WiGRUNT: WiFi-enabled Gesture Recognition Using Dual-attention Network

Yu Gu, *Senior Member, IEEE*, Xiang Zhang, Yantong Wang, Meng Wang, Zhi Liu, *Senior Member, IEEE*, Jianhua Li, and Mianxiong Dong[✉], *Senior Member, IEEE*

Abstract—Gestures constitute an important form of nonverbal communication where bodily actions are used for delivering messages alone or in parallel with spoken words. Recently, there exists an emerging trend of WiFi sensing enabled gesture recognition due to its inherent merits like device-free, non-line-of-sight covering, and privacy-friendly. However, current WiFi-based approaches mainly rely on domain-specific training since they don't know “*where to look*” and “*when to look*”. To this end, we propose WiGRUNT, a WiFi-enabled gesture recognition system using dual-attention network, to mimic how a keen human being intercepting a gesture regardless of the environment variations. The key insight is to train the network to dynamically focus on the domain-independent features of a gesture on the WiFi Channel State Information (CSI) via a spatial-temporal dual-attention mechanism. WiGRUNT roots in a Deep Residual Network (ResNet) backbone to evaluate the importance of spatial-temporal clues and exploit their inbuilt sequential correlations for fine-grained gesture recognition. We evaluate WiGRUNT on the open Widar3 dataset and show that it significantly outperforms its state-of-the-art rivals by achieving the best-ever performance in-domain or cross-domain.

Index Terms—Gesture Recognition, Cross-Domain, CSI, machine learning, Neural Network, WiFi, CSI, Attention.

1 INTRODUCTION

A Gesture is a movement usually of the body or limbs that conveys an idea, sentiment or attitude. It dates back to our hominid ancestors who are believed to be “*better pre-adapted to acquire language-like competence using manual gestures than using vocal sounds*” [1]. Nowadays, while the mankind steps into the information era, the situation remains quite the same under the context of human-computer interaction (HCI) since we still heavily rely on the gesture to deliver messages, commands or even emotions to the computers surrounding us [2]. Thus, there is a compelling demand for an effective gesture recognition system, which can accurately recognize gestures to ensure timely communication and reaction of computers to their users [3].

Gesture recognition has drawn attention from both academic and industrial community in recent years. Various signals have been exploited in prior works, such as vision [4], wearable sensor [5], RFID [6], and doppler radar [7]. As the primary sense of human beings, vision is also the most important signal of computers in gesture recognition [8]. In general, a gesture is captured as a dynamic movement on video, which is then decomposed into a set of features taking individual frames into account for recognition [9]. Vision sensing is convenient since it allows non-invasive and distance gesture recognition, but it usually requires sufficient lightning or its recognition accuracy may be severely damaged. Also, the line-of-sight constraint may demand

deploying multiple cameras working in a cooperative way to eliminate the blind zone, even for covering some small area. Another major option is the wearable sensors such as accelerometer (ACC) or electromyography (EMG) sensors, which have no lightening requirements while delivering an excellent performance on gesture recognition [5]. But wearable sensors, as their name suggests, need users to be in close proximity to ensure valid and reliable sensory data. The same issue also exists for the RFID-based techniques, which attach cheap and passive RFID tags to the cloth or body to gather movement readings [6]. The doppler radar based approaches are non-contact. By exploiting the Doppler Effect to detect movements, they can recover the corresponding gesture via machine learning [7]. However, the radar devices currently are the most costly both in deployment and use compared to other approaches. In summary, current gesture recognition methods using the above signals usually need specialized devices with non-negligible deployment overheads, hindering their practicality and flexibility in daily life. Also, the privacy concerns are unlikely to be overlooked.

To solve these issues, WiFi sensing emerges recently as an efficient alternative for gesture recognition [10]. It can fully leverage the ubiquitous WiFi infrastructure to provide a low-cost, device-free and privacy-friendly solution. The theoretic underpinning is that human beings to the WiFi signal (either 2.4 GHz or 5 GHz) are just like mirrors to the light, i.e., a movement will bounce off the WiFi signal and cause multi-path distortions in channel response. By modeling or mapping such distortions to the corresponding movements, a gesture can be recovered.

WiGest [11] is among the first of WiFi sensing enabled gesture recognition, with a focus on hand gestures. By modeling each gesture to a manually-defined pattern in Received Signal Strength (RSS), it utilizes a matching

- Yu Gu, Xiang Zhang, Yantong Wang, Meng Wang and Jianhua Li are with the School of Computer and Information, Hefei University of Technology, China. E-mail: yugu.bruce@ieee.org, zhangxiang/yantongwang/mengw512@mail.hfut.edu.cn, jhli@hfut.edu.cn
- Zhi Liu is with Department of Computer and Network Engineering, The University of Electro-Communications, Tokyo, Japan. Email: liuzhi@ieee.org
- Corresponding author: Mianxiong Dong is with the Dept. of Sciences and Informatics, Muroran Institute of Technology, Japan. E-mail: mx.dong@csse.muroran-it.ac.jp

method to achieve an average 87.5% accuracy using only one access point (AP). WiGest is quite inspiring. However, it has two major drawbacks, i.e., the coarse-grained RSS indicator and the poor scalability of gestures. WiFinger [12] tackles such demerits via employing the fine-grained Channel State Information (CSI) as the indicator and relying on machine learning to accommodate more gestures, respectively. WiFinger lays the foundation on learning-based WiFi gesture recognition for the following research [12], [13], [14]. However, the learning-based approaches face a critical challenge, i.e., dependence on domain-specific training like locations, orientations and environments.

Recently, WiDar3 [15] aims at this issue and proposes a new feature named Body-coordinate Velocity Profile (BVP) that describes power distribution over different velocities to achieve cross-domain gesture recognition. Similarly, WiHF [16] derives a domain-independent motion change pattern of arm gestures, rendering the unique characteristics and user performing style. Their work are very enlightening, but the handcrafted cross-domain features in [15], [16] are unlikely to fully recover the domain-independent clues of a gesture spread over the spatial-temporal dimension of CSI. In particular, a gesture recorded in CSI usually involves multiple pairs of transmitting and receiving antennas that are placed diffusely to increase the spatial granularity. For each pair of antennas, the signal distortion caused by a gesture is likely distributed over multiple subcarriers (representing different central frequencies). Therefore, there exists a surging demand for an agile learning framework that can automatically and adaptively focus and extract such critical gestural cues scattered in spatial-temporal CSI profiles.

To this end, we propose WiGRUNT, a device-free and privacy-friendly gesture recognition system that dynamically concentrates on the domain-independent features via a dual-attention mechanism. The key idea of WiGRUNT, in a nutshell, is to mimic how a keen person intercepting a gesture regardless of the environment changes by exploring its sequential correlations in the time and space dimension. Firstly, for a human being, vision constitutes the most reliable signal for gesture recognition. Similarly, instead of processing the WiFi signal directly, we propose a novel CSI visualization method that leverages the CSI-ratio method [17] for effective denoising and then fuses all CSI subcarriers from spatially distributed pairs of antennas into normalized time-series images for leveraging cutting-edge techniques developed for vision sensing such as neural network backbones and pre-training. Secondly, a well-trained human being would subconsciously isolate and perceive a gesture from any environments via its the inherent characteristics. Likewise, we also design a dual-attention CSI network (DACN) that can automatically focus on sequential correlations of a gesture despite the domain variations. In particular, we embed a Deep Residual Network (ResNet) backbone, well-recognized for handling the correlation decaying issue, with a fine-tuned spatial-temporal attention module as a human being determining “*where to look*” and “*when to look*” for a gesture. DACN assigns a weight to each pixel denoting how much attention it deserves to form an attention map for each image indicating the distribution of domain-irrelevant clues of a gesture in spatial-temporal dimension. It then combines the images and the correspond-

ing attention maps for gesture recognition.

We evaluate WiGRUNT on the open Widar3 dataset and show that WiGRUNT significantly outperforms its state-of-the-art rivals by achieving the best-ever performance, i.e., 99.67% in-domain recognition accuracy and 96%, 92.6%, and 93.15% recognition accuracy across locations, orientations, and environments, respectively. Moreover, we share all the source codes on <https://github.com/purpleleaves007/WiGRUNT> to facilitate further validation and optimization.

The main contributions of WiGRUNT are summarized as follows:

- To the best of our knowledge, we are among the first to explore the attention mechanism for effective gesture recognition in WiFi sensing, which enables the best-ever in-domain or cross-domain recognition performance compared to the state-of-the-art rivals.
- We propose a simple but effective CSI visualization method to fuse multiple CSI streams from spatially distributed devices into time-series images to provide a fine-grained representation of a gesture.
- We design the dual-attention CSI network (DACN) based on a ResNet backbone to dynamically focus on the domain-independent informative clues of a gesture spread over the spatial-temporal dimension.
- We realize and evaluate WiGRUNT on the open Widar3 dataset and verify that WiGRUNT achieves 99.67% in-domain recognition accuracy and 96%, 92.6%, and 93.15% recognition accuracy across locations, orientations, and environments, respectively.

The rest of this paper is organized as follows: Section 2 reviews some representative prior works for gesture recognition using different signals, especially WiFi. Section 3 introduces the dual-attention CSI network rooted in a ResNet backbone, followed by experimental evaluation on the Widar3 dataset in Section 4. Finally, Section 5 summarizes the paper.

2 RELATED WORK

Half a million years ago, our hominin ancestors began to harvest large animals like mammoths through precisely coordinated hunting using vocal and gestural signals [22]. Since then, gestures have played an important role in our daily life for information and emotion exchange. While human beings enter the information era, a new communicate entity rises, i.e., computers. Primary input devices like mice and keyboards are cumbersome and pose location-constraints for users. As wireless communication blossoms, pervasive computing [23] aims to relieve such limits and allows seamless access to the surrounding computers anytime anywhere, raising a new challenge: how to convey messages, commands or even emotions to the ambient computers? Again, the gestural language naturally becomes an attempting option. However, current gesture recognition solutions under the context of HCI are mainly based on signals like vision [4], [24], [25], wearable sensor [5], [26], [27], RFID [6], [28], [29], and doppler radar [7], [30], [31], where have incurred non-negligible deployment and maintenance overheads, limiting their practicality and scalability in real-world. Recently, WiFi sensing emerges as

TABLE 1: WiFi-enabled Gesture Recognition

Reference	Gesture	Preprocessing	Algorithm	Dataset	User	No. of Gestures	Sample	Performance
WiGest, 2015 [11]	Hand Gesture	Wavelet Denoising	Pattern Encoding; Gesture Matching	Private	3	7	> 1000	87.5% using one AP; 96% using three APs.
WiDraw, 2015 [18]	Hand Trajectory	-	Angle-Of-Arrival (AOA)	Private	-	-	-	Hand tracking error: < 5cm; Recognition accuracy: 91%.
QGesture, 2018 [19]	Hand Trajectory	Phase and Magnitude Correction	Local Extreme Value Detection	Private	5	4	1000	Hand tracking error: < 5.5cm.
Wikey, 2015 [20]	Keystroke	Low Pass Filtering; PCA	Statistical Features; KNN	Private	10	37	12950	Classifying single keys: 96.4%; Recognize keystrokes in sentence: 93.5%.
Tan-WiFinger, 2016 [21]	Finger Gesture	Wavelet based Denoising	STFT; Multi-Dimensional DTW	Private	-	8	-	Over 90% recognition accuracy.
Li-WiFinger, 2016 [12]	Finger Gesture	Low Pass Filtering	Handcraft Feature; KNN	Private	10	9	3465	90.4% average classification accuracy.
WiMu, 2018 [13]	Multi-Person Body Gesture	Conjugate Multiplication Based Denoising	STFT; Jaccard similarity coefficient	Private	10	6	1050	90.9% for 6 simultaneously gestures.
WiSign, 2020 [14]	Sign Language	Bandpass Filtering; PCA	Deep Belief Network; Hidden Markov Model	Private	30	100	300000	Average 72.5% recognition accuracy.
WiDar3, 2019 [15]	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 [16]		Conjugate Multiplication Denoising	Motion Change Pattern Extraction; Dual-Task Neural Network					93.92% in-domain; 90.32%, 79.14% and 89.67% cross domain.
Ours		CSI-Ratio Based Denoising	CSI Visualization; Attention Based Neural Network					99.67% in-domain; 96%, 92.6% and 93.15% cross domain.

a low-cost, device-free and privacy-friendly alternative and draws much attention from the academic society.

Table 1 compares and concludes some representative prior works related to WiFi-sensing enabled gesture sensing and recognition. It can be roughly divided into two categories: modeling-based and learning-based. The former normally relies on manual characterization between signal distortions and gestures, while the latter generally leverages machine learning for gesture recognition.

Modeling-based: WiGest [18] sets up a handcrafted pattern for each gesture in RSS and designs a similarity matching method for recognition. It is an inspiring work but the coarse-grained RSS indicator severely limits its accuracy. WiMU [13] pushes the research further by achieving multi-user gesture recognition using fine-grained CSI. It depends on the exhaustive search between the combinations of known gestures to the collected samples for gesture recognition. The idea is quite interesting, but the way of manually defining patterns naturally leads to the scalability issue. WiDraw [18] uses Angle-Of-Arrival (AOA) measurements for hand tracking, and allows users to draw in the air using bare hands with an average tracking error less than 5 cm. But the practicality issue remains unjustified since it needs over 25 WiFi transceivers surrounding the user. QGesture [19] employs phase information and achieves a similar performance using only two receiving antennas. But it needs to know the initial hand position before tracking.

Learning-based: It can be further divided into shallow learning [12], [20] and deep learning [14], [15], [16]. The former uses handcrafted features to train a shallow learner to classify gestures. Wikey [20] is among the first to realize keystroke recognition based on WiFi sensing. But it is very sensitive to environmental changes. WiFinger [12] uses WiFi CSI to realize the recognition of 9 sign languages, but the user is constrained in the middle of the LoS path between the transmitting and receiving antennas. In general, shallow learning only needs a small training dataset, but its performance is limited. Consequently, deep learning emerges as an effective alternative. For example, WiSign [14] aims at American sign Language recognition, and it uses amplitude and phase CSI profiles processed by a Deep Belief Network (DBN) for recognition. However, the deep learning-based approaches face a critical challenge, i.e., dependence on heavy domain-specific training.

WiDar3 [15] targets on this issue and proposes a domain-independent feature BVP that describes power distribution over different velocities to achieve cross-domain gestures recognition. WiDar3 is among the first to reveal and address such cross-domain issue in gesture recognition and its well-honed dataset lays foundation for a fair comparison among different recognition frameworks. Based on WiDar3 dataset, WiHF [16] derives a domain-independent motion change pattern of arm gestures for obtaining the unique features for cross-domain recognition. Their work are quite enlight-

Fig. 1: System Overview.

ening, but the handcrafted cross-domain features can hardly cover the domain-independent clues of a gesture in spatial-temporal dimension. It drives us to design a new learning framework that can automatically extract and explore such critical gesture cues scattered in CSI profiles.

3 SYSTEM DESIGN

In this part, we first model the CSI-based gesture perception and then introduce the WiGRUNT in detail. Figure 1 provides an overview of the architecture of WiGRUNT, which consists of two modules, i.e., the CSI preprocessing and gesture recognition module. The first module denoises and visualizes raw CSI data as time-series images, while the second module leverages the DACN for gesture recognition.

CSI Preprocessing Module: Upon receiving CSI measurements from spatially distributed WiFi transceiver pairs, our system first denoises the raw data using the CSI-ratio method [17], [32]. Then, it extracts the phase tensor from all subcarriers to form a two-dimensional matrix and visualizes it into time-series images to provide a fine-grained description of gestures in spatial-temporal dimension.

Gesture Recognition Module: The time-series images generated in the first module will be processed in our DACN for gesture recognition. DACN roots in a ResNet backbone mounted with a dual-attention module to exploit the sequential correlations of a gesture for cross-domain recognition.

3.1 Modeling the CSI-based Gesture Perception

WiFi CSI describes the signal's attenuation on its propagation paths, such as scattering, multi-path fading or shadowing, and power decay over distance. In the frequency domain, it can be characterized as [33]:

$$\vec{Y} = \vec{H} \cdot \vec{X} + \vec{N}, \quad (1)$$

where $\vec{Y}$ and $\vec{X}$ are the received and transmitted signal vectors, respectively. $\vec{N}$ is the additive white Gaussian noise,

and $\vec{H}$ is the channel matrix representing CSI. CSI is a superposition of signals of all the propagation paths, and its Channel Frequency Response (CFR) can be represented as:

$$H(f, t) = \sum_{m \in \Phi} a_m(f, t) e^{-j2\pi \frac{d_m(t)}{\lambda}}, \quad (2)$$

where f and t represent center frequency and time stamp, respectively. m is the multi-path component. $a_m(f, t)$ and $d_m(t)$ denote the complex attenuation and propagation length of the m th multi-path component, respectively. Φ denotes the set of multi-path components, and λ is the signal wavelength.

In the case of CSI based gesture recognition, the multi-path component m consists of dynamic and static paths [34]:

$$\begin{aligned} H(f, t) &= H_s(f, t) + H_d(f, t) \\ &= \sum_{m_s \in \Phi_s} a_{m_s}(f, t) e^{-j2\pi \frac{d_{m_s}(t)}{\lambda}} \\ &\quad + \sum_{m_d \in \Phi_d} a_{m_d}(f, t) e^{-j2\pi \frac{d_{m_d}(t)}{\lambda}} \end{aligned} \quad (3)$$

where $H_s(f, t)$ and $H_d(f, t)$ denote the static and dynamic components, respectively. Φ_s represents the set of static paths, e.g., reflected off the walls, furniture and static body parts, while Φ_d denotes the set of dynamic paths, e.g., moving body parts. When to detect gestures, the movements of hands and arms will change the dynamic propagation distance $d_{m_d}(t)$, and thus alter the phase shift $e^{-j2\pi \frac{d_{m_d}(t)}{\lambda}}$ of Φ_d , and finally affect $H(f, t)$. All in all, the gesture can be portrayed by the change of phase shift in CSI.

3.2 CSI Preprocessing

As demonstrated in the previous section, the CSI phase profile can characterize gestures. Unfortunately, for commodity WiFi devices, as the transmitter and receiver are not synchronized, there exists a time-varying random phase offset $e^{-j\theta_{offset}}$:

$$H(f, t) = e^{-j\theta_{offset}} (H_s(f, t) + A(f, t) e^{-j2\pi \frac{d(t)}{\lambda}}) \quad (4)$$

where $A(f, t)$, $e^{-j2\pi \frac{d(t)}{\lambda}}$ and $d(t)$ denote the complex attenuation, phase shift and path length of dynamic components, respectively. This random offset thus prevents us from directly using the CSI phase information.

Therefore, we need to eliminate $e^{-j\theta_{offset}}$. Fortunately, for commodity WiFi card, this random offset remains the same across different antennas on the same WiFi Network Interface Card (NIC) as they share the same RF oscillator. Thus it can be eliminated by the CSI-ratio model [17], [32]:

$$\begin{aligned} H_q(f, t) &= \frac{H_1(f, t)}{H_2(f, t)} \\ &= \frac{e^{-j\theta_{offset}} (H_{s,1} + A_1 e^{-j2\pi \frac{d_1(t)}{\lambda}})}{e^{-j\theta_{offset}} (H_{s,2} + A_1 e^{-j2\pi \frac{d_2(t)}{\lambda}})} \\ &= \frac{A_1 e^{-j2\pi \frac{d_1(t)}{\lambda}} + H_{s,1}}{A_2 e^{-j2\pi \frac{d_1(t) + \Delta d}{\lambda}} + H_{s,2}} \end{aligned} \quad (5)$$

where $H_1(f, t)$ and $H_2(f, t)$ are the CSI of two receiving antennas. When two antennas are close to each other, Δd

Fig. 2: (a) Numerical description of received CSI data; (b) Numerical description of our input map and attention map.

can be regarded as a constant. According to Mobius transformation [17], equation 5 represents transformations such as scaling and rotation of the phase shift $e^{-j2\pi \frac{d_1(t)}{\lambda}}$ of antenna 1 in the complex plane, and these transformations will not affect the changing trend of the phase shift.

TABLE 2: Definitions of basic attention-based recognition neural network (The Superscript * means that the item is a function or process.).

Name	Definition
α	The learning rate of the system
f^*	The function that generate attention map
A	The attention map generated by f
R	The classification result
L	The classification loss
F^*	The processing of the backbone network
La	True label of the input data
P_A	The matrix generated by attention process
W_f	The parameter of function f
CE^*	Generate the classification loss via cross-entropy operation
ImP	The input of attention-based recognition neural network

Therefore, phase P extracted from H_q can be used to describe gestures:

$$P = \text{angle}(H_q) \quad (6)$$

where $\text{angle}(\cdot)$ denotes the phase extraction function.

After removing the random phase offset, we get a 4-dimensional tensor $H \in \mathbb{C}^{N \times M \times K \times T}$ as shown in Fig. 2a. Naturally, the key challenge here is to automatically and adaptively isolate the informative cues of a gesture scattered in such spatial-temporal dimension. As the main sense of our human beings, vision also constitutes a major signal for gesture recognition and draws significant efforts in academia. Hence, instead of processing this high-dimensional data directly, WiGRUNT relies on the visualization method [34] that maps $H \in \mathbb{C}^{N \times M \times K \times T}$ into time-series images ImP (heatmaps in Fig. 2b), which are then fed into our DACN to get the corresponding attention map (also in Fig. 2b) evaluating how much attention an information clue deserves for fine-grained gesture recognition:

$$ImP = \text{Matrix_To_Image}(P) \quad (7)$$

Fig. 3: Basic attention-based recognition neural network

3.3 Gesture Recognition

3.3.1 WiFi Meets Attention: Realize Recognition Using WiFi With Attention Based Neural Network

Psychologists believe that a human-being well-trained in culture would subconsciously keep attention on a gesture during communication by exploring its inherent sequential correlation [35]. Likewise, WiGRUNT aims to mimic such phenomenon by designed a new learning framework that can automatically focus and exploit the inbuilt and domain-independent features of a gesture.

Fig 3 shows a basic structure of an attention-based neural network. In our case, the time-series images ImP includes fine-grained descriptions of a gesture scattered in time and space (including frequency) dimension. For each image of ImP , its pixels should be evaluated for the importance to the recognition of the gesture, so that WiGRUNT can keep focusing on the essential cues of a gesture while suppressing the rest information, i.e.,

$$P_A = ImP \cdot A \quad (8)$$

$$A = f(W_f, ImP) \quad (9)$$

where A denotes an attention map. W_f means the parameter of the function f and W_f is initialized randomly. The dimensions of A and P are the same, and each pixel in A corresponds to the weight of the pixel in ImP . The larger the weight, the more important the pixel in ImP .

Then, P_A is processed by the backbone network F to get the classification result R :

$$R = F(P_A) \quad (10)$$

Fig. 4: Dual-Attention CSI Network Architecture

The classification result R and the true label L_a provided by the training set are processed by cross-entropy to obtain the classification loss L :

$$L = CE(R, L_a) \quad (11)$$

After obtaining the loss, the neural network calculates the partial derivative of the loss L with respect to the parameter W_f of attention function f based on the back-propagation algorithm:

$$\frac{\partial L}{\partial W_f} = \frac{\partial L}{\partial F} \frac{\partial F}{\partial W_f} \quad (12)$$

Consequently, the network updates the parameter W_f based on gradient descent:

$$W_f = W_f - \alpha \frac{\partial L}{\partial W_f} \quad (13)$$

where α represents the learning rate. The process of calculating the loss L and adjusting the parameter W_f continues to iterate until the model converges. In this way, the attention module adaptively learns how to generate an accurate attention map A , which assigns larger weights to the critical pixels in ImP , allowing the network to pay more attention to essential cues when recognizing gestures.

Expanded from the above basic network, in this paper, we propose Dual-Attention CSI Network (DACN) for CSI-based cross-domain gesture recognition task, which utilizes two Temporal-Spatial Attention modules to generate attention maps for the input CSI phase map and the feature map extracted from the backbone network ResNet18, respectively. And we realize zero-effort cross-domain gesture recognition with the help of attention map, they are described in detail later.

3.3.2 Dual-Attention CSI Network based Gesture Recognition

Given an phase map $ImP \in R^{C \times H \times W}$ as input, where C, H, W are the number of channels, height and width of the map, respectively. In WiGRUNT they are set to 3 (R,G,B channel), 224 and 224, respectively. Our DACN sequentially infers a 2D temporal-spatial attention map $A_{tsa} \in R^{1 \times H \times W}$ and a 1D temporal-spatial attention map $A_{tsb} \in R^{C1 \times 1 \times 1}$

TABLE 3: Definitions of symbols in defining the DACN (The Superscript * means that the item is a function or process.).

Name	Definition
A_{tsa}	The attention map that focuses on pixels in ImP
A_{tsb}	The attention map that marks the importance of different channel in feature map
ImP	The heat map of the phase matrix P
ImP'	The output of the first attention processing, and the input of <i>ResNet</i> 18
ImP''	The output of <i>ResNet</i> 18, and the input of the second attention processing
ImP'''	The output of the second attention processing, and the input of recognition processing
σ^*	The sigmoid function
$f^{*(7 \times 7)}$	The convolution operation with the filter size of 7×7
BN^*	The batch normalization processing
MLP^*	Multi-layer perceptron
$AvgPool^*$	Average pooling
$MaxPool^*$	Max pooling
$ResNet18^*$	The operate of <i>ResNet</i> 18 neural network
$Softmax^*$	Classification process

as illustrated in Figure. 4, where $C1$ is 512 if using ResNet18 [36] as backbone network to extract features. The overall process of our DACN can be summarized as:

$$\begin{aligned}
 ImP' &= A_{tsa}(ImP) \otimes ImP \oplus ImP \\
 ImP'' &= ResNet18(ImP') \\
 ImP''' &= A_{tsb}(ImP'') \otimes ImP'' \\
 RecognitionResults &= Softmax(Avgpool(ImP'''))
 \end{aligned} \quad (14)$$

where $\otimes$ and $\oplus$ denotes element-wise multiplication and summation, respectively. Figure 4 depicts the computation process of each attention map. The following describes the details of each attention module.

Temporal-Spatial Attention Module A

We generate temporal-spatial attention map A_{tsa} by utilizing the inter-spatial relationship of inputs [37]. The A_{tsa} directly focuses on 'where' is an informative part, to mark the importance of different pixels in the ImP . DACN uses a standard convolution layer that convolves the input ImP to produce our 2D temporal-spatial attention map. Through the convolution operation, we can calculate the local inter-spatial relationship of inputs. The temporal-spatial attention

Fig. 5: (a) Phase waveforms of different antenna pairs (from top to bottom, the first to sixth antenna pairs); (b) Raw input map; (c) Temporal-spatial attention map A; (d) Input map after temporal-spatial attention A.

A_{tsa} is computed as:

$$A_{tsa}(ImP) = \sigma(f^{7 \times 7}(BN(ImP))) \quad (15)$$

where σ denotes the sigmoid function, $f^{7 \times 7}$ represents a convolution operation with the filter size of 7×7 , and BN indicates the batch normalization operation.

Figure 5a shows the CSI phase waveform after noise reduction. In the above, we have explained that gestures could cause the CSI phase to change. And the more significant the phase waveform changes, the more information it contains. From Figure 5a, we can easily locate more critical periods in time dimension for gesture recognition, and we mark them with black boxes. In the input map Figure 5b, we also use black boxes to mark the same periods. The temporal-spatial attention map A_{tsa} is shown in Figure 5c, and the redder the color means the larger the weight. It can be seen that the critical pixels in the black box recording a gesture from different antenna pairs in different time periods in Figure 5b turn hotter in Figure 5c, which demonstrates the effectiveness of our temporal-spatial attention module A.

In addition to finding important antenna pairs and periods, our attention module can also mark essential subcarriers. As shown in Figure 5c, for the second Transmit-Receive (TR) pair, the importance of the subcarrier with a smaller index is not as important as that of the subcarrier with a larger index between 400-600ms. As shown in Figure 6, we show the phase waveforms of the second and twenty-fifth subcarriers of the second TR pair. It can be seen that in the 25th subcarrier, the waveform fluctuates significantly during 400-600 ms, while in the second subcarrier, the fluctuation is not obvious, meaning that this subcarrier is not sensitive to gestures and thus deserves less attention.

After obtaining the attention map, we multiply the attention map with the input and then add the original input map as shown in Figure 5d. We add the original input map after the multiplication because the multiplication will completely ignore some pixels, and cause subsequent training on the ResNet18 (18 layers) backbone unstable.

Temporal-Spatial Attention Module B

In this module, we produce a temporal-spatial attention map by exploiting the inter-channel relationship of features [38]. The feature map output from the backbone network ResNet18 has 512 channels, and the dimension of each channel matrix is 7×7 . As each channel of a feature map is considered as a feature detector [39], different channels pay different attention to different parts of the input. The A_{tsb} indirectly focuses on 'where' is an informative part,

Fig. 6: Phase waveforms of subcarrier 2 (upper) and 25 (below).

Fig. 7: Feature map before and after temporal-spatial attention B.

that is, to mark the importance of the different channels in the feature map.

To compute the channel attention efficiently, we first aggregate spatial information of a feature map using average-pooling and max-pooling operations and generate two different spatial context descriptors. Both descriptors are then forwarded to a shared multi-layer perceptron (MLP) to produce our 1D attention map A_{tsb} . Finally, we use element-wise summation to merge the output features and generate our attention map through the sigmoid operation. The 1D temporal-spatial attention is computed as:

$$A_{tsb}(ImP) = \sigma(MLP(AvgPool(ImP)) + MLP(MaxPool(ImP))) \quad (16)$$

where σ denote the sigmoid function, $AvgPool$ and $MaxPool$ denotes average pooling and max pooling, respectively.

Figure 7 shows the channel 1, 3 and 212 in feature map before and after weighted by the temporal-spatial attention map B. The information concerned by channel 3 has a higher

TABLE 4: Description of our evaluation dataset

Environments	No. of Users	Gestures	No. of Locations	No. of Orientations	No. of Samples
1st (Classroom)	9	1: Push Pull; 2: Sweep; 3: Clap; 4:Slide; 5: Draw-O(Horizontal); 6: Draw-Zigzag(Horizontal); 7: Draw-N(Horizontal); 8: Draw-Triangle(Horizontal); 9: Draw-Rectangle(Horizontal);	5	5	10125
2nd (Hall)	3	1: Push Pull; 2: Sweep; 3: Clap; 4:Slide; 5: Draw-O(Horizontal); 6: Draw-Zigzag(Horizontal);	5	5	2250
3rd (Office)	4	1: Push Pull; 2: Sweep; 3: Clap; 4:Slide; 5: Draw-O(Horizontal); 6: Draw-Zigzag(Horizontal);	5	5	3000

TABLE 5: Accuracy for WiGRUNT on Widar3.0 dataset (The target label denotes the data for testing while others for training, CL, CO and CE means Cross Location, Orientation and Environment respectively).

Target Label	1	2	3	4	5
Gesture	Indomain	99.67%			
CL	96.22%	93%	95.56%	96.33%	98.89%
CO	89.33%	93.78%	92.22%	96.22%	91.44%
CE	87.56%	97.6%	94.3%	-	-

degree of coincidence with the black boxes in Figure 5, the degree of coincidence of channel 1 is low, and channel 212 focuses on areas that are completely unimportant. Therefore, our attention module assigns the largest weight to channel 3, a smaller weight to channel 1 and near zero weight to channel 212. The weighted feature map will go through an average pooling layer and a softmax layer to obtain the final recognition result.

4 IMPLEMENTATION AND EVALUATION

Dataset: The public dataset WiDar3 [15] is constructed for fair comparisons among different learning-frameworks. Therefore, as in the state-of-the-art prior works like WiHF [16], we also rely on Widar3 to evaluate WiGRUNT. WiDar3 contains 15375 samples collected from 3 environments, and its detailed description is shown 4. In our evaluation, the division of the dataset keeps the same as [16]. In section 4.1 and 4.2, we use 4500 samples (6 users $\times$ 5 positions $\times$ 5 orientations $\times$ 6 gestures $\times$ 5 instances, position means the position of the subject, orientation means the direction the subject faces.) to evaluate the in-domain, cross-location, and orientation performance of WiGRUNT same as other researches. To verify the performance of the system with more users and more gestures, in section 4.5, we also evaluate the in-domain, cross-location, and orientation performance of WiGRUNT with all the data from 1st environment (10125 samples, 9 users $\times$ 5 positions $\times$ 5 orientations $\times$ 9 gestures $\times$ 5 instances).

For in-domain, cross-location, and orientation evaluation, 80% of the data are used as the training set, 20% as the test set, and we perform 5-fold cross-validation. For cross-location evaluation, we choose one position for testing and the remaining four for training each time. In-domain and cross-orientation evaluations are similar to cross-location evaluations. For cross-environments evaluation, we use the data from all 3 environments, which contains 12000 samples (16 users $\times$ 5 positions $\times$ 5 orientations $\times$ 6 gestures $\times$ 5 instances). We use data from two environments for training and the other one environment for testing, and perform 3-fold cross-validation.

Fig. 8: Performance comparison between state-of-the-art methods.

4.1 Overall Performance

On average, WiGRUNT achieves 99.67% recognition accuracy for in-domain gesture recognition, and obtains 96%, 92.6%, and 93.15% accuracy in cross-location, orientation, and environment, respectively.

The cross-domain performance is shown in Table 5. The first interesting observation is that WiGRUNT yields a stable performance across different locations, e.g., the recognition accuracy ranges from 93% to 98.89% with a standard deviation 0.01913. However, the recognition accuracy of different orientations is quite diversified. For instance, WiGRUNT only obtains 89.33% accuracy for orientation 1, while reaches 96% for orientation 4. The same phenomenon also occurs in WiHF and WiDar3 (70.17% in WiHF and close to 78% in WiDar3 for orientation 1). And the reason is that in the case of orientation 1, gestures might be shadowed by the human body. Nevertheless, the performance degradation cross the best and worst orientations of WiGRUNT (6.89%) is much lower than WiHF and WiDar3 (19.89% in WiHF and over 10% in WiDar3, respectively), demonstrating the superiority of WiGRUNT to its state-of-art rivals.

4.2 Comparative Study

Compared to Widar3 and WiHF, WiGRUNT does not need to extract handcrafted features. Thus the preprocessing step is simplified. We show the specific processing flows of these three approaches in Table 6. WiDar3 first denoises the CSI data, then performs time-frequency analysis and motion tracking on the noise-reduced data to generate BVP feature, and finally uses a neural network to recognize gesture with BVP. WiHF is similar to WiDar3. But the extracted feature is the motion change pattern instead of BVP. Considering that the handcrafted feature may lose gestural information, WiGRUNT extracts features adaptively based on the at-

TABLE 6: CSI processing flow compared with state-of-the-art solutions

Method	Step1	Step2	Step3	Step4
Widar3	CSI denoising	Time-Frequency analysis and motion tracking	Body-Coordinate Velocity Profile generation	Gesture recognition
WiHF	CSI denoising and PCA	Time-Frequency analysis and seam carving	Motion Change Pattern generation	Gesture recognition
WiGRUNT	CSI denoising	CSI Visualization	Gesture recognition	-

TABLE 7: Gesture recognition results compared with state-of-the-art solutions.

Method	In-domain	CL	CO	CE
Widar3	92.7%	89.7%	82.6%	92.4%
WiHF with HuFuM	93.92%	90.32%	79.14%	89.67%
WiHF with HuFu	97.65%	92.07%	82.38%	unknown
WiGRUNT	99.67%	96%	92.6%	93.15%

tention mechanism, omitting cumbersome feature extraction steps while simplifying the system implementation.

As shown in Figure 8 and Table 7, WiGRUNT significantly outperforms its state-of-art rivals in both in-domain and cross-domain scenarios, even for WiHF with the HuFu dataset (tailored from the WiDar3 dataset). In the case of using the same dataset (WiHF with HuFuM), WiGRUNT is superior to the current best solutions by 5.75%, 5.68%, 10%, and 0.75% respectively in terms of in-domain, cross-location, orientation, and environment evaluation. We think the reason is that though the handcrafted features are quite ingenious they may be not able to cover all the domain-independent sequential correlations scattered in spatial-temporal dimension. For instance, the BVP feature is subtly designed but it somehow ignores which clues are important and which clues are not. WiHF has solved this issue by directly evaluating these clues, but the features provided by WiHF only focus on the period of the motion change and filter out the information in the remaining period, leading to certain information loss.

4.3 Impact of Different Attention Modules

In this part, we evaluate the performance of WiGRUNT with the combination of different attention modules to evaluate the effectiveness of the two attention modules. The results are shown in Table 8. One interesting observation is that the attention mechanism only provides negligible enhancement for the in-domain scenario, i.e. 0.2% improvement. But for the cross-domain scenarios, it proves to be quite helpful. Moreover, the performance of ResNet with temporal-spatial attention module A and module B is better than the basic network ResNet 18. The temporal-spatial attention module A directly focuses on essential clues that are conducive to gesture recognition in the temporal and spatial dimension via assign weights to different pixels in the input image. The temporal-spatial attention module B indirectly focuses on important information in the temporal and spatial dimension by assigning weights to each channel in the feature map. Furthermore, using two temporal-spatial attention modules simultaneously achieves better performance compared to single module. The experimental results show that both attention modules in WiGRUNT can provide important clues, and these clues can be complementary.

TABLE 8: Gesture recognition results with different attention modules (A and B denotes attention module A and B).

	In-domain	CO	CL
ResNet18	99.47%	90.89%	94.58%
ResNet18+A	99.54%	91.76%	94.84%
ResNet18+B	99.67%	91.91%	95.33%
ResNet18+A+B	99.67%	92.6%	96%

TABLE 9: The impact of using the ImageNet for pre-training.

Method	In-domain	CL	CO	CE
With pre-training	99.67%	96%	92.6%	93.15%
Without pre-training	97.56%	94.33%	86.6%	83.47%

4.4 Impact of Pre-training With ImageNet

In this part, we design experiments to show that pre-training the network using image datasets in the CV field is still effective in our Table 9 illustrates the performance of WiGRUNT while leveraging methods, i.e., with and without pre-training using the super-large-scale pre-training data set ImageNet, respectively. The experimental results are shown in Table 9. The pre-trained model's performance using ImageNet is higher than the model without pre-training, and it is even close to 10% higher in cross-environment gesture recognition. As shown in Table 3, in the case of cross-environment gesture recognition, the number of samples in 2nd and 3rd environments is small, and pre-training can significantly alleviate the model's inability to learn parameters due to the small training dataset.

Pre-training is commonly used in CV due to the existence of many well-constructed and well-labeled image datasets. On the one hand, pre-training acts like an initial parameter setter and plays a vital role in subsequent supervised training. On the other hand, it relieves the over-fitting problem on a small dataset. ImageNet is the world's largest labeled image dataset containing 22,000 categories and 15 million images, and it is a dominant paradigm to initialize the backbones of object detection and segmentation models [40]. It is also an major gain of our CSI visualization method.

4.5 Impact of No. of Gestures and Users

To verify the performance of WiGRUNT in the case of increased number of users/gestures, we evaluate our system's performance with more users and more gestures (the default number is 6). The evaluation results are shown in Table 10, the in-domain accuracy remains above 98% though the number of gestures or users increases to 9. And with the increase in the number of gestures and users, the performance of WiGRUNT has not been significantly affected, and among them, the increase in the number of users has almost no impact on system performance. In some cases, the

Fig. 9: Performance comparison between state-of-the-art methods.

TABLE 10: Impact of number of gestures and users.

No. of Gestures		6	7	8	9
6 users	Indomain	99.67%	99.16%	98.08%	98.12%
	CO	92.6%	91.64%	89.01%	87.99%
	CL	96%	94.99%	91.9%	92.87%
No. of Users		6	7	8	9
6 gestures	Indomain	99.67%	99.45%	99.48%	99.1%
	CO	92.6%	92.25%	93.58%	93.63%
	CL	96%	96.17%	96.72%	96.21%

increase in the number of users/gestures can increase the accuracy. This may be because newly added gestures/users are easier to recognize. We compare the effect of the increase in the number of gestures on the system performance with WiHF [16], and the results are shown in Figure 9, where we can see that WiGRUNT is less affected by the increase in the number of gestures.

5 CONCLUSION

A gesture is defined as “a movement of part of the body, especially a hand or the head, to express an idea or meaning”. While our hominid ancestors have used gestures to communicate with their fellows with or without vocal expression, we in the informatics era also rely on gestures to deliver messages to the computers around us. This paper proposes WiGRUNT, a device-free and noncontact solution leveraging the ubiquitous WiFi infrastructure. It roots in a attention-based ResNet backbone to dynamically focus on informative clues of a gesture spread over spatial-temporal dimension and explore their inherent sequential correlations for cross-domain gesture recognition. WiGRUNT has been evaluated on the open Widar3 dataset and achieves the best-ever performance in-domain or cross-domain compared to its state-of-the-art rivals.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program (2018YFB0803403), the Cultivation Project A for Excellent Young Scholars, HFUT (No.JZ2018HGPA0272.), the National Science Foundation of China (61772169) and the Key Research and Development Plan of Anhui Province (202004b11020018).

REFERENCES

- [1] M. C. Corballis, “The gestural origins of language,” *The origins of language*, pp. 11–23, 2008.
- [2] Y. Song, D. Demirdjian, and R. Davis, “Continuous body and hand gesture recognition for natural human-computer interaction,” *ACM Transactions on Interactive Intelligent Systems*, vol. 2, no. 1, pp. 1–28, 2012.
- [3] S. Mitra and T. Acharya, “Gesture recognition: A survey,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 3, pp. 311–324, 2007.
- [4] Z. Ren, J. Yuan, J. Meng, and Z. Zhang, “Robust part-based hand gesture recognition using kinect sensor,” *IEEE transactions on multimedia*, vol. 15, no. 5, pp. 1110–1120, 2013.
- [5] X. Zhang, X. Chen, Y. Li, V. Lantz, K. Wang, and J. Yang, “A framework for hand gesture recognition based on accelerometer and emg sensors,” *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 41, no. 6, pp. 1064–1076, 2011.
- [6] Y. Zou, J. Xiao, J. Han, K. Wu, Y. Li, and L. M. Ni, “Grfid: A device-free rfid-based gesture recognition system,” *IEEE Transactions on Mobile Computing*, vol. 16, no. 2, pp. 381–393, 2016.
- [7] S. Skaria, A. Al-Hourani, M. Lech, and R. J. Evans, “Hand-gesture recognition using two-antenna doppler radar with deep convolutional neural networks,” *IEEE Sensors Journal*, vol. 19, no. 8, pp. 3041–3048, 2019.
- [8] S. S. Rautaray and A. Agrawal, “Vision based hand gesture recognition for human computer interaction: a survey,” *Artificial intelligence review*, vol. 43, no. 1, pp. 1–54, 2015.
- [9] V. I. Pavlovic, R. Sharma, and T. S. Huang, “Visual interpretation of hand gestures for human-computer interaction: A review,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 19, no. 7, pp. 677–695, 1997.
- [10] Y. Ma, G. Zhou, and S. Wang, “Wifi sensing with channel state information: A survey,” *ACM Computing Surveys (CSUR)*, vol. 52, no. 3, pp. 1–36, 2019.
- [11] H. Abdelnasser, M. Youssef, and K. A. Harras, “Wigest: A ubiquitous wifi-based gesture recognition system,” in *2015 IEEE conference on computer communications (INFOCOM)*. IEEE, 2015, pp. 1472–1480.
- [12] H. Li, W. Yang, J. Wang, Y. Xu, and L. Huang, “Wifinger: talk to your smart devices with finger-grained gesture,” in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016, pp. 250–261.
- [13] R. H. Venkatnarayan, G. Page, and M. Shahzad, “Multi-user gesture recognition using wifi,” in *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, 2018, pp. 401–413.
- [14] L. Zhang, Y. Zhang, and X. Zheng, “Wisign: Ubiquitous american sign language recognition using commercial wi-fi devices,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 3, pp. 1–24, 2020.
- [15] 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 *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, 2019, pp. 313–325.
- [16] C. L. Li, M. Liu, and Z. Cao, “Wihf: Gesture and user recognition with wifi,” *IEEE Transactions on Mobile Computing*, 2020.
- [17] 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,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–26, 2019.
- [18] L. Sun, S. Sen, D. Koutsonikolas, and K.-H. Kim, “Widraw: Enabling hands-free drawing in the air on commodity wifi devices,” in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, 2015, pp. 77–89.
- [19] N. Yu, W. Wang, A. X. Liu, and L. Kong, “Qgesture: Quantifying gesture distance and direction with wifi signals,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 1–23, 2018.
- [20] K. Ali, A. X. Liu, W. Wang, and M. Shahzad, “Keystroke recognition using wifi signals,” in *Proceedings of the 21st annual international conference on mobile computing and networking*, 2015, pp. 90–102.
- [21] S. Tan and J. Yang, “Wifinger: Leveraging commodity wifi for fine-grained finger gesture recognition,” in *Proceedings of the 17th ACM international symposium on mobile ad hoc networking and computing*, 2016, pp. 201–210.

- [22] K. Sterelny, "Language, gesture, skill: the co-evolutionary foundations of language," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 367, no. 1599, pp. 2141–2151, 2012.
- [23] M. Satyanarayanan, "Pervasive computing: Vision and challenges," *IEEE Personal communications*, vol. 8, no. 4, pp. 10–17, 2001.
- [24] X. Zabulis, H. Baltzakis, and A. A. Argyros, "Vision-based hand gesture recognition for human-computer interaction." *The universal access handbook*, vol. 34, p. 30, 2009.
- [25] P. K. Pisharady and M. Saerbeck, "Recent methods and databases in vision-based hand gesture recognition: A review," *Computer Vision and Image Understanding*, vol. 141, pp. 152–165, 2015.
- [26] Z. Lu, X. Chen, Q. Li, X. Zhang, and P. Zhou, "A hand gesture recognition framework and wearable gesture-based interaction prototype for mobile devices," *IEEE transactions on human-machine systems*, vol. 44, no. 2, pp. 293–299, 2014.
- [27] J. Wu and R. Jafari, "Orientation independent activity/gesture recognition using wearable motion sensors," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1427–1437, 2018.
- [28] P. Asadzadeh, L. Kulik, and E. Tanin, "Gesture recognition using rfid technology," *Personal and Ubiquitous Computing*, vol. 16, no. 3, pp. 225–234, 2012.
- [29] R. Parada and J. Melia-Segui, "Gesture detection using passive rfid tags to enable people-centric iot applications," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 56–61, 2017.
- [30] T. Fan, C. Ma, Z. Gu, Q. Lv, J. Chen, D. Ye, J. Huangfu, Y. Sun, C. Li, and L. Ran, "Wireless hand gesture recognition based on continuous-wave doppler radar sensors," *IEEE Transactions on Microwave Theory and Techniques*, vol. 64, no. 11, pp. 4012–4020, 2016.
- [31] G. Li, R. Zhang, M. Ritchie, and H. Griffiths, "Sparsity-driven micro-doppler feature extraction for dynamic hand gesture recognition," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 54, no. 2, pp. 655–665, 2017.
- [32] 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," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 1, pp. 1–27, 2020.
- [33] Y. Gu, X. Zhang, C. Li, F. Ren, J. Li, and Z. Liu, "Your wifi knows how you behave: Leveraging wifi channel data for behavior analysis," in *2018 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2018, pp. 1–6.
- [34] Y. Gu, X. Zhang, Z. Liu, and F. Ren, "Wife: Wifi and vision based intelligent facial-gesture emotion recognition," *arXiv preprint arXiv:2004.09889*, 2020.
- [35] A. Joshi, H. Sierra, and E. Arzuaga, "American sign language translation using edge detection and cross correlation," in *2017 IEEE Colombian Conference on Communications and Computing (COLCOM)*. IEEE, 2017, pp. 1–6.
- [36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [37] X. Zhu, D. Cheng, Z. Zhang, S. Lin, and J. Dai, "An empirical study of spatial attention mechanisms in deep networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6688–6697.
- [38] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
- [39] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.
- [40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

Yu Gu (M'10-SM'12) 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. He is currently a Professor and Dean Assistant with the School of Computer and Information, Hefei University of Technology, Hefei, China. 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.

Xiang Zhang was born in Anhui, China, in 1996. He received the B.E. degree from the Hefei University of Technology, where he is currently pursuing the Ph.D. degree. His research interests include intelligent information processing, and wireless sensing and affective computing.

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.

Meng Wang received the bachelor's degree from Hefei University of Technology, where she is currently pursuing the master's degree. Her current research interests include intelligent information processing, wireless sensing and machine learning.

Zhi Liu (S11-M14-SM19) received the B.E., from the University of Science and Technology of China, China and Ph.D. degree in informatics in National Institute of Informatics. He is currently an Associate Professor at The University of Electro-Communications. His research interest includes video network transmission, vehicular networks and mobile edge computing. He is now an editorial board member of Springer wireless networks and has been a Guest Editor of Mobile Networks & Applications, Springer

Wireless networks and IEICE Transactions on Information and Systems. He is a senior member of IEEE.

Jianhua Li received the BS degree from the Department of Computer Science and Technology, Anqing Normal University, Anhui, China, in 2007 and the PhD degree from the School of Computer Science and Technology, University of Science and Technology of China, in 2013. He also obtained the PhD degree from City University of Hong Kong in 2013. He is now an associate professor in the School of Computer and Information, Hefei University of Technology, Anhui, China. His research interests include multi-core

memory system, emerging nonvolatile memories, and on-chip networks.

Mianxiong Dong received B.S., M.S. and Ph.D. in Computer Science and Engineering from The University of Aizu, Japan. He is the youngest ever Vice President and Professor of Muroran Institute of Technology, Japan. He was a JSPS Research Fellow with School of Computer Science and Engineering, The University of Aizu, Japan and was a visiting scholar with BCCR group at the University of Waterloo, Canada supported by JSPS Excellent Young Researcher Overseas Visit Program from April 2010 to August 2011.

Dr. Dong was selected as a Foreigner Research Fellow (a total of 3 recipients all over Japan) by NEC C&C Foundation in 2011. He is the recipient of IEEE TCSC Early Career Award 2016, IEEE SCSTC Outstanding Young Researcher Award 2017, The 12th IEEE ComSoc Asia-Pacific Young Researcher Award 2017, Funai Research Award 2018 and NISTEP Researcher 2018 (one of only 11 people in Japan) in recognition of significant contributions in science and technology. He is Clarivate Analytics 2019 Highly Cited Researcher (Web of Science).