WiFi-Based Indoor Human Activity Sensing: A Selective Sensing Strategy and A Multi-Level Feature Fusion Approach

Yiyun Zhang, Gongpu Wang, Heng Liu, Wei Gong, and Feifei Gao, *Fellow, IEEE*

***Abstract*—Utilizing communication signals for indoor human activity recognition (HAS) is an important component of integrated sensing and communication (ISAC). The current majority HAS solutions adopt a single sensing strategy and only work in a simple environment. In this paper, we propose a new HAS method named WiSMLF that can flexibly select multiple sensing strategies and then use multi-level feature fusion for sensing. We first use the high frequency energy (HFE) method to categorize human activities into two types: static activities (SAs) and moving activities (MAs). Subsequently, for SAs, we adopt a joint localization and activity recognition sensing strategy, and use a multi-level feature fusion network based on visual geometry group (VGG). For MAs, we adopt a joint activity recognition and moving distance estimation sensing strategy, and use a multi-level feature fusion network based on long short-term memory (LSTM). The experimental results show that WiSMLF outperforms the existing methods especially in complex environments, and can obtain 92% higher accuracy in location, activity recognition, and distance estimation.**

***Index Terms*—Activity recognition, indoor localization, movement distance estimation, multi-level feature fusion, WiFi**

I. Introduction

Integrated sensing and communication (ISAC) is expected to become a key technology for the sixth generation (6G) of mobile communications [1], [2]. The basic idea of WiFi-based ISAC is to utilize WiFi signals to sense the location of objects and human activities in the real world. WiFi-based Indoor human activity sensing (HAS) has a wide range of real world applications, such as smart homes [3], health monitoring [4], and rescue systems [5]. Traditional HAS technologies typically employ vision based approach [6] and contact wearable device [7]. The vision based HAS usually relies on camera and have high recognition accuracy for HAS. However, for non-line-of-sight (NLOS) scenarios, the camera is highly susceptible to being affected by an obstructed view of the subjects. The contact wearable device approach relies on the user wearing sensors, which offers the advantages of continuous monitoring and multifunctionality. However, the contact-based wearable device intensifies the physical load and may be cumbersome for users.

In recent years, several RF-based [8], [9] HAS solutions, such as radars technology [10], radio frequency identification (RFID) [11], and WiFi technology [12], have been developed to overcome such constraints. Radar sensing is very stable and resistant to interference but it is costly to deploy and has limited coverage. RFID sensing technology relies on sticky notes and readers, which makes the coverage too small. In contrast, WiFi wireless networks are widely deployed without additional overhead, and WiFi-based HAS could protect the privacy of users.

Lately, a series of research efforts have been conducted for WiFi-based indoor localization [13], [14] and activity recognition [15], [16]. The authors in [17] designed the WiGEM system that measures the received signal strength indicator (RSSI) of the target at multiple access points (APs) and combines them using triangulation and propagation models for target localization. The work in [18] enhanced the granularity of RSSI values by employing software-defined radio, and achieves a 72% increase in accuracy across four different methods for activity recognition. Though, the system in [18] is easy to deploy, it suffers from significant localization errors due to the limited RSSI resolution offered by commercial devices.

HAS based on channel state information (CSI) information have been proposed, owing to the detailed measurement data obtained from the physical layer of WiFi devices through CSI [19], [20]. In [21], the authors demonstrated the initial potential of utilizing CSI amplitude features for activity recognition. In [22] presented CARM, a system for human activity recognition and monitoring based on CSI, with experimental outcomes showing that CARM attains a notable recognition

This work was supported in part by National Key R&D Program of China and Shandong Province, China (2021YFB3901300), by the National Natural Science Foundation of China under Grant U23A20272, by the Natural Science Foundation of China (62221001 & U22B2004). *(Corresponding author: Gongpu Wang.)*

Yiyun Zhang is with the Beijing Key Laboratory of Transportation Data Analysis and Mining, School of Computer Science and Technology, Beijing Jiaotong University, Beijing 100044, China (e-mail: yiyunzhang@bjtu.edu.cn).

Gongpu Wang is with the Beijing Key Laboratory of Transportation Data Analysis and Mining, School of Computer Science and Technology, Beijing Jiaotong University, Beijing 100044, China, and also with the China (Wuxi) Institute of nternet of Things, Wuxi 214111, China (e-mail: gpwang@bjtu.edu.cn).

Heng Liu is with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China (email: heng_liu_bit_ee@163.com).

Wei Gong is with the School of Computer Science and Technology and the School of Data Science, University of Science and Technology of China, Hefei 230027, China (e-mail: weigong@ustc.edu.cn).

Feifei Gao is with the Institute for Artificial Intelligence Tsinghua University (THUAI), State Key Laboratory of Intelligent Technologies and Systems, Beijing National Research Center for Information Science and Technology (BNRist), Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: feifeigao@ieee.org).

accuracy of 96% and exhibits strong resilience across different environmental conditions. The authors of [23] presented DeepFi, a novel deep-learning-based indoor fingerprinting system utilizing CSI data.

With the development of artificial intelligence (AI) and deep learning (DL), many neural network (NN) backbone architectures have been proposed, including residual networks (ResNet) [24] and visual geometry group (VGG) [25]. They were proved to be robust and effective in various tasks [26], [27], including image classification [28]–[30], similar to WiFi-based localization and activity recognition. This has inspired researchers to address WiFi-based indoor localization and activity recognition problems using neural networks. The authors of [31] proposed a DL framework utilizing a convolutional neural network (CNN) to recognize sign gestures via CSI. The authors of [32] initially presented a method based on long short-term memory (LSTM) networks to preserve the temporal state data of activities, assisting in the extraction of nuanced features for akin activities. Moreover, [33] proposed a comprehensive framework that leverages the synergistic capabilities of CNN and LSTM to achieve the classification of actions.

In light of the preceding analysis, WiFi-based HAS technology demonstrates strong performance in activity recognition and indoor localization tasks. Nevertheless, the varying significance of the same activity at different locations emphasizes the importance of joint tasks. In [34], the authors suggested a dual-task convolutional neural network featuring one-dimensional convolutional layers, aimed at simultaneously handling activity recognition and indoor localization. This approach resulted in achieving accuracy of 95.68% for localization and 88.13% for activity recognition tasks. However, [34] relied only on amplitude data, and there is a need to enhance recognition accuracy.

The authors in [35] proposed to use multi-view fusion strategy for sensing, and use multiple antenna phase and amplitude information for fusion, joint sensing localization and activity recognition. The findings indicate that the precision of activity recognition and indoor localization surpassed 94.38% and 95.68%, respectively. However, it only recognizes activities for stationary positions and cannot identify mobile activities such as walking. The authors of [36] utilized support vector machine (SVM) techniques for the joint tasks of localization and activity recognition. Experimental results indicate an accuracy of up to 90%. The method in [36] can recognize and locate all types of activities, but it employs a single perceptual strategy for different types of activities.

HAS still faces several practical challenges. For example, the shortcomings of model learning in more complex real world environments and the limitations of a single perceptual strategy. Thus, in this paper, we proposed a new HAS method named WiSMLF, which used CSI extracted from 802.11n. WiSMLF can flexibly select multiple sensing strategies and then use multi-level feature fusion approach for sensing. The experimental results show that WiSMLF outperforms the existing methods especially in complex environments, and can obtain 92% higher accuracy in location, activity recognition, and distance estimation. The major contributions of this paper

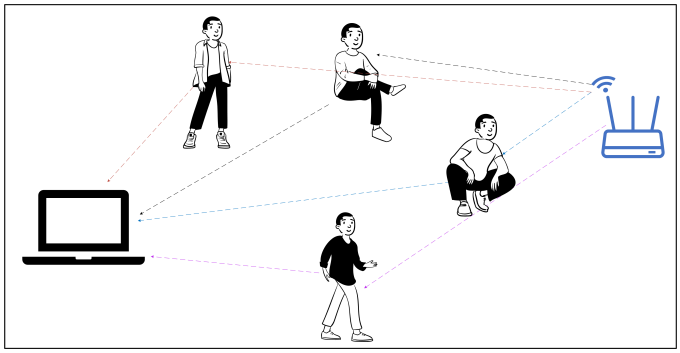

Fig. 1. Signal propagation models for different human activities.

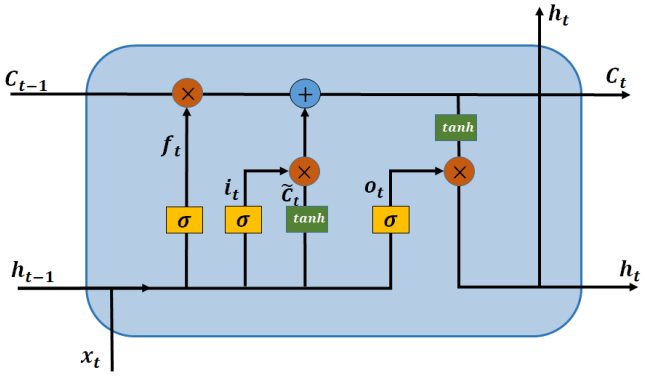

Fig. 2. LSTM network structure.

are listed as follows:

- The proposed WiSMLF leverages high frequency energy (HFE) information to distinguish between MAs and SAs, enabling the design of distinct sensing strategies for each category. This adaptability to various scenarios in real world indoor environments enhances the versatility and robustness of the sensing approach.
- The proposed WiSMLF introduces multi-level feature fusion in WiFi sensing, which allows us to extract richer feature information from the data, thereby maintaining high accuracy when dealing with more complex environmental datasets.
- The proposed WiSMLF is tested on three field-collected datasets with different scene scenarios, all of which demonstrated significant improvements in testing accuracy.

The rest of this paper is organized as follows. In Section II, we provide a brief preliminary work. Section III describes the system architecture in detail. The experimental performance evaluation is provided in Section IV. Finally, we conclude the paper in Section V.

II. Preliminary Work

A. Channel State Information

During the transmission process from the signal transmission end to the signal reception end, human activities and environmental interference often cause signal scattering, environmental attenuation, and distance fading effects, as

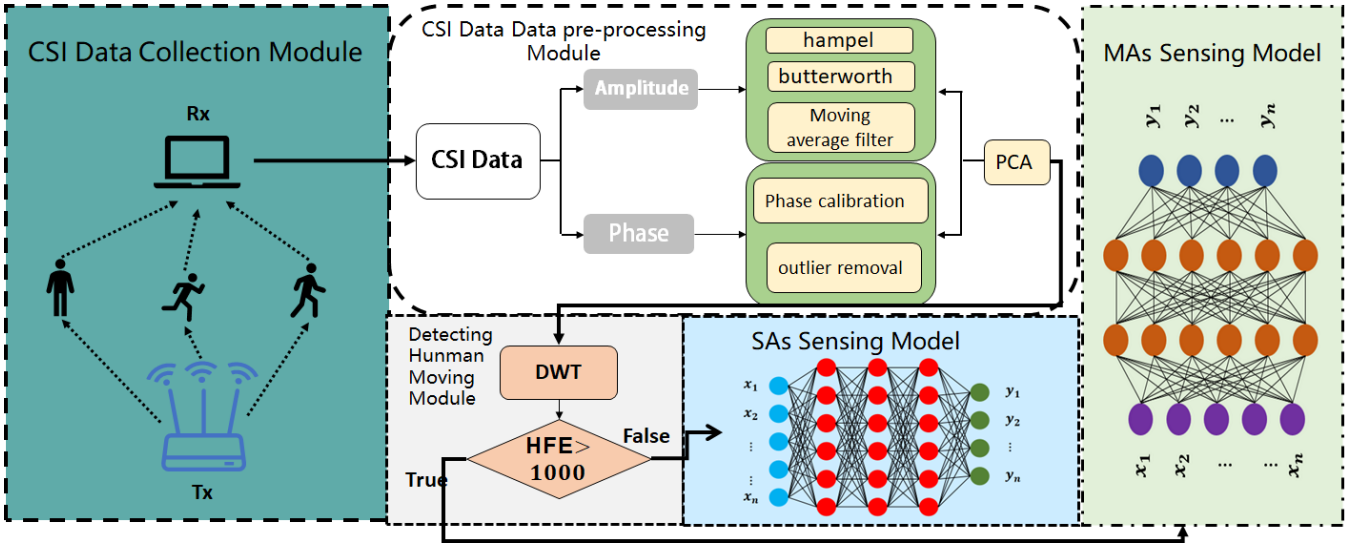

Fig. 3. WiSMLF architecture. WiSMLF architecture consists of four components: data pre-processing module, moving detection module, SAs sensing module and MAs sensing module

shown in Fig. 1. According to IEEE 802.11n standard [37], select commercial WiFi network interface controllers (NICs) have the capability to continuously observe wireless channels. Given that CSI offers subcarrier-level information and enables the accurate capture of signal attenuation, WiSMLF employs CSI to sense human activities.

Assume that the transmitter (Tx) has N_T antennas and the receiver (Rx) has N_R antennas. The transmission is modeled as

$$\boldsymbol{y}_{i,j} = \boldsymbol{H}_{i,j}\boldsymbol{x}_{i,j} + \boldsymbol{n}_{i,j} \tag{1}$$

where $\boldsymbol{x}_{i,j}$ and $\boldsymbol{y}_{i,j}$ denote the transmit signal vector and the receive signal vector, respectively. Moreover, $\boldsymbol{n}_{i,j}$ is the Gaussian white noise from the ith antenna of transmitter T_x to the jth antenna of receiver R_x, and $\boldsymbol{H}_{i,j}$ is the channel matrix between Tx and Rx. Additionally, the estimation of $\boldsymbol{H}_{i,j}$ for N_s subcarriers can be represented as

$$\boldsymbol{H}_{i,j} = [H_{i,j,1}, H_{i,j,2}, \cdots, H_{i,j,N_s}]^T \tag{2}$$

where $H_{i,j,k}$ denotes the CSI of the kth subcarrier. The kth subcarrier in $\boldsymbol{H}_{i,j}$ can be represented as

$$H_{i,j,k} = |H_{i,j,k}|e^{j\angle H_{i,j,k}} \tag{3}$$

where $|H_{i,j,k}|$ and $\angle H_{i,j,k}$ denote the amplitude and phase of the kth subcarrier, respectively.

B. Long Short-Term Memory

The LSTM is a specialized type of recurrent neural network (RNN) that has proven particularly effective in analyzing time series data. Consequently, in recent times, the LSTM model has seen extensive application in HAS.

The LSTM architecture incorporates a gated output approach, which comprises three gates and the LSTM is updated at time t based on its input parameters, as follows

$$\begin{aligned} i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\ \widetilde{C_t} &= tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \\ o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ C_t &= f_t \times C_{t-1} + i_t \times \widetilde{C_t} \\ h_t &= o_t \cdot \tanh(C_t) \end{aligned} \tag{4}$$

where W_f, W_i, W_C, W_o is the weight term, b_f, b_i, b_C, b_o is the bias term, $\sigma(\cdot)$ is the Sigmoid function, $\tanh(\cdot)$ is the tanh function, h_t represents the cell output activation vector, while i_t, f_t, $\widetilde{C_t}$, o_t, C_t are input gates, forget gates, output gates, input modulation gates, and memory gates, respectively. In LSTM networks, various gates manage distinct data flows. The input gate primarily regulates the quantity of new data utilized by the current memory cell, the forget gate chiefly decides what information should be omitted, and the output gate principally governs the volume of information released from the current storage unit. The structure of LSTM is shown in Fig. 2.

III. System Design

The proposed WiSMLF shown Fig. 3 includes 5 modules: the data collection module, the CSI pre-processing module, the human movement detection module, the SAs sensing module, and the MAs sensing module.

A. Data Collection Module

In this module, we collect WiFi CSI. Using a network card equipped with three antennas and a router with two antennas, we are able to obtain 6 CSI streams. Additionally, we utilize CSITOOL to extract CSI data, with each stream containing data from 30 subcarriers. Consequently, we extract 180 sets

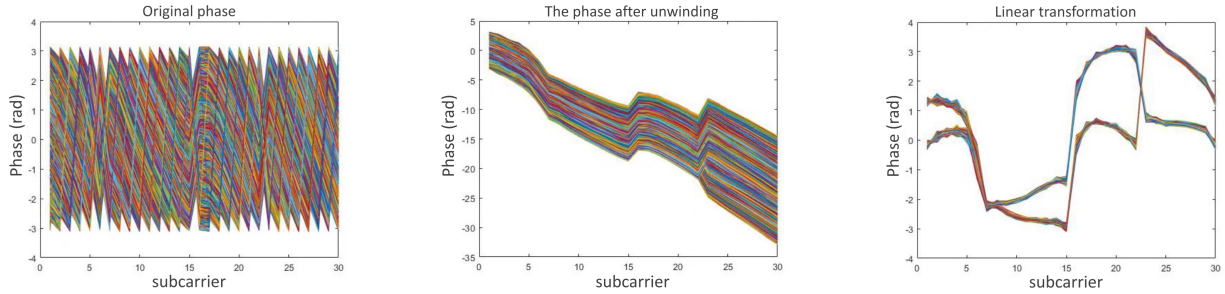

Fig. 4. The phase value of the antenna. (a) Original phase. (b) The phase after unwinding. (c) Phase after linear processing.

of CSI values ($2 \times 3 \times 30$) from each data packet, and the matrix representing the ith receiving antenna is expressed as

$$\boldsymbol{H}_i = \begin{bmatrix} H_{1,1} & H_{1,2} & H_{1,3} & \cdots & H_{1,30} \\ H_{2,1} & H_{2,2} & H_{2,3} & \cdots & H_{2,30} \\ H_{3,1} & H_{3,2} & H_{3,3} & \cdots & H_{3,30} \end{bmatrix}. \tag{5}$$

B. CSI Pre-Processing Module

In this module, we handle the processing of gathered CSI amplitude and phase information.

1) Phase Processing: The phase component of CSI is more vulnerable to variations in the environment compared to the CSI amplitude. Nonetheless, in real-world settings, the acquired CSI phase is frequently altered due to factors such as carrier frequency offset (CFO), sampling frequency offset (SFO), and phase detection delay (PDD) [38]. Taking into account these three types of offsets, the observed phase for the kth subcarrier can be depicted as

$$\hat{\varphi_k} = \varphi_k - 2\pi \frac{\alpha_k}{N}\delta + \beta + Z \tag{6}$$

where φ_k is the true phase information, δ is the timing offset of the receiver, β is the unknown phase offset, and Z is the phase measurement noise. Moreover, α_k denotes the subcarrier index (the range in IEEE 802.11n is $[-28, 28]$), and N is the FFT size. The specific parameter representation of linear phase transformation can be calculated as

$$\begin{aligned} a &= \frac{\hat{\varphi_n} - \hat{\varphi_1}}{\alpha_n - \alpha_1} = \frac{\varphi_n - \varphi_1}{\alpha_n - \alpha_1} - \frac{2\pi}{N}\delta \\ b &= \frac{1}{n}\sum_{1\le j\le n} \hat{\varphi_j} = \frac{1}{n}\sum_{1\le j\le n} \varphi_j - \frac{2\pi\delta}{nN}\sum_{1\le j\le n} \alpha_j + \beta \end{aligned} \tag{7}$$

where n represents the total number of subcarriers received by the device. According to IEEE 802.11n specifications, the download waveforms exhibit symmetry in a 40MHz bandwidth, indicating $\sum_{j=1}^{n} \alpha_j = 0$. Consequently, there holds $c = \frac{1}{n}\sum_{1\le j\le n} \varphi_j + \beta$, which allows for the derivation of a linear combination of the actual phase by subtracting $\tilde{\varphi}$ from the measured phase $\hat{\varphi}$, as

$$\tilde{\varphi_i} = \hat{\varphi_i} - b\alpha_i - c = \varphi_i - \frac{\varphi_n - \varphi_1}{\alpha_n - \alpha_1}\alpha_k - \frac{1}{n}\sum_{j=1}\varphi_j. \tag{8}$$

Since the collected CSI signal exhibits cyclic behavior, resulting in folded CSI phases, it is necessary to unwrap the CSI phases before extraction, and apply linear transformation to it, as shown in Fig. 4.

2) Outlier Removal and Filtering: In the process of WiFi signal propagation, CSI contains both useful information and interference, such as noise. Therefore, data preprocessing is essential to minimize interference.

First, we use Hample filter to detect and remove outliers. The expressions for the Hampel filter are as follows

$$\begin{aligned} &\text{MAD} = \text{Median}(|X(i) - \text{Median}(X)|) \\ &|X(i) - \text{Median}(X)| > \beta \times \text{MAD} \end{aligned} \tag{9}$$

where $\text{Median}(X)$ denotes the median of dataset X, MAD denotes median absolute deviation, and β represents the standard threshold.

Second, because human activities are primarily found in the spectrum of lower frequencies, while noise predominantly occupies the high-frequency spectrum, we apply a Butterworth filter to eliminate noise.

Given that the data is a CSI time series, we further apply a moving average filter to reduce noise in the data.

3) Data Normalization: Since each CSI subcarrier has a different range, we use the Z-cores method to normalize the data, which can speed up the convergence of the gradient descent based model and even improve the accuracy of the model.

4) Principal Component Analysis: We utilize principal component analysis (PCA) to reduce the substantial redundancy in data generated by multiple subcarriers. This technique helps in retaining the most significant features of the high-dimensional data while eliminating noisy and unimportant features. Thus, the data processing speed can be enhanced effectively.

C. Detecting Hunman Moving Module

To enable the selection of suitable sensing strategies, the initial step involves detecting the type of subject movement (judging the motion is SAs or MAs). In particular, the HFE of CSI is employed to identify human movement. To obtain HFE, it's necessary to compute the coefficients of the discrete wavelet transform (DWT), which can be formulated as follows

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3397708

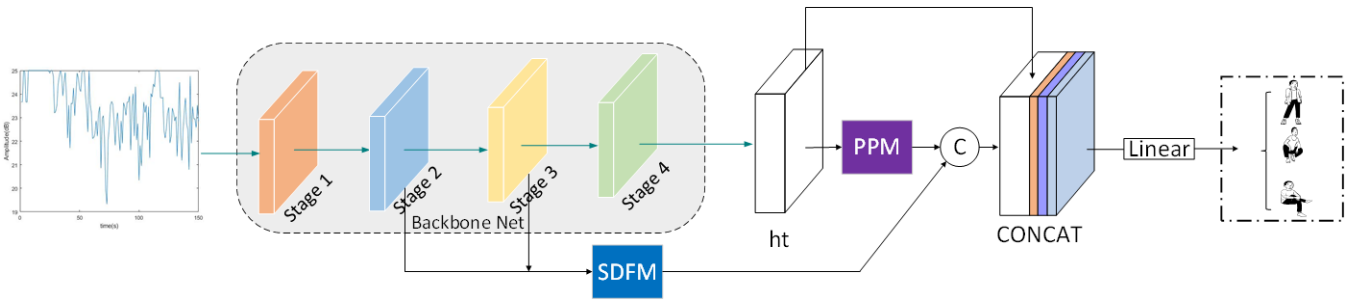

Fig. 5. SAs sensing module.

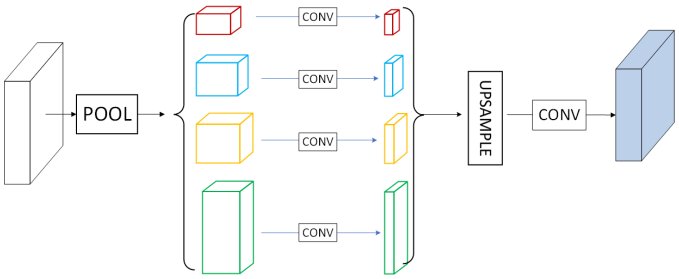

Fig. 6. Pyramid pooling module.

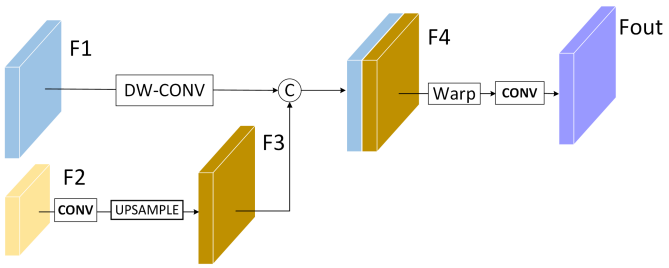

Fig. 7. Supplementary detail feature fusion module.

$$x_{\alpha,L}[n] = \sum_{k=0}^{K-1} x_{\alpha-1,L}[2n-k]g[k]$$
$$x_{\alpha,H}[n] = \sum_{k=0}^{K-1} x_{\alpha-1,L}[2n-k]h[k] \quad (10)$$

where $x[n]$ is the discrete input signal, $x_{\alpha,L}[n]$ and $x_{\alpha,H}[n]$ are the approximation and detail coefficients, respectively. Moreover, $g[n]$ and $h[n]$ are the low-pass filter and high-pass filter (db1 wavelet basis), respectively. Therefore, the HFE is

$$HFE = \sum_{k=1}^{K-1} (x_{\alpha,H}[n])^2. \quad (11)$$

For SAs, the CSI amplitude stays at a low level due to the stable propagation path, causing a small HFE. When the motion becomes MAs, the rapid change in the propagation path can lead to the increase in CSI amplitude, causing a large HFE.

D. SAs Sensing Module

The SAs sensing module illustrated in Fig. 5 is then described to enhance the performance of human activity and location recognition in complex scenarios sensing. The model comprises backbone network, fusion network and classification network. The backbone network utilizes an enhanced VGG16 architecture for efficient feature extraction. In this paper, we divide the feature extraction process of the backbone network into four stages. The structure of the backbone network is detailed in Tab. I, where H and W denote the antenna and subcarrier in the input data, conv denotes the convolution, BN is the batch normalization operation, and ReLU is the activation function.

1) Backbone Network: The multi-level feature fusion primarily comprises a micro-feature branch, a detail-supplementary branch, and an ambient information branch. The output of stage 4 is utilized as the input for the ambient information branch, employing pyramid pooling to aggregate environmental information across different scales. In complex environments, indoor location and activity recognition necessitates sufficient micro-features to capture subtle changes. Thus, the output of stage 1 is designated as the input to the micro-feature branch. The intermediate stages (stage 2 and stage 3) serve as inputs to the detail-supplementary branch, further capturing additional informative features. After refinement through detail supplementation and micro-feature fusion, the feature contains a wealth of precise details and ambient information. The final feature is then input into the classification network for categorization.

2) Ambient Information Branch: In real world scenarios, environments exhibit a complexity that necessitates the acquisition of ambient information for precise location and human activity recognition. To address this problem, we fuse features across different dimensions, associating local features with global features of the environment. Utilizing the pyramid pooling module (PPM) [39], we take into account both minute targets and macroscopic overall objectives by employing global average pooling at various scales to enrich global and local features, thereby enhancing environmental information. Adopting ambient information explicitly integrates features of the target across various scales, effectively mitigating the impact of complex environments on feature extraction. The PPM module utilizes adaptive global pooling with output sizes of 1, 2, 3 and 5 to extract both global and local information. Subsequently, features are down-sampled using 1×1 convo-

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3397708

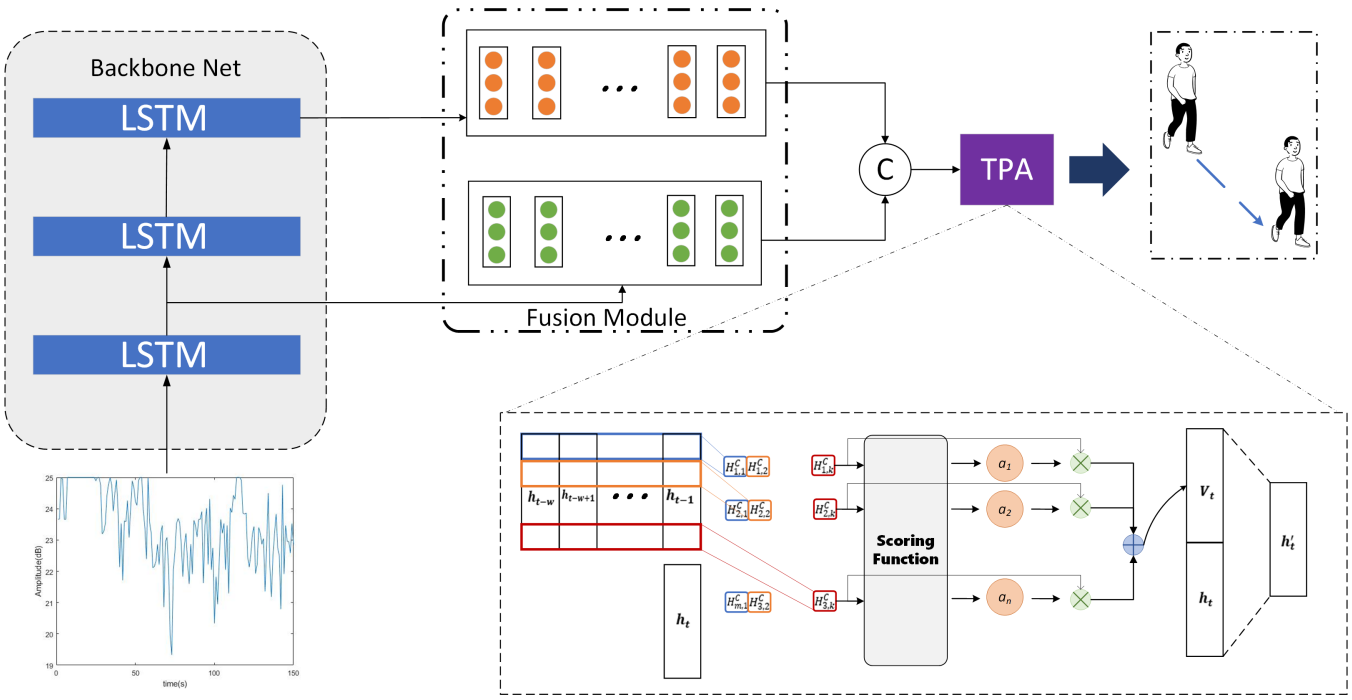

Fig. 8. MAs sensing module.

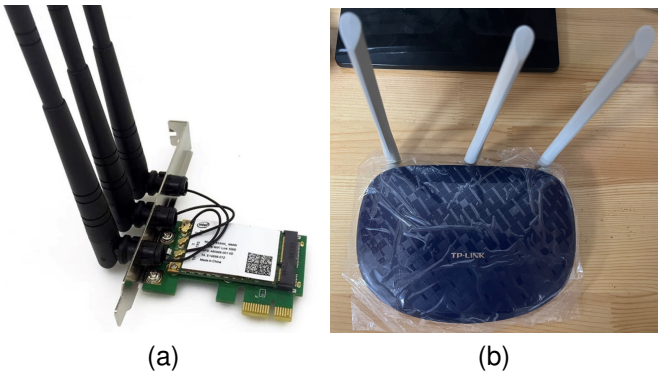

Fig. 9. Experimental equipment. (a) Intel 5300 NIC. (b) TP-Link router.

lution followed by bilinear interpolation for up-sampling to match the size of the original input feature. The input feature is fused with the feature possessing global and local features, and finally processed through conv to obtain feature rich in ambient information. The PPM network structure adopted in this paper is illustrated in Fig. 6.

3) Feature Fusion Module: Within the realm of HAS in complex environments, two key factors should be considered: spatial detail information and category detail information. Each of these facets carries its own set of limitations. Deep network layers excel in extracting a wealth of ambient information but notably lose fine-grained details. Conversely, shallow network layers are adept at capturing detailed information, yet they fall short in understanding the signal variations due to changes in the environment.

Therefore, we propose a supplementary detail feature fusion module (SDFM), as depicted in Fig. 7. This module efficiently integrates higher-level features into lower-level counterparts by establishing spatial relationships across different hierarchical features. Specifically, SDFM takes two distinct hierarchical features, F_1 and F_2, as inputs. We employs a conv 1×1 to reduce the dimensionality of F_2, followed by an upsample to match the size of F_1, resulting in feature F_3. Due to the necessity of establishing feature space relationships, a depthwise separable conv 1×1 is applied to F_1 to enhance the independence of features. Subsequently, F_1 and F_3 are fused. The fused features are then subjected to a warp operation. Finally, a conv 3×3 is applied to amalgamate the features, yielding an output feature F_{out}.

TABLE I
SAs Sensing Backbone Network Structure

Input size	Stage	Layer structure	Up/down sampling	Output Size
$H \times W \times 3$	Stage 1	Conv3 $\times$ 3 BN-ReLU	2	$\frac{H}{2} \times \frac{W}{2} \times 128$
	Stage 2	Conv3 $\times$ 3 BN-ReLU	2	$\frac{H}{4} \times \frac{W}{4} \times 256$
	Stage 3	Conv3 $\times$ 3 BN-ReLU	2	$\frac{H}{8} \times \frac{W}{8} \times 512$
	Stage 4	Dilated-Conv2 $\times$ 2 BN-ReLU	1	$\frac{3H}{32} \times \frac{3W}{32} \times 256$

E. MAs Sensing Module

For MAs sensing module, we adopt a joint activity recognition and movement distance estimation sensing strategy. The MAs sensing module consists of a backbone network

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3397708

Fig. 10. Three different scenarios. (a) Meeting room scenario. (b) Laboratory scenario. (c) Bedroom scenario.

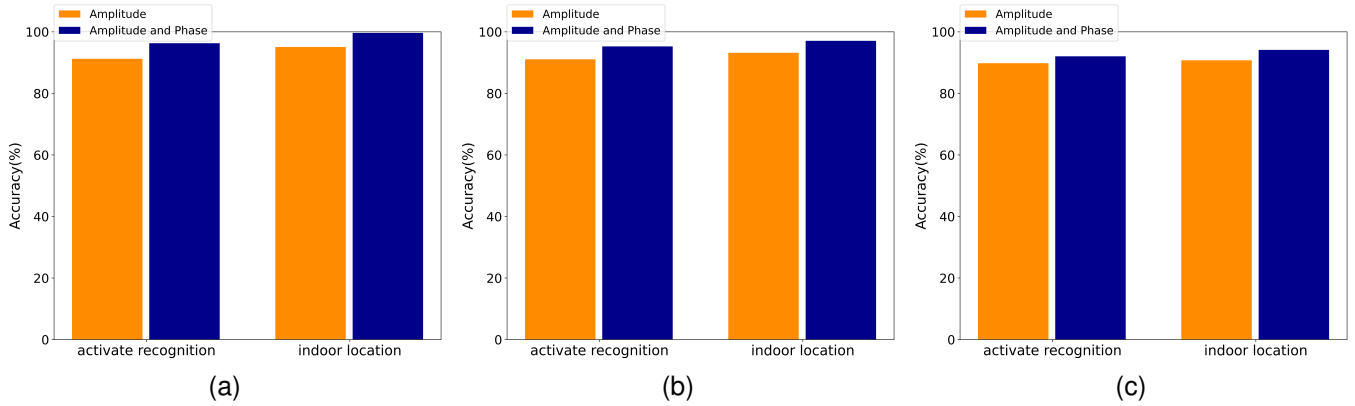

Fig. 11. The accuracy of WiSMLF using amplitude data alone and both amplitude and phase data in these three scenarios for SAs. (a) Meeting room scenario. (b) Laboratory scenario. (c) Bedroom scenario.

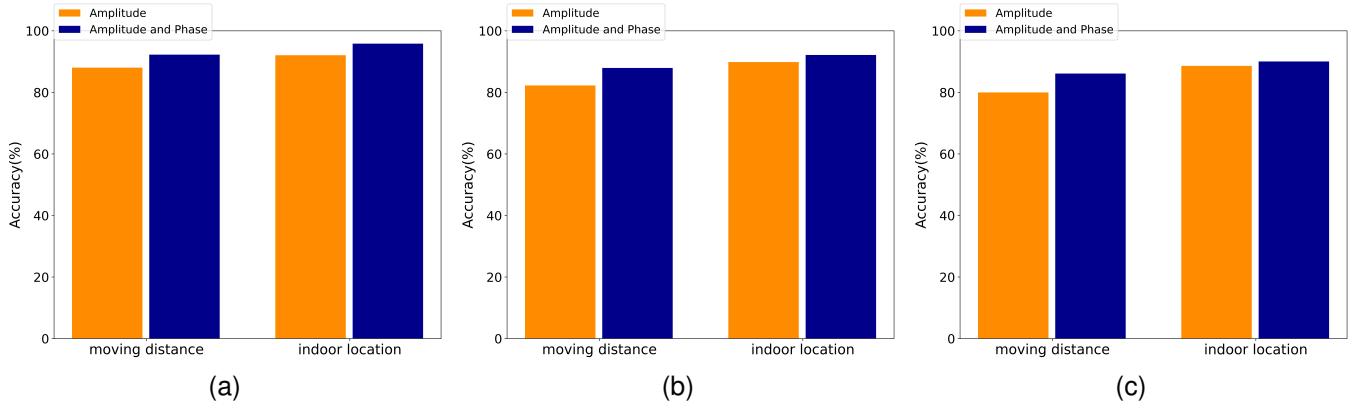

Fig. 12. The accuracy of WiSMLF using amplitude data alone and both amplitude and phase data in these three scenarios for MAs. (a) Meeting room scenario. (b) Laboratory scenario. (c) Bedroom scenario.

module, a fusion network module, and an attention module. The backbone network employs an LSTM structure to achieve efficient feature extraction from time series data. In this paper, we divide the feature extraction process of the backbone network into three stages. The backbone network structure is shown in Tab. II., where W, T, and H denote the number of samples, time steps, and feature numbers in the input data, respectively, with Than representing the activation function.

1) Backbone Network Module: In the field of HAS, certain activities exhibit close temporal relationships, resulting in two key aspects: high-dimensional temporal features and low-dimensional temporal features. Each dimension has its inherent limitations. Deep network layers are adept at extracting complex high-dimensional temporal features, capturing a wealth of details. However, deeper network layers are also more susceptible to environmental noise. In contrast, shallow network layers excel at capturing simpler low-dimensional

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3397708

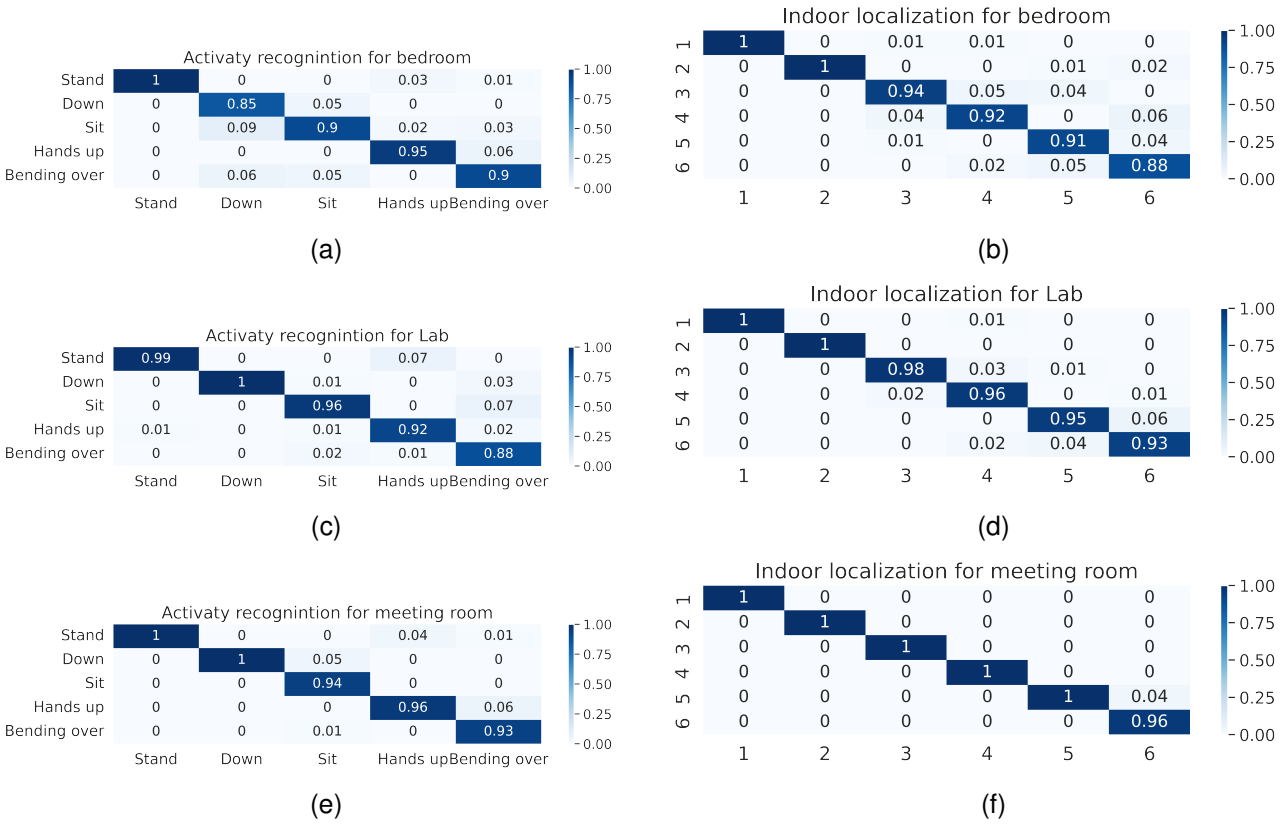

Fig. 13. Normalized confusion matrix for multi-level feature fusion for activity recognition and indoor localization. (a) Normalized confusion matrix for activity recognition in bedroom. (b) Normalized confusion matrix for indoor localization in bedroom. (c) Normalized confusion matrix for activity recognition in Lab. (d) Normalized confusion matrix for indoor localization in Lab. (e) Normalized confusion matrix for activity recognition in meeting room. (f) Confusion matrix for indoor localization in meeting room.

temporal features and exhibit robustness to changes in the environment, but they may overlook some complexities within the data.

To address this challenge, we have adopted a multi-level feature fusion strategy. Within the existing LSTM network framework, we seamlessly integrate the extraction of high-dimensional temporal features.

2) Time Series Feature Fusion Network Module: We concatenate the high-dimensional features from the third layer with the low-dimensional features from the first layer to form a new feature vector, thereby preserving the rich information of the high-dimensional features while also utilizing the stability and robustness of the low-dimensional features.

3) Temporal Pattern Attention Module: the human movement distance estimation, which is closely linked to time series analysis. We integrate the temporal pattern attention (TPA) mechanism [40] to enhance the ability to capture long-term dependencies of the model. The mechanism assists the model in focusing on the most relevant time steps, thereby facilitating a deeper understanding of the time series. The TPA model, as illustrated in Fig. 8. First, we process the feature sequence to obtain a fused hidden state h_i (a column vector) for each time step, with each h_i having a dimension of m. Here, w represents the length of the time window, indicating the selected range of historical data, resulting in a hidden state matrix $H = h_t - w, h_t - w + 1, ..., h_t - 1$. Then, we employs

TABLE II
MAS SENSING BACKBONE NETWORK STRUCTURE

Input size	Layer	LSTM Layer structure	Activation Function	Output Size
$W \times T \times H$	layer 1	32 hidden nods	Tanh	$W \times T \times 32$
	layer 2	64 hidden nods	Tanh	$W \times T \times 64$
	layer 3	32 hidden nods	Tanh	$W \times T \times 32$

k filters to derive the temporal pattern information from each distinct variable, as follow:

$$H_{i,j}^{C} = \sum_{l=1}^{w} H_{i,t-w-1+l} \times C_{j,T-w+l} \tag{12}$$

where the size of the convolution kernel is $1 \times T$, $(T = w)$, H^C is the temporal pattern matrix of the variable within the range of this convolution kernel, and $H_{i,j}^{C}$ represents the result value of the interaction between the ith row vector and the jth convolution kernel.

Subsequently, we compute a weighted sum of the row

TABLE III
EVALUATION OF FIVE DIFFERENT ACTIVITY RECOGNITION METHODS ON THREE DATASETS.

	MLP			ResNet_18			SVM			RF			WiSMLF		
	Bedroom	Lab	Meeting room	Bedroom	Lab	Meeting room	Bedroom	Lab	Meeting room	Bedroom	Lab	Meeting room	Bedroom	Lab	Meeting room
Stand	75.32	79.16	84.27	90.17	92.37	93.58	73.28	82.13	85.79	70.29	75.58	83.77	91.38	95.74	98.93
Down	71.93	76.27	83.15	89.01	90.67	92.93	71.29	81.33	84.12	69.07	71.81	79.85	90.37	94.29	95.08
Sit	72.15	76.27	84.53	89.25	91.54	93.54	70.53	83.25	85.07	70.88	75.29	81.34	90.13	95.03	97.63
Hands up	75.88	76.35	82.17	87.92	89.01	91.39	71.59	78.11	80.76	68.19	73.55	80.89	88.07	91.36	95.27
Bending over	74.11	78.53	81.37	88.17	90.53	92.76	70.07	80.54	82.59	68.95	74.58	82.81	89.92	93.18	93.17

TABLE IV
EVALUATION OF FIVE DIFFERENT INDOOR POSITIONING METHODS ON THREE DATASETS.

	MLP			ResNet_18			SVM			RF			WiSMLF		
	Bedroom	Lab	Meeting room	Bedroom	Lab	Meeting room	Bedroom	Lab	Meeting room	Bedroom	Lab	Meeting room	Bedroom	Lab	Meeting room
Location_1	82.47	87.66	95.26	90.87	95.54	96.66	83.26	86.39	92.11	85.67	89.92	96.27	95.77	98.15	100
Location_2	82.28	86.53	96.13	89.22	94.16	95.51	84.15	86.58	92.92	87.35	88.11	97.81	96.89	98.21	100
Location_3	82.35	86.53	94.01	89.15	94.27	94.39	83.33	84.71	93.78	85.21	89.58	96.39	95.12	97.03	99.92
Location_4	79.53	83.27	92.39	88.07	93.08	93.17	82.58	84.92	92.69	83.13	89.76	95.22	93.21	95.34	99.13
Location_5	79.92	84.62	91.77	87.13	92.71	92.93	82.09	84.07	91.35	83.05	90.53	94.36	93.92	95.26	99.02
Location_6	78.07	82.38	91.05	87.08	92.05	91.22	82.01	83.23	90.92	84.16	87.36	94.04	93.05	94.17	98.58

TABLE V
STATISTICS FOR THE CSI DATASET.

Datasets	NTU-Fi HAR [41]	Widar [42]	Self-Collected Data
platform	Atheros CSI Tool	Intel5300 NIC	Intel5300
Category Number	6	22	HAR 5 IHL 6
Category Names	Box, Circle, Clean,Fall, Run, Walk	Push and Pull,Sweep, Clap, Slide,18 types of Draws	Stand up, Down, Sit down, Hands up, Run, Walk
Data Size	(3,114,500) (antenna, subcarrier, packet)	(22,20,20) (time,x_velocity,y_velocity)	(3,30,30) (antenna, amplitude, phase)
Training Samples	936	34926	8000
Testing Samples	264	8726	2000

vectors of H^C to obtain V_t. Integration of V_t and H^C is then performed to yield the final output h'_t.

IV. EXPERIMENT AND EVALUATION

A. Experiment Setup

The experimental equipment includes a TP-LINK wireless router (TL-WR886N) and a GIGABYTE desktop with Intel 5300 NIC, shown in Fig. 9. Among them, the receiver device is equipped with Ubuntn 14.02 system and CSITool is installed to receive CSI data. The sampling frequency of the receiver is 100HzThe transmitter is placed on the side close to the window, and the receiver is placed on the opposite side close to the wall facing the window. The participants were four adults (two males and two females) with an average age of 24.5 years. Each participant performed the aforementioned activities at the various locations mentioned, with data collected for 10 seconds at each type of location and each activity repeated more than 10 times. During the data collection process, only the participant was present in the room. For the experimental evaluation, we conduct data collection in the following three different scenarios: the meeting room shown in Fig. 10(a); the laboratory shown in Fig. 10(b); and the bedroom shown in Fig. 10(c).

Within these scenarios, we design five distinct activities: standing, sitting, squatting, walking, and running, and collect data from six different locations to comprehensively assess the system performance. It is evident that the environment of the laboratory is more complex than the meeting room, and the environment of the bedroom is more complex than the laboratory.

B. Experimental Results

1) Performance Comparison Using Both Amplitude and Phase: We compare the accuracy of WiSMLF under two conditions: utilize amplitude data alone and incorporate both amplitude and phase data over the datasets of the three scenarios.

Fig. 11 illustrates the accuracy of WiSMLF utilizing amplitude data alone and amplitude with phase data for SAs. It is evident that the accuracy of WiSMLF improves when utilizing both amplitude and phase data, outperforming the use of amplitude data alone. The indoor location and activity recognition accuracy of WiSMLF with both amplitude and phase data reaches 99.76% and 96.40%, respectively, while the one with amplitude data alone only reaches 95.17% and 91.32%. Clearly, the combination of amplitude and phase data contains more valuable information, effectively enhancing the recognition performance of WiSMLF.

Fig. 12 displays the accuracy of WiSMLF employing amplitude data alone and amplitude with phase data in these three scenarios for MAs. Similarly, using both amplitude and phase data leads to significantly higher accuracy compared to using amplitude data alone. Activity recognition and distance estimation accuracy for WiSMLF with both amplitude and phase data reaches 95.93% and 92.37%, respectively, while the one with amplitude data alone achieves activity recognition and distance estimation accuracy of 92.17% and 88.13%. Based on these results, the combination of amplitude and phase data provides WiSMLF with richer information, ensuring its performance.

2) Activity Recognition and Localization Performance in Three Different Scenarios for SAs: For SAs, we compare both

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3397708

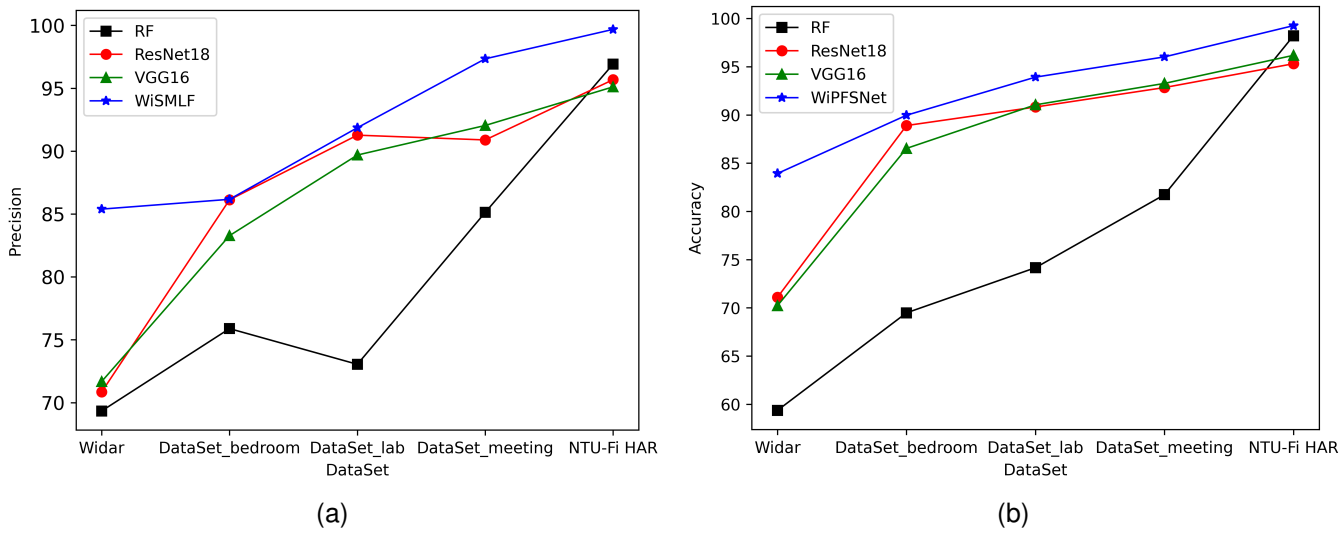

Fig. 14. Comparison of two performance indicators for four methods: (a) Precision comparison. (b) Accuracy comparison.

activity recognition and localization performance of the six methods: the proposed WiSMLF, MLP [43], ResNet18 [24] , and non-deep learning methods like support vector machine (SVM) , random forest (RF) over the datasets collected from the three scenarios.

Tab. III depicts the accuracy of six different activity recognition methods for five distinct activities in the three scenarios. It can be observed that in more complex scenarios, fewer effective features are extracted from the data, resulting in decreased accuracy of recognition performance. In the meeting room scenario, the WiSMLF method achieved an accuracy rate of over 93% for the recognition of five different activities, with the lowest accuracy for the activity of raising a hand, at 93.17%. The accuracy rates for the other two deep learning methods for the five different activities are only around 86%, while the non-deep learning methods can only reach an accuracy of 83%. In the bedroom scenario, the WiSMLF method achieved an accuracy rate of around 90% for the recognition of five different activities, the other two deep learning methods only have an accuracy rate of around 83% for the five different activities, and the non-deep learning methods can only reach 70%. To compare these results, from which we can draw the following observations:

* Deep learning methods perform better than non-deep learning methods across three different datasets.
* Among the deep learning methods, our proposed WiSMLF method shows stable and superior performance compared to the other two methods.
* The more complex the scenario, the higher the interference with the CSI signal. WiSMLF demonstrates strong learning ability in the face of highly interfered CSI signals, thereby improving recognition performance.

Tab. IV describes the accuracy of six different indoor positioning methods for locating six different positions across three scenarios involving five different activities. It is observed that in more complex scenarios, fewer effective features can be extracted from the data, resulting in decreased accuracy of

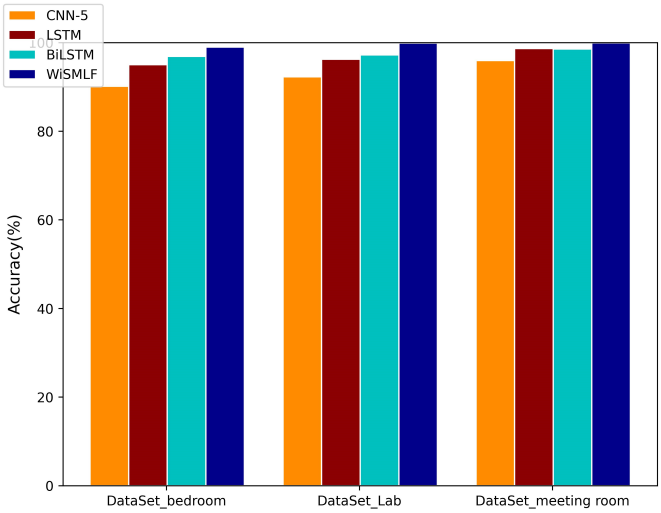

Fig. 15. Accuracy of indoor localization using different networks on datasets with different environment for MAs.

indoor positioning. In the meeting room scenario, the WiSMLF method's accuracy for positioning six different locations exceeds 93%, with the lowest accuracy for position 6 being 98.58%. The other two deep learning methods only achieve about 94% accuracy for positioning the six different locations, while the non-deep learning method can reach an accuracy of 95%. In the bedroom scenario, the WiSMLF method achieves about 93% accuracy for positioning six different locations, while the other two deep learning methods only reach about 88% accuracy for the six different locations, and the non-deep learning method achieves only 83%. Comparing these results, we find that WiSMLF is able to extract richer feature information from CSI signals, ensuring the positioning performance of WiSMLF.

The confusion matrices for joint indoor localization and activity recognition in three different scenarios with multilevel

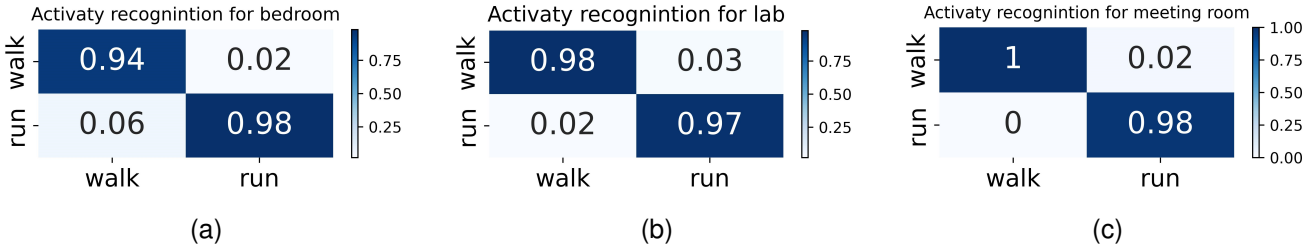

Fig. 16. Normalized confusion matrix for multi-level feature fusion for MAs activity recognition. (a) Normalized confusion matrix for activity recognition in bedroom. (b) Normalized confusion matrix for activity recognition in lab. (c) Normalized confusion matrix for activity recognition in meeting room.

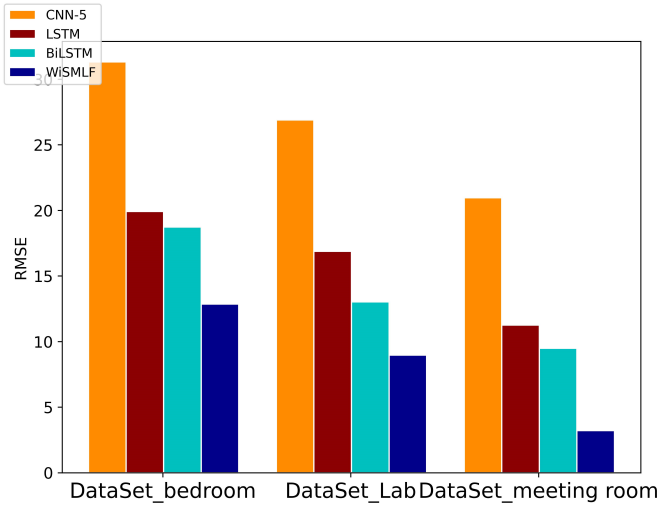

Fig. 17. Accuracy of moving distance estimation using different networks on datasets with different environment for MAs.

feature fusion are shown in Fig. 13. Specifically, Fig. 13(a) and Fig. 13(b) represent the confusion matrices for activity recognition and indoor localization in the bedroom scene, respectively. Fig. 13(c) and Fig. 13(d) represent the confusion matrices for activity recognition and indoor localization in the laboratory scenario, respectively. Fig. 13(e) and Fig. 13(f) represent the confusion matrices for activity recognition and indoor localization in a meeting room scenario, respectively.

3) Performance Comparison with Two Open Datasets: To evaluate the robustness of the proposed model, we conducted performance tests on three datasets: two open datasets used in [41], [42] and one dataset we collected. The detailed information and statistics of these datasets are summarized in Tab. V.

We used the following metrics to measure recognition performance:

1) ***Precision:*** This indicates the proportion of actual positive samples among all samples ($P_{(judge+)}$) that the model has identified as positive ($P_{(correct+)}$). It ensures that the activities identified are mostly correct, reducing false alarms. It can be represented as:

$$Precision = \frac{P_{(correct+)}}{P_{(judge+)}} \times 100\%. \tag{13}$$

2) ***Accuracy:*** This indicates the proportion of samples (both positive and negative) ($P_{(correcttotal)}$) that the model correctly predicted out of the total number of samples ($P_{(total)}$), used to evaluate overall performance. It can be represented as:

$$Accuracy = \frac{P_{(correcttotal)}}{P_{(total)}} \times 100\%. \tag{14}$$

Fig. 14 shows the recognition precision and accuracy of four methods on the datasets. As shown in Fig. 14, across all datasets, WiSMLF maintains a higher recognition accuracy and outperforms the other three methods, indicating WiSMLF higher recognition precision. Even in the presence of more data interference in the Widar and DataSet_bedroom datasets, the proposed WiSMLF accuracy rates reach as high as 88.93% and 92.15% respectively, which is about 10% higher than the other three methods.

As shown in Fig. 14, WiSMLF also maintains a higher recognition precision across all datasets, ensuring that the activities identified are mostly correct, thereby reducing false alarms. This indicates that WiSMLF has high recognition precision, with a recognition precision reaching above 85%. We can observe that the curve of the WiSMLF method is relatively stable, demonstrating strong robustness.

4) Activity Recognition and Moving Distance Performance in Three Different Scenarios for MAs: For MAs, we compare both activity recognition and moving distance performance of the four methods: the proposed WiSMLF, CNN-5 [44], LSTM [45], and BiLSTM [45] over the datasets collected from the three scenarios.

Fig. 15 exhibits the accuracy of the four different methods for activity recognition in the three scenarios. It can be observed that in more complex scenarios, fewer effective features can be extracted from the data, leading to reduced recognition performance accuracy. In the meeting room scenario, WiSMLF achieves an accuracy of 99.92%, while CNN-5 has an accuracy of 95.93%. LSTM and BiLSTM both achieve accuracies around 98%. In the bedroom scenario, WiSMLF achieves an accuracy of 98.96%, while CNN-5 has an accuracy of 90.15%. LSTM and BiLSTM achieve accuracies around 95%. Clearly, as the scene complexity increases, the interference in CSI signals also rises. For CSI signals with higher interference, WiSMLF demonstrates strong learning capabilities. The confusion matrices for MAs activity recognition with multi-level feature fusion in three different scenarios are shown in Fig. 16. Specifically, Fig. 16(a), 16(b), and 16(c) represent the confusion matrices for MAs activity recognition in the

bedroom scenario, laboratory scenario, and meeting room scenario, respectively.

For estimating moving distances, we utilize the Root Mean Squared Error (RMSE) as the metric to measure the performance of methods. RMSE provides a way to quantify the accuracy of model estimates, where a smaller error value indicates that the model estimated results are closer to the actual situation. The formula for RMSE is as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2} \tag{15}$$

where, n is the number of samples, y_i is the ith observed value, and $\hat{y}_i$ is the ith estimated value.

Fig. 17 presents the RMSE for different methods of estimating moving distances under three different scenarios. In the meeting room scenario, WiSMLF RMSE stands at 3.21, while the RMSE for the other three methods is around 14. In the bedroom scenario, WiSMLF RMSE is 12.86, whereas the RMSE for the other three methods is around 25. Additionally, when human activities are closely related to time, the performance of the LSTM network is significantly better than that of the CNN network. The multi-level feature fusion method based on LSTM extracts richer feature information from the data, with an RMSE of 3.21, which is a reduction of 8.05 compared to the conventional LSTM method and 6.28 compared to the BiLSTM method.

Based on these results, WiSMLF can extract richer feature information from time series, ensuring the performance of WiSMLF.

V. Conclusion

We have proposed a multi-level deep learning based approach for HAS by utilizing feature information extracted by backbone networks, such as VGG and LSTM, which demonstrates superior performance in complex environments. We utilize the DWT method to compute HFE, serving the purpose of classifying motion types, thus facilitating the selection of suitable sensing strategies. Subsequently, we employ the sensing strategy to select backbone networks that are tailored to the user activities for feature extraction, which further enhances the practicality of WiSMLF sensing. Experimental results indicate that the proposed WiSMLF is effective and achieves better perceptual results in real-world scenarios compared to the existing perception methods. In future work, we plan to delve into multi-person scenarios for localization and people counting, further enhancing the system adaptability to even more realistic and complex settings.

References

[1] C. De Lima, D. Belot, R. Berkvens, A. Bourdoux, D. Dardari, M. Guillaud, M. Isomursu, E.-S. Lohan, Y. Miao, A. N. Barreto, M. R. K. Aziz, J. Saloranta, T. Sanguanpuak, H. Sarieddeen, G. Seco-Granados, J. Suutala, T. Svensson, M. Valkama, B. Van Liempd, and H. Wymeersch, "Convergent communication, sensing and localization in 6g systems: An overview of technologies, opportunities and challenges," *IEEE Access*, vol. 9, pp. 26 902–26 925, 2021.

[2] C. D. Alwis, A. Kalla, Q.-V. Pham, P. Kumar, K. Dev, W.-J. Hwang, and M. Liyanage, "Survey on 6g frontiers: Trends, applications, requirements, technologies and future research," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 836–886, 2021.

[3] J. Yang, X. Chen, H. Zou, D. Wang, Q. Xu, and L. Xie, "EfficientFi: Toward large-scale lightweight WiFi sensing via csi compression," *IEEE Internet Things J.*, vol. 9, no. 15, pp. 13 086–13 095, Aug. 2022.

[4] J. Yang, Y. Zhou, H. Huang, H. Zou, and L. Xie, "MetaFi: Device-free pose estimation via commodity WiFi for metaverse avatar simulation," in *in Proc. IEEE World Forum Internet Things*, 2022, pp. 1–6.

[5] B. Li, W. Cui, W. Wang, L. Zhang, Z. Chen, and M. Wu, "Two-stream convolution augmented transformer for human activity recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 1, 2021, pp. 286–293.

[6] Y. Zhang, F. Zhang, Y. Jin, Y. Cen, V. Voronin, and S. Wan, "Local correlation ensemble with GCN based on attention features for cross-domain person Re-ID," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 19, no. 2, pp. 1–22, Feb. 2023.

[7] B. Fang, F. Sun, H. Liu, and C. Liu, "3d human gesture capturing and recognition by the IMMU-based data glove," *Neurocomputing*, vol. 277, pp. 198–207, Jan. 2018.

[8] G. Yang, Q. Zhang, and Y.-C. Liang, "Cooperative ambient backscatter communications for green internet-of-things," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 1116–1130, 2018.

[9] G. Yang, R. Dai, and Y.-C. Liang, "Energy-efficient uav backscatter communication with joint trajectory design and resource optimization," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 2, pp. 926–941, 2021.

[10] Z. Wang, Z. Yu, X. Lou, B. Guo, and L. Chen, "Gesture-radar: A dual doppler radar based system for robust recognition and quantitative profiling of human gestures," *IEEE Trans. Hum.-Mach. Syst.*, vol. 51, no. 1, pp. 32–43, Nov. 2020.

[11] J. R. Smith, K. P. Fishkin, B. Jiang, A. Mamishev, M. Philipose, A. D. Rea, S. Roy, and K. Sundara-Rajan, "RFID-based techniques for human-activity detection," *Commun. ACM*, vol. 48, no. 9, pp. 39–44, Sep. 2005.

[12] X. Zhou, W. Liang, I. Kevin, K. Wang, H. Wang, L. T. Yang, and Q. Jin, "Deep-learning-enhanced human activity recognition for internet of healthcare things," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6429–6438, Apr. 2020.

[13] D. Vasisht, S. Kumar, and D. Katabi, "Decimeter-level localization with a single WiFi access point," in *Proc. USENIX Symp. Netw. Syst. Design Implement. (NSDI)*, Santa Clara, CA, 2016, pp. 165–178.

[14] C. Chen, Y. Chen, Y. Han, H.-Q. Lai, and K. R. Liu, "Achieving centimeter-accuracy indoor localization on wifi platforms: A frequency hopping approach," *IEEE Internet Things J.*, vol. 4, no. 1, pp. 111–121, Nov. 2016.

[15] S. Shang, Q. Luo, J. Zhao, R. Xue, W. Sun, and N. Bao, "LSTM-CNN network for human activity recognition using WiFi csi data," in *J. Phys. Conf. Ser.*, vol. 1883, no. 1, Apr. 2021, p. 012139.

[16] F. Hong, X. Wang, Y. Yang, Y. Zong, Y. Zhang, and Z. Guo, "WFID: Passive device-free human identification using WiFi signal," in *Proc. Int. Conf. on Mobile and Ubiquitous Syst. Comput., Networking and Serv.*, 2016, pp. 47–56.

[17] A. Goswami, L. E. Ortiz, and S. R. Das, "WiGEM: A learning-based approach for indoor localization," in *Proc. Conf. on Emerging Networking Exp. and Technol.*, 2011, pp. 1–12.

[18] S. Sigg, M. Scholz, S. Shi, Y. Ji, and M. Beigl, "RF-sensing of activities from non-cooperative subjects in device-free recognition systems using ambient and local signals," *IEEE Trans. Mob. Comput.*, vol. 13, no. 4, pp. 907–920, Feb. 2013.

[19] G. Yang, X. Xu, Y.-C. Liang, and M. D. Renzo, "Reconfigurable intelligent surface-assisted non-orthogonal multiple access," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 5, pp. 3137–3151, 2021.

[20] J. Lin, G. Wang, S. Atapattu, R. He, G. Yang, and C. Tellambura, "Transmissive metasurfaces assisted wireless communications on railways: Channel strength evaluation and performance analysis," *IEEE Trans. Commun.*, vol. 71, no. 3, pp. 1827–1841, 2023.

[21] H. Chen, G. Hu, Z. Lei, Y. Chen, N. M. Robertson, and S. Z. Li, "Attention-based two-stream convolutional networks for face spoofing detection," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 578–593, Jun. 2019.

[22] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Device-free human activity recognition using commercial WiFi devices," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1118–1131, Mar. 2017.

[23] X. Wang, L. Gao, S. Mao, and S. Pandey, "CSI-based fingerprinting for indoor localization: A deep learning approach," *IEEE Trans. Veh. Technol.*, vol. 66, no. 1, pp. 763–776, Mar. 2016.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[26] Y. Peng, X. Tang, Y. Zhou, Y. Hou, J. Li, Y. Qi, L. Liu, and H. Lin, "How to tame mobility in federated learning over mobile networks?" *IEEE Trans. Wireless Commun.*, vol. 22, no. 12, pp. 9640–9657, 2023.

[27] Y. Peng, X. Tang, Y. Zhou, J. Li, Y. Qi, L. Liu, and H. Lin, "Computing and communication cost-aware service migration enabled by transfer reinforcement learning for dynamic vehicular edge computing networks," *IEEE Trans. Mob. Comput.*, vol. 23, no. 1, pp. 257–269, 2024.

[28] C. Duan, P. Yin, Y. Zhi, and X. Li, "Image classification of fashion-MNIST data set based on VGG network," in *Proc. Int. Symp. Educational Technol.*, vol. 19, 2019.

[29] S. Islam, S. I. A. Khan, M. M. Abedin, K. M. Habibullah, and A. K. Das, "Bird species classification from an image using VGG-16 network," in *Proc. Int. Conf. on Comput. and Commun. Manage.*, 2019, pp. 38–42.

[30] A. Mahajan and S. Chaudhary, "Categorical image classification based on representational deep network (RESNET)," in *Proc. Int. conf. Electron. Commun. Aerosp. Technol.*, 2019, pp. 327–330.

[31] Y. Ma, G. Zhou, S. Wang, H. Zhao, and W. Jung, "SignFi: Sign language recognition using WiFi," vol. 2, no. 1, pp. 1–21, Mar. 2018.

[32] S. Yousefi, H. Narui, S. Dayal, S. Ermon, and S. Valaee, "A survey on behavior recognition using WiFi channel state information," *IEEE Commun. Mag.*, vol. 55, no. 10, pp. 98–104, Oct. 2017.

[33] F. Wang, W. Gong, and J. Liu, "On spatial diversity in WiFi-based human activity recognition: A deep learning-based approach," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2035–2047, Sep. 2018.

[34] F. Wang, J. Feng, Y. Zhao, X. Zhang, S. Zhang, and J. Han, "Joint activity recognition and indoor localization with WiFi fingerprints," *IEEE Access*, vol. 7, pp. 80 058–80 068, Jun. 2019.

[35] B. Yan, W. Cheng, Y. Li, X. Gao, and H. Liu, "Joint activity recognition and indoor localization with WiFi sensing based on multi-view fusion strategy," *Digital Signal Process.*, vol. 129, p. 103680, Aug. 2022.

[36] K. Wu, M. Yang, C. Ma, and J. Yan, "CSI-based wireless localization and activity recognition using support vector machine," in *Proc. IEEE Int. Conf. Signal Process., Commun.Comput.*, 2019, pp. 1–5.

[37] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Tool release: Gathering 802.11 n traces with channel state information," vol. 41, no. 1, pp. 53–53, 2011.

[38] Y. Zhuo, H. Zhu, H. Xue, and S. Chang, "Perceiving accurate CSI phases with commodity WiFi devices," in *in Proc. IEEE Int. Conf. Comput. Commun.*, 2017, pp. 1–9.

[39] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

[40] S.-Y. Shih, F.-K. Sun, and H.-y. Lee, "Temporal pattern attention for multivariate time series forecasting," *Mach. Learn.*, vol. 108, pp. 1421–1441, 2019.

[41] J. Yang, X. Chen, H. Zou, C. X. Lu, D. Wang, S. Sun, and L. Xie, "SenseFi: A library and benchmark on deep-learning-empowered WiFi human sensing," *Patterns*, vol. 4, no. 3, Feb. 2023.

[42] Y. Zhang, Y. Zheng, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Widar3. 0: Zero-effort cross-domain gesture recognition with Wi-Fi," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8671–8688, Aug. 2021.

[43] M. W. Gardner and S. Dorling, "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences," *Atmos. Environ.*, vol. 32, no. 14-15, pp. 2627–2636, Aug. 1998.

[44] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[45] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

Yiyun Zhang received the B.Eng. degree from Hebei University of Technology, Tianjin, China, in 2021. She is currently working toward the Ph.D. degree with the School of Computer Science and Technology, Beijing Jiaotong University, Beijing, China.

Her research interests include integrated sensing and communication, WiFi sensing, wireless communication and beamforming design.

Gongpu Wang received the B.Eng. degree in communication engineering from Anhui University, Hefei, Anhui, China, in 2001, the M.Sc. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2004, and the Ph.D. degree from the University of Alberta, Edmonton, AB, Canada, in 2011.

From 2004 to 2007, he was an Assistant Professor with the School of Network Education, Beijing University of Posts and Telecommunications. He is currently a Full Professor with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing. His research interests include wireless communication theory, signal processing technologies, and the Internet of Things.

Heng Liu received the B.E. degree and the Ph.D. degree from the Beijing Institute of Technology, China, in 2017 and 2022, respectively. Now he is a postdoctoral fellow with the School of Cyberspace Science and Technology, Beijing Institute of Technology, China. His main research interests include signal processing, wireless sensor networks and convex optimization.

Wei Gong received his B.S. degree from the Department of Computer Science and Technology at Huazhong University of Science and Technology, M.S. and Ph.D. degrees in the School of Software and Department of Computer Science and Technology at Tsinghua University. He is a Professor with the School of Computer Science and Technology at University of Science and Technology of China. His research interests include backscatter networks, edge system, and IoT applications.

Feifei Gao (Fellow, IEEE) received the B.Eng. degree from Xi'an Jiaotong University, Xi'an, China in 2002, the M.Sc. degree from McMaster University, Hamilton, ON, Canada in 2004, and the Ph.D. degree from National University of Singapore, Singapore in 2007. Since 2011, he joined the Department of Automation, Tsinghua University, Beijing, China, where he is currently a tenured full professor. Prof. Gao's research interests include signal processing for communications, array signal processing, convex optimizations, and artificial intelligence assisted communications. He has authored/coauthored more than 200 refereed IEEE journal papers and more than 150 IEEE conference proceeding papers that are cited more than 18000 times in Google Scholar. Prof. Gao has served as an Editor of IEEE Transactions on Wireless Communications, IEEE Journal of Selected Topics in Signal Processing (Lead Guest Editor), IEEE Transactions on Cognitive Communications and Networking, IEEE Signal Processing Letters (Senior Editor), IEEE Communications Letters (Senior Editor), IEEE Wireless Communications Letters, and China Communications. He has also served as the symposium co-chair for 2019 IEEE Conference on Communications (ICC), 2018 IEEE Vehicular Technology Conference Spring (VTC), 2015 IEEE Conference on Communications (ICC), 2014 IEEE Global Communications Conference (GLOBECOM), 2014 IEEE Vehicular Technology Conference Fall (VTC), as well as Technical Committee Members for more than 50 IEEE conferences.