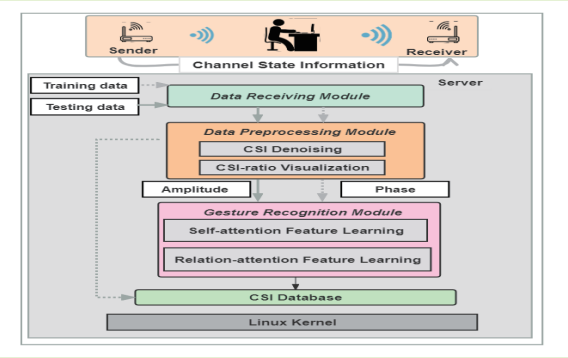
Attention-based gesture recognition using commodity WiFi devices

Yu Gu, *Senior Member, IEEE,* Huan Yan, Xiang Zhang, Yantong Wang, Jinyang Huang, Yusheng Ji, *Fellow, IEEE,* and Fuji Ren, *Senior Member, IEEE*

***Abstract*— The broad spectrum of applications of WiFi sensing technology, like gait and gesture recognition, has received widespread attention in recent years. Though most WiFi sensing systems may achieve impressive performance, the challenge lies in making good use of the amplitude and phase information of the Channel State Information (CSI) retrieved from commodity WiFi devices to carry out sensing tasks. To address this issue, we develop an attention-based framework to properly track the importance of amplitude and phase information to adaptively extract distinguishing features related to gestures. Specifically, we first use the CSI ratio instead of the original CSI as the basic signal, which not only eliminates most of the noise, but also contains the complete information of the CSI signal corresponding to human motion. Then, we use the self-attention module to learn the coarse attention weights of amplitude and phase information of the CSI ratio. Moreover, the relation-attention module is used to integrate features to further refine the attention weight. In this way, we proposed a framework that can adaptively learn distinctive feature representations and thus facilitate ubiquitous gesture recognition. Extensive experiments demonstrate the effectiveness of method for gesture recognition under various conditions on the open Widar3.0 dataset. The proposed method achieves 99.69% in-domain recognition accuracy, 96.95% cross-location recognition accuracy and 93.71% cross-orientation recognition accuracy, outperforming the state-of-the-art solutions.**



***Index Terms*— Gesture Recognition, Channel State Information, WiFi, Deep Learning, Attention.**

I. Introduction

THE development of gesture recognition technology is critical to the development of new human-computer interaction modes, and it also brings significant changes to people's life experiences [1]. In an indoor environment, people only need to set specific gestures, such as waving their arms in the air, and the device can make different responses by recognizing different gestures, such as controlling the corresponding smart home. Based on these advantages, the key is to develop a convenient and intelligent gesture recognition system that can play a major role in the field of human-computer interaction.

At present, in terms of technical means, gesture recognition is mainly divided into: First, based on wearable device, gesture motion information is collected mainly through sensors, and gesture recognition is performed through a specially designed algorithm [2]–[4]. Second, based on computer vision technology, the camera is used to capture images of user gestures, and the gesture status is recognized through image processing technology. However, this technical means will reduce the stability of the system under poor lighting conditions [5], [6]. Recently, WiFi-based sensing technology has been extensively used in indoor environments [7]–[14]. CSI readings retrieved from commercial WiFi devices can provide a wealth of sensory information. Therefore, WiFi-based gesture recognition has acquired extensive research attention. For example, WiGrus [15] uses commercial WiFi devices to capture the intrinsic characteristics of each gesture. Additionally, a Principal Component Analysis (PCA)-based method and the first order difference are employed to reduce the noise and mitigate multipath effects caused by the environment changes. WiGeR [16] designs a novel and agile segmentation and windowing algorithm based on wavelet analysis and short-term energy to reveal specific patterns associated with each gesture. The results show that WiGeR can classify gestures with high accuracy even when the signal passes through multiple walls. WiHF [17] derives the cross-domain motion change pattern of arm gestures from WiFi signals to realize cross-domain gesture recognition and user recognition. Widar3.0 [18] derives and estimates the speed curves of gestures at a lower signal level. These speed curves represent the unique dynamic characteristics of gestures and are domain-independent to realize a zero-effort cross-

Yu Gu and Fuji Ren are with I$^+$ Lab, School of Computer Science and Engineering, University of Electronic Science and Technology of China. E-mail: yugu.bruce@ieee.org, ren2fuji@gmail.com

Huan Yan (corresponding author), Xiang Zhang and Yantong Wang are with the School of Computer and Information, Hefei University of Technology, China. Email: yanhuan@mail.hfut.edu.cn, zhangxiang@mail.hfut.edu.cn, wangyantong912@mail.hfut.edu.cn, hjy@hfut.edu.cn

Yusheng Ji is with the National Institute of Informatics, Tokyo. Email: kei@nii.ac.jp

domain gesture recognition system. Nevertheless, most WiFi-based gesture recognition only uses the amplitude or phase information of the CSI readings, which causes the system to suffer from short sensing range and low accuracy [19], [20]. It is not difficult to observe the reason behind it carefully. When only amplitude information is used, the signal caused by the fine-grained motions is usually several orders of magnitude weaker than the Line-of-Sight (LoS) path signal. When the target is farther away from the transceivers, the reflected signal will become weaker. Moreover, due to the inherent characteristics of the device, the synchronization error between the transceivers further affects the CSI phase information, which is unstable when the human motions are directly sensed [21], [22].

To solve these problems of human motion perception based on WiFi, we first use the amplitude and phase information of the base signal called the CSI ratio as the signal source for network feature learning. It can be concluded from [19] that there are two benefits of using the CSI ratio: 1. The CSI ratio can eliminate the noise of the original amplitude and phase, and can provide a high-SNR base signal, which is more sensitive to human movements; 2. The CSI ratio provides orthogonal complementary amplitude/phase, and contains the complete information of the CSI signal corresponding to the human movement. State-of-the-art approaches only use the amplitude information or phase information of the CSI ratio for sensing tasks. For example, FingerDraw [23] is the first to use the phase information of the CSI ratio to implement the first sub-wavelength finger motion tracking system using ordinary WiFi devices. MultiSense [20] uses multiple antennas widely available on commercial WiFi devices to extract the phase information of the CSI ratio to simultaneously monitor the breathing of multiple people. At present, only FarSense has explored the complementarity of the amplitude and phase information of the CSI ratio. The combination scheme consists of two steps: (1) generating multiple combination candidates by assigning different weights to the I/Q components for combination; (2) selecting one from the candidates as the final extracted respiration pattern [22]. However, we observe that the method is not sufficient to explore the complementarity of the amplitude and phase information of the CSI ratio. Consequently, there is still redundant information in the extracted features.

In this paper, we design and implement a privacy-friendly system to realize gesture recognition with commodity WiFi devices in a contactless manner. The key enabler for such contactless WiFi human sensing is that CSI describes how the wireless signal is transmitted from the transmitter to the receiver. When a person performs daily activities, his/her body movement will cause changes in the reflection of the wireless signal, leading to changes in CSI. We first use the amplitude and phase information of the ratio of the CSI readings of the two antennas for each receiver as the signal source for network feature learning. According to [24], the same random phase offset between the same receiver antennas can be offset by calculating the CSI ratio. Moreover, according to [19], it can also eliminate the uncertain impulse noise in the CSI amplitude, which means that the CSI ratio can provide a base signal with a high signal-to-noise ratio, which is more sensitive to human movement. What we have solved is how to make full use of CSI amplitude and phase information retrieved from commercial WiFi devices to perform sensing tasks. However, CBAM mainly extracts meaningful attention features from the two dimensions of channel and spatial [25], which cannot effectively combine information from different signal levels. Motivated by [26], we introduce the attention mechanism to adaptively learn the importance of the amplitude and phase information of the CSI ratio to explore their complementarity and learn the distinctive features related to gestures. Specifically, we attach two attention modules to the backbone network of ResNet-18 [27]. One is *self-attention module* and the other is *relation-attention module*. For the self-attention module, we introduce the self-attention mechanism to simultaneously capture the importance of the feature map of the amplitude and phase information of the CSI ratio. A coarse attention weight is learned for the extracted features through a fully connected layer and a sigmoid function, and then the refined features for each input signal source are obtained by multiplying the initial features and the learned coarse attention weight. For the relation-attention module, we cascade the two features obtained through the self-attention mechanism to obtain the global feature, and then integrate the global feature and the features extracted by each signal source. The self-attention mechanism is used to model the relationship between local features and global representations to further refine the attention weight. Finally, the feature vector obtained by the two-way refined attention weight is merged to further enhance the feature representation.

The main contributions of this paper are summarized below:

- We propose a scheme that uses the attention mechanism to successfully implement gesture recognition in different locations and directions on commodity WiFi devices. Our proposed scheme can achieve the best recognition performance compared to the state-of-the-art rivals.
- We use the self-attention module and relation-attention module that can explore the amplitude and phase information of the CSI ratio, and better extract the refined feature representations related to the gestures.
- We have implemented gesture recognition systems on COTS WiFi devices, and conducted extensive experiments on the public Widar3.0 dataset. The results demonstrate the effectiveness of our system for gesture recognition in different locations and different directions.

The rest of this paper is structured as follows. Section II reviews some representative works using WiFi for gesture recognition. Section III presents the details of the proposed method. The experimental settings and results are provided in Section IV. Conclusions are provided in Section V.

II. RELATED WORK

In this section, we briefly review some related work on wearable sensor based, camera based, and WiFi-based gesture recognition in the literature.

TABLE I: WiFi-enabled Gesture Recognition

Reference	Gesture	Preprocessing	Algorithm	Dataset	User	No. of Gestures	Sample	Performance
WiGest, 2015 [28]	Hand Gesture	Wavelet Denoising	Pattern Encoding; Gesture Matching	Private	3	7	>1000	87.5% using one AP; 96% using three APs.
Tan-WiFinger, 2016 [29]	Finger Gesture	Wavelet based Denoising	STFT; Muti-Dimensional DTW	Private	-	8	-	Over 90% recognition accuracy.
CrossSense, 2018 [30]	Hand&Finger Gesture	-	Artificial Neural Network	Private	100	40	1,200,000	Over 90% recognition accuracy.
WiADG, 2018 [31]	Hand Gesture	-	Artificial Neural Network	Private	2	6	2500 CSI frames	Over 90% recognition accuracy.
Yang et al., 2019 [32]	Hand&Finger Gesture	-	Artificial Neural Network	Private	10	40	2400	Over 80% recognition accuracy.
Widar3.0, 2019 [18]	Hand and Arm Gesture	Conjugate Multiplication Denoising	BVP Extraction; Neural Network	Public (WiDar3)	16	9	15375	92.7% in-domain; 89.7%, 82.6% and 92.4% cross domain.
WiHF, 2020 [17]		Conjugate Multiplication Denoising	Motion Change Pattern Extraction; Dual-Task Neural Network					92.7% in-domain; 89.07% cross-location and 82.6% cross-orientation
WiGrunt, 2022 [33]		CSI-Ratio Based Denoising	CSI Visualization; Attention Based Neural Network					99.67% in-domain; 96% cross-location and 92.6% cross-orientation.
Ours		CSI-Ratio Based Denoising	CSI Visualization; Attention Based Neural Network					99.69% in-domain; 96.95% cross-location and 93.71% cross-orientation.

A. Wearable sensor based gesture recognition

Most research requires users to wear devices such as inertial sensors, gyroscopes, etc. for gesture recognition. For example, Gozzi et al. [34] used the EMG signals of the upper limb muscles to be recorded during the movement process for gesture recognition. Moin et al. [1] proposed a self-contained, wearable sEMG biosensing system that uses HD computing to process and classify hand gestures. Zhang [35] proposed a gesture recognition system based only on pressure sensors, wrist pressure, which can reflect different pressure changes of different gestures. Although sensor based gesture recognition has achieved good performance in human-computer interaction applications, there are still some limitations in some aspects. For example, some sensing devices need to contact users directly to capture the behavior characteristics related to gesture actions, so that users' gesture actions will affect their behavior habits because of contacting the sensing devices. At the same time, in continuous interaction, frequent wearing of sensing devices will reduce the comfort of user interaction, and ultimately lead to poor user experience. In addition, for some specific sensing devices, although they can improve the user experience during interaction, they cannot be widely used in actual scenarios due to their high cost.

B. Camera based gesture recognition

The camera vision based sensor is a common, suitable and applicable technique because it provides contactless communication between humans and computers [36], [37]. Jiang et al. [38] established a convenient and effective binocular vision system that can accurately extract gesture information from complex environments. Devineau et al. [39] introduced a new Convolutional Neural Network (CNN) to process the position sequence of hand skeleton joints through parallel convolution, and then studied the performance of the model in the task of hand gesture sequence classification.As the visual technology has been relatively mature, the gesture recognition based on computer vision has a high recognition accuracy. However, there are still some shortcomings in the actual scene. For example, the gesture recognition performance deteriorates when the lighting conditions are not good or there are obstacles between the user and the camera. In addition, in areas involving privacy protection, cameras cannot be used to collect images, so gesture recognition cannot be realized. In recent years, a large number of researchers have begun to implement gesture recognition on WiFi devices, which not only solves the limitations of the above two methods, but also realizes gesture recognition with high accuracy.

C. WiFi-based gesture recognition

Recently, some methods of using CSI retrieved from commodity WiFi devices have been proposed for gesture recognition. For example, Abdelnasser et al. [28] early proposed a system where WiFi signal strength changes to perceive air gestures around the user's mobile device. The system recognizes different signal change primitives and constructs independent gesture families based on this, so there is no need for any training process to directly perform gesture recognition. Sheng et al. [29] designed an environmental noise cancellation mechanism to reduce the dynamic impact of signals caused by environmental changes and capture inherent gesture behaviors to deal with individual differences and inconsistencies in gestures to achieve fine-grained gesture recognition. Zhang et al. [30] extended WiFi perception to new environments and larger problem scales in gait recognition and gesture recognition. Widar3.0 [18] extracted the domain-independent feature representation BVP from the CSI, and then feeds it into a well-designed one-fits-all model and only one-time training can achieve good results. In order to adapt to the model well in the case of very limited samples, Ding et al. [40] proposed a Channel–Time–Subcarrier Attention Mechanism (CTS-AM) enhanced few-shot learning method to complete feature representation and recognition tasks. Yang et al. [32] proposed a novel deep Siamese representation learning architecture for one-shot gesture recognition, which is superior to existing solutions in terms of time-space representation learning, and achieves satisfactory results under one-shot conditions. WiADG [31] can accurately and consistently recognize human gestures in the dynamic environment through adversarial domain adaptation, achieving 98% gesture recognition accuracy in the original environment. Nevertheless, these WiFi-based gesture recognition methods only used the amplitude or phase information of the CSI readings, which caused the system to suffer from short sensing range and low accuracy [19], [20]. It is not difficult to observe the reason behind it carefully. When only amplitude information is used, the signal caused by the fine-grained motions is usually several orders of magnitude weaker than the Line-of-Sight (LoS) path signal. When the target is farther away from the transceivers, the reflected signal will become weaker. Moreover, due to the inherent characteristics of the device, the synchronization error between the transceivers further affects the CSI phase information, which is unstable when the human motions are directly sensed [21], [22]. Table I compares and concludes some representative prior works related to WiFi-sensing enabled gesture sensing and recognition.

III. SYSTEM DESIGN

In this section, the proposed methods for gesture recognition are detailed. The system architecture is shown in Figure 1, which mainly includes three modules: data acquisition, data preprocessing and gesture recognition. In the data acquisition module, CSI measurements can be recorded at the receiver with open source tools only in environments where multiple wireless links are deployed. Then, as we study the impact of CSI retrieved from commodity WiFi devices on gesture recognition, the first stage is to introduce the background and challenges of using CSI for gesture recognition. Section III-A introduces the background information about CSI, and then we demonstrated the feasibility of using gesture change patterns for gesture recognition in a WiFi environment. In the data preprocessing module, we use CSI ratio as the input signal, which not only eliminates the noise in CSI amplitude and random offset in CSI phase, but also quantifies the correlation between the dynamic CSI value and human activities. If the original CSI ratio is directly input into the neural network, it will be unfavorable for network learning, so we need to visualize the CSI ratio to generate the amplitude and phase images of CSI ratio. Then the attention-based framework is proposed to identify gestures based on the self-attention and relation-attention module through amplitude and phase information of the CSI ratio. In Section III-B and III-C, the specific details of data preprocessing and gesture recognition framework based on attention mechanism are elaborated respectively.

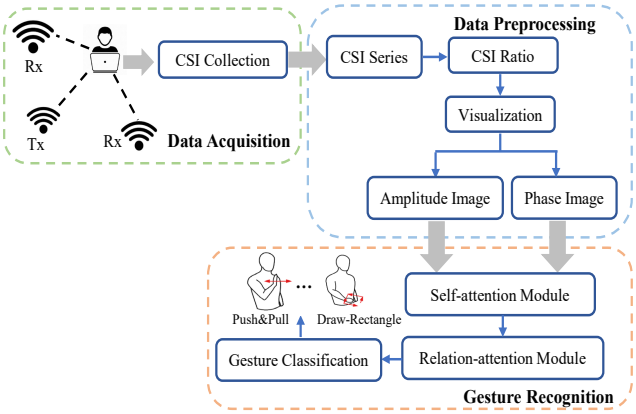
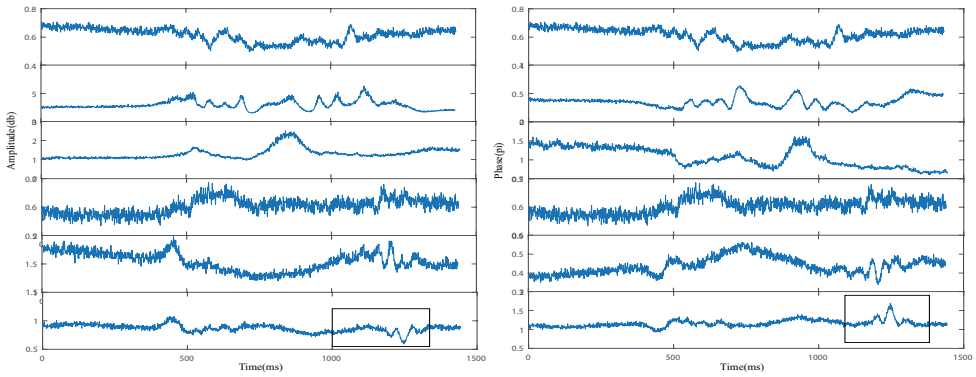


Fig. 1: The system architecture.

A. Background

1) CSI primer: CSI is the information used to estimate the channel characteristics of a communication link. It is used to characterize how the signal propagates from the transmitting to the receiving, combining various effects such as time delay, amplitude attenuation and phase offset, and describing the amplitude and phase of each subcarrier in the frequency domain. Let $X(f,t)$ be the transmitted signal at time t and carrier frequency f in the frequency domain, and $Y(f,t)$ be the received signal of the carrier frequency f at time t in the frequency domain. The relationship of the channel state $H(f,t)$ can be expressed as $Y(f,t) = H(f,t) \times X(f,t)$, where $H(f,t)$ represents the complex-valued Channel Frequency Response (CFR). It is well known that the received signal strength is the superposition of signals from multiple paths when sensing human activity in indoor environments due to multipath effects [21], [41]. Furthermore, the uncertainty of the power amplifier in the Radio Frequency (RF) chain due to the interference of hardware devices often causes impulse noise and burst noise in the CSI amplitude. The difference in carrier frequency between the transceivers results in a time-varying phase shift of each CSI sample, which quickly



(a) (b)

Fig. 2: (a)(b) CSI amplitude and phase for user 1 performing gesture *push*.

accumulates over time and disrupts the phase changes caused by human activity. Therefore, the total CSI can be denoted as follows:

$$H(f,t) = A(f,t)e^{-j\theta_0(f,t)} \sum_{i=1}^{L} a_i(f,t)e^{-j2\pi \frac{d_i(t)}{\lambda}}, \quad (1)$$

where L is the number of propagation paths, a_i and $\frac{d_i(t)}{\lambda}$ are the complex attenuation and propagation delay of the i-th path, respectively. λ is the wavelength for the carrier with frequency f, $A(f,t)$ is the impulse noise in amplitude, and $\theta_0(f,t)$ is the random phase error caused by timing alignment offset, sampling frequency offset and carrier frequency offset. CSI is usually divided into static components and dynamic components. The former is composed of LoS propagation and other reflection paths from static objects in the environment, and the latter is the path reflected by the subject's actions [21]. Thus CSI can be transformed as follows:

$$\begin{aligned} H(f,t) &= A(f,t)e^{-j\theta_0(f,t)}(H_s(f,t) + H_d(f,t)) \\ &= A(f,t)e^{-j\theta_0(f,t)}(H_s(f,t) + \sum_{i\in P_d} a_i(f,t)e^{-j2\pi \frac{d_i(t)}{\lambda}}), \end{aligned} \quad (2)$$

where the constant $H_s(f,t)$ is the sum of all static signals, $H_d(f,t)$ is the sum of all dynamic signals, P_d is the set of dynamic signals (e.g., signals reflected by the subject's actions).

2) Gesture recognition with WiFi CSI: Now let's take a closer look at how to use CSI for gesture recognition. As mentioned above, we describe the wireless signal from the sender through multiple paths to the receiver, and record the physical space characteristics it passes through. As a person moves through physical space, the body reflects and diffracts signals to introduce additional paths (NLoS paths). Therefore, the influence of human activities on the propagation of RF signals will be described by the signals reaching the receiver. By establishing the mapping relationship between the changes of these signals and different human activities, gesture recognition based on CSI can be realized. Figure 2 shows a comparison of CSI amplitude and phase waveform of user 1 performing gesture *push* in the Widar3.0 dataset. Note that here we only give the amplitude and phase results for a subcarrier of 1. We can see that both amplitude and phase gestures affect signal fluctuations, which is consistent with some existing work [42]. Although both CSI amplitude and phase can be mapped to the same gesture, gesture recognition can be achieved by extracting the distinguishing features associated with gesture and then training the classifier. However, the original CSI amplitude is limited by the sensing accuracy, and when the human movement is directly sensed, the synchronization error further affects the instability of CSI phase information. Thus, using CSI to achieve gesture recognition is feasible [17], [42], but it requires a more distinguishable base signal than the original CSI amplitude and phase, and the complementarity of amplitude and phase information can be better explored.

B. Data preprocessing

As mentioned earlier, since the CSI readings retrieved from different antennas of the same receiver contain very similar hardware noise, the amplitude and phase are not ideal CSI signals. Some recent studies use a new base signal called the CSI ratio, which eliminates the noise in the CSI amplitude and the random offset in the CSI phase, and quantifies the correlation between the dynamics of the CSI value and the human activities, thereby improving WiFi sensing performance [19], [20], [23]. As is implied by the name, the CSI ratio is defined as the quotient of the CSI readings of two adjacent antennas of the same receiver, so the CSI ratio is expressed as:

$$\begin{aligned} H_r(f,t) &= \frac{H_1(f,t)}{H_2(f,t)} \\ &= \frac{A(f,t)e^{-j\theta_0(f,t)}(H_{s,1}(f,t) + H_{d,1}(f,t))}{A(f,t)e^{-j\theta_0(f,t)}(H_{s,2}(f,t) + H_{d,2}(f,t))} \\ &= \frac{H_{s,1}(f,t) + \sum_{i1\in P_d} a_{i1}(f,t)e^{-j2\pi \frac{d_{i1}(t)}{\lambda}}}{H_{s,2}(f,t) + \sum_{i2\in P_d} a_{i2}(f,t)e^{-j2\pi \frac{d_{i2}(t)}{\lambda}}}, \end{aligned} \quad (3)$$

where $H_1(f,t)$ and $H_2(f,t)$ represent the CSI readings corresponding to the two receiving antennas of the same receiver, respectively. Since the power ratios of different antennas on the same receiver are the same and share the same clock, that is, $A(f,t)e^{-j\theta_0(f,t)}$ for different antennas on the same receiver is

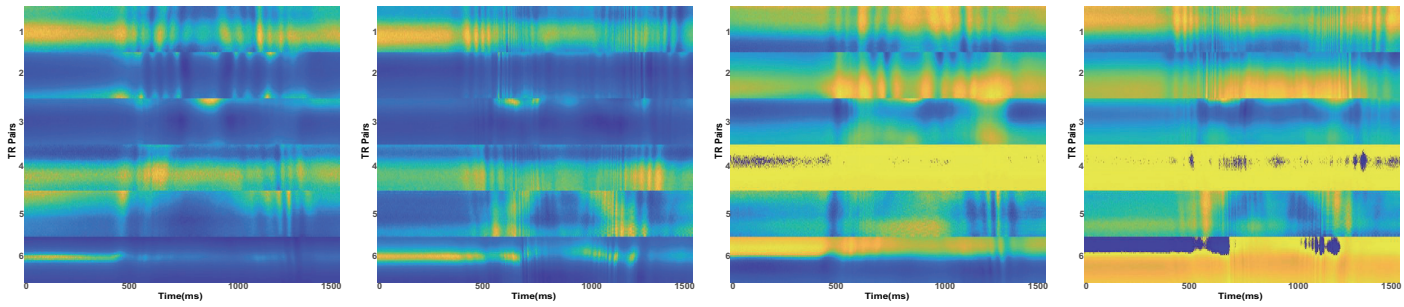


Fig. 3: (a)(c) CSI ratio amplitude and phase image of gesture *push* performed by user 1. (b)(d) CSI ratio amplitude and phase image of gesture *sweep* performed by user 1.

equal. Through the division operation, the random phase shift and the pulse amplitude noise can be canceled.

To explore the complementarity of the amplitude and phase of the CSI ratio, we use deep learning to properly track the importance of the amplitude and phase information of the CSI ratio, and adaptively extract distinguishing features related to gestures. According to the existing work foundation [18], [43], the use of CSI ratio visualization rather than the original CSI ratio value is more suitable for deep learning. Assuming that there are N_t transmitting antennas, N_r receiving antennas, N_s subcarriers, and T packets during signal transmission, the CSI ratio vector $H \in \mathbb{R}^{N_t \times N_r \times N_s \times T}$ can be obtained. Therefore, we can map the amplitude and phase values of the CSI ratio to the pixels of the image (i.e., the amplitude and phase heatmap), and the image dimension is $(N_t \times N_r \times N_s) \times T \times 1$(the image height and width are $N_t \times N_r \times N_s$ and T, respectively, and the depth is 1). We visualize the complex value of the CSI ratio as amplitude and phase images as shown in Figure 3. Through the visual operation, we can observe the difference in amplitude and phase information of the received CSI ratio when user 1 performs different gestures. Intuitively, different gestures have individualized action understanding and presentation styles, and the amplitude and phase images are complementary.

TABLE II: Definitions of symbols in defining the attention network.

Name	Definition
α_1, α_2	Coarse attention weights generated in the self-attention module
β_1, β_2	Fine-grained attention weights generated in the relation-attention module
f	Fully connected layer and sigmoid function operation
F_1, F_2	The deep features generated by the backbone network correspond to different inputs
F_{sa}	Fusion feature
FC	Fully connected layer
σ	Sigmoid function
$concat$	Concatenate operation
$\mathbf{W_1^T}, \mathbf{W_2^T}$	The parameter of the FC layer
R_i	Relation feature
β_i	Relational attention weight corresponding to the relation feature

C. Gesture recognition

After data preprocessing, we use the attention feature learning model to explore the complementarity of the amplitude and phase of the CSI ratio. Figure 4 illustrates the overall structure of the attention network feature learning model. Specifically, the amplitude and phase images of the CSI ratio are first normalized, which is to resize the dimensions of the input image to $224 \times 224 \times 3$ (the width and height of the image are both 224, and the depth is 3). The normalized output is then input into attention feature learning model, which from left to right consists of backbone network for feature extraction, self-attention module for learning the coarse attention weights of the amplitude and phase information of the CSI ratio (III-C.1), and relation-attention module is used to integrate global and local characteristics and further refine the weight of attention (III-C.2).

The model used is a result of the effectiveness of amplitude and phase complementarity. With amplitude and phase images as input, the attention feature learning model can adaptively learn fine-grained distinctive feature representations and thus facilitate ubiquitous gesture recognition under various conditions (i.e., different orientations and locations). We will verify the effectiveness of our proposed method for gesture recognition through extensive experiments in IV. The definition of symbols in the network is shown in Table II.

1) Self-attention feature learning module: Convolutional Neural Networks (CNN) have become the mainstream solution for many tasks (e.g., classification) due to their local perception and parameter sharing for automatic feature learning [44]. Inspired by CNN, given that I_a and I_p are expressed as the amplitude and phase images of the CSI ratio, we first adopted the widely used ReNet18 backbone network with good performance to automatically extract their deep features. However, the learned deep features are not fused and directly used for gesture classification will not be effective. To learn the coarse importance weights of different input data streams, we use a fully connected layer (FC) and a sigmoid function to learn the coarse attention weights of the amplitude and phase information of the CSI ratio, that is:

$$\alpha_i = \sigma(\mathbf{W_1^T} F_i), i = 1, 2, \tag{4}$$

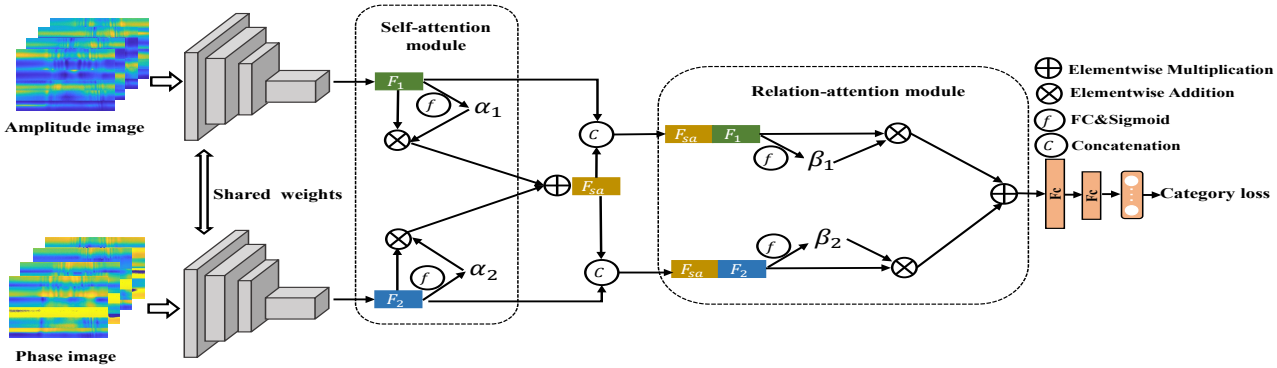


Fig. 4: The process of attention network feature learning.

where α_i is expressed as the coarse attention weight corresponding to the i-th deep feature F_i, σ denote the sigmoid function, and $\mathbf{W_1}$ indicates the parameters of the FC layer. Finally, combining the coarse importance weights and the input features, we can obtain the fusion feature F_{sa} of the amplitude and phase of the CSI ratio, that is:

$$F_{sa} = \sum_{i=1,2} \alpha_i F_i. \tag{5}$$

2) Relation-attention feature learning module: In the self-attention feature learning module, it mainly learns the coarse attention weight to extract the important features (i.e., the local features) in each input data stream (i.e., amplitude and phase). However, if the local features of different input data streams are directly fused and used as the final output features, the bridge of learning complementarity between amplitude and phase information cannot be established. Therefore, we need to integrate local features and global features to learn the complementarity of different input data streams. Specifically, we first concatenate the features of each input stream with the global features, the role of which is to refine the relational attention weight, that is:

$$R_i = concat(F_{sa}, F_i), i = 1, 2, \tag{6}$$

where $concat$ represents the concatenate operation, R_i denotes the relation features of amplitude and phase. In order to further refine the relational attention weights for the amplitude and phase, we again use the FC layer and a sigmoid function to learn, that is:

$$\beta_i = \sigma(\mathbf{W_2^T} R_i), i = 1, 2, \tag{7}$$

where β_i is expressed as relational attention weight corresponding to the i-th relation feature R_i, and $\mathbf{W_2}$ indicates the parameters of the FC layer. Finally, the relation features generated by the amplitude and phase image of the CSI ratio are added together to obtain the final gesture recognition feature, that is:

$$R = \sum_{i=1,2} \beta_i R_i. \tag{8}$$

IV. IMPLEMENTATION AND EVALUATION

We first briefly introduce the training and inference details, as well as the Widar3.0 dataset and implementation details, and compare our method with state-of-the-art methods, and then analyze the importance of each module through ablation experiments. Finally, we perform deep feature visualization to show the effectiveness of the proposed model.

A. Tranining and inference details

Given a gesture recognition CSI dataset, we have a training set D_{tr} with N samples: $D_{tr} = \{x_i^a, x_i^p, y_i\}_{i=1}^N$ where x_i^a and x_i^p represent the amplitude and phase images of the CSI ratio, respectively, and y_i represents the corresponding gesture label. We optimize the network parameters based on the backpropagation algorithm [33] on the training set, and the minimized cross-entropy (CE) loss function that needs to be optimized for the entire network is:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N}\sum_{c=1}^{C} \Pi_{[i=y_{ic}]} log(\mathcal{F}(x_{ic}, \mathbf{w})), \tag{9}$$

where C is the number of gestures, c is the category index, $\Pi_{[i=y_{ic}]}$ outputs 1 when $i = y_{ic}$, and 0 otherwise, $\mathcal{F}$ represents the proposed model with a parameter of $\mathbf{w}$ and the output is the gesture recognition probability. In the test phase, given a gesture CSI reading, we first obtain the corresponding CSI ratio amplitude and phase images according to *data preprocessing module*, then extract the distinguishing features related to the gesture through *self-attention feature learning module* and *relation-attention feature learning module*, and finally predict the final score vector.

B. Dataset and implementation details

Widar3.0 is a CSI-based gesture recognition dataset, which is collected in different locations and directions in each sensing area. Widar3.0 mainly includes two types of datasets. The first is 6 gestures commonly used in human-computer interaction, with a total of 12,000 gesture samples (16 users×5 positions×5 orientations×6 gestures×5 instances). The second is the 0∼9 gesture on the horizontal plane, there are a total of 5,000 gesture samples (2 users×5 positions×5 orientations×10 gestures×5 instances). To enable a fair comparison with WiHF [17], we only use 4,500 gesture samples (6 users×5 positions×5 orientations×6 gestures×5 instances) in the Widar3.0 dataset for in-domain, cross-orientation, and

cross-location gesture recognition. Moreover, we select all gesture data (9 users×5 positions×5 orientations×9 gestures×5 instances) in the 1 environment to verify the performance of the system under more users and more gestures. The sketches of the nine gestures from [18] are plotted in Figure 5.

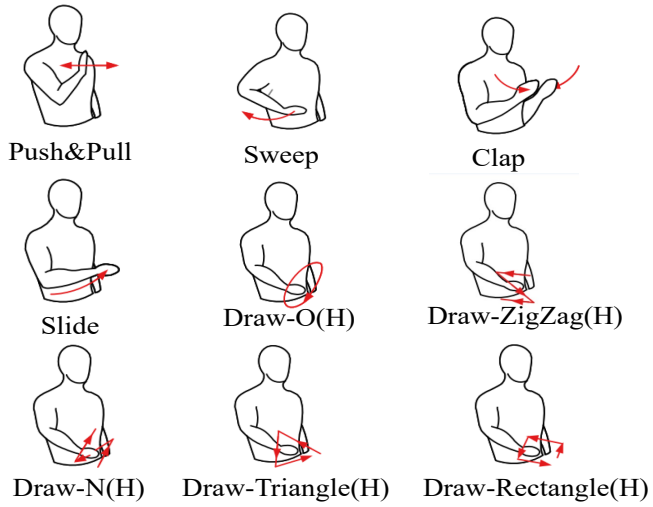


Fig. 5: Sketches of gestures evaluated in the experiment.

To make a fair comparison, we set the in-domain evaluation, cross location evaluation and cross direction evaluation as follows: in-domain evaluation: we randomly select 80% of all data as the training set, and the rest as the test set; cross-location evaluation: we randomly select 4 locations gesture data as the training set, and the rest as the test set; cross-orientation evaluation: we randomly select 4 orientations gesture data as the training set, and the rest as the test set. Regardless of the evaluation, we use 5-fold cross-validation as the final result. Moreover, Resnet-18 [27] has been widely used as a backbone network in the field of object detection [45], [46], image classification [26], [47] and wireless sensing [33], [48] because of its simple and practical advantages. Compared with other networks, Resnet-18 is a relatively small network with low parameters [45]. On the widar3.0 dataset, we refer to previous research work [33], [49], [50], so we choose Resnet-18 as the backbone network. We use Stochastic Gradient Descent (SGD) algorithm [51] with the momentum of 0.9, a learning rate of 0.001, a batch size of 64. The model is trained on a single NVIDIA GTX 2080 Ti using Pytorch for 50 epochs.

C. Overall performance

Table III reports the performance of our method on the gesture recognition task performed using 5-fold cross-validation on the Widar3.0 dataset. From Table III, we have the following observations:1) In the 5-fold cross-validation, the gesture recognition rate fluctuates slightly in the evaluation situation within the domain, while the gesture recognition rate fluctuates greatly in the cross-location and cross-orientation evaluation. The fluctuations in the in-domain, cross-location and cross-orientation evaluation are 0.45%, 6.44% and 7.33%, respectively. This is because the feature gaps in different locations or different orientations are large. 2) The overall gesture recognition performance of the in-domain evaluation is the largest, which is 99.69%, and the cross-orientation evaluation is the second, which is 96.95%, and the cross-location evaluation has the lowest gesture recognition performance, which is 93.71%. The reason is that the data collected in different environments has *domain shift* due to different data distributions [52].

TABLE III: Overall performance for gesture recognition using 5-fold cross-validation on Widar3.0 dataset.

Settings	1	2	3	4	5	Mean
In-domain	99.44%	99.67%	99.56%	99.89%	99.89%	99.69%
Location	97.56%	92.89%	97.67%	97.33%	99.33%	96.95%
Orientation	89.56%	95.33%	94.89%	96.89%	91.89%	93.71%

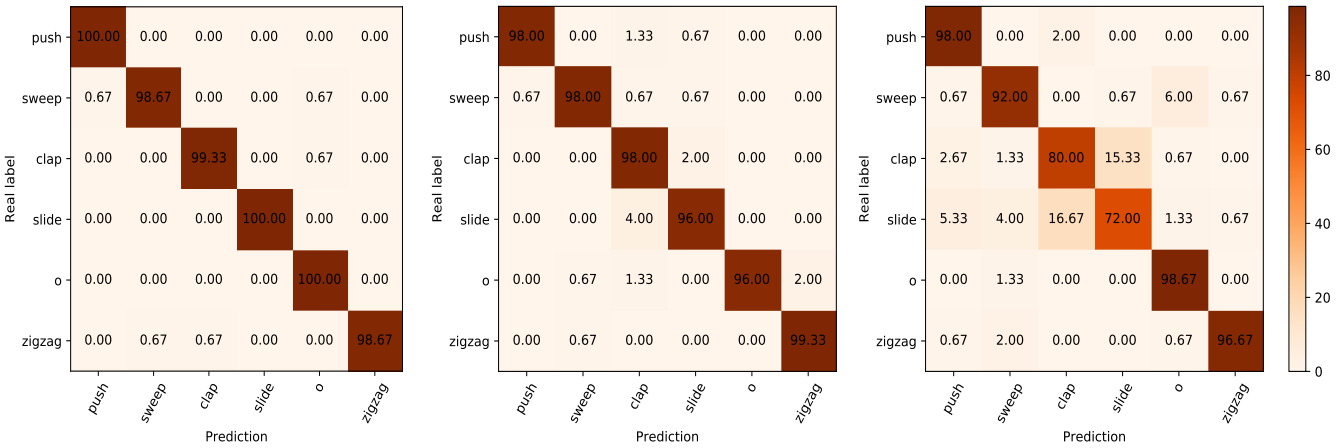
Figure 6 further shows the confusion matrix considering each specific setting. Note that we only show the case where the number of gestures in different settings is 6 and the first fold result in the 5-fold cross-validation. For each domain factor, we calculate the average accuracy of all gesture recognition after 5-fold cross-validation. We can see that our method achieves consistently high performance in different domains, which proves its ability to recognize across domains.

D. Performance comparison for different methods

To comprehensively evaluate the effectiveness of our method, we compared it with state-of-the-art methods of WiHF, FarSense and WiGRUNT. To give a thoroughly analysis, three aspects, including in-domain, cross-location, cross-orientation, the gesture recognition results are compared in Table IV (the default number of gesture types is 6). We can see that our performance is better than WiHF , FarSense and WiGRUNT in terms of in-domain settings, cross-location settings and cross-orientation settings. FarSense [22] uses different I/Q component weights to generate multiple combination candidates for feature extraction, and selects the final feature extraction method from these candidates. However, it does not adaptively combine amplitude and phase information compared to using CNN methods. WiHF [17] derives the cross-domain motion change pattern of arm gestures from WiFi signals to realize cross-domain gesture recognition and user recognition. The above method achieves good performance, but they ignore the complementarity between the amplitude and phase of the CSI ratio. And WiGRUNT also ignores the complementarity of amplitude and phase of the CSI ratio. In contrast, we developed an attention-based framework to adaptively track the importance of the amplitude and phase information of the CSI ratio to propose a distinctive feature representation related to the gestures, which leads to an improvement in performance.

E. Performance comparison of different gestures and number of users

To further demonstrate the effectiveness of our method, we also show the performance of the system in gesture recognition and user identification in the case of more users or more



(a) (b) (c)

Fig. 6: Confusion matrices of different settings: (a) In-domain (99.44%); (b) Cross-location (97.56%); (c) Cross-orientation (89.56%).

TABLE IV: Comparison to the state-of-the-art results on gesture recognition.

Methods	Model types	#Settings		
		In-domain	Location	Orientation
Widar3.0 [18]	CNN+GRU	92.7%	89.07%	82.6%
WiHF [17]	CNN+GRU	97.65%	92.07%	82.38%
FarSense [22]	-	98.12%	87.97%	84.68%
WiGRUNT [33]	CNN	99.67%	96%	92.6%
Ours	CNN	**99.69%**	**96.95%**	**93.71%**

gestures. Note that only the results of gesture recognition are reported in [18] (the default number of gesture types is 6).

Results are shown in Table V. In in-domain setting, we can see that when the number of gestures is 6, the gesture recognition accuracy of our method is 99.69% close to the WiGRUNT gesture recognition result, which is 99.67%. In the number of other gestures, we can find that our method can lead WiGRUNT by 3.92% at most in in-domain setting. In the cross-location and cross-orientation settings, whether it is gesture recognition or user identification, the proposed method consistently achieves higher recognition accuracy than WiGRUNT and WiHF. For example, In in-domain setting, we can see that when the number of users is 6, the user recognition accuracy of our method is 96.7% close to the WiHF user recognition result, which is 96.74%. In the number of other users, we can find that our method can lead WiHF by 4.64% at most in in-domain setting. Moreover, our method can achieve the best performance regardless of the increase in the number of users or the number of gestures.

F. Deep feature visualization

To prove that the proposed method can provide effective gesture-related distinguishing features, we further analyze the effectiveness of the method through deep feature visualization. Specifically, we use t-SNE [53] to map the expression features in the model to a two-dimensional space to visualize the distinguishing ability of different gestures deep features. Figure 7 shows the deep features of the CSI ratio under in-domain, cross-location and cross-orientation setting in our proposed method (i.e., the G distribution in Figure 4). Note that we only use the test data with 6 gestures for deep feature visualization. We can observe that the method proposed by us can include separate data clusters in the feature space under any environment, and gesture samples in the same cluster contain most of the same class labels. Compared with cross-orientation and cross-location, the separability between different gestures in the domain is stronger. In addition, no matter what kind of environment, *clap* and *slide* are easy to approach and misjudge in the feature space.

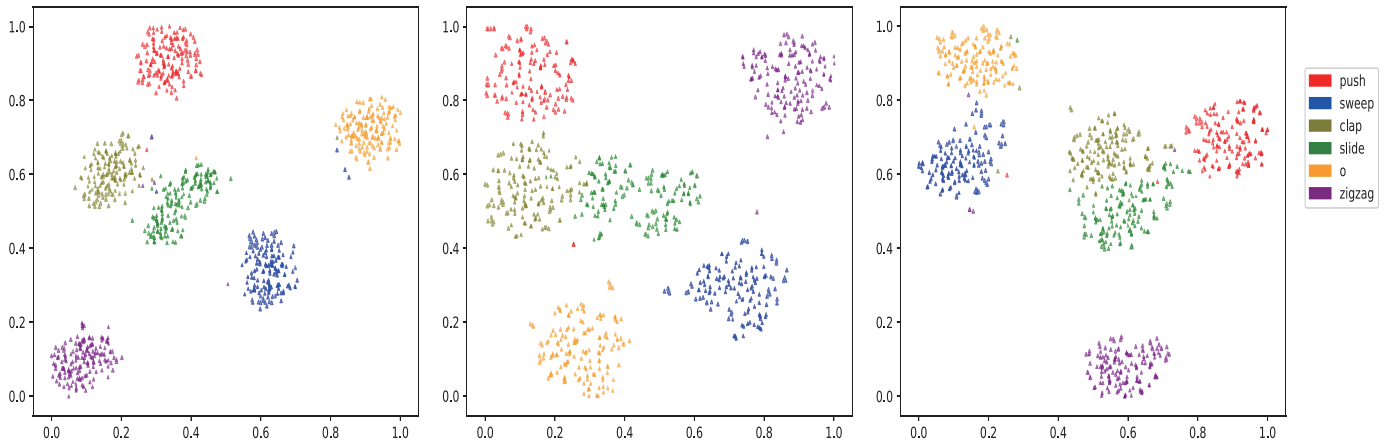
V. CONCLUSION

In this paper, we develop an attention-based framework to properly track the importance of amplitude and phase information of the CSI ratio to adaptively extract distinguishing features related to gestures. Specifically, we first use the self-attention module to learn the coarse attention weights of amplitude and phase information of the CSI ratio. Then we use the relation-attention module to integrate global and local features to further refine the attention weight. Extensive experiments demonstrate the effectiveness of our proposed method for gesture recognition under various conditions (i.e., different locations and orientations) on the open Widar3.0 dataset. The proposed method achieves 99.69% in-domain recognition accuracy, 96.95% cross-location recognition accuracy and 93.71% cross-orientation recognition accuracy, outperforming the state-of-the-art solutions.

Although our proposed attention-based framework can improve gesture recognition performance, it has certain limitations in model deployment. There are two reasons. First, in the actual scene, If we input the whole input signal into the network for learning, there will be some signals unrelated to

TABLE V: Impact of number of gestures and users.

Methods	Settings	#Gesture				#User			
		6	7	8	9	6	7	8	9
Widar3.0 [18]	In-domain	92.7%	-	-	-	-	-	-	-
	Location	89.07%	-	-	-	-	-	-	-
	Orientation	82.6%	-	-	-	-	-	-	-
WiHF [17]	In-domain	97.65%	96.14%	95.33%	93.11%	96.74%	92.56%	93.76%	94.43%
	Location	92.07%	85.81%	84.92%	83.81%	75.31%	66.98%	64.70%	65.26%
	Orientation	82.38%	74.46%	72.72%	74.55%	69.52%	63.26%	55.86%	57.26%
WiGRUNT [33]	In-domain	99.67%	99.16%	98.08%	98.12	99.67%	99.45%	99.48%	99.1%
	Location	96%	94.99%	91.9%	92.87%	96%	96.17%	96.72%	96.21%
	Orientation	92.6%	91.64%	89.01%	87.99%	92.6%	92.25%	93.58%	93.63%
Ours	In-domain	99.69%	99.58%	99.06%	98.77%	96.72%	96.68%	97.2%	97.5%
	Location	96.95%	96.74%	94.21%	94.19%	83.01%	84.73%	82.53%	87.26%
	Orientation	93.71%	92.53%	89.85%	91.91%	89.8%	92.92%	92.21%	92.41%



(a) (b) (c)

Fig. 7: Deep feature visualization on the deep features of the CSI ratio under in-domain, cross-location and cross-orientation setting.

gestures that also participate in the network learning, which will eventually lead to an increase in the amount of network learning. Secondly, in the actual model deployment, model pruning should be used to minimize redundant information and make a balance between performance and model size. In the future, we will consider the problem of gesture action cutting in real-time systems, and at the same time enrich the types of gestures to build a real-time smart home gesture interaction system. In this paper, the amplitude and phase information of CSI ratio are effectively combined to improve gesture recognition performance, which is verified on the public dataset Widar3.0. Therefore, there will be some limitations in the position of the human body relative to the transceiver system and in the presence of obstacles (such as walls). When the position of the transceiver system changes, the main challenge is to extract the position independent feature representation to represent different gesture actions. In the scene where obstacles (such as walls) exist, the pre-processing method using the existing CSI data cannot better eliminate the interference of obstacles to CSI data, so more robust pre-processing methods are needed to eliminate this interference.

ACKNOWLEDGMENT

This work was supported by the Key Research and Development Plan of Anhui Province (202004b11020018), Fundamental Research Fund of Chinese Academy of Medical Sciences (2020-JKCS-002), Open Fund of Anhui Province Key Laboratory of Cyberspace Security Situation Awareness and Evaluation (CSSAE-2021-009).

REFERENCES

[1] A. Moin, A. Zhou, A. Rahimi, A. Menon, S. Benatti, G. Alexandrov, S. Tamakloe, J. Ting, N. Yamamoto, Y. Khan *et al.*, "A wearable

biosensing system with in-sensor adaptive machine learning for hand gesture recognition," *Nat. Electron.*, vol. 4, no. 1, pp. 54–63, 2021.

[2] J. McIntosh, C. McNeill, M. Fraser, F. Kerber, M. Löchtefeld, and A. Krüger, "Empress: Practical hand gesture classification with wrist-mounted emg and pressure sensing," in *2016 Conf. Hum. Factors Comput. Syst. - Proc.*, 2016, pp. 2332–2342.

[3] E. Akan, H. Tora, and B. Uslu, "Hand gesture classification using inertial based sensors via a neural network," in *2017 24th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*. IEEE, 2017, pp. 140–143.

[4] P. Molchanov, S. Gupta, K. Kim, and K. Pulli, "Multi-sensor system for driver's hand-gesture recognition," in *Proc. Int. Conf. Autom. Face Gesture Recognit.*, vol. 1. IEEE, 2015, pp. 1–8.

[5] J. Singha, A. Roy, and R. H. Laskar, "Dynamic hand gesture recognition using vision-based approach for human–computer interaction," *Neural. Comput. Appl.*, vol. 29, no. 4, pp. 1129–1141, 2018.

[6] M. K. Ahuja and A. Singh, "Static vision based hand gesture recognition using principal component analysis," in *MITE*. IEEE, 2015, pp. 402–406.

[7] Y. Gu, X. Zhang, Z. Liu, and F. Ren, "Besense: leveraging wifi channel data and computational intelligence for behavior analysis," *IEEE Comput. Intell. Mag.*, vol. 14, no. 4, pp. 31–41, 2019.

[8] H. Yan, Y. Zhang, Y. Wang, and K. Xu, "Wiact: A passive wifi-based human activity recognition system," *IEEE Sens. J.*, vol. 20, no. 1, pp. 296–305, 2019.

[9] J. Huang, B. Liu, P. Liu, C. Chen, N. Xiao, Y. Wu, C. Zhang, and N. Yu, "Towards anti-interference wifi-based activity recognition system using interference-independent phase component," in *Proc. - IEEE INFOCOM*. IEEE, 2020, pp. 576–585.

[10] X. Cheng and B. Huang, "Csi-based human continuous activity recognition using gmm–hmm," *IEEE Sens. J.*, vol. 22, no. 19, pp. 18 709–18 717, 2022.

[11] A. Pandey, M. Zeeshan, and S. Kumar, "Csi-based joint location and activity monitoring for covid-19 quarantine environments," *IEEE Sens. J.*, 2022.

[12] Y. Wang, L. Yao, Y. Wang, and Y. Zhang, "Robust csi-based human activity recognition with augment few shot learning," *IEEE Sens. J.*, vol. 21, no. 21, pp. 24 297–24 308, 2021.

[13] J. Ding, Y. Wang, H. Si, S. Gao, and J. Xing, "Multimodal fusion-adaboost based activity recognition for smart home on wifi platform," *IEEE Sens. J.*, vol. 22, no. 5, pp. 4661–4674, 2022.

[14] J. Ding, Y. Wang, S. Fu, H. Si, J. Zhang, and S. Gao, "Multiview features fusion and adaboost based indoor localization on wifi platform," *IEEE Sens. J.*, vol. 22, no. 16, pp. 16 607–16 616, 2022.

[15] T. Zhang, T. Song, D. Chen, T. Zhang, and J. Zhuang, "Wigrus: A wifi-based gesture recognition system using software-defined radio," *IEEE Access*, vol. 7, pp. 131 102–131 113, 2019.

[16] M. A. A. Al-qaness and F. Li, "Wiger: Wifi-based gesture recognition system," *ISPRS Int. J. Geo-Inf.*, vol. 5, no. 6, p. 92, 2016.

[17] C. Li, M. Liu, and Z. Cao, "Wihf: Enable user identified gesture recognition with wifi," in *Proc. - IEEE INFOCOM*. IEEE, 2020, pp. 586–595.

[18] Y. Zheng, Y. Zhang, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Zero-effort cross-domain gesture recognition with wi-fi," in *MobiSys'19*, 2019, pp. 313–325.

[19] Y. Zeng, D. Wu, J. Xiong, and D. Zhang, "Boosting wifi sensing performance via csi ratio," *IEEE Pervasive Comput.*, 2020.

[20] Y. Zeng, D. Wu, J. Xiong, J. Liu, Z. Liu, and D. Zhang, "Multisense: Enabling multi-person respiration sensing with commodity wifi," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 3, pp. 1–29, 2020.

[21] Y. Zeng, D. Wu, R. Gao, T. Gu, and D. Zhang, "Fullbreathe: Full human respiration detection exploiting complementarity of csi phase and amplitude of wifi signals," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 3, pp. 1–19, 2018.

[22] Y. Zeng, D. Wu, J. Xiong, E. Yi, R. Gao, and D. Zhang, "Farsense: Pushing the range limit of wifi-based respiration sensing with csi ratio of two antennas," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 3, no. 3, pp. 1–26, 2019.

[23] D. Wu, R. Gao, Y. Zeng, J. Liu, L. Wang, T. Gu, and D. Zhang, "Fingerdraw: Sub-wavelength level finger motion tracking with wifi signals," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 1, pp. 1–27, 2020.

[24] X. Li, D. Zhang, Q. Lv, J. Xiong, S. Li, Y. Zhang, and H. Mei, "Indotrack: Device-free indoor human tracking with commodity wi-fi," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 3, pp. 1–22, 2017.

[25] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[26] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, 2020.

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

[28] H. Abdelnasser, M. Youssef, and K. A. Harras, "Wigest: A ubiquitous wifi-based gesture recognition system," in *Proc. - IEEE INFOCOM*. IEEE, 2015, pp. 1472–1480.

[29] S. Tan and J. Yang, "Wifinger: Leveraging commodity wifi for fine-grained finger gesture recognition," in *MobiHoc*, 2016, pp. 201–210.

[30] J. Zhang, Z. Tang, M. Li, D. Fang, P. Nurmi, and Z. Wang, "Crosssense: Towards cross-site and large-scale wifi sensing," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, 2018, pp. 305–320.

[31] H. Zou, J. Yang, Y. Zhou, L. Xie, and C. J. Spanos, "Robust wifi-enabled device-free gesture recognition via unsupervised adversarial domain adaptation," in *ICCCN*. IEEE, 2018, pp. 1–8.

[32] J. Yang, H. Zou, Y. Zhou, and L. Xie, "Learning gestures from wifi: A siamese recurrent convolutional architecture," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10 763–10 772, 2019.

[33] Y. Gu, X. Zhang, Y. Wang, M. Wang, H. Yan, Y. Ji, Z. Liu, J. Li, and M. Dong, "Wigrunt: Wifi-enabled gesture recognition using dual-attention network," *IEEE Trans. Hum.-Mach. Syst.*, 2022.

[34] N. Gozzi, L. Malandri, F. Mercorio, and A. Pedrocchi, "Xai for myo-controlled prosthesis: Explaining emg data for hand gesture classification," *Knowl. Based Syst.*, p. 108053, 2022.

[35] Y. Zhang, B. Liu, Z. Liu, J. Huang, and R. Sun, "Wristpress: Hand gesture classification with two-array wrist-mounted pressure sensors," in *BSN*. IEEE, 2019, pp. 1–4.

[36] M. Oudah, A. Al-Naji, and J. Chahl, "Hand gesture recognition based on computer vision: a review of techniques," *J. Imaging*, vol. 6, no. 8, p. 73, 2020.

[37] H. Kaur and J. Rani, "A review: Study of various techniques of hand gesture recognition," in *2016 ICPEICES*. IEEE, 2016, pp. 1–5.

[38] D. Jiang, Z. Zheng, G. Li, Y. Sun, J. Kong, G. Jiang, H. Xiong, B. Tao, S. Xu, H. Yu *et al.*, "Gesture recognition based on binocular vision," *Clust. Comput.*, vol. 22, no. 6, pp. 13 261–13 271, 2019.

[39] G. Devineau, F. Moutarde, W. Xi, and J. Yang, "Deep learning for hand gesture recognition on skeletal data," in *Proc. Int. Conf. Autom. Face Gesture Recog.* IEEE, 2018, pp. 106–113.

[40] X. Ding, T. Jiang, Y. Zhong, S. Wu, J. Yang, and J. Zeng, "Wi-fi-based location-independent human activity recognition with attention mechanism enhanced method," *Electronics*, vol. 11, no. 4, p. 642, 2022.

[41] L. Zhang, C. Wang, and D. Zhang, "Wi-pigr: Path independent gait recognition with commodity wi-fi," *IEEE Trans. Mob. Comput.*, 2021.

[42] H. Kong, L. Lu, J. Yu, Y. Chen, L. Kong, and M. Li, "Fingerpass: Finger gesture-based continuous user authentication for smart homes using commodity wifi," in *MobiHoc*. ACM, 2019, pp. 201–210.

[43] L. Jia, Y. Gu, K. Cheng, H. Yan, and F. Ren, "Beaware: Convolutional neural network (cnn) based user behavior understanding through wifi channel state information," *Neurocomputing*, vol. 397, pp. 457–463, 2020.

[44] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, 2018.

[45] R. Li, Y. Wang, F. Liang, H. Qin, J. Yan, and R. Fan, "Fully quantized network for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 2810–2819.

[46] G. Li, X. Li, Y. Wang, S. Zhang, Y. Wu, and D. Liang, "Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation," in *AAAI Conf. Artif. Intell.*, vol. 36, no. 2, 2022, pp. 1306–1313.

[47] H. Yan, Y. Gu, X. Zhang, Y. Wang, Y. Ji, and F. Ren, "Mitigating label-noise for facial expression recognition in the wild," in *IEEE Int. Conf. Multimed. Expo*. IEEE, 2022, pp. 1–6.

[48] F. A. Bhatti, M. J. Khan, A. Selim, and F. Paisana, "Shared spectrum monitoring using deep learning," *IEEE Trans. Cogn. Commun.*, vol. 7, no. 4, pp. 1171–1185, 2021.

[49] Y. Gu and J. Li, "A novel wifi gesture recognition method based on cnn-lstm and channel attention," in *AISS*, 2021, pp. 1–4.

[50] H. Kang, Z. Li, and Q. Zhang, "Communicational and computational efficient federated domain adaptation," *IEEE Trans. Parallel Distrib. Syst.*, 2022.

[51] N. Ketkar and N. Ketkar, "Stochastic gradient descent," *Deep learning with Python: A hands-on introduction*, pp. 113–132, 2017.
[52] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2015, pp. 1180–1189.
[53] M. Wattenberg, F. Viégas, and I. Johnson, "How to use t-sne effectively," *Distill*, vol. 1, no. 10, p. e2, 2016.

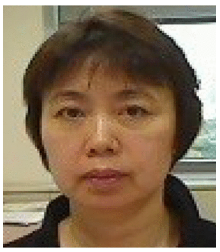
Yu Gu (M'10-SM'12) Yu Gu received the B.E. degree from the Special Classes for the Gifted Young, University of Science and Technology of China, Hefei, China, in 2004, and the D.E. degree from the same university in 2010. In 2006, he was an Intern with Microsoft Research Asia, Beijing, China, for seven months. From 2007 to 2008, he was a Visiting Scholar with the University of Tsukuba, Tsukuba, Japan. From 2010 to 2012, he was a JSPS Research Fellow with the National Institute of Informatics, Tokyo, Japan. Since 2012, he has been a Professor and Dean Assistant with the School of Computer and Information, Hefei University of Technology. He is a Professor of University of Electronic Science and Technology of China from 2023. His current research interests include pervasive computing and affective computing. He was the recipient of the IEEE Scalcom2009 Excellent Paper Award and NLP-KE2017 Best Paper Award. He is a member of ACM and a senior member of IEEE.

Huan Yan received the B.E. degree from Hefei University of Technology, China, in 2017. He is currently pursuing the Ph.D. degree at the School of Computer and Information, Hefei University of Technology, China. His research interests include intelligent information processing and wireless sensing and affective computing.

Xiang Zhang received the B.E. degree from Hefei University of Technology, China, in 2017. He is currently pursuing the Ph.D. degree at the School of Computer and Information, Hefei University of Technology, China. His research interests include intelligent information processing and wireless sensing and affective computing. He has served as a reviewer for Pervasive and Mobile Computing and IEEE Open Journal of the Computer Society.

Yantong Wang received the B.E degree from Shanghai Normal University in 2016. She received master degree from Hefei University of Technology, where she is currently working toward the Ph.D. degree. Her research interest includes affective computing and sensorless sensing.

Jingyang Huang received the B.Eng. degree from Anhui University, China in 2017, and received the Ph.D. degree at School of Cyberspace Security from University of Science and Technology of China in 2022. Now, he is a Lecturer at the School of Computer Science and Information Engineering, Hefei University of Technology (HFUT), and a member of the HFUT-S2AC Group. His research interests lie Human-computer interaction, Wireless sensing, Wireless communication, and Machine learning.



Yusheng ji (IEEE Fellow) received the B.E., M.E., and D.E. degrees in electrical engineering from the University of Tokyo. She joined the National Center for Science Information Systems, Japan, in 1990. She is currently a Professor with the National Institute of Informatics, Japan, and SOKENDAI (the Graduate University for Advanced Studies). Her research interests include network architecture, resource management in wireless networks, and mobile computing. She was a Board Member of Trustees of Institute of Electronics, Information and Communication Engineers (IEICE), Steering Committee Member of Quality Aware Internet SIG, and an Expert Member of IEICE Technical Committees on Internet Architecture, and Communication Quality, a Symposium Co-Chair of IEEE GLOBECOM in 2012 and 2014, a Track Chair of IEEE VTC 2016 Fall and 2017 Fall, an Associate Editor for the IEICE Transactions and IPSJ Journal. She is a Steering Committee Member of Internet and Operation Technologies SIG of IPSJ, an Editor of IEEE TRANSACTIONS OF VEHICULAR TECHNOLOGY, and a TPC Member of major conferences including IEEE INFOCOM, International Conference on Communications (ICC), GLOBECOM, and Wireless Communications and Networking Conference (WCNC), etc.

Fuji Ren received his Ph. D. degree in 1991 from the Faculty of Engineering, Hokkaido University, Japan. From 1991 to1994, he worked at CSK as a chief researcher. In 1994, he joined the Faculty of Information Sciences, Hiroshima City University, as an Associate Professor. Since 2001, he has been a Professor of the Faculty of Engineering, Tokushima University. He is a Chair Professor of University of Electronic Science and Technology of China from 2022. His current research interests include Natural Language Processing, Artificial Intelligence, Affective Computing, Emotional Robot. He is the Academician of The Engineering Academy of Japan and EU Academy of Sciences. He is a senior member of IEEE, Editor-in-Chief of International Journal of Advanced Intelligence, a vice president of CAAI, and a Fellow of The Japan Federation of Engineering Societies, a Fellow of IEICE, a Fellow of CAAI. He is the President of International Advanced Information Institute, Japan.