, resulting in an approximate 2 dBi gain improvement compared to the original flat reflector design [166]. Furthermore, they facilitate co-planar system configurations by preventing unintended direct LOS communication between transmitter and receiver [160].

The antenna's radiating element retains the biquad geometry from system $\mathcal{A}$: a vertically polarized loop of 2.5 mm² solid-core copper wire with 30.5 mm edge lengths ($\frac{1}{4}$ wavelength). The element is attached to the reflector by a centrally soldered copper tube, ensuring a 15.25 mm ($\frac{1}{8}$ wavelength) reflector-to-element spacing. Connection to the ESP32-S3-WROOM-1 module remains unchanged, using an SMA connector soldered directly to the PIFA trace via a short segment of low-loss coaxial cable.



(a) System $\mathcal{B}$ (frontside) (b) System $\mathcal{B}$ (backside)

Figure 4.3: Overview of system $\mathcal{B}$, showing: (a) the improved biquad antenna geometry with side lips, and (b) the internal electronic components, comprising ESP32-S3-DevKitC-1 development board and Nvidia Jetson Orin Nano developer kit. [52]

To further improve robustness against mechanical deformation, system $\mathcal{B}$ incorporates a custom 3D-printed enclosure, which rigidly secures the antenna and protects its components during handling and transportation. Two additional 15 mm nylon standoffs between the reflector and radiating element ensure consistent geometry, maintaining the $\frac{1}{8}$-wavelength spacing (Figure 4.3a).

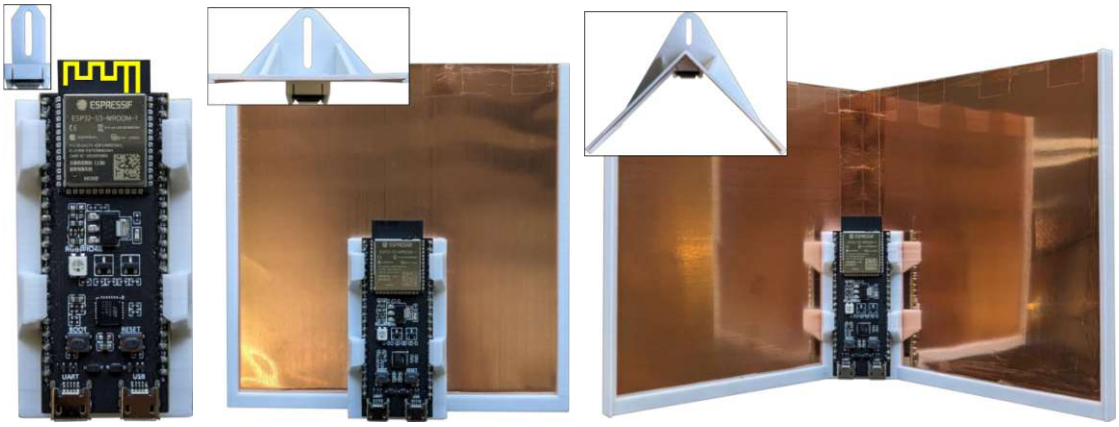
Another advancement of system $\mathcal{B}$ is its expanded computational capabilities. As illustrated in Figure 4.3b, the enclosure provides mounting points not only for the ESP32-S3-DevKitC-1 but also for an *Nvidia Jetson Orin Nano* developer kit[7], a powerful and compact single-board computer enabling real-time on-device inference for WiFi-based PCS **RQ II**. This integration expands the system's autonomy, eliminating the need for external computational resources. For reproducibility, CAD models of all 3D-printed components of system $\mathcal{B}$ are publicly available[8].

4.2.3 Systems $\mathcal{C}_{0-2}$

A notable drawback identified in systems $\mathcal{A}$ and $\mathcal{B}$ is the irreversible modification of the ESP32-S3-WROOM-1's built-in PIFA required to connect external antennas. Scraping, cutting, and soldering an SMA connector directly to the antenna traces permanently disables the original antenna. For applications where preserving the PIFAS's functionality is desirable, reflector-based solutions offer a practical alternative. Reflectors can effectively shape the PIFA's radiation pattern, increasing forward gain and directionality without physically altering the module. Motivated by this concept, two reflector-based systems

[7]Nvidia Jetson Orin Nano Developer Kit, https://nvdam.widen.net/s/zkfqjmtds2/jetson-orin-datasheet-nano-developer-kit-3575392-r2, accessed: 26.03.2025

[8]System $\mathcal{B}$ CAD models, https://zenodo.org/records/15147388, accessed: 26.03.2025



(a) System $\mathcal{C}_0$ (b) System $\mathcal{C}_1$ (c) System $\mathcal{C}_2$

Figure 4.4: Overview of systems $\mathcal{C}_{0-2}$, showing: (a) system $\mathcal{C}_0$, the baseline system relying solely on the ESP32-S3's built-in PIFA, (b) system $\mathcal{C}_1$, PIFA with a plane reflector, and (c) system $\mathcal{C}_2$, PIFA with a 90° corner reflector. [52]

($\mathcal{C}_1$ and $\mathcal{C}_2$) are developed and evaluated alongside a baseline system ($\mathcal{C}_0$) and compared to system $\mathcal{B}$ in a long-range TW HAR scenario [52].

System $\mathcal{C}_0$ (PIFA Baseline) As a baseline, system $\mathcal{C}_0$ employs an unmodified ESP32-S3-DevKitC-1 board, utilizing the built-in omnidirectional PIFA of the ESP32-S3-WROOM-1 module. To facilitate consistent and reproducible measurements, a simple 3D-printed frame is developed, securing the ESP32-S3-WROOM-1 module in an upright position to ensure unrestricted radial emission in the horizontal plane (Figure 4.4a).

System $\mathcal{C}_1$ (PIFA with Plane Reflector) To enhance forward gain and limit backside interference without physical modifications to the PIFA, system $\mathcal{C}_1$ employs a plane reflector (Figure 4.4b). This reflector is constructed from a 123 mm×123 mm copper sheet with a thickness of 0.2 mm, rigidly mounted behind the PIFA using a custom 3D-printed carrier. The reflector is placed 30.5 mm ($\frac{1}{4}$ wavelength) behind the PIFA, ensuring a total phase shift of 360° (180° due to round-trip propagation and 180° from reflection), thereby producing constructive forward interference. This configuration reduces noise from directions behind the antenna and increases forward signal strength without modifying or destroying the PIFA.

System $\mathcal{C}_2$ (PIFA with 90° Corner Reflector) Building upon system $\mathcal{C}_1$, system $\mathcal{C}_2$ incorporates a 90° corner reflector to further focus the antenna's radiation pattern and narrow its beamwidth (Figure 4.4c). The reflector consists of two 123×123 mm copper sheets, each 0.2 mm thick, joined at a 90° angle using conductive copper tape and secured by a custom 3D-printed frame. Similar to system $\mathcal{C}_1$, the spacing between the reflector center and the PIFA is maintained at 30.5 mm ($\frac{1}{4}$ wavelength), creating directional gain enhancement while limiting backside and lateral interference. CAD models for all 3D-printed components used in systems $\mathcal{C}_{0-2}$ are publicly available[9].

[9]Systems $\mathcal{C}_{0-2}$ CAD models, https://zenodo.org/records/15147388, accessed: 26.03.2025

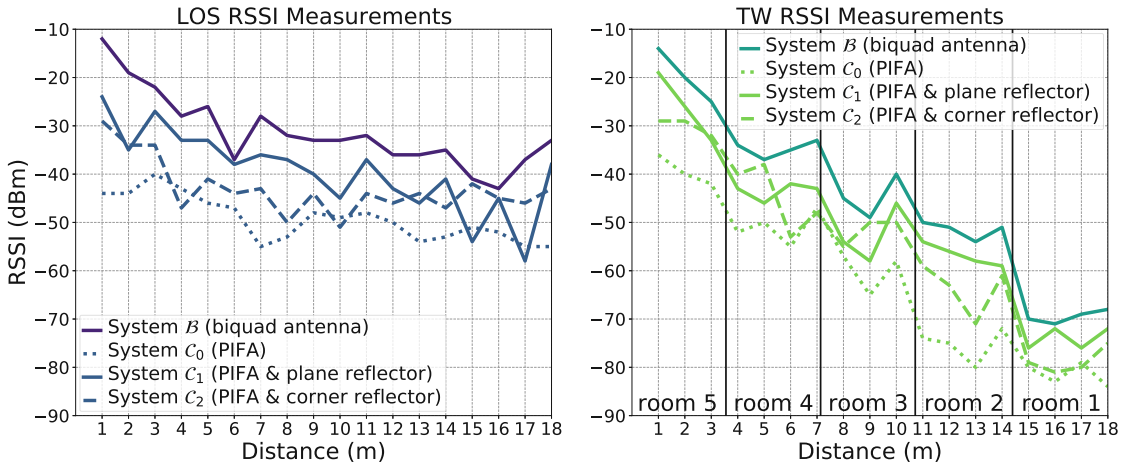


Figure 4.5: Comparison of line-of-sight (LOS) and through-wall (TW) signal strength (RSSI) between systems $\mathcal{B}$, $\mathcal{C}_0$, $\mathcal{C}_1$, and $\mathcal{C}_2$, over a distance of 18 m and five rooms. In the TW scenario, black vertical lines mark the position of walls between rooms. [52] †

Signal Strength Evaluation To evaluate the performance of systems $\mathcal{C}_{0-2}$ and identify viable solutions for long-range TW PCS, RSSI measurements are conducted in both LOS and TW scenarios, spanning distances up to 18 meters across five rooms separated by 25 cm-thick brick walls (Figure 4.5). Each measurement averages the RSSI over 1,000 WiFi packets at 1-meter intervals.

In the LOS scenario, all reflector-based systems outperform the baseline system $\mathcal{C}_0$, clearly demonstrating the directional gain provided by the reflectors. System $\mathcal{C}_1$ (plane reflector) consistently surpasses system $\mathcal{C}_2$ (corner reflector), likely due to the more uniform reflector spacing in the planar geometry, whereas the corner reflector introduces spacing variations that may cause partial destructive interference. Future investigations into parabolic reflectors might resolve this limitation.

In the TW scenario, involving multiple walls, the systems exhibit similar relative performance. System $\mathcal{C}_1$ maintains robust connectivity, with RSSI dropping from -19 dBm (1 m) to -72 dBm (18 m), while the baseline system deteriorates significantly from -36 dBm to -84 dBm, resulting in unstable connectivity and frequent packet loss at long distances. Overall, reflector-based systems achieve a performance improvement in challenging propagation environments without physical modifications to the PIFA, offering a viable alternative to externally connected directional antennas.

However, ultimately, system $\mathcal{B}$, with its dedicated biquad antenna, outperforms all reflector-based configurations in both LOS (-12 dBm at 1 m to -33 dBm at 18 m) and TW (-14 dBm at 1 m to -68 dBm at 18 m) scenarios, reinforcing the benefit of directly employing external directional antennas for robust, long-range WiFi-based PCS. However, systems $\mathcal{C}_1$ and $\mathcal{C}_2$ remain valuable in scenarios demanding preservation of the PIFA's functionality.

Figure 4.6: Overview of system $\mathcal{D}$, showing internal components comprising ESP32-S3-DevKitC-1 module, ALFA Network APA-M25 directional panel antenna, and Nvidia Jetson Orin Nano developer kit. [49]

4.2.4 System $\mathcal{D}$

Building upon insights from previous systems, system $\mathcal{D}$, proposed in [49], aims at enhancing reproducibility, robustness, and practical applicability by exclusively employing COTS components. Earlier designs, such as Systems $\mathcal{A}$ and $\mathcal{B}$, rely on manually modified ESP32-S3-WROOM-1 modules and custom antennas, resulting in limited scalability and reproducibility. System $\mathcal{D}$ resolves these issues by using standardized hardware while retaining the directional gain essential for long-range TW sensing.

At the core of system $\mathcal{D}$ (see Figure 4.6) is the ESP32-S3-DevKitC-1**U**[10], featuring the ESP32-S3-WROOM-1U[11] module. As shown in Figure 4.1, unlike the ESP32-S3-WROOM-1 variant used in prior systems, this module includes an I-PEX MHF I antenna connector, bypassing the built-in PIFA antenna and enabling straightforward attachment of external antennas without manual modifications. System $\mathcal{D}$ integrates a COTS directional antenna, the *ALFA Network APA-M25*[12], a dual-band (2.4 GHz and 5 GHz) panel antenna priced around USD 20. It offers a gain of 8 dBi at 2.4 GHz and a horizontal beamwidth of approximately 66°, similar in performance to the custom biquad antenna in system $\mathcal{B}$ (gain of 10–12 dBi, beamwidth 70°). The APA-M25 antenna provides comparable sensing capabilities in a robust, compact, and mass-produced form factor, greatly enhancing reproducibility and ease of deployment.

Furthermore, like system $\mathcal{B}$, system $\mathcal{D}$ integrates an Nvidia Jetson Orin Nano devleopment

[10] ESP32-S3-DevKitC-1U, https://docs.espressif.com/, accessed: 26.03.2025
[11] ESP32-S3-WROOM-1U Datasheet, https://www.espressif.com/, accessed: 26.03.2025
[12] ALFA APA-M25 antenna, https://www.alfa.com.tw/products/apa-m25, accessed: 26.03.2025

kit within a compact housing (Figure 4.6), enabling real-time, on-device inference. The electronic components, including the ESP32-S3-DevKitC-1U, Jetson Orin Nano, and directional antenna, are securely enclosed within a rigid, compact, and easily replicable 3D-printed housing. This housing is approximately half the size of system $\mathcal{B}$, improving portability and practical deployment. System $\mathcal{D}$'s modular design facilitates versatile antenna configurations. By substituting the ESP32-S3-DevKitC-1U with the ESP32-S3-DevKitC-1, which is a drop-in replacement, one could revert to the built-in PIFA if required. Moreover, alternative external antennas can easily be connected via the I-PEX MHF I connector to accommodate different use cases or experimental scenarios. Due to its dual-band support (2.4 GHz and 5 GHz), the ALFA APA-M25 antenna opens potential avenues for future ESP32 hardware revisions that support advanced WiFi standards beyond IEEE 802.11b/g/n, such as IEEE 802.11ac (WiFi 5), further extending the versatility of system $\mathcal{D}$. The CAD models of all 3D-printed components of system $\mathcal{D}$ are publicly available[13].

To evaluate the practical signal strength of system $\mathcal{D}$, a LOS experiment comparing its RSSI performance against system $\mathcal{B}$ is conducted across a distance of 25 meters. Both systems show similar performance at short ranges (1–10 m). Between 10–15 m, system $\mathcal{B}$ slightly outperforms system $\mathcal{D}$, which aligns with expectations given its higher theoretical antenna gain of approximately 10–12 dBi [166], compared to the 8 dBi gain of system $\mathcal{D}$'s COTS antenna. However, at greater distances (15–25 m), system $\mathcal{D}$ demonstrates superior signal stability, slightly outperforming system $\mathcal{B}$. This result highlights that despite its lower nominal gain, the COTS directional antenna in system $\mathcal{D}$ achieves practically equivalent or even better long-range performance compared to the custom-built biquad antenna of system $\mathcal{B}$. Given its additional advantages, including compactness, flexibility, ease of replication, and exclusive use of COTS components, system $\mathcal{D}$ ultimately offers the most balanced trade-off among all WiFi-based PCS systems introduced, making it highly attractive for practical deployments.

4.2.5 WiFi System Setup and CSI Capturing

All proposed systems are deployed in a point-to-point configuration, consisting of a single transmitter and a single receiver with their antennas oriented toward one another. The monitored environment lies between the two devices, ensuring optimal coverage of the sensing area. Although the hardware is identical on both ends, the devices are configured to perform distinct roles: the transmitter continuously emits WiFi packets at a fixed rate of 100 Hz (corresponding to a 10 ms packet interval), while the receiver captures them and extracts the corresponding CSI. This rate is chosen specifically to support the recognition of macroscopic human activities such as walking, sitting, and standing. Research shows that these activities contain frequency components up to 80 Hz [37, 116], motivating the use of a conservative 100 Hz rate in most WiFi-based HAR studies [45]. Although higher sending rates can improve temporal resolution, as required in tasks like WiFi-based hand gesture recognition, where rates often exceed 1000 Hz [167, 168, 169], the added

[13] System $\mathcal{D}$ CAD models, https://zenodo.org/records/10715595, accessed: 26.03.2025

benefits are typically offset by increased packet loss and computational demands [45] in macroscopic activity recognition. Communication between the nodes is facilitated by Espressif's ESP-NOW[14], a lightweight wireless communication standard that operates independently of existing WiFi infrastructure. CSI extraction is performed at the receiver using a modified version of the official Espressif CSI Tool integrated into the ESP-IDF framework. This standardized setup is used for all experiments conducted.

4.2.6 Limitations

A central goal of **RQ I** is enabling long-range TW PCS, which aligns with the use of the 2.4 GHz band supported by the ESP32 family. Since the signal power loss in free space is proportional to the square of the carrier frequency, the 2.4 GHz band has a range advantage over higher-frequency bands (i.e., 5 GHz and above) [170]. In addition, higher-frequency bands are shown to suffer from greater penetration loss through dense building materials. While the difference is negligible for light materials such as drywall, denser materials like glass ($\approx$ 1.2 dB $\uparrow$), wood ($\approx$ 3.3 dB $\uparrow$), concrete ($\approx$ 3.6 dB $\uparrow$), or brick ($\approx$ 10.2 dB $\uparrow$) introduce significantly more attenuation at 5 GHz [170, 171]. This trend intensifies at 6 GHz and beyond, where signals are easily blocked by walls and obstacles [172, 173, 171], limiting the effective sensing range and practical utility of systems that rely on higher-frequency bands in TW scenarios.

While the 2.4 GHz band offers favorable propagation characteristics, it is more congested than newer bands, increasing susceptibility to interference. It also supports only up to 40 MHz bandwidth, resulting in a lower subcarrier count compared to 80 or 160 MHz configurations available in 5 GHz and 6 GHz systems. Although higher subcarrier density is shown to improve performance in LOS scenarios [174], it also increases computational overhead, which may be undesirable for real-time inference on edge devices (**RQ II**).

In line with **RQ I**, which emphasizes minimal system cost and complexity, the proposed systems adopt a Single Input Single Output (SISO) / single-link configuration. This choice is partly imposed by the ESP32's limitation to a single transmit antenna. While Multiple Input Single Output (MISO) or multi-link setups could enhance spatial diversity and feature robustness [74, 109], they incur increased system cost, complexity, synchronization requirements, and computational overhead. Furthermore, the use of phase-based features is limited in SISO systems due to challenges in correcting CFO and SFO. Although multi-link systems can address this via complex conjugate multiplication [175], prior work indicates that effectively eliminating phase noise remains difficult and that the utility of phase features is limited in practice, with the result that most works rely on amplitude features for WiFi-based PCS [45].

The proposed systems are explicitly designed for long-range, TW PCS, with an emphasis on minimal system cost and complexity. These design constraints introduce potential limitations, such as lower spectral resolution and reduced spatial diversity, which may restrict the use of advanced features and impact performance in short-range or LOS scenarios. Nevertheless, these trade-offs are well-aligned with the objectives of **RQ I**.

[14] ESP-NOW, https://docs.espressif.com, accessed: 26.03.2025

Summary

Directional sensing with low-cost COTS WiFi hardware shows potential for enabling long-range TW PCS by improving signal strength and stability under challenging propagation conditions while minimizing system complexity, thereby addressing the constraints posed in **RQ I**. The proposed systems, based on the ESP32 platform, leverage either directional antennas or passive reflectors to increase forward signal gain. Reflector-based systems preserve the internal PIFA and offer a low-cost, non-invasive solution, whereas systems using external antennas achieve consistently higher signal quality. Evaluation of signal strength confirms the effectiveness of the proposed directional sensing approach, with external antenna configurations demonstrating superior performance. The proposed systems are used to collect the datasets presented in Chapter 5 and define the hardware constraints for the real-time, embedded CSI processing approaches explored in Chapter 6, thereby informing the design objectives of **RQ II**.

CHAPTER 5

Datasets

This chapter provides an overview of influential publicly available WiFi PCS datasets, highlighting their characteristics, applications, and contributions to the field. Following this review, the chapter introduces five new datasets, designed specifically for WiFi-based presence detection, HAR, localization, and imaging tasks (**RQ IV**). These novel datasets address challenging scenarios such as long-range and TW sensing (**RQ I**), in addition to standard LOS conditions. Furthermore, the datasets enable the evaluation of cross-domain generalization capabilities (**RQ III**). Each dataset is described in detail, covering data characteristics, recording environments, and employed WiFi systems.

5.1 Public Datasets

Over the past decade, indoor WiFi-based PCS research has evolved from early single-environment CSI datasets to extensive, diverse datasets that support fine-grained HAR, gesture recognition, pose estimation, and localization. Early datasets such as StanWiFi [176] and SignFi [177] establish initial benchmarks for CSI-based HAR and gesture recognition using Intel Wireless Link 5300 NICs. Later, WiAR [178] expands application scopes by incorporating joint activity recognition and indoor localization across multiple environments. The Widar3.0 [74] dataset marks a milestone through its multi-receiver setup and scale for enabling cross-domain gesture recognition. Person-in-WiFi [94] further advances the field with fine-grained 2D pose estimation and segmentation, while Schäfer et al.[180] demonstrate that COTS 802.11ac hardware with Nexmon firmware yields high-resolution CSI for fine-grained HAR. Baha et al.[179] introduce a dataset explicitly exploring LOS and NLOS CSI variations, featuring activities from multiple subjects to facilitate robust HAR. RF-Net [55] further diversifies research directions by providing a dataset tailored for meta-learning applications, enabling effective one-shot HAR across new, unseen indoor environments. In 2022, OPERAnet [181] combines CSI with additional RF and vision-based modalities to evaluate HAR and localization techniques comprehensively. More recently, NTU-Fi [43] and WiFi-80 MHz [182] improve sensing fidelity by increasing subcarrier resolution and integrating data from diverse

Dataset	Year	Description
StanWiFi [176]	2017	HAR dataset featuring 6 activities captured from 6 subjects in a controlled lab environment using the Intel Wireless Link 5300 via the Linux 802.11n CSI Tool at 5 GHz.
SignFi [177]	2018	Dataset for sign language recognition with 276 sign gestures and video supervision, acquired using the Intel Wireless Link 5300 with the Linux 802.11n CSI Tool.
WiAR [178]	2019	HAR dataset featuring 16 activities by 10 subjects across 3 environments, captured with the Intel Wireless Link 5300 via the Linux 802.11n CSI Tool on both 2.4 GHz and 5 GHz channels at approximately 100 Hz.
Person-in-WiFi [94]	2019	Dataset for 2D pose estimation and human segmentation, collected over 16 indoor scenes using the Intel Wireless Link 5300 with the Linux 802.11n CSI Tool and synchronized RGB annotations.
Baha et al. [179]	2020	HAR dataset comprising 5 activities performed by 30 subjects across LOS and NLOS indoor environments, captured using Intel 5300 NICs at 2.4 GHz with a sampling rate of 320 Hz.
RF-NET [55]	2020	Dataset for one-shot HAR, featuring 6 activities collected across multiple indoor environments to support meta-learning approaches.
Schäfer et al. [180]	2021	HAR dataset acquired using Asus RT-AC86U routers running Nexmon firmware on 802.11ac 80 MHz channels; captures 256 subcarriers to demonstrate sensing on routers, smartphones, and IoT devices.
Widar3.0 [74]	2022	Dataset for gesture recognition and localization, featuring 22 gestures from 17 users across 8 locations and 75 domains; captured at high sampling rates (1000 Hz) using the Intel Wireless Link 5300 via the Linux 802.11n CSI Tool.
OPERAnet [181]	2022	Multimodal dataset for HAR and localization combining CSI with additional RF modalities (Passive WiFi Radar via SDR and UWB) and Kinect vision; includes 8 hours of annotated measurements from 6 subjects performing 6 daily activities across 2 rooms.
NTU-Fi [43]	2023	High-resolution HAR and gait dataset using Atheros AR9k NICs with the Atheros CSI Tool; offers 114 subcarriers per antenna over a 40 MHz channel, collected from 20 subjects in a single lab at 1000 Hz.
WiFi-80 MHz [182]	2023	Multi-environment dataset for HAR, person identification, and people counting, captured from Netgear APs with Nexmon on 802.11ac (80 MHz) channels featuring 256 subcarriers (242 usable) from 10 subjects across 7 indoor settings.
MM-Fi [183]	2023	Multimodal dataset for HAR and gesture recognition using the Intel Wireless Link 5300 with the Linux 802.11n CSI Tool; notable for increased subject diversity and environmental variability, with additional modalities such as RGB, depth, LiDAR, and radar.
Person-in-WiFi 3D [98]	2024	Multi-person 3D pose dataset using the Intel Wireless Link 5300 via the Linux 802.11n CSI Tool; captures 3D keypoints for up to 3 persons at 300 Hz over 30 subcarriers, with Kinect-synchronized ground truth.
WiMANS [44]	2024	Multimodal, multi-person dataset for HAR, localization, and pose estimation, collected using a TP-Link N750 with the Atheros CSI Tool.

Table 5.1: Overview of large-scale public WiFi-based PCS datasets.

environments. The latest advances include Person-in-WiFi 3D [98], which achieves multi-person 3D pose estimation using multi-static setups, as well as MM-Fi [183] and WiMANS [44], which further enhance dataset scale, subject diversity, and multimodal annotations. Together, these datasets demonstrate clear improvements in scale, resolution, and application breadth, progressing from basic HAR in controlled settings to robust, multi-person sensing across diverse indoor domains. A summary of the datasets discussed in this section is provided in Table 5.1. Furthermore, an overview of additional small- to mid-sized WiFi datasets can be found in the comprehensive survey by Wang et al. [184].

Dataset	Year	Application	Scenario	Labels	#Packets
TOA [40]	2023	HAR, PD	LOS, TW	CSI, activity, room	316,862
Wallhack1.8k [48]	2024	HAR, G	LOS, TW	CSI, activity	500,093
HALOC [49]	2024	L	LOS	CSI, 3D traj.	118,679
WiFiCam [51]	2024	I	TW	CSI, RGB image	57,103
3DO [50]	2025	HAR, L, G	TW	CSI, activity, 3D traj.	1,292,727

Table 5.2: Overview of proposed datasets. Applications and scenarios are abbreviated as human activity recognition (HAR), presence detection (PD), localization (L), imaging (I), cross-domain generalization (G), line-of-sight (LOS), and through-wall (TW).

5.2 Proposed Datasets

Existing publicly available WiFi-based PCS datasets predominantly focus on short-range, LOS scenarios recorded using NIC-based WiFi systems (Intel or Atheros), thereby neglecting challenging yet economically beneficial long-range and TW scenarios [43, 44]. Recently, the ESP32 microcontroller has emerged as a promising and cost-effective alternative for WiFi-based PCS, but public datasets based on the ESP32 remain severely underrepresented, with only a handful of peer-reviewed datasets becoming available [185, 186, 187]. Furthermore, existing datasets either omit TW scenarios completely or provide limited short-range examples [179, 185], restricting the assessment of WiFi sensing under realistic long-range and TW conditions.

To address these critical research gaps five datasets specifically designed to validate the ESP32's feasibility for long-range and TW sensing scenarios are collected, advancing the field beyond current benchmarks. An overview of the proposed datasets is provide in Table 5.2. The *Through-Wall Office Activities (TOA)* and *Wallhack1.8k* datasets allow controlled direct comparisons of LOS versus TW performance, while Wallhack1.8k further enables evaluation of cross-system generalization between different ESP32-based WiFi systems. The *3-Days Office (3DO)* dataset explicitly isolates static, dynamic, and temporal environmental variations, facilitating focused research into cross-domain generalization. Furthermore, the proposed datasets offer unique label combinations previously unavailable in public WiFi datasets, such as TW CSI paired with synchronized RGB video (*WiFiCam* dataset), TW CSI combined with 3D trajectories and activity labels (3DO), and long-range LOS CSI with precise 3D trajectory labels (*HAllway LOCalization (HALOC)*). Collectively, these datasets enrich the existing WiFi dataset landscape, enabling targeted exploration of economically advantageous long-range and TW WiFi-based PCS scenarios.

5.2.1 Through-Wall Office Activities Dataset

The Through-Wall Office Activities (TOA) dataset, introduced in [40], is created to evaluate the feasibility of long-range WiFi-based PCS (presence detection and HAR) using ESP32 microcontrollers. Specifically, it demonstrates that a modified ESP32-based

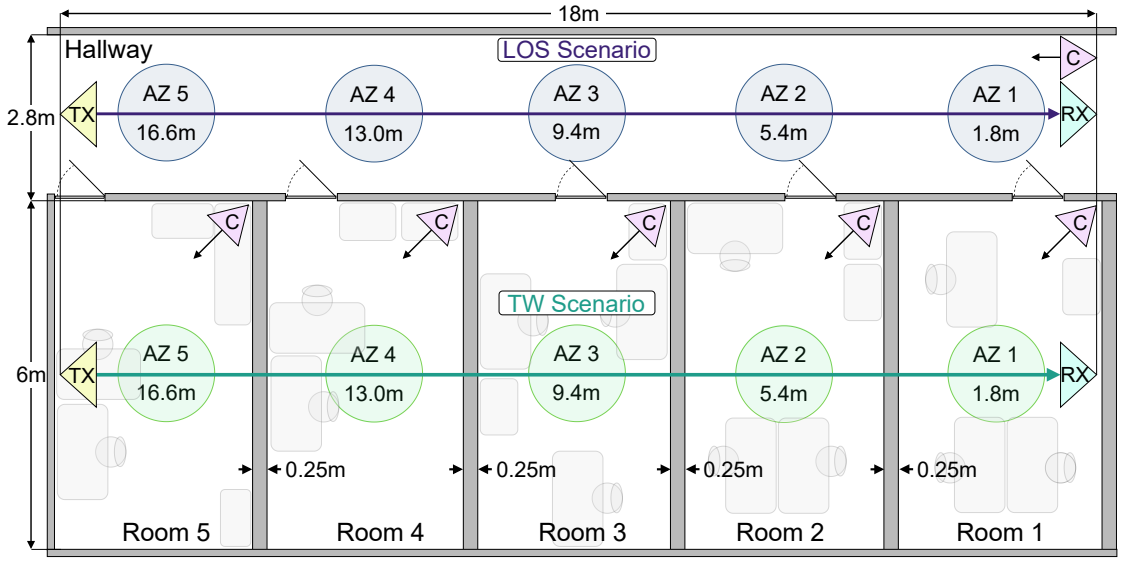


Figure 5.1: Recording environment of the TOA and Wallhack1.8k datasets, showing the transmitter and receiver placement in LOS and through-wall (TW) scenarios. Transmitter (TX), receiver (RX), and camera (C). [52] †

system employing directional high-gain antennas (system $\mathcal{A}$) reliably supports challenging TW scenarios over distances that significantly exceed the typical capabilities of a standard ESP32 with a built-in PIFA.

The dataset is recorded in a carefully structured office environment featuring an 18-meter-long hallway flanked by five uniformly sized office rooms (approximately 3.5 m × 6.0 m each), as depicted in Figure 5.1. The rooms, separated by 25 cm-thick brick walls, introduce substantial attenuation and multipath effects, creating challenging propagation conditions. This controlled layout allows direct comparison between LOS and TW scenarios at identical transmitter-receiver distances, an aspect not provided by any other public dataset. Such direct comparisons facilitate a deeper understanding of signal propagation and model robustness to TW interference.

In the LOS scenario (Figure 5.1, top), the transmitter and receiver are positioned facing each other at opposite ends of the hallway, separated by 18 meters. In the TW scenario (bottom), the devices are similarly aligned but placed outside rooms 5 and 1, respectively, again separated by 18 meters, with signals traversing four brick walls. In both scenarios, antenna alignment is optimized by micro-adjusting the position of the receiver to achieve the maximum RSSI. To facilitate precise temporal trimming of CSI packets, RGB cameras synchronized with the receiver are placed in the hallway (LOS scenario) and within each of the five rooms (TW scenario).

Data collection involves a subject performing two distinct activities, walking and walking combined with arm-waving, for two minutes each within five activity zones (AZ 1–AZ 5, 1.5 m radius), located at 1.8 m, 5.4 m, 9.4 m, 13.0 m, and 16.6 m distances from

the receiver (see Figure 5.1). Additionally, five minutes of background signal (no person present) are recorded per scenario. Raw CSI packets undergo preprocessing, where a Hampel filter [188] removes signal outliers from the CSI amplitude data. The cleaned data is then segmented into 400-packet windows (≈ 4-second intervals at a 100 Hz sampling rate), creating 776 labeled time-frequency amplitude spectrograms (52 L-LTF subcarriers × 400 packets) from an original total of 316,862 WiFi packets. Each spectrogram is annotated with two labels: activity class (0: *no presence*, 1: *walking*, 2: *walking + arm-waving*) and presence detection indicating the corresponding room ID (0–5).

The TOA dataset enables direct cross-scenario evaluations, providing the first publicly available benchmark specifically designed for ESP32-based long-range TW sensing. The dataset is publicly available on Zenodo[1].

5.2.2 Wallhack1.8k Dataset

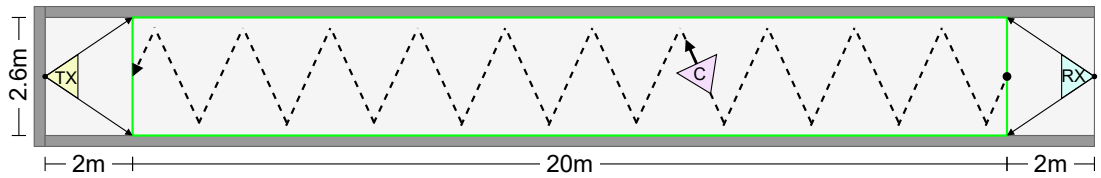
The Wallhack1.8k dataset, introduced in [48], extends the TOA dataset by introducing two distinct WiFi systems, thus enabling the evaluation of cross-system and cross-scenario (LOS ↔ TW) generalization in WiFi-based HAR. The dataset is initially created to investigate image-based data augmentation techniques aimed at enhancing model robustness across different environmental and hardware conditions. Additionally, Wallhack1.8k serves as a benchmark in [52] for designing and evaluating systems $\mathcal{B}$ and $\mathcal{C}_{0-2}$ in long-range LOS and TW sensing scenarios.

Wallhack1.8k shares the exact recording environment, transmitter-receiver placements, and data collection procedures of the TOA dataset (see Figure 5.1 and the TOA dataset description). The primary innovation of Wallhack1.8k lies in the introduction of two new WiFi systems: system $\mathcal{B}$ (biquad antenna) and system $\mathcal{C}1$ (PIFA with plane reflector). Because of the identical experimental setup and data acquisition procedures, the Wallhack1.8k and TOA datasets together uniquely enable direct comparisons across three distinct ESP32-based WiFi systems ($\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}1$), supporting comprehensive cross-system evaluations.

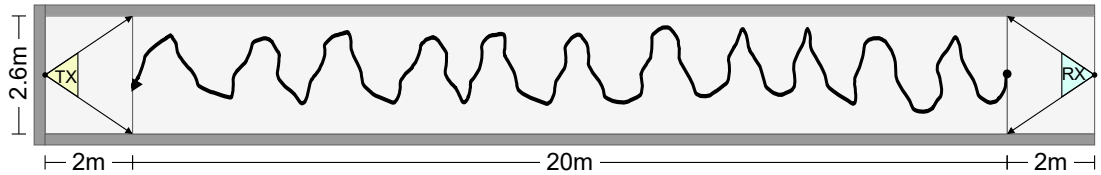
The Wallhack1.8k dataset comprises a total of 1,806 time-frequency amplitude spectrograms, derived from the CSI of 500,093 WiFi packets. The data is systematically divided into four subsets, representing each combination of WiFi system ($\mathcal{B}$ or $\mathcal{C}_1$) and scenario (LOS or TW). Each spectrogram and WiFi packet is labeled with activity classes 0: *no presence*, 1: *walking*, 2: *walking + arm-waving*.

Wallhack1.8k fills an important gap by explicitly supporting the evaluation of cross-system generalization in controlled long-range LOS and TW scenarios, an aspect not addressed by existing public WiFi datasets. The dataset, including raw WiFi packet sequences, derived spectrograms, and labels, is publicly available on Zenodo[2].

[1] TOA Dataset, https://zenodo.org/record/8021099, accessed: 15.04.2025
[2] Wallhack1.8k Dataset, https://zenodo.org/records/15147388, accessed: 15.04.2025



(a) Point-to-point transmitter-receiver arrangement in the recording environment.



(b) Test trajectory.

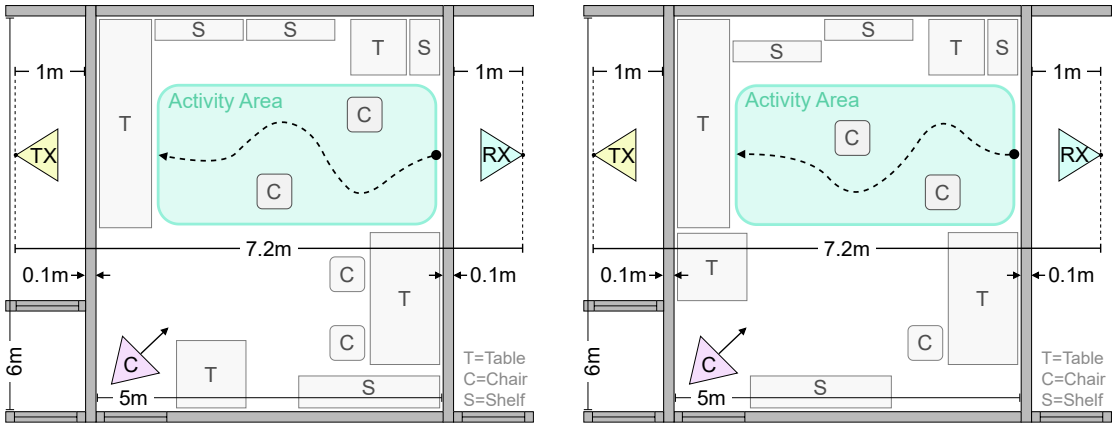
Figure 5.2: (a) HALOC experimental setup showing the point-to-point transmitter-receiver arrangement in the recording environment, with the recording area highlighted in green and the approximate shape of walking trajectories (dotted line). (b) Walking trajectory of the test sequence (black). Transmitter (TX), receiver (RX), and camera (C). [49] †

5.2.3 Hallway Localization Dataset

The HAllway LOCalization (HALOC) dataset, introduced in [49], is created to investigate the feasibility of using ESP32-based WiFi systems for long-range indoor localization tasks. Specifically, the dataset is employed to demonstrate that system $\mathcal{D}$ (ESP32-S3-DevKitC-1U + ALFA APA-M25 antenna) reliably captures person-centric information necessary for accurate single-person localization at significantly greater distances than systems relying on the ESP32's PIFA.

Data collection takes place in a long indoor hallway environment, illustrated in Figure 5.2. The experimental setup consists of a point-to-point arrangement of transmitter and receiver devices placed 24 m apart, ensuring full horizontal beam coverage of a central recording area measuring approximately 2.6 m × 20 m. Such an extended and structured environment provides unique conditions for evaluating long-range localization, going beyond the short-range scenarios typically covered by existing public WiFi datasets.

The dataset comprises CSI measurements and synchronized 3D location trajectories of a single participant. Data collection involves the subject performing multiple walking sequences (six in total, each lasting 4–5 minutes) within the defined recording area. During these sequences, the person moves in a zig-zag trajectory, systematically covering the entire sensing area, as depicted in Figure 5.2. Concurrently, WiFi packets are captured at a rate of 100 Hz using system $\mathcal{D}$, while a chest-mounted camera simultaneously records egocentric video for ground-truth trajectory estimation. The 3D walking trajectories are estimated from the videos with ORB-SLAM3 [189] at 30 Hz, then upsampled to 100 Hz by piecewise-linear interpolation of the original 3D coordinates and uniform resampling to match the WiFi packet sending rate. An example of an extracted walking trajectory appears in Figure 5.2b (black line).



(a) 3DO setup on days 1 and 2.

(b) 3DO setup on day 3.

Figure 5.3: 3DO recording environment layout over three consecutive days: (a) fixed layout on days 1 and 2, and (b) layout on day 3, featuring static environmental variations due to furniture rearrangement. Transmitter (TX), receiver (RX), and camera (C). [50]

In total, the dataset contains 118,679 raw WiFi packets accompanied by per-packet absolute 3D locations expressed in meters relative to the transmitter position. The data is explicitly divided into training (4 sequences), validation (1 sequence), and test (1 sequence) sets, facilitating reproducible experiments and benchmarking of regression models.

HALOC uniquely extends existing WiFi localization datasets by providing a long-range LOS scenario with detailed absolute 3D trajectory annotations. As such, it serves as the first publicly available ESP32-based dataset of its kind, enabling targeted research into long-range localization methods using cost-effective WiFi hardware. The dataset is publicly available on Zenodo[3].

5.2.4 3-Days Office Dataset

The 3-Days Office (3DO) dataset, introduced in [50], provides a foundation for investigating model robustness and cross-domain generalization in WiFi-based person-centric sensing. Specifically, it represents the first publicly available dataset explicitly designed to isolate and systematically label static, dynamic, and temporal environmental variations within a realistic TW scenario. This setup allows researchers to study how these distinct variations affect WiFi-based HAR and localization models, addressing a critical yet underexplored topic.

The dataset is recorded using system $\mathcal{D}$ operating with a 100 Hz packet-sending rate. Figure 5.3a illustrates the recording environment, which consists of a central furnished office room (6 m × 5 m) where activities take place, flanked by two adjacent rooms housing the WiFi transmitter and receiver. The transmitter and receiver are arranged in

[3]HALOC Dataset, https://zenodo.org/records/10715595, accessed: 15.04.2025

a point-to-point configuration separated by 7.2 m, with two plasterboard walls (each 10 cm thick) introducing TW signal propagation challenges.

A unique characteristic of the 3DO dataset is the explicit isolation of static environmental variations over three consecutive recording days. On the first two days (Figure 5.3a), the room layout remains constant, allowing analyses of only dynamic (variations in activity execution) and temporal (time-induced signal drift) changes between days. On day three, furniture and smaller items within the central room are rearranged (Figure 5.3b), introducing controlled static environmental changes. Throughout the entire experiment, the positions of the transmitter, receiver, and activity area (4 m × 2.5 m) remain fixed to ensure consistent conditions for evaluating domain shifts.

Five repeated sequences per activity class, *walking*, *sitting*, and *lying*, are recorded each day, yielding a total of 42 sequences (3 sequences are excluded due to corruption). Each sequence lasts five minutes. During these activities, the subject moves freely within the activity area while performing specific tasks: walking involves continuous movement avoiding furniture, sitting involves alternating between two chairs with random limb movements, and lying simulates a fall scenario with sliding and struggling motions. In addition to the activity sequences, one ten-minute sequence of *no presence* is recorded for each day, resulting in a total of 45 sequences comprising the 3DO dataset.

To capture ground-truth trajectories, an egocentric camera mounted on the subject's chest is used. As with the HALOC dataset, videos are subsequently processed using ORB-SLAM3 [189] to obtain accurate 3D trajectories at 30 Hz, which are then linearly up-sampled to 100 Hz to match the packet sending rate. Activity labels (0: *no presence*, 1: *walking*, 2: *sitting*, 3: *lying*) are manually annotated from video frames using visual cues. The resulting dataset contains 1,292,727 raw WiFi packets, each annotated with synchronized activity class labels and corresponding 3D location data.

Due to its explicit separation of static, dynamic, and temporal variations in a TW scenario, the 3DO dataset serves as a unique benchmark to study and improve model generalization. Researchers can directly train models using data from day one and evaluate them on days two and three, thereby quantifying robustness to domain shifts induced by temporal and static environmental variations. The 3DO dataset is publicly available on Zenodo[4], supporting further advancements in WiFi-based PCS research.

5.2.5 WiFiCam Dataset

The WiFiCam dataset, introduced in [51], represents the first publicly available dataset specifically designed for synthesizing RGB images directly from CSI captured in a TW scenario. It is originally created to demonstrate and validate a novel multimodal VAE-based approach for CSI-based TW imaging.

Data is recorded in a 3.8 m × 5.3 m office environment, as shown in Figure 5.4. A single-room TW scenario is set up, with the transmitter and receiver placed outside

[4]3DO Dataset, https://zenodo.org/records/10925351, accessed: 15.04.2025

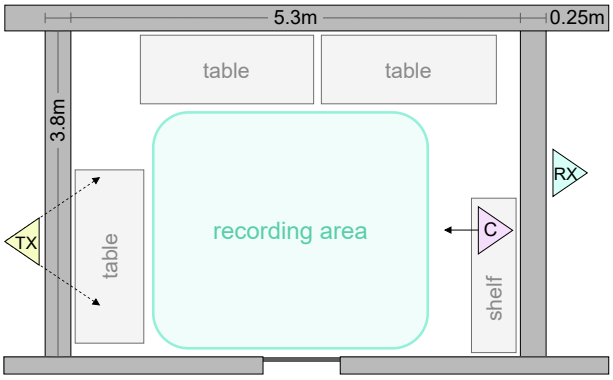


Figure 5.4: WiFiCam dataset experimental setup for through-wall imaging, showing the arrangement of transmitter, receiver, and camera in the office recording environment. Transmitter (TX), receiver (RX), and camera (C). [51]

opposite walls of the room, facing each other. The system used is system $\mathcal{D}$, which transmits WiFi packets at a fixed rate of 100 Hz. An RGB camera positioned inside the room captures images at a resolution of 640 × 480 pixels and a frame rate of 30 Hz, synchronized with the WiFi packet collection.

During data collection, a subject continuously performs walking activities within a predefined recording area inside the room over a duration of ten minutes. Raw data synchronization is achieved by connecting both the receiver and the camera to a notebook that concurrently logs WiFi packets and RGB frames with timestamps. For accurate modality alignment, each WiFi packet is paired with the closest RGB frame based on timestamp proximity, resulting in approximately three WiFi packets per image. The dataset consists of 57,413 raw WiFi packets and 18,261 synchronized RGB images. Unlike other datasets discussed here, WiFiCam does not provide explicit activity or trajectory annotations, focusing instead on the raw image–CSI pairing. This focus makes the dataset particularly suitable for tasks such as direct CSI-to-image synthesis, cross-modal learning, and visual HAR from TW CSI data. The dataset is publicly available on Zenodo[5].

5.2.6 Temporal Resolution

All proposed datasets are captured at a fixed packet sending rate of 100 Hz (corresponding to a 10 ms packet interval). However, the packet sampling rate, i.e., the actual frequency of packet arrivals at the receiver, can vary due to time-varying transmission delays and packet loss. To assess this, Figure 5.5 presents the distribution of sampling intervals between consecutive packets across the datasets listed in Table 5.2. The distributions are tightly centered around the nominal interval of 10 ms. Moreover, the global median sampling interval across all datasets (comprising approximately 2.3 million packets) is 9.999 ms, corresponding to an effective sampling rate of 100.01 Hz. This consistency across LOS and TW scenarios, varied environments, different WiFi systems, and transmitter-

[5]WiFiCam Dataset, https://zenodo.org/records/11554280, accessed: 15.04.2025

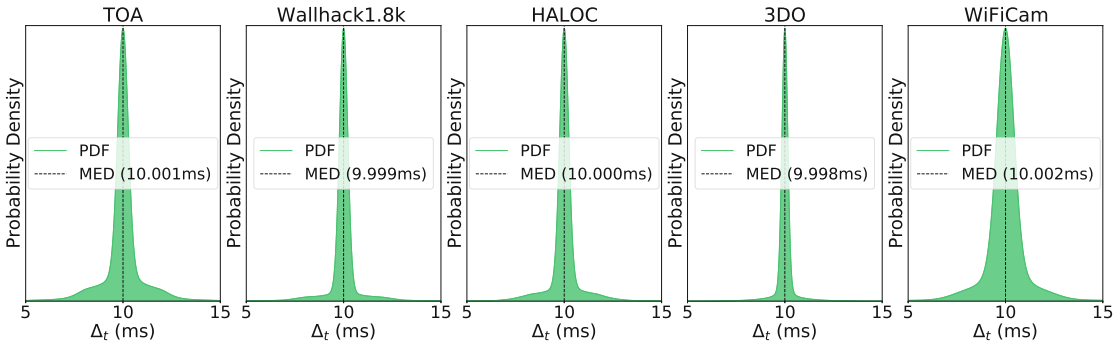


Figure 5.5: Probability density of packet sampling intervals between consecutive WiFi packets Δ_t for the proposed datasets listed in Table 5.2.

receiver distances, confirms that the packet sampling rate remains stable under domain variations.

Summary

The proposed datasets extend WiFi-based PCS to underexplored but practically important scenarios such as long-range and TW sensing. Captured with the proposed systems, they introduce novel combinations of modalities, scenarios, and labels for tasks including presence detection, HAR, localization, and image synthesis. By capturing domain variations across hardware, environments, and time, they enable systematic evaluation of cross-domain generalization, addressing the challenges in **RQ I**, **RQ III**, and **RQ IV**. They further define the sensing conditions and data characteristics that guide the design and evaluation of embedded, real-time CSI processing methods in Chapter 6, informing the contributions related to **RQ II**.

CHAPTER 6

Methodology

This chapter presents the methodological contributions, addressing key challenges in enabling practical and robust WiFi-based PCS. It begins with feasibility studies that evaluate the long-range and TW sensing capabilities of the WiFi systems proposed. Building on these results, the chapter introduces methods for the efficient processing of CSI, explores strategies to improve model robustness under domain shift, and concludes with an investigation of CSI-based TW imaging. Together, these works form a methodological framework for advancing CSI-based PCS across diverse sensing tasks and domains.

6.1 Feasibility of Long-Range and Through-Wall PCS

RQ I explores how long-range TW PCS can be realized using low-cost, COTS WiFi hardware, aiming to minimize complexity and deployment cost. Current PCS methods predominantly target short-range LOS scenarios, limiting their scalability and practical applicability in real-world environments, especially when signals must traverse walls or span considerable distances. To overcome these challenges, a methodology that enhances the widely available and cost-effective ESP32-S3 microcontroller with directional sensing capabilities is proposed. By designing optimized WiFi systems, PCS capabilities are extended to challenging long-range and TW domains. Comprehensive evaluations using TOA, Wallhack1.8k, and HALOC datasets empirically verify the feasibility and robustness of the proposed methodology for PCS tasks such as presence detection, HAR [40, 52], and person localization [49].

6.1.1 Evaluation Methodology

To systematically verify the feasibility of long-range and TW PCS using the proposed WiFi systems, comprehensive evaluations are conducted across three core person-centric sensing tasks: presence detection, HAR, and localization. Specifically, evaluations focus on four optimized WiFi systems ($\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}_1$, and $\mathcal{D}$), all of which enhance the cost-effective ESP32-S3 microcontroller with directional sensing capabilities. Each system addresses

inherent hardware constraints, such as limited range and sensitivity, by integrating either custom directional antennas or passive reflector solutions.

Evaluations leverage the systematically collected datasets, TOA, Wallhack1.8k, and HALOC, to ensure coverage of a variety of sensing scenarios. Specifically, presence detection and HAR tasks are evaluated using subsets extracted from the TOA dataset (System $\mathcal{A}$) and the Wallhack1.8k dataset (Systems $\mathcal{B}$ and $\mathcal{C}_1$). Both datasets feature structured environments explicitly designed for direct comparisons of LOS and TW performance, spanning distances of up to 18 meters and including signal propagation through multiple walls. The HALOC dataset (System $\mathcal{D}$) complements these evaluations by providing data for assessing localization capabilities over a 20-meter LOS hallway scenario.

To maintain consistency, reproducibility, and comparability of results, all evaluations employ standard CNNs based on the lightweight and efficient *EfficientNetV2* architecture [190]. *EfficientNetV2* is designed to achieve high classification accuracy while using fewer computational resources and training faster than previous, standard CNN architectures such as *ResNet* [191]. The evaluations leverage amplitude-based CSI spectrograms due to their robustness to noise and signal instability.

Data For the assessment of presence detection and HAR performance, CSI amplitude spectrogram subsets are extracted from the TOA and Wallhack1.8k datasets as detailed in Table 6.1.

The TOA dataset is divided into four subsets: T_P$_{L\mathcal{A}}$ and T_P$_{T\mathcal{A}}$ for presence detection, and T_A$_{L\mathcal{A}}$ and T_A$_{T\mathcal{A}}$ for HAR, in LOS and TW scenarios respectively. The Wallhack1.8k dataset is divided into eight subsets: W_P$_{L\mathcal{B}}$, W_P$_{T\mathcal{B}}$, W_P$_{L\mathcal{C}_1}$, and W_P$_{T\mathcal{C}_1}$ for presence detection, and W_A$_{L\mathcal{B}}$, W_A$_{T\mathcal{B}}$, W_A$_{L\mathcal{C}_1}$, and W_A$_{T\mathcal{C}_1}$ for HAR. Subsets follow the naming convention X_Y$_{Z\mathcal{W}}$, where X denotes the dataset (T: TOA, W: Wallhack1.8k), Y indicates the task (P: presence detection, A: HAR), Z indicates the scenario (L: LOS, T: TW), and $\mathcal{W}$ indicates the system ($\mathcal{A}$, $\mathcal{B}$, or $\mathcal{C}_1$). All subsets are further partitioned into training, validation, and test splits at an 8:1:1 ratio.

For evaluating the feasibility of long-range localization, the HALOC dataset, recorded with System $\mathcal{D}$, is employed using its predefined training, validation, and test splits, consisting of four training sequences, one validation sequence, and one test sequence.

6.1.2 Presence Detection

The presence detection task involves classifying the spatial location of a person based on patterns in CSI amplitude spectrograms. It is formulated as a 6-class classification problem, where the classes correspond to discrete spatial zones, including five room locations and a background class representing no presence. This task is evaluated on both LOS and TW scenarios over an 18-meter range, using the TOA and Wallhack1.8k datasets. In the TW scenario (T_P$_{T\mathcal{A}}$, W_P$_{T\mathcal{B}}$, W_P$_{T\mathcal{C}1}$), the person is located in one of five adjacent rooms, while in the LOS scenario (T_P$_{L\mathcal{A}}$, W_P$_{L\mathcal{B}}$, W_P$_{L\mathcal{C}1}$), presence is detected at hallway segments aligned with those room positions.

Dataset	Task	Scenario	System	Rooms	Classes	Sampling Interval	Samples
T_P$_{L\mathcal{A}}$	PD	LOS	$\mathcal{A}$	1	6	$\approx$ 4 s (400 packets)	392
T_P$_{T\mathcal{A}}$	PD	TW	$\mathcal{A}$	5	6	$\approx$ 4 s (400 packets)	384
T_A$_{L\mathcal{A}}$	HAR	LOS	$\mathcal{A}$	1	3	$\approx$ 4 s (400 packets)	392
T_A$_{T\mathcal{A}}$	HAR	TW	$\mathcal{A}$	5	3	$\approx$ 4 s (400 packets)	384
W_P$_{L\mathcal{B}}$	PD	LOS	$\mathcal{B}$	1	6	$\approx$ 4 s (400 packets)	458
W_P$_{L\mathcal{C}_1}$	PD	LOS	$\mathcal{C}_1$	1	6	$\approx$ 4 s (400 packets)	461
W_P$_{T\mathcal{B}}$	PD	TW	$\mathcal{B}$	5	6	$\approx$ 4 s (400 packets)	450
W_P$_{T\mathcal{C}_1}$	PD	TW	$\mathcal{C}_1$	5	6	$\approx$ 4 s (400 packets)	437
W_A$_{L\mathcal{B}}$	HAR	LOS	$\mathcal{B}$	1	3	$\approx$ 4 s (400 packets)	458
W_A$_{L\mathcal{C}_1}$	HAR	LOS	$\mathcal{C}_1$	1	3	$\approx$ 4 s (400 packets)	461
W_A$_{T\mathcal{B}}$	HAR	TW	$\mathcal{B}$	5	3	$\approx$ 4 s (400 packets)	450
W_A$_{T\mathcal{C}_1}$	HAR	TW	$\mathcal{C}_1$	5	3	$\approx$ 4 s (400 packets)	437

Table 6.1: TOA and Wallhack1.8k subsets for assessing presence detection and HAR performance in LOS and TW scenarios.

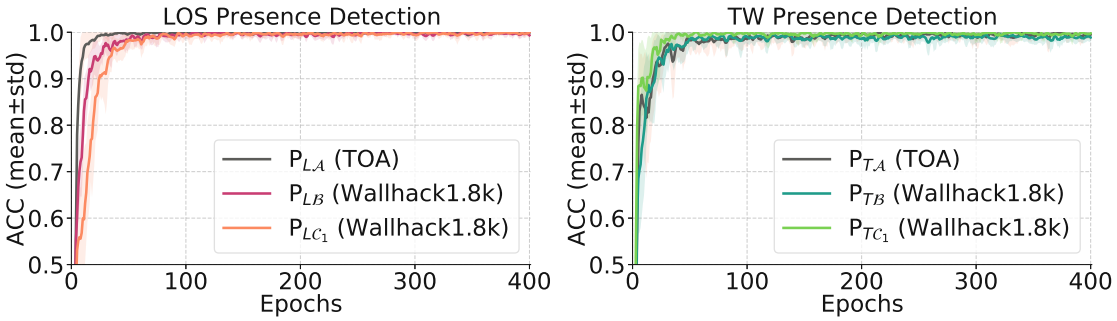


Figure 6.1: LOS and TW presence detection accuracy on TOA and Wallhack1.8k validation subsets, measured across ten independent training runs with random initialization, spanning 400 epochs. [40, 52] †

Model Training Presence detection models based on the *EfficientNetV2 small* architecture are trained using the TOA and Wallhack1.8k subsets described in Table 6.1. All models are trained from scratch for 400 epochs using the Adam optimizer, a learning rate of 1×10^{-4}, and a batch size of 16. A balanced sampler is employed to address class imbalance within training sets. Data augmentation consists of random circular shifts applied to CSI amplitude spectrograms along the temporal axis. Each model configuration undergoes ten independent training runs with random initialization, and evaluation metrics, including precision, recall, F1-score, and classification accuracy (ACC), are reported as means and standard deviations across these runs on their respective test subsets. Model names follow the convention P$_{Z\mathcal{W}}$, where Z denotes the scenario (L: LOS, T: TW), and $\mathcal{W}$ indicates the system ($\mathcal{A}$, $\mathcal{B}$, or $\mathcal{C}_1$).

Figure 6.1 shows training progress for presence detection models trained on the TOA and Wallhack1.8k datasets in LOS and TW scenarios. LOS models (left) (i.e., P$_{L\mathcal{A}}$, P$_{L\mathcal{B}}$, P$_{L\mathcal{C}_1}$)

Model	Testset	Precision ↑	Recall ↑	F1-Score ↑	ACC ↑
$P_{L\mathcal{A}}$	T_$P_{L\mathcal{A}}$	98.67 ±1.8	98.61 ±1.9	98.64 ±1.8	98.46 ±2.1
$P_{T\mathcal{A}}$	T_$P_{T\mathcal{A}}$	98.14 ±1.6	97.72 ±2.2	97.93 ±1.9	97.89 ±2.0
$P_{L\mathcal{B}}$	W_$P_{L\mathcal{B}}$	97.67 ±1.9	97.30 ±2.4	97.48 ±2.1	97.83 ±1.9
P_{LC_1}	W_P_{LC_1}	99.00 ±1.6	98.69 ±1.9	98.85 ±1.7	98.91 ±1.5
$P_{T\mathcal{B}}$	W_$P_{T\mathcal{B}}$	96.84 ±2.1	97.60 ±1.5	97.22 ±1.8	96.89 ±2.3
P_{TC_1}	W_P_{TC_1}	98.44 ±2.1	98.32 ±2.2	98.38 ±2.1	98.64 ±1.8

Table 6.2: Performance of LOS and TW presence detection models on TOA ($P_{L\mathcal{A}}$, $P_{T\mathcal{A}}$) and Wallhack1.8k ($P_{L\mathcal{B}}$, P_{LC_1}, $P_{T\mathcal{B}}$, P_{TC_1}) datasets, reported as mean and standard deviation across ten independent training runs with random initialization.

demonstrate rapid convergence, reaching near-optimal accuracy within approximately 30 epochs. Conversely, TW models (right) (i.e., $P_{T\mathcal{A}}$, $P_{T\mathcal{B}}$, P_{TC_1}) exhibit slower convergence, requiring up to approximately 120 epochs to reach stable, high accuracy. Ultimately all models achieve consistent validation accuracy, highlighting their ability to effectively learn from the underlying data.

Results Table 6.2 summarizes the performance of presence detection models on the TOA and Wallhack1.8k test subsets. On the TOA dataset, models trained with system $\mathcal{A}$ data achieve consistently high accuracy in both LOS and TW scenarios, with model $P_{L\mathcal{A}}$ slightly outperforming $P_{T\mathcal{A}}$ (98.46% vs. 97.89%). These results indicate that system $\mathcal{A}$ effectively captures PCI, maintaining high reliability even in challenging TW scenarios.

On Wallhack1.8k, models using system $\mathcal{C}_1$ data (i.e., P_{LC_1} and P_{TC_1}) consistently outperform their system $\mathcal{B}$ counterparts (i.e., $P_{L\mathcal{B}}$, $P_{T\mathcal{B}}$). In the LOS scenario, accuracy for system $\mathcal{C}_1$ reaches 98.91%, compared to 97.83% for system $\mathcal{B}$. Similarly, in the TW scenario, system $\mathcal{C}_1$ achieves an accuracy of 98.64%, while system $\mathcal{B}$ achieves 96.89%. Notably, the accuracy drop from LOS to TW is smaller for system $\mathcal{C}_1$ (0.27 percentage points) compared to system $\mathcal{B}$ (0.94 percentage points). Although a performance gap between systems $\mathcal{B}$ and $\mathcal{C}_1$ is measurable, it remains modest, making a definitive statement about system superiority difficult due to minor dataset variations.

Overall, the high performance across all three evaluated systems ($\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}_1$), with mean accuracies of 98.40% in LOS scenarios and 97.80% in TW scenarios, demonstrates their capability to reliably capture PCI in long-range TW sensing settings, effectively supporting presence detection tasks.

6.1.3 Human Activity Recognition

The HAR task shares the same physical environment and propagation settings as presence detection but focuses on identifying the type of activity being performed. It is defined as a 3-class classification problem, distinguishing between *no presence*, *walking*, and *walking + arm-waving*. Evaluation is conducted under both LOS and TW conditions using the same TOA and Wallhack1.8k subsets (T_$A_{L\mathcal{A}}$, T_$A_{T\mathcal{A}}$, W_$A_{L\mathcal{B}}$, W_$A_{T\mathcal{B}}$,

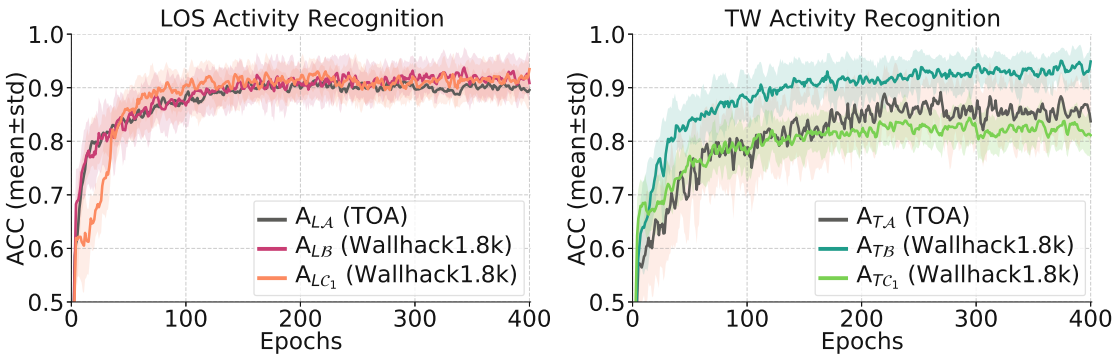


Figure 6.2: LOS and TW HAR accuracy on TOA and Wallhack1.8k validation subsets, measured across ten independent training runs with random initialization, spanning 400 epochs. [40, 52] †

W_A$_{LC1}$, W_A$_{TC1}$). This task aims to benchmark a systems' ability to capture coarse and fine-grained human motion patterns in the CSI.

Model Training The training procedure for HAR models mirrors that of the presence detection task. All models are based on the *EfficientNetV2 small* architecture and trained from scratch for 400 epochs using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 16. A balanced sampler eliminates class imbalance, and spectrograms are augmented through random circular shifts along the temporal axis. Each model is trained ten times with random initialization, and the evaluation metrics, precision, recall, F1-score, and classification accuracy (ACC), are reported as means and standard deviations across the corresponding test subsets. HAR models follow the naming convention A$_{Z\mathcal{W}}$, where Z denotes the scenario (L: LOS, T: TW) and $\mathcal{W}$ refers to the system ($\mathcal{A}$, $\mathcal{B}$, or $\mathcal{C}_1$).

Figure 6.2 shows the training behavior of HAR models trained on the TOA and Wallhack1.8k datasets across LOS (i.e., A$_{L\mathcal{A}}$, A$_{L\mathcal{B}}$, A$_{LC_1}$) and TW (i.e., A$_{T\mathcal{A}}$, A$_{T\mathcal{B}}$, A$_{TC_1}$) scenarios. In the LOS scenario, models converge after approximately 200 epochs, with final validation accuracies plateauing near 90%. This slower and less stable convergence, compared to the presence detection task, highlights the increased difficulty of distinguishing between activities. In the TW scenario, models require the full 400 epochs to converge, with increased run-to-run variance and early divergence in performance across systems, which persists throughout training and highlights the increased difficulty of HAR in the presence of multi-wall signal attenuation and multipath effects.

Results Table 6.3 presents the performance of HAR models trained on the TOA and Wallhack1.8k datasets. On TOA, system $\mathcal{A}$ achieves 97.43% accuracy in the LOS scenario (A$_{L\mathcal{A}}$) and 88.16% in the TW scenario (A$_{T\mathcal{A}}$), indicating a substantially larger performance gap than observed in the presence detection task. This suggests that the HAR task, which requires distinguishing between more subtle motion patterns, is more sensitive to signal degradation under TW conditions. Notably, the TW model exhibits

Model	Testset	Precision ↑	Recall ↑	F1-Score ↑	ACC ↑
$A_{L\mathcal{A}}$	T_$A_{L\mathcal{A}}$	97.80 ±0.9	98.04 ±0.9	97.92 ±0.9	97.43 ±1.1
$A_{T\mathcal{A}}$	T_$A_{T\mathcal{A}}$	90.94 ±3.8	87.92 ±4.3	89.39 ±4.0	88.16 ±4.6
$A_{L\mathcal{B}}$	W_$A_{L\mathcal{B}}$	89.98 ±2.5	89.05 ±2.2	89.51 ±2.3	89.35 ±2.3
A_{LC_1}	W_A_{LC_1}	91.12 ±3.9	90.42 ±4.0	90.77 ±4.0	90.22 ±4.3
$A_{T\mathcal{B}}$	W_$A_{T\mathcal{B}}$	92.81 ±3.3	91.20 ±3.9	91.99 ±3.5	92.00 ±3.5
A_{TC_1}	W_A_{TC_1}	86.56 ±5.0	86.18 ±4.9	86.37 ±4.9	86.82 ±4.7

Table 6.3: Performance of LOS and TW HAR models on TOA (A_L, $A_{T\mathcal{A}}$) and Wallhack1.8k ($A_{L\mathcal{B}}$, A_{LC_1}, $A_{T\mathcal{B}}$, A_{TC_1}) datasets, reported as mean and standard deviation across ten independent training runs with random initialization.

higher run-to-run variance (±4.6 percentage points), occasionally reaching LOS-level performance.

On the Wallhack1.8k dataset, both systems $\mathcal{B}$ and $\mathcal{C}_1$ perform similarly in the LOS scenario. The model trained on data from system $\mathcal{C}_1$ (A_{LC_1}) achieves a slightly higher accuracy of 90.22% compared to 89.35% for system $\mathcal{B}$ ($A_{L\mathcal{B}}$). This outcome is somewhat counterintuitive, as amplitude spectrograms captured by system $\mathcal{B}$ exhibit more pronounced signal variations in response to human activities in the LOS path, as shown in Figure 6.3. It is hypothesized that while system $\mathcal{B}$ may produce higher-magnitude signal fluctuations, the relative structure of activity-induced patterns remains similar across both systems. Consequently, models trained on data from system $\mathcal{C}_1$ can achieve comparable performance without necessarily benefiting from stronger amplitude variations.

In the TW scenario (the scenario these systems were designed for), however, system $\mathcal{B}$ clearly outperforms system $\mathcal{C}_1$. The model $A_{T\mathcal{B}}$ achieves 92.00% accuracy, compared to 86.82% for A_{TC_1}. This performance gap aligns with the diverging validation behavior observed during training and reflects system $\mathcal{B}$'s enhanced ability to capture activity-relevant signal variations under severe attenuation and multipath conditions introduced by walls.

Taken together, the results confirm that all three systems ($\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}_1$) are capable of reliably capturing person-centric information relevant for HAR in long-range scenarios. Furthermore, with mean accuracies of 92.33% in LOS and 88.99% in TW scenarios, these systems demonstrate practical feasibility for HAR applications in challenging environments.

6.1.4 Localization

The localization task, presented by the HALOC dataset, aims to evaluate the feasibility of long-range person localization using CSI amplitude spectrograms (i.e., fingerprint-based indoor localization). Specifically, it is formulated as a regression problem, where the goal is to predict the 3D coordinates of a walking person relative to the receiver. Recordings are conducted in a 2.6 m × 20 m indoor hallway environment using system $\mathcal{D}$.

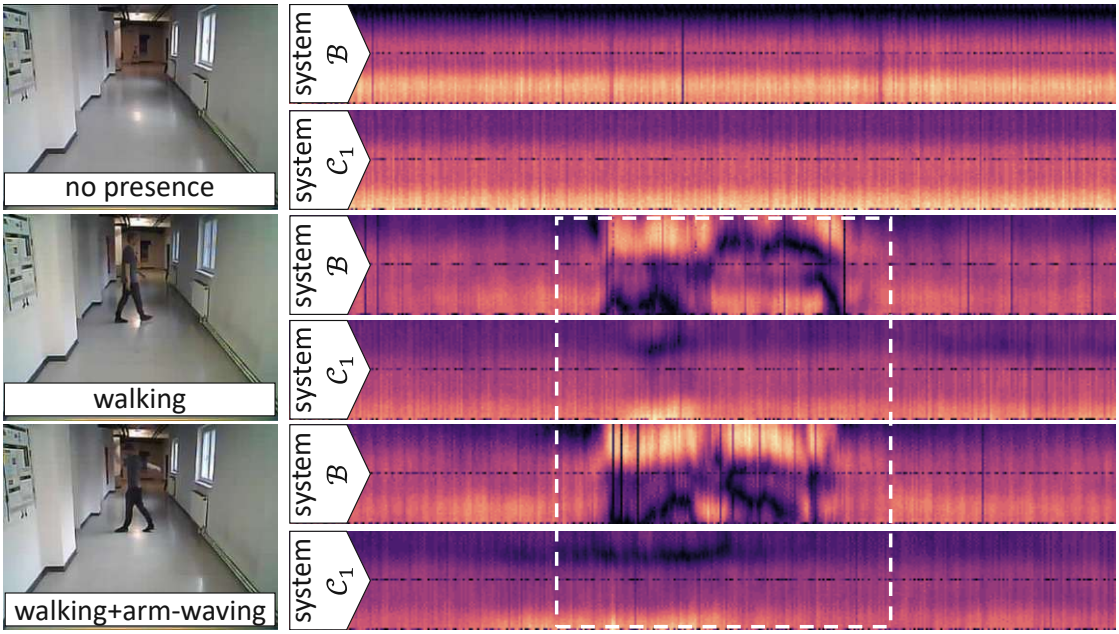


Figure 6.3: LOS CSI amplitude spectrograms from the Wallhack1.8k dataset of the classes *no presence*, *walking*, and *walking + arm-waving*, captured with systems $\mathcal{B}$ and $\mathcal{C}_1$ at a distance of approximately 9.4 m (corresponding to the center of room 3 in the TW scenario). The amplitude spectrograms show the amplitudes of 52 L-LTF subcarriers over a time interval of $\approx$ 4 seconds (400 WiFi packets). Highlighted areas (dotted lines) show the varying signal characteristics between systems. [48] †

Model Training To assess localization performance, an *EfficientNetV2 small* regression model is trained on the HALOC dataset using the original split of four training, one validation, and one test sequence. The model takes as input amplitude spectrograms of size 52×351, constructed from 52 L-LTF subcarriers over 351 WiFi packets ($\approx$ 3.51 seconds), which is identified as the optimal width through hyperparameter search. Training is conducted for 200 epochs using the AdamW optimizer with a learning rate and weight decay of 1×10^{-3}. A cosine annealing scheduler adjusts the learning rate over time. Data augmentation includes random channel-wise amplitude perturbations (± 0.2), pixel-wise dropout ($p = 0.2$), and column-wise dropout ($p = 0.2$) with channel mean replacement. Model selection is based on the best validation Root Mean Squared Error (RMSE), and evaluation is performed on the held-out test sequence.

Results The trained model achieves an RMSE of 0.197 m on the HALOC test sequence, confirming its ability to predict 3D locations with high spatial precision. As illustrated in Figure 6.4, the predicted positions (colored) closely follow the actual walking trajectory (black), accurately capturing both linear motion and directional changes. These results highlight the viability of system $\mathcal{D}$ for long-range localization, offering a low-cost and deployable solution for continuous person tracking in indoor environments.

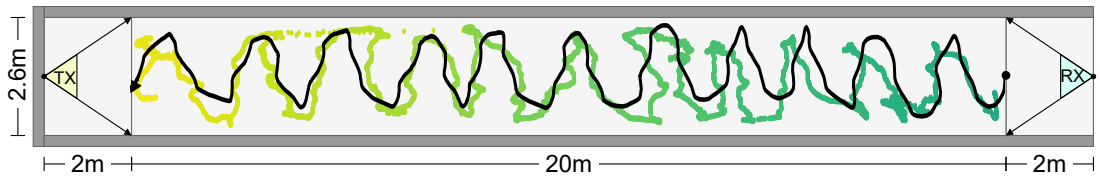


Figure 6.4: Walking trajectory of the test sequence (black) and locations predicted by the regression model (colored). Transmitter (TX) and receiver (RX). [49]

6.1.5 Discussion

The conducted evaluation explicitly validates that low-cost, single-link COTS WiFi systems equipped with directional sensing effectively overcome inherent range and sensitivity limitations of default omnidirectional configurations, thereby addressing **RQ I**. All evaluated WiFi systems ($\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}_1$, and $\mathcal{D}$) demonstrate robust long-range TW PCS performance across presence detection, HAR, and localization tasks, achieving reliable sensing with minimal complexity and cost.

For presence detection, systems $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}_1$ exhibit consistently high accuracy on the TOA and Wallhack1.8k datasets. Specifically, mean accuracies of 98.40% in LOS scenarios and 97.80% in TW scenarios are achieved, covering distances up to 18 meters and penetrating four 25 cm-thick brick walls. These results strongly support the feasibility of long-range TW presence detection using the proposed directional WiFi systems.

The HAR task, which involves distinguishing subtle human movements, also shows strong performance, albeit somewhat reduced due to its increased complexity. Mean accuracies across the TOA and Wallhack1.8k datasets reach approximately 92.33% in LOS scenarios and around 89.99% in TW scenarios. Despite inherent signal degradation and multipath effects in TW scenarios, the proposed systems maintain robust sensing capabilities.

Finally, the localization capability demonstrated by system $\mathcal{D}$ further emphasizes the practicality of the approach. Evaluated on the HALOC dataset in a 20-meter LOS hallway scenario, system $\mathcal{D}$ achieves precise indoor tracking with a test RMSE of 0.197 m. The high fidelity of the predicted trajectories underscores the effectiveness of directional sensing for precise person localization over significant spatial scales.

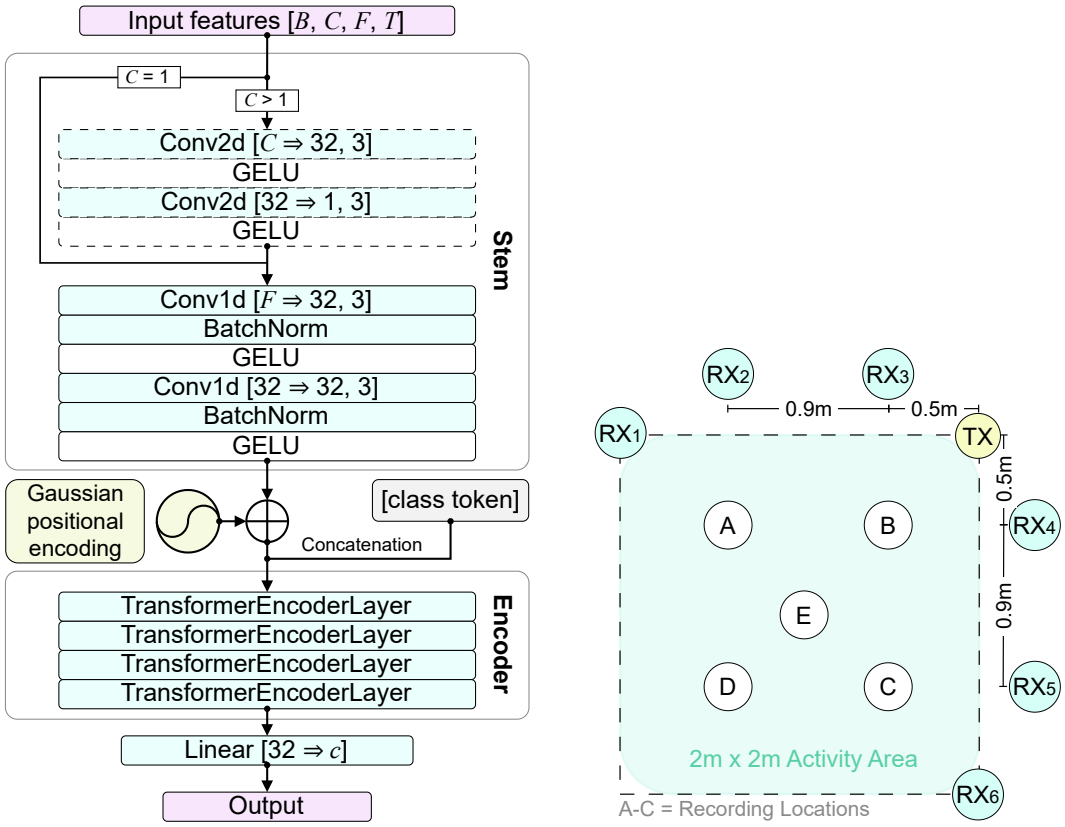
6.2 Efficient Architectures for the Processing of CSI

RQ II addresses the challenge of designing efficient deep learning architectures capable of processing WiFi CSI in real-time on low-power edge devices while capturing CSI's unique temporal and spectral characteristics. Practical PCS deployments, especially in privacy-aware smart environments and healthcare applications, necessitate continuous real-time inference directly on resource-constrained embedded hardware [192]. However, traditional CSI-based sensing often relies on generic vision architectures such as CNNs [191, 190] that, despite their adaptability and strong feature extraction capabilities [177, 74, 71], incur substantial computational overhead due to their high parameter count and have translational-equivariance priors that are misaligned with the non-shift-invariant nature of CSI [46]. On the other hand, existing architectures tailored to the processing of RF/WiFi data [46, 55] rely on computationally-intensive preprocessing steps such as PCA and STFT, rendering them unsuitable for real-time inference on edge devices. Recently, Transformer-based architectures have demonstrated potential for WiFi sensing tasks by effectively modeling long-range temporal dependencies in CSI data [104, 96]. For example, the two-stream Transformer approach (*THAT*) [104] separately captures time-over-subcarrier and subcarrier-over-time relationships. However, this and similar architectures often incorporate multi-stream or multi-stage pipelines, increasing complexity, parameter count, and inference latency. Therefore, despite their theoretical advantages, existing specialized Transformer models remain impractical for real-time PCS deployments at the edge.

To address **RQ II**, *WiFlexFormer* [53], a novel Transformer-based architecture explicitly designed to achieve real-time CSI-based PCS on low-power embedded devices is introduced. Unlike previous architectures, *WiFlexFormer* features a streamlined, parameter-efficient structure specifically tailored to exploit the intrinsic temporal and spectral characteristics of WiFi CSI. Consequently, it achieves exceptionally low parameter count ($\approx$ 50k) and inference latency ($\approx$ 10 milliseconds on an *Nvidia Jetson Orin Nano*, as integrated in System $\mathcal{D}$) without sacrificing competitive HAR performance. Comprehensive evaluations on public datasets, such as Widar3.0 [74] and 3DO, demonstrate that *WiFlexFormer* matches or surpasses the performance of generic vision and state-of-the-art architectures tailored to RF/WiFi data, yet with up to three orders of magnitude fewer parameters and faster inference. Thus, *WiFlexFormer* represents a critical advancement in enabling scalable, practical, and real-time deployable WiFi-based PCS solutions, laying the groundwork for further exploration of efficient model architectures suitable for widespread edge deployment.

6.2.1 WiFlexFormer

The *WiFlexFormer* architecture, illustrated in Figure 6.5a, comprises an initial stem module followed by a Transformer encoder. This design is motivated by the unique characteristics of WiFi CSI. In the proposed approach, the stem module is specifically tailored to CSI-based features, performing dimensionality reduction on the input and enabling the use of a compact, BERT-like Transformer architecture that achieves robust performance with a low parameter count. Transformer encoders can capture long-



(a) WiFlexFormer architecture (b) Widar3.0 setup

Figure 6.5: (a) The proposed *WiFlexFormer* architecture. Convolution parameters are denoted as: [input channels ⇒ number of filters, kernel size]. The final linear layer has 32 input features and c output features, the number of classes. Only the output at the position of the class token is used for the prediction, the remaining positions are discarded. (b) Widar3.0 recording setup featuring a single transmitter and six receivers. [53]

range dependencies and global context [103], yet their extension to CSI-based feature learning is not straightforward. *WiFlexFormer* introduces a novel BERT-like design with a class token, a strategy not previously proposed for CSI, which distinguishes it from other methods that merely employ standard Transformer layers. Furthermore, the *WiFlexFormer* does not rely on CSI preprocessing and consists solely of efficient, well-supported operations, ensuring compatibility with resource-constrained platforms while matching or exceeding the performance of larger RF-specific and generic vision architectures. The PyTorch implementation of *WiFlexFormer* is publicly available[1].

Input Features *WiFlexFormer* expects generic real-valued input features in the shape $[B, C, F, T]$, where B, C, F, and T correspond to the batch, channel, frequency, and time dimensions, respectively. This allows for the processing of common WiFi features

[1]WiFlexFormer, https://github.com/StrohmayerJ/WiFlexFormer, accessed: 15.04.2025

such as CSI, DFS, amplitude, phase, and their derivations. To handle complex-valued inputs like CSI, real and imaginary parts are separated and stored in two real-valued channels, as proposed in [46]. This results in an input tensor shape of $[B,\ 2,\ F,\ T]$. For unprocessed inputs such as CSI or amplitude, the dimensionality of F corresponds to the number of subcarriers, while for features resulting from STFT, such as DFS, F corresponds to the number of frequency bins.

Stem The stem of *WiFlexFormer* is designed to handle various input features and perform initial feature extraction and dimensionality reduction. It consists of two main components: a 2D stem for multi-channel inputs and a 1D stem for further processing. For inputs with multiple channels ($C > 1$), such as DFS features where each subcarrier generates a 2D spectrogram, a 2D stem comprising two convolutional layers with kernel size $(1, 3)$ is applied. The first convolutional layer maintains the number of input channels while the second reduces it to 1. Both use GELU activation functions.

Following the 2D stem (or directly for single-channel inputs like amplitude features), a 1D stem consisting of two blocks, each containing a 1D convolutional layer (kernel size 3, 32 filters), batch normalization, GELU activation, and dropout (rate 0.1), is applied. This 1D stem reduces the frequency dimension from its input size to a fixed dimension of 32.

This flexible architecture allows *WiFlexFormer* to handle various input types: amplitude features $[B, 1, F, T]$ (F is the number of subcarriers) and DFS features $[B, C, F, T]$ (C is the number of subcarriers, F is the number of frequency bins). The stem's design replaces heuristic preprocessing steps, providing an end-to-end learnable approach for feature extraction and noise reduction. The temporal receptive field of 5 in the 1D stem helps accumulate information from adjacent positions, enhancing the model's ability to handle noisy inputs.

Positional Encoding and Class Token To encode the temporal dimension, a Gaussian positional encoding is applied to the stem's output, following the method described in [104]. The resulting encoded features are combined with a class token before being input into a four-layer Transformer encoder. This token serves as an aggregator of global information, and its output will be used for the final classification. The use of a class token rather than direct feature aggregation, as in [104], prevents blurring temporal relationships.

Encoder Each layer of the encoder contains 16 attention heads and a feedforward dimension of 64. The final prediction is generated by processing the class token output through a linear classification head. *WiFlexFormer* can be trained end-to-end and remains relatively lightweight, containing only $\approx$ 50k parameters (depending on the input shape). Furthermore, it is designed to work directly with CSI amplitude features and, therefore, does not rely on complicated or slow feature extraction methods.

6.2.2 Evaluation Setup

The WiFi-based PCS capabilities of *WiFlexFormer* are evaluated on two publicly available datasets. These datasets encompass a variety of systems, transmitter-receiver configurations, and recording environments and include both micro- and macroscopic human activities. Additionally, the evaluation covers diverse scenarios, such as LOS and TW sensing. Given that cross-domain generalization remains a significant challenge in WiFi-based sensing [47], our evaluation is structured to measure performance in both in-domain and cross-domain contexts. To provide a comprehensive assessment, *WiFlexFormer* is compared against a range of state-of-the-art vision architectures, as well as architectures specifically designed for processing RF signals, such as WiFi. The evaluation metrics include both HAR performance and inference speed. Finally, to optimize inference speed, various subcarrier sub-sampling strategies using amplitude and DFS features are explored. DFS features are investigated alongside amplitude features since they are a popular choice among existing methods [46, 55] and due to their potential robustness to environmental variations. However, DFS features require more computational resources for extraction.

Data The 3DO dataset is specifically designed to enable controlled studies of model generalization in TW HAR scenarios by isolating environmental variation. It captures three macroscopic activities (*walking*, *sitting*, and *lying*) performed by a single participant over three consecutive days. By holding the participant constant, 3DO removes inter-subject variability and focuses solely on dynamic, static, and temporal domain shifts introduced across days. The WiFi transceivers (system $\mathcal{D}$) remain fixed throughout, operating in a static TW scenario. Day 1 establishes a static in-domain baseline. Day 2 introduces dynamic variation through altered activity execution and natural hardware drift. Day 3 adds static environmental variation by rearranging furniture in the sensing space, resulting in a challenging test domain involving all three sources of domain shift. This design allows for the evaluation of generalization under controlled domain shifts, complementing the cross-subject variability assessed with Widar3.0-G6.

The Widar3.0 WiFi PCS dataset [74] features CSI recordings of 22 human hand gestures performed by 16 participants in three different indoor environments. Since not all 22 gestures are consistently performed in all three environments, a subset of Widar3.0, referred to as Widar3.0-G6 [75], is often utilized instead of the full dataset. This subset includes 6 gestures that are performed across all three environments by 15 users, resulting in a total of 11,250 hand gesture samples. The recording setup, shown in Figure 6.5b, consists of one 5.825 GHz WiFi transmitter (TX) and six receivers (RX_n), each equipped with an *Intel WiFi Link 5300* wireless NIC that has three antennas. The CSI of 90 subcarriers (3 antennas $\times$ 30 subcarriers) is collected at each receiver using the Linux CSI Tool [193], utilizing a packet sending rate of 1,000 Hz.

The 3DO dataset, although based on a single participant, is intentionally chosen to isolate environmental variations from inter-person variability. This controlled setting allows assessing the model's ability to generalize to dynamic and static environmental changes. Complementing this, the Widar3.0-G6 offers cross-subject and cross-receiver evaluation, ensuring that the evaluation captures both controlled and real-world variability.

Model Training *WiFlexFormer* is compared against standard vision architectures and specialized architectures for the processing of RF signals such as WiFi. The vision architectures include *EfficientNetV2s*[190], *ResNet18*[191], and *ShuffleNetV2x0.5*[194], while the RF-specific architectures include *RF-Net*[55], which uses a dual-path architecture processing DFS features in both time and frequency domains with attention-based temporal mechanisms; *SLNet*[46], which combines neural network-based super-resolution spectrogram enhancement with polarized convolution for DFS feature processing; and *THAT*[104], which employs a two-stream (time-over-frequency and frequency-over-time) architecture capturing amplitude-based features with convolution-augmented transformers.

For model training, 3DO and Widar3.0-G6 datasets are utilized, employing both CSI amplitude and DFS features. For the 3DO dataset, a 3:1:1 split on day 1 data is performed for training, validation, and testing to evaluate in-domain performance. Data from days 2 and 3 are reserved for testing cross-domain generalization under dynamic and static environmental variations, respectively. CSI of the 52 L-LTF subcarriers is used, with samples extracted over a window of 351 WiFi packets ($\approx$ 3.51 seconds at a 100 Hz packet sending rate) a duration determined empirically through hyperparameter search.

The Widar3.0-G6 dataset is employed to assess cross-receiver generalization. It is split into two subsets: one containing data from receivers RX_{1-3}, used for training with an 8:2 training-validation split, and the other from receivers RX_{4-6}, reserved for testing. To align with the single-link nature of the 3DO dataset, only the CSI from antenna 1 at each receiver is used, resulting in a selection of 30 subcarriers. Temporal sub-sampling is performed at 100 Hz, with the sampling window length set to 369 packets based on the longest sample length post-sub-sampling, while shorter samples are zero-padded to this length.

Both amplitude and DFS features are extracted from CSI data. DFS features are computed on a per-subcarrier basis using STFT with a Gaussian window and a segment and FFT length of 125 WiFi packets. A frequency band-pass filter from -60 Hz to 60 Hz, as proposed in [46], is applied, resulting in 121 frequency bins. The input shapes for amplitude features are [B, 1, 52, 351] for the 3DO dataset and [B, 1, 30, 369] for the Widar3.0-G6 dataset. For DFS features, the input shapes are [B, 52, 121, 351] and [B, 30, 121, 369] for the 3DO and Widar3.0-G6 datasets, respectively. Furthermore, for *SLNet*, real and imaginary parts of the complex-valued DFS features are stored separately in an additional dimension. The remaining models are fed with the absolute value of the computed DFS features.

For the HAR task, each architecture and feature configuration is trained from scratch in three independent runs with different random seeds, over 10 epochs. The AdamW optimizer [195] with a learning rate of 1×10^{-3} and a weight decay of 1×10^{-3} is used, optimizing for cross-entropy loss. To address class imbalances in the datasets, a balanced random sampler is employed. Training is conducted with a batch size of 32, and no data augmentation is applied, allowing the evaluation of stand-alone generalization capabilities

Model	Inference Time [ms]	
	Amplitude	DFS
RF-Net [55]	-	427.20 ±13.
SLNet [46]	-	322.31 ±7.3
EfficientNetV2s [190]	67.72 ±0.6	68.66 ±0.8
THAT [104]	37.87 ±0.5	-
ShuffleNetV2x0.5 [194]	22.16 ±0.6	22.59 ±0.5
ResNet18 [191]	9.67 ±0.2	12.01 ±0.4
WiFlexFormer [proposed]	9.26 ±0.2	11.06 ±0.6

Table 6.4: Inference time comparison between models for amplitude and DFS features. Inference time is reported as the mean inference time over 1,000 iterations (excluding 100 warm-up iterations) on an Nvidia Jetson Orin Nano using a batch size of 1.

of each architecture. For each run, the best model, with respect to validation loss, is selected for evaluation on the test sets.

6.2.3 Results

Model performance is measured using standard metrics such as recall, precision, F1-score, and accuracy (ACC), computed on the test datasets for each model. To account for variability between runs, the mean and standard deviation of these metrics across three independent training runs are reported, providing a more robust performance measure.

Inference Time To provide context for the HAR performance results, the inference times of all models are first evaluated, reflecting their parameter count and computational efficiency. Inference time is measured using amplitude and DFS features on an Nvidia Jetson Orin Nano single-board computer with 8 GB of VRAM. A batch size of 1 is used, resulting in input shapes of [1, 1, 52, 351] and [1, 52, 121, 351] for amplitude and DFS features, respectively. For each configuration, 100 warm-up iterations are performed preceding the measurement of the mean inference time over 1,000 iterations. The results, presented in Table 6.4, show that *WiFlexFormer* achieves the lowest inference times for both feature types, with a mean inference times of 9.26 ms for amplitude features and 11.06 ms for DFS features. The second-fastest model, *ResNet18*, achieves mean inference times of 9.67 ms for amplitude features and 12.01 ms for DFS features. In comparison, specialized models such as *RF-Net* (427.20 ms for DFS), *SLNet* (322.31 ms for DFS), and *THAT* (37.87 ms for amplitude) exhibit significantly higher inference times due to their larger parameter counts. These findings demonstrate that *WiFlexFormer* provides more efficient inference across both feature types, making it particularly suitable for real-time edge applications where low latency is essential.

HAR Performance on 3DO Table 6.5 presents the in- and cross-domain HAR performance for all models on the 3DO dataset using amplitude features. Day 1 represents in-domain performance, while days 2 and 3 reflect cross-domain performance under dynamic and static environmental variations, respectively.

Model	Params	D	Precision ↑	Recall ↑	F1-Score ↑	ACC ↑
EfficientNetV2s [190]	20.18 M	1	96.64 ±0.2	98.70 ±0.2	97.66 ±0.2	99.11 ±0.1
ResNet18 [191]	11.17 M	1	97.98 ±0.3	99.19 ±0.1	98.58 ±0.2	99.38 ±0.1
THAT [104]	7.96 M	1	97.90 ±0.6	98.01 ±0.6	97.95 ±0.6	98.01 ±0.6
ShuffleNetV2x0.5 [194]	0.34 M	1	96.51 ±0.3	99.03 ±0.1	97.75 ±0.2	99.36 ±0.1
WiFlexFormer [proposed]	0.05 M	1	97.45 ±1.0	98.43 ±0.6	97.94 ±0.8	98.41 ±0.7
EfficientNetV2s [190]	20.18 M	2	77.31 ±1.4	79.79 ±2.1	78.53 ±1.7	80.01 ±2.2
ResNet18 [191]	11.17 M	2	78.79 ±0.3	82.85 ±0.3	80.77 ±0.3	83.07 ±0.3
THAT [104]	7.96 M	2	87.88 ±2.7	88.01 ±2.8	87.95 ±2.8	88.03 ±2.8
ShuffleNetV2x0.5 [194]	0.34 M	2	77.91 ±1.9	80.19 ±2.5	79.03 ±2.2	80.25 ±2.5
WiFlexFormer [proposed]	0.05 M	2	83.37 ±3.1	85.16 ±3.5	84.26 ±3.3	85.26 ±3.6
EfficientNetV2s [190]	20.18 M	3	75.27 ±1.5	77.77 ±2.4	76.49 ±1.9	78.44 ±2.3
ResNet18 [191]	11.17 M	3	76.55 ±0.8	77.94 ±1.1	77.24 ±0.9	78.70 ±1.1
THAT [104]	7.96 M	3	75.60 ±0.3	75.60 ±0.3	75.60 ±0.3	75.61 ±0.3
ShuffleNetV2x0.5 [194]	0.34 M	3	70.20 ±4.4	70.02 ±5.9	70.10 ±5.2	70.49 ±5.9
WiFlexFormer [proposed]	0.05 M	3	85.74 ±2.4	86.47 ±2.9	86.10 ±2.7	86.98 ±2.9

Table 6.5: In- and cross-domain activity recognition performance on the **3DO dataset** using **amplitude features**. Column D indicates the day of data collection. All models are trained on day 1 data with a 3:1:1 training-validation-test split. Amplitude features from all 52 subcarriers are used as input. Results are presented as mean and standard deviation across three independent runs with random initialization.

For in-domain performance (day 1), all models perform similarly, with vision-based models such as *ResNet18* and *ShuffleNetV2x0.5* slightly outperforming specialized models. However, *WiFlexFormer* achieves a competitive accuracy of 98.41%, outperforming the specialized model *THAT*, while using only a fraction of the parameters (0.05 M vs. 7.96 M).

In the cross-domain evaluation on day 2, which introduces dynamic variations, *WiFlexFormer* demonstrates strong generalization capabilities, achieving 85.26% accuracy, second only to *THAT* at 88.03%. In contrast, vision-based models show a noticeable drop in accuracy, with *ResNet18* reaching only 83.07%.

On day 3, which adds the challenge of static environmental variation, *WiFlexFormer* outperforms all other models with an accuracy of 86.98%, highlighting its robustness in challenging cross-domain scenarios. Notably, *THAT* experiences a significant drop in accuracy to 75.61%, while vision-based models like *ResNet18* struggle to maintain performance, achieving only 78.70%.

Overall, considering both in-domain and cross-domain performance, *WiFlexFormer*, using amplitude features, emerges as the best-performing model, offering superior generalization at a dramatically lower parameter count (0.05 M) compared to models like *EfficientNetV2s* (20.18 M) and *ResNet18* (11.17 M), making it highly efficient for real-world WiFi-based HAR applications.

Model	Params	D	Precision ↑	Recall ↑	F1-Score ↑	ACC ↑
RF-Net [55]	349.57 M	1	87.46 ±0.9	88.18 ±1.0	87.82 ±1.0	88.19 ±1.0
SLNet [46]	146.27 M	1	85.04 ±4.8	87.53 ±4.3	86.27 ±4.6	87.50 ±4.3
EfficientNetV2s [190]	20.19 M	1	97.33 ±0.1	97.66 ±0.0	97.50 ±0.1	97.67 ±0.0
ResNet18 [191]	11.33 M	1	94.60 ±0.9	94.94 ±1.0	94.77 ±0.9	94.92 ±1.0
ShuffleNetV2x0.5 [194]	0.36 M	1	97.24 ±0.5	97.50 ±0.4	97.37 ±0.5	97.48 ±0.4
WiFlexFormer [proposed]	0.06 M	1	85.49 ±1.4	92.70 ±0.4	88.95 ±0.9	92.83 ±0.4
RF-Net [55]	349.57 M	2	62.62 ±2.1	62.89 ±2.1	62.75 ±2.1	62.89 ±2.1
SLNet [46]	146.27 M	2	62.94 ±12.	63.84 ±12.	63.39 ±12.	63.87 ±12.
EfficientNetV2s [190]	20.19 M	2	86.42 ±3.4	86.58 ±3.3	86.50 ±3.4	86.58 ±3.3
ResNet18 [191]	11.33 M	2	75.03 ±0.0	75.15 ±0.1	75.09 ±0.1	75.16 ±0.1
ShuffleNetV2x0.5 [194]	0.36 M	2	71.11 ±1.2	71.20 ±1.2	71.15 ±1.2	71.21 ±1.2
WiFlexFormer [proposed]	0.06 M	2	75.59 ±1.4	79.77 ±1.4	77.62 ±1.4	79.91 ±1.4
RF-Net [55]	349.57 M	3	59.69 ±1.6	59.57 ±1.6	59.63 ±1.6	59.58 ±1.6
SLNet [46]	146.27 M	3	75.53 ±7.2	76.88 ±7.8	76.20 ±7.5	76.91 ±7.8
EfficientNetV2s [190]	20.19 M	3	73.26 ±4.4	73.61 ±4.5	73.44 ±4.5	73.62 ±4.5
ResNet18 [191]	11.33 M	3	69.26 ±6.9	69.30 ±7.1	69.28 ±7.0	69.30 ±7.1
ShuffleNetV2x0.5 [194]	0.36 M	3	71.35 ±3.4	71.47 ±3.5	71.41 ±3.5	71.47 ±3.5
WiFlexFormer [proposed]	0.06 M	3	71.10 ±2.3	73.85 ±3.1	72.45 ±2.7	74.18 ±3.2

Table 6.6: In- and cross-domain activity recognition performance on the **3DO dataset** using **DFS features**. Column D indicates the day of data collection. All models are trained on day 1 data with a 3:1:1 training-validation-test split. Amplitude features from all 52 subcarriers are used as input. Results are presented as mean and standard deviation across three independent runs with random initialization.

Table 6.6 shows the in- and cross-domain HAR performance for all models on the 3DO dataset using DFS features. While *WiFlexFormer* is outperformed by larger vision models such as *EfficientNetV2s* and *ShuffleNetV2x0.5*, this is expected due to their high parameter count, which allows them to make better use of the dense DFS features across all subcarriers. In contrast, *WiFlexFormer* prioritizes strong feature compression in its design to remain computationally efficient, which limits its ability to leverage the full richness of DFS data. Despite this, *WiFlexFormer* delivers a competitive in-domain accuracy of 92.83%.

In cross-domain evaluations, especially on day 2, *WiFlexFormer* demonstrates reasonable generalization, achieving 79.91% accuracy, outperforming larger models such as *ResNet18* and *ShuffleNetV2x0.5*, and coming close to *EfficientNetV2s*. On day 3, which introduces static environmental variations, *WiFlexFormer* continues to show stability, with an accuracy of 74.18%, higher than *ResNet18* and comparable to other models, while *SLNet* and *EfficientNetV2s* exhibit a larger drop in performance.

One notable observation is the high run-to-run variance exhibited by the larger models, such as *SLNet*, across all days, indicating instability during training, especially when faced with out-of-distribution samples. In contrast, *WiFlexFormer* consistently shows lower variance, suggesting that its low parameter count may provide a natural regularization

Model	Params	Precision ↑	Recall ↑	F1-Score ↑	ACC ↑
EfficientNetV2s [190]	20.18 M	52.26 ±2.0	52.08 ±0.5	52.16 ±1.2	51.98 ±0.5
ResNet18 [191]	11.17 M	51.26 ±2.1	51.50 ±1.9	51.38 ±2.0	51.38 ±1.8
THAT [104]	8.48 M	50.60 ±1.3	49.96 ±0.6	50.26 ±0.6	49.84 ±0.6
ShuffleNetV2x0.5 [194]	0.35 M	52.26 ±0.6	51.89 ±0.8	52.07 ±0.7	51.74 ±0.8
WiFlexFormer [proposed]	0.04 M	49.29 ±0.3	49.59 ±0.4	49.44 ±0.4	49.38 ±0.4

Table 6.7: Cross-receiver gesture recognition performance on the **Widar3.0-G6 dataset** using **amplitude features**. All models are trained on data from receivers 1-3 with an 8:2 training-validation split and tested on receivers 4-6. Amplitude features from all 30 subcarriers are used as input. Results are presented as mean and standard deviation across three independent runs with random initialization.

effect, leading to more stable training and performance across different domains.

Interestingly, none of the models, including the high-parameter models, are able to fully leverage DFS features for HAR, as the overall performance with DFS is notably lower than with amplitude features. This suggests that DFS features may not be optimal for this TW sensing scenario. The complex signal scattering and the highly noisy phase information involved in DFS computation likely contribute to the lower performance.

From an efficiency perspective, this outcome is promising: amplitude features, which have a lower dimensionality and do not require computationally expensive preprocessing, outperform DFS features requiring computationally expensive preprocessing, such as STFT, across all models. For *WiFlexFormer*, this is particularly advantageous, as it achieves better HAR performance with simpler, faster-to-process amplitude features, making it an ideal solution for WiFi-based real-time sensing applications.

HAR Performance on Widar3.0-G6 Table 6.7 presents the cross-receiver gesture recognition performance using amplitude features on the Widar3.0-G6 dataset. Overall, all models perform similarly, with *EfficientNetV2s* achieving the highest accuracy of 51.98%. Vision-based models generally outperform specialized models such as *THAT* and *WiFlexFormer*, but the performance differences remain small. For instance, *WiFlexFormer* trails *EfficientNetV2s* by only 2.6% accuracy, despite the latter having a 500x larger parameter count, demonstrating that *WiFlexFormer* offers competitive accuracy with a much smaller model.

The results using DFS features, as shown in Table 6.8, tell a similar story. *EfficientNetV2s* again achieves the highest accuracy with 51.34%, outperforming specialized models like *RF-Net* (48.11%) and *SLNet* (50.63%). *WiFlexFormer* delivers a competitive accuracy of 49.72%, trailing *EfficientNetV2s* by only 1.62%. Notably, *WiFlexFormer* outperforms *RF-Net* by 1.61% and comes close to *SLNet*, despite their significantly larger parameter counts.

Overall, the performance across models on the Widar3.0-G6 dataset is quite close, consequently, there is no significant advantage in using DFS features over amplitude

Model	Params	Precision ↑	Recall ↑	F1-Score ↑	ACC ↑
RF-Net [55]	120.58 M	48.90 ±1.2	48.11 ±0.4	48.49 ±0.5	48.11 ±0.4
SLNet [46]	88.88 M	50.63 ±0.5	50.76 ±0.1	50.70 ±0.2	50.63 ±0.0
EfficientNetV2s [190]	20.19 M	49.87 ±0.0	51.52 ±0.2	50.68 ±0.1	51.34 ±0.2
ResNet18 [191]	11.26 M	49.92 ±0.3	51.00 ±0.3	50.45 ±0.3	50.84 ±0.2
ShuffleNetV2x0.5 [194]	0.35 M	49.76 ±0.5	50.13 ±0.4	49.94 ±0.4	50.08 ±0.4
WiFlexFormer [proposed]	0.05 M	49.36 ±0.8	49.82 ±0.2	49.59 ±0.5	49.72 ±0.1

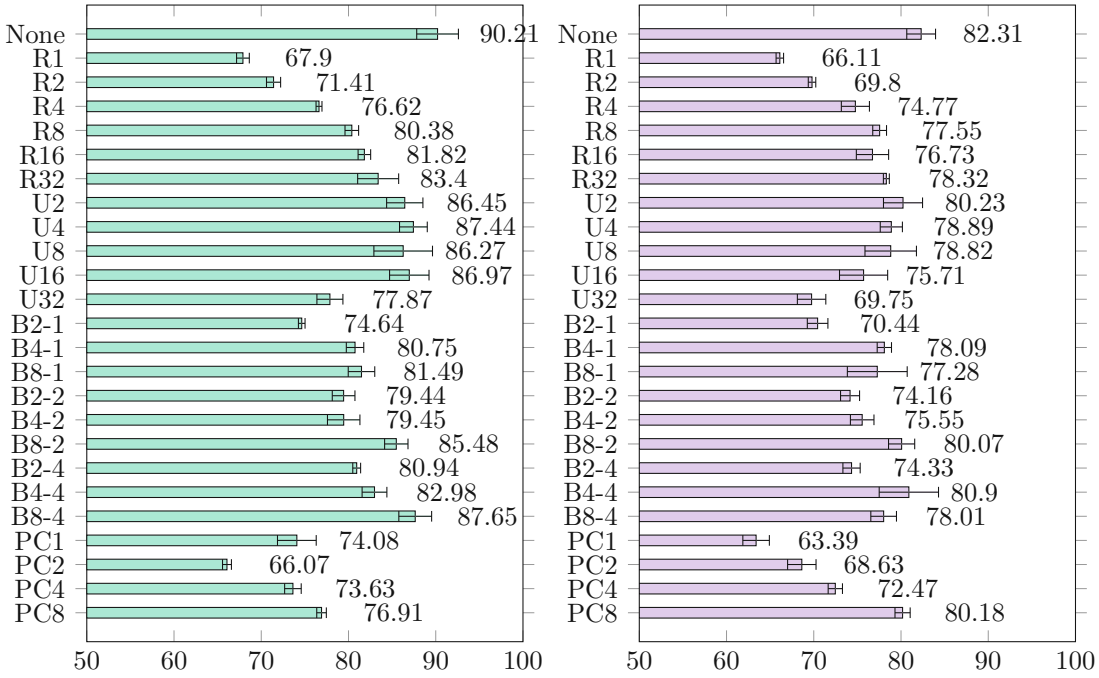
Table 6.8: Cross-receiver gesture recognition performance on the **Widar3.0-G6 dataset** using **DFS features**. All models are trained on data from receivers 1-3 with an 8:2 training-validation split and tested on receivers 4-6. Amplitude features from all 30 subcarriers are used as input. Results are presented as mean and standard deviation across three independent runs with random initialization.

features in terms of cross-receiver generalization. Given that amplitude features require no computationally expensive preprocessing or parameter tuning, they remain a more efficient option for cross-domain generalization tasks.

Subcarrier Selection Prior work on subcarrier selection primarily focuses on LOS scenarios, leaving open questions about their performance in more challenging signal propagation conditions. Evaluations on the 3DO dataset [50] indicate that insights from LOS conditions may not easily translate to TW scenarios. To address this gap, subcarrier sub-sampling strategies that reduce computational complexity while maintaining model accuracy in a TW scenario are explored. Processing CSI-based features from all subcarriers incurs high computational costs, particularly for DFS features, which require per-subcarrier STFT preprocessing.

As shown in [75], considering all subcarriers is neither efficient nor necessary given the high correlation among them. To reduce computational overhead while preserving accuracy, several subcarrier sub-sampling methods, including random sampling, band-restricted random sampling, uniform sampling, and projection-based sampling via PCA, are evaluated. The results for amplitude and DFS features, presented in Figures 6.6a and 6.6b, report mean accuracy and standard deviation across days 1–3, capturing both in-domain and cross-domain HAR performance.

For amplitude features, the highest accuracy is achieved using all subcarriers (None). Although uniform sampling of every 4th subcarrier (U4) and band-restricted sampling using eight bands with four subcarriers per band (B8-4) yield comparable results, the reduction in preprocessing and inference time for amplitude features is negligible, making sub-sampling unnecessary. For DFS features, the best accuracy is also obtained using all subcarriers, which is to be expected. However, subcarrier sub-sampling strategies, such as uniform sampling of every 2nd subcarrier (U2) or band-restricted sub-sampling with four bands and four subcarriers per band (B4-4), achieve similar accuracy while reducing the number of STFT computations to half and one-fourth, respectively. These strategies are a potential way to further reduce inference time, especially when using DFS features.



(a) Accuracy (mean ± std across days 1-3) (b) Accuracy (mean ± std across days 1-3)

Figure 6.6: Comparative analysis of subcarrier selection strategies for (a) **amplitude features**, and (b) **DFS features** using the **3DO dataset**. Strategies include: (1) None: use of all subcarriers; (2) Rn: random selection of n subcarriers; (3) Un: uniform selection of every nth subcarrier; (4) Bn-m: division into n subcarrier bands with random selection of m subcarriers from each band; and (5) PCn: selection of the first n principal components. [53]

6.2.4 Discussion

The evaluation results of *WiFlexFormer* confirm its suitability for real-time PCS on low-power edge devices, addressing the core aspects of **RQ II**. The architecture's lightweight design, with only ≈ 50k parameters, achieves inference times of ≈ 10 milliseconds on an Nvidia Jetson Orin Nano. This significant reduction in computational complexity relative to traditional CNN-based [190, 191, 194] and specialized RF/WiFi architectures [55, 104, 46], renders *WiFlexFormer* uniquely suitable for embedded contexts requiring real-time inference, such as gesture recognition or continuous health monitoring applications [45].

The compact and computationally efficient nature of *WiFlexFormer* also naturally supports advanced deployment strategies, such as test-time training [196] or adaptation [54]. The architecture's rapid inference capability facilitates quick, online fine-tuning to adapt to new domains and dynamic environmental conditions, critical in scenarios like smart homes where environmental contexts frequently change. Furthermore, Although this evaluation primarily benchmarks relative performance, there remains considerable potential to further enhance the accuracy of *WiFlexFormer*, pre-training, data augmentation strate-

gies [48, 50], or ensemble methods [197] (feasible due to the low parameter count). These strategies, could enhance accuracy and and cross-domain generalization performance without substantially increasing inference latency.

Lastly, the systematic exploration of subcarrier sub-sampling techniques identifies additional pathways to reduce computational overhead. Specifically, uniform and band-restricted random sub-sampling demonstrate potential for further complexity reduction, highlighting additional optimizations for resource-constrained edge deployments. Collectively, these results establish *WiFlexFormer* as a robust and highly efficient architecture that precisely meets the demands outlined in RQ II, thereby advancing practical, scalable, and real-time CSI-based person-centric sensing systems.

6.3 Cross-Domain Generalization

While recent advances in WiFi-based PCS demonstrate strong performance in controlled environments, deploying these systems effectively in real-world scenarios remains challenging due to limited cross-domain generalization. Even minor environmental changes, differences in hardware setups, or variations in user behaviors can introduce significant shifts in CSI distributions, leading to substantial performance degradation when models are deployed in previously unseen domains [47]. Such domain sensitivity severely restricts the scalability, robustness, and practical deployment of existing WiFi-based PCS approaches, highlighting the necessity of addressing domain variations explicitly.

Unlike optical sensing methods, which benefit from established benchmarks and standardized evaluation procedures for assessing generalization, WiFi-based sensing lacks equivalent systematic methodologies and dedicated public datasets. Consequently, strategies specifically designed to enhance model robustness across different scenarios, system configurations, and environmental conditions are required. **RQ III** addresses this challenge by exploring how WiFi-based PCS models can achieve robustness against real-world domain variations, thus enabling practical, scalable, and reliable deployments.

To comprehensively address the challenges posed by domain variability, three complementary works are presented, each contributing uniquely toward answering **RQ III**:

The first work [48], discussed in Section 6.3.1, investigates data-centric approaches, specifically assessing the effectiveness of image-based data augmentation techniques for enhancing generalization. Leveraging the Wallhack1.8k dataset, it systematically evaluates virtual sample generation strategies for two critical and underexplored generalization problems: cross-scenario (LOS versus TW) and cross-system (system $\mathcal{C}_1$ versus system $\mathcal{B}$) generalization.

Building upon this evaluation, the second work [50], discussed in Section 6.3.2, adopts a controlled TW sensing scenario using the 3DO dataset. It systematically analyzes a variety of preprocessing techniques, including feature selection, feature scaling, dimensionality reduction, and data augmentation, to quantify their effectiveness in mitigating environmental and temporal domain shifts in TW scenarios. This structured evaluation provides detailed insights into the relative strengths and limitations of different generalization techniques.

Finally, the third work [54], discussed in Section 6.3.3, contributes Domain-Adversarial Test-Time Adaptation (DATTA), a novel framework that integrates domain-adversarial training and test-time adaptation. By combining these methods, DATTA explicitly targets domain-invariant feature learning during training while adapting dynamically at inference, effectively bridging the gap between offline training and real-world deployment conditions, where new target-domain samples are typically unavailable beforehand.

6.3.1 Data Augmentation for Cross-System and Cross-Scenario Generalization

Robust cross-domain generalization remains a critical challenge in WiFi-based PCS. Existing generalization approaches include domain-invariant feature extraction, transfer learning, domain adaptation, big data, and virtual sample generation [47]. Among these, virtual sample generation is appealing due to its simplicity, computational efficiency, and independence from target-domain data [121]. For instance, prior studies employ noise injection and dropout-based methods to simulate realistic perturbations [120, 49], or leverage generative models such as VAEs and GANs for richer but computationally intensive data synthesis [122, 123, 124]. However, most existing works either require access to both source and target domains, which is impractical in real-world scenarios, or do not explicitly address the unique challenges posed by cross-scenario (LOS vs. TW) or cross-system (hardware variation) generalization [47].

Addressing this critical research gap, the underlying work [48], systematically explores simple yet effective image-based augmentation strategies tailored explicitly for CSI amplitude spectrograms. Leveraging the Wallhack1.8k dataset, designed to enable controlled comparisons across scenarios (LOS and TW) and distinct WiFi systems (system $\mathcal{C}_1$ and $\mathcal{B}$), an ablation study is conducted to assess how individual augmentation techniques and their combinations influence model robustness to domain variations. By quantifying the impact of each technique, this study provides a practical foundation for developing simple domain-agnostic augmentation strategies for CSI data, thus addressing the real-world applicability concerns highlighted by **RQ III**.

CSI Data Augmentations

Motivated by the effectiveness of common image-based data augmentation techniques for enhancing a model's robustness to domain variations [198], this evaluation systematically investigates their effectiveness when applied directly to CSI amplitude spectrograms. Specifically, the augmentations *randomCircularRotation*, *randomResizedCrop*, *randomAmplitude*, and *randomContrast*, defined in the following, are evaluated individually and in combination, aiming to assess their impact on cross-scenario and cross-system generalization. Each augmentation targets specific characteristics of CSI data, simulating realistic temporal or amplitude variations on a subcarrier and channel-level, known to occur in practice [125]. Augmentation are applied probabilistically ($p = 0.5$) to maintain data diversity without overwhelming the original training signals.

randomCircularRotation Treating the spectrogram as a 2D time-frequency array, circular rotations along the time axis shift the array elements along the positive time direction, with elements going beyond the array's boundary wrapping around to the opposite side, becoming the first element. A random number of circular rotations $n \sim \mathcal{U}(1, w)$, with $w = 400$ being the spectrogram width, is applied to the spectrogram.

randomResizedCrop For this augmentation, the spectrogram is either cropped along the time axis and stretched back to its original width w or compressed along the time

axis and re-scaled to w. Both the crop and the compression factor are randomly sampled from $\mathcal{U}(\frac{w}{2}, w)$. The cropping and re-scaling of the spectrogram correspond to a slowdown in time, while the compression and re-scaling correspond to a speed-up, resulting in new samples with activities being carried out at varying speeds.

randomAmplitude As observed in [125], CSI amplitude can vary significantly even in static environments and system settings. To replicate this behavior, the amplitude of spectrograms is scaled randomly on a per-channel basis. The magnitude of the augmentation factor is randomly sampled from $\mathcal{U}(0.75, 1.25)$.

randomContrast Furthermore, it is also found that CSI amplitude variation occurs differently depending on the subcarrier index [125]. Following the approach of the *randomContrast* augmentation, this behavior is replicated by randomly scaling the contrast of spectrograms on a per-subcarrier basis. The magnitude of the augmentation factor is randomly sampled from $\mathcal{U}(0.75, 1.25)$.

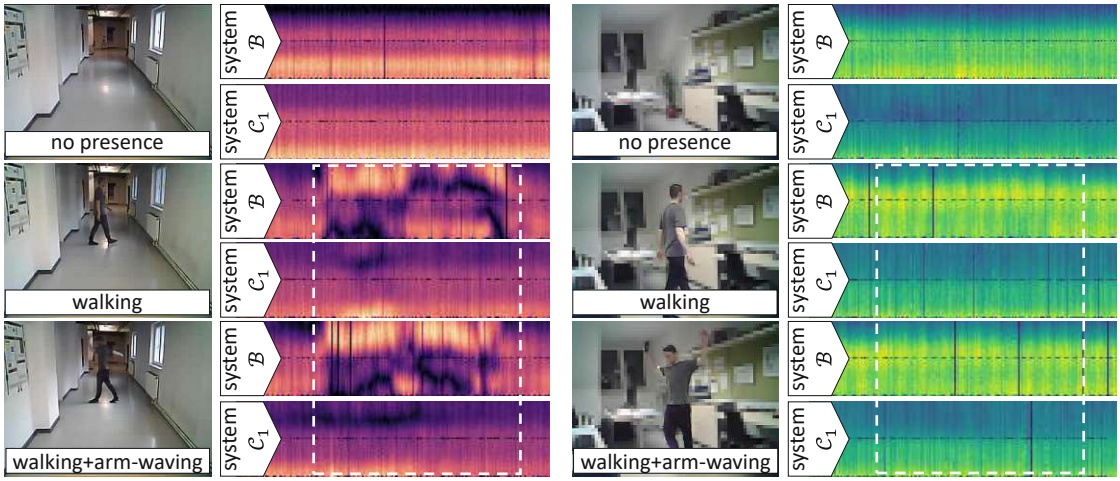
Evaluation Setup

To evaluate the impact of data augmentation techniques on cross-scenario (LOS $\leftrightarrows$ TW) and cross-system ($\mathcal{C}_1 \leftrightarrows \mathcal{B}$) generalization performance, an ablation study is performed. The study begins with a baseline model trained without any augmentation. Subsequently, separate models are trained with individual augmentation strategies, and the resulting changes in accuracy are measured. Augmentations that yield performance improvements over the baseline are then combined to train a final model, which is also evaluated in comparison to the baseline configuration.

Data The evaluation utilizes the Wallhack1.8k dataset to systematically evaluate model generalization across heterogeneous scenarios and WiFi systems. The dataset captures three human activities (*no presence*, *walking*, *walking + arm-waving*) in both LOS and TW signal propagation scenarios using two distinct WiFi systems: $\mathcal{C}_1$ and $\mathcal{B}$. To facilitate a structured evaluation, the established HAR subset (W_A$_{L\mathcal{B}}$, W_A$_{L\mathcal{C}_1}$, W_A$_{T\mathcal{B}}$, W_A$_{T\mathcal{C}_1}$), detailed in Table 6.1, are utilized for training and testing.

The Wallhack1.8k dataset is particularly well-suited for studying cross-domain generalization. As illustrated in Figure 6.7, the amplitude spectrograms reveal clear differences in CSI signal characteristics across scenarios and systems. LOS recordings with system $\mathcal{B}$, for example, exhibit strongly pronounced activity-induced amplitude variations due to LOS path obstruction, while these effects are diminished in TW scenarios and less visible in recordings from system $\mathcal{C}_1$. These signal variations present meaningful generalization challenges, making Wallhack1.8k an ideal testbed for evaluating the impact of data augmentation on model robustness under domain shifts.

Model Training To evaluate cross-scenario and cross-system generalization, the experiments rely on the *EfficientNetV2 small* architecture [190], a lightweight feature extractor commonly used for classification tasks. All models are trained from scratch to eliminate any influence from prior knowledge, such as pre-training on ImageNet. The input to each



(a) LOS scenario (b) TW scenario

Figure 6.7: Visual comparison of signal characteristics between scenarios and systems ($\mathcal{B}$ and $\mathcal{C}_1$) on the Wallhack1.8k dataset. Highlighted areas (dotted rectangle) show the varying signal characteristics between scenarios and systems. [48] †

model consists of CSI amplitude spectrograms constructed from 52 L-LTF subcarriers over 400 consecutive WiFi packets, corresponding to ≈ 4-second intervals. Training is conducted for 400 epochs using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 16. To address class imbalance, a balanced random sampler is applied. Each configuration is independently trained ten times with random initialization, retaining the model instance that achieves the highest validation accuracy in each run. Final results are reported as the mean and standard deviation of classification accuracies across the ten selected models.

Results

Cross-Scenario Generalization (LOS ⇆ TW) Table 6.9 summarizes the results of the ablation study conducted using data collected with system $\mathcal{C}_1$. In the LOS → TW direction, a baseline accuracy of 37.5% is achieved. Among the individual data augmentation strategies, *randomCircularRotation* and *randomResizedCrop* contribute to noticeable accuracy improvements, increasing accuracy by 6.1 and 12.0 percentage points, respectively. However, when both augmentations are applied in combination, the model's accuracy decreases slightly by 1.1 percentage points relative to the baseline, indicating a potentially conflicting interaction between the two augmentations.

In contrast, the reverse direction (TW → LOS) yields a substantially higher baseline accuracy of 65.0%, suggesting that generalization from TW to LOS is more tractable than the inverse. In this setting, only the *randomAmplitude* augmentation produces a measurable gain, improving accuracy by 1.1 percentage points. Other augmentations fail to yield further improvement, highlighting the asymmetric nature of generalization between these signal propagation scenarios.

Augmentation	LOS → TW	change	TW → LOS	change
none (baseline)	37.5 ±5.3	-	65.0 ±5.5	-
randomCircularRotation	43.6 ±9.8	+6.1	61.3 ±6.4	-3.7
randomResizedCrop	49.5 ±11.4	+12.0	54.8 ±8.4	-10.2
randomAmplitude	34.1 ±7.5	-3.4	66.1 ±4.8	+1.1
randomContrast	32.5 ±3.2	-5.0	63.9 ±5.9	-1.1
combined	36.4 ±7.1	-1.1	66.1 ±4.8	+1.1

Table 6.9: Cross-scenario generalization performance of HAR models, trained on amplitude spectrograms collected with system $\mathcal{C}_1$ (mean±std accuracy over ten runs). X → Y indicates model training on data from scenario X and testing on data from scenario Y.

Augmentation	LOS → TW	change	TW → LOS	change
none (baseline)	36.4 ±5.1	-	53.0 ±10.1	-
randomCircularRotation	39.6 ±9.6	+3.2	69.6 ±3.2	+16.6
randomResizedCrop	41.8 ±7.4	+5.4	63.3 ±6.9	+10.3
randomAmplitude	41.1 ±5.2	+4.7	61.7 ±6.2	+8.7
randomContrast	40.2 ±5.3	+3.8	57.0 ±8.1	+4.0
combined	58.9 ±7.7	+22.5	68.7 ±2.9	+15.7

Table 6.10: Cross-scenario generalization performance of HAR models (accuracy), trained on amplitude spectrograms collected with system $\mathcal{B}$ (mean±std accuracy over ten runs). X → Y indicates model training on data from scenario X and testing on data from scenario Y.

Ablation results for system $\mathcal{B}$, reported in Table 6.10, follow similar trends but exhibit distinct performance dynamics. When generalizing from LOS to TW, the baseline model achieves an accuracy of 36.4%. All individual augmentation strategies lead to modest accuracy improvements, and combining all augmentations results in a notable gain of 22.5 percentage points over the baseline. This indicates a strong cumulative effect of the applied augmentations in this challenging direction.

In the reverse direction (TW → LOS), all augmentation techniques continue to yield positive impacts. Notably, the *randomCircularRotation* augmentation stands out, increasing accuracy by 16.6 percentage points. However, combining all augmentations in this setting does not provide additional benefit. Instead, the model achieves a slightly reduced improvement of 15.7 percentage points over the baseline. This outcome reflects the observation made for system $\mathcal{C}_1$, where augmentation combinations do not necessarily yield additive gains and may introduce interference between augmentations.

Cross-System Generalization ($\mathcal{C}_1 \leftrightarrows \mathcal{B}$) Table 6.11 summarizes the ablation study results for cross-system generalization in the LOS scenario. When models trained on data from system $\mathcal{C}_1$ are tested on system $\mathcal{B}$ ($\mathcal{C}_1 \rightarrow \mathcal{B}$), all applied data augmentations lead to

Augmentation	$\mathcal{C}_1 \rightarrow \mathcal{B}$	change	$\mathcal{B} \rightarrow \mathcal{C}_1$	change
none (baseline)	36.1 ±6.0	-	35.2 ±4.7	-
randomCircularRotation	48.7 ±13.5	+12.6	38.9 ±2.5	+3.7
randomResizedCrop	38.7 ±6.7	+2.6	36.5 ±3.0	+1.3
randomAmplitude	36.5 ±4.1	+0.4	35.2 ±4.9	0.0
randomContrast	36.5 ±4.8	+0.4	32.4 ±3.9	-2.8
combined	57.2 ±15.7	+21.1	50.7 ±11.8	+18.5

Table 6.11: Cross-system generalization performance of HAR models, trained on LOS amplitude spectrograms (mean±std accuracy over ten runs). X → Y indicates model training on data from system X and testing on data from system Y.

Augmentation	$\mathcal{C}_1 \rightarrow \mathcal{B}$	change	$\mathcal{B} \rightarrow \mathcal{C}_1$	change
none (baseline)	34.7 ±6.5	-	30.0 ±3.9	-
randomCircularRotation	39.6 ±6.8	+4.9	34.5 ±5.7	+4.5
randomResizedCrop	39.3 ±3.2	+4.3	31.4 ±10.5	+1.4
randomAmplitude	36.4 ±6.1	+1.7	33.0 ±4.8	+3.0
randomContrast	33.8 ±6.6	-0.9	33.6 ±4.4	+3.6
combined	39.6 ±8.2	+4.9	52.3 ±13.8	+22.3

Table 6.12: Cross-system generalization performance of HAR models, trained on TW amplitude spectrograms (mean±std accuracy over ten runs). X → Y indicates model training on data from system X and testing on data from system Y.

improvements over the baseline accuracy of 36.1%. Among them, *randomCircularRotation* yields the largest increase, enhancing performance by 12.6 percentage points. Furthermore, combining all augmentations results in a substantial improvement of 21.1 percentage points compared to the baseline, indicating a strong cumulative benefit.

In the reverse direction ($\mathcal{B} \rightarrow \mathcal{C}_1$), a similar trend is observable. While individual augmentations produce moderate gains in accuracy, the combined application of all augmentation techniques leads to a pronounced improvement of 18.5 percentage points over the baseline. These results suggest that, in LOS scenarios, data augmentation strategies can consistently enhance cross-system generalization in both directions.

Table 6.12 presents the corresponding results for the TW scenario. When transferring from system $\mathcal{C}_1$ to system $\mathcal{B}$, most augmentation techniques yield improvements over the baseline accuracy of 34.7%, with the exception of *randomContrast*. However, combining the beneficial augmentations does not produce a cumulative gain; the overall accuracy improvement remains at 4.9 percentage points, matching the result obtained using *randomCircularRotation* alone. This finding indicates limited complementarity among the augmentations in this direction.

A clearer and more consistent improvement is observed in the opposite direction ($\mathcal{B} \rightarrow \mathcal{C}_1$). All individual augmentation strategies result in moderate accuracy increases relative to the baseline of 30.0%, and their combination leads to a substantial improvement of 22.3 percentage points. These results confirm that, especially in the TW setting, carefully selected data augmentation strategies can enhance model robustness across heterogeneous hardware systems.

Discussion

The conducted evaluation reveals notable asymmetries in cross-domain generalization. Specifically, models consistently exhibit higher baseline accuracy and better generalization when transitioning from TW to LOS scenarios compared to the reverse. This phenomenon is evidenced by baseline accuracies notably above random chance (65.0% and 53.0%) in the TW-to-LOS generalization for systems $\mathcal{C}_1$ and $\mathcal{B}$, respectively. A likely explanation is that TW scenarios inherently contain richer propagation variability, enabling models to generalize more effectively toward less complex LOS environments.

Moreover, system-specific differences emerge clearly. System $\mathcal{B}$, characterized by higher-sensitivity directional antennas and gain, shows more favorable responsiveness to data augmentations, indicating a narrower domain gap between LOS and TW conditions compared to system $\mathcal{C}_1$. This result underscores the importance of hardware characteristics in determining model robustness to environmental variations.

In the cross-system evaluations, domain gaps are smaller in LOS scenarios, as indicated by similar baseline accuracies and consistent augmentation-induced improvements across systems. In contrast, significant asymmetry arises in TW scenarios, where models trained on system $\mathcal{B}$ data generalize substantially better to system $\mathcal{C}_1$ than vice versa (improvements of 22.3 vs. 4.9 percentage points). This pattern suggests indicates greater difficulty in adapting models trained on system $\mathcal{C}_1$ to more sensitive or higher-fidelity CSI captured by system $\mathcal{B}$.

Overall, these results highlight that achieving effective cross-scenario and cross-system generalization in WiFi-based PCS through data augmentation requires carefully selecting and combining domain-appropriate augmentation techniques. While no single technique universally enhances robustness, specific augmentation combinations improve performance in targeted cross-domain shifts. In regard to **RQ III**, this evaluation provides clear evidence that carefully chosen, lightweight data augmentation methods represent a promising, practically feasible approach for improving domain robustness in WiFi-based PCS deployments without requiring prior access to target-domain data.

6.3.2 Mitigating Environmental and Temporal Domain Shifts in Through-Wall Scenarios

Real-world deployment of WiFi-based PCS systems faces significant challenges due to subtle yet impactful domain shifts, including dynamic variations in human behavior, static environmental changes (e.g., rearranged furniture), and temporal fluctuations arising from environmental or hardware instabilities [125]. While domain shifts due to hardware differences or scenario transitions (e.g., LOS to TW) are addressed via data augmentation techniques in [48], environmental and temporal domain shifts remain less understood and more difficult to model explicitly, as they occur unpredictably and are challenging to replicate experimentally. Particularly in TW sensing scenarios, where signal propagation is heavily affected by attenuation and multipath interference from intervening structures [199], such subtle shifts can substantially degrade model performance. Although recent efforts address the cross-domain generalization dynamics of WiFi-based PCS models, most studies focus on LOS or same-room conditions [47], leaving TW generalization severely underexplored.

Addressing this critical gap, [50] systematically evaluates the effectiveness of preprocessing techniques, namely, feature extraction, feature scaling, dimensionality reduction, and data augmentation, in mitigating environmental and temporal domain shifts explicitly within TW scenarios. Leveraging the 3DO dataset (see Section 5.2.4), which captures synchronized CSI and 3D trajectory data over three consecutive days with controlled static, dynamic, and temporal domain variations, this evaluation isolates specific domain shifts relevant to practical PCS deployments. By evaluating these techniques across both HAR and localization tasks, this study complements [48], and contributes to **RQ III** by identifying preprocessing techniques which can further enhance model robustness to subtle yet realistic domain shifts in real-world TW scenarios.

Methods

This study considers a range of techniques aimed at improving model generalization in TW scenarios. The analysis begins with a comparison of CSI-based feature representations, including amplitude, phase, first-order differences, and PSD, to determine which features most effectively support model robustness. Subsequently, feature scaling approaches such as max-min scaling and z-normalization are examined, followed by dimensionality reduction techniques including PCA, ICA, and UMAP. Finally, the evaluation includes perturbation-based data augmentation methods applied to CSI amplitude spectrograms in the image domain. Each method is described in more detail in the following sections.

CSI Feature Extraction In the domain of WiFi-based PCS, the primary features extracted from CSI are the amplitude and phase. As domain shifts impact these features directly, models trained on such data may face generalization problems across unseen domains. To address this, *first-order difference* (temporal difference) features based on amplitude or phase are proposed in [39, 72], capturing the change between consecutive time steps (WiFi packets) rather than absolute values, thus potentially enhancing domain

generalization. First-order difference features are defined as follows:

$$h_{\Delta}[t] = h[t] - h[t-1]. \tag{6.1}$$

Applying Equation 6.1 to the CSI time series h within the CSI matrix, defined in Equation 2.7, yields the first-order difference time series h_{Δ}, from which $\mathcal{H}_{\Delta}[t]$, and subsequently the first-order difference amplitude $\mathcal{A}_{\Delta}[t]$ and phase matrices $\mathcal{P}_{\Delta}[t]$, can be extracted.

Power Spectral Density (PSD) [73] offers another alternative for feature extraction. Unlike time-domain features, PSD quantifies the signal's power distribution over frequency, emphasizing periodic components such as repetitive motion patterns while attenuating high-frequency noise and low-frequency drifts. By applying the FFT to a time window of CSI measurements, it captures the dominant frequency components caused by human activities, which are typically less sensitive to static environmental variations. The PSD is computed for each subcarrier across a window size w, resulting in the PSD matrix $\mathcal{PSD}[t]$:

$$h_{PSD}[t] = \frac{|\mathrm{FFT}(h[t])|^2}{w}. \tag{6.2}$$

Feature Scaling Two standard feature scaling techniques for improving model robustness in CSI-based sensing tasks are considered: *max-min scaling* and *z-normalization*. Max-min scaling, also known as min-max normalization, transforms input features to a fixed range, typically between 0 and 1, by subtracting the minimum value and dividing by the feature range. This normalization ensures that all input features contribute equally during learning and can support improved convergence in gradient-based optimization. In the context of CSI, max-min scaling is applied to the feature matrix $F[t]$ as defined in Equation 6.3, where min_F and max_F denote the minimum and maximum values in F, respectively:

$$F[t]' = \frac{F[t] - min_F}{max_F - min_F} \tag{6.3}$$

Z-normalization, also referred to as standardization, scales features to have zero mean and unit variance by subtracting the mean and dividing by the standard deviation. This technique is particularly effective for models sensitive to input variance and for algorithms assuming normally distributed data. Z-normalization is applied to the feature matrix $F[t]$ as shown in Equation 6.4, where μ_F and σ_F represent the mean and standard deviation of the feature matrix:

$$F[t]' = \frac{F[t] - \mu_F}{\sigma_F} \tag{6.4}$$

Dimensionality Reduction Dimensionality reduction techniques such as PCA [200], ICA [201], and UMAP [202] are foundational across various disciplines, primarily for their capacity to distill complex datasets into a more manageable form. For WiFi-based PCS, these dimensionality reduction techniques offer promising strategies for dealing

with high-dimensional CSI data and eliminating subcarriers-specific noise, potentially enhancing model performance and generalization.

PCA compresses data by projecting them onto a new coordinate system defined by principal components, which are directions of maximum variance. This approach not only reduces the dimensionality but also manages to eliminate noisy OFDM subcarriers [45], thereby enhancing model performance. ICA distinguishes itself by separating multivariate signals into independent non-Gaussian components. This is especially beneficial in multi-person scenarios within WiFi-based sensing, where it helps identify original signal sources from complex mixtures (blind source separation problem) [110]. UMAP, on the other hand, offers a non-linear approach to dimensionality reduction, effectively maintaining both the local and global structure of high-dimensional data. This method is valuable for exploring complex patterns within data, facilitating insights into intricate relationships that linear techniques like PCA might overlook. Applied to WiFi-based indoor localization, UMAP shows potential in enhancing model performance [203, 117].

Data Augmentation To improve model generalization in TW scenarios, four random perturbation-based data augmentation techniques are considered: *random magnitude*, *circular rotation* along the time axis, *horizontal flipping*, and *dropout*. These methods are intended to address the challenge of temporal variability in CSI signals, which may result from hardware drift or environmental changes and are known to impair model robustness [125]. The *random magnitude* augmentation introduces a global scaling factor s to the feature matrix $F[t]$, simulating noise variations that may occur over time. As defined in Equation 6.5, s is drawn from a uniform distribution, and x controls the perturbation range:

$$F[t]^{'} = F[t]s, \qquad \text{with} \quad s \sim \mathcal{U}(1-x, 1+x) \tag{6.5}$$

Circular rotation applies a temporal shift to the entries in $F[t]$, rotating the sequence along the time axis. This augmentation alters the alignment of temporal activity patterns without modifying their internal structure, thereby promoting invariance to temporal shifts.

Horizontal flipping reverses the order of packets in $F[t]$, generating temporally inverted spectrograms. Analogous to left-right flips in image classification, this transformation increases dataset diversity by simulating mirrored activity sequences.

Dropout is implemented at both the subcarrier and packet levels. Unlike conventional dropout, where elements are zeroed, this variant replaces dropped elements with the global feature mean μ_F to preserve the overall signal structure while introducing localized noise [49]. The transformation is defined in Equation 6.6, where $M[t]$ is a binary dropout mask sampled from a Bernoulli distribution, and $\neg M[t]$ is its complement. The Hadamard product is denoted by $\odot$:

$$F[t]^{'} = D_{\mu}(F[t], p) = F[t] \odot M[t] + \neg M[t]\mu_F \tag{6.6}$$

(a) *walking* (b) *sitting* (c) *lying*

Figure 6.8: Day 1 examples of the activity classes *walking*, *sitting*, and *lying*. [50]

Subcarrier-wise dropout samples each element of $M[t]$ independently, while packet-wise dropout assigns a single binary value to each column of $M[t]$. Both variants aim to mimic noise on the subcarrier- or packet-level.

Evaluation Setup

To systematically evaluate the effectiveness of methods for enhancing model generalization under environmental and temporal domain shifts, an ablation study is conducted focusing on two key WiFi-based PCS tasks: 3D person localization and HAR. Both tasks are jointly addressed by adapting the *EfficientNetV2 small* architecture [190], extending it with an additional regression head to predict the spatial coordinates of the monitored individual alongside activity classification.

The evaluation is conducted in two distinct phases. Initially, the study assesses the generalization performance of various CSI features introduced previously: raw amplitude and phase, first-order amplitude and phase differences, and PSD. The best-performing CSI representation is subsequently selected as a baseline for further optimization. In the second phase, the baseline undergoes additional processing steps, including feature scaling, dimensionality reduction, and targeted data augmentations, aiming to further enhance model robustness against domain shifts.

Data The evaluation is conducted using the 3DO dataset (introduced in Section 5.2.4), which captures synchronized WiFi CSI and ground truth labels for HAR and localization in a controlled TW scenario situated in an office environment. The dataset is collected using System $\mathcal{D}$ in a fixed point-to-point transmitter-receiver configuration spanning two interior walls. Designed to analyze model generalization under real-world variability, the 3DO dataset includes three consecutive days of recordings that isolate static, dynamic, and temporal domain shifts. Static variation is introduced on day 3 by rearranging furniture and objects, dynamic variation arises from differing activity executions, and temporal effects occur naturally over time due to changes in environmental conditions or WiFi system behavior [125]. The HAR task is formulated as a three-class classification problem involving the activities *walking*, *sitting*, and *lying*, with visual examples shown in Figure 6.8. The localization task is framed as a regression problem for predicting the subject's 3D coordinates. To assess model robustness under domain shift, training is

performed exclusively on data from day 1, which serves as the baseline condition without domain shifts. An 8:2 split is used for training and validation. Testing is conducted on the data from days 2 and 3, allowing evaluation under dynamic, temporal, and static variations, respectively.

Model Training Model training leverages data exclusively from day 1, representing a controlled baseline environment free from domain shifts. This data is partitioned into training and validation subsets following an 8:2 ratio, ensuring a representative distribution of activities (*walking*, *sitting*, and *lying*). A balanced random sampler is applied to mitigate class imbalance within the dataset. The modified dual-task architecture is optimized using AdamW [195], combined with a cosine annealing learning rate scheduler [204]. The joint training objective combines Mean Squared Error (MSE) for regression and Cross-Entropy (CE) for classification, formulated as:

$$\mathcal{L} = \text{MSE} + \alpha \text{CE}, \tag{6.7}$$

where the hyperparameter α balances task contributions. A systematic hyperparameter search is conducted, identifying $\alpha = 0.4$, learning rate 1×10^{-3}, batch size $b = 32$, and window size $w = 351$ (≈ 3.51 seconds at a 100 Hz packet sending rate) as optimal. For each evaluated configuration, three independent training runs of 25 epochs each are performed, with the best-performing model selected based on validation loss. Model generalization performance is then evaluated on data from days 2 and 3, representing distinct static, dynamic, and temporal domain shifts.

Metrics Final results are reported using metrics appropriate for each task: RMSE for localization, along with precision, recall, F1-score, and classification accuracy (ACC) for HAR. Results are summarized as mean and standard deviation across the independent runs, providing a clear measure of consistency and reliability for each method under consideration.

Results

CSI Feature Extraction Table 6.13 compares the generalization performance of various CSI features: amplitude ($\mathcal{A}$), phase ($\mathcal{P}$), first-order differences ($\mathcal{A}_\Delta$, $\mathcal{P}_\Delta$), and power spectral density ($\mathcal{PSD}$). Across all feature types, models exhibit performance degradation on days 2 and 3 compared to the day-1 baseline, underscoring the challenge posed by dynamic, static, and temporal domain shifts inherent in the 3DO dataset.

Notably, amplitude features ($\mathcal{A}$) demonstrate superior robustness, consistently yielding the lowest localization errors (RMSE) and highest classification accuracy (ACC) on both days 2 and 3. Specifically, compared to phase features ($\mathcal{P}$), amplitude reduces localization errors by 33.75% and 4.94% on days 2 and 3, respectively, while improving classification accuracy by 34.39 percentage points on day 2 and 13.80 percentage points on day 3. This highlights the relative sensitivity of phase-based features to TW scenarios, where multipath effects and environmental noise influence phase stability.

Model	Day	RMSE [m] ↓	Precision ↑	Recall ↑	F1-Score ↑	ACC ↑
$\mathcal{A}$	1 (val.)	0.364 ±0.00	97.81 ±0.49	99.12 ±0.10	98.46 ±0.23	99.55 ±0.11
$\mathcal{P}$	1 (val.)	0.512 ±0.02	74.34 ±8.16	91.66 ±2.93	81.96 ±6.24	91.50 ±2.92
$\mathcal{A}_\Delta$	1 (val.)	0.596 ±0.02	79.71 ±5.39	92.50 ±2.03	85.58 ±3.98	93.03 ±1.99
$\mathcal{P}_\Delta$	1 (val.)	0.708 ±0.01	62.48 ±1.32	80.18 ±1.84	70.22 ±1.16	80.23 ±1.70
$\mathcal{PSD}$	1 (val.)	0.415 ±0.01	92.81 ±0.76	97.56 ±0.39	95.13 ±0.56	97.97 ±0.47
$\mathcal{A}$	2 (test)	0.587 ±0.02	79.38 ±2.59	82.94 ±3.22	81.12 ±2.89	83.36 ±3.21
$\mathcal{P}$	2 (test)	0.886 ±0.03	46.33 ±4.45	48.78 ±2.50	47.49 ±3.50	48.97 ±2.63
$\mathcal{A}_\Delta$	2 (test)	0.689 ±0.02	63.35 ±2.79	67.81 ±2.50	65.51 ±2.66	68.18 ±2.53
$\mathcal{P}_\Delta$	2 (test)	0.948 ±0.02	38.73 ±3.30	40.46 ±1.36	39.55 ±2.32	40.73 ±1.39
$\mathcal{PSD}$	2 (test)	0.588 ±0.01	76.46 ±0.87	80.11 ±1.47	78.24 ±1.15	80.57 ±1.50
$\mathcal{A}$	3 (test)	0.904 ±0.01	77.52 ±2.87	80.63 ±4.67	79.04 ±3.71	81.16 ±4.57
$\mathcal{P}$	3 (test)	0.951 ±0.02	64.74 ±3.56	66.81 ±6.62	65.72 ±5.04	67.36 ±6.55
$\mathcal{A}_\Delta$	3 (test)	0.972 ±0.03	71.35 ±0.07	76.92 ±0.30	74.03 ±0.14	77.64 ±0.30
$\mathcal{P}_\Delta$	3 (test)	0.962 ±0.03	64.06 ±1.71	73.81 ±2.50	68.58 ±1.93	74.27 ±2.44
$\mathcal{PSD}$	3 (test)	0.939 ±0.01	67.97 ±0.46	71.69 ±0.37	69.78 ±0.36	72.21 ±0.24

Table 6.13: Generalization performance of models trained on amplitude ($\mathcal{A}$), phase ($\mathcal{P}$), first-order difference of amplitude ($\mathcal{A}_\Delta$), first-order difference of phase ($\mathcal{P}_\Delta$), and PSD features ($\mathcal{PSD}$). All models are trained on day 1 data, representing the baseline, and tested on days 2 and 3. Metrics are presented as the mean and standard deviation across three independent training runs.

PSD features also demonstrate competitive generalization performance, closely following amplitude with only marginally higher RMSE and slightly lower ACC on both test days. Interestingly, first-order difference features ($\mathcal{A}_\Delta$ and $\mathcal{P}_\Delta$), although theoretically less sensitive to environmental variations [39, 72], do not outperform raw amplitude features. Particularly, $\mathcal{P}_\Delta$ performs the weakest among all features, affirming observations made in prior LOS studies [45].

Overall, these results emphasize the efficacy and robustness of amplitude features ($\mathcal{A}$) for TW scenarios, as they offer the best compromise between generalization performance and computational efficiency. Consequently, amplitude features are selected as the baseline for subsequent evaluations involving feature scaling, dimensionality reduction, and data augmentation techniques.

Feature Scaling Table 6.14 presents the performance of models trained using two feature scaling strategies: max-min scaling ($\mathcal{A}_{mm}$) and z-normalization ($\mathcal{A}_z$), both applied to amplitude features. Compared to the baseline model trained on raw amplitude ($\mathcal{A}$), max-min scaling yields consistent improvements in both localization and classification performance across days 2 and 3. Z-normalization also improves localization performance relative to the baseline but shows a slight reduction in classification accuracy on both test days.

Model	Day	RMSE [m] ↓	Precision ↑	Recall ↑	F1-Score ↑	ACC ↑
$\mathcal{A}$ (baseline)	1 (val.)	0.364 ±0.00	97.81 ±0.49	99.12 ±0.10	98.46 ±0.23	99.55 ±0.11
$\mathcal{A}_{mm}$	1 (val.)	0.363 ±0.01	97.41 ±0.28	98.70 ±0.20	98.05 ±0.23	99.26 ±0.24
$\mathcal{A}_{z}$	1 (val.)	0.356 ±0.01	98.04 ±0.17	98.71 ±0.10	98.37 ±0.06	99.45 ±0.05
$\mathcal{A}_{PCA\_42}$	1 (val.)	0.360 ±0.01	97.58 ±0.16	98.60 ±0.15	98.08 ±0.06	99.49 ±0.04
$\mathcal{A}_{ICA\_24}$	1 (val.)	0.770 ±0.28	64.54 ±25.4	69.18 ±25.6	66.73 ±25.4	69.33 ±25.4
$\mathcal{A}_{UMAP\_48}$	1 (val.)	0.467 ±0.01	93.54 ±1.14	97.78 ±0.33	95.61 ±0.76	98.19 ±0.19
$\mathcal{A}_{AUG}$	1 (val.)	0.347 ±0.00	97.92 ±0.76	98.93 ±0.21	98.42 ±0.48	99.32 ±0.20
$\mathcal{A}_{mmAUG}$	1 (val.)	0.350 ±0.01	97.70 ±0.36	98.86 ±0.18	98.27 ±0.23	99.25 ±0.11
$\mathcal{A}_{zAUG}$	1 (val.)	0.343 ±0.00	97.91 ±0.11	98.82 ±0.06	98.37 ±0.08	99.33 ±0.14
$\mathcal{A}$ (baseline)	2 (test)	0.587 ±0.02	79.38 ±2.59	82.94 ±3.22	81.12 ±2.89	83.36 ±3.21
$\mathcal{A}_{mm}$	2 (test)	0.542 ±0.06	80.48 ±3.45	83.19 ±3.58	81.81 ±3.52	83.54 ±3.65
$\mathcal{A}_{z}$	2 (test)	0.573 ±0.02	78.53 ±0.82	82.77 ±2.03	80.59 ±1.34	83.10 ±2.04
$\mathcal{A}_{PCA\_42}$	2 (test)	0.601 ±0.05	76.46 ±1.84	80.72 ±2.82	78.53 ±2.30	81.07 ±2.88
$\mathcal{A}_{ICA\_24}$	2 (test)	0.796 ±0.16	61.79 ±12.8	65.13 ±13.8	63.42 ±13.3	65.40 ±14.0
$\mathcal{A}_{UMAP\_48}$	2 (test)	0.720 ±0.02	72.23 ±1.44	72.06 ±2.10	72.14 ±1.77	72.23 ±2.11
$\mathcal{A}_{AUG}$	2 (test)	0.500 ±0.02	82.20 ±1.46	86.23 ±1.37	84.17 ±1.42	86.52 ±1.38
$\mathcal{A}_{mmAUG}$	2 (test)	0.503 ±0.01	84.22 ±0.19	87.60 ±0.72	85.88 ±0.44	88.00 ±0.75
$\mathcal{A}_{zAUG}$	2 (test)	0.527 ±0.02	82.80 ±1.91	86.18 ±2.21	84.45 ±2.05	86.51 ±2.22
$\mathcal{A}$ (baseline)	3 (test)	0.904 ±0.01	77.52 ±2.87	80.63 ±4.67	79.04 ±3.71	81.16 ±4.57
$\mathcal{A}_{mm}$	3 (test)	0.887 ±0.01	81.91 ±7.59	83.31 ±8.15	82.60 ±7.86	83.89 ±8.07
$\mathcal{A}_{z}$	3 (test)	0.896 ±0.01	77.67 ±2.19	80.00 ±2.36	78.82 ±2.24	80.69 ±2.29
$\mathcal{A}_{PCA\_42}$	3 (test)	0.948 ±0.02	66.18 ±6.89	68.32 ±8.07	67.23 ±7.46	68.89 ±8.01
$\mathcal{A}_{ICA\_24}$	3 (test)	1.012 ±0.04	52.99 ±15.6	53.46 ±15.5	53.19 ±15.5	53.83 ±15.8
$\mathcal{A}_{UMAP\_48}$	3 (test)	0.913 ±0.01	74.56 ±2.36	79.24 ±1.78	76.83 ±2.09	79.91 ±1.72
$\mathcal{A}_{AUG}$	3 (test)	0.871 ±0.00	77.93 ±4.00	79.68 ±3.74	78.79 ±3.84	80.31 ±3.64
$\mathcal{A}_{mmAUG}$	3 (test)	0.872 ±0.02	79.71 ±1.97	82.10 ±2.75	80.88 ±2.35	82.76 ±2.76
$\mathcal{A}_{zAUG}$	3 (test)	0.880 ±0.02	79.94 ±2.05	82.02 ±1.82	80.97 ±1.93	82.61 ±1.86

Table 6.14: Effects on model generalization performance of max-min scaling ($\mathcal{A}$), z-normalization ($\mathcal{A}_z$), PCA ($\mathcal{A}_{PCA\_42}$), ICA ($\mathcal{A}_{ICA\_24}$), UMAP ($\mathcal{A}_{UMAP\_48}$), data augmentation $\mathcal{A}_{AUG}$, max-min scaling with data augmentation ($\mathcal{A}_{mmAUG}$) and z-normalization with data augmentation ($\mathcal{A}_{zAUG}$). All models are trained on day 1 data, and tested on days 2 and 3. Metrics are presented as the mean and standard deviation across three independent training runs.

While max-min scaling leads to marginally better generalization than z-normalization in most scenarios, the performance gap between the two methods remains small. These results suggest that both scaling techniques can support generalization under environmental and temporal domain shifts, though neither clearly outperforms the other across all metrics. Given the limited difference, performance variability across training runs may account for the observed fluctuations.

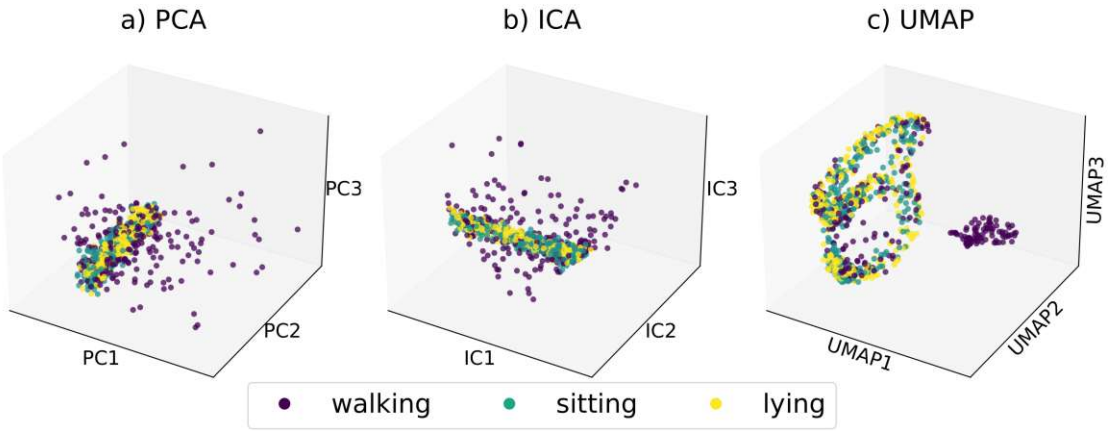


Figure 6.9: Comparison of dimensionality reduction techniques on day 1 data, showing a) PCA, b) ICA, and c) UMAP projections down to three dimensions (visualizing 0.5% of day 1 samples). It can be seen that neither PCA nor ICA effectively separates activity clusters. In contrast, UMAP distinguishes most samples associated with the *walking* activity from the combined cluster of *sitting* and *lying* activities.

Dimensionality Reduction The performance of models trained on amplitude features subjected to dimensionality reduction using PCA ($\mathcal{A}_{PCA\_d}$), ICA ($\mathcal{A}_{ICA\_d}$), and UMAP ($\mathcal{A}_{UMAP\_d}$) is detailed in Table 6.14. The number d in the model names signifies the reduced dimensionality, which was optimized through a hyperparameter search within the range $d \in \{2, 3, 4, ..., 52\}$, utilizing day 1 validation data to determine the optimal value for each method. This search identifies an optimal dimensionality of 42 for PCA, 24 for ICA, and 48 for UMAP.

The evaluation of dimensionality reduction techniques shows that none of the applied methods enhances performance metrics beyond the baseline. PCA and UMAP demonstrate comparable performance levels, with PCA having a slight edge on day 2 and UMAP on day 3. In contrast, ICA leads to significant performance degradation, with a reduction in accuracy of 21.55% on day 2 and 33.67% on day 3 relative to the baseline. Figure 6.9 illustrates the three-dimensional projections of day 1 data using PCA, ICA, and UMAP, showcasing UMAP's ability to distinguish *walking* samples from the conjoined clusters of *sitting* and *lying* more effectively than PCA and ICA. Despite this advantage in data visualization, UMAP does not lead to better localization and classification performance compared to PCA on the validation set. UMAP's emphasis on preserving local structures for visualization might lead to a loss of predictive information, in contrast to PCA's approach of retaining global variance, which could be more relevant for certain predictive tasks. While some studies have reported improved model performance in WiFi-based PCS tasks with these methods [160, 117], the obtained results indicate that these findings do not translate to TW scenarios.

Data Augmentation The evaluation of data augmentation techniques on amplitude features includes random amplitude perturbations, dropout, circular rotation, and horizontal flipping, aiming to improve model robustness and generalization in TW scenarios. Optimal parameters for each technique are determined through a comprehensive hyperparameter search. The analysis reveals that random amplitude perturbations and dropout generally do not yield consistent improvements over the baseline model performance on days 2 and 3. The limited effectiveness of these augmentations is attributed to the temporal stability of the WiFi system used, as indicated by stable CSI amplitude statistics across all days (day 1: 12.90 ±2.33, day 2: 12.90 ±2.27, day 3: 12.77 ±2.55). Consequently, such perturbations diverge from the inherent data distribution, resulting in degraded test performance. Nevertheless, a WiFi system experiencing less temporal stability could potentially benefit from these augmentations.

Conversely, random circular rotations and horizontal flipping demonstrate substantial improvements in both RMSE and ACC. As presented in Table 6.14, applying circular rotations with a magnitude of ±12.5% (±43 samples), referred to as $\mathcal{A}_{AUG}$, achieves the highest improvement, reducing RMSE by 14.82% on day 2 and 3.65% on day 3. In terms of classification accuracy, circular rotations at ±12.5% result in an increase of 3.98% on day 2, while rotations at ±6.25% achieve the highest ACC improvement of 2.24% on day 3. These findings suggest the optimal rotation magnitude might be dataset-specific, indicating the necessity of tuning this parameter carefully.

To explore potential additive effects, circular rotations at ±12.5% are combined with horizontal flipping. This combined augmentation results in RMSE reductions of 14.31% on day 2 and 2.10% on day 3. However, classification performance shows mixed results: a 3.42% increase in accuracy on day 2 but a slight decrease of 1.66% on day 3 compared to circular rotation alone. These outcomes highlight the complex interactions between augmentation techniques, suggesting that combining multiple augmentations does not always guarantee improved performance.

Finally, combining circular rotations at ±12.5% with max-min scaling ($\mathcal{A}_{mmAUG}$) and z-normalization ($\mathcal{A}_{zAUG}$) shows distinct performance patterns. Max-min scaling combined with circular rotations maintains stable localization performance while achieving the highest classification scores on day 2, with F1 and ACC reaching 85.88% and 88.00%, respectively. On day 3, $\mathcal{A}_{mmAUG}$ provides noticeable improvements over the baseline, though it does not surpass max-min scaling alone ($\mathcal{A}_{mm}$) in classification performance. These results suggest that the effectiveness of feature scaling methods may be context-dependent when paired with specific augmentation techniques.

Discussion

The presented evaluation demonstrates that plain CSI amplitude features consistently achieve superior generalization performance under environmental and temporal domain shifts in TW scenarios. This finding aligns with previous observations in LOS contexts [45], suggesting amplitude-based representations as inherently robust to subtle domain variations. The evaluation further highlights that simple feature scaling techniques (max-min

scaling and z-normalization) and straightforward temporal augmentations (random circular rotations and horizontal flipping) yield notable performance improvements without introducing significant computational overhead or complexity.

In contrast, dimensionality reduction techniques, such as PCA, ICA, and UMAP, do not enhance performance in the evaluated TW settings. This result marks a clear divergence from their known effectiveness in LOS environments [45], indicating that TW scenarios possess distinct propagation characteristics that render standard dimensionality reduction strategies less effective. Consequently, this highlights an important avenue for future research, namely the development of dimensionality reduction techniques explicitly tailored to the unique constraints and signal dynamics inherent in TW environments.

While the current study provides foundational insights, future work should expand the scope of evaluation to a broader variety of domain variations, including diverse antenna configurations, different transmitter-receiver placements, and subject diversity in physiological characteristics. Additionally, combining amplitude features with other types of CSI information, such as phase or derivative features, may offer further generalization improvements, as indicated by prior research [104]. Evaluating these combinations alongside alternative preprocessing methods across additional publicly available datasets will be crucial to validating the generalizability of the observed results beyond the specific experimental conditions employed here.

Overall, these findings advance the understanding of generalization in WiFi-based PCS, particularly under complex TW conditions. With respect to **RQ III**, this study identifies practical and computationally efficient preprocessing techniques that enhance the cross-domain generalization capabilities of WiFi-based PCS models at the training stage.

6.3.3 Domain-Adversarial Test-Time Adaptation

Preprocessing techniques such as data augmentation, feature scaling, and dimensionality reduction are simple ways to improving model robustness to known domain shifts [48, 50]. To address domain variability more effectively, prior research explores Domain-Adversarial Training (DAT)[132], which utilizes adversarial feature alignment techniques to extract domain-invariant representations. DAT-based architectures demonstrate success in reducing model sensitivity to source-specific domain characteristics, thereby improving cross-domain generalization [133, 71, 205]. However, these approaches are fundamentally limited to static adaptation and do not accommodate the rapid domain variations frequently encountered during real-world deployments. In practice, environments evolve continuously, i.e., objects move, signal characteristics change, and user behavior varies over time, causing the underlying distribution of CSI to drift in ways that static methods cannot anticipate or correct. This constraint makes models highly vulnerable to test-time domain shifts, particularly in WiFi-based PCS, where signal propagation is inherently unstable due to the dynamic nature of human behavior.

To address this challenge, adaptive strategies are required that not only generalize across domains during training but also adjust to evolving CSI distributions at test time. To this end, Test-Time Training (TTT) methods propose continuous adaptation by updating model weights during inference using auxiliary self-supervised tasks [196, 206, 207, 208]. While powerful, TTT requires substantial architectural modifications and training complexity. In contrast, Test-Time Adaptation (TTA) methods adapt pre-trained models during inference without altering the training procedure, offering a more streamlined and computationally efficient solution [209, 210, 211]. Despite their success in computer vision domains, TTA approaches remain underexplored in WiFi-based sensing contexts.

Motivated by these observations, the proposed Domain-Adversarial Test-Time Adaptation (DATTA) framework [54] integrates the complementary strengths of DAT and TTA specifically for WiFi-based PCS. DATTA leverages DAT to learn robust, domain-invariant features during initial training, and subsequently applies TTA for real-time adaptation to evolving CSI distributions at test time. To maintain adaptation stability and prevent catastrophic forgetting during prolonged test-time adaptation, DATTA incorporates a random weight-resetting mechanism, periodically restoring model weights toward their original domain-invariant state [210]. Additionally, DATTA builds upon the efficient *WiFlexFormer* architecture [53], ensuring low computational overhead and real-time inference suitable for embedded edge deployment. By combining these components, DATTA addresses **RQ III**, offering robust adaptation to complex, unpredictable domain shifts that go beyond static preprocessing techniques previously discussed.

DATTA Framework

The proposed DATTA framework combines DAT with TTA to enable robust cross-domain generalization in WiFi-based PCS. Through DAT, the model learns domain-invariant features by leveraging data from diverse domains, achieving offline adaptation to varied

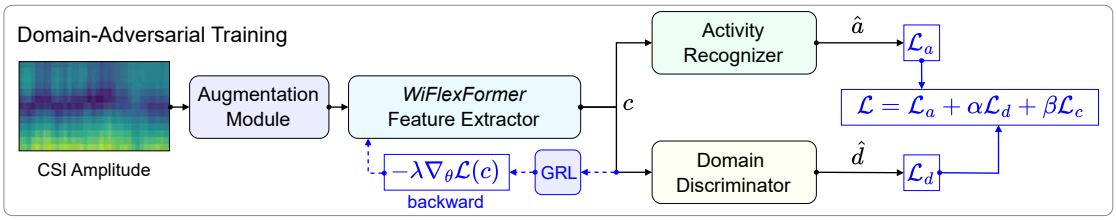


Figure 6.10: Overview of the domain-adversarial training approach used in DATTA. [54]

data characteristics. The DAT architecture builds on the adversarial structure for domain-invariant feature learning outlined in [71], with a *WiFlexFormer*-based central feature extractor [53] to ensure efficient, real-time PCS. To further enhance cross-domain generalization, a specialized augmentation module, tailored to the unique properties of WiFi CSI, is incorporated. Despite DAT's effectiveness, residual domain shifts can still occur at test time due to environmental changes that lead to distribution shifts in data. To account for these shifts as well, the TTA framework from [209] is adapted for WiFi-based PCS to align the feature distributions of target and source domains in real-time at test time. Additionally, random weight resetting [212] is leveraged to prevent catastrophic forgetting of learned domain invariance during prolonged domain shifts, ensuring sustained model stability. A PyTorch implementation of DATTA is publicly available[2].

Domain-Adversarial Training Figure 6.10 illustrates the DAT architecture used in DATTA, consisting of the *Feature Extractor*, *Activity Recognizer*, and *Domain Discriminator*. The feature extractor processes CSI amplitude data to learn domain-invariant features by generating representations that are informative for activity recognition while disregarding domain-specific aspects. During training, the extracted features are fed to the activity recognizer and domain discriminator. The domain discriminator enforces domain invariance by applying an adversarial loss, pushing the feature extractor to produce features that are indistinguishable across domains. After training, the domain discriminator is discarded, leaving a model optimized for cross-domain sensing.

Model Input/Output: The DAT architecture takes CSI amplitude spectrograms $s \in \mathcal{S}$ as input, each associated with an activity label $a \in \mathcal{A}$ and a domain label $d \in \mathcal{D}$, where $\mathcal{A}$ and $\mathcal{D}$ represent the sets of activities and domains, respectively. The output is the predicted activity label $\tilde{a} \in \mathcal{A}$.

Augmentation Module: Before feature extraction, raw CSI amplitude spectrograms are processed by the augmentation module, which applies a set of realistic random augmentations to increase data variability. Among these augmentations are amplitude perturbations, circular rotations along the temporal axis, as well as pixel- and row-wise dropout with mean replacement [50].

Feature Extractor: The feature extractor is based on the *WiFlexFormer* [53] architecture which is chosen due to its lightweight design and simplicity. It consists of a convolutional

[2]DATTA, https://github.com/StrohmayerJ/DATTA, accessed: 15.04.2025

stem followed by Gaussian positional encoding and a transformer encoder with class token and a linear classification layer. Its architecture is designed for both, amplitude and DFS features, however, experiments are conducted in the amplitude mode to minimize the parameter count and inference time. The resulting model is comparatively small with only $\approx$ 40k parameters.

Activity Recognizer: The activity recognizer, acting as the classification head, consists of two linear layers with a ReLU activation function in between and $|\mathcal{A}|$ output channels. It takes the class token embeddings c as input and outputs the predicted activity probabilities $\hat{a}$. Its loss $\mathcal{L}_a$, the main training objective, is the cross-entropy between the predicted and true activities:

$$\mathcal{L}_a = -\sum_{k=1}^{|\mathcal{A}|} a_k \log(\hat{a}_k). \tag{6.8}$$

Domain Discriminator: The domain discriminator architecture mirrors the activity recognizer with two linear layers with a ReLU activation function in between and $|\mathcal{D}|$ output channels. It takes the class token embeddings c as input and outputs a prediction over domains. The domain loss $\mathcal{L}_d$ is then computed using the cross-entropy between the predicted domain probabilities $\hat{d}$ and the true domain labels d:

$$\mathcal{L}_d = -\sum_{k=1}^{|\mathcal{D}|} d_k \log(\hat{d}_k). \tag{6.9}$$

Furthermore, to facilitate efficient training without having to freeze model weights alternately, a GRL is utilized. As shown in Figure 6.10, during the forward pass, the GRL acts as an identity function, allowing the features to flow unchanged to the domain discriminator. However, during backpropagation, it multiplies the gradients $\nabla_\theta \mathcal{L}(s)$ by $-\lambda$ to reverse them, thus, returning $-\lambda \nabla_\theta \mathcal{L}(s)$ to the feature extractor:

$$\text{Forward Pass:} \qquad \text{GRL}(s) = s, \tag{6.10}$$

$$\text{Backward Pass:} \qquad \text{GRL}(s) = -\lambda \nabla_\theta \mathcal{L}(s), \tag{6.11}$$

where λ is a scaling parameter that controls the strength of the adversarial signal. By reversing the gradients, the feature extractor is encouraged to produce features that are indistinguishable across domains, thus learning domain-invariant representations.

To perform DAT without overwhelming the feature extractor in the early phase of training, dynamic scaling of λ is performed as follows:

$$\lambda = \Big(\frac{2}{1 + e^{-10p}} - 1\Big)\gamma, \tag{6.12}$$

where $p \in [0, 1]$ represents the training progress and γ is a scaling parameter to control the adversarial signal strength. This allows the feature extractor to focus on learning

robust features for the primary task of activity recognition in the beginning and, by gradually increasing the strength of the adversarial signal, enables a smooth transition to domain-invariant feature learning.

Domain-Adversarial Loss: The loss function in the DAT architecture leverages adversarial training to balance activity recognition and domain-invariant feature learning while penalizing overconfidence in a specific class through a confidence control mechanism. This is achieved by combining task-specific ($\mathcal{L}_a$) and domain-specific ($\mathcal{L}_d$) losses with the axillary loss $\mathcal{L}_c$, representing the *Confidence Control Constraint* (CCC) from [71]. $\mathcal{L}_c$ penalizes predictions that are overly certain by adding a penalty for class probabilities approaching 0 or 1. Here, $\hat{a}_{ik}$ represents the predicted probability for activity class $k \in \mathcal{A}$ of the i-th sample, ensuring that each class prediction is regularized:

$$\mathcal{L}_c = -\sum_{k=1}^{|\mathcal{A}|} \log(\hat{a}_{ik}) + \log(1 - \hat{a}_{ik}). \tag{6.13}$$

Combined with task- and domain-specific losses, the final domain-adversarial loss function minimized during DAT is given by:

$$\mathcal{L} = \mathcal{L}_a + \alpha\mathcal{L}_d + \beta\mathcal{L}_c, \tag{6.14}$$

where α and β are weighting parameters, used for controlling the strengths of the adversarial signal and the CCC, respectively.

Test-Time Adaptation While DAT is effective in learning domain-invariant features, substantial domain shifts still lead to a drop in performance during test time. In order to reduce the impact of such domain shifts, TTA is employed, allowing off-the-shelf pre-trained models to adapt online to new target domains without requiring additional labeled data. Building upon the framework proposed by Lin et al. [209], TTA is adopted from RGB video to CSI amplitude spectrograms to further enhance the generalization of DAT during test time by performing feature distribution alignment, i.e., aligning source statistics of the model with online estimates of the target statistics. Additionally, to prevent overfitting to the target distribution during prolonged adaptation, i.e., catastrophic forgetting [212], random weight resetting is implemented, following the approach proposed by Wang et al. [210]. Specifically, a subset of model weights is reverted to their original source values to keep them closer to the domain-invariant feature space learned with DAT.

Feature Map Alignment: To address the distribution shift, the statistics of feature maps are aligned, i.e., matching the means and variances of training and test CSI amplitude spectrograms. Let $\phi_l(s;\theta)$ represent the feature map of the l-th layer of network ϕ, computed for a spectrogram s with parameters θ. Each feature map is a matrix of dimensions (t_l, f_l), where t_l and f_l correspond to the time steps (WiFi packets) and frequency bins (subcarriers), respectively.

Computing the mean of the l-th layer features for a dataset $\mathcal{S}$ across the time dimension results in a mean vector of size f_l, which can be expressed as:

$$\mu_l(\mathcal{S};\theta) = \mathbb{E}_{s\in\mathcal{S}}\mathbb{E}_{t\in[1,t_l]}\Big[\phi_l(x;\theta)[t]\Big], \tag{6.15}$$

and the variance of the l-th layer features is given by:

$$\sigma_l^2(\mathcal{S};\theta) = \mathbb{E}_{s\in\mathcal{S}}\mathbb{E}_{t\in[1,t_l]}\Big[(\phi_l(x;\theta)[t] - \mu_l(\mathcal{S};\theta))^2\Big]. \tag{6.16}$$

For the remainder of this work, the mean and variance computed on the training set are denoted with $\bar{\mu}_l$ and $\bar{\sigma}_l^2$. When training data is unavailable, these statistics can be estimated from batch norm layers as well, though with a small decrease in performance [209].

At test time, updates are performed iteratively, adjusting the discrepancy between the test statistics of a batch $\mathcal{B}$ of selected layers L with those computed during training:

$$\mathcal{L}_{\text{TTA}} = \sum_{l\in L} ||\mu_l(\mathcal{B};\theta) - \bar{\mu}_l||_2 + ||\sigma_l^2(\mathcal{B};\theta) - \bar{\sigma}_l^2||_2. \tag{6.17}$$

In the conducted experiments, optimal results are obtained by selecting L to contain only the first out of four transformer encoder layers that compose the *WiFlexFormer* encoder.

Given the low inference time of *WiFlexFormer*, TTA is performed for the most realistic application scenario: online, on data received in a stream. Hence, $|\mathcal{B}| = 1$ is chosen and target statistics are continuously evaluated using exponential moving averages instead of repeatedly computing statistics for the constantly growing test set. In other words, given the CSI amplitude spectrogram s_i, received in iteration i, the mean and variance estimates are updated as follows:

$$\hat{\mu}_l^{(i)} = \alpha \cdot \mu_l(s_i;\theta) + (1-\alpha)\cdot \hat{\mu}_l^{(i-1)}, \tag{6.18}$$

$$\hat{\sigma}_l^{2(i)} = \alpha \cdot \sigma_l^2(s_i;\theta) + (1-\alpha)\cdot \hat{\sigma}_l^{2(i-1)}, \tag{6.19}$$

where $1-\alpha$ denotes the momentum. As a starting point, the source statistics are selected, i.e., $\mu_l^{(0)} = \bar{\mu}_l$ and $\sigma_l^{2(0)} = \bar{\sigma}_l^2$. Without modifications to Equation 6.17 and with no extensive recomputation necessary, $\mathcal{L}_{\text{TTA}}$ is computed using these estimates instead.

Weight Resetting In order to avoid catastrophic forgetting, in each iteration, a subset of the current weights is reset to their original values from $\bar{\theta}$, the source models's parameters. Specifically, consider the weights of layer l in iteration i, denoted as $\theta_l^{(i)}$. To randomly reset these, a Boolean mask m_l is defined with $\dim(\theta_l^{(i)}) = \dim(m_l)$, where each element of the mask is sampled from a Bernoulli distribution with reset rate p. The updated parameter vector $\theta_l^{(i)}$ is then:

$$\theta_l^{(i)} = m_l \odot \bar{\theta}_l + (1 - m_l) \odot \theta_l^{(i)}, \tag{6.20}$$

where $\odot$ denotes element-wise multiplication.

Comparison to TTT and Other TTA Methods Unlike TTT, TTA requires no changes to the architecture, making it suitable for off-the-shelf models. Additionally, inference with TTA is arguably faster than with TTT, as TTT often involves a computationally expensive reconstruction task as the secondary objective [208, 213]. Compared to other TTA approaches that align features by adjusting only the running statistics of normalization layers [214, 211], the proposed method updates the entire parameter vector θ up to the highest layer in L, offering greater flexibility during adaptation compared to the others. However, this can also be problematic since continuously adapting entire parameter vectors will unlearn the learned domain-invariant feature transformation eventually. To overcome this, weight resetting is employed, allowing the model to retain its original form. With this approach, while TTA does increase inference time compared to the original *WiFlexFormer*, practical deployment in real-time applications remains feasible.

Evaluation Setup

DATTA's effectiveness is evaluated through a detailed ablation study of its components, including the augmentation module, loss function elements, and discriminator input in DAT, as well as random weight resetting in TTA to assess its impact on model stability. Experiments are based on publicly available data adapted for DAT and TTA. Additional, the inference time on edge hardware is measured to assess the computational impact of TTA and weight resetting on real-time performance, addressing practical deployment feasibility.

Data *Widar3.0-G6 Dataset*: The Widar3.0 [74] WiFi HAR dataset features CSI recordings of 22 human hand gestures performed by 16 participants in three different indoor environments. Since not all 22 gestures are consistently performed in all three environments, a subset of Widar3.0, referred to as Widar3.0-G6 [75], is often utilized instead of the full dataset. This subset includes 6 hand gestures (*push and pull*, *sweep*, *clap*, *slide*, *draw circle* and *draw zigzag*) that are performed across all three environments by 16 users, resulting in a total of 68,246 hand gesture samples. The recording setup, shown in Figure 6.5b, consists of one 5.825 GHz WiFi transmitter (TX) and six receivers (RX_n), each equipped with an *Intel WiFi Link 5300* wireless NIC that has three antennas. The CSI of 90 subcarriers (3 antennas $\times$ 30 subcarriers) is collected at each receiver using the *Linux CSI Tool* [193], utilizing a packet sending rate of 1,000 Hz.

Widar3.0-G6D Dataset: To evaluate cross-domain generalization, the Widar3.0-G6 dataset is preprocessed by extracting CSI from 30 subcarriers of the first antenna at each receiver and applying temporal sub-sampling to 100 Hz for improved computational efficiency. Only samples between 120 and 220 WiFi packets ($\approx$ 1.2–2.2 seconds) are retained, with shorter samples zero-padded to standardize length. Amplitude features are then extracted and normalized using min-max scaling. The resulting Widar3.0-G6D dataset comprises 58,648 samples and is split into two disjoint subsets based on unique room-participant combinations (domains): a training subset with 7 domains (2 environments, 7 persons) and a test subset with 9 domains (1 environment, 9 persons), as shown in Table 6.15. For reproducibility, a Python script to generate the Widar3.0-G6D dataset is provided[2].

Subset	Activities	Environments	Persons	Domains	Samples
Train	6	2	7	7	19,586
Val	6	2	7	7	4,896
Val_{TTA}	6	1	9	9	3,417
Test	6	1	9	9	30,749
Total	6	3	16	16	58,648

Table 6.15: Overview of the distribution of activities, environments, participants, domains, and activity samples across the Widar3.0-G6D subsets used for training, validation, and testing.

Model Training Four model configurations are evaluated to assess the impact of DATTA and its components: the baseline *WiFlexFormer* (W), its TTA variant (W_{TTA}), the DAT model (W_{DAT}), and the final DATTA model (W_{DATTA}). All model backbones are trained on *Train*, validated on *Val*, with TTA hyperparameters tuned on Val_{TTA}, and tested on the shuffled dataset *Test*, if not stated otherwise. Note that shuffling *Test* poses an extreme case of high-frequency domain shifts.

The baseline W uses the vanilla *WiFlexFormer* architecture without DAT or TTA. W_{DAT} incorporates DAT with loss weights $\alpha = 0.3$, $\beta = 0.2$, and GRL-scaling $\gamma = 8$. Adding TTA to W and W_{DAT} produces W_{TTA} and W_{DATTA}, allowing the model's first layer to adapt during inference. The complete DATTA framework, W_{DATTA}, is tuned using Bayesian and grid search. For all models using TTA, random weight resetting with $p = 1 \times 10^{-4}$ is evaluated to enhance stability. The augmentation module's effectiveness is further assessed by training each model with and without augmentations.

Results

Augmentation Module The first component in the DAT pipeline to be evaluated is the augmentation module. Inspecting Table 6.16, which depicts all cross-domain HAR results, reveals the critical role of data augmentation for handling cross-domain variations. Comparing the baseline model W with and without augmentation, a substantial improvement in F1-Score, from 40.62% to 49.32%, can be observed.

Its impact is even more pronounced in W_{DAT}, the DAT model: without augmentation, W_{DAT} achieves an F1-Score of only 42.02%, reflecting poor generalization and significant overfitting to domain-specific features. Hence, augmentation is crucial for DAT as without it, the model overfits early and fails to learn domain-invariant features at later stages of training when the adversarial signal strength is increased. However, with augmentation, W_{DAT} attains an F1-Score of 65.66%, showing that sufficient data variability is essential to facilitating domain-invariant feature learning.

Confidence Control Constraint Next, the impact of the CCC is assessed, which is designed to prevent overconfidence in specific class predictions to encourage a more balanced feature representation across classes. By discouraging extreme confidence in

Model	Data Augmentation	Weight Resetting	ACC ↑	F1-Score ↑
W (baseline)	-		38.69 ±3.17	40.62 ±2.93
	✓		47.75 ±4.31	49.32 ±4.45
W_{TTA}	-	-	42.76 ±4.21	44.89 ±3.93
	✓	-	51.70 ±4.40	53.20 ±4.26
W_{DAT}	-		39.54 ±2.80	42.02 ±2.62
	✓		64.53 ±1.87	65.66 ±1.85
W_{DATTA}	-	-	45.26 ±5.83	48.23 ±5.11
	✓	-	65.90 ±1.53	67.29 ±1.43
	✓	✓	66.92 ±1.54	68.13 ±1.45

Table 6.16: Cross-domain HAR performance on the Widar3.0-G6D dataset averaged over three runs, comparing models trained with conventional training (W), test-time adaptation (W_{TTA}), domain-adversarial training (W_{DAT}), and the proposed combined approach, domain-adversarial test-time adaptation (W_{DATTA}).

Model	CCC	ACC ↑	F1-Score ↑
W_{DAT}	-	62.81 ± 1.09	63.81 ± 1.09
W_{DAT}	✓	64.53 ± 1.87	65.66 ± 1.85
W_{DATTA}	-	65.62 ± 1.08	66.78 ± 1.08
W_{DATTA}	✓	66.92 ± 1.54	68.13 ± 1.45

Table 6.17: Ablation study on the Confidence Control Constraint (CCC), weighted with $\beta = 0.2$.

particular classes, CCC helps stabilize training, making the model more adaptable to unseen domains. By means of a hyperparameter search, the optimal CCC weight $\beta = 0.2$ is identified and to evaluate its effectiveness, an ablation study comparing models trained with and without CCC (i.e., $\beta = 0$) is conducted. Observing the results, given in Table 6.17, shows that including a CCC improves performance in both the W_{DAT} and W_{DATTA} models. For W_{DAT}, enabling CCC leads to an increase in F1-Score from 63.81% to 65.66%, and for W_{DATTA}, from 65.62% to 68.13%, reinforcing that CCC enhances cross-domain generalization.

Random Weight Resetting Following this, the impact of random weight resetting during TTA is evaluated which is introduced to prevent catastrophic forgetting. The results, given in Table 6.16, indicate that random weight resetting promotes cross-domain generalization, however only when the model is domain-invariant to some degree, as achieved through DAT with data augmentation.

For W_{DATTA} with augmentation, enabling random weight resetting boosts the F1-Score to 68.13%, up from 67.29% without resetting, achieving the highest performance within

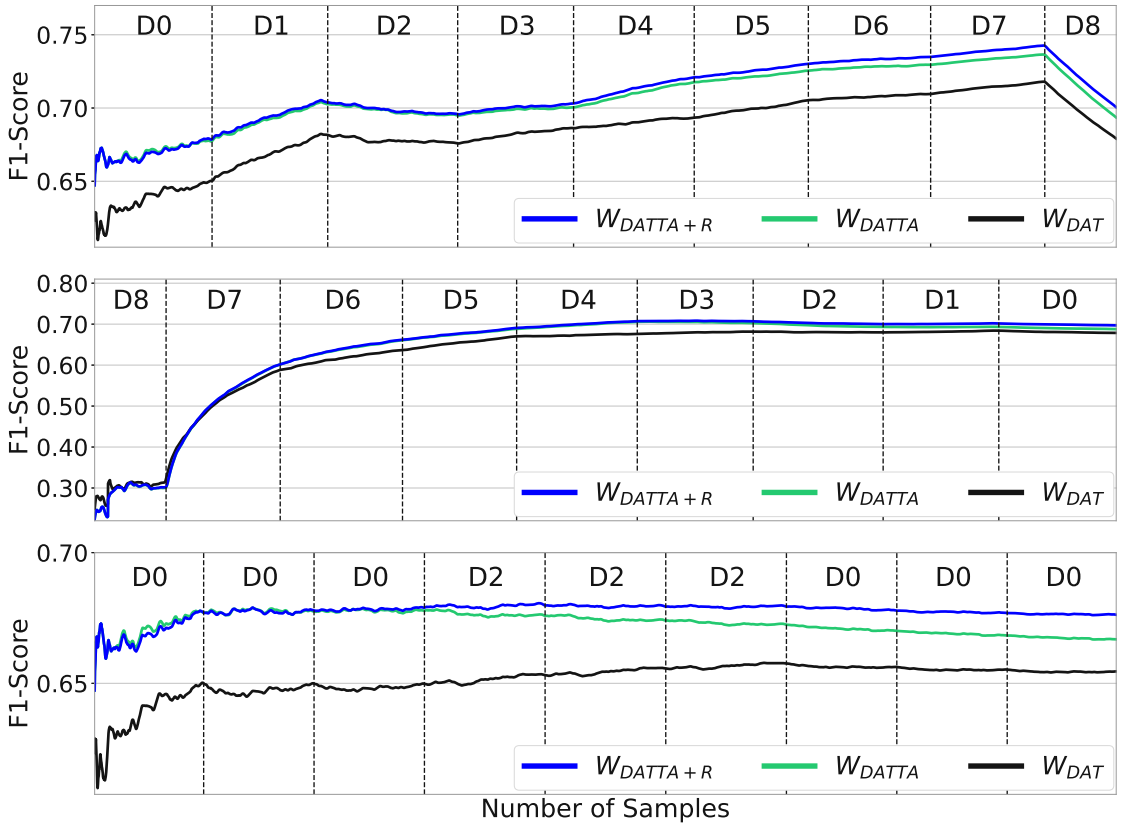


Figure 6.11: TTA performance across continuous test domain sequences. From top to bottom: (1) ascending domain order (D0 to D8), (2) descending domain order (D8 to D0), and (3) alternating domain order with prolonged domains D0 and D2. Depicted are F1-Scores computed with a rolling window of 100 samples for DATTA models with weight resetting ($W_{\text{DATTA+R}}$, blue), without weight resetting (W_{DATTA}, green), and the baseline DAT model without TTA (W_{DAT}, black). [54]

our model set. However, in other cases, it has limited or even negative effects. For instance, applying weight resetting to W_{DATTA} without augmentation or to W_{TTA} leads to reduced performance. It is hypothesized that this is due to the weights of the base model W and W_{DAT} without augmentation not being sufficiently domain-invariant yet, causing weight resets to revert beneficial adaptations during TTA. In contrast, W_{DATTA}, trained with augmentation, and thus exhibiting stronger domain invariance, allows weight resetting to maintain proximity to this invariant space during TTA. Consequently, it prevents drifting due to over-adaptation to specific test domains, avoiding catastrophic forgetting and stabilizing performance across diverse domains.

Cross-Domain Adaptation Moving forward, Figure 6.11 illustrates the performance of three models: DATTA with weight resetting ($W_{\text{DATTA+R}}$), DATTA without weight resetting (W_{DATTA}), and DAT without TTA as the baseline (W_{DAT}), across three domain sequences to evaluate adaptability and resilience to domain shifts.

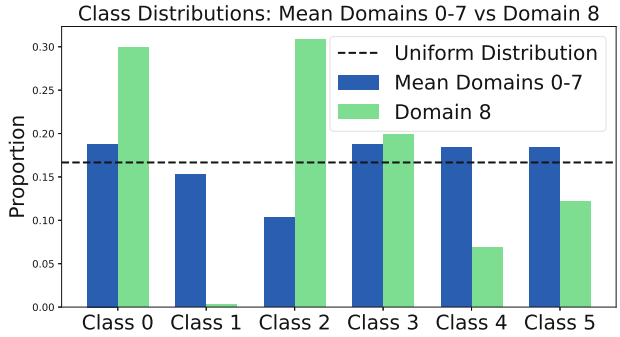


Figure 6.12: Comparison of class distributions between the mean distribution of domains 0–7 and the highly skewed distribution of domain 8.

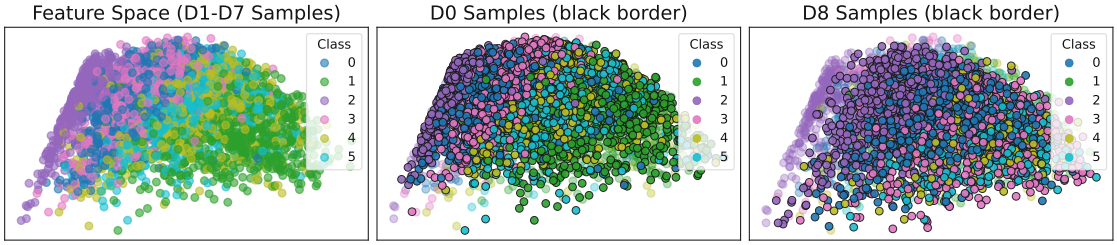


Figure 6.13: Comparison of feature spaces for domains D1–D7, D0, and D8, highlighting the low intra-class cohesion and poor separability of D8 samples.

In the first experiment (top plot), where domains are processed one after another (D0 to D8) to simulate smooth and realistic domain shifts, TTA consistently improves in performance across domains, showing effective incremental adaptation except for the last domain, D8, where it experiences a significant drop. On closer analysis, it is found that D8 has a highly skewed class distribution (with class 1 containing only 8 samples, approximately 0.3% of the total $\approx$ 2,700 samples in D8) and substantially different data statistics compared to the other domains, as shown in Figure 6.12. While these factors may hinder adaptability, it is hypothesized that mislabeled activities are the primary cause of poor performance, potentially compounded by the steady, linear decline in accuracy, which suggests systematic misclassification. To investigate this, PCA analysis is conducted. As shown in Figure 6.13, domain D8 (right) diverges considerably from the nominal feature space defined by domains D1–D7 (left), exhibiting low intra-class cohesion and poor class separability, indicating high label noise and ambiguity. In contrast, a well-behaved domain such as D0 (middle) demonstrates clear and consistent clusters of activities. These factors explain the accuracy drop in D8 due to systematic misclassification. However, TTA effectively adapts to other domains, confirming its capability to overcome participant-induced domain gaps.

In the second experiment (middle plot), domains are processed in reversed order (D8 to D0) to test the ability to recover from domains with stark differences in underlying data statistics. After finishing adaptation to the ill-posed domain, D8, TTA quickly recovers and exceeds baseline performance, demonstrating its resilience in recovering from challenging domains.

Model		ACC ↑	F1-Score ↑
W	+ViTTA [209]	47.76 ±4.33	49.54 ±4.43
W	+DAT [71]	61.87 ±1.39	63.16 ±1.55
W_{TTA}	+DAT [71]	64.48 ±1.58	65.87 ±1.41
W_{DAT}	+ViTTA [209]	64.37 ±2.05	65.60 ±1.94
W	+DAT [71]+ViTTA [209]	61.66 ±0.81	63.04 ±0.92
W_{DATTA}		<u>66.92</u> ±1.54	<u>68.13</u> ±1.45

Table 6.18: SotA comparison against DAT [71] and ViTTA [209]. (W and W_{TTA}) + DAT use the discriminator input $c \oplus \hat{a}$, (W and W_{TTA}) + ViTTA, employ the ℓ_1-loss, and initialize targets with zero, with slightly tuned hyperparameters. Note that all models are trained with augmentation, and W_{DATTA} additionally uses weight resetting.

The third experiment (bottom plot) assesses resilience to catastrophic forgetting by alternating between domains D0 and D2, simulating antagonistic domain shifts. As shown previously, D2 (like D8) negatively affects performance, unlike other domains. This configuration thus represents a challenging setting. Here, $W_{\text{DATTA+R}}$ maintains stable performance across shifts, showcasing robustness in retaining learned features. Conversely, W_{DATTA} shows continual performance degradation, and gradual loss of domain invariance, highlighting the effectiveness of weight resetting to prevent catastrophic forgetting.

Overall, both $W_{\text{DATTA+R}}$ and W_{DATTA} significantly outperform the baseline W_{DAT}, confirming TTA's effectiveness in enhancing cross-domain generalization. Additionally, $W_{\text{DATTA+R}}$ outperforms W_{DATTA} across experiments, validating the role of random weight resetting in preventing performance degradation under repeated, prolonged domain shifts.

State of the Art Comparison To compare the proposed DATTA, with existing SotA methods, it is evaluated against the default DAT model as proposed in [71] and the video-based TTA variant in [209]. As this represents the first port of this video-based TTA variant to the WiFi domain, all parameters are retained to align as closely as possible with the original work, tuning only the most crucial ones: the layers for statistic alignment (choosing the first layer instead of the last two) and the learning rate (1×10^{-6}).

Shown in Table 6.18 are the results of this SotA evaluation; note that all variants were trained using the proposed data augmentation module. When combining ViTTA [209] with W or W_{DAT}, the F1-Score stays almost unchanged (+0.4%/−0.1%), indicating that employing the method as-is has no benefit. Consequently, evaluating it against the proposed TTA variants, namely W_{TTA} and W_{DATTA} without weight resetting, performance is worse (49.54% versus 52.2% and 67.29%), suggesting that adopting ViTTA to the WiFi domain is not straightforward, requiring changes in the loss function, target initialization, and extensive parameter tuning.

Similarly, when combining DAT [71] with W or W_{TTA}, performance is significantly improved over W, though it remains below the proposed DAT variants (63.16% vs. 65.66% and 67.29%), i.e., W_{DAT} and W_{DATTA}. Here, the key difference is the input to the

Model	Parameters	Weight Resetting	Inference Time [ms] RTX 2070	Inference Time [ms] Jetson Orin Nano
W	40.80 k		<u>1.98</u> ±0.21	<u>9.24</u> ±0.25
W_{DAT}	41.97 k		2.05 ±0.23	10.30 ±0.18
W_{TTA}	40.80 k	-	9.86 ±1.42	56.73 ±1.29
	40.80 k	✓	19.52 ±2.27	111.39 ±3.93
W_{DATTA}	41.97 k	-	10.05 ±1.40	57.63 ±1.31
	41.97 k	✓	20.21 ±2.05	115.39 ±3.76

Table 6.19: Inference time comparison for the base model W, TTA model W_{TTA}, DAT model W_{DAT}, and DATTA model W_{DATTA}. Column R indicates the use of random weight resetting. Mean inference time is reported over 1,000 iterations (after 100 warm-up iterations) with batch size 1 on an Nvidia RTX 2070 GPU and a Jetson Orin Nano single-board computer.

discriminator: class token embeddings c concatenated with ([71]) and without (proposed) activity logits $\hat{a}$. It is hypothesized that this is due to $\hat{a}$ inherently including domain information, namely priors over activities, e.g., *lying* being more likely performed in the bedroom than in the office. Hence, by including $\hat{a}$, the activity recognizer is penalized for learning these priors, making predictions less accurate.

In a final evaluation, both SotA methods are combined naively (W+DAT+ViTTA) and compared against DATTA W_{DATTA}: as before, the impact of ViTTA is negligible, resulting in an improvement over the SotA by 8.1%.

Inference Time Table 6.19 compares the inference times for the baseline model W, W_{DAT}, W_{TTA}, and W_{DATTA} on both an *Nvidia RTX 2070* GPU and an *Nvidia Jetson Orin Nano* single-board computer. On the RTX 2070, W achieves the fastest inference time at 1.98 ms per sample, with a minor increase to 2.05 ms for W_{DAT} due to the DAT component. Models incorporating TTA show a notable rise in inference time due to additional adaptation steps, with W_{TTA} reaching 9.86 ms and W_{DATTA} reaching 10.05 ms, a $\approx$ 5x increase over the baseline. With random weight resetting, inference time increases further to approximately 19.52 ms for W_{TTA} and 20.21 ms for W_{DATTA}, roughly a 10x increase.

Switching from the RTX 2070 to the Jetson Orin Nano incurs an additional $\approx$ 5x increase in inference time across models, with W_{DATTA} achieving around 115.39 ms per sample with weight resetting. However, this latency remains sufficient for HAR, achieving $\approx$ 9 frames per second. Although weight resetting adds to inference time, it significantly improves model stability by preventing catastrophic forgetting, which is crucial for sustained performance across domain shifts. Notably, the underlying TTA code is not optimized for speed, and further optimizations are likely to reduce adaptation time.

Discussion

The conducted evaluation validates the effectiveness of DATTA in enhancing cross-domain generalization for WiFi-based PCS. By integrating DAT and TTA, DATTA achieves substantial improvements over existing state-of-the-art methods, with an 8.1% increase in F1-Score on the HAR task, notably outperforming both standalone DAT and traditional TTA approaches adapted from the video domain. The augmentation module, utilizing the effective preprocessing techniques from [50], facilitates DAT, emphasizing its essential role in enabling domain-invariant feature learning. Without effective augmentation, models quickly overfit to domain-specific features, limiting their adaptability.

The employed random weight-resetting mechanism further improves model robustness by effectively mitigating catastrophic forgetting during prolonged adaptation periods. This method stabilizes adaptation performance across diverse and challenging domain sequences, as demonstrated by its resilience to extreme domain shifts (e.g., domain D8, characterized by severe class imbalance and label noise). The effectiveness of random weight resetting depends on initial domain invariance, underscoring the importance of a robust base model established via DAT with augmentation.

Experiments evaluating incremental and antagonistic domain sequences confirm the practical advantage of DATTA over baseline methods. Notably, in scenarios involving alternating domain shifts, weight resetting preserves performance stability, whereas DATTA without resetting experiences steady performance degradation. This outcome emphasizes the critical role of periodic weight resets in maintaining stable domain invariance.

In terms of computational efficiency, DATTA, through the integration of *WiFlexFormer* [53], achieves inference latency suitable for real-time edge deployments. Despite additional computation from TTA and weight resetting mechanisms, DATTA achieves inference times of $\approx$ 10 ms per sample on an Nvidia RTX 2070 GPU and $\approx$ 115 ms per sample on the Nvidia Jetson Orin Nano single board computer, translating to $\approx$ 9 frames per second (**RQ II**). These latencies remain sufficient for practical localization and HAR scenarios, especially considering further potential optimizations in the adaptation implementation.

Overall, DATTA effectively addresses the limitations of previous static preprocessing approaches, responding to **RQ III** by demonstrating robust, real-time adaptability to domain shifts at test time. Its combination of DAT, TTA, targeted augmentations, and random weight resetting constitutes a comprehensive framework that advances the feasibility of practical, domain-invariant WiFi-based PCS deployments in complex, real-world scenarios.

6.4 CSI-based Through-Wall Imaging

Despite advances in model robustness and generalization, a limitation of WiFi-based PCS remains its lack of visual interpretability. Unlike optical sensing modalities [215], CSI encodes signal propagation characteristics in an abstract format that offers little intuitive insight into human activities or environmental context. This limitation constrains the usability of WiFi sensing in applications where visual understanding is critical, such as security monitoring, or human behavior analysis.

To address this challenge, researchers have begun exploring methods to synthesize dense visual representations directly from WiFi signals. Early work primarily focuses on semantic segmentation and pose estimation. Techniques such as *WiSIA*[139] and *Wi-Seg*[145] employ conditional GANs or encoder-decoder architectures to extract masks or silhouettes from CSI. Other efforts, including *Person-in-WiFi*[94] and *DensePose-from-WiFi*[142], estimate human poses from WiFi signals, though typically constrained to intra-room scenarios. More recent approaches extend these capabilities to richer outputs: *Wi2Vi*[140] and *CSI2Image*[141] aim to generate RGB video frames using autoencoders or GANs, and *Wi-Depth* [147] reconstructs depth maps for 3D scene understanding. However, these methods either assume LOS conditions or limit output to semantic representations, leaving the problem of CSI-to-image translation in TW settings largely unexplored.

This gap is addressed by introducing the first approach for directly synthesizing RGB images from WiFi CSI captured in TW scenarios [51]. Drawing inspiration from multimodal learning frameworks used in brain decoding [216], the method employs a generalized multimodal VAE [217] tailored to crossmodal translation from CSI amplitude spectrograms to RGB images. This architecture enables aligned generation of visual representations from abstract WiFi signals, facilitating human interpretability and unlocking image-based downstream tasks such as activity annotation or visual inspection beyond visual line of sight, an without the violation of visual privacy [20].

By leveraging WiFi's penetrative properties in combination with generative modeling, this approach expands the interpretability and application scope of WiFi-based PCS. In the context of **RQ IV**, it represents a key contribution toward making CSI semantically meaningful and visually intuitive, laying the foundation for more accessible and trustworthy real-world deployments.

6.4.1 WiFiCam Architecture

Figure 6.14 provides an overview of the proposed architecture for synthesizing images from TW WiFi CSI, including both the training and inference stages. The approach relies on the MoPoE-VAE [217], a multimodal variational autoencoder designed to learn a posterior distribution over a joint latent variable $\mathbf{z} \in \mathbb{R}^D$ conditioned on both CSI and image modalities. A corresponding decoder reconstructs images from sampled latent vectors derived from this joint distribution. A PyTorch implementation of the proposed architecture is publicly available[3].

[3]WiFiCam, https://github.com/StrohmayerJ/wificam, accessed: 15.04.2025

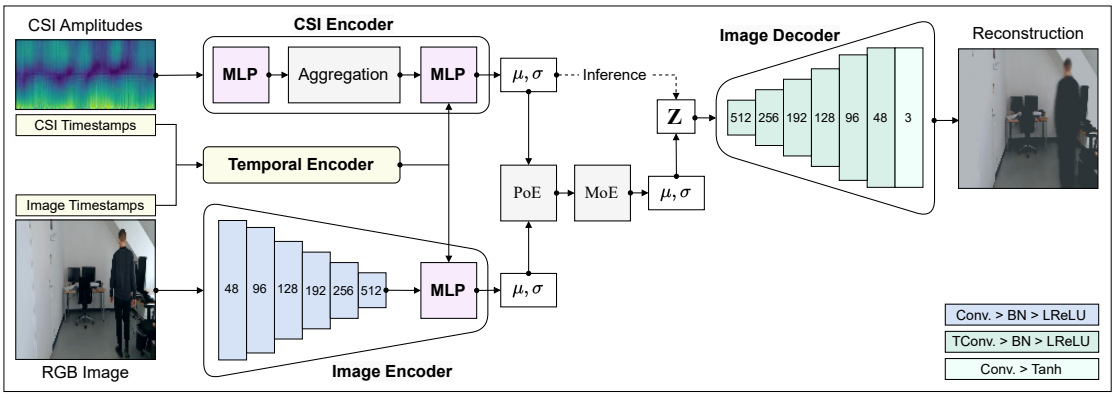


Figure 6.14: Proposed WiFiCam architecture for WiFi CSI-based image synthesis. [51]

Given a set of samples of images $\mathbb{X} := \{\mathbb{X}^i\}_{i=1}^N$, each paired with a fixed time interval of WiFi packets (the central packet determines the corresponding image), i.e., $\mathbb{X}^i = \{\mathbf{X}_I^i,\ \mathbf{X}_W^i\}$, the aim is to maximize the log-likelihood of the data at hand:

$$\log p_\theta(\mathbb{X}) = \log p_\theta(\{\mathbb{X}^i\}_{i=1}^N) \tag{6.21}$$

Within VAEs, this is achieved by maximizing a lower bound to this objective, the so-called evidence lower bound. For a given sample $\mathbb{X}^i$, this lower bound on the marginal log-likelihood takes on the following form:

$$\mathcal{L}(\theta, \phi; \mathbb{X}^i) := \mathbb{E}_{q_\phi(\mathbf{z}|\mathbb{X}^i)}\big[\log\left(p_\theta(\mathbb{X}^i|\mathbf{z})\right)\big] - \beta D_{KL}(q_\phi(\mathbf{z}|\mathbb{X}^i)||p_\theta(\mathbf{z})), \tag{6.22}$$

where D_{KL} denotes the Kullback-Leibler divergence [218] between the approximated posterior and the assumed, in our case, Gaussian prior, and β being an additional weight parameter (see Higgins et al. [219]), which promotes disentanglement of the latent variable $\mathbf{z}$, in the case that $\beta > 1$. During inference, when image data is missing, i.e., $\mathbb{X}_W^i := \mathbb{X}^i \setminus \{\mathbf{X}_I^i\}$, a valid lower bound is still to be obtained on the joint probability $\log p_\theta(\mathbb{X}^i)$. However, when using $\mathcal{L}(\theta, \phi; \mathbb{X}_W^i)$, it only yields a lower bound on $\log p_\theta(\mathbb{X}_W^i)$. Hence, to obtain a correct lower bound on the joint probability, the following adapted lower bound is used:

$$\mathcal{L}_W(\theta, \phi_W; \mathbb{X}^i) := \mathbb{E}_{\tilde{q}_{\phi_W}(\mathbf{z}|\mathbb{X}_W^i)}\big[\log\left(p_\theta(\mathbb{X}^i|\mathbf{z})\right)\big] - \beta D_{KL}(\tilde{q}_{\phi_W}(\mathbf{z}|\mathbb{X}_W^i)||p_\theta(\mathbf{z})). \tag{6.23}$$

In the general naive case with M different modalities, approximating a lower bound of the joint probability in any case of missing modalities requires the optimization of 2^M different models, one for each subset contained within the powerset, posing a drastic scalability issue. Compared to prior literature, Sutter et al. [217] circumvent this by modeling the joint posterior approximation as a so-called Mixture of Products of Experts (MoPoE), combining Product of Experts (PoE) [220] and Mixture of Experts (MoE) [221] through abstract mean functions [222], to enable efficient retrieval of the joint posterior for all subsets. A detailed description of this original approach can be found in [217].

While the task at hand involves only two modalities (WiFi CSI and images) and focuses exclusively on reconstructing images from WiFi CSI, the methodology is designed to remain extensible to additional modalities in future work. To this end, the framework retains the multimodal structure of MoPoE-VAE. However, custom variational autoencoders are introduced for inference from WiFi CSI, and the loss function is adapted to disregard image sequence decoding, as it is not required in the current setting.

Image Reconstruction

Given the stark difference of modalities, two architectural different models are required for learning the unimodal posterior distributions $q_{\phi_I}(\mathbf{z}|\mathbb{X}_I^i)$ and $q_{\phi_W}(\mathbf{z}|\mathbb{X}_W^i)$:

Image VAE For the encoding and decoding of images, convolutional and transpose-convolutional layers are employed, respectively. Additionally, each of these layers is followed with batch normalization and a Leaky ReLU activation. Before inferring the distribution parameters of the latent variable $\mathbf{z}$ given images using the encoder, the input is rescaled from a resolution of 640×480 down to 128×128 pixels in order to reduce computational complexity and apply normalization using per-channel means and standard deviations. Next, six consecutive convolutional blocks increase the channel size from three (RGB) to 512, followed by a simple Multi-Layer Perceptron (MLP) to map from the flattened output of the convolutions to the unimodal distribution parameters. Decoding is performed by taking a latent vector $\mathbf{z}$ and reversing the process using transpose-convolutions.

CSI VAE The CSI VAE first extracts amplitude information from the raw WiFi CSI and applies a MLP to embed each sample in a sequence, producing a more expressive feature representation. To handle the temporal dimension, several aggregation strategies are considered: uniform feature weighting, Gaussian feature weighting, and concatenation, where the sequence is flattened without further modification. These aggregation methods are evaluated through an ablation study. Following aggregation, a second MLP estimates the unimodal distribution parameters for the WiFi CSI. Reconstruction of the CSI from the latent variable $\mathbf{z}$ is omitted, as it is not required for the task at hand.

During training, the unimodal distribution parameters predicted by the two VAEs are first combined into subsets using the PoW approach and then aggregated across the powerset using a MoE formulation. Given the predicted parameters $\boldsymbol{\mu}, \boldsymbol{\sigma} \in \mathbb{R}^D$, a latent vector $\mathbf{z}$ is sampled using the reparameterization trick: $\mathbf{z} = \mu + \sigma \odot \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$. At inference time, only the mode of the approximated posterior distribution produced by the CSI VAE is used for reconstruction, specifically the predicted mean.

Aggregation Options In the CSI VAE, several aggregation strategies are evaluated to identify the most informative representation for predicting an image corresponding to the central WiFi packet in a given sequence. Let $\tilde{\mathbf{X}}_W \in \mathbb{R}^{L \times H}$ denote the embedded WiFi CSI amplitudes after the first encoder MLP, where L is the sequence length and H the hidden dimension. One baseline approach applies a uniform weighing, assigning equal importance to all packets and thereby disregarding packet order or the prominence

of the central packet. To emphasize temporal structure, a Gaussian weighing centered at the middle of the sequence is considered, using $\mu = \sigma = L/2$ to prioritize packets near the center and reduce permutation invariance. Both of these aggregation strategies may result in temporal instability, such as flickering between frames. To address this, a third strategy based on concatenation is introduced, in which packet features are left unaggregated to preserve full permutation sensitivity. Depending on the chosen aggregation method, the input to the second encoder MLP lies either in $\mathbb{R}^H$ (for weighed approaches) or $\mathbb{R}^{LH}$ (for concatenation).

Temporal Encoding A final adjustment to the architecture addresses the issue of temporal instability through the use of temporal encoding. Although concatenation helps reduce abrupt changes between consecutive frames, minor jittering artifacts may still persist. To further mitigate this, sinusoidal temporal encoding is introduced, following a strategy similar to that used in NeRF [223]. This approach augments the input features with explicit time information. For a given timestamp t and a set of frequencies L, the temporal encoding is defined as:

$$T(t, F) = \Big[\sin\Big(\frac{2^0 \pi t}{3L}\Big), \cos\Big(\frac{2^0 \pi t}{3L}\Big), \ldots, \sin\Big(\frac{2^{F-1} \pi t}{3L}\Big), \cos\Big(\frac{2^{F-1} \pi t}{3L}\Big)\Big]. \tag{6.24}$$

In this formulation, the timestamp t is scaled by three times the window size L to incorporate contextual timing information from before and after the current window. When applying temporal encoding, the encoded time features are concatenated with the image or WiFi features produced by the CNN or MLP, respectively. This combined feature vector is then passed through the corresponding MLP to predict the distribution parameters, as illustrated in Figure 6.14.

6.4.2 Evaluation Setup

To assess the relationship between reconstruction fidelity and architectural design choices, an ablation study evaluating both quantitative metrics and qualitative image characteristics across different model configurations, is conducted.

Data The evaluation of the proposed TW image synthesis approach relies on the WiFiCam dataset (see Section 5.2.5), which is specifically designed to enable the direct translation of WiFi CSI into RGB images in a TW scenario. Unlike traditional person-centric datasets that include explicit activity or trajectory labels, WiFiCam focuses on temporally aligned CSI–image pairs, making it well-suited for studying visual interpretability and image synthesis from CSI in realistic indoor sensing conditions. The dataset is recorded in a furnished office room measuring 3.8 m $\times$ 5.3 m, with a point-to-point WiFi setup based on system $\mathcal{D}$ placed outside opposite walls to establish a single-room TW configuration. An RGB camera placed inside the room captures 640 $\times$ 480 images at 30 Hz while WiFi packets are transmitted at 100 Hz. Synchronization is performed via timestamps, resulting in approximately three WiFi packets per image.

The dataset comprises 57,413 WiFi packets and 18,261 RGB images collected during a ten-minute sequence of continuous walking. Data cleaning removes samples at the start

Model	PSNR ↑	SSIM ↑	RMSE ↓	FID ↓
UW	20.03 ±.05	0.734 ±.00	8.02 ±.04	142.95 ±2.4
GW	20.02 ±.07	0.734 ±.00	8.00 ±.05	141.00 ±5.0
C	20.39 ±.05	0.748 ±.00	7.83 ±.05	127.33 ±4.7
C+T	20.39 ±.04	0.749 ±.00	7.80 ±.02	125.62 ±3.2

Table 6.20: Quantitative ablation study results for the aggregation options *uniform weighing* (UW), *Gaussian weighing* (GW), *concatenation* (C), and *concatenation* with *temporal encoding* (C+T). The metrics reported represent the mean and standard deviation across ten independent training runs.

and end of the recording that do not correspond to the target activity. For training, validation, and testing, the dataset is split at an 8:1:1 ratio. CSI amplitude spectrograms are constructed using the magnitudes of 52 L-LTF subcarriers over a 151-packet window ($\approx$ 1.5 seconds), and each spectrogram is paired with the RGB image closest in time to the center packet. This forms the input-output pairs used for model training and evaluation.

Model Training Models are trained using the proposed WiFiCam architecture for CSI-based image synthesis, incorporating the following architectural variants: *uniform weighing*, *Gaussian weighing*, *concatenation*, and *concatenation with temporal encoding*.

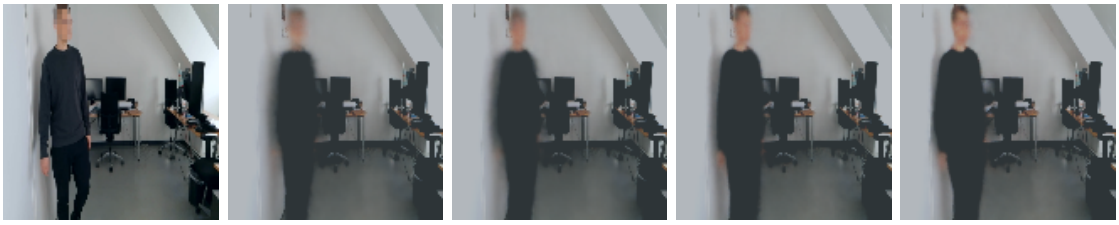
A hyperparameter search is performed using the *concatenation with temporal encoding* configuration on the validation subset to identify optimal training parameters. The search explores batch sizes $b \in \{16, 32, 64, 128, 256, 512\}$, window sizes $L \in \{51, 101, 151, 201, 251, 301\}$, and KL divergence weights $\beta \in \{1, 2, 4, 6\}$. The optimal values found are $b = 32$, $L = 151$, and $\beta = 1$, and these settings are subsequently used across all model variants.

For each WiFi packet, amplitudes of the 52 L-LTF subcarriers are extracted from the CSI matrix and used to construct $52 \times L$ spectrograms as input to the CSI VAE. Raw RGB images are resized to 128×128 pixels and normalized using per-channel dataset statistics. All models are trained using the Adam optimizer with a learning rate of 1×10^{-3}. Each configuration is trained independently ten times for 50 epochs, and the model achieving the lowest validation loss is selected for evaluation on the test subset.

Metrics To evaluate image reconstruction fidelity, the assessment employs widely used metrics from compression and generative modeling literature, including Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Root Mean Squared Error (RMSE), and Fréchet Inception Distance (FID)[224].

6.4.3 Results

Quantitative Results Table 6.20 presents the quantitative results of the ablation study, comparing MoPoE-VAE variants using *uniform weighing*, *Gaussian weighing*, *concatenation*, and *concatenation with temporal encoding*. Evaluation is based on PSNR, SSIM, RMSE, and FID, averaged across ten independent training runs. Among the



(a) ground truth (b) UW (c) GW (d) C (e) C+T

Figure 6.15: Comparison of reconstruction fidelity between MoPoE-VAE models employing the aggregation options (b) *uniform weighing* (UW), (c) *Gaussian weighing* (GW), (d) *concatenation* (C), and (e) *concatenation* with *temporal encoding* (C+T). This visual comparison highlights the improvements in image clarity and reduction of artifacts. [51]

configurations, *uniform weighing* and *Gaussian weighing* yield the weakest performance, characterized by lower PSNR and SSIM values and elevated RMSE and FID scores.

Introducing *concatenation* leads to notable improvements across all metrics, indicating that preserving temporal order enhances the model's ability to reconstruct images more accurately. This is further supported by qualitative improvements, including enhanced sharpness and perceptual fidelity. The *concatenation with temporal encoding* configuration surpasses all other variants, yielding marginally higher PSNR and SSIM, along with lower RMSE and FID values compared to *concatenation*. In particular, the improved FID score suggests reduced perceptual distance and enhanced image quality.

Although the addition of temporal encoding produces only subtle quantitative gains, it contributes to improved frame coherence and visual sharpness. Overall, the results indicate that while *concatenation with temporal encoding* offers the strongest reconstruction performance, qualitative assessment remains essential for a complete evaluation, especially when differences in numerical metrics are modest.

Qualitative Results The qualitative analysis supports the quantitative findings presented earlier. Figure 6.15 illustrates how different aggregation strategies affect image reconstruction fidelity. The *uniform weighing* model (Figure 6.15b) produces noticeably blurred images, consistent with its lower PSNR and higher RMSE values reported in Table 6.20. This outcome is attributed to a lack of temporal focus, which appears to diminish perceptual clarity. The *Gaussian weighing* model (Figure 6.15c) exhibits a slight visual improvement despite comparable quantitative metrics, likely due to the increased weighting of central WiFi packets, those temporally closest to the target image, which contribute more prominently to the reconstruction.

In Figure 6.15d, the *concatenation* model further improves reconstruction quality. By preserving full temporal resolution through feature concatenation, the model learns to emphasize relevant signal patterns, resulting in sharper images and more coherent frame-to-frame transitions. These improvements are also reflected in Figure 6.16, where the reduction of spatiotemporal discontinuities is evident. Although *concatenation* outperforms both *uniform weighing* and *gaussian weighing* quantitatively, minor artifacts

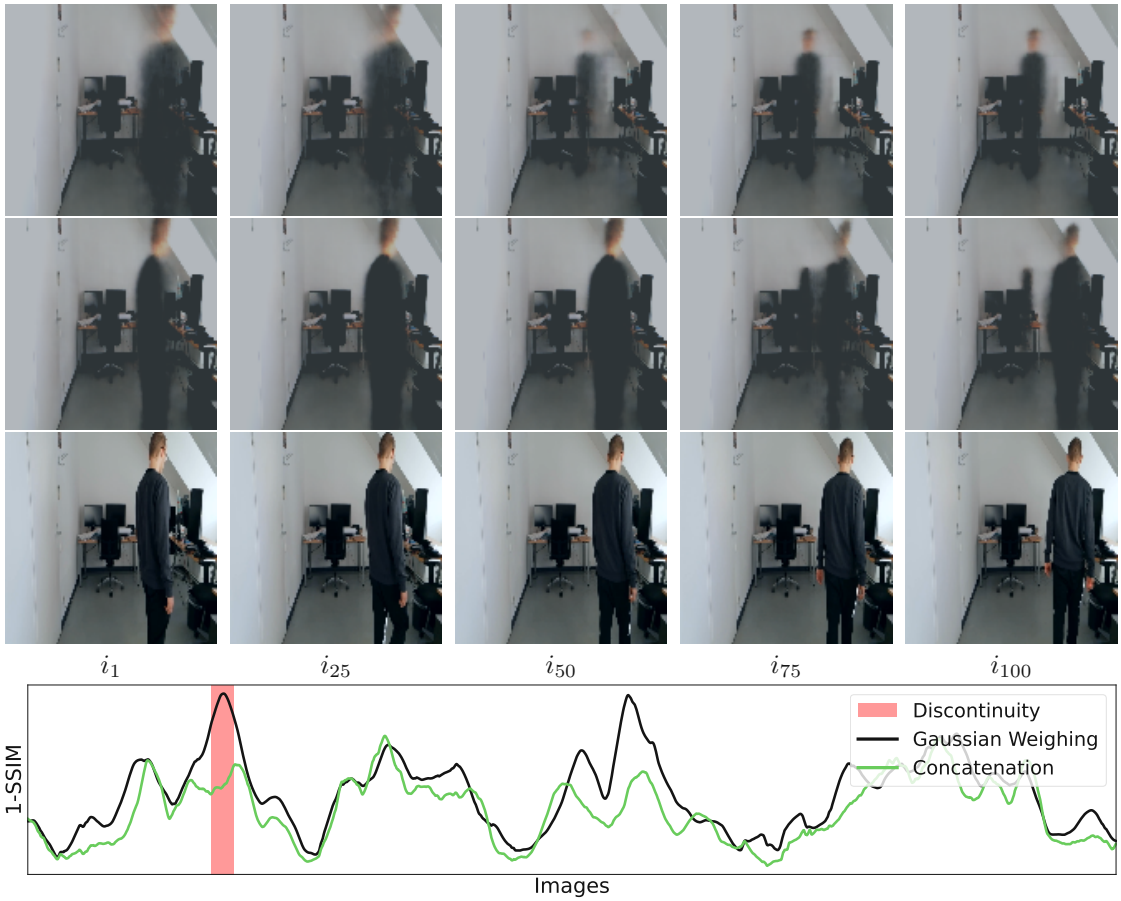


Figure 6.16: Example showing the elimination of spatiotemporal discontinuities in a sequence of 100 test images with high 1-SSIM, highlighted in red, through the aggregation option *concatenation*. From top to bottom, the rows show *Gaussian weighing*, *concatenation*, and ground truth, respectively. [51] †

and temporal jitter remain visible in video sequences, consistent with the marginal differences observed in FID scores.

The best results are achieved by the *concatenation with temporal encoding* model (Figure 6.15e). The addition of cyclic temporal encoding further reduces jitter, enhances frame consistency, and lowers perceptual distance as reflected in FID. Reconstructed video sequences[3] further illustrate these improvements by highlighting the temporal stability achieved by each configuration.

Overall, the qualitative findings confirm that the proposed method enables effective TW visual monitoring using WiFi CSI. The ability to reconstruct coherent and visually meaningful images without conventional cameras highlights the feasibility of this approach and its potential for image-based downstream tasks. This work represents a significant advancement in improving the interpretability of WiFi CSI and offers a solid foundation for further exploration in CSI-to-image translation.

6.4.4 Discussion

This work demonstrates the feasibility of synthesizing person-centric RGB images from WiFi CSI captured in TW scenarios, marking the first approach to enable visual monitoring across walls using WiFi alone. An extensive ablation study evaluates several WiFiCam architecture variants and reveals that temporal feature concatenation, combined with cyclic temporal encoding, yields the most consistent and perceptually accurate image reconstructions. This configuration outperforms alternatives in both quantitative metrics and qualitative visual fidelity, producing sharper, temporally coherent frames suitable for image- and video-based downstream tasks.

By directly translating abstract CSI amplitude spectrograms into intuitive visual representations, the proposed approach improves the interpretability of WiFi-based sensing. It bridges the semantic gap between RF signals and human-understandable imagery, making WiFi sensing more accessible and practically useful in scenarios such as TW security monitoring or visual activity annotation, domains where conventional camera systems are often intrusive or unfeasible.

In relation to **RQ IV**, this work provides a key advancement in making CSI interpretable by design, rather than relying solely on performance metrics or latent activations. It shows that CSI can not only support predictive tasks but also enable generative, human-interpretable outputs that enrich the usability and transparency of WiFi-based PCS systems.

While the WiFiCam model (i.e., MoPoE-VAE+C+T) achieves strong results, its generalization to unseen domains or subjects remains unexplored. The domain adaptation techniques introduced earlier, particularly the DATTA framework (see Section 6.3.3), offer promising pathways to enhance robustness by enabling domain-invariant representation learning and domain-adaptation at test time without requiring labeled data.

Looking forward, the inherently multimodal design of the proposed WiFiCam architecture opens opportunities to integrate additional sensing modalities such as radar, acoustic signals, or structural vibration data, to complement WiFi and enhance performance in degraded or ambiguous conditions. These extensions could further expand the applicability of TW image synthesis in complex real-world settings, establishing a robust and interpretable alternative to conventional visual systems.

Summary

This chapter presents a unified methodology to enable practical, robust, and interpretable WiFi-based PCS across long-range and TW sensing scenarios. It addresses the core challenges identified in **RQ I–IV** through four complementary contributions.

First, to address **RQ I**, the chapter demonstrates that single-link directional antenna systems based on low-cost COTS WiFi devices enable reliable long-range and TW PCS in partitioned indoor environments. The proposed systems demonstrate robust presence detection, HAR, and localization performance across distances up to 20 meters and multiple walls, offering a scalable alternative to conventional short-range solutions while minimizing system complexity and cost.

Second, to address **RQ II**, the chapter introduces *WiFlexFormer*, a lightweight Transformer-based architecture optimized for real-time inference on edge devices. With only $\approx$ 50k parameters and $\approx$ 10 ms inference latency, *WiFlexFormer* maintains competitive HAR performance while supporting advanced adaptation strategies such as test-time training and subcarrier sub-sampling, enabling efficient CSI processing in embedded environments.

Third, in response to **RQ III**, the chapter systematically investigates algorithmic strategies to improve cross-domain generalization. Data augmentation and preprocessing techniques are shown to improve model generalization across hardware, scenario, environmental, and temporal variations without requiring access to target-domain data. Furthermore, the proposed DATTA framework introduces a novel integration of domain-adversarial training and test-time adaptation, enabling real-time domain adaptation on the edge at test time. This approach outperforms prior methods and provides a scalable foundation for domain-invariant WiFi-based PCS in dynamic, unpredictable environments.

Finally, to address **RQ IV**, the chapter presents the first approach to synthesize RGB images directly from WiFi CSI in TW settings. By leveraging a multimodal VAE with temporal encoding, the proposed method enables coherent, interpretable visual reconstructions from RF signals, bridging the semantic gap between CSI and human-readable imagery. This represents a major step toward interpretable and camera-free monitoring applications and opens new opportunities for image-based downstream tasks.

CHAPTER 7

Discussion

This chapter examines the broader implications, limitations, and future directions of WiFi-based PCS. While recent advances improve sensing capabilities through novel system designs, datasets, and methodologies, achieving scalable and practical adoption still requires addressing persistent challenges in cross-domain generalization, dataset availability, multi-person interference, and dual-use concerns. Further progress may result from the development of specialized communication protocols, the integration of complementary sensing modalities, and the expansion of application domains.

7.1 Current Limitations

Despite recent advancements, WiFi-based PCS still faces significant barriers to widespread practical adoption, particularly regarding cross-domain generalization, dataset availability, and multi-person interference.

Cross-Domain Generalization While substantial progress exists toward improving cross-domain generalization in WiFi-based PCS, it remains one of the most persistent challenges. Due to the inherent sensitivity of CSI to domain variations, including changes in physical environment, hardware configuration, or user physiology, model performance tends to deteriorate in unseen domains [47]. To address this, recent research explores a variety of strategies including domain-invariant feature extraction, physics-informed data augmentation, transfer and few-shot learning, and large-scale multi-domain training [47]. These approaches aim to either suppress domain-specific artifacts or improve model adaptability across diverse conditions. Despite notable gains, no existing method consistently generalizes across the full range of real-world deployment scenarios. Static models still require fine-tuning or retraining when faced with even minor domain shifts, highlighting the fragility of existing solutions and the need for complementary, dynamic strategies to improve robustness [207, 209]. However, achieving truly domain-agnostic performance remains an open problem. A promising path forward may lie in the creation of large-scale datasets that comprehensively capture real-world domain variability, coupled

with domain-invariant feature learning and dynamic test-time adaptation methods, exemplified by DATTA [54].

Data Scarcity and Standardization Limited availability of diverse and standardized WiFi datasets substantially constrains the evaluation and improvement of PCS models. Acquiring extensive, annotated WiFi datasets involves complex setups, human participants, and resource-intensive labeling procedures, often dependent on additional visual modalities. Recent datasets partially mitigate these issues by increasing environmental variability [183, 44], yet current benchmarks remain insufficient to represent the full complexity of real-world conditions [77]. Additionally, the limited hardware support for CSI extraction further restricts research and deployment efforts [42]. Establishing standardized data collection protocols and promoting broader hardware compatibility are essential steps toward robust and generalizable PCS systems.

Signal Interference in Multi-Person Scenarios Research on WiFi-based PCS typically targets single-user scenarios [125], but practical deployments can involve multiple individuals interacting simultaneously within the same environment. Such scenarios complicate PCS tasks, as person-specific perturbations are mixed up in the WiFi signal, causing interference and making signal isolation challenging. Consequently, existing PCS approaches tailored to single-user scenarios perform poorly under multi-person conditions. While methods employing blind source separation, such as ICA in systems like *MultiSense* [110], provide partial solutions, comprehensive approaches to robustly handle general multi-person sensing without relying on complex MIMO setups [98] remain to be developed. Although emerging multi-person datasets such as WiMANS [44] are valuable resources, methodological advancements in blind source separation needed to fully address this limitation.

7.2 Dual-Use Potential

The possibility of non-consensual monitoring using WiFi signals is illustrated by recent studies demonstrating passive sensing from outside a building. Hernandez and Bulut [160], for example, show that occupancy and movement direction can be detected through walls using low-cost COTS WiFi devices. Through a co-planar arrangement of transmitter and receiver devices on the exterior of a building and statistical CSI aggregation, their system enables presence detection without physical access. Abedi and Vasisht's *Wi-Peep* [225] further demonstrates that a flying drone equipped with a low-cost WiFi transceiver can localize WiFi-enabled devices, such as smartphones, inside a building. Since mobile phones are typically carried by individuals, this approach indirectly enables localization of persons from outside the premises. These examples highlight the feasibility of passive, infrastructure-free WiFi sensing for adversarial purposes, motivating the need for safeguards. These could come in the form of privacy-by-design strategies such as CSI randomization at the physical or medium access control layers which can introduce unpredictable perturbations in the signal to prevent unauthorized systems from extracting stable CSI [226]. This technique preserves connectivity for intended devices while degrading the CSI's utility for passive eavesdroppers, effectively obfuscating

activity-related signal patterns without affecting communication quality. Furthermore, hardware-based privacy measures like intelligent reflective surfaces [227] dynamically control electromagnetic wave propagation, effectively enabling targeted privacy zoning (e.g., for bedrooms or bathrooms) without disabling WiFi connectivity.

7.3 On the Future of Wireless Sensing

While the potential of WiFi-based PCS is evident, its full scope remains largely unexplored. Future progress may be driven by new communication protocols optimized for PCS, integration with complementary modalities for more accurate and interpretable inference, and deployment in diverse application domains where scalability, privacy, and cost are critical.

Novel WiFi Standards Emerging wireless communication standards offer new opportunities for extending the capabilities of WiFi-based PCS. For example, IEEE 802.11ah (WiFi HaLow)[1], designed for long-range, low-power data transmission, operates in sub-GHz frequency bands (863-870 MHz in Europe), enabling improved penetration of building materials and coverage of larger indoor areas. These characteristics make it a promising candidate for wide-area PCS deployments, such as whole- or cross-building PCS, using minimal infrastructure. Looking further ahead, there is significant potential in the development of wireless communication protocols tailored to the requirements of PCS. Unlike general-purpose standards, such protocols could be designed to maximize sensitivity to human-induced signal perturbations, increase information density, and support richer spatiotemporal resolution, thereby enhancing the versatility and accuracy of PCS systems across diverse applications. The *IEEE 802.11bf Task Group*[2] explicitly aims to develop such an amendment, proposing PCS-oriented modifications to the IEEE 802.11 family of standards. The 802.11bf amendment, expected to become operational by the end of 2025, defines enhancements to the medium and access control and physical layers, facilitating capabilities such as explicit signaling of sensing functionality, structured sensing measurements, and standardized feedback protocols [228].

6G Integrated Sensing and Communication Looking beyond WiFi-centric standards, the forthcoming sixth-generation (6G) mobile networks are expected to integrate native sensing and communication (ISAC) capabilities, enabled by ultra-wide bandwidths and sub-THz operation [229]. These capabilities promise unprecedented spatial resolution and multi-user tracking in LOS conditions, unlocking new possibilities for high-precision, privacy-aware PCS in complex environments such as smart city infrastructures [230]. In smart buildings, ISAC could support camera-free presence detection, fall detection, and occupancy-aware automation, contributing to both energy efficiency and occupant safety [231]. In healthcare, continuous and contactless monitoring of vital signs and mobility patterns could facilitate early risk detection, longitudinal health assessment, and rehabilitation support [232, 233]. Human–computer interaction stands to benefit from fine-grained gesture, pose, and even emotion recognition for immersive, privacy-preserving

[1] IEEE 802.11ah, https://www.wi-fi.org/discover-wi-fi/wi-fi-certified-halow, Accessed: 24.07.2025
[2] IEEE 802.11bf, https://www.ieee802.org/11/Reports/tgbf_update.htm, Accessed: 24.07.2025

interfaces [234]. Safety systems may leverage ISAC to locate individuals through smoke, fog, or occlusion, where optical sensors such as LiDAR and depth cameras fail [235, 236]. Industrial and logistics settings could use joint sensing and communication for zone-based worker safety, contactless robot collaboration, and real-time asset localization [237]. In public and commercial spaces, ISAC may enable population analytics, personalized services, and queue estimation while safeguarding user anonymity [238]. Across these domains, 6G ISAC envisions a new class of ambient, camera-free intelligence that fuses connectivity with awareness. Notably, recent WiFi-based systems have already demonstrated the feasibility of many such applications using COTS hardware and narrowband sensing, providing a practical foundation upon which future ISAC systems can build.

Multimodal Sensing Multimodal sensing, i.e., the fusion of WiFi CSI with complementary sensor modalities (images, radar, sound, etc.), offers another direction for advancing WiFi-based PCS [239]. Each modality captures different aspects of human behavior, and their integration can help overcome the individual limitations of unimodal systems [215]. For example, WiFi-based sensing preserves visual privacy and possesses wall-penetration capabilities, but on the other hand lacks spatial resolution and semantic richness. In contrast, vision-based systems offer high spatial and contextual resolution but are limited by LOS, lighting, and privacy constraints. Assuming the violation of visual privacy is not an issue, fusing these modalities could yield more accurate, robust, and interpretable representations of person-centric information, especially in complex multi-person, or TW sensing scenarios.

Beyond performance gains, multimodal sensing can simplify annotation and labeling by leveraging modalities with well-established ground truth extraction methods, such as vision. Furthermore, it facilitates the learning of cross-modal representations that transfer person-centric semantics to otherwise opaque RF signals (such as CSI), as explored in [51]. In addition to demonstrating promising results for vision-based inference from CSI, the study shows that translating between visual and RF representations can also enhance the interpretability of CSI, highlighting the potential of vision to contextualize and semantically enrich RF-based sensing for downstream tasks. Recent datasets, such as MM-Fi [183], further illustrate the growing interest in multimodal PCS, providing benchmarks that enable joint training and evaluation across modalities. The availability of such datasets enables PCS systems to utilize hybrid architectures that can intelligently integrate complementary sensing modalities, offering not only improved performance but also more transparent and trustworthy inferences.

Application Scope The versatility of WiFi-based sensing opens largely untapped opportunities for integration into existing wireless infrastructure, setting the stage for pervasive, unobtrusive monitoring across a broad spectrum of domains. While current research focuses primarily on home and healthcare scenarios [42, 240, 4], future applications are poised to extend far beyond, addressing challenges in domains where conventional sensing solutions are limited by cost, scalability, or privacy concerns. In environmental contexts, WiFi signals could support smart agriculture through crop monitoring [241, 242] and enable scalable livestock and wildlife tracking [243]. Similar methods may facili-

tate fault detection and safety analytics in industrial settings by capturing vibration patterns and worker-machine interactions [10]. In automotive applications, in-cabin sensing using WiFi could provide non-intrusive monitoring of driver alertness and passenger well-being [244, 161, 245]. Smart environments may benefit from occupancy-aware building management systems that leverage ambient WiFi signals for real-time energy optimization [246, 247], while public infrastructure could integrate WiFi-based anomaly detection for early warning of hazards such as fire or smoke [8, 9]. Finally, the penetrative nature of WiFi signals holds promise for public safety and disaster response, where future systems may enable the detection of survivors in obstructed environments during emergency operations [11].

WiFi-based PCS holds the potential to become an ubiquitous, privacy-aware alternative to traditional sensing technologies. While current systems face challenges related to generalization, scalability, and ethical deployment, ongoing advancements in wireless protocols, multimodal fusion, and adaptive learning pave the way toward robust, interpretable, and unobtrusive sensing. As these capabilities mature, WiFi-based PCS may transform everyday environments into intelligent spaces that sense and respond to human behavior, seamlessly, securely, and at scale.

CHAPTER 8

Conclusion

WiFi-based PCS has emerged as a promising alternative to traditional optical modalities by enabling passive, device-free monitoring through existing wireless infrastructure. Leveraging intrinsic properties such as cost-efficiency, unobtrusiveness, and the ability to penetrate building materials, WiFi-based PCS is particularly suited to large-scale indoor monitoring applications. Despite these appealing attributes, widespread practical adoption remains constrained by critical challenges, notably the scarcity of public datasets, limited effective sensing range of COTS WiFi systems, computational inefficiencies that hinder on-device real-time inference, poor cross-domain generalization, and the absence of intuitive, visually interpretable data representations, which restrict downstream usability.

To address the scarcity of publicly available datasets and provide a robust foundation for experimentation, five CSI-based PCS datasets were contributed: TOA, Wallhack1.8k, HALOC, 3DO, and WiFiCam. Each dataset was designed to address distinct evaluation requirements, including long-range and TW sensing, cross-scenario and cross-system variation, environmental and temporal domain shifts, and synchronized multimodal data for image synthesis.

Leveraging this infrastructure, remaining limitations were addressed through four core contributions, each grounded in extensive empirical evaluation and directly aligned with the central research questions.

To address **RQ I**, the feasibility of long-range TW PCS using low-cost COTS WiFi systems was established through comprehensive experiments leveraging the proposed systems and datasets. Evaluations of systems $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}1$ on the TOA and Wallhack1.8k datasets demonstrated consistently high presence detection accuracy exceeding 98% across both LOS and TW scenarios. For the HAR task, average accuracies of 92.33% in LOS and 88.99% in TW conditions were achieved, with system $\mathcal{B}$ notably outperforming $\mathcal{C}1$ under challenging TW conditions. Additionally, system $\mathcal{D}$ achieved sub-meter localization accuracy (0.197 m RMSE) over a 20-meter LOS corridor using CSI amplitude features,

validating its suitability for long-range indoor tracking. Importantly, these results were obtained using a standard CNN backbone not explicitly tailored to CSI, suggesting that significantly higher performance may be achievable with advanced architectures tailored to the characteristics of WiFi CSI. This motivates the development of dedicated architectures, such as the *WiFlexFormer* presented subsequently.

To answer **RQ II**, addressing the requirement of computational efficiency in real-world PCS deployments, *WiFlexFormer*, a lightweight Transformer-based architecture tailored to the efficient processing of WiFi CSI, was introduced. Unlike generic vision or RF-specific architectures, *WiFlexFormer* was tailored to exploit the temporal and spectral characteristics of CSI, achieving competitive HAR performance with only $\approx$ 50k parameters. Extensive evaluations demonstrated that it enables real-time inference on embedded platforms such as the *Nvidia Jetson Orin Nano*, reaching latencies of $\approx$ 10 ms, three orders of magnitude smaller in model size than many existing baselines. Its compact design not only supports real-time PCS but also opens the door for complementary, dynamic adaptation to domain shifts at test time without sacrificing computational constraints. By bridging the gap between sensing capability and edge deployment feasibility, *WiFlexFormer* represents a key enabling step for practical and scalable WiFi-based PCS.

In response to **RQ III**, the critical and persistent challenge of poor cross-domain generalization, a key obstacle preventing the reliable deployment of WiFi-based PCS in real-world environments, was addressed. Models trained in a single domain often suffer significant performance degradation when applied to unseen settings due to the high sensitivity to changes in environment, hardware, or subject morphologies. To address this issue, multiple strategies were explored, including data augmentation, feature selection, feature scaling, and dimensionality reduction. Experiments using the Wallhack1.8k and 3DO datasets revealed that while some of these methods, particularly amplitude-based features and temporal augmentations, improve robustness, they remain insufficient for generalization across complex, real-world domain shifts. To overcome these limitations, DATTA, a novel framework that integrates domain-invariant feature learning during training with dynamic adaptation at test time, was introduced. Leveraging the proposed lightweight *WiFlexFormer* backbone, DATTA demonstrated substantial gains over baseline and state-of-the-art approaches, achieving up to 8.1% improvements in F1-score while enabling real-time inference on edge devices such as the *Nvidia Jetson Orin Nano*. These findings highlight that robust generalization requires not only static domain-invariant modeling, but also dynamic, context-aware adaptation strategies.

Finally, to address **RQ IV**, a novel approach for enabling TW visual monitoring without the use of conventional cameras by directly synthesizing RGB images from WiFi CSI, was proposed. Representing the first demonstration of image synthesis from CSI captured in a TW scenario, this work expands the functional scope of WiFi-based sensing toward intuitive visual monitoring applications. The proposed method leverages a multimodal VAE architecture jointly adapted to CSI amplitude spectrograms and RGB images, enabling the reconstruction of semantically meaningful images solely from CSI at test time. By exploiting WiFi's intrinsic wall-penetrating capability, the approach offers a visual

privacy-preserving alternative to optical systems for monitoring indoor environments across room boundaries. Experimental results on the WiFiCam dataset confirmed the technical viability of this approach, while also opening the door to visually-driven downstream tasks such as activity labeling and behavioral analysis. This contribution marks a significant advancement in the application space of WiFi-based PCS, highlighting the potential for intuitive and camera-free visual monitoring using WiFi signals.

Bibliography

[1] Julian Strohmayer, Jennifer Lumetzberger, Thomas Heitzinger, and Martin Kampel. Person-centric sensing in indoor environments. In *Scanning Technologies for Autonomous Systems*, pages 303–341. Springer, 2024.

[2] Damien Bouchabou, Sao Mai Nguyen, Christophe Lohr, Benoit LeDuc, and Ioannis Kanellos. A survey of human activity recognition in smart homes based on iot sensors algorithms: Taxonomies, challenges, and opportunities with deep learning. *Sensors*, 21(18), 2021.

[3] Biying Fu, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Sensing technology for human activity recognition: A comprehensive survey. *IEEE Access*, PP:1–1, 01 2020.

[4] Jingyang Hu, Hongbo Jiang, Tianyue Zheng, Jingzhi Hu, Hongbo Wang, Hangcheng Cao, Zhe Chen, and Jun Luo. M 2-fi: Multi-person respiration monitoring via handheld wifi devices. In *IEEE INFOCOM 2024-IEEE Conference on Computer Communications*, pages 1221–1230. IEEE, 2024.

[5] Shengjie Li, Zhaopeng Liu, Qin Lv, Yanyan Zou, Yue Zhang, and Daqing Zhang. Wilife: Long-term daily status monitoring and habit mining of the elderly leveraging ubiquitous wi-fi signals. *ACM Trans. Comput. Healthcare*, 6(1), January 2025.

[6] Md Touhiduzzaman, Steven M Hernandez, Peter E Pidcoe, and Eyuphan Bulut. Wi-pt-hand: Wireless sensing based low-cost physical rehabilitation tracking for hand movements. *ACM Transactions on Computing for Healthcare*, 6(1):1–25, 2025.

[7] Hongbo Jiang, Chao Cai, Xiaoqiang Ma, Yang Yang, and Jiangchuan Liu. Smart home based on wifi sensing: A survey. *IEEE Access*, 6:13317–13325, 2018.

[8] Shuxin Zhong, Yongzhi Huang, Rukhsana Ruby, Lu Wang, Yu-Xuan Qiu, and Kaishun Wu. Wi-fire: Device-free fire detection using wifi networks. In *2017 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2017.

[9] Junye Li, Aryan Sharma, Deepak Mishra, and Aruna Seneviratne. Fire detection using commodity wifi devices. In *2021 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE, 2021.

[10] Sirui Jian, Shigemi Ishida, and Yutaka Arakawa. Initial attempt on wi-fi csi based vibration sensing for factory equipment fault detection. In *Adjunct proceedings of the 2021 international conference on distributed computing and networking*, pages 163–168, 2021.

[11] Yogesh Dasgaonkar and Radhe Shyam Sharma. Through the rubble: Discovering the invisible. In *2024 18th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pages 1074–1079. IEEE, 2024.

[12] Beddiar Romaissa, Brahim Nini, Mohammad Sabokrou, and Abdenour Hadid. Vision-based human activity recognition: a survey. *Multimedia Tools and Applications*, 79, 11 2020.

[13] Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[15] Shahriar Shakir Sumit, Dayang Rohaya Awang Rambli, and Seyedali Mirjalili. Vision-based human detection techniques: A descriptive review. *IEEE Access*, 9:42724–42761, 2021.

[16] Julian Strohmayer, Jakob Knapp, and Martin Kampel. Efficient models for real-time person segmentation on mobile phones. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 651–655, 2021.

[17] Ce Zheng, Wenhan Wu, Taojiannan Yang, Sijie Zhu, Chen Chen, Ruixu Liu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ArXiv*, abs/2012.13392, 2019.

[18] Ronald Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, 2010.

[19] Julia Offermann van Heek, Wiktoria Wilkowska, and Martina Ziefle. Cultural impact on perceptions of aging, care, and lifelogging technology: A comparison between turkey and germany. *International Journal of Human–Computer Interaction*, 37(2):156–168, 2021.

[20] Katrin Arning and Martina Ziefle. "get that camera out of my house!" conjoint measurement of preferences for video-based healthcare monitoring systems in private and public places. In *Inclusive Smart Cities and e-Health: 13th International Conference on Smart Homes and Health Telematics, ICOST 2015, Geneva, Switzerland, June 10-12, 2015, Proceedings 13*, pages 152–164. Springer, 2015.

[21] M. Alwan, P.J. Rajendran, S. Kell, D. Mack, S. Dalal, M. Wolfe, and R. Felder. A smart and passive floor-vibration based fall detector for elderly. In *2006 2nd International Conference on Information and Communication Technologies*, volume 1, pages 1003–1007, 2006.

[22] Xueyi Wang, Joshua Ellul, and George Azzopardi. Elderly fall detection systems: A literature survey. *Frontiers in Robotics and AI*, 7, 2020.

[23] Chao Bian, Bing Ye, Anna Hoonakker, and Alex Mihailidis. Attitudes and perspectives of older adults on technologies for assessing frailty in home settings: a focus group study. *BMC Geriatrics*, 21, 05 2021.

[24] Rui Min, Jongmoo Choi, Gérard Medioni, and Jean-Luc Dugelay. Real-time 3d face identification from a depth camera. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 1739–1742, 2012.

[25] Carlos A. Luna, Cristina Losada-Gutiérrez, David Fuentes-Jimenez, and Manuel Mazo. People re-identification using depth and intensity information from an overhead camera. *Expert Systems with Applications*, 182:115287, 2021.

[26] Vincent Weidlich. Thermal infrared face recognition. *Cureus*, 13, 03 2021.

[27] Julia Offermann, Wiktoria Wilkowska, and Martina Ziefle. Cultural impact on perceptions of aging, care, and lifelogging technology: A comparison between turkey and germany. *International Journal of Human-Computer Interaction*, 09 2020.

[28] Scott Beach, Richard Schulz, Julie Downs, Judith Matthews, Bruce Barron, and Katherine Seelman. Disability, age, and informational privacy attitudes in quality of life technology applications: Results from a national web survey. *ACM Trans. Access. Comput.*, 2(1), may 2009.

[29] Ha Manh Do, Karla Conn Welch, and Weihua Sheng. Soham: A sound-based human activity monitoring framework for home service robots. *IEEE Transactions on Automation Science and Engineering*, 19(3):2369–2383, 2022.

[30] Arindam Ghosh, Amartya Chakraborty, Dhruv Chakraborty, Mousumi Saha, and Sujoy Saha. Ultrasense: A non-intrusive approach for human activity identification using heterogeneous ultrasonic sensor grid for smart home environment. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–22, 2023.

[31] Maria Valero, Fangyu Li, Liang Zhao, Chi Zhang, José M. Garrido, and Zhu Han. Vibration sensing-based human and infrastructure safety/health monitoring: A survey. *Digit. Signal Process.*, 114:103037, 2021.

[32] Gierad Laput, Yang Zhang, and Chris Harrison. Synthetic sensors: Towards general-purpose sensing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 3986–3999, New York, NY, USA, 2017.

[33] Reza Shahbazian and Irina Trubitsyna. Human sensing by using radio frequency signals: A survey on occupancy and activity detection. *IEEE Access*, 11:40878–40904, 2023.

[34] Xinyu Li, Yuan He, and Xiaojun Jing. A survey of deep learning-based human activity recognition in radar. *Remote. Sens.*, 11:1068, 2019.

[35] Guozhen Zhu, Yuqian Hu, Chenshu Wu, Wei-Hsiang Wang, Beibei Wang, and KJ Liu. Experience paper: Scaling wifi sensing to millions of commodity devices for ubiquitous home monitoring. *arXiv preprint arXiv:2506.04322*, 2025.

[36] Jian Liu, Hongbo Liu, Yingying Chen, Yan Wang, and Chen Wang. Wireless sensing for human activity: A survey. *IEEE Communications Surveys & Tutorials*, 22(3):1629–1645, 2020.

[37] Wei Wang, Alex X. Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. Device-free human activity recognition using commercial wifi devices. *IEEE Journal on Selected Areas in Communications*, 35(5):1118–1131, 2017.

[38] Xu Feng, Khuong An Nguyen, and Zhiyuan Luo. A review of open access wifi fingerprinting datasets for indoor positioning. *IEEE Access*, 12:167970–167989, 2024.

[39] Kamran Ali, Mohammed Alloulah, Fahim Kawsar, and Alex X. Liu. On goodness of wifi based monitoring of sleep vital signs in the wild. *IEEE Transactions on Mobile Computing*, 22(1):341–355, 2023.

[40] Julian Strohmayer and Martin Kampel. Wifi csi-based long-range through-wall human activity recognition with the esp32. In *International Conference on Computer Vision Systems*, pages 41–50. Springer, 2023.

[41] Sihao Li, Zhe Tang, Kyeong Soo Kim, and Jeremy S Smith. On the use and construction of wi-fi fingerprint databases for large-scale multi-building and multi-floor indoor localization: A case study of the ujiindoorloc database. *Sensors*, 24(12):3827, 2024.

[42] Robert Schumann, Frédéric Li, and Marcin Grzegorzek. Wifi sensing with single-antenna devices for ambient assisted living. In *Proceedings of the 8th international Workshop on Sensor-Based Activity Recognition and Artificial Intelligence*, pages 1–8, 2023.

[43] Jianfei Yang, Xinyan Chen, Han Zou, Chris Xiaoxuan Lu, Dazhuo Wang, Sumei Sun, and Lihua Xie. Sensefi: A library and benchmark on deep-learning-empowered wifi human sensing. *Patterns*, 4(3), 2023.

[44] Shuokang Huang, Kaihan Li, Di You, Yichong Chen, Arvin Lin, Siying Liu, Xiaohui Li, and Julie A McCann. Wimans: A benchmark dataset for wifi-based multi-user activity sensing. In *European Conference on Computer Vision*, pages 72–91. Springer, 2024.

[45] Steven M. Hernandez and Eyuphan Bulut. Wifi sensing on the edge: Signal processing techniques and challenges for real-world systems. *IEEE Communications Surveys & Tutorials*, 25(1):46–76, 2023.

[46] Zheng Yang, Yi Zhang, Kun Qian, and Chenshu Wu. {SLNet}: A spectrogram learning neural network for deep wireless sensing. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, pages 1221–1236, 2023.

[47] Chen Chen, Gang Zhou, and Youfang Lin. Cross-domain wifi sensing with channel state information: A survey. *ACM Computing Surveys*, 55(11):1–37, 2023.

[48] Julian Strohmayer and Martin Kampel. Data augmentation techniques for cross-domain wifi csi-based human activity recognition. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 42–56. Springer, 2024.

[49] Julian Strohmayer and Martin Kampel. Wifi CSI-based long-range person localization using directional antennas. In *The Second Tiny Papers Track at ICLR 2024*, 2024.

[50] Julian Strohmayer and Martin Kampel. On the generalization of wifi-based person-centric sensing in through-wall scenarios. In *Pattern Recognition*, pages 194–211, Cham, 2025. Springer Nature Switzerland.

[51] Julian Strohmayer, Rafael Sterzinger, Christian Stippel, and Martin Kampel. Through-wall imaging based on wifi channel state information. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 4000–4006, 2024.

[52] Julian Strohmayer and Martin Kampel. Directional antenna systems for long-range through-wall human activity recognition. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 3594–3599, 2024.

[53] Julian Strohmayer, Matthias Wödlinger, and Martin Kampel. Wiflexformer: Efficient wifi-based person-centric sensing. *arXiv preprint arXiv:2411.04224*, 2024.

[54] Julian Strohmayer, Rafael Sterzinger, Matthias Wödlinger, and Martin Kampel. Datta: Domain-adversarial test-time adaptation for cross-domain wifi-based human activity recognition. *arXiv preprint arXiv:2411.13284*, 2024.

[55] Shuya Ding, Zhe Chen, Tianyue Zheng, and Jun Luo. Rf-net: A unified meta-learning framework for rf-enabled one-shot human activity recognition. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, pages 517–530, 2020.

[56] 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. *IEEE Std 802.11-2024 (Revision of IEEE Std 802.11-2020)*, pages 1–5956, 2025.

[57] Ieee standard for telecommunications and information exchange between systems - lan/man specific requirements - part 11: Wireless medium access control (mac) and physical layer (phy) specifications: High speed physical layer in the 5 ghz band. *IEEE Std 802.11a-1999*, pages 1–102, 1999.

[58] Ieee standard for information technology– local and metropolitan area networks–specific requirements– part 11: Wireless lan medium access control (mac)and physical layer (phy) specifications amendment 5: Enhancements for higher throughput. *IEEE Std 802.11n-2009 (Amendment to IEEE Std 802.11-2007 as amended by IEEE Std 802.11k-2008, IEEE Std 802.11r-2008, IEEE Std 802.11y-2008, and IEEE Std 802.11w-2009)*, pages 1–565, 2009.

[59] 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. *IEEE Std 802.11ax-2021 (Amendment to IEEE Std 802.11-2020)*, pages 1–767, 2021.

[60] Yongsen Ma, Gang Zhou, and Shuangquan Wang. Wifi sensing with channel state information: A survey. *ACM Computing Surveys (CSUR)*, 52(3):1–36, 2019.

[61] Angelos Vlavianos, Lap Kong Law, Ioannis Broustis, Srikanth V. Krishnamurthy, and Michalis Faloutsos. Assessing link quality in ieee 802.11 wireless networks: Which is the right metric? In *2008 IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications*, pages 1–6, 2008.

[62] Zheng Yang, Zimu Zhou, and Yunhao Liu. From rssi to csi: Indoor localization via channel response. *ACM Comput. Surv.*, 46(2), dec 2013.

[63] Ambili Thottam Parameswaran, Mohammad Iftekhar Husain, Shambhu Upadhyaya, et al. Is rssi a reliable parameter in sensor localization algorithms: An experimental study. In *Field failure data analysis workshop (F2DA09)*, volume 5. IEEE Niagara Falls, NY, USA, 2009.

[64] Ahmed Abdel Ghany, Bernard Uguen, and Dominique Lemur. A robustness comparison of measured narrowband csi vs rssi for iot localization. In *2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall)*, pages 1–5, 2020.

[65] Siamak Yousefi, Hirokazu Narui, Sankalp Dayal, Stefano Ermon, and Shahrokh Valaee. A survey on behavior recognition using wifi channel state information. *IEEE Communications Magazine*, 55(10):98–104, 2017.

[66] Ramjee Prasad. *OFDM for wireless communications systems*. Artech House, 2004.

[67] IEEE. 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. *IEEE Std 802.11-2016 (Revision of IEEE Std 802.11-2012)*, pages 1–3534, 2016.

[68] Daniel Halperin, Wenjun Hu, Anmol Sheth, and David Wetherall. 802.11 with multiple antennas for dummies. *ACM SIGCOMM Computer Communication Review*, 40(1):19–25, 2010.

[69] Josiah W. Smith. Complex-valued neural networks for data-driven signal processing and signal understanding, 2023.

[70] Akira Hirose and Shotaro Yoshida. Generalization characteristics of complex-valued feedforward neural networks in relation to signal coherence. *IEEE Transactions on Neural Networks and learning systems*, 23(4):541–551, 2012.

[71] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, et al. Towards environment independent device free human activity recognition. In *Proceedings of the 24th annual international conference on mobile computing and networking*, pages 289–304, 2018.

[72] Enjie Ding, Xiansheng Li, Tong Zhao, Lei Zhang, Yanjun Hu, et al. A robust passive intrusion detection system with commodity wifi devices. *Journal of Sensors*, 2018, 2018.

[73] Xingang Wang, Yufei Wang, and Dong Wang. A real-time csi-based passive intrusion detection method. In *2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, pages 1091–1098, 2020.

[74] Yi Zhang, Yue Zheng, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. Widar3.0: Zero-effort cross-domain gesture recognition with wi-fi. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8671–8688, 2022.

[75] Weiying Hou and Chenshu Wu. Rfboost: Understanding and boosting deep wifi sensing via physical data augmentation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8:1 – 26, 2024.

[76] Heju Li, Xukai Chen, Haohua Du, Xin He, Jianwei Qian, Peng-Jun Wan, and Panlong Yang. Wi-motion: A robust human activity recognition using wifi signals, 2018.

[77] Sheng Tan, Yili Ren, Jie Yang, and Yingying Chen. Commodity wifi sensing in ten years: Status, challenges, and opportunities. *IEEE Internet of Things Journal*, 9(18):17832–17843, 2022.

[78] Paramvir Bahl and Venkata N Padmanabhan. Radar: An in-building rf-based user location and tracking system. In *Proceedings IEEE INFOCOM 2000. Conference on computer communications. Nineteenth annual joint conference of the IEEE computer and communications societies (Cat. No. 00CH37064)*, volume 2, pages 775–784. Ieee, 2000.

[79] Daniel Halperin, Wenjun Hu, Anmol Sheth, and David Wetherall. Tool release: Gathering 802.11n traces with channel state information. *ACM SIGCOMM CCR*, 41(1):53, Jan. 2011.

[80] Ahmed E. Kosba, Ahmed Saeed, and Moustafa Youssef. Rasid: A robust wlan device-free passive motion detection system. In *2012 IEEE International Conference on Pervasive Computing and Communications*, pages 180–189, 2012.

[81] Kaishun Wu, Jiang Xiao, Youwen Yi, Min Gao, and Lionel M. Ni. Fila: Fine-grained indoor localization. In *2012 Proceedings IEEE INFOCOM*, pages 2210–2218, 2012.

[82] Jiang Xiao, Kaishun Wu, Youwen Yi, Lu Wang, and Lionel M Ni. Fimd: Fine-grained device-free motion detection. In *2012 IEEE 18th International conference on parallel and distributed systems*, pages 229–235. IEEE, 2012.

[83] Stephan Sigg, Shuyu Shi, Felix Büsching, Yusheng Ji, and Lars Wolf. Leveraging rf-channel fluctuation for activity recognition: Active and passive systems, continuous and rssi-based signal features. *ACM International Conference Proceeding Series*, 12 2013.

[84] Yan Wang, Jian Liu, Yingying Chen, Marco Gruteser, Jie Yang, and Hongbo Liu. E-eyes: Device-free location-oriented activity identification using fine-grained wifi signatures. *Proceedings of the Annual International Conference on Mobile Computing and Networking, MOBICOM*, 09 2014.

[85] Moustafa Youssef and Ashok Agrawala. The horus wlan location determination system. In *Proceedings of the 3rd International Conference on Mobile Systems, Applications, and Services*, MobiSys '05, page 205–218, New York, NY, USA, 2005. Association for Computing Machinery.

[86] Moustafa Youssef, Matthew Mah, and Ashok Agrawala. Challenges: device-free passive localization for wireless environments. In *Proceedings of the 13th Annual ACM International Conference on Mobile Computing and Networking*, MobiCom '07, page 222–229, New York, NY, USA, 2007. Association for Computing Machinery.

[87] Hai Zhu, Enlai Dong, Mengmeng Xu, Hongxiang Lv, and Fei Wu. Commodity wi-fi-based wireless sensing advancements over the past five years. *Sensors*, 24(22):7195, 2024.

[88] Souvik Sen, Božidar Radunovic, Romit Roy Choudhury, and Tom Minka. You are facing the mona lisa: spot localization using phy layer information. In *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, MobiSys '12, page 183–196, New York, NY, USA, 2012. Association for Computing Machinery.

[89] Wei Wang, Alex X. Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. Understanding and modeling of wifi signal based human activity recognition. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, MobiCom '15, page 65–76, New York, NY, USA, 2015.

[90] Iftikhar Ahmad, Arif Ullah, and Wooyeol Choi. Wifi-based human sensing with deep learning: Recent advances, challenges, and opportunities. *IEEE Open Journal of the Communications Society*, 5:3595–3623, 2024.

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

[92] Jie Zhang, Zhanyong Tang, Meng Li, Dingyi Fang, Petteri Nurmi, and Zheng Wang. Crosssense: Towards cross-site and large-scale wifi sensing. In *Proceedings of the 24th annual international conference on mobile computing and networking*, pages 305–320, 2018.

[93] Zhenghua Chen, Le Zhang, Chaoyang Jiang, Zhiguang Cao, and Wei Cui. Wifi csi based passive human activity recognition using attention based blstm. *IEEE Transactions on Mobile Computing*, 18(11):2714–2724, 2019.

[94] Fei Wang, Sanping Zhou, Stanislav Panev, Jinsong Han, and Dong Huang. Person-in-wifi: Fine-grained person perception using wifi. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5452–5461, 2019.

[95] Wenjun Jiang, Hongfei Xue, Chenglin Miao, Shiyang Wang, Sen Lin, Chong Tian, Srinivasan Murali, Haochen Hu, Zhi Sun, and Lu Su. Towards 3d human pose construction using wifi. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, pages 1–14, 2020.

[96] Mingze Yang, Hai Zhu, Runzhe Zhu, Fei Wu, Ling Yin, and Yuncheng Yang. Witransformer: A novel robust gesture recognition sensing model with wifi. *Sensors*, 23(5):2612, 2023.

[97] Fei Luo, Salabat Khan, Bin Jiang, and Kaishun Wu. Vision transformers for human activity recognition using wifi channel state information. *IEEE Internet of Things Journal*, 11(17):28111–28122, 2024.

[98] Kangwei Yan, Fei Wang, Bo Qian, Han Ding, Jinsong Han, and Xing Wei. Person-in-wifi 3d: End-to-end multi-person 3d pose estimation with wi-fi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 969–978, 2024.

[99] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

[100] Jürgen Schmidhuber. Learning complex, extended sequences using the principle of history compression. *Neural computation*, 4(2):234–242, 1992.

[101] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[102] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[103] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[104] Bing Li, Wei Cui, Wei Wang, Le Zhang, Zhenghua Chen, and Min Wu. Two-stream convolution augmented transformer for human activity recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 286–293, 2021.

[105] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[106] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[107] Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. Zero-effort cross-domain gesture recognition with wi-fi. In *Proceedings of the 17th annual international conference on mobile systems, applications, and services*, pages 313–325, 2019.

[108] Yunhao Bai, Zejiang Wang, Kuangyu Zheng, Xiaorui Wang, and Junmin Wang. Widrive: Adaptive wifi-based recognition of driver activity for real-time and safe takeover. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pages 901–911. IEEE, 2019.

[109] Fangxin Wang, Wei Gong, and Jiangchuan Liu. On spatial diversity in wifi-based human activity recognition: A deep learning-based approach. *IEEE Internet of Things Journal*, 6(2):2035–2047, 2019.

[110] Youwei Zeng, Dan Wu, Jie Xiong, Jinyi Liu, Zhaopeng Liu, and Daqing Zhang. Multisense: Enabling multi-person respiration sensing with commodity wifi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(3):1–29, 2020.

[111] Biyun Sheng, Fu Xiao, Letian Sha, and Lijuan Sun. Deep spatial–temporal model based cross-scene action recognition using commodity wifi. *IEEE Internet of Things Journal*, 7(4):3592–3601, 2020.

[112] Pengli Hu, Chengpei Tang, Kang Yin, and Xie Zhang. Wigr: a practical wi-fi-based gesture recognition system with a lightweight few-shot network. *Applied Sciences*, 11(8):3329, 2021.

[113] Hua Kang, Qian Zhang, and Qianyi Huang. Context-aware wireless-based cross-domain gesture recognition. *IEEE Internet of Things Journal*, 8(17):13503–13515, 2021.

[114] Dan Wu, Youwei Zeng, Ruiyang Gao, Shenjie Li, Yang Li, Rahul C Shah, Hong Lu, and Daqing Zhang. Witraj: Robust indoor motion tracking with wifi signals. *IEEE Transactions on Mobile Computing*, 22(5):3062–3078, 2021.

[115] Wei Wang, Alex X Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. Device-free human activity recognition using commercial wifi devices. *IEEE Journal on Selected Areas in Communications*, 35(5):1118–1131, 2017.

[116] Zhenguo Shi, Qingqing Cheng, J Andrew Zhang, and Richard Yi Da Xu. Environment-robust wifi-based human activity recognition using enhanced csi and deep learning. *IEEE Internet of Things Journal*, 9(24):24643–24654, 2022.

[117] Maximilian Stahlke, George Yammine, Tobias Feigl, Bjoern M. Eskofier, and Christopher Mutschler. Indoor localization with robust global channel charting: A time-distance-based approach. *IEEE Transactions on Machine Learning in Communications and Networking*, 1:3–17, 2023.

[118] Yan Liu, Anlan Yu, Leye Wang, Bin Guo, Yang Li, Enze Yi, and Daqing Zhang. Unifi: A unified framework for generalizable gesture recognition with wi-fi signals using consistency-guided multi-view networks. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 7(4), jan 2024.

[119] Dazhuo Wang, Jianfei Yang, Wei Cui, Lihua Xie, and Sumei Sun. Airfi: Empowering wifi-based passive human gesture recognition to unseen environment via domain generalization. *IEEE Transactions on Mobile Computing*, 23(2):1156–1168, 2024.

[120] Kaixuan Gao, Huiqiang Wang, Hongwu Lv, and Wenxue Liu. Toward 5g nr high-precision indoor positioning via channel frequency response: A new paradigm and dataset generation method. *IEEE Journal on Selected Areas in Communications*, 40(7):2233–2247, 2022.

[121] Omer Gokalp Serbetci, Ju-Hyung Lee, Daoud Burghal, and Andreas F. Molisch. Simple and effective augmentation methods for csi based indoor localization, 2023.

[122] Xi Chen, Hang Li, Chenyi Zhou, Xue Liu, Di Wu, and Gregory Dudek. Fido: Ubiquitous fine-grained wifi-based localization for unlabelled users via domain adaptation. In *Proceedings of The Web Conference 2020*, WWW '20, page 23–33, New York, NY, USA, 2020. Association for Computing Machinery.

[123] Xi Chen, Hang Li, Chenyi Zhou, Xue Liu, Di Wu, and Gregory Dudek. Fidora: Robust wifi-based indoor localization via unsupervised domain adaptation. *IEEE Internet of Things Journal*, 9(12):9872–9888, 2022.

[124] Xinyi Li, Liqiong Chang, Fangfang Song, Ju Wang, Xiaojiang Chen, Zhanyong Tang, and Zheng Wang. Crossgr: Accurate and low-cost cross-target gesture recognition using wi-fi. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5(1), mar 2021.

[125] Hoonyong Lee, Changbum R. Ahn, and Nakjung Choi. Toward single occupant activity recognition for long-term periods via channel state information. *IEEE Internet of Things Journal*, pages 1–1, 2023.

[126] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2018.

[127] Zhanjun Hao, Juan Niu, Xiaochao Dang, and Danyang Feng. Wi-cal: A cross-scene human motion recognition method based on domain adaptation in a wi-fi environment. *Electronics*, 11(16):2607, 2022.

[128] Jie Zhang, Yang Li, Haoyi Xiong, Dejing Dou, Chunyan Miao, and Daqing Zhang. Handgest: Hierarchical sensing for robust-in-the-air handwriting recognition with commodity wifi devices. *IEEE Internet of Things Journal*, 9(19):19529–19544, 2022.

[129] Zijian Zhao, Zhijie Cai, Tingwei Chen, Xiaoyang Li, Hang Li, and Guangxu Zhu. Knn-mmd: Cross domain wi-fi sensing based on local distribution alignment. *arXiv preprint arXiv:2412.04783*, 2024.

[130] Zijian Zhao, Tingwei Chen, Zhijie Cai, Xiaoyang Li, Hang Li, Qimei Chen, and Guangxu Zhu. Crossfi: A cross domain wi-fi sensing framework based on siamese network. *arXiv preprint arXiv:2408.10919*, 2024.

[131] Naiyu Zheng, Yuanchun Li, Shiqi Jiang, Yuanzhe Li, Rongchun Yao, Chuchu Dong, Ting Chen, Yubo Yang, Zhimeng Yin, and Yunxin Liu. Adawifi, collaborative wifi sensing for cross-environment adaptation. *IEEE Transactions on Mobile Computing*, 24(2):845–858, 2025.

[132] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[133] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

[134] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

[135] Xue Ding, Ting Jiang, Yi Zhong, Yan Huang, and Zhiwei Li. Wi-fi-based location-independent human activity recognition via meta learning. *Sensors*, 21(8):2654, 2021.

[136] Jianfei Yang, Han Zou, Yuxun Zhou, and Lihua Xie. Learning gestures from wifi: A siamese recurrent convolutional architecture. *IEEE Internet of Things Journal*, 6(6):10763–10772, 2019.

[137] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[138] Xie Zhang, Chengpei Tang, Kang Yin, and Qingqian Ni. Wifi-based cross-domain gesture recognition via modified prototypical networks. *IEEE Internet of Things Journal*, 9(11):8584–8596, 2021.

[139] Chenning Li, Zheng Liu, Yuguang Yao, Zhichao Cao, Mi Zhang, and Yunhao Liu. Wi-fi see it all: generative adversarial network-augmented versatile wi-fi imaging. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, pages 436–448, 2020.

[140] Mohammad Hadi Kefayati, Vahid Pourahmadi, and Hassan Aghaeinia. Wi2vi: Generating video frames from wifi csi samples. *IEEE Sensors Journal*, 20(19):11463–11473, 2020.

[141] Sorachi Kato, Takeru Fukushima, Tomoki Murakami, Hirantha Abeysekera, Yusuke Iwasaki, Takuya Fujihashi, Takashi Watanabe, and Shunsuke Saruwatari. Csi2image: Image reconstruction from channel state information using generative adversarial networks. *IEEE Access*, 9:47154–47168, 2021.

[142] Jiaqi Geng, Dong Huang, and Fernando De la Torre. Densepose from wifi, 2022.

[143] Jianfei Yang, Yunjiao Zhou, He Huang, Han Zou, and Lihua Xie. Metafi: Device-free pose estimation via commodity wifi for metaverse avatar simulation, 2022.

[144] Yunjiao Zhou, He Huang, Shenghai Yuan, Han Zou, Lihua Xie, and Jianfei Yang. Metafi++: Wifi-enabled transformer-based human pose estimation for metaverse avatar simulation. *IEEE Internet of Things Journal*, 10(16):14128–14136, 2023.

[145] Zhiguo Chen, Jiaren Xiao, and Bing Luo. Human segmentation using commercial wi-fi device. In *2023 3rd International Conference on Neural Networks, Information and Communication Engineering (NNICE)*, pages 299–302, 2023.

[146] Yichao Wang, Yili Ren, and Jie Yang. Multi-subject 3d human mesh construction using commodity wifi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–25, 2024.

[147] Guanyu Cao, Takuya Maekawa, Kazuya Ohara, and Yasue Kishino. Reconstructing depth images of moving objects from wi-fi csi data, 2025.

[148] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023.

[149] Ariel Luzzatto and Motti Haridim. *Wireless transceiver design: mastering the design of modern wireless equipment and systems*. John Wiley & Sons, 2016.

[150] Zhi Ning Chen, Xianming Qing, Terence Shie Ping See, and Wee Kian Toh. Antennas for wifi connectivity. *Proceedings of the IEEE*, 100(7):2322–2329, 2012.

[151] Yaxiong Xie, Zhenjiang Li, and Mo Li. Precise power delay profiling with commodity wifi. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, MobiCom '15, page 53–64, New York, NY, USA, 2015. ACM.

[152] Zhiping Jiang, Tom H. Luan, Xincheng Ren, Dongtao Lv, Han Hao, Jing Wang, Kun Zhao, Wei Xi, Yueshen Xu, and Rui Li. Eliminating the barriers: Demystifying wi-fi baseband design and introducing the picoscenes wi-fi sensing platform, 2021.

[153] Francesco Gringoli, Matthias Schulz, Jakob Link, and Matthias Hollick. Free your csi: A channel state information extraction platform for modern wi-fi chipsets. In *Proceedings of the 13th International Workshop on Wireless Network Testbeds, Experimental Evaluation & Characterization*, WiNTECH '19, page 21–28, 2019.

[154] Francesco Gringoli, Marco Cominelli, Alejandro Blanco, and Joerg Widmer. Ax-csi: Enabling csi extraction on commercial 802.11 ax wi-fi platforms. In *Proceedings of the 15th ACM Workshop on Wireless Network Testbeds, Experimental Evaluation & CHaracterization*, pages 46–53, 2022.

[155] Steven M. Hernandez and Eyuphan Bulut. Lightweight and Standalone IoT Based WiFi Sensing for Active Repositioning and Mobility. In *21st International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM) (WoWMoM 2020)*, Cork, Ireland, June 2020.

[156] Muhammad Atif, Shapna Muralidharan, Heedong Ko, and Byounghyun Yoo. Wi-ESP—A tool for CSI-based Device-Free Wi-Fi Sensing (DFWS). *Journal of Computational Design and Engineering*, 7(5):644–656, 05 2020.

[157] Anisha Natarajan, Vijayakumar Krishnasamy, and Munesh Singh. Design of a low-cost and device-free human activity recognition model for smart led lighting control. *IEEE Internet of Things Journal*, 11(4):5558–5567, 2024.

[158] Steven M Hernandez. Wifi sensing at the edge towards scalable on-device wireless sensing systems. 2023.

[159] Ajit Kumar Sahoo, Vaishnavi Kompally, and Siba K Udgata. Wi-fi sensing based real-time activity detection in smart home environment. In *2023 IEEE Applied Sensing Conference (APSCON)*, pages 1–3, 2023.

[160] Steven M. Hernandez and Eyuphan Bulut. Adversarial occupancy monitoring using one-sided through-wall wifi sensing. In *ICC 2021 - IEEE International Conference on Communications*, pages 1–6, 2021.

[161] Zhanjun Hao, Guowei Wang, and Xiaochao Dang. Car-sense: vehicle occupant legacy hazard detection method based on dfws. *Applied Sciences*, 12(22):11809, 2022.

[162] Rikesh Makwana and Talal Shaikh. Touchless biometric user authentication using esp32 wifi module. In *Proceedings of International Conference on Information Technology and Applications: ICITA 2021*, pages 527–537. Springer, 2022.

[163] Sahoo Ajit Kumar, K Akhil, and Siba K Udgata. Wi-fi signal-based through-wall sensing for human presence and fall detection using esp32 module. In *Intelligent Systems: Proceedings of ICMIB 2021*, pages 459–470. Springer, 2022.

[164] Mohammad Zeeshan, Ankur Pandey, and Sudhir Kumar. Csi-based device-free joint activity recognition and localization using siamese networks. In *2022 14th International Conference on COMmunication Systems & NETworkS (COMSNETS)*, pages 260–264, 2022.

[165] Hyuckjin Choi, Manato Fujimoto, Tomokazu Matsui, Shinya Misaki, and Keiichi Yasumoto. Wi-cal: Wifi sensing and machine learning based device-free crowd counting and localization. *IEEE Access*, 10:24395–24410, 2022.

[166] B Singh and A Singh. A novel biquad antenna for 2.4 ghz wireless link application: a proposed design. *International Journal of Electronics & Communication Technology*, 3(1):174–176, 2012.

[167] Jincao Zhu, Youngbin Im, Shivakant Mishra, and Sangtae Ha. Calibrating time-variant, device-specific phase noise for cots wifi devices. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*, SenSys '17, New York, NY, USA, 2017. Association for Computing Machinery.

[168] Nan Yu, Wei Wang, Alex X Liu, and Lingtao Kong. Qgesture: Quantifying gesture distance and direction with wifi signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):1–23, 2018.

[169] Lei Zhang, Yixiang Zhang, and Xiaolong Zheng. Wisign: Ubiquitous american sign language recognition using commercial wi-fi devices. *ACM Trans. Intell. Syst. Technol.*, 11(3), apr 2020.

[170] Losses Through Common. Propagation losses through common building materials 2.4 ghz vs 5 ghz. *E10589, Magis Network, Inc*, 2002.

[171] Dipankar Shakya, Mingjun Ying, Theodore S. Rappaport, Hitesh Poddar, Peijie Ma, Yanbo Wang, and Idris Al-Wazani. Wideband penetration loss through building materials and partitions at 6.75 ghz in fr1(c) and 16.95 ghz in the fr3 upper mid-band spectrum, 2024.

[172] Swetank Kumar Saha, Viral Vijay Vira, Anuj Garg, and Dimitrios Koutsonikolas. A feasibility study of 60 ghz indoor wlans. In *2016 25th international conference on computer communication and networks (ICCCN)*, pages 1–9. IEEE, 2016.

[173] Moinak Ghoshal, Shravan Bhoopasamudram Krishna, Francesco Gringoli, Joerg Widmer, and Dimitrios Koutsonikolas. A first look at 160 mhz wifi 6/6e in action: Performance and interference characterization. In *2024 IFIP Networking Conference (IFIP Networking)*, pages 489–495. IEEE, 2024.

[174] Haobin Guan, Aryan Sharma, Deepak Mishra, and Aruna Seneviratne. Experimental accuracy comparison for 2.4 ghz and 5ghz wifi sensing systems. In *ICC 2023-IEEE International Conference on Communications*, pages 4755–4760. IEEE, 2023.

[175] Jun Wan, Ting Jiang, Xinyi Zhou, and Danlan Huang. Wi-locind: Location-independent respiration sensing based on wifi csi. In *2024 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6, 2024.

[176] Siamak Yousefi, Hirokazu Narui, Sankalp Dayal, Stefano Ermon, and Shahrokh Valaee. A survey on behavior recognition using wifi channel state information. *IEEE Communications Magazine*, 55(10):98–104, 2017.

[177] Yongsen Ma, Gang Zhou, Shuangquan Wang, Hongyang Zhao, and Woosub Jung. Signfi: Sign language recognition using wifi. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(1), mar 2018.

[178] Linlin Guo, Lei Wang, Chuang Lin, Jialin Liu, Bingxian Lu, Jian Fang, Zhonghao Liu, Zeyang Shan, Jingwen Yang, and Silu Guo. Wiar: A public dataset for wifi-based activity recognition. *IEEE Access*, 7:154935–154945, 2019.

[179] Alsaify Baha'A, Mahmoud M Almazari, Rami Alazrai, and Mohammad I Daoud. A dataset for wi-fi-based human activity recognition in line-of-sight and non-line-of-sight indoor environments. *Data in Brief*, 33:106534, 2020.

[180] Jörg Schäfer, Baldev Raj Barrsiwal, Muyassar Kokhkharova, Hannan Adil, and Jens Liebehenschel. Human activity recognition using csi information with nexmon. *Applied Sciences*, 11(19):8860, 2021.

[181] Mohammud J Bocus, Wenda Li, Shelly Vishwakarma, Roget Kou, Chong Tang, Karl Woodbridge, Ian Craddock, Ryan McConville, Raul Santos-Rodriguez, Kevin Chetty, et al. Operanet, a multimodal activity recognition dataset acquired from radio frequency and vision-based sensors. *Scientific data*, 9(1):474, 2022.

[182] Francesca Meneghello, Nicolò Dal Fabbro, Domenico Garlisi, Ilenia Tinnirello, and Michele Rossi. A csi dataset for wireless human sensing on 80 mhz wi-fi channels. *IEEE Communications Magazine*, 61(9):146–152, 2023.

[183] Jianfei Yang, He Huang, Yunjiao Zhou, Xinyan Chen, Yuecong Xu, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing, 2023.

[184] Fei Wang, Tingting Zhang, Bintong Zhao, Libao Xing, Tiantian Wang, Han Ding, and Tony Xiao Han. A survey on wi-fi sensing generalizability: Taxonomy, techniques, datasets, and future research prospects. *arXiv preprint arXiv:2503.08008*, 2025.

[185] Zhe-Yu Lim, Lee-Yeng Ong, and Meng-Chew Leow. Radio frequency-based human activity dataset collected using esp32 microcontroller in line-of-sight and non-line-of-sight indoor experiment setups. *Data in Brief*, 57:111101, 2024.

[186] Hari Prabhat Gupta, Salla Jahnavi, Mansi Bhavikbhai, and Rahul Mishra. Channel state information dataset for multi-human activity recognition in indoor environments, 2024.

[187] Thuan VA Tong, Binh Bui-Thanh, and Phuoc Nguyen TH. Human activity recognition using wireless signals and low-cost embedded devices. In *2024 Tenth International Conference on Communications and Electronics (ICCE)*, pages 7–12. IEEE, 2024.

[188] Ronald Pearson, Y. Neuvo, J. Astola, and Moncef Gabbouj. Generalized hampel filters. *EURASIP Journal on Advances in Signal Processing*, 2016, 08 2016.

[189] Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE transactions on robotics*, 37(6):1874–1890, 2021.

[190] Mingxing Tan and Quoc V. Le. Efficientnetv2: Smaller models and faster training, 2021.

[191] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[192] Qiong Wu, Xu Chen, Zhi Zhou, and Junshan Zhang. Fedhome: Cloud-edge based personalized federated learning for in-home health monitoring. *IEEE Transactions on Mobile Computing*, 21(8):2818–2832, 2022.

[193] Daniel Halperin, Wenjun Hu, Anmol Sheth, and David Wetherall. Tool release: Gathering 802.11n traces with channel state information. *Computer Communication Review*, 41:53, 01 2011.

[194] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018.

[195] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[196] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020.

[197] Nikita Dvornik, Cordelia Schmid, and Julien Mairal. Diversity with cooperation: Ensemble methods for few-shot classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3723–3731, 2019.

[198] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[199] Hualei Zhang, Zhu Wang, Zhuo Sun, Wenchao Song, Zhihui Ren, Zhiwen Yu, and Bin Guo. Understanding the mechanism of through-wall wireless sensing: A model-based perspective. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 6(4), jan 2023.

[200] Ian T Jolliffe. *Principal component analysis for special types of data*. Springer, 2002.

[201] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

[202] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. arxiv 2018. *arXiv preprint arXiv:1802.03426*, 1802.

[203] Zhendong Xu, Baoqi Huang, Bing Jia, Wuyungerile Li, and Hui Lu. A boundary aware wifi localization scheme based on umap and knn. *IEEE Communications Letters*, 26(8):1789–1793, 2022.

[204] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[205] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019.

[206] Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei A. Efros. Test-time training with masked autoencoders, 2022.

[207] Renhao Wang, Yu Sun, Yossi Gandelsman, Xinlei Chen, Alexei A. Efros, and Xiaolong Wang. Test-time training on video streams, 2023.

[208] Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. Learning to (Learn at Test Time): RNNs with Expressive Hidden States, August 2024. arXiv:2407.04620.

[209] Wei Lin, Muhammad Jehanzeb Mirza, Mateusz Kozinski, Horst Possegger, Hilde Kuehne, and Horst Bischof. Video Test-Time Adaptation for Action Recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22952–22961, Vancouver, BC, Canada, June 2023. IEEE.

[210] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual Test-Time Domain Adaptation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7191–7201, New Orleans, LA, USA, June 2022. IEEE.

[211] Jing Ma. Improved self-training for test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23701–23710, 2024.

[212] Sayna Ebrahimi, Franziska Meier, Roberto Calandra, Trevor Darrell, and Marcus Rohrbach. Adversarial Continual Learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, volume 12356, pages 386–402. Springer International Publishing, Cham, 2020. Series Title: Lecture Notes in Computer Science.

[213] Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei Efros. Test-Time Training with Masked Autoencoders. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 29374–29385. Curran Associates, Inc., 2022.

[214] M. Jehanzeb Mirza, Jakub Micorek, Horst Possegger, and Horst Bischof. The Norm Must Go On: Dynamic Unsupervised Domain Adaptation by Normalization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14745–14755, New Orleans, LA, USA, June 2022. IEEE.

[215] Julian Strohmayer and Martin Kampel. A compact tri-modal camera unit for rgbdt vision. In *2022 the 5th International Conference on Machine Vision and Applications (ICMVA)*, ICMVA 2022, page 34–42, New York, NY, USA, 2022.

[216] Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: toward real-time reconstruction of visual perception, 2023.

[217] Thomas M. Sutter, Imant Daunhawer, and Julia E. Vogt. Generalized multimodal ELBO. In *International Conference on Learning Representations*, 2020.

[218] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[219] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

[220] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[221] Yuge Shi, Siddharth N, Brooks Paige, and Philip Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[222] Frank Nielsen. On the jensen–shannon symmetrization of distances relying on abstract means. *Entropy*, 21(5), 2019.

[223] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[224] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018.

[225] Ali Abedi and Deepak Vasisht. Non-cooperative wi-fi localization & its privacy implications. In *Proceedings of the 28th Annual International Conference On Mobile Computing And Networking*, pages 570–582, 2022.

[226] Marco Cominelli, Felix Kosterhon, Francesco Gringoli, Renato Lo Cigno, and Arash Asadi. Ieee 802.11 csi randomization to preserve location privacy: An empirical evaluation in different scenarios. *Computer Networks*, 191:107970, 2021.

[227] Paul Staat, Simon Mulzer, Stefan Roth, Veelasha Moonsamy, Markus Heinrichs, Rainer Kronberger, Aydin Sezgin, and Christof Paar. Irshield: A countermeasure against adversarial physical-layer wireless sensing. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1705–1721. IEEE, 2022.

[228] Rui Du, Haocheng Hua, Hailiang Xie, Xianxin Song, Zhonghao Lyu, Mengshi Hu, Narengerile, Yan Xin, Stephen McCann, Michael Montemurro, Tony Xiao Han, and Jie Xu. An overview on ieee 802.11bf: Wlan sensing. *IEEE Communications Surveys & Tutorials*, 27(1):184–217, 2025.

[229] Kai Wu, Zhongqin Wang, Shu-Lin Chen, J. Andrew Zhang, and Y. Jay Guo. Isac: From human to environmental sensing, 2025.

[230] Bang Chul Jung. The 6g mobile network as a smart sensor platform [mobile radio]. *IEEE Vehicular Technology Magazine*, 20(2):6–12, 2025.

[231] Aryan Kaushik, Rohit Singh, Ming Li, Honghao Luo, Shalanika Dayarathna, Rajitha Senanayake, Xueli An, Richard A Stirling-Gallacher, Wonjae Shin, and Marco Di Renzo. Integrated sensing and communications for iot: Synergies with key 6g technology enablers. *IEEE Internet of Things Magazine*, 7(5):136–143, 2024.

[232] Pi-Yun Chen, Hsu-Yung Lin, Zi-Heng Zhong, Neng-Sheng Pai, Chien-Ming Li, and Chia-Hung Lin. Contactless and short-range vital signs detection with doppler radar millimetre-wave (76–81 ghz) sensing firmware. *Healthcare Technology Letters*, 11(6):427–436, 2024.

[233] Giovanni Diraco, Gabriele Rescio, and Alessandro Leone. Radar-based activity recognition in strictly privacy-sensitive settings through deep feature learning. *Biomimetics*, 10(4):243, 2025.

[234] Hira Hameed, Mostafa Elsayed, Jaspreet Kaur, Muhammad Usman, Chong Tang, Nour Ghadban, Julien Le Kernec, Amir Hussain, Muhammad Imran, and Qammer H Abbasi. Rf sensing enabled tracking of human facial expressions using machine learning algorithms. *Scientific Reports*, 14(1):27800, 2024.

[235] Chris Xiaoxuan Lu, Stefano Rosa, Peijun Zhao, Bing Wang, Changhao Chen, John A Stankovic, Niki Trigoni, and Andrew Markham. See through smoke: robust indoor mapping with low-cost mmwave radar. In *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*, pages 14–27, 2020.

[236] Ali Alnoman, Ahmed Shaharyar Khwaja, Alagan Anpalagan, and Isaac Woungang. Emerging ai and 6g-based user localization technologies for emergencies and disasters. *IEEE Access*, 12:197877–197906, 2024.

[237] Danny Kai Pin Tan, Jia He, Yanchun Li, Alireza Bayesteh, Yan Chen, Peiying Zhu, and Wen Tong. Integrated sensing and communication in 6g: Motivations, use cases, requirements, challenges and future directions. In *2021 1st IEEE International Online Symposium on Joint Communications & Sensing (JC&S)*, pages 1–6, 2021.

[238] Am Wolfsmantel and Bernhard Niemann. Next-generation positioning within 6g. *A Fraunhofer 6G white paper*, 2023.

[239] Xinyan Chen and Jianfei Yang. X-fi: A modality-invariant foundation model for multimodal human sensing. In *The Thirteenth International Conference on Learning Representations*, 2025.

[240] Sijie Ji, Yaxiong Xie, and Mo Li. Sifall: Practical online fall detection with rf sensing. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, pages 563–577, 2022.

[241] Jian Ding and Ranveer Chandra. Towards low cost soil sensing using wi-fi. In *The 25th annual international conference on mobile computing and networking*, pages 1–16, 2019.

[242] Prasanta Kr Sen, Mrigank Sharad, and Ram Babu Roy. Enhancing horticulture field security: Intruder detection utilizing wi-fi-csi technology with esp32 modules. In *International Conference on Agriculture-Centric Computation*, pages 24–33. Springer, 2024.

[243] Samuel Vieira Ducca, Artur Jordão, and Cintia Borges Margi. Detection and classification of animal crossings on roads using iot-based wifi sensing. In *2023 IEEE Latin-American Conference on Communications (LATINCOM)*, pages 1–6. IEEE, 2023.

[244] Shihong Duan, Tianqing Yu, and Jie He. Widriver: Driver activity recognition system based on wifi csi. *International Journal of Wireless Information Networks*, 25:146–156, 2018.

[245] Sakila S Jayaweera, Beibei Wang, and KJ Ray Liu. Robust in-car child presence detection using commercial wifi. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pages 1799–1801, 2024.

[246] Muhammad Salman, Young-Duk Soe, and Youngtae Noh. Wifi-enabled occupancy monitoring in smart buildings with a self-adaptive mechanism. In *Proceedings of the 38th ACM/SIGAPP symposium on applied computing*, pages 759–762, 2023.

[247] Muhammad Salman, Lismer Andres Caceres-Najarro, Young-Duk Seo, and Young-tae Noh. Wisom: Wifi-enabled self-adaptive system for monitoring the occupancy in smart buildings. *Energy*, 294:130420, 2024.