Asia Pacific Science Press

Journal of Advances in Engineering and Technology
ISSN(O):3005-7507
ISSN(P):3005-7493
Vol. 3 No.3 (2026)

A Comprehensive Review in Wi-Fi-Based Indoor Human Activity Detection

Zengqian Song, Jingming Li*

School of Civil Engineering and Architecture, Nanyang Normal University, Nanyang, Henan, 473061, China

Corresponding author: Jingming Li, jmli@nynu.edu.cn

Abstract: Driven by the rising demand for privacy-friendly, unobtrusive indoor monitoring, Wi-Fi passive human sensing has developed into a widely studied technical paradigm. Compared with vision-based and wearable sensing techniques, Wi-Fi sensing exhibits unique advantages including robust illumination, privacy protection, wall penetration, device-free operation, and low hardware cost. Driven by advanced deep learning, this field has advanced substantially over the past decade, evolving from coarse-level activity classification to fine-grained tasks including 3D human skeleton estimation, dense body pose matching, and multi-user activity sensing. However, several fundamental bottlenecks remain unresolved, including severe performance degradation in cross-domain and cross-environment scenarios, insufficient spatial resolution for motion sensing, limited capability for multi-user signal disentanglement, and the lack of standardized benchmark datasets. This survey provides a systematic overview of cutting-edge research on Wi-Fi-enabled indoor human activity detection. We classify mainstream technologies along three dimensions: signal processing pipelines, learning paradigms, and application granularity, and further dissect core challenges including environmental adaptability, data scarcity, and system scalability. Finally, we highlight promising future research directions covering domain generalization, cross-modal foundation model distillation, multi-user collaborative sensing, and real-time on-device deployment.

Keywords: Wi-Fi Sensing; CSI; Human Activity Recognition; Indoor Sensing; Device-Free Sensing

Published: Aug 14, 2026

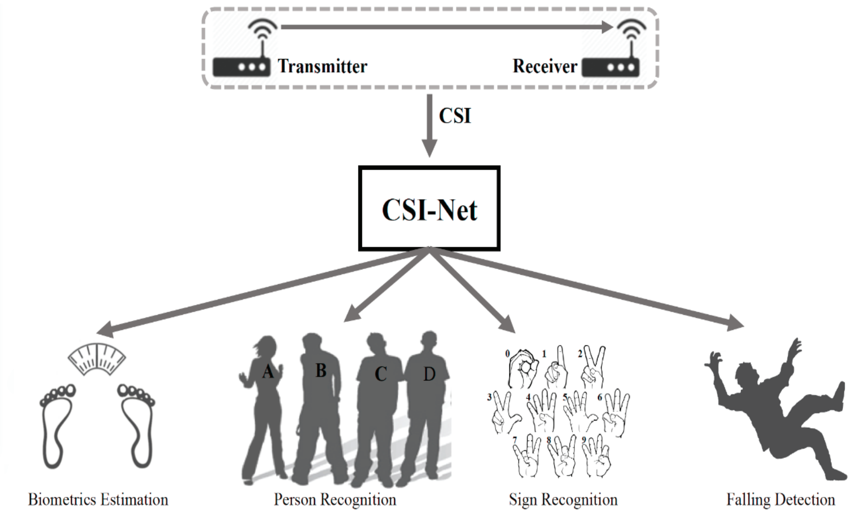
DOI: https://doi.org/10.62177/jaet.v3i3.1613

1. Introduction

1.1 Research Background and Indoor Application Demand

Recent years have witnessed growing research interest in indoor human activity detection, driven by widespread demands for intelligent systems spanning health monitoring, smart homes, elderly care, security surveillance, and human-computer interaction (HCI). Applications such as fall detection for elderly living alone [1,2] contextual awareness for energy-efficient smart homes, and hands-free gesture-based device control [3] require reliable, continuous monitoring of human movement and behavior within indoor spaces. Figure 1 shows a unified deep learning architecture for body characterization and activity recognition. The fundamental principle of Wi-Fi sensing lies in human motion perturbs the multipath propagation of Wi-Fi radio signals, causing measurable variations in the wireless channel characteristics [4,5]. Specifically, Channel State Information (CSI) provides a fine-grained and stable representation of the wireless channel that has proven highly effective for sensing human activities. Compared to earlier wireless features such as Received Signal Strength Indicator (RSSI), CSI offers significantly higher temporal and frequency resolution, enabling the detection of subtle motions including finger-level gestures [6].

Figure 1. System overview [6].



The research community has explored Wi-Fi sensing across a broad spectrum of tasks. Early work focused on coarse-grained activity classification, such as distinguishing between walking, sitting, and falling [1,4]. Subsequent research progressively pushed the granularity of Wi-Fi sensing toward finer tasks including hand gesture recognition [7], sample-level action recognition [2], 3D human pose estimation[2,8,9], dense body pose correspondence [10], and fine-grained hand skeleton construction [11]. The breadth and depth of this body of work underscores the potential of Wi-Fi as a general-purpose indoor sensing modality and motivates the need for a comprehensive survey.

1.2 Significance of Wireless Sensing Compared

To understand the unique value proposition of Wi-Fi-based sensing, it is instructive to compare it with the two dominant paradigms: vision-based and wearable-based sensing.

Over the past decade, vision-based sensing supported by computer vision techniques have achieved remarkable progress in human pose estimation and gesture recognition [12–14]Virtual Reality (VR. Unbiased data processing techniques have further improved 2D pose estimation accuracy on benchmark datasets [13], while multi-view approaches combined with mirror reflection provide additional views for resolving depth ambiguity [15]. Vision-based methods applied to structured environments such as sports analytics deliver position and pose estimations with good spatial and temporal resolution [16].

Despite these strengths, vision-based sensing suffers from several fundamental limitations. First, cameras require strict line-of-sight conditions; self-occlusion between fingers, inter-person occlusion in crowded scenes, and object occlusion during hand-object interaction represent persistent challenges that degrade detection and pose estimation accuracy [17–19]. Second, vision systems are inherently sensitive to ambient lighting; performance degrades significantly in low-light environments or under strong illumination variation [14]. Third, depth cameras such as Kinect and RealSense add hardware cost and complexity, and their field of view is limited [12,20]. Fourth and critically, vision-based systems capture identifiable visual appearance data, which raises significant privacy concerns [1,10,21]. These limitations motivate the exploration of non-visual sensing alternatives.

Wearable sensing, typically implemented through sensorized gloves with joint angle sensors or IMUs placed on limbs, offers high accuracy under controlled conditions by directly measuring body motion [22,23]. Bayesian inference methods have demonstrated that acceptable hand pose reconstruction is achievable even with a limited number of sensors by exploiting a priori knowledge of hand synergies.

However, wearable sensing imposes a persistent user burden. Sensorized gloves require physical wearing, calibration, and maintenance, which is unnatural and inconvenient for long-term or unobtrusive monitoring [12,22,23]. Furthermore, wearable systems are inherently limited to the wearer and cannot provide passive observation of multiple users in a shared environment.

Wi-Fi-based sensing occupies a complementary position relative to both vision and wearable approaches. Its principal advantages can be summarized as follows:

Table 1. Advantages of Wi-Fi sensing technology

Advantages	Key Benefits	Applications	Sources
Privacy preservation	Eliminates the risk of sensitive personal information leakage; ensures full compliance with privacy protection regulations; avoids unauthorized collection of identifiable personal data	Privacy-sensitive environments including residential homes, healthcare facilities, private office spaces, and public rest areas	[1], [10], [21], [24]
Illumination invariance	Delivers consistent, reliable sensing performance regardless of lighting fluctuations, from complete darkness to strong direct illumination; maintains stable performance in dynamic lighting environments	Nighttime monitoring, dark environment sensing, areas with frequent lighting changes, outdoor low-light surveillance, and indoor 24/7 continuous monitoring	[2], [10], [21], [25]
Non-line-of-sight and wall-penetrating capability	Enables continuous monitoring and sensing of targets located behind visual occlusions, outside the direct line of sight of the sensing device; supports sensing in cluttered, multi-obstacle environments	Through-wall monitoring, indoor occupancy detection in multi-room spaces, disaster rescue trapped person localization, and surveillance in cluttered industrial or residential environments	[1], [21], [24]
Device-free and non-intrusive	Eliminates the user burden associated with wearable or attached sensing devices; requires no active user cooperation for normal sensing operation; does not interfere with users' daily activities	Long-term daily activity monitoring, elderly fall detection, gesture recognition for smart home control, and unobtrusive in-home healthcare monitoring	[5], [11], [26]
Cost-effectiveness and ubiquity	Requires minimal additional hardware investment; can be deployed using existing Wi-Fi infrastructure; significantly reduces overall system deployment and maintenance costs	Large-scale smart city sensing, residential smart home deployments, enterprise office monitoring, and public space occupancy sensing	[11], [21] [24], [26]
Multi-user potential	Enables parallel sensing and analysis of multiple users' activities within a single space; supports simultaneous monitoring of multiple targets without performance degradation	Multi-user smart home interaction, crowded space occupancy monitoring, multi-person activity recognition, and group behavior analysis	[5], [27]

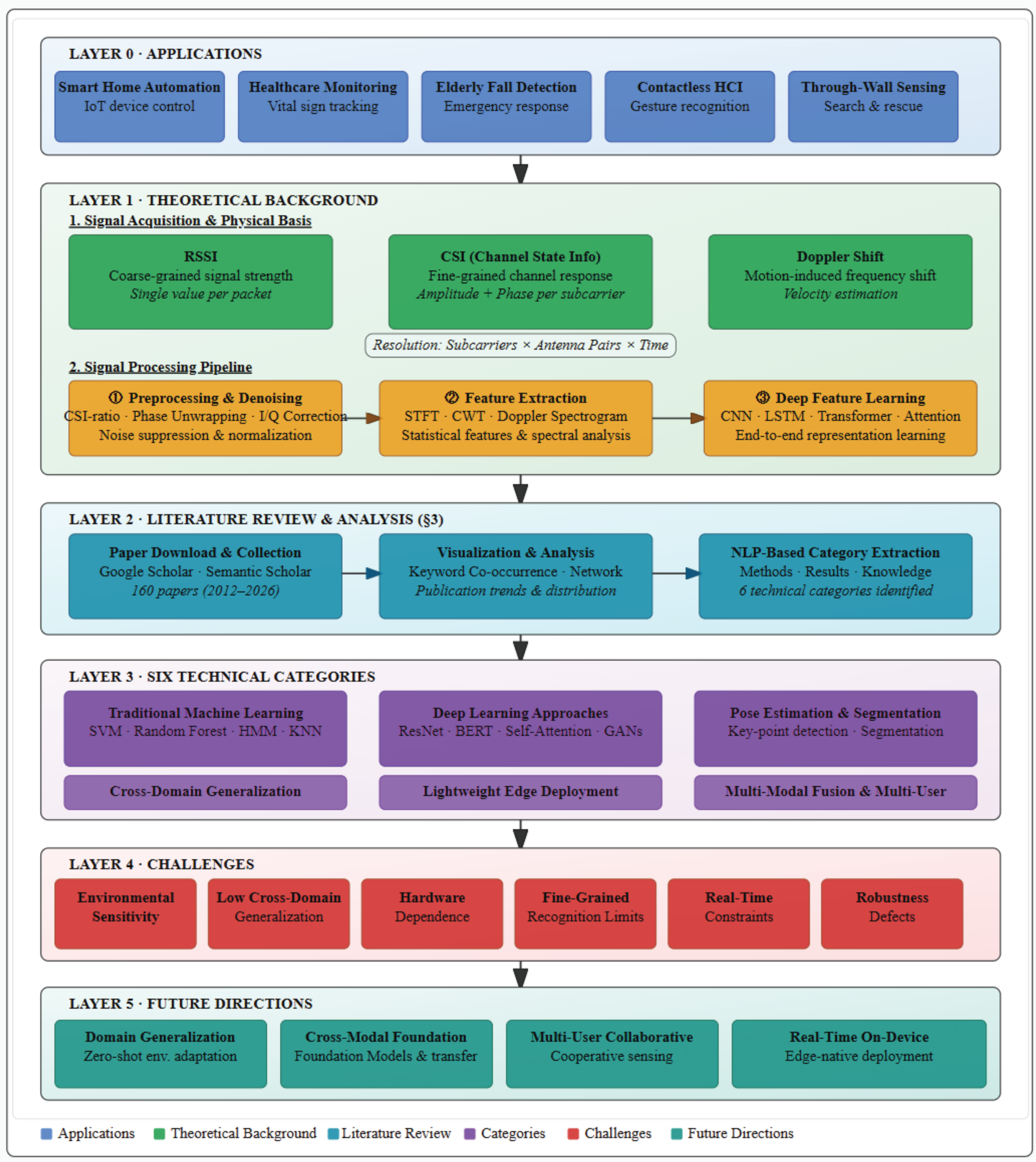
These advantages are particularly valuable for applications in elderly fall detection, smart home automation, healthcare monitoring, and contactless HCI, where privacy, reliability under adverse conditions, and unobtrusiveness are paramount requirements.

1.3 Overall Structure of This Review

This article systematically surveys indoor human detection techniques built upon Wi-Fi sensing, with its structure outlined below. Section 2 presents the methods used for literature review and analysis. Section 3 surveys the mainstream technical approaches, categorizing methods along three dimensions: (i) signal processing and feature extraction pipelines, (ii) learning paradigms ranging from hand-crafted feature classification to end-to-end deep learning, and (iii) task granularity from coarse-grained activity classification to fine-grained pose estimation and gesture recognition. Section 4 discusses representative application scenarios, including smart home automation, healthcare monitoring, elderly fall detection, HCI, and through-wall sensing. Section 5 systematically analyzes the core challenges facing the field, including cross-domain generalization, environmental sensitivity, multi-user disentanglement, data scarcity, and computational constraints for edge deployment.

Finally, Section 6 identifies key future research directions and concludes the review. The general framework of the paper is illustrated in Figure 2.

Figure 2. Research framework.



2. Theoretical Background

2.1 Basic Characteristics of Wi-Fi Signals

Wi-Fi-based human sensing relies on the fine-grained characterization of wireless channel conditions, which can be captured through several signal-level representations. The most widely adopted features in Wi-Fi sensing include RSSI, CSI, and their derived components—phase, amplitude, and Doppler shift. Understanding the physical meaning and sensing capability of each of these representations is essential for designing robust activity detection systems.

RSSI is the simplest and most readily available wireless signal feature, representing the overall received power of a Wi-Fi signal at the antenna level. Early Wi-Fi sensing systems predominantly relied on RSSI for human presence detection and activity classification due to its ease of extraction from standard Wi-Fi Network Interface Cards (NICs) [26]. However, RSSI provides only a scalar measurement per subcarrier and is inherently coarse-grained. In complex indoor environments, RSSI-based sensing degrades significantly because it cannot distinguish among multipath components and is highly susceptible to slow-fading and time-varying channel dynamics [28]. RSSI alone is generally insufficient for fine-grained tasks such as gesture recognition or pose estimation.

As a fine-grained channel characterization tool, CSI gradually replaces coarse RSSI and becomes the core signal carrier for high-precision human motion capture. CSI is essentially the sampling value of the channel frequency response across multiple orthogonal subcarriers [28]. In contrast to the single-value RSSI, CSI captures the complex gain for each individual subcarrier and for each transmit-receive antenna pair [4]. CSI provides finer-grained and more stable measurements of the wireless channel compared with RSSI, enabling its use in a wide range of sensing tasks including activity recognition, gesture detection, and vital sign monitoring [29].

Doppler shift arises from the relative motion between the signal transmitter, receiver, and reflecting objects in the environment. When a human body moves within the Wi-Fi signal propagation space, the frequency of the reflected signal is shifted in proportion to the radial velocity of the moving reflector. The Doppler effect manifests in CSI as time-varying phase changes across consecutive packets, and can be explicitly estimated through techniques [30]. Doppler-based features, including Doppler spectrograms and Range-Doppler Indicators, have proven effective for capturing the temporal dynamics of human motion and are widely employed in gesture recognition and activity classification [27].

2.2 Indoor Multipath Effect and Wi-Fi Human Sensing Mechanism

Indoor Wi-Fi signal propagation is fundamentally shaped by the multipath effect, which arises from the reflection, diffraction, and scattering of radio waves by walls, furniture, and other objects in the environment. The multipath structure of Wi-Fi signals in indoor environments has been extensively modeled and studied, as shown in Table 2.

Table 2. Wi-Fi Human Sensing Mechanisms and Indoor Multipath Propagation

Sensing Mechanism	Representative Literature
The human body as a dynamic reflector	[1], [4]
Sensing model	[31]
Angle of arrival and spatial sensing	[8], [9]
Through-wall sensing capability	[28]

Human body as a dynamic reflector. The human body, composed largely of water and biological tissues with distinct dielectric properties at Wi-Fi frequency bands, reflects a significant portion of incident Wi-Fi energy. The dielectric characteristics of biological tissues influence how Wi-Fi signals are absorbed, reflected, and transmitted through the body [1]. The dynamic state of the human subject can be modeled by the deployment variables including the positions and orientations of Wi-Fi transceivers, the room layout, and the arrangement of static objects [31].

Beyond the raw CSI tensor, Wi-Fi-based sensing systems can exploit spatial information encoded in the signal. The two-dimensional angle of arrival of signals reflected from different parts of the human body can be estimated using antenna array signal processing techniques algorithm applied to L-shaped antenna arrays at the receiver [8]. This approach enables environment-independent pose estimation even in non-line-of-sight scenarios [9]. Wi-Fi signals possess the ability to penetrate common building materials including drywall, plaster, and wood, albeit with moderate signal attenuation. The attenuation of Wi-Fi signals through typical building materials is relatively small [28].

2.3 Core Theoretical Basis for Human Activity Perception

The core challenge in Wi-Fi-based human activity perception is to establish a reliable mapping between the time-varying CSI measurements and the underlying human activity or gesture class. Raw Wi-Fi CSI data is inherently noisy, containing both environmental noise and hardware-induced artifacts. Several preprocessing techniques have been developed to enhance the quality of CSI before feature extraction. The CSI-ratio method is a widely adopted denoising approach that computes the ratio of CSI between two receivers, effectively canceling out common-mode hardware distortions while preserving human-induced variations [30].

The extraction of discriminative features from CSI data streams forms the theoretical foundation of Wi-Fi-based activity recognition [4]. Several signal processing methods have been extensively employed for this purpose as listed in Table 3.

Table 3. Common time-frequency analysis and feature extraction methods

Methods	Core Mechanism	Key Functional Capabilities
Short-Time Fourier Transform (STFT) [30]	Applied to CSI time series to generate time-frequency representations; divides the CSI signal into short overlapping windows and computes the Fourier transform within each window to reveal the temporal evolution of frequency-domain characteristics	Produces a two-dimensional image-like representation that captures both the spectral content and the temporal dynamics of CSI variations caused by human motion
Discrete Wavelet Transform (DWT) [32]	Decomposes the CSI signal into multiple frequency bands at different resolution levels to provide a multi-scale representation of the signal; balances temporal and spatial resolution	Captures both transient high-frequency signal components and sustained low-frequency patterns of CSI signals; delivers multi-scale signal characterization
Doppler spectrograms [27]	Constructed by extracting Doppler frequency shift information from processed CSI phase data; generated by applying FFT along the time axis of CSI-ratio processed signals, with Doppler spectra from multiple receivers concatenated along the time axis to produce fused images with multi-angle motion information	Explicitly represents the velocity profile of human motion over time; captures distinctive velocity signatures of different human motions; provides multi-angle fused motion features

CNNs have become the most widely used deep learning model for Wi-Fi sensing, owing to their ability to capture spatial patterns in CSI spectrograms and Doppler features [33]. Long Short-Term Memory (LSTM) recurrent neural networks are particularly suited for modeling the temporal dependencies inherent in CSI time series [10]. Hybrid architectures combining CNNs and LSTMs leverage both the spatial feature extraction capability of CNNs and the temporal modeling strength of LSTMs, achieving superior performance in gesture recognition and pose estimation tasks [5].

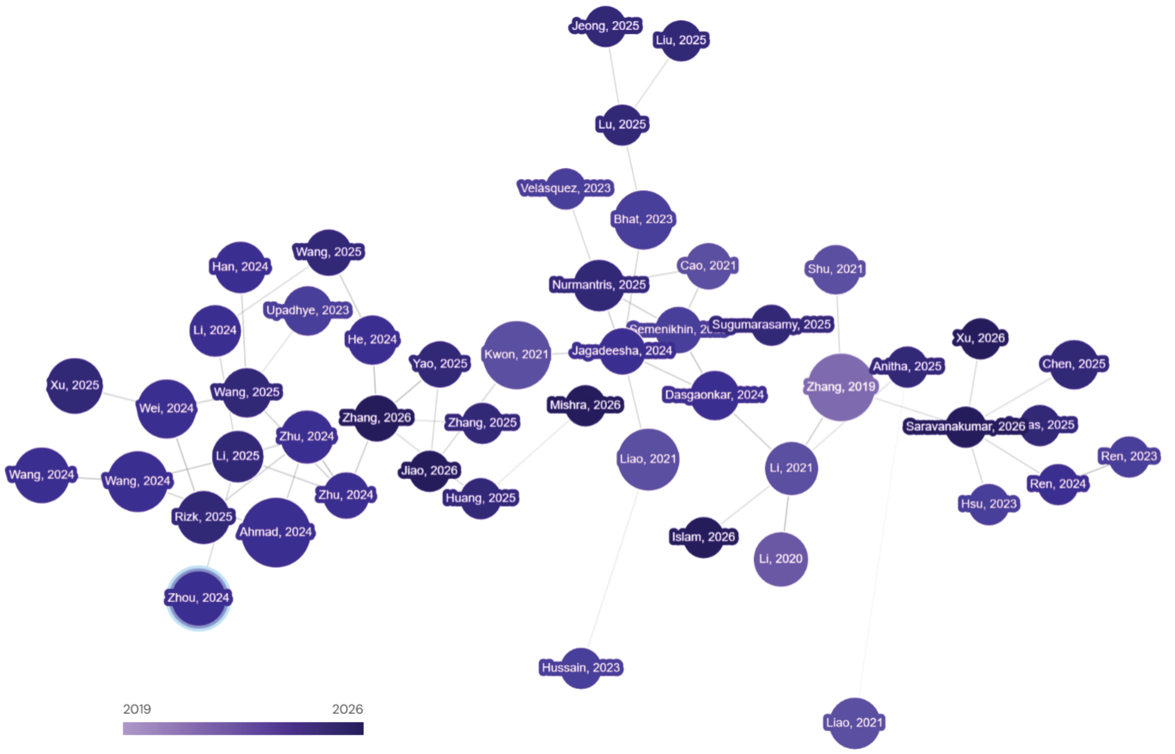
In series-level recognition, a complete CSI distortion series corresponding to one activity instance is classified as a single action label, typically using dynamic time warping for sequence matching [2]. Temporal U-Net architectures have been proposed to address the challenge of sample-level recognition by learning action features from time-serial Wi-Fi distortions while leveraging neighboring context.

Attention mechanisms are used for identifying the most informative components of the high-dimensional CSI tensor. Variational dual-path attention networks (VDAN) have been designed to explicitly model the frequency-time sparsity of CSI signals and utilize variational inference for noise-robust feature selection. Attention-based approaches represent a principled fusion of physical signal priors and data-driven representation learning. Wi-Fi latent domain mining uses unsupervised clustering to automatically discover latent domain factors responsible for distributional shifts [30]. This latent-domain framework provides a theoretically grounded approach to mitigating the domain shift problem that has long limited the practical deployment of Wi-Fi sensing systems.

3. Methods

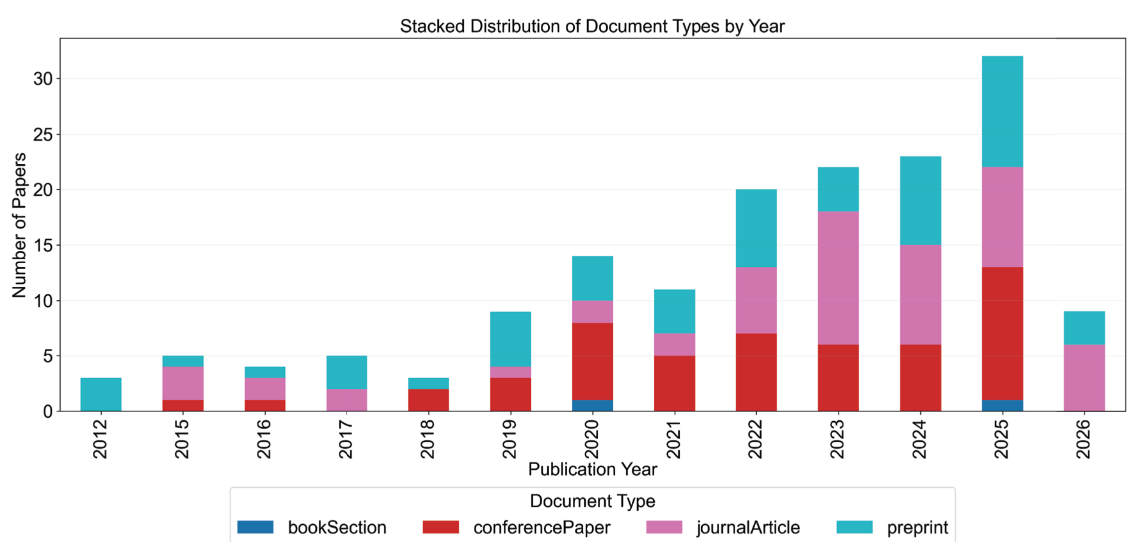
The field of Wi-Fi based indoor human activity detection has undergone rapid evolution since its inception. Papers within the past ten years have been downloaded from databases, such as Google Scholar, Semantic Scholar, ScienceDirect. Key words are Wi-Fi, building environment, indoor human detection. The graphic network in Figure 3 illustrates the connections of the 160 papers.

Figure 3. Literature connections of the downloaded papers.



This stacked bar chart in Figure 4 illustrates the annual output of four categories of academic documents related to Wi-Fi through-wall human sensing from 2012 to 2026. The horizontal axis represents publication years, and the vertical axis denotes the total number of papers published each year. The stacked segments stand for four document types: book section (dark blue), conference paper (red), journal article (pink), and preprint (cyan). The total height of each bar corresponds to the aggregate annual publications, while each colored segment reflects the quantity and proportional share of each document category.

Figure 4. Stacked distribution of document types by year



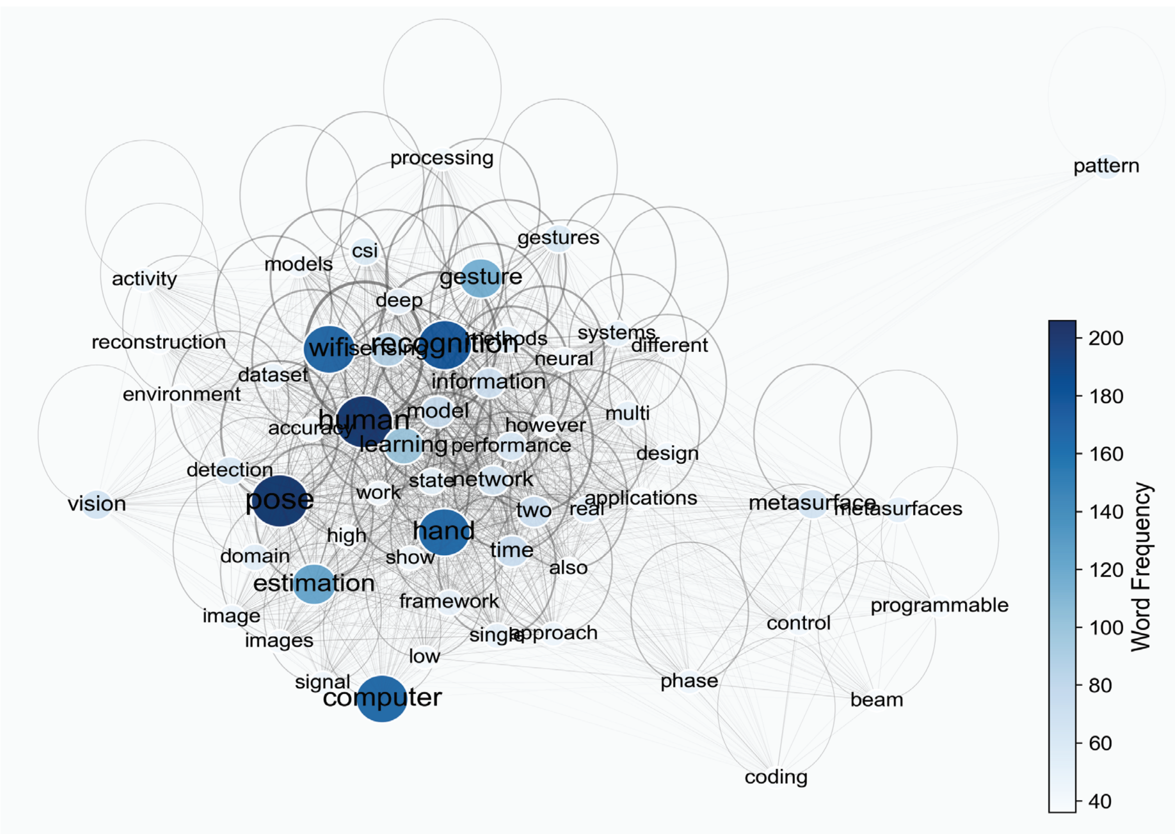
Based on a comprehensive analysis of the literature, we identify four distinct phases of development as listed in Table 4. The total number of Wi-Fi through-wall human sensing publications has expanded steadily year by year, accompanied by rising research enthusiasm. The publication landscape has evolved from an early system relying solely on preprints into a diversified structure: conference papers remain the primary output form, journal articles show robust growth, preprints provide regular supplementary sharing, and book sections contribute limited formal academic summaries.

Table 4. Distinct phases of Wi-Fi based indoor human activity detection development

Phase	Year	Representative Literature
Early Wi-Fi Sensing and Gesture Recognition	2015–2017	[4] [26]
Deep Learning Empowerment and Pose Estimation	2018-2020	[1], [2], [25], [34]
Fine-Grained 3D Pose Estimation and Benchmarking	2021-2023	[8], [11], [10], [24], [9], [33], [35]
Robustness, Multi-User Sensing, and Foundation Models	2024-Present	[5], [29], [27], [36], [37], [30], [31]

Figure 5 extracts keywork co-occurrence from the abstracts and visualizes research hotspots, word frequencies and topic correlations. The darkest, largest central nodes including pose, human, recognition, Wi-Fi, hand and computer form the absolute core of this field, surrounding by a densely connected main research cluster covering CSI signal processing, deep learning neural network models, and mainstream human sensing tasks such as human activity detection, gesture recognition and pose estimation, with frequent references to computer vision and image comparison as a common benchmark.

Figure 5. Top 60 keyword co-occurrence graph



Using Natural Language Processing (NLP), the methods, results, and knowledge of these articles are extracted. Six primary technical categories are identified, which are traditional machine learning, deep learning methods, Fine-Grained pose estimation and continuous segmentation, Cross-Domain generalization and robustness enhancement sensing, lightweight model and Edge-Deployment optimization, and Multi-Modal Fusion. Section 4 explained more detailed analysis of the downloaded papers.

4. Results

This chapter presents a systematic taxonomy of Wi-Fi-based indoor human activity detection methods. We classify existing approaches into six broad categories according to their algorithmic paradigm and deployment characteristics: (1) traditional machine learning methods; (2) deep learning methods; (3) fine-grained pose estimation and continuous segmentation, (4) cross-domain generalization and robustness enhancement sensing, (5) lightweight and edge-deployable sensing methods designed for resource-constrained environments, and (6) Multi-Modal fusion, multi-user sensing and new modeling. For each category, we elaborate the technical evolution, representative algorithms, performance characteristics, and applicable conditions.

4.1 Traditional Machine Learning Methods for Gesture Recognition

The earliest generation of Wi-Fi-based activity recognition systems relied on a three-stage pipeline: signal preprocessing, hand-crafted feature extraction, and classification via traditional machine learning algorithms. The paradigms for scenarios demanding low computational overhead and model interpretability are listed in Table 5.

Table 5. Features of traditional machine learning methods

Method	Core Signal Feature	Algorithm	Applications	Papers
RSSI-based Gesture Recognition	WiFi RSSI, basic signal primitives: step, slope, spike, fluctuation waveform	Extract time-domain change primitives of RSSI, combine wavelet filtering for signal denoising, and use traditional machine learning algorithms such as SVM and Random Forest to realize gesture classification and recognition	Device-free in-air gesture interaction, simple smart home control, basic interactive operation of multimedia players	[26]
Hand-Crafted Feature + ML Classification	CSI amplitude and phase information, manually designed statistical features: mean, variance, energy, entropy and other time-domain statistical features	Use sliding window mechanism to complete time-domain feature extraction, combine Dynamic Time Warping to realize sequence matching, and rely on SVM, k-NN, Random Forest, Hidden Markov Model	Human coarse-grained behavior recognition, personnel identity verification, basic human fall detection	[4,28]

RSSI-based gesture recognition requires no hardware modification, is compatible with standard WiFi devices, needs no specialized training, supports through-wall sensing, and enjoys extremely low deployment cost and strong universality, yet it suffers from limited, single recognizable gesture types, lower recognition accuracy compared with CSI-based methods, and signals that are highly susceptible to environmental occlusion, human movement and multipath interference, leading to poor stability.

By contrast, the scheme combining hand-crafted features and machine learning classification features low computational overhead, fast inference speed, highly interpretable features and low demand for labeled training data, making it suitable for lightweight deployment on edge terminals, but its features rely on manual empirical design with strong specificity and weak generalization, resulting in poor cross-scenario and cross-environment adaptability, and it can only realize coarse-grained behavior recognition rather than fine-grained action identification.

4.2 Deep Learning Methods

The advent of deep learning has fundamentally transformed Wi-Fi-based human activity detection. Deep architectures automatically learn feature representations from raw or minimally processed CSI data, eliminating the need for manual feature engineering and enabling significantly more complex sensing tasks including fine-grained 3D pose estimation and multi-user activity tracking, as listed in Table 6.

Table 6. Deep learning end-to-end basic sensing

Method	Core Signal Feature	Algorithms	Applications	Papers
Fine-Grained Finger Gesture Recognition	5GHz WiFi subcarrier-level fine-grained CSI amplitude and phase information, capturing subtle signal distortions corresponding to tiny finger movements	Complete adaptive environmental noise removal, model the inherent behavior characteristics of gestures, realize multi-link signal separation, and combine SVM and GAN to complete fine-grained gesture classification	Finger-level Human-Computer Interaction, VR virtual interaction, fine control of smart home, somatosensory game interaction	[38]

Method	Core Signal Feature	Algorithms	Applications	Papers
Unified DNN for Body Characterization	Raw subcarrier-level CSI amplitude and phase full time-series signals, no manually filtered features	Build an end-to-end unified network based on ResNet-18, realize multi-task joint learning of human biometrics estimation and pose recognition, and completely abandon the manual feature design process	Human biometrics estimation, personnel identity recognition, sign language gesture recognition, home human fall detection	[1]
LSTM / Recurrent Networks	CSI continuous time-series sequence features, focusing on the dynamic change law of amplitude and phase in the time domain	Model time-series dependencies based on LSTM-RNN, capture the long-range time-series correlation characteristics of CSI data streams, and support CNN-LSTM hybrid architecture to integrate spatial and time-series features	Human daily behavior recognition, dynamic gesture recognition, wireless signal time-series pattern detection	[4]

CNNs are the dominant architecture for Wi-Fi-based gesture recognition and activity classification. CSI-Net [1] jointly solved multiple body characterization tasks using a shared backbone. Recurrent architectures, particularly LSTM networks, are well-suited for Wi-Fi sensing because CSI data is inherently temporal and sequential. LSTM networks capture long-range temporal dependencies in CSI sequences, modeling the dynamic evolution of signal perturbations caused by human motion [4]. Tan et al. [38] uses an explicit environmental noise Even using 5 GHz band with larger bandwidth, the system must be recalibrated when the environment changes substantially.

4.3 Pose Estimation and Continuous Segmentation

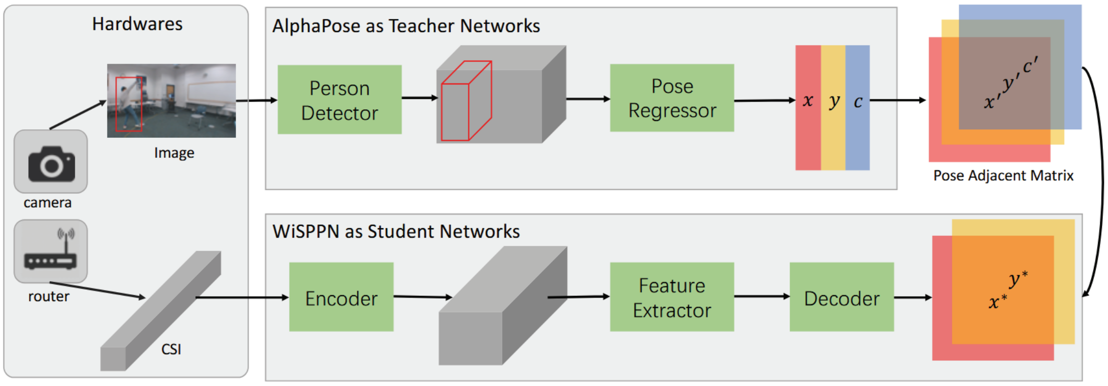
Wi-Fi-based 3D human pose estimation represents the frontier of fine-grained sensing, moving beyond activity classification to reconstruct the full skeletal structure of a human body. Representative studies are listed in Table 7.

Table 7. Fine-grained pose estimation and continuous segmentation

Methods	Core Signal Feature	Algorithms	Application	Papers
FCN-Based Single-Person Pose Estimation	CSI signal + Doppler shift feature, equipped with 3-antenna transceiver synchronization system, matched with camera key point annotation information	Use ResNet as the backbone network, combine the position attention module to embed key point coordinates and displacement information, and realize pixel-level pose estimation through the full convolutional network	Single-person fine-grained pose estimation, through-wall human monitoring, no-light/dark environment human sensing	[25]
Temporal U-Net for Sample-Level Recognition	CSI time-domain continuous distortion sampling data, focusing on subtle signal change features at the single-sample level	Adapt the U-Net network in the image segmentation field to the time-series version, realize per-sample classification of CSI distortion signals, and complete end-to-end time-series feature autonomous learning	Human continuous behavior monitoring, real-time action semantic segmentation, home high-precision fall detection and warning	[2]
3D AoA-Based Pose Tracking	CSI two-dimensional AoA spectrum, including azimuth and elevation spatial features, supporting three-dimensional spatial positioning	Complete two-dimensional AoA signal separation based on L-shaped antenna array and MUSIC algorithm, model the independent motion trajectory of limbs, and realize limb-joint linkage mapping and overall motion modeling	Three-dimensional free posture tracking, VR/AR immersive interaction, medical action training, smart home high-end somatosensory interaction	[8], [9]

Limited by the lack of pixel-level joint mapping for wireless signals, early Wi-Fi pose estimation methods suffered from low reconstruction accuracy. To tackle this gap, Wang et al. [25] put forward WiSPPN, an FCN-driven framework dedicated to single-person skeletal tracking via commodity Wi-Fi in Figure 6. The system uses a 3-antenna Wi-Fi sender and a 3-antenna receiver to generate CSI data, synchronized with camera-captured videos for key point annotations. Breaking the limitation of series-level integral classification, Temporal U-Net [2] realizes frame-by-frame semantic segmentation of continuous CSI sequences, enabling more granular real-time action perception. Whereas prior work categorizes an entire CSI distortion series into a single action label, temporal U-Net classifies every individual CSI sample in the series.

Figure 6. WiSPPN System Framework [25].



Different from static pose reconstruction schemes, Winect [8] builds a 3D tracking framework that supports free-form human activity perception with off-the-shelf Wi-Fi devices. Built upon the static pose tracking framework of Winect, GoPose [9] further optimizes the AoA signal decoupling algorithm to support tracking of unrestricted, dynamic human movements. It leverages the 2D AoA spectrum of signals reflected from the human body combined with deep learning to model the complex relationship between 2D AoA spectrums and 3D skeletons. However, the system relies on customized multi-element antenna arrays, raising the hardware deployment threshold; it only supports predefined limb motion geometric models, failing to adapt to unconstrained extreme free movements.

4.4 Generalization and Robustness Enhancement Sensing

Models trained in one environment often fail when deployed in another due to environmental sensitivity of CSI. Several deep learning methods specifically address this challenge, representative papers are listed in Table 8.

Table 8. Cross-domain generalization and robustness enhancement sensing

Methods	Core Signal Feature	Algorithms	Applications	Papers
Cross-Domain Gesture Recognition	CSI time-frequency features, mapped to text semantic space, realizing cross-domain correlation between wireless signals and semantic information	Rely on HMM to model time-series dynamic features, CNN to extract spectrum features, combine word2vec word embedding and verb attribute encoding, to realize cross-domain projection from radio frequency signals to text semantics	Behavior recognition in scenarios with scarce labeled data, rapid deployment of new gesture sets, low-cost scenario-based model iteration	[34]
Attention-Based Cross-Domain Gesture Recognition	Multi-receiver fused CSI Doppler spectrum, converted into multi-angle semantic image features	Based on ResNet-18 backbone network, fuse SMSA and improved CBAM, to realize multi-angle signal feature fusion	Complex cross-environment gesture recognition, high robustness smart home non-intrusive interaction, multi-scenario general somatosensory control	[27]
Domain-Consistent Representation Learning	CSI time-series signal, construct mask reconstruction self-supervised features and contrastive learning dual-view features	Adopt time-series consistent contrastive learning combined with uniformity regularization, rely on GCN + Transformer decoder, combine task prompts to realize topology-constrained skeleton generation	Cross-environment general 2D/3D human pose estimation, multi-scenario model migration deployment, complex environment robust sensing	[36]
Geometry-Conditioned Cross-Layout Pose	CSI signal + calibrated transceiver geometric position embedding features, fused with spatial layout information	Build a unified three-dimensional space coordinate system, take the transceiver position as conditional encoding, realize the disentanglement of human motion features and equipment layout features, and eliminate layout interference	Cross-scenario three-dimensional pose estimation, large-scale intelligent interaction deployment, multi-layout general non-intrusive sensing system	[31]

Cross-domain gesture recognition [34] maps CSI time-frequency features into a text semantic space by combining HMM time-series modeling, CNN spectrum extraction, and word2vec verb embedding to project radio signals to text semantics. It recognizes unseen gestures without extra labeled data, reducing data collection and iteration costs by leveraging existing text corpora; however, recognition precision degrades drastically when expanding unseen motion categories, and tedious cross-domain parameter tuning is required to restore model performance. To convert multi-receiver fused CSI Doppler spectra into multi-angle semantic image features, the attention-based cross-domain gesture recognition scheme [27] integrates a ResNet-18 backbone, SMSA module and improved CBAM channel self-attention mechanism for multi-view signal fusion, achieving outstanding accuracy and environmental robustness. Despite these strengths, it demands expensive multi-receiver hardware and bears heavy computational overhead caused by stacked attention modules. Domain-consistent representation learning [36] builds self-supervised mask reconstruction and contrastive dual-view features from CSI time-series signals. It reduces labeled data reliance via self-supervision, and boosts environmental invariance, but requires lengthy pre-training, and poses high real-world deployment barriers. Geometry-conditioned cross-layout pose [31] embeds CSI signals with calibrated transceiver geometric coordinates and spatial layout data, constructing a unified 3D coordinate system to encode device layout conditions and decouple human motion from equipment placement artifacts, yet it mandates checkerboard-based device calibration and depends on expensive multi-layout labeled datasets for model training.

4.5 Lightweight Model and Edge-End Deployment

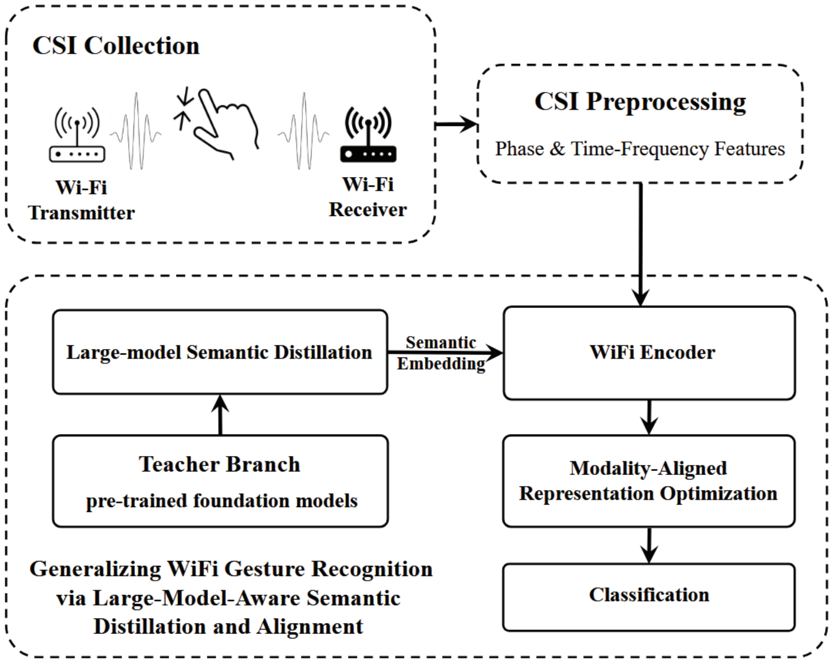
Focusing on lightweight, low-latency inference optimized for edge and mobile device deployment, studies utilize dual-path processing frameworks to divide feature extraction and prediction workflows, extract refined multi-dimensional wireless signal features paired with targeted filtering modules to counter signal noise. Table 9 listed the studies attain high task-specific accuracy at the cost of unique structural, training, or hardware restrictions that limit broad general applicability.

Table 9. Lightweight model and edge-end deployment optimization

Methods	Core Signal Feature	Algorithms	Applications	Papers
VDAN	CSI time-frequency sparse features, focusing on effective signal components in low SNR environments	Adopt dual-path architecture of feature refinement and classification decoupling, rely on variational inference to realize noise-robust feature selection, and model CSI signal sparsity through attention weights	Resource-constrained edge equipment sensing, low SNR complex environment gesture recognition, interpretable wireless sensing system	[39]
Large-Model Semantic Distillation (GLSDA)	CSI ratio phase sequence + Doppler spectrum cross-modal features	Dual-path encoding architecture, combined with multi-scale semantic encoder and pre-trained large model cross-modal attention, construct lightweight student network through robust dual distillation algorithm	General gesture interaction interface, real scene AIoT intelligent sensing, resource-constrained edge terminal real-time inference	[37]
mmWave Tiny CNN for On-Device Inference	60GHz millimeter wave radar Range-Doppler Image, based on 802.11ad protocol pulse Doppler signal	Build ultra-lightweight CNN-LSTM hybrid network, solve the signal frame loss problem through RDI sequence restoration, and adapt to device-end real-time inference	Smartphone end gesture recognition, mobile human-computer interaction, non-contact device intelligent control	[40]

VDAN [39 leverages CSI time-frequency sparse features with variational inference for noise-resistant feature screening, delivering robust low-SNR accuracy with efficient inference and transparent attention weights yet suffers partial joint optimization loss and weak system integration. GLSDA [37] processes cross-modal CSI phase-Doppler data via pre-trained model distillation to create edge-friendly student networks, achieveing 92.80% cross-direction recognition with drastically compressed model size for AIoT deployment but faces complex, power-heavy distillation training. As illustrated in Figure 7, Ren et al. [40] extracts images through a minimal-parameter CNN-LSTM hybrid optimized for local device real-time inference, offering ultra-small parameters, sub-100ms mobile latency, and high spatial radar resolution yet is limited by narrow mobile antenna space and exclusive reliance on dedicated 60GHz radar hardware.

Figure 7. GLSDA framework for WiFi-based gesture recognition [37].



4.6 Multi-Modal Fusion, Multi-User Sensing and New Modeling

Targeting hybrid multi-modal wireless signals, studies tackle indoor human sensing tasks like pose tracking, gesture recognition and signal imaging, each integrating specialized neural network frameworks to fix real-world sensing challenges such as environmental domain shifts and multi-target interference, as listed in Table 10. The 3D hand skeleton construction [11] method relies on a depth camera to provide cross-modal supervision for model training, though it is limited by palm signal occlusion that impairs the perception of subtle fingertip movements. Such method is widely adopted in accessible intelligent interaction scenarios, including high-precision finger tracking, sign language recognition, somatosensory game interaction, and medical rehabilitation monitoring. Focusing on robust gesture recognition in complex environments, GesFi [30] faces inherent drawbacks of unstable training performance and gradient oscillation in adversarial learning, and its domain alignment accuracy is heavily dependent on data distribution and clustering effects.

Table 10. Fusion of Multi-modal Signals, Multi-user Activity Sensing and Innovative Modeling Strategies

Methods	Core Signal Feature	Algorithms	Applications	Papers
3D Hand Skeleton Construction	CSI + millimeter wave multi-modal phase and amplitude features, combined with depth supervision cross-modal information	Based on HandNet multi-task learning network, synchronously realize two-dimensional hand mask segmentation and 21-joint three-dimensional skeleton prediction, embed domain generalization module to improve scene adaptability	High-precision finger tracking, sign language recognition, somatosensory games, medical rehabilitation monitoring, accessible intelligent interaction	[11]
Transformer-Based Spatiotemporal Pose	CSI high-dimensional spatiotemporal fusion features, including time-domain dynamic change and spatial structure correlation information	Build dual-branch ViSTA-Former network, model time-domain dependency and human body structure correlation respectively, add velocity modeling branch to capture short-term displacement change of key points	Smart home elderly care monitoring, elderly behavior monitoring, non-invasive health monitoring, home fall warning	[41]
Latent Domain Mining (GesFi)	Denoised CSI ratio feature + STFT time-frequency visualization representation	Based on class-level adversarial learning to suppress redundant gesture semantics, mine latent domain distribution differences through unsupervised clustering, and realize adaptive adversarial domain alignment	High robustness cross-environment gesture recognition, multi-scenario non-adaptive large-scale deployment, complex dynamic environment sensing	[30]

Methods	Core Signal Feature	Algorithms	Applications	Papers
Weakly-Supervised Domain Adaptation (Adapose)	CSI time-series frame features, construct mapping consistency constraint representation	Introduce mapping consistency loss function to align domain distribution difference, realize supervised pose transfer, and optimize model performance through regression scale consistency regularization	Cross-site general pose estimation, smart city human sensing, rapid deployment of new environments	[24]
Multi-User Activity Sensing (WiMANS)	2.4G/5G dual-band CSI signal + synchronous video multi-modal annotation data, including multi-user mixed signal features	Based on CNN, LSTM, Transformer, GAN and other architectures, complete benchmark test of multi-user CSI samples, synchronously model personnel identity, position and behavior characteristics	Multi-user smart home monitoring, indoor security monitoring, multi-person behavior situation analysis	[5]
Multimodal VAE for Through-Wall Imaging	CSI time-domain signal features, realize cross-modal mapping from wireless signals to visual images	Complete CSI-to-image generative conversion based on multi-modal VAE architecture, optimize the quality of generated images through uniform/ Gaussian weighting in the time-domain dimension, and support downstream visual tasks	Through-wall visual monitoring, cross-room human behavior perception, camera-free visual monitoring scenarios	[21]
DensePose from WiFi	Linear fitting denoised CSI phase and amplitude pure signal features	Combine FCN and U-Net network, map WiFi signals to UV coordinates of 24 human body regions, and realize whole-body dense pose correspondence modeling	Human body dense pose estimation, privacy-level home human monitoring, multi-target non-intrusive sensing scenarios	[10]

Designed for universal human pose estimation, Adapose [24] is unable to achieve complete zero-label adaptation and demands a small number of weakly labeled target-domain data for model fine-tuning. It excels at cross-site universal human pose estimation, smart city human sensing, and rapid model deployment in unfamiliar new environments. Targeting daily indoor human monitoring, the transformer-based spatiotemporal pose method [41] is plagued by high computational overhead, slow inference speed and high deployment costs due to its complex network structure and rigorous hyperparameter tuning requirements, yet it is reliably applied to smart home elderly care monitoring, daily elderly behavior tracking, non-invasive health monitoring and home fall warning scenarios. In terms of multi-person perception, WiMANS [5] delivers lower multi-user behavior recognition accuracy than conventional video-based approaches, and it requires massive workloads and high iteration costs for dataset collection and annotation, making it applicable to multi-user smart home monitoring, indoor security surveillance and multi-person behavioral situation analysis.

Devoted to privacy-preserving and camera-free human perception, multimodal VAE for through-wall imaging [21] relies on paired CSI-image training data, and its limited generated image resolution hinders the performance of high-precision downstream visual tasks. It enables camera-free perception and is deployed for through-wall visual monitoring, cross-room human behavior sensing and privacy-protective visual surveillance. Similarly, the DensePose [10] from WiFi method achieves non-intrusive human sensing but has lower spatial resolution than optical cameras; it also requires extra phase denoising preprocessing and involves cumbersome data collection and high-difficulty annotation work. This method is mainly applied to full-body dense pose estimation, privacy-oriented home human monitoring and non-intrusive multi-target human sensing scenarios.

5. Discussions and limitations

Despite the substantial progress documented in Section 4, Wi-Fi-based indoor human activity detection still faces fundamental challenges that limit its widespread practical deployment. This section systematically categorizes and discusses the existing problems identified across the literature.

5.1 Environmental Interference Sensitivity

Wi-Fi CSI is highly sensitive to environmental variables, making the detection performance inherently fragile when the physical environment changes. The fundamental principle of Wi-Fi sensing relies on the modulation of wireless signal propagation paths by human body movement; however, any environmental change alters the multipath propagation characteristics in ways that are indistinguishable from human-induced signal variations [4,24]. Environmental variations drastically distort CSI distribution and trigger severe domain shifts. AdaPose [24] explicitly summarized direct deployment of pre-trained models to new environments delivers far from optimal performance. The Temporal U-Net framework for sample-level action recognition notes the challenge of distinguishing action-induced Wi-Fi distortions from ambient signal variations in the temporal domain [2]. Cross-domain gesture recognition studies consistently report performance degradation in untrained environments, even with attention mechanisms and multi-angle feature fusion [27]. These findings underscore that environmental robustness remains a fundamental bottleneck: the same Wi-Fi signal pattern can correspond to entirely different physical scenarios, and the inverse mapping from signal to activity is inherently ambiguous without environmental priors.

5.2 Low Generalization Ability

The generalization capability of Wi-Fi sensing models across different environments, device layouts, and user populations represents one of the most critical limitations identified in the literature. Multiple pose estimation studies, including the work of Chen et al. [36], point out two core obstacles restricting model generalization: the significant pose distribution differences between source and target environments and the structural fidelity gap of decoded poses violating human skeletal anatomical topology.

The domain generalization challenge extends beyond pose estimation to gesture recognition. Research from Zhang et al. [30] exposes the defects of traditional adversarial domain adaptation: these methods overly depend on artificially divided physical domain labels that cannot reflect real data distribution patterns, bringing classification conflicts and manifold distortion that fundamentally undermine adversarial generalization performance. Wi-Fringe addresses the related data scarcity problem by leveraging text semantics to enable zero-shot recognition of previously unseen activities, but its accuracy drops from 90% for two unseen activities to 61% for six unseen activities, indicating limited scalability of cross-domain knowledge transfer [34]. The cross-domain gesture recognition system by Liu et al. achieves 97.61% cross-domain accuracy, yet this still represents a 2.11% degradation from in-domain performance, and the system was evaluated on the Widar3 dataset where only three environments are considered [27]. These results demonstrate that while significant progress has been made in cross-domain adaptation, Wi-Fi sensing systems still cannot achieve environment-invariant performance comparable to vision-based systems.

5.3 Hardware Dependence

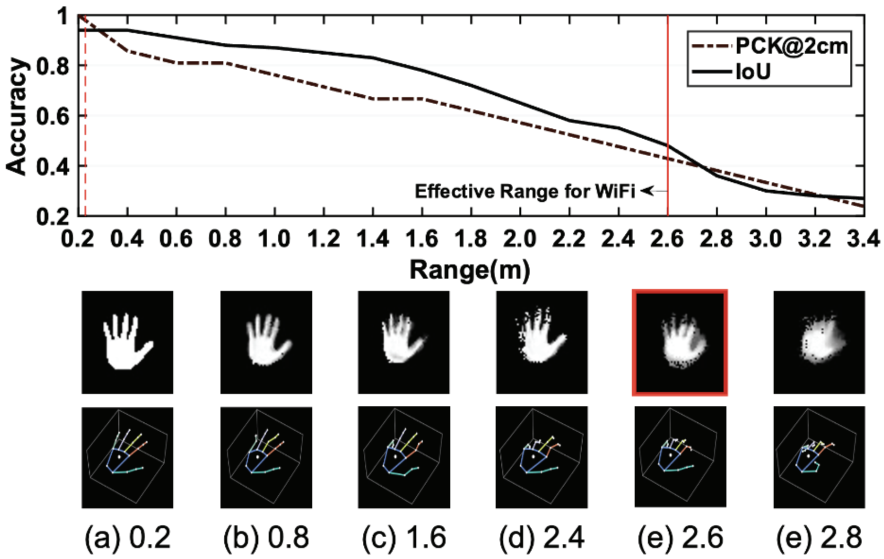
Despite claims of using “commodity Wi-Fi devices,” existing systems exhibit substantial hardware dependence that limits practical deployment flexibility. The majority of Wi-Fi sensing research relies on specific network interface cards that provide fine-grained CSI access through non-standard firmware modifications such as the 802.11n CSI Tool. Such NICs are no longer commercially available, making replication of prior results difficult and preventing deployment of modern Wi-Fi equipment.

Antenna configuration significantly affects sensing performance. GoPose and Winect require specific antenna arrays to achieve the 2D angle-of-arrival estimation needed for limb-level tracking [8,9]. Ren et al.’s mmWave gesture recognition system makes use of 802.11ad 60 GHz technology. While this technology delivers higher spatial resolution compared to traditional 2.4/5 GHz Wi-Fi, it is only supported by a limited number of devices [40]. The mmWave-based seated upper body pose estimation system uses dual 24 GHz FMCW radars placed in close proximity, which, while effective, represents specialized hardware rather than commodity Wi-Fi [42]. This limitation becomes critical in multi-user scenarios where consistent temporal sampling.

5.4 Complex Activity Recognition Difficulty

While coarse-grained activities (walking, sitting, falling) have been recognized with high accuracy, fine-grained and complex activity recognition remains challenging. The fundamental difficulty stems from the ambiguity between signal patterns: complex activities involving multiple body parts, subtle hand gestures, or interactions between multiple users produce overlapping and entangled signal reflections that are difficult to disentangle. Winect addresses this for free-form activity by separating entangled limb signals using 2D angle-of-arrival estimation, but it requires sophisticated antenna arrays and is limited in multi-user scenarios, as shown in Figure 8. HandFi tackles the challenge of palm-as-dominant-reflector by developing custom loss functions that capture low-level hand information from Wi-Fi signals, but it requires cross-modality supervision from depth cameras during offline training [11].

Figure 8.Quantitative results of HandFi at different sensing distances and the inferred hand mask at the corresponding sensing distances [11].



Multi-user activity recognition represents an additional layer of complexity. The WiMANS benchmark demonstrates that while Wi-Fi-based models perform well for multi-user identification and localization, activity recognition accuracy degrades substantially when multiple users perform simultaneous activities [5]. The Heterogeneous Sensor Data Fusion approach by Jing et al. highlights that single-modality sensing suffers from limited performance in complex scenarios, and even multi-sensor fusion struggles with high false alarm rates caused by environmental noise [43].

The lack of adequate training data compounds these difficulties. Wi-Fringe identifies that collecting sufficient training examples for all possible activities is prohibitively expensive, and proposes zero-shot learning using text semantics as a partial remedy, but the accuracy of unseen activity recognition remains limited [34].

5.5 Real-Time Performance Limitation

Real-time performance is essential for interactive applications such as human–computer interaction, fall alerting, and smart home control. While some systems achieve near real-time inference, most deep learning-based Wi-Fi sensing models incur substantial computational overhead. CSI-Net uses ResNet-based architectures for unified body characterization and pose recognition, which, while achieving high accuracy, requires GPU-level computational resources for practical inference. The Transformer-based approaches offer improved accuracy but introduce quadratic complexity with respect to sequence length, making real-time processing of long CSI sequences challenging [33,41].

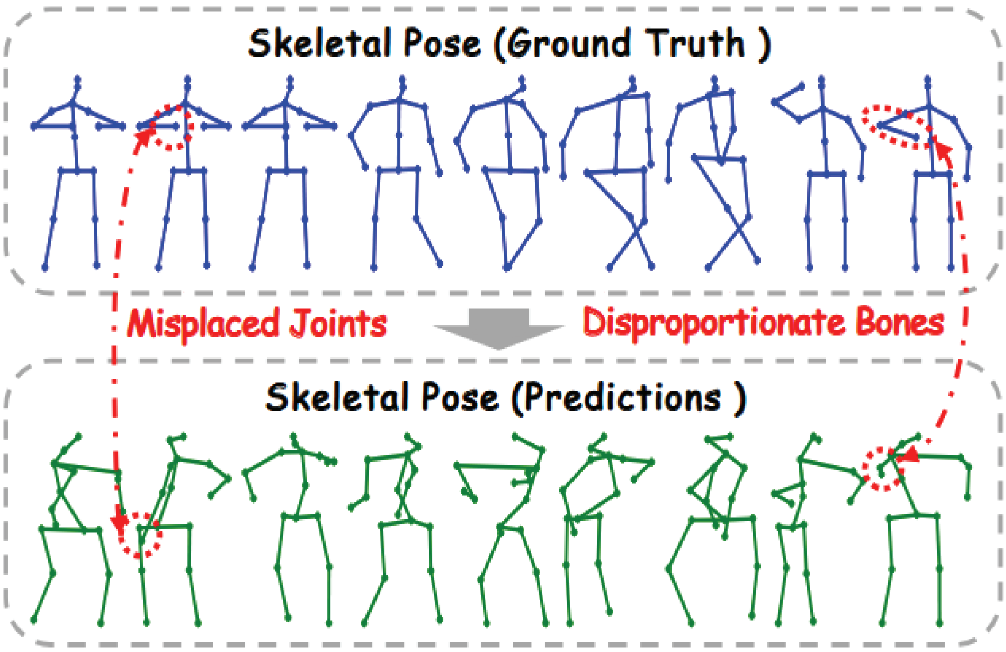
The Temporal U-Net framework explicitly introduces sample-level action recognition as a critical technique for real-time action recognition, noting that previous series-level approaches inherently cannot support real-time detection because they require the entire action sequence before classification [2]. The fundamental temporal constraint that the system must wait for the complete action to end before it can classify it, is unacceptable for applications requiring immediate response such as fall alerting.

5.6 Robustness Defects

Beyond environmental sensitivity and generalization, Wi-Fi sensing systems exhibit robustness defects across multiple dimensions. VDAN addresses the challenge of high-dimensional noise in CSI data and resource constraints on edge devices, noting that prevailing end-to-end models tightly couple feature extraction with classification, overlooking the inherent time-frequency sparsity of CSI and leading to noise amplification [39]. Under noisy conditions (5 dB SNR), the VDAN-preprocessed system maintains 89% accuracy, demonstrating that robustness to signal noise is achievable but requires explicit architectural design.

Pose estimation robustness is affected by the structural fidelity gap: decoded poses may violate human anatomical constraints, producing physically impossible joint configurations [36]. DT-Pose addresses this by incorporating human topology structure into Graph Convolution layers, but this requires prior knowledge of human skeletal topology and does not generalize to arbitrary articulated structures, as shown in Figure 9.

Figure 9. represents the predictions of the MetaFi++ method and corresponding ground truth [36].



6. Future Research Trends and Prospects

Addressing the challenges outlined in Section 5 requires coordinated advances across multiple dimensions. This section summarizes future optimization directions identified in the literature, corresponding to each challenge category.

6.1 Advanced Representation Learning and Topology-Aware Decoding

The next generation of Wi-Fi sensing algorithms must move beyond straightforward feature-to-label regression toward principled representation learning. DT-Pose suggests a broader future direction: self-supervised pre-training of Wi-Fi signal representations on large-scale unlabeled datasets, followed by task-specific fine-tuning. The incorporation of human topology structure into the decoding process ensures physically plausible output and represents a generalizable strategy for any articulated-body sensing task. Transformer-based architectures have shown promise in capturing long-range temporal dependencies in CSI sequences [33,41]. Future work should explore lightweight attention mechanisms, such as linear attention or sparse attention, to reduce the quadratic complexity that currently limits real-time applicability.

6.2 Domain Generalization and Latent Domain Discovery

Environmental interference sensitivity and cross-domain generalization are intimately related challenges that require advances in domain generalization methodology. AdaPose demonstrates that domain adaptation with mapping consistency loss is effective when target-domain labels are available, even in limited quantities [24]. However, fully unsupervised adaptation remains challenging. PerceptAlign introduces geometry-conditioned learning that explicitly encodes calibrated transceiver positions as conditional variables, forcing the network to disentangle human motion from deployment layouts [31]. The attention-based cross-domain gesture recognition system by Liu et al. demonstrates that fusing multi-angle Doppler spectra with multi-semantic spatial attention and self-attention-based channel mechanisms can extract domain-independent features [27]. Multi-receiver signal fusion combined with attention-based feature selection is a promising general strategy for anti-interference design.

6.3 Edge-Friendly Architectures for Resource-Constrained Devices

Practical Wi-Fi sensing systems must operate on edge devices with limited computational resources. VDAN [39]decouples feature refinement from classification and explicitly models the time-frequency sparsity of CSI through variational inference, achieving noise-robust feature selection with low computational overhead suitable for resource-constrained wireless sensing systems. Separating preprocessing from classification allows the classification head to be replaced or updated independently, facilitating adaptation to new tasks without retraining the entire pipeline. The social impact is significant: with falls ranking second among unintentional injury deaths in the elderly, and approximately 30% of fall victims being over age 65, reliable non-intrusive fall detection systems have the potential to save lives and enable independent living [44–46]. Smart home monitoring, human–computer interaction, and through-wall sensing represent additional domains where Wi-Fi sensing offers distinct advantages over cameras and wearable sensors. Future research should explore hardware-aware neural architecture search specifically for Wi-Fi sensing tasks, as well as model quantization and pruning techniques that preserve the fine-grained temporal sensitivity required for activity recognition while reducing parameter count and memory footprint.

6.4 Foundation Models and Zero-Shot Sensing

The goal of Wi-Fi sensing research is environment-invariant, user-independent activity recognition that can be deployed without retraining. Wi-Fringe takes a first step toward this goal by demonstrating zero-shot gesture recognition through cross-domain knowledge transfer from text semantics to RF signatures [34]. While the current approach is limited in scale, the fundamental insight opens a pathway to leveraging large language models for zero-shot Wi-Fi sensing. The development of Wi-Fi sensing foundation models represents a longer-term prospect. HandFi demonstrates the concept of a foundation model for Wi-Fi hand sensing: by constructing 3D hand skeletons rather than recognizing predefined gestures, it enables flexible downstream application development without model retraining [11]. Future multi-scene generalization research should pursue: (1) large-scale Wi-Fi sensing datasets with standardized collection protocols; (2) self-supervised pre-training on unlabeled Wi-Fi CSI data from diverse environments; (3) semantic alignment between Wi-Fi representations and natural language activity descriptions; and (4) federated learning frameworks that enable model improvement across distributed deployments without centralized data collection, preserving user privacy while enhancing generalization.

7. Conclusions

As a device-free sensing scheme with ubiquitous hardware support, Wi-Fi indoor human detection delivers unique advantages that neither vision nor wearable sensors can match. This review has systematically examined the current research landscape, covering signal acquisition methods, deep learning architectures, application scenarios, and the fundamental challenges that continue to shape the field.

Over the past decade, Wi-Fi-based activity detection has evolved from simple RSSI-based presence detection to sophisticated 3D pose estimation and fine-grained finger gesture recognition using CSI from commodity Wi-Fi devices. The transition from hand-crafted feature engineering to end-to-end deep learning has been the primary driver of performance improvements, with unified architectures such as CSI-Net demonstrating that a single learned representation can simultaneously serve multiple sensing tasks including biometrics estimation, person recognition, fall detection, and hand sign recognition.

The most mature application domain is elderly fall detection, where Wi-Fi sensing offers a unique combination all time continuous monitoring, privacy protection, and zero user burden that no existing sensing modality can simultaneously provide.

Despite these achievements, six fundamental challenges remain unresolved:

(1) Environmental interference sensitivity: Wi-Fi CSI is inherently sensitive to environmental changes that are indistinguishable from human-induced signal variations, requiring costly recalibration.

(2) Low generalization ability: Cross-domain and cross-layout performance degradation remains substantial, with models exhibiting coordinate overfitting to specific transceiver deployments.

(3) Hardware dependence: Most systems rely on discontinued NICs with non-standard firmware, specialized antenna arrays, or mmWave hardware that limits broad deployment.

(4) Complex activity recognition difficulty: Fine-grained gestures and multi-user scenarios remain challenging due to

overlapping signal reflections and insufficient training data.

(5) Real-time performance limitation: Deep learning models incur substantial computational overhead, and series-level recognition cannot support immediate-response applications.

(6) Robustness defects: High-dimensional noise, structural fidelity violations in decoded poses, and multi-user channel contention degrade reliability under practical conditions.

In summary, Wi-Fi-based indoor human activity detection has matured from a proof-of-concept technology to a practically viable sensing modality with demonstrated performance across diverse tasks and scenarios. The challenges identified in this review are not fundamental limitations of the modality itself, but rather engineering and algorithmic obstacles that are amenable to systematic research. With continued advances in deep learning architectures, domain generalization methodology, and hardware-standardized benchmarking, Wi-Fi sensing has the potential to become a ubiquitous component of next-generation smart environments, providing privacy-preserving, non-invasive, and continuous human activity monitoring without the cost, intrusiveness, or deployment constraints of existing solutions.

Funding

The project is funded by Nanyang Normal University Doctoral Special Project (No.2025ZX014) and Nanyang Science and Technology Plan Project (25KJGG041). Zengqian Song would like to thank Nanyang Normal University (No.SPCP2026419) for their support.

Conflict of Interests

The authors declare that there is no conflict of interest regarding the publication of this paper.

Reference

[1] F. Wang, J. Han, S. Zhang, X. He, D. Huang, CSI-Net: Unified Human Body Characterization and Pose Recognition, (2019). https://doi.org/10.48550/arXiv.1810.03064.

[2] F. Wang, Y. Song, J. Zhang, J. Han, D. Huang, Temporal Unet: Sample Level Human Action Recognition using WiFi, (2019). https://doi.org/10.48550/arXiv.1904.11953.

[3] H. Abdelnasser, M. Youssef, K.A. Harras, WiGest: a ubiquitous WiFi-based gesture recognition system, (2015). https://doi.org/10.48550/ARXIV.1501.04301.

[4] S. Yousefi, H. Narui, S. Dayal, S. Ermon, S. Valaee, A Survey of Human Activity Recognition Using WiFi CSI, (2017). https://doi.org/10.48550/arXiv.1708.07129.

[5] S. Huang, K. Li, D. You, Y. Chen, A. Lin, S. Liu, X. Li, J.A. McCann, WiMANS: A Benchmark Dataset for WiFi-based Multi-user Activity Sensing, (2024). https://doi.org/10.48550/arXiv.2402.09430.

[6] S. Ji, X. Zhang, Y. Zheng, M. Li, Construct 3D hand skeleton with commercial WiFi, in: Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems, ACM, Istanbul Turkiye, 2023: pp. 322–334. https://doi.org/10.1145/3625687.3625812.

[7] J.-L. Huang, Y.-S. Wang, Y.-P. Zou, K.-S. Wu, L.M. Ni, Ubiquitous WiFi and acoustic sensing: principles, technologies, and applications, J. Comput. Sci. Tech. 38 (2023) 25–63. https://doi.org/10.1007/s11390-023-3073-5.

[8] Y. Ren, Z. Wang, S. Tan, Y. Chen, J. Yang, Winect: 3D Human Pose Tracking for Free-form Activity Using Commodity WiFi, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 5 (2021) 1–29. https://doi.org/10.1145/3494973.

[9] Y. Ren, Z. Wang, Y. Wang, S. Tan, Y. Chen, J. Yang, GoPose: 3D Human Pose Estimation Using WiFi, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 6 (2022) 1–25. https://doi.org/10.1145/3534605.

[10] J. Geng, D. Huang, F.D. la Torre, DensePose From WiFi, (2022). https://doi.org/10.48550/arXiv.2301.00250.

[11] S. Ji, X. Zhang, Y. Zheng, M. Li, Construct 3D Hand Skeleton with Commercial WiFi, in: Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems, 2023: pp. 322–334. https://doi.org/10.1145/3625687.3625812.

[12] E. Barsoum, Articulated Hand Pose Estimation Review, (2016). https://doi.org/10.48550/arXiv.1604.06195.

[13] J. Huang, Z. Zhu, F. Guo, G. Huang, D. Du, The Devil is in the Details: Delving into Unbiased Data Processing for Human Pose Estimation, (2020). https://doi.org/10.48550/arXiv.1911.07524.

[14] L. Yuan, S. Chang, X. Nie, Z. Huang, Y. Zhou, Y. Chen, J. Feng, S. Yan, Towards Accurate Human Pose Estimation in Videos of Crowded Scenes, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020: pp. 4630–4634. https://doi.org/10.1145/3394171.3416299.

[15] Q. Fang, Q. Shuai, J. Dong, H. Bao, X. Zhou, Reconstructing 3D Human Pose by Watching Humans in the Mirror, (2021). https://doi.org/10.48550/arXiv.2104.00340.

[16] S. Phon-Amnuaisuk, K.T. Murata, L.-O. Kovavisaruch, T.-H. Lim, P. Pavarangkoon, T. Mizuhara, Visual-based Positioning and Pose Estimation, (2022). https://doi.org/10.48550/arXiv.2204.09232.

[17] J. Huang, Z. Zhu, F. Guo, G. Huang, The devil is in the details: delving into unbiased data processing for human pose estimation, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA, 2020: pp. 5699–5708. https://doi.org/10.1109/CVPR42600.2020.00574.

[18] E. Barsoum, Articulated hand pose estimation review, (2016). https://doi.org/10.48550/ARXIV.1604.06195.

[19] S. Chang, L. Yuan, X. Nie, Z. Huang, Y. Zhou, Y. Chen, J. Feng, S. Yan, Towards accurate human pose estimation in videos of crowded scenes, in: Proceedings of the 28th ACM International Conference on Multimedia, ACM, Seattle WA USA, 2020: pp. 4630–4634. https://doi.org/10.1145/3394171.3416299.

[20] O. Köpüklü, A. Gunduz, N. Kose, G. Rigoll, Real-time Hand Gesture Detection and Classification Using Convolutional Neural Networks, (2019). https://doi.org/10.48550/arXiv.1901.10323.

[21] J. Strohmayer, R. Sterzinger, C. Stippel, M. Kampel, Through-Wall Imaging based on WiFi Channel State Information, in: 2024 IEEE International Conference on Image Processing (ICIP), 2024: pp. 4000–4006. https://doi.org/10.1109/ICIP51287.2024.10647775.

[22] M. Bianchi, P. Salaris, A. Bicchi, Synergy-based Hand Pose Sensing: Reconstruction Enhancement, (2012). https://doi.org/10.48550/arXiv.1206.0555.

[23] M. Bianchi, P. Salaris, A. Bicchi, Synergy-Based Hand Pose Sensing: Optimal Glove Design, (2012). https://doi.org/10.48550/arXiv.1206.0556.

[24] Y. Zhou, J. Yang, H. Huang, L. Xie, AdaPose: Towards Cross-Site Device-Free Human Pose Estimation with Commodity WiFi, (2024). https://doi.org/10.48550/arXiv.2309.16964.

[25] F. Wang, S. Panev, Z. Dai, J. Han, D. Huang, Can WiFi Estimate Person Pose?, (2019). https://doi.org/10.48550/arXiv.1904.00277.

[26] R. Liu, C. Pan, J. Zeng, H. Ren, K. Wang, L. Kong, J. Wang, WiFi-based Cross-Domain Gesture Recognition Using Attention Mechanism, (2025). https://doi.org/10.48550/ARXIV.2512.04521.

[27] C. Wang, L. Tang, M. Zhou, Y. Ding, X. Zhuang, J. Wu, Indoor Human Fall Detection Algorithm Based on Wireless Sensing, Tsinghua Sci. Technol. 27 (2022) 1002–1015. https://doi.org/10.26599/TST.2022.9010011.

[28] X. Zhang, Z. Ye, J. Zhang, X. Tian, Z. Liang, S. Yu, VST-pose: a velocity-integrated spatiotem-poral attention network for human WiFi pose estimation, (2025). https://doi.org/10.48550/arXiv.2507.09672.

[29] X. Zhang, H. Yan, J. Huang, B. Liu, Y. Feng, J. Liu, M. Li, F. Zhang, Z. Liu, Beyond Physical Labels: Redefining Domains for Robust WiFi-based Gesture Recognition, (2026). https://doi.org/10.48550/ARXIV.2601.03825.

[30] S. Jia, Y. Lu, B. Liu, X. Zhang, P. Zhao, X. Tang, Y. Wei, J. Huang, H. Yan, Z. Liu, Breaking Coordinate Overfitting: Geometry-Aware WiFi Sensing for Cross-Layout 3D Pose Estimation, (2026). https://doi.org/10.48550/ARXIV.2601.12252.

[31] G. Panyin, Z. Guidong, Z. Zhigang, C. Xiao, D. Xuan, Research on Fall Detection System Based on Commercial Wi-Fi Devices, 21 (2023).

[32] J. Yang, X. Chen, D. Wang, H. Zou, C.X. Lu, S. Sun, L. Xie, SenseFi: A Library and Benchmark on Deep-Learning-Empowered WiFi Human Sensing, (2023). https://doi.org/10.48550/arXiv.2207.07859.

[33] M.T. Islam, S. Nirjon, Wi-Fringe: Leveraging Text Semantics in WiFi CSI-Based Device-Free Named Gesture Recognition, (2019). https://doi.org/10.48550/arXiv.1908.06803.

[34] D. Wang, J. Yang, W. Cui, L. Xie, S. Sun, AirFi: Empowering WiFi-based Passive Human Gesture Recognition to

Unseen Environment via Domain Generalization, (2022). https://doi.org/10.48550/arXiv.2209.10285.

[35] Y. Chen, J. Guo, Towards Robust and Realistic Human Pose Estimation via WiFi Signals, (2026). https://doi.org/10.48550/arXiv.2501.09411.

[36] F.-Q. Cui, Y.-T. Guo, T. Zheng, J. Huang, Generalizing WiFi Gesture Recognition via Large-Model-Aware Semantic Distillation and Alignment, (2025). https://doi.org/10.48550/ARXIV.2510.13390.

[37] Sheng Tan, Jie Yang, Fine-grained gesture recognition using WiFi, in: 2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), IEEE, San Francisco, CA, USA, 2016: pp. 257–258. https://doi.org/10.1109/INFCOMW.2016.7562082.

[38] N. Zhang, Variational dual-path attention network for CSI-based gesture recognition, (2026). https://doi.org/10.48550/arXiv.2601.13745.

[39] Y. Ren, J. Lu, A. Beletchi, Y. Huang, I. Karmanov, D. Fontijne, C. Patel, H. Xu, Hand gesture recognition using 802.11ad mmWave sensor in the mobile device, (2022). https://doi.org/10.48550/arXiv.2211.07090.

[40] X. Zhang, Z. Ye, J. Zhang, X. Tian, Z. Liang, S. Yu, VST-Pose: A Velocity-Integrated Spatiotem- poral Attention Network for Human WiFi Pose Estimation, (n.d.).

[41] B. Zhang, Z. Zhou, B. Jiang, R. Zheng, SUPER: Seated Upper Body Pose Estimation using mmWave Radars, in: 2024 IEEE/ACM Ninth International Conference on Internet-of-Things Design and Implementation (IoTDI), 2024: pp. 181–191. https://doi.org/10.1109/IoTDI61053.2024.00020.

[42] C. Jing, Y. Bowen, Y. Xiaoxia, Z. Hao, Research on fall detection algorithm based on heterogeneous sensor data fusion, 31 (2025) 1–13.

[43] C. Wang, L. Tang, M. Zhou, Y. Ding, X. Zhuang, J. Wu, Indoor human fall detection algorithm based on wireless sensing, Tsinghua Sci. Technol. 27 (2022) 1002–1015. https://doi.org/10.26599/TST.2022.9010011.

[44] G. Panyin, Z. Guidong, Z. Zhigang, C. Xiao, D. Xuan, Research on fall detection system based on commercial wi-fi devices, 21 (2023) 1–9.

[45] H. Jian, H. Chen, A portable fall detection and alerting system based on k-NN algorithm and remote medicine, China Commun. 12 (2015) 23–31. https://doi.org/10.1109/CC.2015.7114066.