Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 4598460, 14 pages
https://doi.org/10.1155/2022/4598460

Research Article

Device-Free Human Activity Recognition Based on Dual-Channel Transformer Using WiFi Signals

Zhihao Gu,[1] **Taiwei He**,[1] **Ziqi Wang**,[2] **and Yuedong Xu**[1]

[1]*School of Information Science and Technology, Fudan University, Shanghai 200433, China*
[2]*Electrical and Computer Engineering Department, University of California, Los Angeles, CA 90095, USA*

Correspondence should be addressed to Yuedong Xu; ydxu@fudan.edu.cn

Received 17 February 2022; Revised 30 May 2022; Accepted 31 May 2022; Published 28 June 2022

Academic Editor: Pan Tang

Human activity recognition plays a significant role in smart building applications, healthcare services, and security monitoring. In particular, WiFi-based indoor wireless sensing system becomes increasingly popular due to its noninvasiveness. This work presents the design and implementation of DARMS, a Device-free human Activity Recognition and Monitoring System that can be deployed with low-cost commodity WiFi devices. DARMS is a passive wireless sensing system, and it can accurately distinguish various daily activities without the user wearing any sensor. DARMS makes two key technical contributions. First, an effective signal processing methodology is designed to extract the CSI features both in the time domain and frequency domain. Second, a dual-channel neural network that combines temporal and frequency information is proposed to achieve fine-grained activity recognition. In our experiments, DARMS shows outstanding performance in different indoor environments, with an average accuracy of 96.9% for fall detection and 93.3% for human activity recognition.

1. Introduction

Human activity recognition is an important task, and it plays a crucial role in smart building applications [1–3], healthcare services [4–6], and security monitoring [7–9]. Many efforts have been devoted to building the activity recognition system based on various sensing devices such as high-resolution cameras or wearable sensors. But these sensing systems still have limitations in privacy security and deployment cost. For example, setting up a camera in a private space, such as toilet or bedroom, may seriously violate the privacy of users. Wearable sensors are usually limited by battery capacity and have to be recharged frequently. A promising approach is to employ pervasive WiFi signals for human activity recognition. WiFi-based passive wireless sensing system does not require users to wear any equipment, and it is becoming ubiquitous due to its noninvasive feature.

WiFi channel state information (CSI) portrays the characteristics of wireless channel. In indoor environment, CSI is affected by human movements through multipath effects and thus carrying motion features. Many WiFi-based sensing systems have been dedicated to designing and implementing various signal processing methods to combine the substantial human movement information carried by CSI.

Since the success of deep learning in computer vision, various accurate and reliable image classification models have been proposed. Inspired by these models, lots of activity recognition systems transplant these mature models for visual tasks into wireless sensing tasks. These studies have been dedicated to transforming the raw CSI or features extracted from CSI into images. Then, feeding these pictures generated by CSI into the image classification model to obtain activity recognition results. However, this approach ignores the essential difference between visual images and

wireless signals. The design of neural network lacks fine-grained temporal or frequency modeling of human activities in the wireless signal space. In this paper, we propose DARMS, a Device-free human Activity Recognition and Monitoring System that can be deployed with low-cost commodity WiFi devices. In summary, our contributions are threefold:

(i) We analyze the CSI model for passive activity recognition and design an effective signal processing methodology that can extract the CSI features both in the time domain and frequency domain

(ii) A novel neural network with the backbone of dual-channel convolution-enhanced transformer is proposed to achieve fine-grained activity recognition. Instead of raw CSI, the neural network takes the CSI features as model input, which can utilize the human motion information carried by CSI data more effectively

(iii) We design and implement DARMS with Intel 5300 NIC, which is a low-cost commodity WiFi device. The experiments conducted in different environments show that DARMS achieves an average accuracy of 96.9% for fall detection and 93.3% for human activity recognition

The rest of the paper is organized as follows. We provide a short review of related work in Section 2 and give an overview of our system design in Section 3. Section 4 analyzes the CSI model for passive activity recognition, and Section 5 describes the methodology of data acquisition and processing. In Section 6, we elaborate the design of neural network in DARMS. Section 7 presents the experimental settings and evaluation results, followed by a conclusion in Section 8.

2. Related Work

Many efforts have been devoted to building device-free human activity recognition system based on WiFi signals. They can be divided into two categories: model-based method and deep learning-based method.

2.1. Model-Based Method. Model-based activity recognition systems explicitly build physical models between wireless signals and human movements [10–16]. Many published works extract the characteristics from CSI amplitude and phase to distinguish various movements. CARM [15, 16] builds the model of WiFi-based human activity recognition and employs a hidden Markov model-based approach to identify various activities accurately. Two theoretical models are proposed in CARM, one is the CSI-speed model which quantifies the relation between CSI dynamics and human movement speeds, and the other is the CSI-activity model which establish the relation between movement speeds and human activities. Niu et al. model the wireless sensing to reveal the impact of static multipaths in activity recognition, and a novel method that exploits multipath effect is proposed to improve the wireless sensing performance [10]. WiFall [11] analyzes radio propagation model to build the correlations between different radio signal variations and activities. It uses CSI amplitude-related features to characterize falls and other daily movements. DeFall [13] is a passive fall detection system using WiFi signals. It probes the distinctive patterns of speed and acceleration in human falls. DeFall compares the patterns of the real-time speed and acceleration estimates against the template to detect falls. WiAct [14] takes the Doppler frequency shift information as the input for extreme learning machine to classify different activities. These works have in common that they extract the signal features on the basis of wireless sensing models. Then, employing traditional machine learning classifiers (e.g., SVM, random forest, and ELM) or setting threshold to distinguish the features of various activities. However, for activity recognition task, it is difficult to handle features with too high dimension and complex structure only based on traditional signal model, restricting them to achieve fine-grained activity recognition with higher accuracy. Thus, DARMS proposes a novel dual-channel neural network to process the CSI features extracted from the signal processing algorithm.

2.2. Deep Learning-Based Method. Many wireless sensing systems using WiFi signals rely on data-driven approaches to learn the complex relationship between signal characteristics and human activities [2, 6, 17–26]. Widar 3.0 [17] uses convolutional neural network (CNN) and recurrent neural network (RNN) to mine the properties of CSI features in the spatial dimension and temporal dimension. It can distinguish eight gestures with an average accuracy of 92.9%. Chen et al. propose an attention-based bidirectional long short-term memory (ABLSTM) network for passive human activity recognition [21]. The ABLSTM is employed to learn representative features in two directions from raw sequential CSI measurements. CSAR [24] proposes a channel selection mechanism to actively select available WiFi channels with better quality. It also employs LSTM to combine the CSI features for activity recognition. Instead of using CNN and RNN, STFNets [18] propose a new foundational neural network block. It integrates a widely used time-frequency analysis method, the short-time Fourier transform, into data processing to learn features directly in the frequency domain. In RaGAM [25], the statistical characteristics of RSS data in the time domain are fed into a probabilistic neural network to achieve target intrusion sensing based on WiFi. Li et al. design a neural network with two-stream structure to capture both time-over-channel and channel-over-time features of CSI [22]. We also notice the concept of "dual-view" [26] that merges the time series image of XOZ plane and that of YOZ plane collected by millimeter-wave radar as the input of neural network, which is different from the meaning of "dual-channel" in DARMS. Compared with previous works, the design of neural network in DARMS takes the signal characteristics into consideration. Instead of raw CSI, DARMS takes the temporal and frequency features of CSI as model input, which can utilize the human motion information carried by CSI data more effectively. Besides, inspired by state-of-the-art neural network structures, a novel neural network built with the backbone of convolution-enhanced transformer is proposed in DARMS to improve the performance of activity recognition.

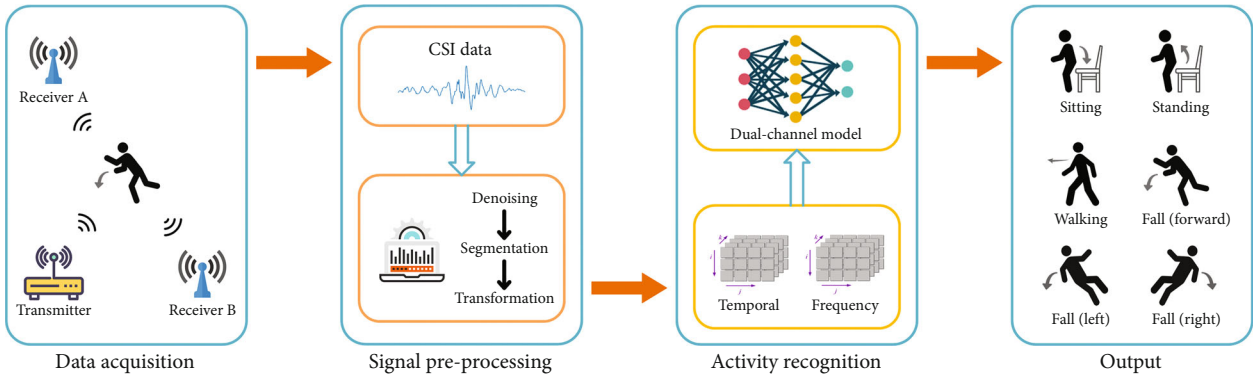


Figure 1: System overview. DARMS is composed of three key components, hardware module for data acquisition, software module for signal preprocessing, and neural network for activity recognition.

3. System Overview

Figure 1 illustrates the architecture of DARMS. It is composed of three key components: data acquisition hardware module, signal preprocessing software module, and activity recognition module. The hardware and the software components work as a unity to transform the low-cost commodity WiFi devices into a wireless sensing platform.

3.1. Data Acquisition Hardware. This component is aimed at collecting the WiFi signals that can be used to recognize human activities. During the data collection process, we use one transmitter and two vertically placed receivers to record human movements. Each of them is equipped with a low-cost commodity WiFi device (Intel 5300 NIC, cost ~ $5). Then, we use the Linux CSI Tool [27] to extract CSI from the collected WiFi packets.

3.2. Signal Preprocessing Software. In this module, DARMS sanitizes the CSI data so that they can be fed into the proposed deep learning model. Specifically, we first perform data denoising to remove the phase offset and amplitude shift of raw CSI data. Then, we propose an activity indicator to extract the CSI segment that contains human activities. Finally, we transform each CSI segment into the representation that can be fed into the neural network.

3.3. Activity Recognition. This component is designed to recognize human activities using the processed CSI data that generated from the signal preprocessing module. To achieve the goal, we design a dual-channel neural network that considers both time and frequency domain information to achieve fine-grained human activity recognition. The details of the proposed deep learning model are described in Section 6.

4. CSI Model for Passive Activity Recognition

CSI portrays the channel state between the transmitter and the receiver, which is sensitive to environmental changes and human body's movements. Thus, we will model the impact of human activity on CSI first.

4.1. Multipath Effect. For indoor environment, walls, ceiling, furniture, human body, etc., can reflect wireless signals. Thus, multiple transmission paths exist between the transmitter and receiver. Taking the multipath effect into consideration, the CSI of a subcarrier with center frequency f at time t could be modeled as follows [15, 17, 28]:

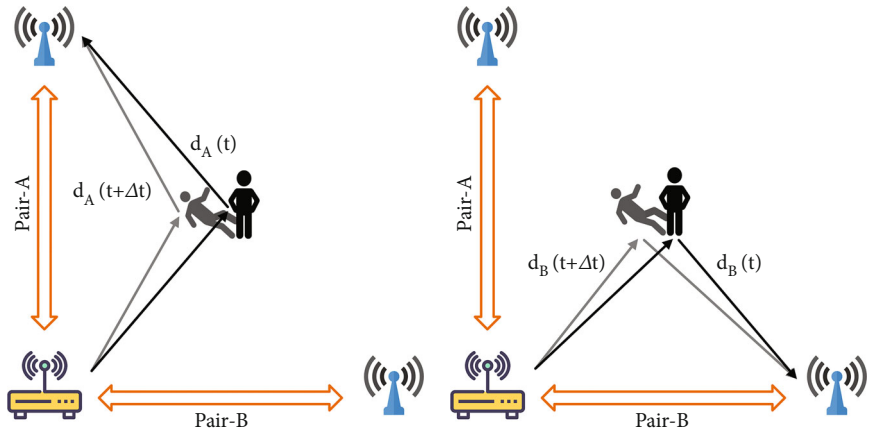
$$\begin{cases} H(f,t) = e^{-j2\phi(f,t)}[h_s(f) + h_d(f,t)], \\ h_d(f,t) = \sum_{l=1}^{N} \alpha_l(f,t) e^{-j2\pi f \tau_l(t)}, \end{cases} \tag{1}$$

where $h_s(f)$ describes the channel state of static part, including the line-of-sight (LoS) signal and the signals reflected by ambient objects, $h_d(f,t)$ represents the channel state of dynamic part, corresponding to the signals reflected by moving human body. $\phi(f,t)$ is the phase shift caused by carrier frequency offset, sampling frequency offset and timing alignment offset. $\alpha_l(f,t)$ and $\tau_l(t)$ denote the amplitude attenuation and propagation delay of the signal transmitted through l_{th} path, respectively. According to Equation (1), the amplitude of CSI is determined by the sum of multipath signals, including the signal transmitted through LoS path and the signal reflected by objects or human body. The movement of the human body will change the propagation path of the signal. Thus, the fluctuation of CSI amplitude in the time domain portrays the impact of movements on the signals reflected by human body.

4.2. Doppler Effect. When a person moves between a pair of transmitter and receiver, his movement will lead to Doppler effect that shifts the frequency of received signals. The Doppler frequency shift $f_D(t)$ is as follows:

$$f_D(t) = \frac{\Delta d_{\text{path}}}{\Delta t \cdot c} f_0, \tag{2}$$

where Δd_{path} represents the change in the length of signal propagation path in the duration Δt, c denotes the propagation speed of WiFi signal in the air, and f_0 is the carrier frequency of the signal. According to Equation (2), the



(a) Reflection path of pair-A (b) Reflection path of pair-B

Figure 2: For the Tx-Rx pairs placed vertically, the changes in the length of reflection path are unequal.

Doppler information in the frequency domain contains the speed of length change in signal propagation path. In indoor environment, the propagation path can be divided into the static part and dynamic part. While the dynamic part contains the signals reflected by the moving human body and environmental noise. Thus, human movements will change the path length of the signals reflected by human body. The speed of human body is highly correlated with the speed of length change in reflection path. Thus, the Doppler information extracted from CSI data can be used to infer human activity from another perspective. After taking the Doppler effect into consideration, the dynamic part of CSI $h_d(f,t)$ in Equation (1) can be transformed as follows:

$$h_d(f,t) = \sum_{l=1}^{N} \alpha_l(f,t) e^{-j2\pi f \int_{-\infty}^{t} f_{D_l}(x)dx}. \tag{3}$$

According to Equation (1), CSI can be divided into static part $h_s(f)$ and dynamic part $h_d(f,t)$, but we only interested in the dynamic part $h_d(f,t)$ that contains the information of human body's movements. Besides, the phase shift $e^{-j2\phi(f,t)}$ hinders us from directly extracting the $h_d(f,t)$. To eliminate the phase shift, we calculate the conjugate product of CSI:

$$\begin{aligned} |H(f,t)|^2 = {} & h_s(f)h_s(f)^* + h_s(f)\sum_{l=1}^{N} \alpha_l(f,t) e^{j2\pi f \int_{-\infty}^{t} f_{D_l}(x)dx} \\ & + h_s(f)^* \sum_{l=1}^{N} \alpha_l(f,t) e^{-j2\pi f \int_{-\infty}^{t} f_{D_l}(x)dx} \\ & + h_d(f,t)h_d(f,t)^*. \end{aligned} \tag{4}$$

The conjugate product of CSI consists of four terms. In general, the power of signals transmitted through static path is much larger than that of the signals reflected by moving body, i.e., $|h_s| \gg |h_d|$. Thus, the first term, denotes the conjugate product of static path components, is the largest term. The fourth term denotes the conjugate product of dynamic path components, is much smaller than the first three term and can be ignored, while the second and third terms contain the information of human body movements that can be used to realize activity recognition.

4.3. Necessity of Multiview Perception. To achieve accurate and find-grained activity recognition, deploying multiple Tx-Rx pairs at different angles is necessary. As depicted in Figure 2, a volunteer stands between two Tx-Rx pairs, if the volunteer is falling down towards pair-A, the length of reflection path will change accordingly. The length change of pair-A $\Delta d_{\text{path}A} = |d_A(t+\Delta t) - d_A(t)|$ and that of pair-B $\Delta d_{\text{path}B} = |d_B(t+\Delta t) - d_B(t)|$ are unequal. According to Equation (2), the Doppler effect of CSI received by pair-A is different with that of pair-B. Thus, we can take advantage of the various Doppler effect on multiple Tx-Rx pairs to sense the direction of human movement.

5. Data Acquisition and Processing

5.1. Data Acquisition. To detect human activity using WiFi signals, we first collect the channel state information through the CSI Tool [27]. The CSI Tool logs CSI data on 30 subcarriers for each signal transmission link. We combine all collected CSI data to form a CSI matrix **H** with the dimension of $N_{\text{sub}} \times T \times N_{\text{stream}}$, where N_{sub} represents the number of subcarriers that can extract CSI data, T represents the total number of received packets. N_{stream} is the number of spatial streams in MIMO system which equals the number of transmitting antennas times that of receiving antennas.

5.2. CSI Preprocessing. Due to the raw CSI data contains a lot of noise, thus it cannot be fed into the neural network directly. Figure 3 illustrates our preprocessing workflow to extract the features from the CSI data. The preprocessing module can be divided into three parts: denoising, segmentation and transformation.

5.2.1. Denoising. In this stage, DARMS extracts the amplitude of CSI by calculating the modulus of CSI complex. Figure 3(a) presents the amplitude of raw CSI data (only one subcarrier is plotted for clarity). Then, DARMS applies

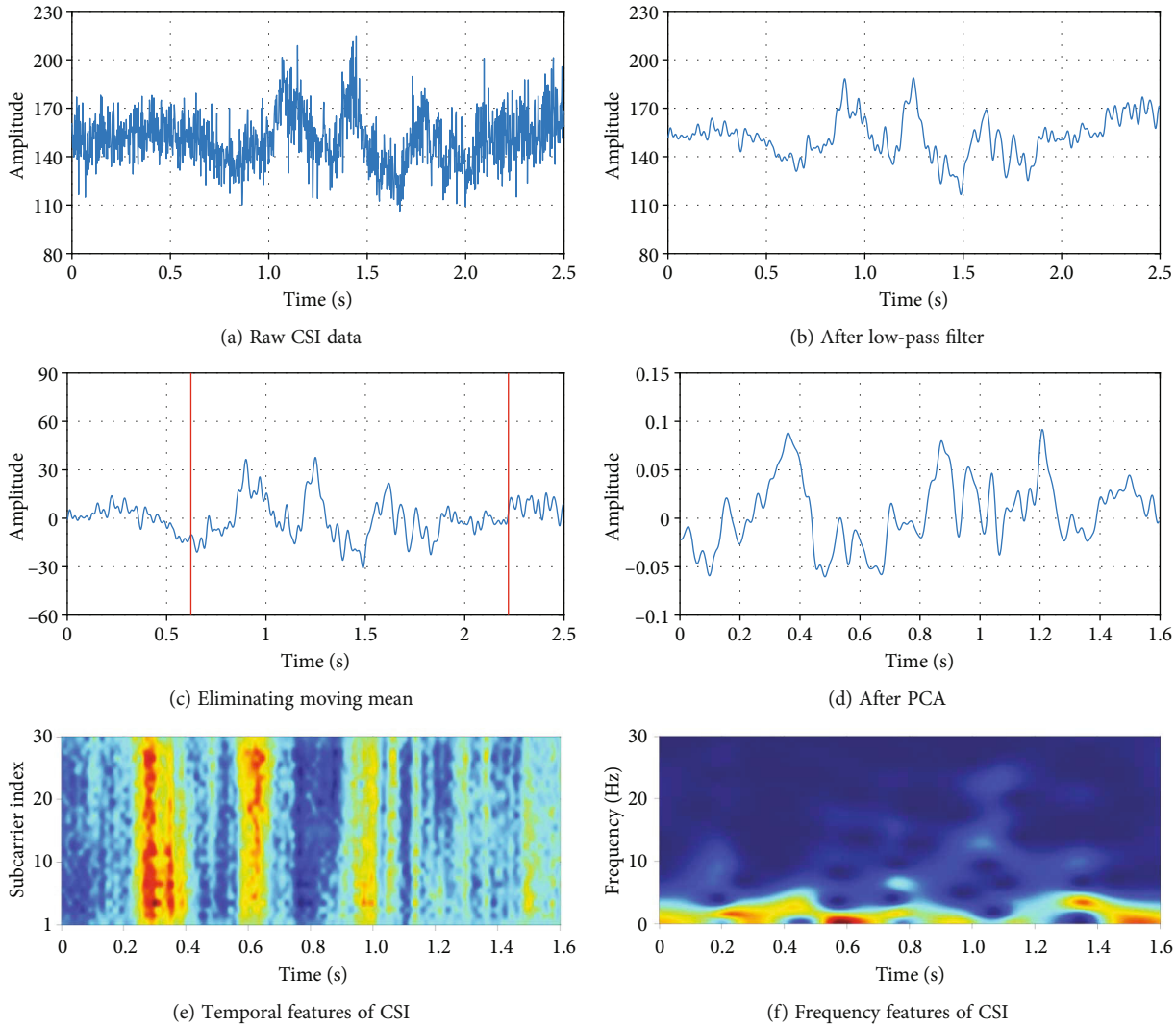


Figure 3: The preprocessing workflow of DARMS.

a 6-order Butterworth low-pass filter (the cutoff frequency is 30 Hz) on each CSI stream to remove high-frequency noise. The result after the filter is plotted in Figure 3(b). To remove the static part of CSI data, DARMS calculates the moving average for each CSI stream with a 1-second sliding window, i.e., the length of sliding window is 500 if the sample rate equals 500 Hz. Figure 3(c) shows the amplitude of CSI data after eliminating the moving mean.

5.2.2. Segmentation. To extract the CSI segment that records human body movements, we design an activity indicator to detect the beginning and end of each movement. In the experiments, one can observe that CSI variance is sensitive to environmental changes. For instance, we ask a volunteer to sit on a chair, the collected CSI after denoising is depicted in Figure 4(a) (only presents the CSI of one subcarrier among all 180 subcarriers for simplicity). When the volunteer is stationary (stand still and sit still), the amplitude of CSI is relatively stable. While the volunteer is sitting down, the processed CSI fluctuates violently. Thus, we propose a variance-based indicator for movement segmentation. The center of the segment p is decided by the following:

$$p = \arg\max_{s} \sum_{i,k} \sum_{j=s}^{s+L} \mathrm{std}\left(h_{i,j-L/2,k}, h_{i,j+L/2,k}\right), \tag{5}$$

where L represents the length of CSI segment that records movements, $h_{i,j,k}$ represents the j^{th} CSI data of i^{th} subcarrier that received by k^{th} spatial stream, and $\mathrm{std}(h_{i,j-L/2,k}, h_{i,j+L/2k})$ is the standard deviation of the CSI segment from $h_{i,j-L/2,k}$ to $h_{i,j+L/2,k}$.

Figure 4(b) shows the amplitude of activity indicator, DARMS takes the maximum point of the indicator as the center of the activity. Then, according to the predefined segment length L, the beginning and end of the activity are $p - \mathrm{L}/2$ and $p + \mathrm{L}/2$, respectively. For example, in Figure 3(c), the data between the two red lines is the CSI segment extracted by the activity indicator.

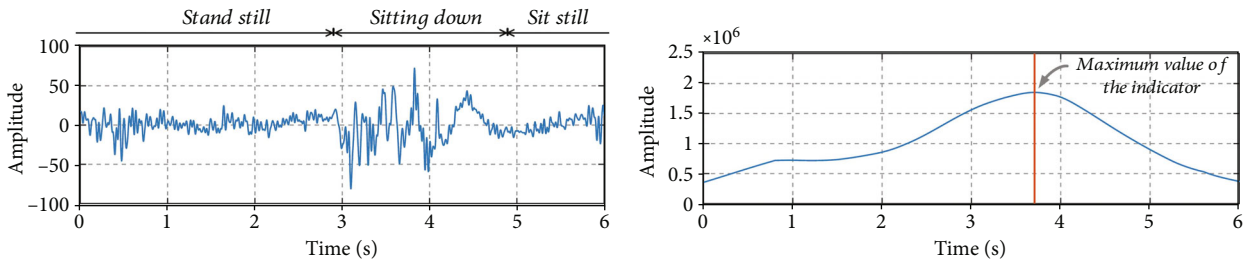


Figure 4: (a) The amplitude of processed CSI. (b) The amplitude of the activity indicator reaches the peak while the volunteer is sitting down.

Input: The raw CSI data $\mathbf{H} \in \mathbb{C}^{N_{\text{sub}} \times T \times N_{\text{stream}}}$.

The element $h_{i,j,k}$ in $\mathbf{H}$ represents the j^{th} CSI data of i^{th} subcarrier that received by k^{th} spatial stream.

Output: Temporal features T_c, frequency features F_c.

1: Extracting the moving average of CSI amplitude. $h^m_{i,j,k} = |h_{i,j,k}| - 1/W \sum_{t=j-W/2}^{j+W/2} |h_{i,t,k}|$, where W is the length of sliding window.

2: Utilizing a six-order Butterworth low-pass filter along the second dimension of $\mathbf{H}$ to remove high-frequency noise. $h^l = \text{Lowpass}(h^m)$

3: Calculating the activity indicator for h^l based on Equation (5) and finding the peak index of activity indicator p.

4: Extracting the CSI segment h^s from h^l, in which $h^s = [h^l_{i,p-L/2,k}, h^l_{i,p+L/2,k}]$ and L represents the length of CSI segment.

5: Taking h^s as the output of temporal CSI features $T_c \in \mathbb{R}^{N_{sub} \times L \times N_{stream}}$.

6: Doing PCA for T_c.

7: Calculating the Doppler spectrum D of T_c based on STFT, in which $D \in \mathbb{R}^{N_f \times L \times N_{stream}}$ and N_f represents the number of FFT points in STFT.

8: Discarding the frequency bins above 30 Hz in D. Then obtaining the output of frequency features $F_c \in \mathbb{R}^{N_{\text{freq\_bin}} \times L \times N_{\text{stream}}}$, where $N_{\text{freq\_bin}}$ is the number of frequency bins below 30 Hz.

9: return T_c and F_c

Algorithm 1: The pseudocode of preprocessing module.

5.2.3. Transformation. After the segmentation module, we obtain the CSI segment that contains human movement information in the time domain. Figure 3(e) vividly shows the heatmap of whole CSI segment, including the CSI data of 180 subcarriers. To extract its Doppler frequency shift profile, we need to transform the CSI data into frequency domain. To reduce the size of output, we apply a principal component analysis (PCA) on CSI streams of each antenna (each antenna has 30 CSI streams) and only the prominent dynamic component is retained. Thus, we can reduce the dimension of CSI matrix from $N_{\text{sub}} \times L \times N_{\text{stream}}$ to $1 \times L \times N_{\text{stream}}$. Figure 3(d) plots the results of PCA (only one CSI stream is plotted in the figure for simplicity). Then, we conduct short-time Fourier transform to extract the Doppler spectrum for each row of the new CSI matrix. We also discard the frequency bins that above 30 Hz to further compress the size of Doppler spectrum. The result of transformation step is depicted in Figure 3(f).

Algorithm 1 summarizes the workflow of signal preprocessing module, which takes the raw CSI data $\mathbf{H}$ as input and returns the temporal features T_c and frequency features F_c extracted from CSI data.

6. Neural Network Design

Figure 5 illustrates the architecture of our dual-channel neural network. The CSI data after preprocessing is split into time and frequency streams. The CNN layer extracts discriminative features from the two streams. Then, the features will be fed into the swin transformer to explore the sequence information. Finally, the dual-channel features are aggregated and put into the multilayer perceptron for the final output.

6.1. Model Inputs. After the preprocessing stage, we obtain the CSI matrix both in the time domain and frequency domain that will be fed into the temporal channel and frequency channel, respectively. The input of temporal channel T_c is the tensor with dimension of $N \times T \times k$, and the input of frequency channel is the Doppler spectrum of all spatial streams, represented by F_c with dimension of $M \times L \times k$, where N is the number of subcarriers per spatial stream, T is the length of CSI segment, k is the number of all spatial streams, and M and L represent the number of frequency bins of STFT and the length of Doppler spectrum. To adapt to the subsequent 2D

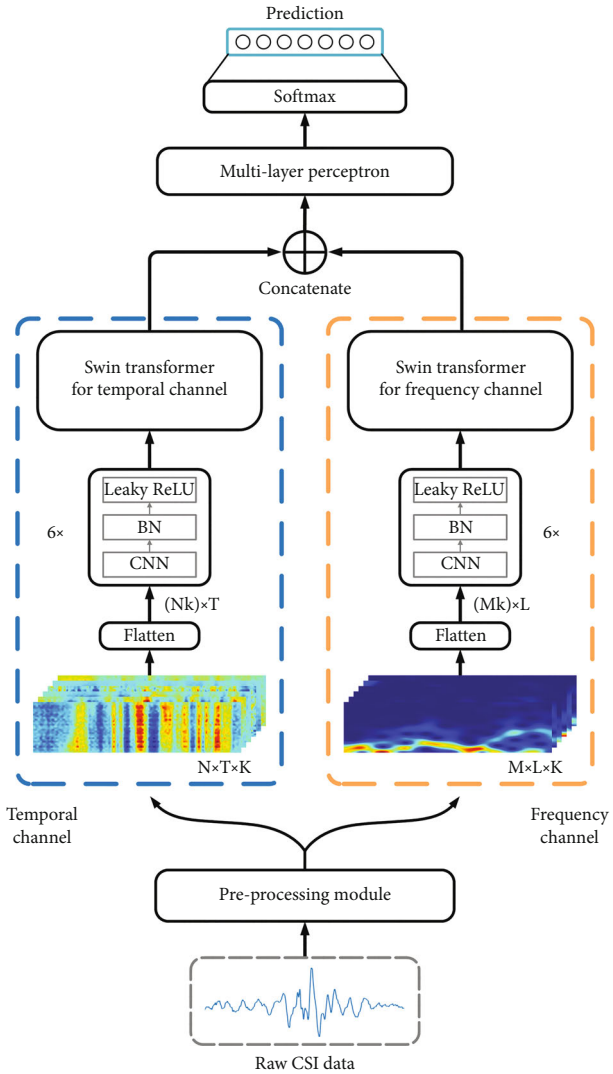


FIGURE 5: The overview of dual-channel neural network. It combines temporal and frequency features of CSI to classify various human activities.

convolutional network, we flatten the input data. The dimension of temporal channel is converted to $(Nk) \times T$, and that of frequency channel is transformed to $(Mk) \times L$ before feeding into the neural network.

6.2. Dual-Channel Convolution-Enhanced Transformer. This layer extracts discriminative features of CSI both from the temporal channel and frequency channel.

6.2.1. CNN Module. Convolutional neural network (CNN) is a powerful technique to extract features from complex inputs. Thus, we first use CNN module to process the input data. Specifically, we use a stacked six layer 2D CNN module. After each convolutional layer, we employ a batch normalization (BN) layer to normalize the mean and variance of the output, followed by a leaky rectified linear unit (Leaky ReLU) to add nonlinearity to the model.

6.2.2. Transformer Module. Then, we employ the swin transformer (tiny version) [22] to process the feature maps generated by CNN module. The swin transformer is composed of patch partition, linear embedding, path merging, and repeated swin transformer blocks. The structure of swin transformer block is illustrated in Figure 6. A swin transformer block is made up of two sequentially stacked subblocks. The first part contains a window-based multihead self-attention (MSA) module [29], followed by a multilayer perceptron (2-layer) which takes Gaussian error linear units (GELU) [30] as its activation function. A LayerNorm (LN) [31] is applied before each MSA module and each MLP, and a residual connection [32] is applied after each module. The second part is built by replacing the window-based MSA module by the MSA module based on shifted windows, with other layers keeping the same with the first part.

6.3. Aggregation Layer. We use a multilayer perceptron (MLP) to aggregate the features encoded by the dual-channel swin transformer. Assume that the features extracted from the temporal channel and frequency channel with the dimension of $d_t \times 1$ and $d_f \times 1$ (after flatten). Before feeding into the MLP, we concatenate the two output vectors to generate the final feature vector with the dimension of $(d_t + d_f) \times 1$. Then, the MLP takes the final feature vector to identify different human activities.

6.4. Loss Function. To minimize the error between the model outputs and the corresponding ground truth, we employ a standard cross-entropy loss as the loss function L, which is defined as follows:

$$L = -\sum_{i}^{M} y_i \log \left[f\left(\theta; \sigma_t; \sigma_f\right) \right], \tag{6}$$

where $f(\cdot)$ denotes the distribution of activity prediction and y_i denotes the corresponding ground truth, θ is the weight of model, σ_t and σ_f are the input of temporal and frequency channel, and M is the total number of activity categories. We optimize the loss function through Adam algorithm [33].

7. Evaluation

In this section, we present the implementation and performance of DARMS in detail.

7.1. Experimental Setup. We evaluate the performance of DARMS in a rectangular meeting room with area of 52 m^2 and a smaller office room of 36 m^2. The layouts of the rooms are shown in Figure 7, with some furniture and cabinets inside, making each of them a rich multipath environment. We build the system prototype using three ThinkPad X200 laptops, and each of them is equipped with an Intel 5300 NIC. One laptop is used as transmitter and the other two are receivers. The CSI Tool [27] developed by Halperin et al. is loaded on all three computers for transmitting WiFi packets and collecting CSI. For the transmitter, we employ one antenna to send 500 packets per second and the WiFi carrier frequency is set on

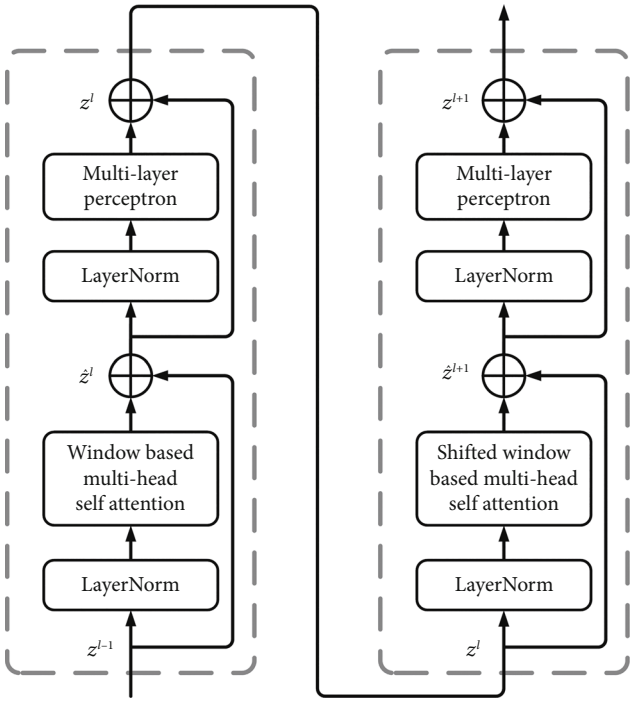


Figure 6: The overview of swin transformer block.

channel 64 (5.32 GHz). For the two receivers, each of them is equipped with three antennas to received WiFi packets. In DARMS, the two Tx-Rx pairs are placed perpendicularly to achieve maximum sensitivity and the distance between each Tx-Rx pair is 3 m. All antennas are placed at the same height (70 cm) from the ground.

7.1.1. Dataset. We recruit 21 volunteers (16 males and 5 females) with different heights (range from 155 cm to 185 cm) and weights (range from 45 kg to 80 kg) in the experiments. Volunteers are asked to perform one of the six movements that happen frequently in daily life. These movements are depicted in Figure 1. Besides, if there is no one in the action zone depicted in Figure 7, it will be called "empty". In total, we collect CSI data for 3050 sets of movements (empty: 500 sets, walking: 500 sets, standing up: 500 sets, sitting down: 500 sets, fall to the left: 350 sets, fall to the right: 350 sets, and fall to the front: 350 sets) to form the dataset.

7.1.2. Model Settings. The input of the neural network is the CSI data after preprocessing module. It is composed of temporal part and frequency part, both of them with the dimension of $30 \times 800 \times 6$. When implementing the stacked six-layer CNN module, 2D convolution operation is used to process the CSI data of temporal channel and frequency channel. The numbers of the convolutional filters in these layers are set as 8, 64, 32, 16, 8, and 1, respectively. The kernel size of all convolutional layers is 3. In the CNN module, Leaky ReLU with the negative slope set as 0.02 is employed as the activation function. In the swin transformer, we adopt the Swin-T, which is detailedly described in [34]. To optimize the loss function, we employ the Adam algorithm. The learning rate of the optimizer is set as 0.00005, and the exponential decay rates β_1 and β_2 are 0.9 and 0.999.

7.2. Experimental Results and Analysis. We now evaluate how the performance of DARMS is affected by the system parameter settings. For each activity, we randomly select 80% of the data (2440 sets in total) to form the training group, while the remaining part (610 sets in total) to form the test group.

7.2.1. Evaluation Metrics. To evaluate the performance of DARMS more accurately and comprehensively, we measure its performance from the accuracy of fall detection and that of activity recognition, i.e., the proportion of correctly recognized activities among all predictions.

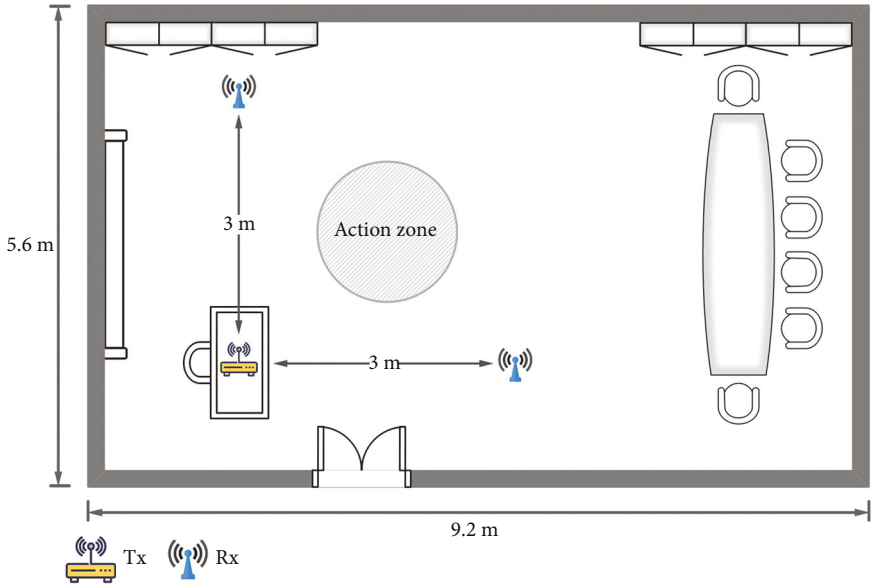
(i) *Fall Detection*. All the test cases are divided into two parts: fall movements (fall to the left, fall to the right, and fall to the front) and nonfall movements (empty, walking, sitting down, and standing up).

(ii) *Activity Recognition*. All the test cases are divided into seven activities: empty, walking, sitting down, standing up, fall to the left, fall to the right, and fall to the front.

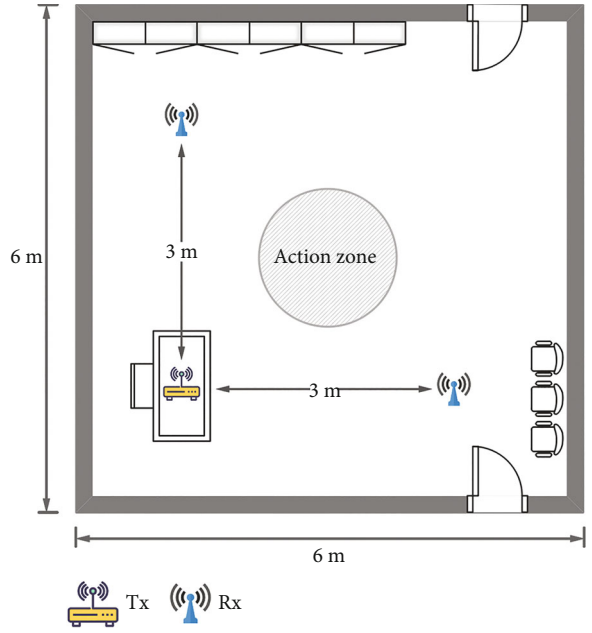
7.2.2. Impact of Array Size. To understand the impact of the array size, we repeat the experiments to test DARMS's performance with one, two, and three antennas of each receiver. The accuracy of fall detection and activity recognition is plotted in Figure 8(a). For fall detection, the accuracy is 88.2% with one antenna, 91.8% with two antennas, and 96.9% with all three antennas. Meanwhile, the accuracy of activity recognition is 83.9%, 89.3%, and 93.3% with one, two, and three antennas. One can clearly observe that larger array size brings better accuracy both in fall detection and activity recognition, while using more antennas increases the size of data fed into the neural network. Thus, the size of the model will be larger, which will consume more computing resources. There is a tradeoff between computing complexity and sensing accuracy.

7.2.3. Impact of Segment Length. The segment length is a key factor in sensing accuracy. We conduct experiments to test DARMS's performance with different segment length settings. In the experiments, we set the segment length as 600, 800, and 1000. The results are plotted in Figure 8(b). For fall detection, the accuracy is 94.6%, 96.9%, and 97.0% with the segment length is 600, 800, and 1000. Meanwhile, the accuracy of activity recognition is 91.6%, 93.3%, and 93.6%. It is clearly that the increase in segment length from 600 to 800 brings obvious performance enhancement. While the increasing from 800 to 1000 only brings slight improvement. It is worth emphasizing that the length of CSI segment determines the input size of neural network, i.e., longer segment length means larger model. Thus, to accelerate the computation, we set the segment length to 800 in the remaining experiments.

7.2.4. Impact of Tx-Rx Pair Layout. To evaluate the impact of Tx-Rx pair layout, we conduct experiments to compare



(a) The layout of room A



(b) The layout of room B

FIGURE 7: Testbed environment. The experiments are conducted in room A and room B.

DARMS's performance with only using Tx-Rx pair-A, pair-B, and both of them. The result is shown in Figure 8(c). The accuracy of fall detection is 92.0% with only using pair-A, 91.1% with pair-B, and 96.9% with both of them, while for activity recognition, the accuracy is 77.2%, 76.2%, and 93.3%, respectively. Obviously, one more perception link brings significant enhancement. Thus, it is important to employ multi-view perception links to achieve accurate and fine-grained wireless sensing.

7.2.5. Impact of Denoising Algorithm. To study the performance of denoising algorithm proposed in DARMS, we compare the performance of four denoising methods. In method 1, we do not apply any preprocessing algorithm to denoise the CSI data, i.e., we remove the denoising step in the signal preprocessing module. The remaining parts of the preprocessing module, including the step of segmentation and transformation, keep the same. In method 2, we utilize the Savitzky-Golay filter to process the noisy CSI data, which can effectively preserve the envelope of the raw waveform [35, 36]. Method 3 applies the singular value decomposition (SVD) to denoise the raw CSI data, which can eliminate the background CSI and effectively extract the channel information of signals reflected by human bodies [37]. Method 4 is the denoising algorithm proposed in DARMS. The results are listed in Table 1. For method 1, the accuracy of fall detection is 85.6% and that of activity recognition is 81.3%. Meanwhile, the accuracy of fall

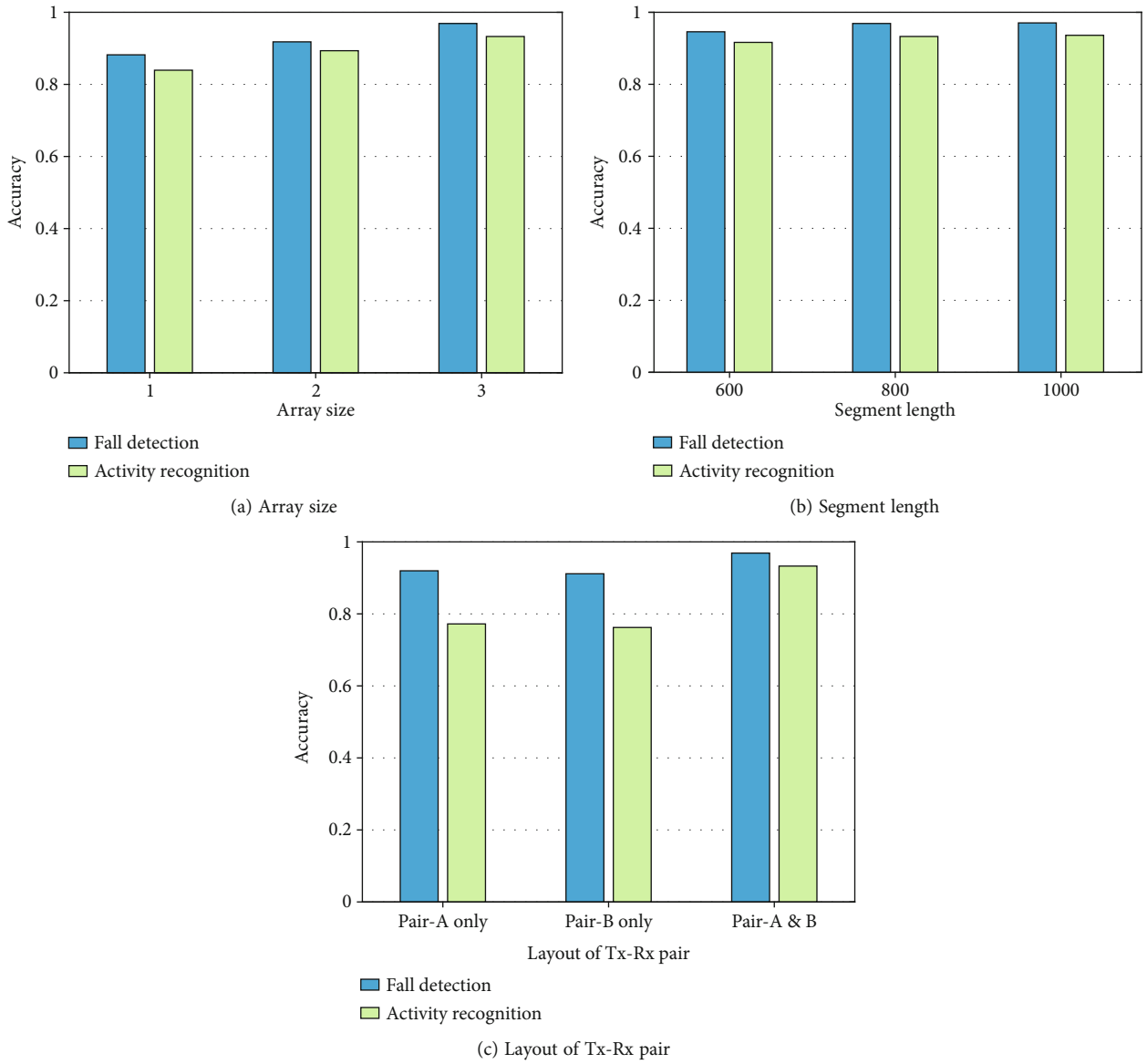


Figure 8: Factors affecting the system performance. (a) Array size. (b) Segment length. (c) Layout of Tx-Rx pair.

Table 1: Comparison: system performance with various denoising algorithms.

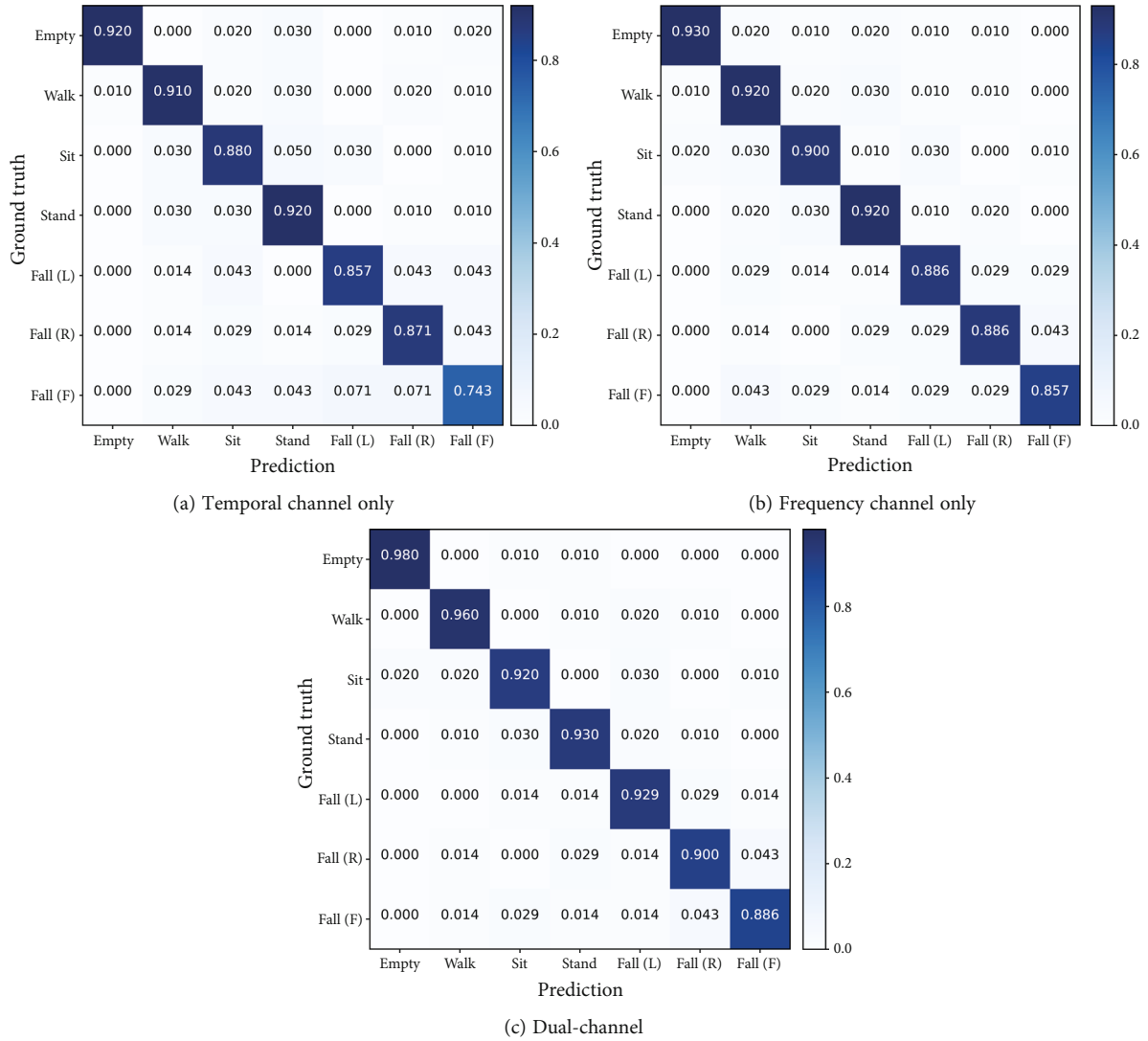
	Fall detection	Activity recognition
Without denoising	85.6%	81.3%
SG filter	95.8%	93.0%
SVD	94.9%	92.2%
DARMS	96.9%	93.3%

detection and activity recognition is 95.8% and 93.0% with method 2. When taking method 3 as the denoising algorithm, the accuracy is 94.9% and 92.2%, respectively. Compared with the methods 2 and 3, the preprocessing workflow proposed by DARMS can slightly improve the performance of fall detection and activity recognition, with the accuracy of 96.9% and 93.3%.

7.2.6. Ablation Study. We conduct the ablation study to evaluate the contribution of each module in the dual-channel neural network. In each test, we ablate a specific component from the full model, including the temporal channel, frequency channel, and transformer block. Table 2 summarizes the experimental results. The accuracy of fall detection is 96.1% while only applying frequency channel 95.4% with only temporal channel and 96.9% with dual-channel. For activity recognition, the accuracy is 90.3%, 87.9%, and 93.3%, respectively. The confusion matrix of activity recognition is depicted in Figure 9. It can be observed that the ablation on frequency module incurs a bigger decline than the ablation of temporal channel. One can conclude that dual-channel structure does truly help DARMS to achieve a better recognition accuracy. We also test the model

Table 2: The results of ablation study compared with the full model of DARMS.

Model	Fall detection		Activity recognition	
	Accuracy	Variation	Accuracy	Variation
DARMS	96.9%	-	93.3%	-
(i) Removing temporal channel	96.1%	-0.8%	90.3%	-3.0%
(ii) Removing frequency channel	95.4%	-1.5%	87.9%	-5.4%
(iii) Removing transformer block	92.2%	-4.7%	86.9%	-6.4%



(a) Temporal channel only

(b) Frequency channel only

(c) Dual-channel

Figure 9: The confusion matrix of activity recognition. (a) Only employ temporal channel. (b) Only employ frequency channel. (c) Employ dual-channel.

Table 3: Comparison: system performance with various backbones.

	Fall detection	Activity recognition
SVM	93.6%	89.8%
ResNet	95.7%	91.2%
CNN+LSTM	96.3%	92.6%
DARMS	96.9%	93.3%

performance if the transformer block is removed. It is worth emphasizing that the output size of CNN module is too large to be directly fed into the multilayer perceptron. Thus, we use a stacked three-layer CNN module to replace the transformer block. In the three-layer CNN module, each convolutional layer is followed by a max pooling layer (the kernel size and stride equal 2) to downsample the output. Compared with the full model of DARMS, the ablation on transformer block incurs an accuracy decline of 4.7% with fall detection and 6.4% with

activity recognition. It reveals that the feature extraction ability of transformer block is better than the traditional CNN module in this classification task, which is critical to the improvement of model performance.

7.2.7. Comparison with Various Backbone. We compare DARMS's performance with four CSI-based activity recognition approaches. In the first approach, we design a set of Gabor filters to extract features from the preprocessed CSI data. Then, the features are used to train an SVM classifier (conducted by LIBSVM library [38], using Gaussian kernel) to distinguish different activities. In the second approach, we transform the processed CSI data into grey images and then feeding the images into the neural network (employing the ResNet-101 as the backbone) to generate the prediction of activity. In the third approach, the transformer module in DARMS is replaced with long short-term memory (LSTM) network and the rest parts (preprocessing module and dual-channel structure) are the same as DARMS. The hidden dimension of the LSTM module is 1024. The fourth approach is the methodology of DARMS. Table 3 shows the performance of four approaches, from which we can observe that DARMS achieves the best performance among all tested algorithms.

7.2.8. Cross-Environment Performance. We conduct a set of experiments to evaluate the cross-environment performance of DARMS. In the first test, we train the model with the dataset collected in room A and test the model with the dataset collected in room B. In the second test, we swap the dataset used for training and validation. Table 4 shows the results. The model trained by the dataset collected in room A achieves the accuracy of 92.6% for fall detection and 88.5% for activity recognition. If the model is trained by the dataset of room B, the accuracy of fall detection and activity recognition is 90.8% and 87.3%, respectively. The performance of DARMS only slightly drops when the environment changes.

7.2.9. Cross-Person Performance. We also evaluate the impact of user's body shape on system performance. In the evaluation of DARMS, we recruit 21 volunteers (16 males and 5 females) with various heights (range from 155 cm to 185 cm) and weights (range from 45 kg to 80 kg) to conduct the experiments. In the experiment, we randomly divide the CSI dataset into two sets. The CSI data in set 1 is collected by 11 volunteers, and the data in set 2 is collected by the other volunteers. In the first test, we utilize the data of set 1 to train the model and take the data of set 2 to test its performance. In the second test, the dataset used for training and validation is swapped. It is worth noting that we do not divide the CSI dataset into more groups or train the neural network for each volunteer in this experiment. Because the total number of CSI data collected by a volunteer (or several volunteers) is insufficient to train the neural network thoroughly. Too little training dataset may lead to severe overfitting. Table 5 lists the results. The model trained by set 1 achieves the accuracy of 95.0% for fall detection and 92.1% for activity recognition. If the model is trained by set 2, the accuracy of fall detection and activity recognition slightly decreases to 93.3% and 91.2%. DARMS still maintains the recognition accuracy at a high level when tested by untrained users.

Table 4: Cross-environment performance of DARMS.

	Fall detection	Activity recognition
Room A (train)	92.6%	88.5%
Room B (train)	90.8%	87.3%

Table 5: Cross-person performance of DARMS.

	Fall detection	Activity recognition
Set 1 (train)	95.0%	92.1%
Set 2 (train)	93.3%	91.2%

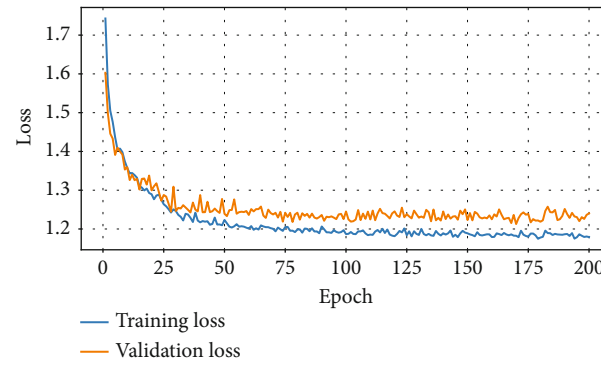


Figure 10: Curves of training loss and validation loss.

7.2.10. Loss Curve. Figure 10 depicts the training loss curve and the validation loss curve of 200 epochs. It can be observed that the validation loss hits the bottom and almost stops decreasing when the deep learning model has been trained around 150 epochs. Thus, in the evaluation of DARMS, we terminate the training when it reaches 150 epochs to avoid overfitting.

8. Conclusion

In this paper, we propose DARMS, a passive wireless sensing system only employs low-cost commodity WiFi devices. To achieve fine-grained activity recognition and monitoring, we carefully analyze the impact of human movement on CSI and design an effective signal processing method to extract the movement information both in the time domain and frequency domain. A novel neural network based on dual-channel transformer is proposed to combine substantial CSI features to improve the performance of human activity recognition. The experimental results demonstrate that DARMS can classify various human activities with high precision (this work was partially done when the third author was affiliated with Fudan University).

It is worth noting that DARMS still has some limitations. For the current prototype of DARMS, its performance has only been evaluated in different rooms with similar geometric setups of Tx-Rx pairs, while the accuracy across various geometric setups is another important criterion for the performance of cross-domain activity recognition. It will be our future work to design a system that can adapt to different geometric setups.

Data Availability

The experimental data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

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

Acknowledgments

This work is supported in part by Guangdong Provincial Key R&D Program under Grant No. 2020B010166003 and in part by Yiwu Research Institute, Fudan University.

References

[1] J. Yang, H. Zou, H. Jiang, and L. Xie, "Device-free occupant activity sensing using WiFi-enabled IoT devices for smart homes," *IEEE Internet of Things Journal*, vol. 5, no. 5, pp. 3991–4002, 2018.

[2] H. Zou, Y. Zhou, J. Yang, H. Jiang, L. Xie, and C. J. Spanos, "DeepSense: device-free human activity recognition via autoencoder long-term recurrent convolutional network," in *IEEE International Conference on Communications*, Kansas City, MO, USA, 2018.

[3] E. Cecchet, A. Acharya, T. Molom-Ochir, A. Trivedi, and P. Shenoy, "WiFiMon: a mobility analytics platform for building occupancy monitoring and contact tracing using WiFi sensing," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, New York, 2020.

[4] X. Zheng, J. Wang, L. Shangguan, Z. Zhou, and Y. Liu, "Design and implementation of a CSI-based ubiquitous smoking detection system," *IEEE/ACM Transactions on Networking*, vol. 25, no. 6, pp. 3781–3793, 2017.

[5] Z. Lin, S. Lyu, H. Cao et al., "HealthWalks: sensing fine-grained individual health condition via mobility data," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 4, 2020.

[6] Z. Wang, Z. Gu, J. Yin, Z. Chen, and Y. Xu, "Syncope detection in toilet environments using Wi-Fi channel state information," in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, New York, 2018.

[7] B. Korany, C. R. Karanam, H. Cai, and Y. Mostofi, "XModal-ID: using WiFi for through-wall person identification from candidate video footage," in *The 25th Annual International Conference on Mobile Computing and Networking*, New York, 2019.

[8] A. Trivedi, C. Zakaria, R. Balan, A. Becker, G. Corey, and P. Shenoy, "WiFitrace: network-based contact tracing for infectious diseases using passive WiFi sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 1, 2021.

[9] X. Guo, B. Liu, C. Shi, H. Liu, Y. Chen, and M. C. Chuah, "WiFi-enabled smart human dynamics monitoring," in *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*, New York, 2017.

[10] K. Niu, F. Zhang, J. Xiong, X. Li, E. Yi, and D. Zhang, "Boosting fine-grained activity sensing by embracing wireless multipath effects," in *Proceedings of the 14th International Conference on emerging Networking EXperiments and Technologie*, New York, 2018.

[11] Y. Wang, K. Wu, and L. M. Ni, "WiFall: device-free fall detection by wireless networks," *IEEE Transactions on Mobile Computing*, vol. 16, no. 2, pp. 581–594, 2017.

[12] S. Palipana, D. Rojas, P. Agrawal, and D. Pesch, "FallDeFi: ubiquitous fall detection using commodity Wi-Fi devices," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 4, pp. 1–25, 2018.

[13] Y. Hu, F. Zhang, C. Wu, B. Wang, and K. J. R. Liu, "DeFall: environment-independent passive fall detection using WiFi," *IEEE Internet of Things Journal*, vol. 9, no. 11, pp. 8515–8530, 2022.

[14] H. Yan, Y. Zhang, Y. Wang, and K. Xu, "WiAct: a passive WiFi-based human activity recognition system," *IEEE Sensors Journal*, vol. 20, no. 1, pp. 296–305, 2020.

[15] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Understanding and modeling of WiFi signal based human activity recognition," in *Proceedings of the 21st annual international conference on mobile computing and networking*, pp. 65–76, New York, 2015.

[16] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Device-Free human activity recognition using commercial WiFi devices," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 5, pp. 1118–1131, 2017.

[17] Y. Zheng, Y. Zhang, K. Qian et al., "Zero-effort cross-domain gesture recognition with Wi-Fi," in *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, New York, 2019.

[18] S. Yao, A. Piao, W. Jiang et al., "Stfnets: learning sensing signals from the time-frequency perspective with short-time fourier neural networks," in *The World Wide Web Conference*, New York, 2019.

[19] S. Ding, Z. Chen, T. Zheng, and J. Luo, "RF-net: a unified meta-learning framework for RF-enabled one-shot human activity recognition," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, New York, 2020.

[20] Y. Tian, G. H. Lee, H. He, C. Y. Hsu, and D. Katabi, "RF-based fall monitoring using convolutional neural networks," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, 2018.

[21] Z. Chen, L. Zhang, C. Jiang, Z. Cao, and W. Cui, "WiFi CSI based passive human activity recognition using attention based BLSTM," *IEEE Transactions on Mobile Computing*, vol. 18, no. 11, pp. 2714–2724, 2019.

[22] B. Li, W. Cui, W. Wang, L. Zhang, Z. Chen, and M. Wu, "Two-stream convolution augmented transformer for human activity recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 1, pp. 286–293, 2021.

[23] H. Choi, M. Fujimoto, T. Matsui, S. Misaki, and K. Yasumoto, "Wi-CaL: WiFi sensing and machine learning based device-

free crowd counting and localization," *IEEE Access*, vol. 10, pp. 24395–24410, 2022.

[24] F. Wang, W. Gong, J. Liu, and K. Wu, "Channel selective activity recognition with WiFi: a deep learning approach exploring wideband information," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 1, pp. 181–192, 2020.

[25] M. Zhou, Y. Lin, N. Zhao, Q. Jiang, X. Yang, and Z. Tian, "Indoor WLAN intelligent target intrusion sensing using ray-aided generative adversarial network," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 4, no. 1, pp. 61–73, 2020.

[26] C. Yu, Z. Xu, K. Yan, Y.-R. Chien, S.-H. Fang, and H.-C. Wu, "Noninvasive human activity recognition using millimeter-wave radar," in *IEEE Systems Journal*, pp. 1–12, IEEE, 2022.

[27] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Tool release: gathering 802.11n traces with channel state information," *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 1, 2011.

[28] Z. Gu, T. He, J. Yin, Y. Xu, and J. Wu, "TyrLoc: a low-cost multi-technology MIMO localization system with a single RF chain," in *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, pp. 228–240, New York, 2021.

[29] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[30] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," 2016, http://arxiv.org/abs/1606.08415.

[31] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, http://arxiv.org/abs/1607.06450.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, Nevada, 2016.

[33] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," 2014, http://arxiv.org/abs/1412.6980.

[34] Z. Liu, Y. Lin, Y. Cao et al., "Swin transformer: hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[35] R. W. Schafer, "What is a Savitzky-Golay filter?," *IEEE Signal Processing Magazine*, vol. 28, no. 4, pp. 111–117, 2011.

[36] D. Wu, D. Zhang, C. Xu, Y. Wang, and H. Wang, "WiDir: walking direction estimation using wireless signals," in *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing*, New York, 2016.

[37] J. Y. Chang, K. Y. Lee, Y. L. Wei, K. C. Lin, and W. Hsu, "Location-independent WiFi action recognition via vision-based methods," in *Proceedings of the 24th ACM international conference on Multimedia*, New York, 2016.

[38] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.