

Article

Wi-CSNet: A Spatio-Temporal Model for CSI-Based Human Activity Recognition

Zhongjian Gao ^{1,2}, Ruige Zhang ^{1,2,3,*}, Yuwei Cai ^{1,2}, Lianhui Zheng ⁴, Han Yang ⁵ and Yao Li ^{1,2}

¹ College of Mechanical and Electrical Engineering, Sanming University, Sanming 365004, China

² Key Laboratory of Universities in Fujian Province for Intelligent Control of Equipment, Sanming 365004, China

³ Strait Institute of Technology, Sanming University, Sanming 365004, China

⁴ New Engineering Industry College, Putian University, Putian 351100, China

⁵ Practice Division of the Academic Affairs Office, Sanming University, Sanming 365004, China

* Correspondence: ruig_zhang@126.com; Tel.: +86-18005985939

Abstract

Human Activity Recognition (HAR) based on Channel State Information (CSI) has attracted considerable attention as a privacy-preserving sensing paradigm. However, CSI-based HAR faces several challenges, including environmental noise, long-range temporal dependencies, and the anisotropic structure of CSI tensors. To address these challenges, this paper presents Wi-CSNet, a lightweight CSI-oriented framework that integrates Discrete Wavelet Transform (DWT) preprocessing, asymmetric-stride convolutions, and a Cross-Scanning State Space Duality (CS-SSD) block derived from Mamba2. DWT preprocessing compresses temporal signals while preserving motion-related trends and reducing input dimensionality. Asymmetric-stride convolutions balance feature scales across heterogeneous CSI dimensions, while the lightweight CS-SSD module captures global dependencies with only a 0.63% parameter overhead. Extensive experiments demonstrate that Wi-CSNet achieves accuracies of 97.81% on HHI, 99.92% on UT-HAR, and 100% on NTU-HAR. These results confirm the effectiveness and robustness of Wi-CSNet for fine-grained CSI-based activity recognition in complex environments.

Keywords: HAR; Wi-Fi sensing; CSI; DWT; CS-SSD

1. Introduction

With the rapid advancement of the Internet of Things (IoT) and smart sensing technologies, HAR has been widely applied in motion tracking, elderly care, smart homes, and remote health monitoring systems [1–3]. Traditional HAR methodologies predominantly rely on vision-based sensors or wearable devices, both of which entail inherent limitations [4,5]. Vision-based systems are plagued by severe privacy concerns and suffer from performance degradation under low-illumination conditions or in the presence of occlusions. Wearable solutions require users to carry hardware devices continuously, which often leads to physical discomfort [5].

In recent years, Wi-Fi-based sensing has gained significant traction as a prominent research frontier, owing to its robust privacy preservation and minimal deployment overhead [3]. The application of Received Signal Strength Indicator (RSSI) to indoor positioning was first explored in 2000 [6]. This was followed by a pivotal milestone the following year: the official release of the CSI Tool [7]. While RSSI is easily accessible, its coarse-

Academic Editor: Gianluigi Ferrari

Received: 18 May 2026

Revised: 26 July 2026

Accepted: 5 August 2026

Published: 6 August 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

grained nature and inherent instability render it inadequate for increasingly sophisticated sensing requirements. In contrast, Channel State Information (CSI) provides a more fine-grained representation by capturing both amplitude and phase variations across individual subcarriers. The high sensitivity of CSI ensures that even minute perturbations, such as light taps, generate discriminative signatures in the signal profile [8].

Despite its advantages, the practical deployment of CSI in long-sequence HAR tasks faces several critical challenges. Wi-Fi signals are highly susceptible to both static obstacles and dynamic environmental factors [9]. After undergoing reflection, scattering, and diffraction, the signals received by the receiver form a complex superposition of multi-path components [9]. Consequently, raw CSI data encapsulates a mixture of environmental noise and user-specific behavioral signatures. Direct input of such raw data into recognition models inevitably increases computational overhead and compromises classification accuracy.

The primary challenges can be summarized as follows:

1. Effective denoising and compression: How to effectively suppress environmental noise and multi-path interference while preserving motion-relevant features?
2. Spatiotemporal Structural asymmetry in CSI data: CSI data typically exhibit a pronounced “long-strip” asymmetry, where the temporal dimension is significantly more extensive than the fixed subcarrier dimension. This “long-strip” structure demands algorithms capable of efficiently capturing both temporal dynamics and inter-subcarrier correlations.
3. Efficiency: The model needs to tackle the inherent trade-off among model parameter complexity, computational efficiency, recognition accuracy, and robustness (or generalization capability).

To address these challenges, this paper proposes a novel CSI-based HAR method named Wi-CSNet. The contributions of this paper can be summarized as follows:

1. DWT-based Feature Compression for CSI HAR: We incorporate a DWT-based pre-processing stage to suppress high-frequency noise and reduce sequence length before classification. This component is intended as a CSI-specific implementation choice rather than a new wavelet operator.
2. Asymmetric-Stride Convolution for Anisotropic CSI Tensors: We use temporally conservative downsampling in the early backbone layers, with temporal stride set to 1 and subcarrier stride set to 2, so that long temporal dependencies are not discarded prematurely.
3. Lightweight CS-SSD Integration: We augment a ResNet18 backbone with a Cross-Scanning State Space Duality module that scans the feature map along temporal and subcarrier directions and fuses the two paths with dynamic gating. This design adapts Mamba2-style state-space modeling to CSI tensors.

The remainder of this paper is organized as follows: Section 2 reviews the related literature. Section 3 introduces the preliminaries. Section 4 details the proposed Wi-CSNet framework. Section 5 presents the experimental results and analysis. Finally, Section 6 concludes the work and outlines future research directions.

2. Related Works

As established in the introduction, raw CSI measurements are inherently susceptible to stochastic noise and exhibit significant redundancy due to environmental interference, such as multi-path propagation and signal attenuation [10]. The complexity and unpredictability of the wireless environment significantly impact the accuracy and flexibility of activity recognition [3]. In response to these challenges, extensive research has focused on CSI preprocessing, feature engineering, and architectural optimization.

During the preprocessing stage, noise suppression is achieved through various smoothing strategies. Common techniques include moving-average filtering [11], Hampel filtering for outlier elimination [12], Savitzky–Golay filtering [5], and Butterworth low-pass filtering [13]. To alleviate the high-dimensional redundancy inherent in CSI data, dimensionality reduction techniques such as Principal Component Analysis (PCA) [5] and Linear Discriminant Analysis (LDA) [14] are widely adopted to project raw signals onto a lower-dimensional manifold.

Beyond time-domain filtering, a highly effective strategy for isolating motion-induced fluctuations from environmental clutter involves domain transformation. By transitioning signals into the frequency or time-frequency domains, researchers can more effectively extract periodic and transient motion characteristics. Representative techniques in this category include the Fast Fourier Transform (FFT) [15] for spectral analysis and the Short-Time Fourier Transform (STFT) [16] for capturing non-stationary temporal dynamics.

Despite their utility, these conventional methodologies exhibit inherent technical constraints. Time-domain filtering is often restricted to attenuating predefined noise profiles. In terms of dimensionality reduction, PCA and LDA rely exclusively on linear projections, thereby failing to capture the complex nonlinear dependencies intrinsic to CSI samples. Meanwhile, FFT and STFT rely on fixed basis functions and exhibit limited adaptability to non-stationary CSI signals. These methods inevitably encounter a fundamental trade-off between temporal and spectral resolution.

To overcome these limitations, DWT has emerged as a superior alternative. As reported in [16] DWT provides a balanced time-frequency representation while enabling data compression without significant information loss. Thus, this work incorporates DWT into the CSI preprocessing pipeline to achieve robust noise suppression and efficient data dimensionality reduction.

Regarding feature extraction and model architecture, existing Wi-Fi CSI-based methods are broadly categorized into Machine Learning (ML) and Deep Learning (DL) frameworks [17]. Conventional ML-based approaches typically employ Support Vector Machines (SVM) [18], k-Nearest Neighbors (KNN) [19] and Hidden Markov Models (HMM) [20]. These methods rely heavily on manual feature engineering (e.g., statistical and spectral features) and require researchers to possess extensive feature extraction experience [21]. Moreover, these approaches often exhibit suboptimal generalization when confronted with complex gestures or diverse environmental conditions.

In contrast, DL methods have gradually become the mainstream owing to their outstanding performance in automatic feature extraction [3]. CNNs originally conceived for computer vision, have been successfully adapted for Wi-Fi sensing. For instance, SignFi [22] utilized CNNs to recognize 276 sign language categories, while E2EDLF [23] employed a three-layer CNN architecture to extract behavioral features directly from CSI images. To further enhance robustness, HHI-AttentionNet [24] integrated Depthwise Separable Convolutions (DS-Conv) with an Antenna-Frame-Subcarrier Attention Module (AFSAM), enabling the model to adaptively weigh feature importance across multiple physical dimensions.

However, serialized Wi-Fi CSI measurements for HAR typically exhibit significant long-term temporal dependencies. The standard CNN models are inherently constrained by their strict local receptive fields. To mitigate this limitation, bidirectional long short-term memory (BLSTM) networks [25] have been introduced into the HAR domain, enabling the simultaneous modeling of both historical and future contextual information. Similarly, H2HI-Net [26] utilizes temporal convolution network (TCN) and bidirectional gated recurrent units (Bi-GRUs) to automatically extract temporal features of activities from HAR data. Nevertheless, sequence-to-sequence (Seq2Seq) models still struggle to fully capture the correlation features between CSI subcarriers [22].

To harmonize spatial and temporal modeling, researchers have explored hybrid architectures. For instance, Reference [21] coupled convolutional layers with LSTMs to develop a CNN-ABLSTM framework, achieving a benchmark accuracy of 97.42% on the HHI dataset. Furthermore, given the unparalleled capacity of Self-Attention mechanisms in capturing long-range dependencies, the THAT framework [27] utilizes a two-stream architecture that integrates convolutions with Transformers. This approach facilitates the concurrent extraction of multi-scale temporal dynamics and subcarrier-channel correlations. Despite their competitive performance, recurrent networks suffer from vanishing gradients and inherent sequential bottlenecks over long packet sequences, while Transformers scale with a prohibitive quadratic complexity that strictly restricts the model's lightweight deployment.

In recent years, the emergence of Mamba has pioneered novel trajectories for long-sequence processing. Mamba2 [28] introduces the Selective Space Duality (SSD) mechanism, which facilitates sophisticated long-sequence modeling by dynamically distilling relevant state information. Notably, this architecture achieves superior computational efficiency compared to Transformer-based models [28]. Building upon these advancements, this paper proposes a hybrid framework that synergizes the SSD mechanism with spatial feature extraction networks (e.g., CNNs). By concurrently capturing the inter-subcarrier spatial correlations and long-range temporal dependencies, our approach strikes an optimal balance among model complexity, efficiency, accuracy, and robustness.

3. Preliminaries

3.1. Channel State Information

In 2011, Halperin et al. [7] released the CSI Toolkit, which modifies the firmware of Intel 5300 network interface card (NIC) drivers to enable commodity Wi-Fi devices to sample fine-grained CSI. CSI characterizes the channel properties of a communication link. The relationship between transmitted and received signals in the frequency domain can be described by Equation (1):

$$Y_k = H_k \times X_k + n_k \quad (1)$$

In the equation, Y_k denotes the received signal in the frequency domain, X_k is the transmitted signal, H_k represents the channel matrix gain; and n_k accounts for the environmental noise. When configured for a 20 MHz channel bandwidth, the CSI Toolkit captures 30 out of the 56 subcarriers, yielding two-dimensional CSI. These 30 subcarriers can be expressed as follows:

$$H = [H_1, H_2, \dots, H_k, \dots, H_N] \quad (2)$$

where N denotes the number of subcarriers. Each subcarrier in CSI contains amplitude and phase relationships, which can be represented as follows:

$$H_k = \|H_k\| e^{j\angle H_k} \quad (3)$$

where $\|H_k\|$ and $\angle H_k$ are the amplitude and phase of the k -th subcarrier, respectively.

3.2. Data Processing

Raw CSI measurements are captured as complex-valued matrices, containing both amplitude and phase components. Theoretically, the CSI phase is considered more robust as it depends primarily on the initial phase and the propagation path length, making it a promising candidate for passive sensing. However, in practical deployment, raw phase data is highly unstable due to hardware imperfections, including sampling frequency offsets (SFO), carrier frequency offsets (CFO), and non-linear phase errors. Given the high computational complexity and inconsistent performance of existing phase calibration

methods, our framework prioritizes the utilization of CSI amplitude information for behavior recognition.

To suppress high-frequency environmental noise while reducing temporal redundancy, a multi-level DWT is applied to the raw CSI streams. Figure 1 illustrates the comparative evolution of the CSI amplitude profile before and after DWT processing, highlighting the clarified motion signatures.

In the Human-to-Human Interaction (HHI) dataset [29], the original recordings vary between 1600 and 2100 packets. To ensure input consistency, we excluded steady-state intervals and uniformly truncated all sequences to a fixed length of 1536 packets. As shown in Figure 1, the raw CSI data exhibits significant noise (Row 1). A moving-average smoothing over adjacent points is first applied (Row 2), halving the temporal dimension to a baseline length of $L_{in} = 768$. To further eliminate high-frequency noise, a 2-level Discrete Wavelet Transform (DWT) with a Daubechies 4 (db4, filter length $N = 8$) basis and periodic padding is implemented. The step-by-step dimension reduction is strictly governed by the dyadic downsampling formula:

$$L_{out} = \frac{L_{in} + N - 1}{2} \quad (4)$$

Consequently, the Level-1 decomposition compresses the 768 packets into $L_{out1} = 387$ points (Row 3), and the Level-2 decomposition further decimates these into exactly $L_{out2} = 197$ packets (Row 5). This mathematical convergence produces an optimized, noise-free temporal profile of 197 frames, matching the input requirement of our subsequent infrastructure.

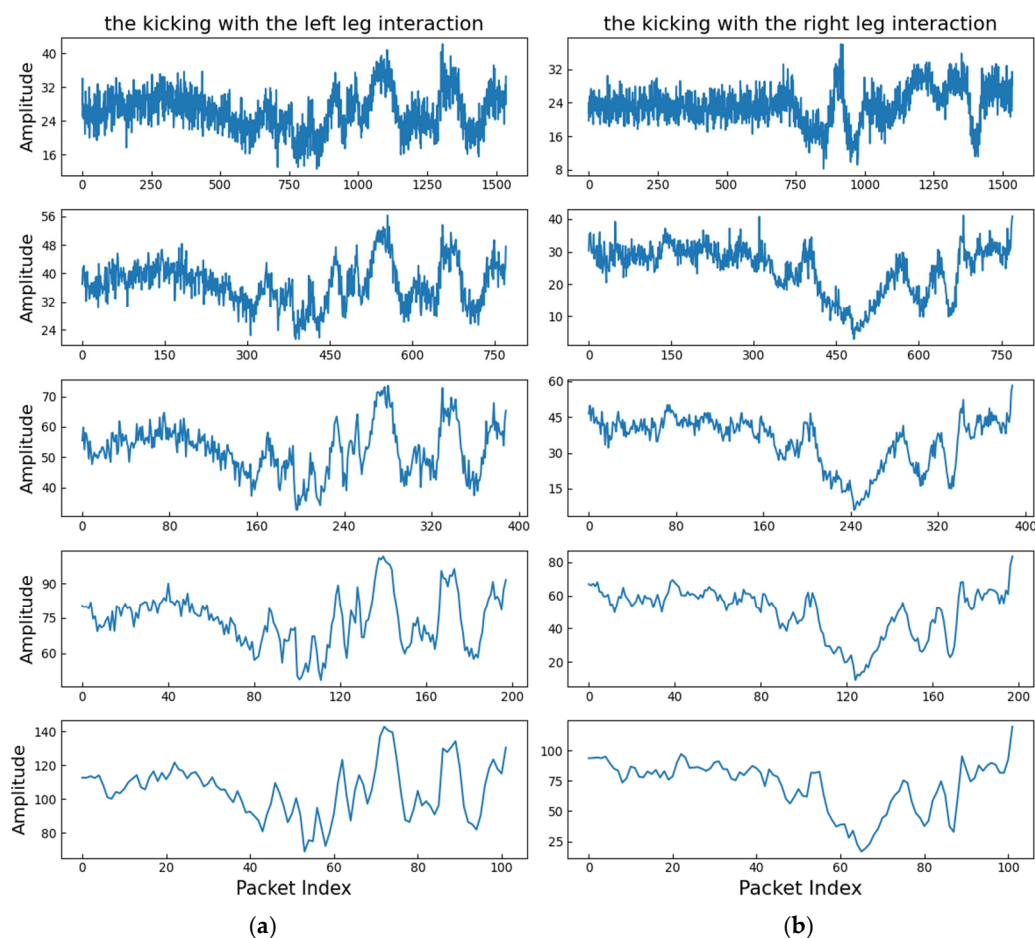


Figure 1. (a) CSI amplitude data for ‘the kicking with the left leg’ interaction; (b) CSI amplitude data for ‘the kicking with the right leg’ interaction. The first row displays the raw CSI data, while the second row presents the results smoothed by averaging adjacent points in the time series. Rows 3–

5 illustrate the outcomes obtained after applying first- to second-order DWT decompositions, respectively.

3.3. Mamba

Mamba is an emerging architecture based on Structured State-Space Models (SSMs), which aims to efficiently capture complex dependencies in sequential data [30]. The classical SSM maps a 1D input sequence $x(t)$ to an output $y(t)$ through a latent state $h(t)$:

$$h'(t) = Ah(t) + Bx(t) \quad (5)$$

$$y(t) = Ch(t) \quad (6)$$

where A is the state transition matrix, B and C are the input matrix and output matrix respectively, and $h'(t)$ denotes the rate of change in the state vector $h(t)$ with respect to time. To process digital signals, the system is discretized using a step size Δ . The specific formulas for discretizing the continuous parameters A and B (using the Zero-Order Hold method) are as follows:

$$\bar{A} = \exp(\Delta A) \quad (7)$$

$$\bar{B} = (\Delta A)^{-1}(e^{(\Delta \times A)} - I)\Delta B \quad (8)$$

To accommodate diverse input sequences, Mamba introduces a selection mechanism known as the Selective State Space Model (SSSM, or S6). In the S6 formulation, the model parameters are no longer fixed. Instead, the parameters B , C , and Δ are generated dynamically through learned projections of the input. This context-awareness allows the model to selectively propagate or suppress information along the temporal axis. According to [28], the SSSM can be regarded as a special form of linear attention under specific constraints.

Building upon this, the State Space Duality (SSD) framework proposed in Mamba2 [28] establishes a theoretical bridge between selective SSMs and attention mechanisms. By reformulating the recurrence of SSMs into a structured matrix multiplication form, SSD enables our Wi-CSNet to execute temporal modeling of CSI signals. This integration harnesses the massive parallelization capabilities typically associated with Transformers while preserving the linear scalability and low-latency inference characteristic of State Space Models. Consequently, Wi-CSNet achieves a superior balance between computational throughput and the ability to capture long-range behavioral dependencies in complex wireless environments.

4. Methodology

The comprehensive architecture of the proposed Wi-CSNet is illustrated in Figure 2. To effectively process the characteristic “long-strip” structure of CSI data, the framework adopts a modified ResNet18 as the foundational backbone, augmented with localized refinement and global context modeling.

A critical refinement in our backbone design is the implementation of asymmetric convolutional strides. Specifically, the initial Conv2D layer and Layer 2 are reconfigured with a stride of (1, 2). Unlike standard symmetric downsampling, this design ensures that the temporal resolution remains intact during the early feature extraction stages. This prevents the loss of discriminative temporal dynamics, thereby facilitating the identification of micro-gestures and subtle behavioral variations within the CSI sequences. Inspired by the SimAM [31], we integrate a parameter-free attention module into Layer 2. This enables the network to infer 3D attention weights based on neuron energy functions, enhancing feature discriminability without introducing additional parametric overhead.

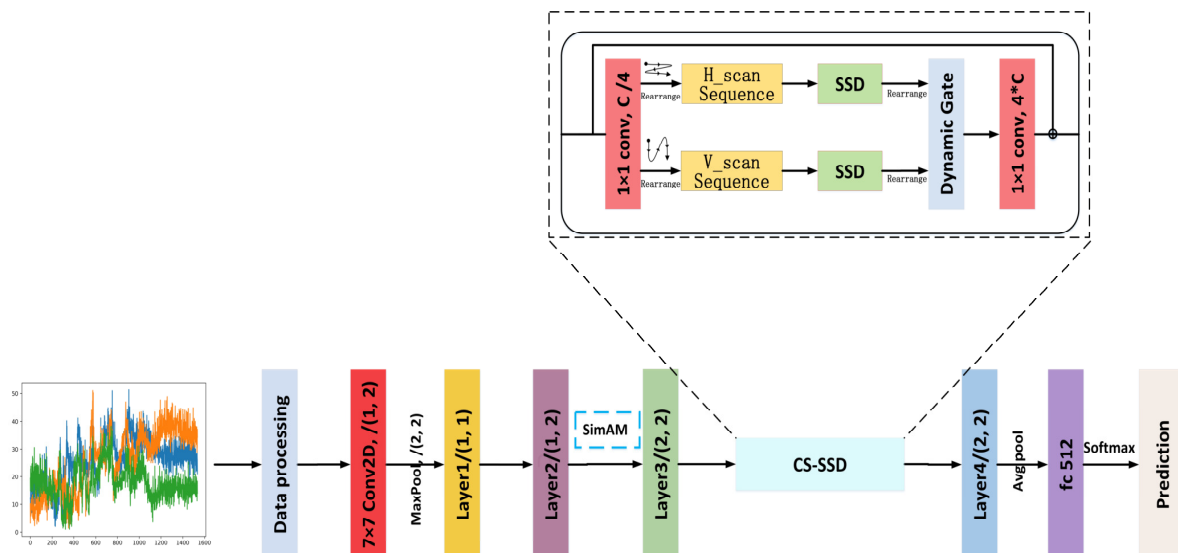


Figure 2. Overall architecture of the Wi-CSNet model.

Following Layer 3, we introduce the CS-SSD (Cross-Scanning State Space Duality) module to bridge the gap between local feature extraction and global dependency modeling. The module first utilizes a 1×1 convolution layer for channel-wise bottlenecking (reduction ratio $r = 4$). This bottleneck design significantly lowers the computational complexity of the subsequent state-space operations while preserving essential feature abstractions.

Inspired by the Cross-Scan Module (CSM) proposed in VMamba [32], we implement a dual-path scanning strategy to fully exploit the spatiotemporal correlations in the “long strip” CSI data. The compressed feature maps are bidirectionally unfolded along the subcarrier axis and temporal axes, then parsed through the State-Space Duality (SSD) operator.

As illustrated in the following workflow, the feature maps are first decomposed into two orthogonal trajectories (horizontal for temporal dynamics and vertical for subcarrier relationships) and converted into independent 1D sequences, respectively. These sequences are subsequently processed by the SSD operator, which facilitates long-range context aggregation while maintaining linear computational complexity.

Rather than employing static integration, the multi-directional features extracted from the dual SSD paths are synthesized via a Dynamic Gated (DG) fusion mechanism. After the State-Space Duality (SSD) operators process the sequences, the resulting 1D latent representations of the horizontal and vertical paths are first rearranged back into their original 2D spatial dimensions. This structural restoration is crucial, as it allows the model to leverage spatial context for weight estimation.

The DG unit then employs a Global Average Pooling (GAP) layer across these reconstructed feature maps to distill the global channel-wise statistics. This is followed by a 1×1 convolution layer configured with two output channels, which effectively models the inter-path dependencies and captures the relative significance of each scanning trajectory. A Softmax-activated bottleneck is then applied to these channels to generate a pair of normalized importance weights (ω_h, ω_v), representing the learned confidence scores for the horizontal (temporal $\in \mathcal{R}^{B \times (H \times W) \times C}$) and vertical (subcarrier $\in \mathcal{R}^{B \times (W \times H) \times C}$) trajectories, respectively. The fused feature map is calculated as:

$$X_{fused} = \omega_h \times X_h + \omega_v \times X_v \quad (9)$$

where ω_h and ω_v are the respective dynamic weights for the horizontal and vertical paths, and $X_h, X_v \in \mathcal{R}^{B \times C \times H \times W}$ represent the corresponding reconstructed feature maps.

As the final stage of the CS-SSD module, the fused features are projected back to the original channel dimension via an output projection layer. A residual connection is maintained to facilitate stable gradient flow, ensuring that the global context distilled by the SSD is effectively merged with the hierarchical mid-level features from the ResNet backbone. The implementation details of Wi-CSNet on the HHI dataset are shown in Table 1.

Table 1. Implementation details of Wi-CSNet for the HHI dataset (Labels = 12).

Layer Name	Output Dimension	Stride	Number of Parameters
Input layer	(None, 3, 197, 60)	\	0
Conv2D + BN	(None, 64, 197, 30)	(1, 2)	9536
Maxpooling 2D	(None, 64, 99, 15)	(2, 2)	0
Layer1	(None, 64, 99, 15)	(1, 1)	147,968
Layer2	(None, 128, 99, 8)	(1, 2)	525,568
SimAM	(None, 128, 99, 8)	\	0
Layer3	(None, 256, 50, 4)	(2, 2)	2,099,712
CS-SSD	(None, 256, 50, 4)	\	71,172
Layer4	(None, 512, 25, 2)	(2, 2)	8,393,728
AveragePooling 2D	(None, 512, 1, 1)	\	0
Dense Layer	(512, 12)	\	6156

As evidenced by Table 1, the horizontal and vertical strides in the Conv2D and Layer 2 stages are intentionally non-identical, formalizing our asymmetric stride CNN architecture. This configuration prioritizes temporal resolution by maintaining a unit stride ($H = 1$) in the early layers. Wi-CSNet has a total of approximately 11.25 M parameters. The proposed CS-SSD module adds only 71,172 parameters, a marginal increase of 0.63% over the ResNet-18 backbone. This confirms that the CS-SSD mechanism enables efficient global dependency modeling with a much lighter footprint than Transformers.

5. Experimental Results

All experiments were conducted on a personal computer equipped with an Intel® Core i9-13900K processor (Intel Corporation, Santa Clara, CA, USA; 3.0 GHz) and an NVIDIA GeForce RTX™ 4090 GPU with 24 GB VRAM (NVIDIA Corporation, Santa Clara, CA, USA). The proposed Wi-CSNet was implemented using the PyTorch 2.11 framework. The network was optimized using Stochastic Gradient Descent with Momentum (SGDM), with an initial learning rate of 0.01 and a momentum coefficient of 0.9. To enhance the model's generalization capabilities while preserving the integrity of the temporal sequences, we only used RandomErasing as the data augmentation technique. This approach effectively simulates signal occlusions without disrupting the crucial chronological order of the CSI packets. This experiment consistently involves three datasets, namely HHI, UT-HAR [15], and NTU-HAR [33]. The hyperparameter configurations used for the three datasets are summarized in Table 2.

Table 2. Hyperparameter configurations across the three datasets.

Hyperparameter	HHI	UT-HAR	NTU-HAR
Input Tensor	(32, 3, 197, 60)	(32, 3, 253, 30)	(32, 3, 255, 114)
Output Label	12	7	6
Split protocol	10-fold cross-validation	5-fold cross-validation	5-fold cross-validation
Epoch	150	150	150
Optimized	SGDM	SGDM	SGDM
Batch size	32	32	32

The input tensor dimensions and output label categories of the three datasets are summarized in Table 2. To ensure consistent and reproducible evaluation across all experiments, we uniformly adopted the StratifiedKFold strategy with shuffle = True and random_state = 42, thereby preserving class balance in each data split while guaranteeing reproducibility. For a fair and direct comparison with previous studies, 5-fold cross-validation was used for UT-HAR and NTU-HAR, while a 10-fold split was maintained for HHI.

5.1. Experimental Results on the HHI Dataset

The human-to-human interaction (HHI) dataset [29] contains recordings of 40 pairs of different subjects performing 12 distinct interaction behaviors in an indoor environment. Each pair of subjects conducted 10 trials for each interaction behavior, resulting in a total of 4800 trials recorded across all 40 pairs (i.e., 40 pairs $\times$ 12 interaction behaviors $\times$ 10 trials). Table 3 presents the corresponding action statistics for the 12 labels in the HHI dataset.

Table 3. The corresponding action statistics for the 12 labels in the HHI dataset.

Label	Interaction	Label	Interaction
I1	Approaching	I7	Kicking with the right leg
I2	Departing	I8	Pointing with the left hand
I3	Handshaking	I9	Pointing with the right hand
I4	High five	I10	Punching with the left hand
I5	Hugging	I11	Punching with the right hand
I6	Kicking with the left leg	I12	Pushing

Following the methodology detailed in Section 3.2, the raw CSI data underwent a systematic two-stage refinement process. First, all sequences were uniformly truncated to a fixed length of 1536 packets. This step eliminated inconsistent steady-state intervals, ensuring that the model focuses exclusively on the active interaction phases. Subsequently, we applied DWT to these truncated signals. Table 4 summarizes the comparative performance of the baseline ResNet18, the Asymmetric Stride ResNet18 (AS-ResNet18), and our proposed Wi-CSNet across various input dimensions.

Table 4. Performance comparison of ResNet18, AS-ResNet18, and Wi-CSNet on the HHI dataset.

Models	Input Dimension	Train Time(s)	ACC(%)	FLOPs (G)	Rank
ResNET18	(32, 3, 256, 180)	1141	96.79	/	7
ResNET18	(32, 3, 387, 60)	398	97.27	/	5
ResNET18	(32, 3, 197, 60)	319	96.25	0.47	10
ResNET18	(32, 1, 197, 180)	423	96.29	/	9
AS-ResNET18	(32, 3, 387, 60)	629	96.55	/	8
AS-ResNET18	(32, 3, 197, 60)	408	97.39	1.529	4
AS-ResNET18	(32, 1, 197, 180)	783	97.58	/	2
Wi-CSNet	(32, 3, 387, 60)	796	96.81	/	6
Wi-CSNet	(32, 1, 197, 180)	1257	97.45	/	3
Wi-CSNet	(32, 3, 197, 60)	449	97.81	1.564	1

As shown in Table 4, Wi-CSNet achieved the highest classification accuracy of 97.81% with an input size of (32, 3, 197, 60), ranking first among all evaluated models. Compared with ResNet18 and AS-ResNet18 using the same input size, Wi-CSNet improved accuracy by 1.56 and 0.42 percentage points, respectively, demonstrating its stronger feature extraction and classification capabilities. Although Wi-CSNet required a training time of 449

s, slightly longer than ResNet18 (319 s) and AS-ResNet18 (408 s), and incurred 1564 M FLOPs, its computational cost was only marginally higher than that of AS-ResNet18 (1529 M) while achieving superior recognition accuracy. Overall, this result underscores that the CS-SSD module achieves a superior balance between model complexity and classification performance. The Performance comparisons of three models are shown in Figure 3.

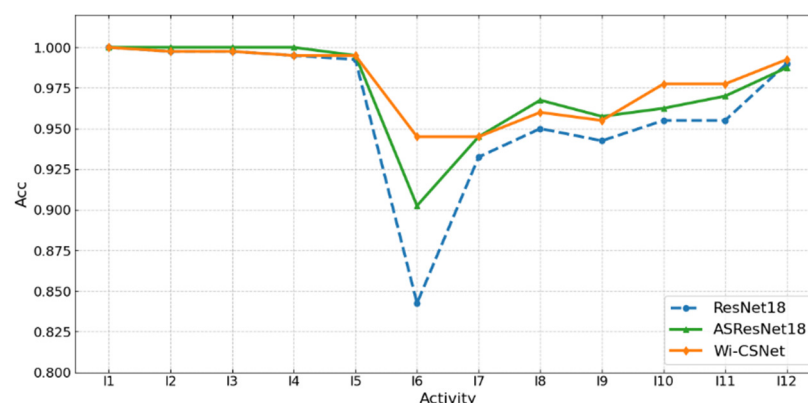


Figure 3. Performance comparisons of three models using 12 labels.

As illustrated in Figure 3, while all three models demonstrate robust discriminative power across most categories, a noticeable performance bottleneck persists within the I6–I11 cluster. Specifically, activities I6 (Kicking with left leg) and I7 (Kicking with right leg) exhibit the lowest recognition accuracies.

This performance decline is primarily attributed to the inherent ambiguity in mirror-interactive actions. The pairs—Kicking (I6/I7), Pointing (I8/I9), and Punching (I10/I11)—differ only in the lateral direction of the limbs. Because CSI measurements are sensitive to the overall displacement of the human body mass, the subtle phase and amplitude variations between left and right limb movements result in highly similar signal envelopes (as shown in Figure 1), leading to inter-class confusion.

Furthermore, Figure 3 reveals that lower-limb interactions (Legs) exhibit significantly higher misclassification rates compared to upper-limb interactions (Hands). This discrepancy is likely rooted in environmental physics: lower-limb movements are more susceptible to ground multipath reflections. To evaluate the competitiveness of the proposed Wi-CSNet, we compared its performance against several state-of-the-art CSI-based activity recognition frameworks on the HHI dataset. The comparative results are summarized in Table 5.

Table 5. Performance comparison with state-of-the-art models on the HHI dataset.

Method	Input Dimension	ACC (%)
E2EDLF [23]	(None, 180, 256)	86.3
CSI-IANet [34]	(None, 1, 180, 256)	91.3
HHI-AttentionNet [24]	(None, 3, 512, 30)	95.47
CSI-IRNet [35]	\	96.18
H2HI-NET [26]	(None, 180, 256)	96.39
CNN-ABLSTM [21]	(None, 500, 90)	97.42
Wi-CSNet	(None, 3, 197, 60)	97.81

The results demonstrate that Wi-CSNet outperforms existing models, achieving a peak accuracy of 97.81%. Notably, unlike recurrent models (H2HI-Net) that suffer from vanishing gradients in long sequences, or attention-based models (CNN-ABLSTM) with

quadratic computational costs, Wi-CSNet leverages the linear scalability of the SSD module. This allows it to capture global dependencies more efficiently, providing a superior trade-off between recognition precision and inference throughput.

5.2. Multi-Dataset Experiments

To verify the performance of the Wi-CSNet framework on different datasets, we extended our experimentation to the UT-HAR and NTU-HAR datasets. The UT-HAR dataset involves seven types of daily activities (e.g., Bed, Fall, Pickup, Run) performed by six subjects. Following the standardized preprocessing pipeline established in the repository (ermongroup/Wifi_Activity_Recognition), we obtained 4973 valid samples. Each sample initially consisted of 500 time steps across 90 subcarrier components, consistent with the parameters reported in the SenseFi benchmark [33].

Similarly, the NTU-HAR dataset was obtained directly from the Google Drive link provided in the SenseFi [33] literature. The dataset encompasses six distinct activity categories: box, circle, clean, fall, run, and walk. A total of 1200 CSI samples were collected. The raw data structure is multi-dimensional, comprising 3 antennas, 114 subcarriers, and 2000 time steps.

By applying the unified preprocessing pipeline detailed in Section 3.2, we transformed the raw data into optimized input tensors. For UT-HAR, the resulting tensor shape was (3, 253, 30), while for NTU-HAR, it was (3, 255, 114). The recognition performance of Wi-CSNet on these two datasets is illustrated in Figure 4.

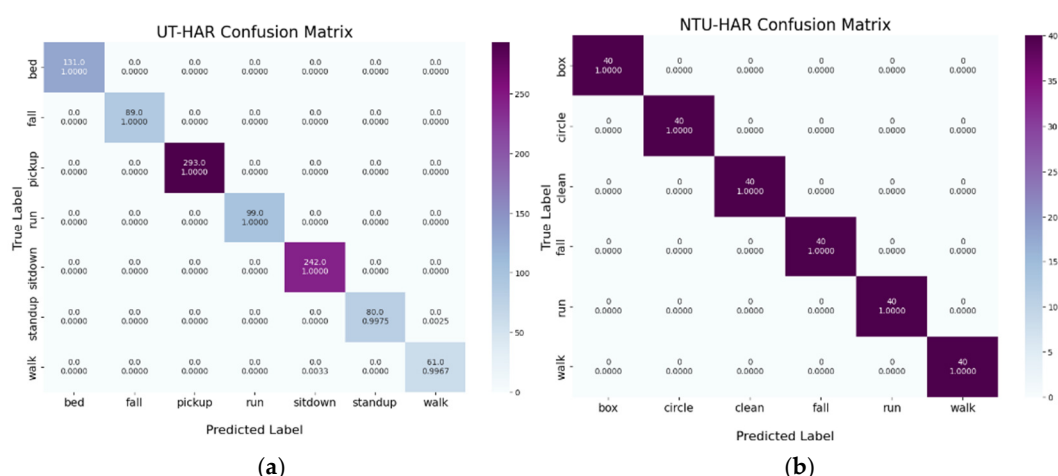


Figure 4. Recognition Performance of Wi-CSNet on UT-HAR and NTU-HAR Datasets. (a) UT-HAR Confusion Matrix, (b) NTU-HAR Confusion Matrix.

We employed StratifiedKFold cross-validation to ensure the statistical significance of our performance metrics. This approach preserves the underlying class distribution, ensuring that the category proportions within each fold remain consistent with the overall dataset. As illustrated in Figure 4a, the UT-HAR dataset exhibits a certain degree of class imbalance. Our analysis reveals a positive correlation between sample size and recognition accuracy. For instance, the minority classes ‘walk’ and ‘stand up’ achieve slightly lower accuracy compared to the well-represented majority classes. In contrast, as shown in Figure 4b, Wi-CSNet achieves near-perfect discrimination on the NTU-HAR dataset, attaining 100% recognition accuracy across all activity categories. The results comparison of existing approaches on the two datasets is presented in Table 6.

Table 6. Results comparison of existing approaches on UT-HAR and NTU-HAR datasets.

Approaches	UT-HAR	NTU-HAR
------------	--------	---------

	ACC (%)	Param (M)	ACC (%)	Param (M)
ViT [36]	95.0	10.58	92.8	10.58
THAT [27]	98.3	49.34	99.2	49.34
RGANet [37]	99.4	0.44	99.24	0.44
ERPANet [38]	99.4	23.55	99.6	23.55
Wi-CSNet	99.9	11.25	100	11.25

As evidenced in Table 6, Wi-CSNet consistently outperforms existing methods across both benchmarks, setting a new state-of-the-art for CSI-based recognition. While RGANet [37] utilizes a smaller parameter footprint, it incurs a significant training overhead. Specifically, reference [37] reports a training time of 18,890 s for RGANet, which is substantially higher than the 5172 s required by ResNet-18 under the same experimental setting. In contrast, our model maintains a training efficiency highly comparable to ResNet18, as shown in Table 4.

5.3. Ablation Experiments

To quantify the individual contributions of the proposed components to the overall system performance, we conducted a systematic “step-down” ablation study across the three benchmark datasets. Starting from the full Wi-CSNet architecture, we sequentially removed the CS-SSD module, the SimAM module, and the Asymmetric Strides (AS) configuration until the model reverted to a standard ResNet18 baseline. The quantitative results of this decomposition are summarized in Table 7.

Table 7. Ablation experiment results across three public datasets.

Dataset	HHI		UT-HAR		NTU-HAR	
Models	Data Size	ACC (%)	Data Size	ACC (%)	Data Size	ACC (%)
Wi-CSNet	(3, 197, 60)	97.81	(3, 253, 30)	99.92	(3, 255, 114)	100
- CS-SSD	(3, 197, 60)	97.47	(3, 253, 30)	99.89	(3, 255, 114)	100
- SimAM	(3, 197, 60)	97.39	(3, 253, 30)	99.84	(3, 255, 114)	100
- AS	(3, 197, 60)	96.25	(3, 253, 30)	99.87	(3, 255, 114)	100

On the HHI dataset, the introduction of the CS-SSD module resulted in a significant performance leap (+0.34%). While standard convolutions are spatially constrained to local receptive fields, the CS-SSD block facilitates long-range dependency modeling across the dual temporal and subcarrier axes. The integration of the SimAM module after Layer 2 provided a further (0.08%) improvement. Without extra parameters, SimAM computes 3D attention weights via neuron energy functions and boosts feature discriminability. The most substantial improvement (1.14%) was observed when replacing uniform strides with Asymmetric Strides. This confirms our primary hypothesis: for “long-strip” CSI signals, maintaining a higher resolution along the temporal axis in the early stages is critical for capturing the fine-grained time-domain characteristics necessary to distinguish complex human-to-human interactions.

On the UT-HAR dataset, Wi-CSNet achieves a peak accuracy of 99.92%. Notably, the ablation study reveals an interesting observation: AS-ResNet18 (99.84%) performs marginally worse than the baseline ResNet18 (99.87%). We attribute this phenomenon to the relatively short sequence length (500 packets) of the UT-HAR dataset. For sequences with limited temporal depth, standard downsampling is sufficient for effective feature extraction, whereas asymmetric strides may introduce slight redundancy—especially for simpler, shorter-duration activities.

For the NTU-HAR dataset, the consistent 100% accuracy achieved across all configurations limits the depth of our ablation analysis, but serves as strong evidence of the

stability and robustness of our proposed pipeline. This result demonstrates that Wi-CSNet generalizes effectively without overfitting, even when the input data is highly refined via DWT preprocessing.

6. Conclusions

This paper proposes Wi-CSNet, a lightweight hybrid architecture for CSI-based human activity recognition. By integrating DWT preprocessing, asymmetric-stride CNN, and the Mamba2 SSD mechanism, Wi-CSNet effectively addresses noise, long-range temporal dependencies, and structural asymmetry in CSI data. Extensive experiments on three public benchmarks demonstrate that Wi-CSNet achieves state-of-the-art accuracy with minimal parameter overhead (0.63% over ResNet18) and training time increase.

The success of Wi-CSNet underscores the potential of state space models for HAR tasks, combining the parallel training efficiency of CNNs with the global modeling capability of SSD. Future work will explore:

- (1) Model compression and quantization for deployment on edge devices.
- (2) Integration of phase information after robust sanitization to further disambiguate mirror actions (e.g., left/right kicks).
- (3) Real-time implementation in smart home environments.

Author Contributions: Conceptualization, R.Z. and Z.G.; methodology, R.Z. and Z.G.; software, Z.G.; validation, Z.G.; formal analysis, R.Z. and Z.G.; investigation, Z.G.; resources, Z.G.; data curation, R.Z.; writing—original draft preparation, Z.G.; writing—review and editing, R.Z.; visualization, H.Y.; supervision, R.Z.; project administration, L.Z.; funding acquisition, Y.C. and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by Fujian Provincial Natural Science Foundation of China under Grant 2022J011183 and 2022J011182, in part by National Foundation Cultivation Program of Sanming University under Grant PYT2303, in part by Project of Sanming University under Grant KH25079, in part by Sanming University High-Level Talent Research Start-up Fund under Grant RD21010P.

Data Availability Statement: Previously reported HHI data were used to support this study and are available at <https://data.mendeley.com/datasets/3dhn4xnjxw/1> (accessed on 6 May 2026). These prior studies (and datasets) are cited at relevant places within the text as references [29]. UT-HAR and NTU-HAR can be accessed via the following link: https://drive.google.com/drive/folders/1R0R8SIVbLI1iUFQCzh_mH90H_4CW2iwt (accessed on 6 May 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Diraco, G.; Rescio, G.; Siciliano, P.; Leone, A. Review on human action recognition in smart living: Sensing technology, multimodality, real-time processing, interoperability, and resource-constrained processing. *Sensors* **2023**, *23*, 5281.
2. Dwivedi, C.C.P.; Jeon, G.; Zhang, Y. (Eds.) *Nanosensors as Robust Non-Invasive Diagnostic Tools for Remote Health Monitoring*; CRC Press: Boca Raton, FL, USA, 2025.
3. Liu, J.; Huang, Y.; Shi, X.; Xiong, R.; Zhang, J.; Mi, T.; Qiu, R.C. TRIS-HAR: Transmissive Reconfigurable Intelligent Surfaces-assisted Human Activity Recognition Using State Space Models. *IEEE Internet Things J.* **2025**, *12*, 33421–33436.
4. Joarder, M.K.; Nafee, M.W.; Mohtasim, M.; Sanam, T.F.J. A Robust Signal Processing Framework for CSI-Based Bangla Sign Language Recognition via Doppler-Based Subcarrier Selection, Complex ICA, and Energy-Based Fusion. *IEEE Access*. **2025**, *13*, 189735–189750.
5. Showmik, I.A.; Sanam, T.F.; Imtiaz, H. Human Activity Recognition from Wi-Fi CSI Data Using Principal Component-Based Wavelet CNN. *Digit. Signal Process.* **2023**, *138*, 104056.

6. Bahl, P.; Padmanabhan, V.N. RADAR: An in-building RF-based user location and tracking system. In *Proceedings of the Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies*; Societies: Tel Aviv, Israel, 2000; pp. 775–784.
7. Halperin, D.; Hu, W.; Sheth, A.; Wetherall, D. Tool release: Gathering 802.11 n traces with channel state information. *ACM SIGCOMM Comput. Commun. Rev.* **2011**, *41*, 53.
8. Shen, X.; Ni, Z.; Liu, L.; Yang, J.; Ahmed, K. WiPass: 1D-CNN-based smartphone keystroke recognition using WiFi signals. *Pervasive Mob. Comput.* **2021**, *73*, 101393.
9. Wen, F.; Liang, C. Fine-grained indoor localization using single access point with multiple antennas. *IEEE Sens. J.* **2014**, *15*, 1538–1544.
10. Sai, S.; Sharma, D.; Peelam, M.S.; Chamola, V.; Guizani, M.; Niyato, D. Machine Learning Techniques for Wi-Fi CSI-based Recognition and Sensing: A Comprehensive Review. *IEEE Internet Things J.* **2026**, *Early access*.
11. Xiao, F.; Chen, J.; Xie, X.; Gui, L.; Sun, L.; Wang, R. SEARE: A system for exercise activity recognition and quality evaluation based on green sensing. *IEEE Trans. Emerg. Top. Comput.* **2018**, *8*, 752–761.
12. Quy, T.D.; Lin, C.-Y.; Shih, T.K. Enhanced human activity recognition using wi-fi sensing: Leveraging phase and amplitude with attention mechanisms. *IEEE Trans. Emerg. Top. Comput.* **2025**, *25*, 1038.
13. Arshad, S.; Feng, C.; Liu, Y.; Hu, Y.; Yu, R.; Zhou, S.; Li, H. Wi-chase: A WiFi based human activity recognition system for sensorless environments. In *Proceedings of the 2017 IEEE 18th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, Macau, China, 12–15 June 2017; pp. 1–6.
14. Tian, Y.; Chen, C.; Zhang, Q.; Li, Y.; Li, S.; Ding, X. Multidimensional information recognition algorithm based on CSI decomposition. *IEEE Internet Things J.* **2023**, *10*, 9234–9248.
15. Yousefi, S.; Narui, H.; Dayal, S.; Ermon, S.; Valaee, S. A survey on behavior recognition using WiFi channel state information. *IEEE Commun. Mag.* **2017**, *55*, 98–104.
16. Lai, Y.-R.; Lin, T.-Y.; Liao, J.-K.; Li, C.-Y. WiFi CSI-based Fall Detection from Wearable Devices. In *Proceedings of the 2024 International Computer Symposium (ICS)*, Taipei, Taiwan, 24–26 October 2024; pp. 114–120.
17. Chen, C.; Zhou, G.; Lin, Y. Cross-domain WiFi sensing with channel state information: A survey. *ACM Comput. Surv.* **2023**, *55*, 1–37.
18. Farhana Thariq Ahmed, H.; Ahmad, H.; Phang, S.K.; Vaithilingam, C.A.; Harkat, H.; Narasingamurthi, K. Higher Order Feature Extraction and Selection for Robust Human Gesture Recognition using CSI of COTS Wi-Fi Devices. *Sensors* **2019**, *19*, 2959.
19. Zhang, Y.; Yuan, B.; Yang, Z.; Li, Z.; Liu, X. Wi-nn: Human gesture recognition system based on weighted knn. *Appl. Sci.* **2023**, *13*, 3743.
20. Yan, H.; Zhang, Y.; Wang, Y.; Xu, K. WiAct: A passive WiFi-based human activity recognition system. *IEEE Sens. J.* **2019**, *20*, 296–305.
21. He, Z.; Zhang, X.; Wang, Y.; Lin, Y.; Gui, G.; Gacanin, H. A robust CSI-based Wi-Fi passive sensing method using attention mechanism deep learning. *IEEE Internet Things J.* **2023**, *10*, 17490–17499.
22. Ma, Y.; Zhou, G.; Wang, S.; Zhao, H.; Jung, W. SignFi: Sign language recognition using WiFi. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2018**, *2*, 1–21.
23. Alazrai, R.; Hababeh, M.; Baha’A, A.; Ali, M.Z.; Daoud, M.I. An end-to-end deep learning framework for recognizing human-to-human interactions using Wi-Fi signals. *IEEE Access* **2020**, *8*, 197695–197710.
24. Shafiqul, I.M.; Jannat, M.K.A.; Kim, J.-W.; Lee, S.-W.; Yang, S.-H. HHI-AttentionNet: An Enhanced Human-Human Interaction Recognition Method Based on a Lightweight Deep Learning Model with Attention Network from CSI. *Sensors* **2022**, *22*, 6018.
25. Chen, Z.; Zhang, L.; Jiang, C.; Cao, Z.; Cui, W. WiFi CSI based passive human activity recognition using attention based BLSTM. *IEEE Trans. Mob. Comput.* **2018**, *18*, 2714–2724.
26. Abdel-Basset, M.; Hawash, H.; Moustafa, N.; Mohammad, N. H2HI-Net: A dual-branch network for recognizing human-to-human interactions from channel-state information. *IEEE Internet Things J.* **2021**, *9*, 10010–10021.
27. Li, B.; Cui, W.; Wang, W.; Zhang, L.; Chen, Z.; Wu, M. Two-stream convolution augmented transformer for human activity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vancouver, BC, Canada, 2–9 February 2021; pp. 286–293.
28. Dao, T.; Gu, A. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv* **2024**, arXiv:2405.21060.
29. Alazrai, R.; Awad, A.; Baha’A, A.; Hababeh, M.; Daoud, M.I. A dataset for Wi-Fi-based human-to-human interaction recognition. *Data Brief* **2020**, *31*, 105668.

30. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2024**, arXiv:2312.00752.
31. Yang, L.; Zhang, R.-Y.; Li, L.; Xie, X. Simam: A simple, parameter-free attention module for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning*; PMLR: London, UK, 2021; pp. 11863–11874.
32. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; Liu, Y. Vmamba: Visual state space model. *arXiv* **2024**, arXiv:2401.10166.
33. Yang, J.; Chen, X.; Zou, H.; Lu, C.X.; Wang, D.; Sun, S.; Xie, L. SenseFi: A library and benchmark on deep-learning-empowered WiFi human sensing. *Patterns* **2023**, *4*, 100703.
34. Kabir, M.H.; Rahman, M.H.; Shin, W. CSI-IANet: An inception attention network for human-human interaction recognition based on CSI signal. *IEEE Access* **2021**, *9*, 166624–166638.
35. Islam, M.S.; Kabir, M.H.; Hasan, M.A.; Shin, W. Wi-MIR: A CSI dataset for Wi-Fi based multi-person interaction recognition. *IEEE Access* **2024**, *12*, 67256–67272.
36. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
37. Hu, J.; Ge, F.; Cao, X.; Yang, Z. RGANet: A Human Activity Recognition Model for Extracting Temporal and Spatial Features from WiFi Channel State Information. *Sensors* **2025**, *25*, 918.
38. Yang, S.; Yang, J.; Yang, C.; Yan, W.; Li, P. Human Activity Recognition using WiFi Signal Features and Efficient Residual Packet Attention Network. *IEEE Sens. Lett.* **2025**, *9*, 1–4.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.