WiGRUNT: WiFi-Enabled Gesture Recognition Using Dual-Attention Network

Yu Gu, Xiang Zhang, Yantong Wang, Meng Wang, *Graduate Student Member, IEEE*, Huan Yan, Yusheng Ji, *Fellow, IEEE*, Zhi Liu, *Senior Member, IEEE*, Jianhua Li, and Mianxiong Dong, *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 remote sensing, non-line-of-sight covering, and privacy-friendly. However, current WiFi-based approaches mainly reply 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 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*—Attention, channel state information (CSI), cross-domain, gesture recognition, neural network, WiFi.**

I. 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], radio-frequency identification (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 noninvasive and distant 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 noncontact. 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, remote sensing, and privacy-friendly solution. The theoretic underpinning is that human beings to the WiFi signal (either

Manuscript received November 1, 2021; revised January 28, 2022; accepted March 13, 2022. Date of publication April 18, 2022; date of current version July 14, 2022. This work was supported in part by the Key Research and Development Plan of Anhui Province under Grant 202004b11020018, and in part by the Fundamental Research Fund of Chinese Academy of Medical Sciences under Grant 2020-JKCS-002. This article was recommended by Associate Editor L. L. Chen. *(Corresponding author: Xiang Zhang.)*

Yu Gu, Xiang Zhang, Yantong Wang, Meng Wang, Huan Yan, and Jianhua Li are with the S^2 AC Laboratory, School of Computer and Information and Key Laboratory of Knowledge Engineering with Big Data, Hefei University of Technology, Anhui 230002, China (e-mail: yugu.bruce@ieee.org; zhangxiang@mail.hfut.edu.cn; yantongwang@mail.hfut.edu.cn; mengw512@mail.hfut.edu.cn; yanhuan@mail.hfut.edu.cn; jhli@hfut.edu.cn).

Yusheng Ji is with the National Institute of Informatics, Tokyo 101-8430, Japan (e-mail: kei@nii.ac.jp).

Zhi Liu is with the Department of Computer and Network Engineering, The University of Electro-Communications, Tokyo 182-8585, Japan (e-mail: liu@ieee.org).

Mianxiong Dong is with the Department of Sciences and Informatics, Muroran Institute of Technology, Hokkaido 050-8585, Japan (e-mail: mx.dong@ieee.org).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/THMS.2022.3163189.

Digital Object Identifier 10.1109/THMS.2022.3163189

2.4 or 5 GHz) are just like mirrors to the light, i.e., a movement will bounce off the WiFi signal and cause multipath 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 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]. However, the learning-based approaches face a critical challenge, i.e., dependence on domain-specific training like locations, orientations, and environments.

Recently, WiDar3 [14] 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 [15] 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 [14], [15] 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 diffusedly 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 remote sensing 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. First, 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 [16] 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 pretraining. Second, a well-trained human being would subconsciously isolate and perceive a gesture from any environments via its the inherent characteristics. When a person is recognizing a gesture, he or she focuses on the period when a gesture occurs (temporal domain), pay attention to the hand rather than the face, and observes the gesture from the front rather than the back (spatial domain). 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 corresponding 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., 99.71% in-domain recognition accuracy and 96.62%, 93.85%, and 93.73% 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.

- We 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 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 99.71% in-domain recognition accuracy and 96.62%, 93.85%, and 93.73% recognition accuracy across locations, orientations, and environments, respectively.

The rest of this article is organized as follows. Section II reviews some representative prior works for gesture recognition using different signals, especially WiFi. Section III introduces the dual-attention CSI network rooted in a ResNet backbone, followed by experimental evaluation on the Widar3 dataset in Section IV. Finally, Section V concludes this article.

II. Related Work

WiFi-sensing-enabled gesture sensing and recognition [17], [18] 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 [19] builds a handcraft pattern for each gesture in RSS and designs a similarity matching method

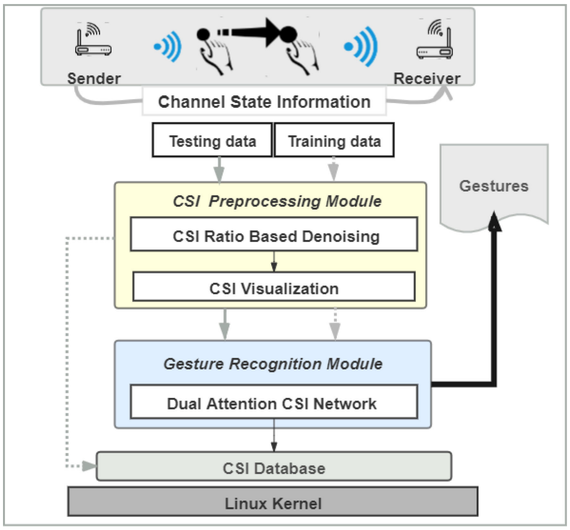

Fig. 1. System overview.

for recognition. It is an inspiring work but the coarse-grained RSS indicator severely limits its accuracy. WiMU [20] pushes the research further by achieving multiuser 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 [19] 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 [21] 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: Instead of handcraft patterns, we could rely on learning data for automatic pattern recognition, which can be further divided into shallow learning [12], [22] and deep learning [13]–[15], [23]. The former uses handcrafted features to train a shallow learner to classify gestures. Wikey [22] 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 nine 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 [13] aims at American sign language recognition, and it uses amplitude and phase CSI profiles processed by a Deep Belief Network for recognition. However, the deep learning-based approaches face a critical challenge, i.e., dependence on heavy domain-specific training.

WiDar3 [14] 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 [15] derives a domain-independent motion change pattern of arm gestures for obtaining the unique features for cross-domain recognition. Their work are quite enlightening, 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. ABLSTM [24] proposed an attention-based framework to perform human activity recognition from scratch, and they believe that handcrafted features require expert knowledge and thus inevitably lose implicit knowledge. Similar to this idea, we also think that manually picking features is far from efficient for this particular mission. Therefore, we focus on exploring attention-based neural networks that can adaptively learn discriminative features from data in an automatic way. Different from this work, our solution not only learns features in the time domain, but also in frequency domain and space domain.

TABLE I
DEFINITIONS OF BASIC ATTENTION-BASED RECOGNITION NEURAL NETWORK

Name	Definition
α	The learning rate of the system
f^*	The function that generate attention map
M	The attention map generated by f
R	The classification result
L	The classification loss
N^*	The processing of the backbone network
La	True label of the input data
P_M	The matrix generated by attention process
W_f	The parameter of function f
CE^*	Generate the classification loss via cross-entropy operation
$\mathbf{T}$	The input of attention-based recognition neural network

The superscript * means that the item is a function or process.

III. SYSTEM DESIGN

In this part, we first model the CSI-based gesture perception and then introduce the WiGRUNT in detail. Fig. 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 recogntion.

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 [16], [25]. 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. And definitions used in this paper are shown in Table I.

A. Modeling the CSI-Based Gesture Perception

WiFi CSI describes the signal's attenuation on its propagation paths, such as scattering, multipath fading or shadowing, and power decay over distance. In the frequency domain, it can be characterized as [26], [27]

$$\vec{Y} = \vec{H} \cdot \vec{X} + \vec{N} \tag{1}$$

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

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

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

In the case of CSI-based gesture recognition, the multipath component m consists of dynamic and static paths

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

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.

B. CSI Preprocessing

As demonstrated in the previous section, the gesture can be portrayed by the change of phase-shift in CSI. Unfortunately, for commodity WiFi devices, as the transmitter and receiver are not synchronized, there exists a time-varying random phase offset $e^{-j\theta_{\text{offset}}}$

$$\begin{aligned} H(f,t) &= e^{-j\theta_{\text{offset}}}(H_s(f,t) + H_d(f,t)) \\ &= e^{-j\theta_{\text{offset}}}(H_s(f,t) + A(f,t) e^{-j2\pi \frac{d(t)}{\lambda}}) \end{aligned} \tag{4}$$

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 $e^{-j\theta_{\text{offset}}}$. 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 [16], [25]

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

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, $\triangle d$ can be regarded as a constant. According to Mobius transformation [16], (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.

CSI ratio is based on two observations [16].

1) For commodity WiFi cards such as the widely used Intel 5300, the time-varying phase offset is the same across different antennas on a Wi-Fi card as they share the same RF oscillator [28], [29].
2) When the target moves a short distance (up to 10 cm), the difference of the two reflection path lengths at two close-by antennas can be considered as a constant [30].

Both conditions can be easily satisfied in the gesture recognition task. For our case, the antennas used in the Widar3 dataset are close enough, and the length of hand gestures is around a few centimeters. Therefore, we can directly apply this model to Widar3 for hand gesture recognition.

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

$$P = angle(H_q) \tag{6}$$

where $angle(\cdot)$ denotes the phase extraction function. For a complex $z = abs(z) \cdot e^{i \cdot \theta}$, we can use $angle(\cdot)$ to obtain the phase of z, $\theta = angle(z)$.

After removing the random phase offset, we get a four-dimensional tensor $H \in C^{N \times M \times K \times T}$, as shown in Fig. 2(a) [10], where M, K, and T represent the number of receive antennas, transmit antennas, subcarriers, and packets, respectively. 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 [23] that maps $H \in C^{N \times M \times K \times T}$ into time-series images $\mathbf{T}$ [heatmaps in Fig. 2(b)], which are then fed into our DACN to get the corresponding attention map [also in Fig. 2(b)] evaluating how much attention an information clue deserves for fine-grained gesture recognition

$$\mathbf{T} = Matrix\_To\_Image(P). \tag{7}$$

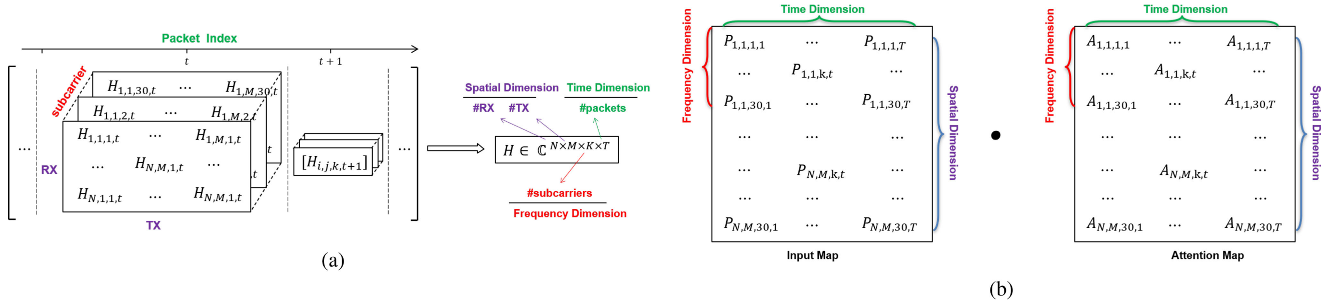
Packet Index
t
t + 1
subcarrier
RX
TX
$H_{1,1,30,t}$ ⋯ $H_{1,M,30,t}$
$H_{1,1,2,t}$ ⋯ $H_{1,M,2,t}$
$H_{1,1,1,t}$ ⋯ $H_{1,M,1,t}$
⋯ $H_{N,M,1,t}$ ⋯
$H_{N,1,1,t}$ ⋯ $H_{N,M,1,t}$
$[H_{i,j,k,t+1}]$
Spatial Dimension
#RX #TX
Time Dimension
#packets
$H \in \mathbb{C}^{N\times M\times K\times T}$
#subcarriers
Frequency Dimension
(a)
Time Dimension
Frequency Dimension
Spatial Dimension
$P_{1,1,1,1}$ ⋯ $P_{1,1,1,T}$
⋯ $P_{1,1,k,t}$ ⋯
$P_{1,1,30,1}$ ⋯ $P_{1,1,30,T}$
$P_{N,M,k,t}$
$P_{N,M,30,1}$ ⋯ $P_{N,M,30,T}$
Input Map
$A_{1,1,1,1}$ ⋯ $A_{1,1,1,T}$
⋯ $A_{1,1,k,t}$ ⋯
$A_{1,1,30,1}$ ⋯ $A_{1,1,30,T}$
$A_{N,M,k,t}$
$A_{N,M,30,1}$ ⋯ $A_{N,M,30,T}$
Attention Map
(b)

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

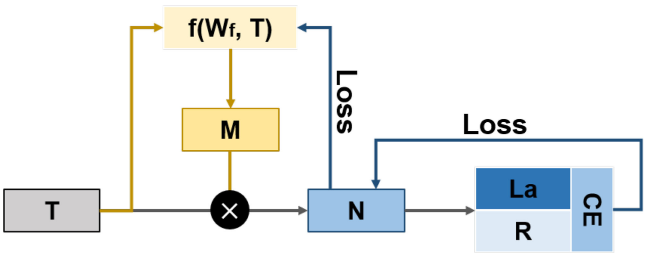

Fig. 3. Basic attention-based recognition neural network.

C. Gesture Recognition

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 [31]. Likewise, WiGRUNT aims to mimic such phenomenon by designing 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 $\mathbf{T}$ includes fine-grained descriptions of a gesture scattered in time and space (including frequency) dimension. For each image of $\mathbf{T}$, 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.,

$$P_M = \mathbf{T} \otimes M \tag{8}$$

$$M = f(W_f, \mathbf{T}) \tag{9}$$

where M denotes an attention map. W_f means the parameter of the function f and W_f is initialized randomly, $\otimes$ denotes the element-wise multiplication. The dimensions of M and P are the same, and each pixel in M corresponds to the weight of the pixel in $\mathbf{T}$. The larger the weight, the more important the pixel in $\mathbf{T}$.

Then, P_M is processed by the backbone network N to get the classification result R

$$R = N(P_M). \tag{10}$$

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

$$L = CE(R, La). \tag{11}$$

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 backpropagation algorithm

$$\frac{\partial L}{\partial W_f} = \frac{\partial L}{\partial F}\frac{\partial F}{\partial W_f}. \tag{12}$$

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

$$W_f = W_f - \alpha\frac{\partial L}{\partial W_f} \tag{13}$$

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 $\mathbf{T}$, allowing the network to pay more attention to essential cues when recognizing gestures.

Expanded from the above basic network, in this article, we propose 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.

2) Dual-Attention CSI Network-Based Gesture Recognition: Given a phase map $\mathbf{T} \in R^{C\times H\times W}$ 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 2-D temporal-spatial attention map $\mathbf{M_a} \in R^{1\times H\times W}$ and a 1D temporal-spatial attention map $\mathbf{M_b} \in R^{C_1\times 1\times 1}$ as illustrated in Fig. 4, where C_1 is 512 if using ResNet18 [32] as backbone network to extract features. As shown in Fig. 4 and Table II, the overall process of our DACN can be summarized as

$$\begin{aligned}\mathbf{F} &= \mathbf{M_a}(\mathbf{T}) \otimes \mathbf{T}\\ \mathbf{F}' &= \mathbf{ResNet}(\mathbf{F})\\ \mathbf{F}'' &= \mathbf{M_b}(\mathbf{F}') \otimes \mathbf{F}'\end{aligned} \tag{14}$$

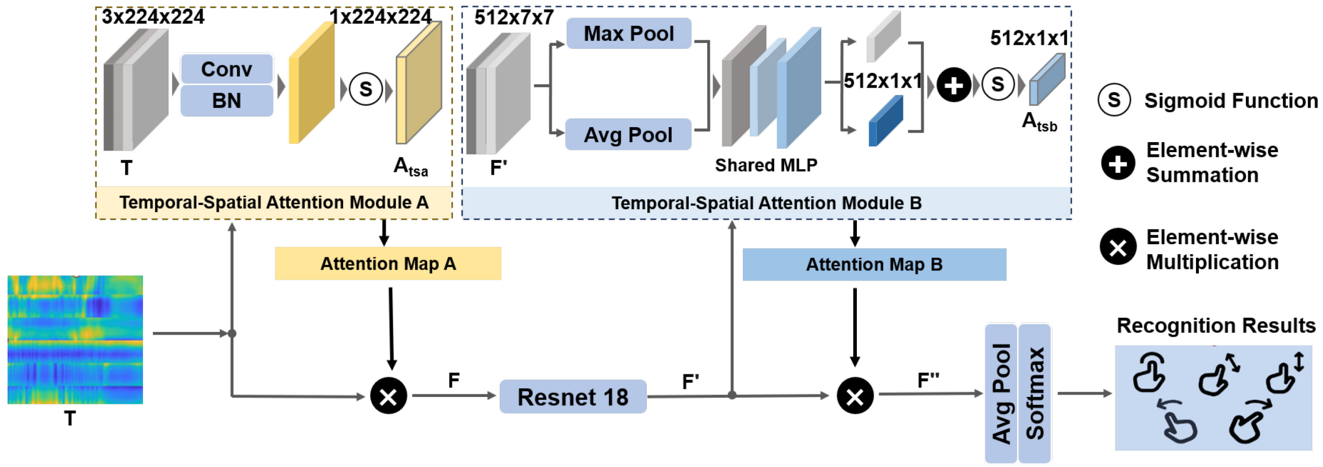

Fig. 4. Dual-attention CSI network architecture.

TABLE II
DEFINITIONS OF SYMBOLS IN DEFINING THE DACN

Name	Definition
$\mathbf{M_a}$	The attention map that focuses on pixels in $\mathbf{T}$
$\mathbf{M_b}$	The attention map that marks the importance of different channel in feature map
$\mathbf{T}$	The heat map of the phase matrix P
$\mathbf{F}$	The output of the first attention processing, and the input of $ResNet$ 18
$\mathbf{F'}$	The output of $ResNet$ 18, and the input of the second attention processing
$\mathbf{F''}$	The output of the second attention processing, and the input of recognition processing
σ^*	The sigmoid function
$f^{*(7\times7)}$	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
ResNet*	The operate of $ResNet$ 18 neural network
$Softmax^*$	Classification process

The superscript * means that the item is a function or process.

where $\otimes$ denotes the element-wise multiplication. During multiplication, the attention values are broadcasted (copied) accordingly: values of $\mathbf{M_b} \in R^{C_1\times1\times1}$ are broadcasted along the spatial dimension ($H \times W$), and vice versa. $\mathbf{M_a}$ assigns a weight to each pixel of the image T, and $\mathbf{M_b}$ assigns a weight to each channel of the feature F output by the $\mathbf{ResNet}$. F'' is the final refined feature, and DACN uses it to perform the final recognition. The following describes the details of each attention module.

The basic principle of how the attention module weights the importance of features is to first integrate the data into the dimension of concern to generate the attention map, then multiply the produced attention map with the original input, and, finally, to achieve the purpose of learning attention by continuously reducing the loss through back-propagation. Our attention modules A and B generate attention maps on the spatial domain and channel domain of the data, respectively, and their purpose is to help the network gradually focus on more important features and suppress uncertain features during the training process. Module A acts on the input image, and its purpose is to help the network learn to pay attention to which parts of the image are more important (because our preprocessing step integrates the time domain, frequency domain, and spatial domain information of WiFi into one image, the role of module A is to focus on the more valuable parts of these three domains). Module B is applied to the output of the features by the backbone network. The output features usually include multiple channels, and the information contained in different channels is also varied. Module B helps the network learn to focus on which channel is more vital.

Temporal-Spatial Attention Module A: We generate temporal-spatial attention map $\mathbf{M_a}$ by utilizing the interspatial relationship of inputs [33]. The $\mathbf{M_a}$ directly focuses on "where" is an informative part, to mark the importance of different pixels in the $\mathbf{T}$. DACN uses a standard convolution layer that convolves the input $\mathbf{T}$ to produce our 2-D temporal-spatial attention map. Its convolution kernel size is 7, which is larger than the size 3 convolution kernel used in the backbone network. The larger the convolution kernel, the larger the image area covered by the convolution operation (the convolution kernel of size 7 convolves a 7×7 size image), the larger the range of capturing spatial relationships. Through the convolution operation, we can calculate the local interspatial relationship of inputs. The temporal-spatial attention $\mathbf{M_a}$ is computed as

$$\mathbf{M_a}(\mathbf{T}) = \sigma(f^{7\times7}(BN(\mathbf{T}))) \tag{15}$$

where σ denotes the sigmoid function, $f^{7\times7}$ represents a convolution operation with the filter size of 7×7, and BN indicates the batch normalization operation.

Fig. 5(a) 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 Fig. 5(a), we can easily locate more critical periods in time dimension for gesture recognition, and we mark them with black boxes. In the input map Fig. 5(b), we also use black boxes to mark the same periods. The temporal-spatial attention map $\mathbf{M_a}$ is shown in Fig. 5(c), 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 Fig. 5(b)

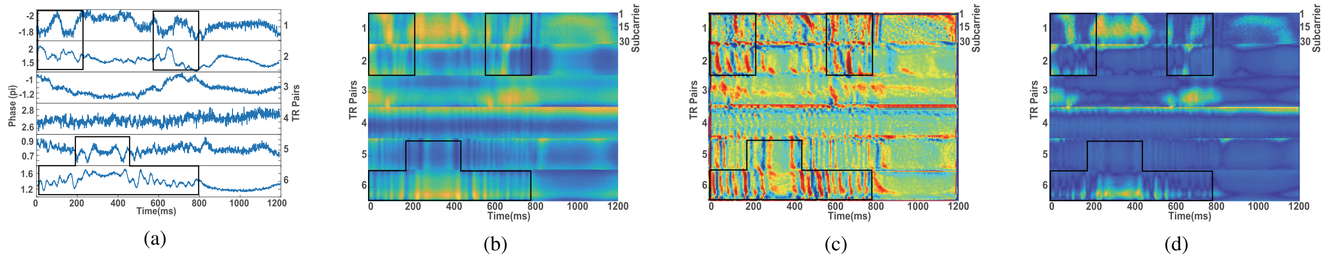

Fig. 5. (a) Phase waveforms of different antenna pairs (from top to bottom, the first to sixth antenna pairs). (b) Raw input map. (c) Temporal-spatial attention map A. (d) Input map after temporal-spatial attention A.

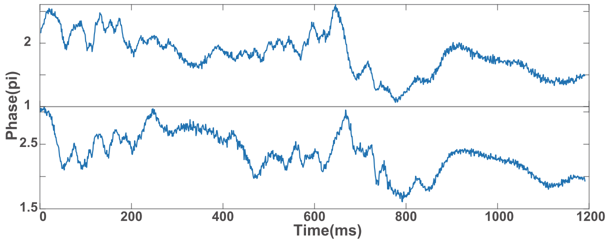

Fig. 6. Phase waveforms of subcarrier 2 (upper) and 25 (below).

turn hotter in Fig. 5(c), 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 Fig. 5(c), 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 and 600 ms. As shown in Fig. 6, we show the phase waveforms of the 2nd and 25th 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, as shown in Fig. 5(d).

Temporal-Spatial Attention Module B: In this module, we produce a temporal-spatial attention map by exploiting the interchannel relationship of features [34]. 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 [35], different channels pay different attention to different parts of the input. The $\mathbf{M_b}$ indirectly focuses on “where” is an informative part, 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 multilayer perceptron (MLP) to produce our 1-D attention map $\mathbf{M_b}$. Finally, we use element-wise summation to merge the output features and generate our attention map through the sigmoid operation. The 1-D temporal-spatial attention is computed as

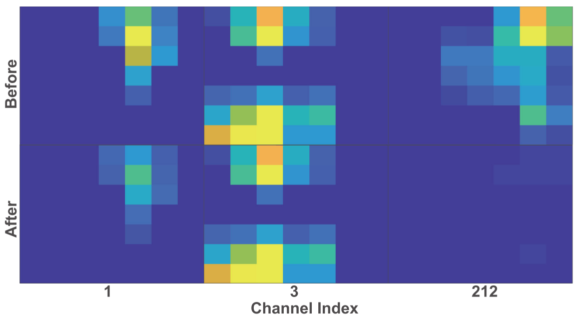

Fig. 7. Feature map before and after temporal-spatial attention B.

$$\mathbf{M_b}(\mathbf{F}') = \sigma(MLP(AvgPool(\mathbf{F}')) + MLP(MaxPool(\mathbf{F}'))) \tag{16}$$

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

Fig. 7 shows 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 degree of coincidence with the black boxes in Fig. 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.

IV. Implementation and Evaluation

Dataset: The public dataset WiDar3 [14] is constructed for fair comparisons among different learning-frameworks. Therefore, as in the state-of-the-art prior works like WiHF [15], we also rely on Widar3 to evaluate WiGRUNT. WiDar3 contains 15 375 samples collected from three environments, and its detailed description is shown in Section III. In our evaluation, the evaluation setting keeps the same as [15]. In Sections IV-A and IV-B, we use 4500 samples (6 users × 5 positions × 5 orientations × 6 gestures × 5 instances; position means the position of the subject, orientation means the direction the subject faces) to

TABLE III
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 IV
ACCURACY FOR WiGRUNT ON WIDAR3.0 DATASET

Target Label	1	2	3	4	5
Gesture	Indomain	99.71%			
CL	97.33%	93%	96.56%	97.67%	98.56%
CO	90.89%	95.67%	92.56%	96.89%	93.22%
CE	87.90%	97.82%	95.47%	-	-

The target label denotes the data for testing while others for training, CL, CO, and CE means cross-location, orientation, and environment, respectively.

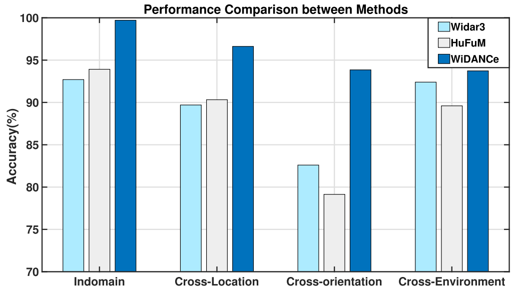

Fig. 8. Performance comparison between state-of-the-art methods.

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 IV-F, we also evaluate the in-domain, cross-location, and orientation performance of WiGRUNT with all the data from 1st environment (10 125 samples, 9 users × 5 positions × 5 orientations × 9 gestures × 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 five-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 three environments, which contains 12 000 samples (16 users × 5 positions × 5 orientations × 6 gestures × 5 instances). We use data from two environments for training and the other one environment for testing, and perform three-fold cross-validation.

A. Overall Performance

On average, WiGRUNT achieves 99.71% recognition accuracy for in-domain gesture recognition, and obtains 96.62%, 93.85%, and 93.73% accuracy in cross-location, orientation, and environment, respectively.

The cross-domain performance is shown in Table IV. The first interesting observation is that WiGRUNT yields a stable performance across different locations, e.g., the recognition accuracy ranges from 93% to 98.56% with a standard deviation 0.019. However, the recognition accuracy of different orientations is quite diversified. For instance, WiGRUNT only obtains 90.89% and 92.56% accuracy for orientation 1 and 3, respectively, while reaches 96.89% 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 orientations 1 and 3, gestures might be shadowed by the human body. Nevertheless, the performance degradation cross of the best and worst orientations of WiGRUNT (6%) 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-the-art rivals.

B. 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 V. 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 attention mechanism, omitting cumbersome feature extraction steps while simplifying the system implementation.

As shown in Fig. 8 and Table VI, WiGRUNT significantly outperforms its state-of-the-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 5.79%, 6.3%, 11.25%, and 1.33%, 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 not be 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

TABLE V
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 VI
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%	unknow
WiGRUNT	**99.71%**	**96.62%**	**93.85%**	**93.73%**

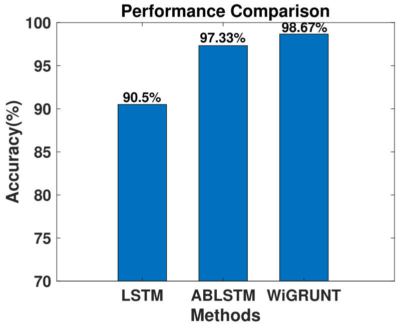

Fig. 9. Performance comparison with other attention-based work.

TABLE VII
IMPACT OF DIFFERENT AMOUNTS OF TRAINING DATA. 1/4 MEANS USE 1/4 OF ALL TRAINING DATA

Method	1/4	1/2	3/4	1
In-domain	93.78%	98.33%	99.00%	99.71%
CL	76.11%	89.33%	94.56%	96.62%
CO	77.32%	89.76%	92.22%	93.85%

TABLE VIII
GESTURE RECOGNITION RESULTS WITH DIFFERENT ATTENTION MODULES

	In-domain	CO	CL
ResNet18	99.47%	90.89%	94.58%
ResNet18+A	99.62%	92.71%	96.18%
ResNet18+B	99.67%	91.91%	95.33%
ResNet18+A+B	**99.71%**	**93.85%**	**96.62%**

A and B denotes attention modules A and B.

WiHF only focus on the period of the motion change and filter out the information in the remaining period, leading to certain information loss.

We also compare WiGRUNT with previous attention-based activity recognition research ABLSTM [24]. But unfortunately, ABLSTM [24] requires that the input data of the network should be with the same length, where the dataset used in our article, i.e., Widar3, fails to meet such a condition. Therefore, we tend to the dataset used in ABLSTM for a fair comparison, which can be publicly accessed in https://github.com/ermongroup/Wifi_Activity_Recognition. For this dataset, each person performs one activity for 20 s during data collection. Note that, at the beginning and the end of an activity, the person remains stationary so that the activity itself can be easily isolated. We implemented both schemes on this dataset and performed 10-fold cross-validation, where the experimental results are shown in Fig. 9. Compared to LSTM and ABLSTM, WiGrunt still yields the best performance by archiving 98.67% accuracy on average. This result also confirms our intuitive thinking that key features of human activities have also been scattered into different subcarriers on spatially distributed receiving antennas.

C. Impact of Different Amounts of Training Data

In this part, we design experiments to explore the impact of different amounts of training data. For in-domain, cross-location, and cross-orientation settings, we use 1/4, 1/2, and 3/4 of all training data to train our network, and then use the same test set to verify performance, and the experimental results are shown in Table VII.

It can be seen that the performance of the scheme based on deep learning depends on the amount of data. The larger the amount of data, the more discriminative features the network can learn, and the more powerful performance the model can achieve. Pretraining can reduce the amount of data required for training, and we believe that the few-shot learning technology used in other fields is also useful in wireless sensors-based human activity recognition, because the data collection is harder than computer vision.

D. 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 VIII.

One interesting observation is that the attention mechanism only provides a negligible enhancement for the in-domain scenario, i.e., 0.24% improvement, and we believe that the improvement of the attention mechanism is limited because the ResNet18 can obtain very high accuracy. But for the more challenging cross-domain scenarios (especially for the cross-orientation scenario), it proves to be quite helpful (2.96% and 2.04% improvements for cross-orientation and cross-location scenarios). Moreover, the performance of ResNet with temporal-spatial attention module A and module B is better than the basic network ResNet 18. The box-plot is shown in Fig. 10. It can be seen that in the in-domain case, the quartile distance of the results obtained by the attention-based method is very small, which means that the attention-based scheme is more robust.

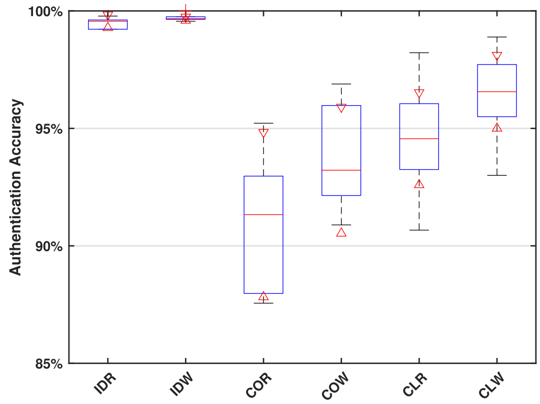

Fig. 10. Box-plot of ResNet18 and WiGRUNT. ID, CL, and CO indicate in-domain, cross-location, and cross-orientation, respectively, and R and W represent ResNet18 and WiGRUNT, respectively.

TABLE IX
IMPACT OF USING THE IMAGENET FOR PRETRAINING

Method	In-domain	CL	CO	CE
With pre-training	**99.71%**	**96.62%**	**93.85%**	**93.73%**
Without pre-training	97.76%	94.73%	86.96%	83.47%

For the cross-location and cross-orientation cases, although the attention-based scheme has a small improvement for the maximum value, it has a large improvement for the minimum value, and the quartile distance is smaller than the results obtained by ResNet18.

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.

E. Impact of Pretraining With ImageNet

In this part, we design experiments to show that pretraining the network using image datasets in the CV field is still effective. Table IX illustrates the performance of WiGRUNT while leveraging pre-training, 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 IX. The pretrained model's performance using ImageNet is higher than the model without pretraining, and it is even close to 10% higher in cross-environment gesture recognition. As shown in Table III, in the case of cross-environment gesture recognition, the number of samples in the 2nd and 3rd environments is small, and pretraining can significantly alleviate the model's inability to learn parameters due to the small training dataset.

Pretraining is commonly used in CV due to the existence of many well-constructed and well-labeled image datasets. On the one hand, pretraining acts like an initial parameter setter and plays a vital role in subsequent supervised training. On the other hand, it relieves the overfitting 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 [36]. It is also a major gain of our CSI visualization method.

TABLE X
IMPACT OF NUMBER OF GESTURES AND USERS

No. of Gestures		6	7	8	9
6 users	Indomain	99.71%	99.26%	98.32%	98.08%
	CO	93.85%	92.12%	89.22%	89.01%
	CL	96.62%	95.45%	92.25%	93.58%
No. of Users		6	7	8	9
6 gestures	Indomain	99.71%	99.45%	99.48%	99.1%
	CO	93.85%	92.86%	94.11%	93.95%
	CL	96.62%	96.21%	96.72%	96.17%

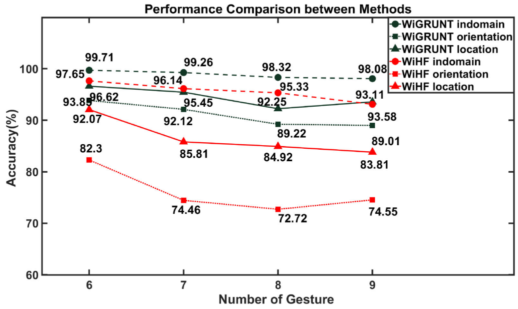

Fig. 11. Performance comparison between state-of-the-art methods.

F. Impact of Number 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 X; 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 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 [15], and the results are shown in Fig. 11, where we can see that WiGRUNT is less affected by the increase in the number of gestures.

V. CONCLUSION

This article proposes WiGRUNT, a remote sensing and noncontact solution leveraging the ubiquitous WiFi infrastructure. It roots in an 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.

Our current design still pays little attention on anti-interference, and deep learning-based methods generally require a large amount of training data. For future work, we would focus on extending WiGRUNT from laboratory to practical applications in three steps. First, we intend to extend Widar3 in various real-world scenarios under certain actual circumstances such as environmental layouts, human, and device interference. Second, we would focus on signal processing to enhance the anti-interference ability. Finally, we will explore few-shot learning and domain adaption methods that are suitable for WiFi perception, so that our solution can be more adaptable to practical scenarios. We also consider using blind source separation and beamforming technology to meet the needs of multiuser scenarios.

Also, hand gestures are manually segmented in Widar3. But its various real-world applications like sign language require automatic gesture segmentation, which constitutes another open problem calling for further attention.

References

[1] M. C. Corballis, "The gestural origins of language," *Wiley Interdiscipl. Rev.: Cogn. Sci.*, vol. 1, no. 1, pp. 2–7, 2010.

[2] Y. Song, D. Demirdjian, and R. Davis, "Continuous body and hand gesture recognition for natural human–computer interaction," *ACM Trans. Interactive Intell. Syst.*, vol. 2, no. 1, pp. 1–28, 2012.

[3] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Trans. Systems, Man, Cybernetics, Part C: Appl. Rev.*, vol. 37, no. 3, pp. 311–324, May 2007.

[4] Z. Ren, J. Yuan, J. Meng, and Z. Zhang, "Robust part-based hand gesture recognition using kinect sensor," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1110–1120, Aug. 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 Trans. Syst., Man, Cybern. Part A: Syst. Humans*, vol. 41, no. 6, pp. 1064–1076, Nov. 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 Trans. Mobile Comput.*, vol. 16, no. 2, pp. 381–393, Feb. 2017.

[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 J.*, vol. 19, no. 8, pp. 3041–3048, Apr. 2019.

[8] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human–computer interaction: A survey," *Artif. Intell. Rev.*, 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 Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 677–695, Jul. 1997.

[10] Y. Ma, G. Zhou, and S. Wang, "Wifi sensing with channel state information: A survey," *ACM Comput. Surv.*, 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 *Proc. IEEE Conf. Comput. Commun.*, 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 *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2016, pp. 250–261.

[13] L. Zhang, Y. Zhang, and X. Zheng, "Wisign: Ubiquitous american sign language recognition using commercial wi-fi devices," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 3, pp. 1–24, 2020.

[14] Y. Zheng *et al.*, "Zero-effort cross-domain gesture recognition with Wi-Fi," in *Proc. 17th Annu. Int. Conf. Mobile Syst., Appl., Serv.*, 2019, pp. 313–325.

[15] C. L. Li, M. Liu, and Z. Cao, "WiHF: Gesture and user recognition with WiFi," *IEEE Trans. Mobile Comput.*, vol. 21, no. 2, pp. 757–768, Feb. 2022.

[16] 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 Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 3, no. 3, pp. 1–26, 2019.

[17] Q. Pu, S. Gupta, S. Gollakota, and S. Patel, "Whole-home gesture recognition using wireless signals," in *Proc. 19th Annu. Int. Conf. Mobile Comput. Netw.*, 2013, pp. 27–38.

[18] Y. Wang, J. Liu, Y. Chen, M. Gruteser, J. Yang, and H. Liu, "E-eyes: Device-free location-oriented activity identification using fine-grained wifi signatures," in *Proc. 20th Annu. Int. Conf. Mobile Comput. Netw.*, 2014, pp. 617–628.

[19] L. Sun, S. Sen, D. Koutsonikolas, and K.-H. Kim, "Widraw: Enabling hands-free drawing in the air on commodity wifi devices," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw.*, 2015, pp. 77–89.

[20] R. H. Venkatnarayan, G. Page, and M. Shahzad, "Multi-user gesture recognition using wifi," in *Proc. 16th Annu. Int. Conf. Mobile Syst., Appl., Serv.*, 2018, pp. 401–413.

[21] N. Yu, W. Wang, A. X. Liu, and L. Kong, "Qgesture: Quantifying gesture distance and direction with wifi signals," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 1, pp. 1–23, 2018.

[22] K. Ali, A. X. Liu, W. Wang, and M. Shahzad, "Keystroke recognition using wifi signals," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw.*, 2015, pp. 90–102.

[23] Y. Gu, X. Zhang, Z. Liu, and F. Ren, "WiFE: WiFi and vision based intelligent facial-gesture emotion recognition," 2020, *arXiv:2004.09889*.

[24] Z. Chen, L. Zhang, C. Jiang, Z. Cao, and W. Cui, "Wifi CSI based passive human activity recognition using attention based BLSTM," *IEEE Trans. Mobile Comput.*, vol. 18, no. 11, pp. 2714–2724, Nov. 2019.

[25] D. Wu *et al.*, "Fingerdraw: Sub-wavelength level finger motion tracking with wifi signals," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 4, no. 1, pp. 1–27, 2020.

[26] 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 *Proc. IEEE Glob. Commun. Conf.*, 2018, pp. 1–6.

[27] 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, Nov. 2019.

[28] M. Kotaru, K. Joshi, D. Bharadia, and S. Katti, "Spotfi: Decimeter level localization using wifi," in *Proc. ACM Conf. Special Int. Group Data Commun.*, 2015, pp. 269–282.

[29] X. Li, S. Li, D. Zhang, J. Xiong, Y. Wang, and H. Mei, "Dynamic-music: Accurate device-free indoor localization," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2016, pp. 196–207.

[30] 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 Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 3, pp. 1–19, 2018.

[31] A. Joshi, H. Sierra, and E. Arzuaga, "American sign language translation using edge detection and cross correlation," in *Proc. IEEE Colombian Conf. Commun. Comput.*, 2017, pp. 1–6.

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

[33] X. Zhu, D. Cheng, Z. Zhang, S. Lin, and J. Dai, "An empirical study of spatial attention mechanisms in deep networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6688–6697.

[34] 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.

[35] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Euro. Conf. Comput. Vis.*, 2014, pp. 818–833.

[36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.