AFERSM-Net: Joint Network for Gesture Recognition and Location Classification

Xu Lu, Zexiao Cai, Xiongwei Huang, Cheng Zhou, and Jun Liu

***Abstract*—With the increasing deployment of wireless communication systems and smart devices, WiFi-based gesture recognition and indoor location classification have gained attention. These technologies extract gesture and location features from WiFi channel state information (CSI). However, the signal is susceptible to interference from the environment during CSI data acquisition to produce multipath effect noise, and the amplitude change with the change of location often affects the extraction and recognition of gesture features. To address these problems, Auxiliary Feature Extraction based Residual Shrinkage Multi-tasking Network (AFERSM-Net) is proposed for gesture recognition and position classification of one-dimensional multivariate time series. AFERSM-Net is a hybrid architecture that combines CNN for spatial feature extraction and LSTM networks for capturing temporal dependencies. Firstly, a reasonable threshold is set adaptively by the shrinkage module to dynamically identify and eliminate the transformed environmental noise. Secondly, the feature extraction module is used to focus on and extract location-independent gesture features to reduce the influence of location-independent features. Finally, the gesture features extracted by the feature extraction module are fused with the shared features of the residual shrinkage multi-tasking network as an aid. Its module fusion is mainly used to improve the accuracy of gesture recognition and solve the problem of insufficient model generalization ability. We evaluated this network on a dual-labeled gesture and location dataset, and the gesture recognition accuracy was 97.84% and the location classification accuracy was 98.92%, which outperformed other advanced network frameworks.**

Link to graphical and video abstracts, and to code:
https://latamt.ieeer9.org/index.php/transactions/article/view/9515

***Index Terms*—Gesture recognition, Location classification, Noise reduction, Feature extraction, Deep learning.**

I. Introduction

WIFI-based human feature extraction, leveraging different subcarrier amplitudes from channel state information (CSI), has been widely applied to tasks like person identification [1], [2], trajectory prediction [3], [4], gesture recognition [5], [6], and location classification [7], [8]. Unlike camera-based methods, WiFi technology avoids issues like privacy concerns, occlusion, lighting, and computational complexity, making it a popular research area. Current methods, including CNN [9], [10], RNN [11], GAN [12], and LSTM [13].These algorithms extract features from both the time and spatial dimensions of CSI data, and then output the corresponding single classification tasks results. However, WiFi data collection is prone to environmental interference, leading to multipath effect noise, which negatively impacts feature extraction and classification accuracy.

To address the problem of multipath effect noise interference, solutions based on professional knowledge were proposed in the literature [14], [15]. Wang et al. [14] proposed a noise reduction method based on signal waveform classification for Raman-based distributed temperature sensors.Leshem et al. [15] demonstrated that increasing the distance between antenna elements and building a large enough linear antenna array can help mitigate specular multipath and channel noise.However, the challenge remains in setting an appropriate threshold to eliminate noise in different environments without requiring signal processing expertise. In real-world applications, multiple features are often needed to solve practical problems, and single-task networks fail to meet these needs. To address the multitasking requirements of WiFi, Wang et al. [16] released a CSI dataset with gesture and location labels, which is shown in Fig. 1. They also proposed a 1DResNet model with hard parameter sharing to combine activity recognition and location classification. However, due to the complex effects of gesture magnitude and location changes, the gesture recognition accuracy of the 1DResNet model was suboptimal, which led us to investigate and address this issue.

In order to solve the above existing problems, AFERSM-Net is proposed in this paper. By leveraging WiFi signals, AFERSM-Net eliminates the need for additional hardware, making it a practical choice for deployment in locations where installing new sensors may be impractical or cost-prohibitive. The model was evaluated using a dataset collected in the literature. [16].The experimental results demonstrate that the proposed network proposed achieves excellent performance in multitasking. The contributions are summarized as follows:

1) A residual shrinkage multi-tasking network is proposed, where the shrinkage module is integrated into the residual module and a gradient descent algorithm is applied to automatically identify and eliminate multipath noise features.

2) A Multiscale Fully Convolutional SE-LSTM (MFCs-LSTM) feature extraction module is developed, which extracts spatial feature vectors of different poses using the SE module and multiscale convolution. These spatial features are then

The associate editor coordinating the review of this manuscript and approving it for publication was Oscar Mauricio Caicedo (*Corresponding author: Jun Liu*).

X. Lu is with the Guangdong Polytechnic Normal University, Guangdong Provincial Key Laboratory of Intellectual Property & Big Data, and Pazhou Lab, Guangzhou, China (e-mail: bruda@126.com).

Z. Cai, X. Huang, C. Zhou, and Jun Liu are with the School of Computer Science, Guangdong Polytechnic Normal University, Guangdong, China (e-mails: purecade_cai@163.com, 1158191782@qq.com, 2471868101@qq.com, and liujun7700@163.com).

processed by an LSTM network to capture temporal characteristics.

3) AFERSM-Net is proposed to enhance gesture recognition accuracy by fusing feature parameters from the feature extraction module, combining shared feature vectors from the gesture task output layer with gesture spatio-temporal vectors extracted by the auxiliary module.

4) Through experimental studies, it is found that the network model proposed in this paper achieves 97.84% and 98.92% accuracy in gesture recognition and location classification tasks, respectively, which is better than other advanced multitasking algorithms.

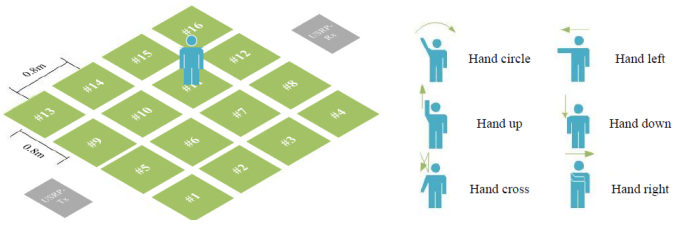


Fig. 1. Location and gesture acquisition scenarios.

II. Related Work

WiFi-based gesture recognition and location classification algorithms are generally divided into shallow learning [1], [2] and deep learning [9], [11] approaches. These algorithms extract features from the amplitude changes of WiFi signals passing through the human body, but the resulting features are often contaminated by ambient multipath effect noise, which adversely affects the training performance. Moreover, most current research focuses on single-task optimization, failing to address the need for multi-task solutions in real-life applications. This section analyzes the current state of multipath effect noise elimination and multi-label sample classification.

A. Multipath Effect Noise Cancellation Method

WiFi CSI datasets suffer from signal distortion due to environmental multipath effects during acquisition. Chen et al. [17] introduced a signal filtering mechanism, which reduces the impact of multipath effects by feeding the row vector averages into a parallel capsule network.However, this method is prone to missing data and overfitting in small datasets. Yang et al. [18] proposed NISE and PSE algorithms to amplify activity signals and reduce inactive ones, but they do not effectively eliminate multipath noise and may increase location-related noise. In contrast, the proposed method uses a residual shrinkage module [19] to train soft thresholds for CSI data, which are then inserted into the residual module as a nonlinear transformation to eliminate noise.

B. Multi-label Sample Classification Methods

Since the joint-label CSI dataset includes two or more labels, feature overlap often occurs, which hinders feature extraction.To mitigate this, Hao et al. [20] proposed the WiPg network, integrating CNN with GAN for body size-independent yoga feature extraction. Ma et al. [21] used a 2D CNN to extract location-independent fall action features, followed by 1D CNN for temporal features and LSTM optimization to improve fall classification. Li et al. [22] addressed environmental variables by training a person activity model, transferring it to the target domain using semi-supervised migration, and fine-tuning with DADA-AD for accuracy enhancement.The approaches mentioned above improve classification accuracy by removing or mitigating the impact of other label features on the classification task. However, in real-world scenarios, there is often a need to handle multiple labeling tasks simultaneously, which presents an additional challenge. Alitaleshi et al. [23] introduced a clustering-assisted extreme learning machine for multi-label classification. In contrast, we address multi-task classification accuracy issues by using a pruned residual shrinkage multi-task network, enhanced by the MFCs-LSTM auxiliary module to learn and fuse gesture features.

III. Method

A. Network Architecture

CSI is susceptible to multipath effect noise caused by environmental factors during the acquisition process.When multi-label CSI dataset input into a multi-task network, the weights often become biased towards tasks with easier feature extraction, affecting model generalization. To address the joint task of gesture recognition and location classification, the AFERSM-Net is proposed in this paper, as shown in Fig. 2.

AFERSM-Net consists of two main components: the auxiliary module and the residual shrinkage multi-task network. The residual shrinkage multi-task network implements two branching tasks by sharing weight parameters, while the auxiliary module trains gesture weight parameters to enhance gesture feature information. For gesture weight parameter training, CSI samples are first input in parallel to two different scales of FCNs with SE module in the auxiliary module to obtain location-independent gesture space features. These features are then reshaped and input to the LSTM network to extract temporal features. For training with shared weight parameters, CSI samples are downsampled using a 1D convolution (size 7) in the residual shrinkage network, then normalized. The processed samples are fed into three residual structures with shrinkage modules, where feature matrix summation and soft thresholding yield common gesture and location feature information. The common feature information is then input into both the gesture and location output layers. The location layer outputs classification results directly, while the gesture layer fuses gesture spatio-temporal features from the auxiliary module with shared features to enhance and output gesture classification results.

B. Multipath Noise Processing

Indoor WiFi signals can be affected by obstacles such as walls and ceilings, resulting in multiple reflection paths during signal propagation. The receiver will continuously receive signals from multiple reflection paths, and this phenomenon is called multipath effect lu2022 characteristics. According to

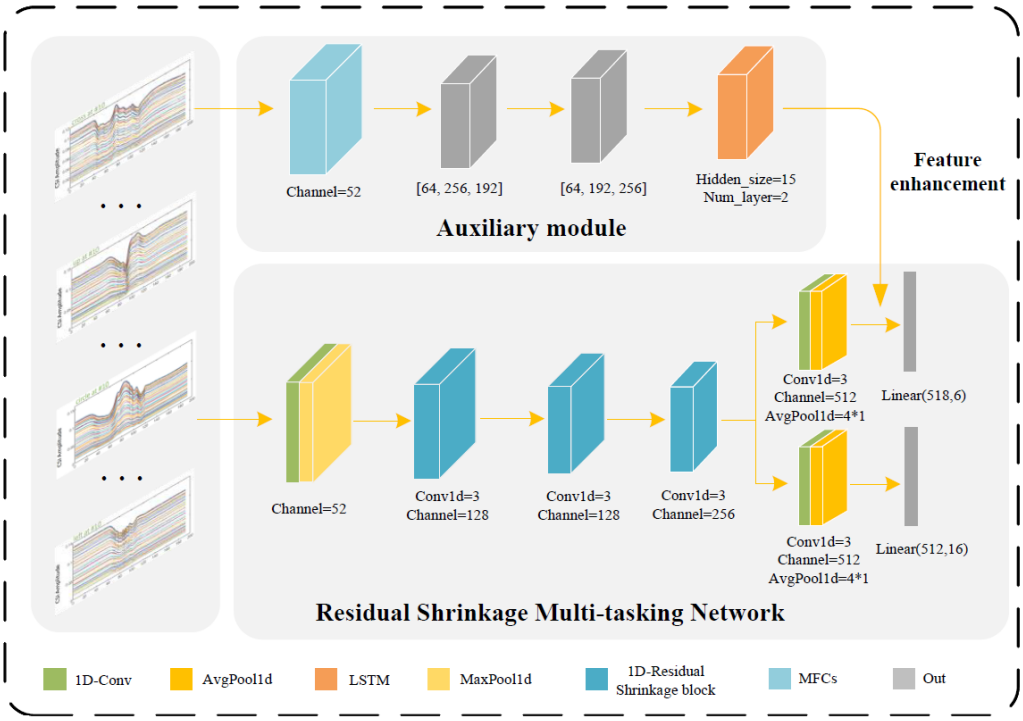


Fig. 2. AFERSM-Net Architecture - The proposed AFERSM-Net comprises a residual shrinkage multi-task network for shared-weight learning for gesture and location classification tasks and an auxiliary module for enhancing gesture features.

the Friis free-space transmission formula, it can be derived that

$$p_r(d) = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 (d^2 + 4h)^2} \tag{1}$$

where $p_r(d)$ is the received power, P_t is the transmitted power, G_r is the gain of the receiver antenna, G_t is the gain of the transmitter antenna, λ is the wavelength of the signal, d is the distance between the transmitter and the receiver, and h is the distance from the radiating point on the ground and the wall to the direct path.

When a human body responds within the signal propagation range, it reflects and scatters the signal, altering the signal propagation equation.he signal propagation equation is

$$p_r(d) = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 (d^2 + 4h + \Delta)^2} \tag{2}$$

Δ is the approximate change of the path length caused by the human body.

Equation (2) shows that the received signal power variation reflects human position, gesture changes, and environmental noise. Under the IEEE 802.11n protocol, CSI represents a combination of multi-carrier channel responses, which can be expressed in the following form.

$$H(f) = [H(f_1), H(f_2), ..., H(f_k)] \tag{3}$$

Due to environmental reflections, received signals exhibit multipath signal attenuation. In this paper, the proposed network decompose each subcarrier signal by wavelet, apply soft thresholding to eliminate near-zero features, and then wavelet reconfiguration. This soft thresholding can be expressed as follows.

$$y = \begin{cases} x - \tau & , x > \tau \\ 0 & , -\tau \leq x \geq \tau \\ x + \tau & , x < -\tau \end{cases} \tag{4}$$

where x is the input feature, y is the output feature, and τ is the threshold value, which is a positive parameter. This soft threshold is assembled into a trainable module through several neural network layers, and is automatically learned using the gradient descent algorithm, which is shown in Fig. 3. The automatically determined soft threshold is inserted into the residual module of 1DResNet [1, 1, 1] as a nonlinear transformation layer. The soft threshold module, as a sub-network of the residual network, first absolutizes the input features and performs a global pooling operation using GAP to obtain a one-dimensional vector. The one-dimensional vector is input to the FC network, and then a scaling parameter is obtained by BN, RELU operation, and finally input to the FC network. Normalizing the scaling parameter to the range of (0, 1) by the Sigmoid function can be expressed as

$$\alpha = \frac{1}{1 + e^{-z}} \tag{5}$$

where z is the output of the second layer FC network in the residuals module and α is the corresponding scaling parameter. The scaling parameter α is then multiplied by the average value of $|x|$ to obtain the threshold value. In summary, the threshold values used in the residual shrinkage module are expressed as

$$\tau = \alpha \cdot average|x_{i,j,c}| \tag{6}$$

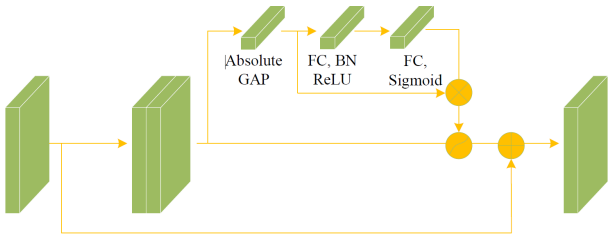


Fig. 3. Shrinkage Module - The soft threshold module is integrated into the residual module of 1DResNet [1, 1, 1] as a nonlinear transformation layer that computes scaling parameters using GAP, FC layers, BN, and ReLU operations.

where τ is the threshold value and i, j, and c are the indicators for the width, height, and channel of the feature map x, respectively. The threshold values can be kept within a reasonable range so that the output of the soft threshold is not all zero.

The residual module consists of two convolutional layers $f(x)$ and a shortcut branch x. As shown in the figure below, by feeding the sample x into the residual module $g(x)$, its result y is output.

$$g(x) = f(x) + x \tag{7}$$

The residual shrinkage multi-tasking network extracts location and gesture features from CSI samples with 52 input channels (Fig. 4). Unlike other methods, it uses the second convolutional layer output as input to the shrinkage module to train the soft thresholding. Then the output features are soft-thresholding and finally summed with the shortcut branch x to obtain the noise-reduced human features. To address overfitting due to the small number of CSI samples and high model complexity, the network reduces residual modules to three through pruning. Finally, after extracting the denoised common features, six gesture classification results and 16 location classification results are output using two fully connected layers, and their classification accuracies are better than 1Dresnet [1, 1, 1, 1].

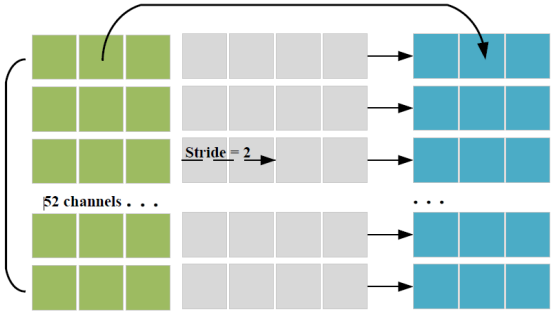


Fig. 4. Multi-channel 1D Convolution - The residual shrinkage multi-tasking network extracts location and gesture features from CSI samples using 52 input channels, employing the second convolutional layer output for soft threshold training and noise reduction, with three residual modules to mitigate overfitting.

C. Feature Extraction

To reduce the impact of location-independent features on gesture classification accuracy, the proposed network introduce an MFCNs-LSTM module to separately extract gesture features. The number of input channels is 52, and the module is shown in Fig. 5. Both FCNs networks at different scales contain three convolutional blocks, each containing one convolutional layer with the number of channels (128, 256, 128), respectively. The size of the convolutional kernels of the first FCNs network is (7, 5, 3) and the size of the convolutional kernels of the second FCNs network is (5, 3, 1), respectively. The first two convolutional blocks contain two SE modules and the latter two convolutional blocks contain two average pooling layers.

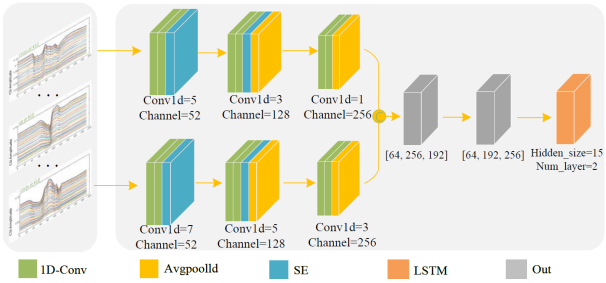


Fig. 5. Auxiliary Module - The auxiliary module comprises two FCN networks with three convolutional blocks, featuring distinct kernel sizes and incorporating SE modules and average pooling layers for enhanced feature extraction.

The SE module plays a critical role in adaptively recalibrating channel-wise feature responses, allowing the network to selectively focus on gesture-relevant spatial features while suppressing location-independent features. Specifically, the SE module achieves this by applying a squeeze operation to aggregate global spatial information into channel-wise statistics, followed by an excitation operation that generates adaptive weights for each channel. This attention mechanism enhances the discriminative power of the feature extraction process, helping to isolate gesture-specific features effectively. The inclusion of SE modules improves the network's ability to handle variations in location-independent features, contributing to higher classification accuracy.

The network uses two FCN blocks of different scales to capture the differences in the duration of each gesture, and then selectively focuses on gesture space features and suppresses location-independent features by the SE module. Finally, the gesture space features are fed into the LSTM network to extract their temporal features.

In order to explicitly characterize the temporal dependencies between CSI sequences, this paper uses LSTM networks to learn their dependencies and extract the temporal features of the gestures. LSTM networks [24] deal with the long-term and short-term relationships through four functions, namely $Sigmoid(\sigma)$, $hyperbolictangent(tanh)$, $multiplication(\times)$ and $sum(+)$. The parameters x_t, o_t and c_t are denoted as input, output and memory units, respectively. The output

function at time t can be expressed as

$$f_t = \sigma(\omega_f \times [o_{t-1}, x_t] + b_f) \tag{8}$$
$$i_t = \sigma(\omega_i \times [o_{t-1}, x_t] + b_i) \tag{9}$$
$$\tilde{c}_t = tanh(\omega_c \times [o_{t-1}, x_t] + b_c)c_t \tag{10}$$
$$\tilde{o}_t = \sigma(\omega_o \times [o_{t-1}, x_t] + b_o) \tag{11}$$
$$o_t = \tilde{o}_t \times tanh(c_t) \tag{12}$$

where ω_f, ω_i, ω_c, ω_o and b_f, b_i, b_c, b_o are the parameters required to construct the LSTM network.

In this paper, 52 subcarriers of CSI samples are input to two FCNs networks as multivariate features. The CSI samples are outputted as two 3D tensors of size [64, 128, 192] after feature extraction, compression, extraction and average pooling operations in three convolutional blocks. The two 3D tensors are stitched on $dim = 1$ to obtain a 3D tensor of size [64, 256, 192], and finally the 3D tensor of size [64, 192, 256] is obtained by converting $dim = 1$ and $dim = 2$ through the permutation function. The CSI samples are extracted with features of spatial dimension by MFCNs network to obtain the 3D output tensor. The 3D output tensor is input to LSTM network with feature dimension of 256, hidden layer dimension of 15, recurrent neural network layers of 2, and output category number of 2. By training the LSTM network, the gesture features independent of the position state are finally obtained.

D. Auxiliary Enhancement

The hard-sharing mechanism in multi-task learning shares weight parameters across tasks. While it is suitable for CSI samples with correlated gesture and location signals, the complexity of gesture amplitude variations due to location influences makes feature extraction difficult. This results in the residual shrinkage multi-task network's weights being biased towards the location task, diminishing the advantages of multi-task learning.

To address the above issues, this paper proposes an enhanced gesture branching task based on auxiliary feature extraction. First, CSI samples are input to the shared layer of the residual shrinkage multi-task network, generating n multi-channel 1D vectors of length 6. These vectors are subjected to a 3×1 1D convolution operation followed by BN, ReLU, and average pooling operations to produce gesture vector features. The output gesture vector features are reshaped into 2D vectors of column size 512. Similarly, CSI samples pass through the MFCs-LSTM network, outputting n 1D vectors of length 6. These vectors are reshaped into 2D vectors of column size 512, and then the two 2D vectors are concatenated by column to form 2D vectors of size 518. This 2D vector is fed into a fully connected layer to output the gesture category. The network improves gesture classification performance while maintaining focus on location classification. By fusing gesture features from the MFCs-LSTM with the gesture classification task output, the network's generalization capability is enhanced.

IV. Experimental Comparisons and Analysis

A. Equations

The experimental dataset in this paper is a CSI dataset collected using the literature [16], involving six hand movements, i.e., gestures up, down, left, right, drawing circles, and hitting cross movements performed by a user in 16 positions, as shown in Fig. 1. In each position, each activity was repeated 15 times, and excluding the bad samples, a total of 1394 sample data were collected, as shown in Fig. 6.

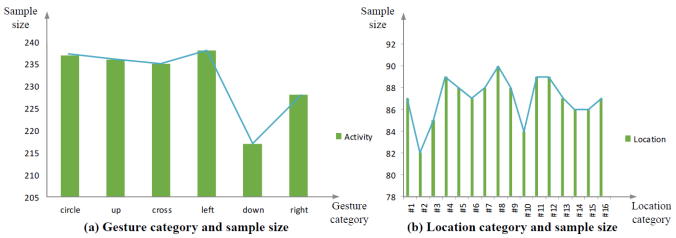


Fig. 6. The CSI dataset from literature [16] includes six hand gestures performed at 16 positions.

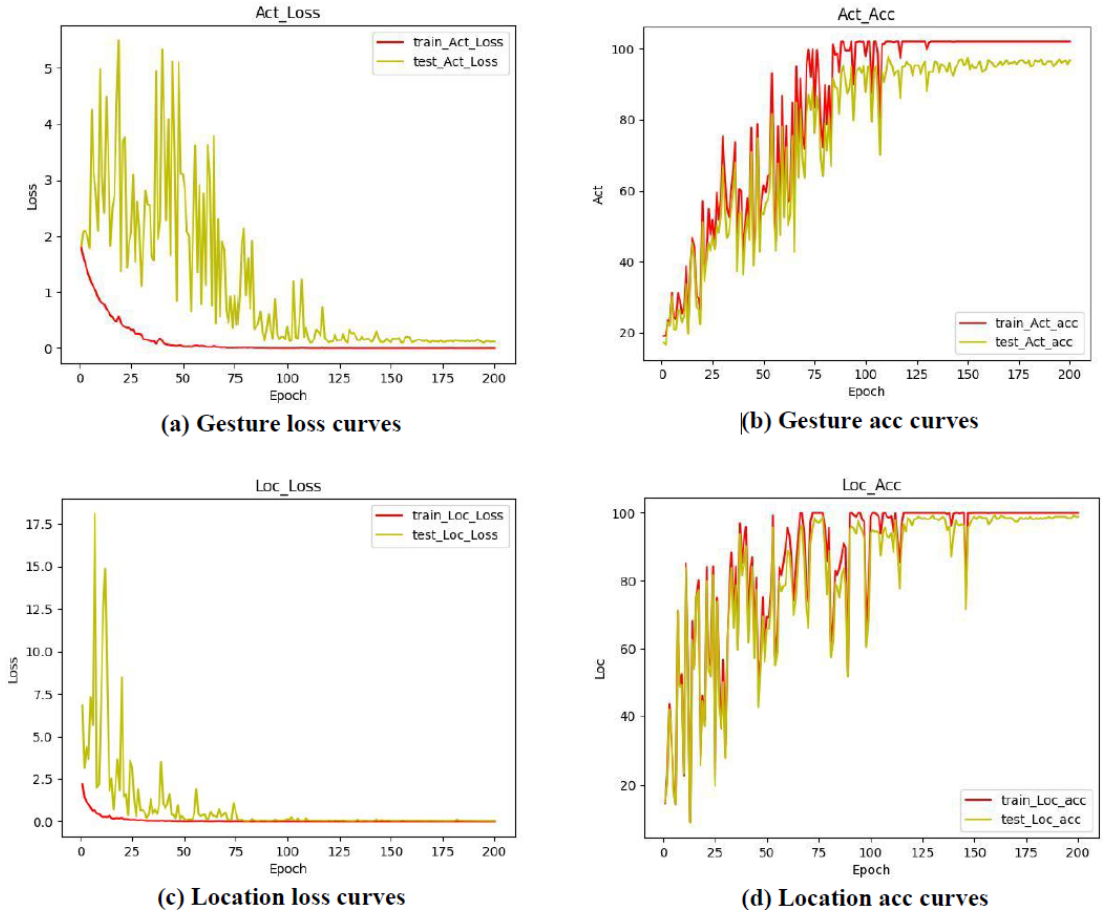
B. Experimental Setup

Each sample data contains 52 subcarrier signals, and 1116 samples were used as the training set data, and then 278 samples with the same conditions as the training set were evenly selected from all samples as the test set data.Each sample data contains 52 subcarrier signals, and 1116 samples were used as the training set data, and then 278 samples with the same conditions as the training set were evenly selected from all samples as the test set data.The proposed model is trained using a Adam optimizer with an initial learning rate of 0.001. The batch size is set to 64, and the loss function used for training is cross-entropy loss.

C. Learning Curves

The configuration of the experimental environment in this paper is shown in Table I.

Fig. 7 shows that during gesture recognition training, the average error loss gradually decreases as the number of iterations epoch increases and accuracy stabilizes after three oscillations. In the process of gesture recognition test, the loss oscillates twice after the 100th epoch before stabilizing, with accuracy peaking at 97.84%. In the training process of location classification, with the increase of epochs, the average error loss decreases and tends to zero; accuracy finally reaches the peak after several oscillations. In the test process, the loss tends to coincide with the training loss after several oscillations, and the test accuracy also approaches the training accuracy after several oscillations. In testing, loss aligns with training loss, and accuracy approaches training accuracy after several oscillations. The observed fluctuations in loss and accuracy during early stages are likely due to small sample inputs in a complex network structure, which disrupt parameter updates and hinder convergence.



(a) Gesture loss curves **(b) Gesture acc curves**

(c) Location loss curves **(d) Location acc curves**

Fig. 7. Loss and Acc curves - (a) and (b) depict the loss and accuracy curves for gesture recognition, while (c) and (d) illustrate the loss and accuracy curves for location classification.

TABLE I
EXPERIMENTAL ENVIRONMENT CONFIGURATION

OS	Ubuntu 20.04
GPU	NVIDIA GeForce GTX 2080Ti
Deep learning framework	Pytorch1.10.0
Deep learning framework	Python3.7.0

D. Experimental Result

This study uses classification evaluation metrics such as confusion matrix, accuracy, precision, recall, specificity, and F1 scores.

Fig. 8(a) and Fig. 8(b) show the confusion matrix of the present network for gesture recognition and location classification respectively. The proposed network obtains 97.84% accuracy for gesture recognition and 98.92% accuracy for location classification. The errors in gesture recognition are mainly in "gesture down" and "gesture right," while location classification errors are concentrated in locations 7, 8, and 12. The results demonstrate that the proposed network achieves excellent performance in both tasks.

We further calculated the precision rate, recall rate, specificity, and F1 scores based on the confusion matrix and presented the results in Tables II and III. From Table II, we can see that each gesture recognition evaluation index achieves better results, among which the precision rate of gesture playing cross is 1.0, the recall rate of gesture up is 1.0, the specificity of gesture playing cross is 1.0, and the F1 scores of gesture up is 0.989. Compared with gesture up and gesture playing cross, the evaluation index of gesture drawing circle is slightly lower. From Table III, it can be seen that the evaluation metrics for location classification are better than those for gesture recognition. Each evaluation metric for each location is 1.0 or close to 1.0, except for location 4, where the accuracy is lower. From the above table, it can be concluded that the present network achieves very good performance in both gesture recognition and location classification tasks.

In this paper, the experiments were trained and tested on the above dataset, using a network framework based on one-dimensional ResNet [1, 1, 1, 1] [16], 1DResNet [1, 1, 1, 1]+ [16], DRSN [19], LSTM-FCNs [25], to compare the results with the proposed AFERSM-Net. The epoch of each network module is set to 200, batch_size to 128, Learning rate to 0.008, and gamma to 0.8, and the experimental results are shown in the following figure. In order to compare the differences between the training results of the two-dimensional convolutional and the one-dimensional convolutional, we use the GAF algorithm to transform the one-dimensional CSI dataset into RGB images. Then RGB images are fed into the classical network EfficientNetV2 for training and comparison. As shown in Table IV, this proposed network achieves 97.84%

TABLE II
LOCATION CLASSIFICATION EVALUATION INDICATORS

Location		#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16
NMTS	Precision	0.954	1.0	1.0	1.0	1.0	1.0	1.0	0.917	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	Recall	1.0	1.0	1.0	1.0	0.937	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	Specificity	0.995	1.0	1.0	1.0	1.0	1.0	1.0	0.995	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	F1 score	0.977	1.0	1.0	1.0	0.968	1.0	1.0	0.957	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.971
UMTS	Precision	1.0	1.0	1.0	1.0	0.421	1.0	1.0	0.367	0.739	1.0	1.0	0.875	1.0	1.0	1.0	0.556
	Recall	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.714	1.0	0.727	0.833	1.0	0.684	0.4375	0.909
	Specificity	1.0	1.0	1.0	1.0	0.893	1.0	1.0	0.91	1.0	0.971	1.0	1.0	0.995	1.0	1.0	0.962
	F1 score	1.0	1.0	1.0	1.0	0.592	1.0	1.0	0.536	0.726	0.85	0.842	0.853	0.933	0.812	0.608	0.689
KDMTS	Precision	1.0	1.0	1.0	0.937	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	Recall	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.944	1.0	1.0	1.0	1.0
	Specificity	1.0	1.0	1.0	0.995	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	F1 score	1.0	1.0	1.0	0.967	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.971	1.0	1.0	1.0	1.0
Wimuse	Precision	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.814
	Recall	1.0	1.0	1.0	0.933	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.778	1.0	1.0	1.0	1.0
	Specificity	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.975
	F1 score	1.0	1.0	1.0	0.965	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.875	1.0	1.0	1.0	0.897
AFERSM-Net (ours)	Precision	1.0	1.0	1.0	0.895	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.947	1.0	1.0	1.0
	Recall	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.938	1.0	1.0	0.944	1.0	1.0	1.0	0.944
	Specificity	1.0	1.0	1.0	0.992	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.996	1.0	1.0	1.0
	F1 score	1.0	1.0	1.0	0.945	1.0	1.0	1.0	1.0	0.968	1.0	1.0	0.971	0.973	1.0	1.0	0.971

TABLE III
GESTURE CLASSIFICATION EVALUATION INDEX

Activity		Up	Down	Left	Right	Circle	Cross
NMTS	Precision	0.965	0.911	0.843	1.0	0.944	0.976
	Recall	0.903	0.968	1.0	0.939	0.971	0.836
	Specificity	0.994	0.984	0.955	1.0	0.989	0.994
	F1 score	0.933	0.939	0.914	0.968	0.957	0.901
UMTS	Precision	0.896	0.761	0.939	1.0	0.739	0.86
	Recall	0.838	1.0	0.721	0.696	0.971	0.877
	Specificity	0.984	0.947	0.988	1.0	0.936	0.959
	F1 score	0.866	0.864	0.815	0.821	0.839	0.868
KDMTS	Precision	0.906	0.885	0.807	0.682	1.0	0.875
	Recall	0.909	0.407	0.847	0.903	0.931	0.857
	Specificity	0.952	0.994	0.966	0.88	0.979	0.983
	F1 score	0.833	0.567	0.857	0.682	0.9	0.889
Wimuse	Precision	0.769	0.944	0.867	0.549	0.871	0.923
	Recall	0.838	1.0	0.721	0.696	0.971	0.877
	Specificity	0.984	0.947	0.988	1.0	0.936	0.959
	F1 score	0.866	0.864	0.815	0.821	0.839	0.868
AFERSM-Net (ours)	Precision	0.979	0.979	0.959	0.98	0.975	1.0
	Recall	1.0	1.0	1.0	1.0	0.907	0.957
	Specificity	0.996	0.996	0.991	0.996	0.996	1.0
	F1 score	0.989	0.989	0.979	0.99	0.94	0.978

and 98.92% accuracy in the gesture recognition and location classification tasks, respectively. These evaluation metrics outperforming other algorithms and proving that the proposed algorithm is more suitable for this CSI dataset.

To comprehensively evaluate the performance of the proposed network, we conducted experiments using five-fold cross-validation. This approach allowed us to train and validate the model multiple times on different subsets of the data, providing an effective test of the model's stability and generalization ability. As shown in Table V, the proposed network consistently performed well in each fold, maintaining high accuracy across different training-validation splits. This result confirms that the proposed network maintains stable and accurate gesture recognition and location classification on unseen data, highlighting its robustness and stability.

TABLE IV
LOCATION CLASSIFICATION EVALUATION INDICATORS

Network Model	Activity Recognition	Indoor Localization
1D ResNet[1,1,1,1]	88.13%	95.68%
1D ResNet[1,1,1,1]+	89.57%	95.68%
DRSN	88.85%	96.76%
LSTM-FCNS	62.94%	94.24%
GAF+EfficientNetV2	83.03%	87.63%
AFERSM-Net (ours)	**97.84%**	**98.92%**

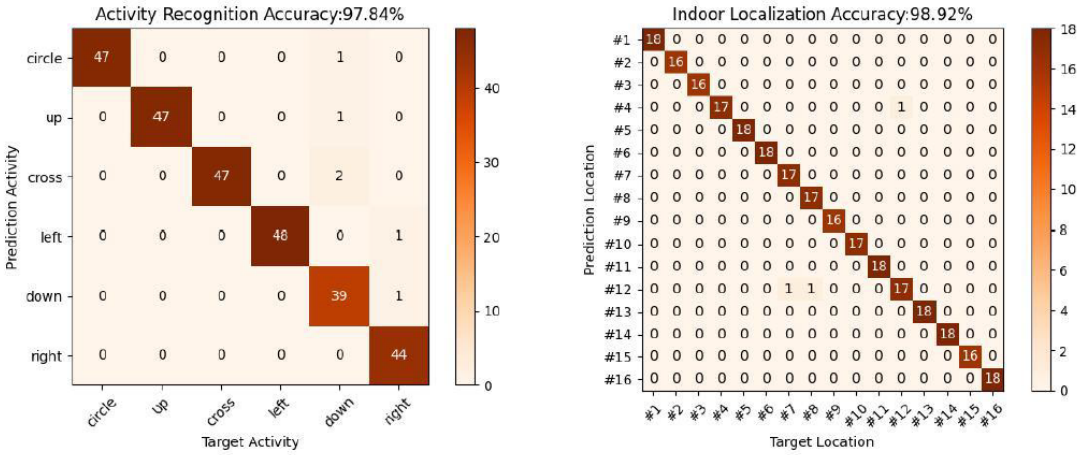
TABLE V
ACCURACY FOR FIVE-FOLD CROSS-VALIDATION

Task	Average	Fold1	Fold2	Fold3	Fold4	Fold5
AR	92.24%	89.60%	93.18%	92.83%	93.90%	91.72%
IL	98.76%	98.20%	98.92%	98.92%	98.56	99.28%

AR and IL are the abbreviations of Activity Recognition and Indoor Localization.

E. Ablation Experiment

In this paper, ablation experiments are used to verify the performance of Shrinkage Block, MFCNs-LSTM in the network. Residual shrinkage multi-tasking network is 1Dresnet [1, 1, 1] joined with Shrinkage Block, AFERSM-Net is residual shrinkage multi-tasking network fused with MFCs-LSTM module. The ablation experiments are used to analyze and



(a) Gestural confusion matrix **(b) Location confusion matrix**

Fig. 8. Confusion matrix - (a) Gesture recognition confusion matrix showing errors mainly in "gesture down" and "gesture right" with 97.84% accuracy; (b) Location classification confusion matrix showing errors mainly in locations 7, 8, and 12 with 98.92% accuracy.

verify the results of the effect of different modules on gesture recognition and location classification. From the table, it can be seen that the original network model 1DResNet [1, 1, 1] has 92.32% accuracy for gesture recognition and 96.68% accuracy for location classification. The residual shrinkage multi-tasking network with the addition of the Shrinkage Block achieves 95.16% accuracy for gesture recognition and 98.20% accuracy for location classification. Furthermore, the residual shrinkage multi-tasking network with the fused feature extraction module achieves 97.84% accuracy for gesture recognition and 98.92% accuracy for location classification. From the above analysis, it can be concluded that the addition of Shrinkage Block module improves the accuracy of gesture recognition and location classification; the addition of MFCNs-LSTM network module improves the accuracy of gesture recognition.

TABLE VI
IMPACT OF INDIVIDUAL MODULES ON THE PERFORMANCE OF AFERSM-NET IN ABLATION EXPERIMENTS

Network Model	Activity Recognition	Indoor Localization
1D ResNet[1,1,1]	92.32%	96.68%
Residual Shrinkage Multi-tasking Network	95.16%	98.20%
Residual Shrinkage Module	97.02%	76.41%
AFERSM-Net (ours)	**97.84%**	**98.92%**

F. Contrast Experiment

Following ablation studies that assessed individual components, we conducted comparative experiments to validate the effectiveness of the proposed network. Specifically, we compared this approach against several established models, including NMTS [26],UMTS [26],KDMTS [26] and Wimuse [26], to evaluate their performance on ARIL Dateset.These experiments clearly demonstrate the advantages of the proposed approach, particularly its performance in multi-task models. As shown in Tables II, III, and VII, the proposed network achieves 97.84% accuracy in Activity Recognition and 98.92% in Indoor Localization, showcasing its effectiveness and consistency across tasks. These results not only validate the superiority of our approach but also provide new insights for future research in multi-task learning.

TABLE VII
PERFORMANCE COMPARISON OF AFERSM-NET WITH EXISTING METHODS IN MULTI-TASK CLASSIFICATION

Network Model	AR	IL	GPU Memory Usage	Time Cost
NMTS	63.05%	73.98%	378MB	691s
UMTS	65.25%	76.43%	378MB	1368s
KDMTS	56.09%	66.26%	668MB	1233s
Wimuse	93.95%	97.31%	670MB	1651s
AFERSM-Net (ours)	**97.84%**	**98.92%**	**1852MB**	**701s**

AR and IL are the abbreviations of Activity Recognition and Indoor Localization.

V. CONCLUSION

In this paper, we propose AFERSM-Net. This network performs fusion operation with MFCs-LSTM module after inserting Shrinkage Block module on the basis of 1Dresnet [1, 1, 1]. This combination effectively removes multipath noise, alleviates gesture feature extraction challenges, and enhances network generalization. It outperforms other multi-task network models in both gesture recognition and location classification accuracy. However, it is limited by the fact that its performance has only been validated on a specific dataset, and its generalization ability on diverse datasets and real-world environments remains untested. To address this, We will validate its generalization capability on diverse datasets and explore integrating it with other deep learning architectures or signal processing methods to improve effectiveness. Finally, AFERSM-Net shows great potential for integration into a broad range of IoT applications. Future research will explore the adaptation of this model for different IoT domains, such as healthcare and large-scale industrial monitoring, addressing challenges like heterogeneous sensor networks and the need for real-time analysis.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China (62176067); Key-Area Research and Development Program of Guangdong Province (2023B0303020001); Scientific and Technological Planning Project of Guangzhou (202103000040, 2023B03J1378); Research project of Guangdong Polytechnic Normal University (22GPNUZDJS14).

REFERENCES

[1] Y. Liu, W. Zhou, M. Xi, S. Shen, and H. Li, "Multi-modal context propagation for person re-identification with wireless positioning," *IEEE Transactions on Multimedia*, vol. 24, pp. 3060–3073, 2021. doi:10.1109/TMM.2021.3092579.

[2] Y. Cao, Z. Zhou, C. Zhu, P. Duan, and J. Li, "A lightweight deep learning algorithm for wifi-based identity recognition," *IEEE Internet of Things Journal*, vol. 8, pp. 17449–17459, 2021. doi:10.1109/JIOT.2021.3078782.

[3] H. Yang, Z. Xia, J. Shin, J. Hua, Y. Mao, and S. Zhong, "A comprehensive study of trajectory forgery and detection in location-based services," *IEEE Transactions on Mobile Computing*, vol. 23, pp. 3228–3242, 2024. doi:10.1109/TMC.2023.3273411.

[4] S. Zamboni, Z. T. Kefato, S. Girdzijauskas, N. Christoffer, and L. D. Col, "Pedestrian trajectory prediction with convolutional neural networks," *Pattern Recognition*, vol. 121, pp. 108252–108263, 2021. doi:10.1016/j.patcog.2021.108252.

[5] S. D. Regani, B. Wang, and K. Liu, "Wifi-based device-free gesture recognition through-the-wall," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021. doi:10.1109/ICASSP39728.2021.9414894.

[6] Q. Y, P. S, and Z. Li, "Cross-domain extendable gesture recognition system using wifi signals," *Electronics Letters*, vol. 59, 2023. doi:/10.1049/ell2.12931.

[7] C. S. Álvarez Merino, H. Q. Luo-Chen, E. J. Khatib, and R. Barco, "Wifi ftm, uwb and cellular-based radio fusion for indoor positioning," *Sensors*, vol. 21, p. 7020, 2021. doi:10.3390/s21217020.

[8] H. Gu, J. Yang, G. Gui, and H. Gacanin., "Triplet matchnet based indoor position method using csi fingerprint similarity comparison," *IEEE Transactions on Vehicular Technology*, vol. 72, pp. 16905–16910, 2023. doi:10.1109/TVT.2023.3289631.

[9] J. Liu, B. Jia, L. Guo, B. Huang, L. Wang, and T. Baker, "Ctsloc: An indoor localization method based on cnn by using time-series rssi," *Cluster Computing*, pp. 1–12, 2021. doi:10.1007/s10586-021-03458-2.

[10] X. Song, X. Fan, C. Xiang, Q. Ye, and G. Fang, "A novel convolutional neural network-based indoor localization framework with wifi fingerprinting," *IEEE Access*, vol. 7, pp. 110698–110709, 2019. doi:10.1109/access.2019.2933921.

[11] J. Li, T. Jiang, J. Yu, X. Ding, Y. Zhong, and Y. Liu, "An wifi-based human activity recognition system under multi-source interference," in *Springer*, vol. 878, pp. 937–944, 2022. doi:10.1007/978-981-19-0390-8_118.

[12] P. F. Moshiri, H. Navidan, R. Shahbazian, S. A. Ghorashi, and D. Windridge, "Using gan to enhance the accuracy of indoor human activity recognition," *China Communications*, 2020. doi:arxiv-2004.11228.

[13] Z. Tang, Q. Liu, M. Wu, W. Chen, and J. Huang, "Wifi csi gesture recognition based on parallel lstm-fcn deep space-time neural network," *China Communications*, vol. 18, no. 3, pp. 205–215, 2021. doi:10.23919/jcc.2021.03.016.

[14] H. Wang, S. Wang, X. Wang, T. Liu, and Y. Wang, "Rdts noise reduction: A fast method study based on signal waveform type," *Optical Fiber Technology*, vol. 65, no. 1, p. 102594, 2021. doi:10.1016/j.yofte.2021.102594.

[15] A. Leshem and U. Erez, "The interference channel revisited: Aligning interference by adjusting antenna separation," *IEEE Transactions on Signal Processing*, vol. 69, pp. 1874–1884, 2021. doi:10.1109/tsp.2021.3063442.

[16] F. Wang, J. Feng, Y. Zhao, X. Zhang, and J. Han, "Joint activity recognition and indoor localization with wifi fingerprints," *IEEE Access*, vol. 34, pp. 158–170, 2019. doi:10.1109/access.2019.2923743.

[17] Y. Chen, C. Wang, K. Xiong, and Z. Huang, "Synchronized perturbation elimination and doa estimation via signal selection mechanism and parallel deep capsule networks in multipath environment," *Chinese Journal of Aeronautics*, vol. 34, no. 4, pp. 158–170, 2021. doi:10.1016/j.cja.2021.01.016.

[18] J. Yang, Y. Liu, Z. Liu, Y. Wu, T. Li, and Y. Yang, "A framework for human activity recognition based on wifi csi signal enhancement," *International Journal of Antennas and Propagation*, vol. 2021, 2021. doi:10.1155/2021/6654752.

[19] M. Zhao, S. Zhong, X. Fu, B. Tang, and M. Pecht, "Deep residual shrinkage networks for fault diagnosis," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 7, pp. 4681–4690, 2020. doi:10.1109/tii.2019.2943898.

[20] Z. Hao, J. Niu, X. Dang, and Z. Qiao, "Wipg: Contactless action recognition using ambient wi-fi signals," *Sensors*, vol. 22, no. 1, p. 402, 2022. doi:10.3390/s22010402.

[21] Y. Ma, S. Arshad, S. Muniraju, E. Torkildson, and G. Zhou, "Location- and person-independent activity recognition with wifi, deep neural networks, and reinforcement learning," *ACM Transactions on Internet of Things*, vol. 2, pp. 1–25, 2021. doi:10.1145/3424739.

[22] Y.-S. Chen, Y.-C. Chang, and C.-Y. Li, "A semi-supervised transfer learning with dynamic associate domain adaptation for human activity recognition using wifi signals," *Sensors*, vol. 21, no. 24, p. 8475, 2021. doi:10.3390/s21248475.

[23] A. Alitaleshi, H. Jazayeriy, and J. Kazemitabar, "Affinity propagation clustering-aided two-label hierarchical extreme learning machine for wifi fingerprinting-based indoor positioning," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 6, pp. 3303–3317, 2022. doi:10.1007/s12652-022-03777-1.

[24] S. Xu and X. Yang, "A long term memory recognition framework on multi-complexity motion gestures," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2018. doi:10.1109/ICDAR.2017.41.

[25] F. Karim, S. Majumdar, H. Darabi, and S. Harford, "Multivariate lstm-fcns for time series classification," *Neural Networks*, vol. 116, pp. 237–245, 2018. doi:10.1016/j.neunet.2019.04.014.

[26] X. Zhang, C. Tang, Y. An, and K. Yin, "Wifi-based multi-task sensing," *Mobile and Ubiquitous Systems: Computing, Networking and Services*, vol. 419, pp. 169–189, 2022. doi:10.1007/978-3-030-94822-1_10.

Xu Lu is a Professor in the School of Computer Science, Guangdong Polytechnic Normal University, China. He received the B.S. degrees from Nanchang University, Jiangxi, China, in 2006, and the M.E. and Ph.D. degree from the Guangdong University of Technology, Guangdong, China, in 2009 and 2015, respectively. His research interests include artificial intelligence and smart system.

Zexiao Cai is currently pursuing a master's degree in Electronic Information at the Interdisciplinary Research Institute,Guangdong Polytechnic Normal University.His main research areas include healthcare and artificial intelligence.

Xiongwei Huang is a university faculty member in the Department of Electronic Information Engineering at Guangzhou Institute of Technology. He graduated from Guangdong Technical Normal University in 2023 with a master's degree in electronic information, and his interested research interests are sensor networks and deep learning.

Cheng Zhou is currently pursuing a master's degree in Systems Engineering at the School of Computer Science, Guangdong Polytechnic Normal University. His main research directions include the Internet of Things and artificial intelligence.

Jun Liu received the Ph.D degree in control science and engineering in 2015 from Guangdong University of Technology, Guangzhou, China. He is currently working as a associate professor in Guangdong Polytechnic Normal University, Guangzhou, China. His research interests mainly include sense and localization, intelligent mobile robot, IIoTs, collaborative sense and data optimization.