Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2022.017654

ARTICLE

Action Recognition Based on CSI Signal Using Improved Deep Residual Network Model

Jian Zhao[1,*], Shangwu Chong[1], Liang Huang[1], Xin Li[1], Chen He[1] and Jian Jia[2,*]

[1]School of Information Science and Technology, Northwest University, Xi'an, 710127, China

[2]School of Mathematics, Northwest University, Xi'an, 710127, China

[*]Corresponding Authors: Jian Zhao. Email: zjctec@nwu.edu.cn; Jian Jia. Email: jiajian@nwu.edu.cn

Received: 28 May 2021 Accepted: 09 July 2021

ABSTRACT

In this paper, we propose an improved deep residual network model to recognize human actions. Action data is composed of channel state information signals, which are continuous fine-grained signals. We replaced the traditional identity connection with the shrinking threshold module. The module automatically adjusts the threshold of the action data signal, and filters out signals that are not related to the principal components. We use the attention mechanism to improve the memory of the network model to the action signal, so as to better recognize the action. To verify the validity of the experiment more accurately, we collected action data in two different environments. The experimental results show that the improved network model is much better than the traditional network in recognition. The accuracy of recognition in complex places can reach 92.85%, among which the recognition rate of raising hands is up to 96%. We combine the improved residual deep network model with channel state information action data, and prove the effectiveness of our model for classification through experimental data.

KEYWORDS

Action recognition; residual deep network; network model; channel state information

1 Introduction

Action recognition technology is an intelligent model technology. It uses the deep neural network to analyze the human action data, and then recognizes human actions through the deep neural network algorithm [1]. Traditional action recognition [2,3] can be roughly divided into wearable recognition technology and non-contact recognition technology according to the difference between contact and non-contact. Wearable action recognition technology is mainly to wear sensors on the human body, receive signals through the contact between the sensor and the body, and then perform action recognition through the analysis of these data. Common sensors mainly include acceleration sensors, pressure sensors, gyroscopes, and tilt switches. Non-contact [4] identification technology is to identify by using cameras [5], radar [6], etc. Although the technology of sensor-based behavior recognition is mature and accurate, there is a serious drawback that the sensor must be worn on the body at all times during the recognition process. This is only suitable

for the data acquisition and research stage, if you want to apply it to real life behavior recognition is not realistic. Non-contact recognition is mostly based on video recognition technology, using high-definition video recognition technology to recognize behavior through the human action in the video frame, but in the process of recognition, the portrait information of the person is also collected by the computer. If the blind spot is blocked, the accuracy is greatly reduced. The behavior recognition based on radar technology has a short wavelength and cannot cover a large area.

Action recognition [7] based on channel state information (CSI) uses common WIFI signals [8]. Nowadays, social wireless networks are ubiquitous. This not only avoids the wearing problems caused by wearable sensors, but also greatly saves the cost of equipment to a certain extent. The WIFI signal has the function of penetrating the wall. At the same time, because the WIFI signal is affected by the action of the human body, the channel multipath characteristics of the WIFI signal will change when the object or the human body moves. As the amplitude of the motion is different, the amplitude and size of the channel will also be different. Since the WIFI signal is not affected by temperature, light, etc., and there is no need to see the subject's appearance characteristics clearly during the recognition process, the drawbacks of using the camera are perfectly avoided. WIFI signal has the advantage of non-intrusiveness, and its application in daily life greatly protects people's privacy. Among multiple technologies such as ultra-wideband, radar, and wireless networks, recognition technology based on wireless networks has attracted more and more attention.

At present, the behavior recognition technology based on wireless signals mainly includes two aspects, one is based on the strength of the wireless signal (RSS), and the other is based on the channel state information (CSI) of the wireless signal. RSS provides coarse-grained information about communication links, and CSI signals provide more fine-grained physical layer information. There is a great correlation between the amplitude and phase of each sub-carrier, which can make a sufficient fine-grained reflection of the subtle changes of the human body.

This paper combines deep neural network with CSI signals [9]. In deep neural networks, we use CNN feature extraction capabilities to extract action feature information from CSI signals and organize these feature sets and send them to the network. Analyze the characteristics of CSI data through the powerful computing power of neural network. Because the signal characteristics of different actions are different, we can use deep learning to correspond these characteristics to different actions, so as to achieve the effect of different action recognition [10]. Due to the easy optimization of the residual network, the accuracy can be improved by adding a certain depth, and the jump connection of the internal residual module alleviates the problem of gradient disappearance caused by the neural network. In the framework, we selected the RestNet network as the basic framework of the network. The knowledge of the wide network is used at the front end of the network, and the feature extraction of the CSI data is carried out by continuously increasing the width of the CNN network. Broadening the width of the network has been proven to improve the performance of the network to a certain extent. At the same time, we know that the optimal width of the initial width of the RestNet network is 64. This will be further explained in the feature extraction network later. In order to further enhance the performance of the network, we introduce the network knowledge of the shrinkage network and threshold of the attention mechanism. The attention mechanism can enhance the connection between the front and back ends of the network. The threshold here is obtained by automatic calculation of the neural network, avoiding the trouble of traditional manual calculation, and using the threshold to reduce or eliminate signals that are irrelevant or weakly related to the main information in the feature to

a certain extent, so that the final recognition effect of the network is enhanced. Since our network is to improve the network by combining the attention mechanism and the concept of widening the network, we named the network Attention-Wild-RestNet (AWRSN) network.

The contributions of this paper are as follows: This paper uses attention mechanism and threshold method to improve the residual module on the basis of the traditional residual network. According to the data characteristics of different channels, the threshold parameters of each channel are calculated adaptively, and the network is adjusted by the threshold values. We use the threshold module to improve the identity connection of the traditional residual network and design three network models by broadening the network. The experimental data show that the improved network model has better performance. The experiment confirmed that, under the conditions of the experimental environment and the state of the equipment, the accuracy of the volunteers' actions when they were standing was higher than when they were sitting down.

Other organizations of the paper are as follows: The second part introduces the historical situation and overview of CSI signals, and introduces the related concepts of threshold and attention mechanism. The third part introduces the related knowledge of CNN feature extraction and classic residual network, and then introduces the specific framework structure of our network model. The fourth part introduces the experimental scene and CSI data collection, and shows the corresponding experimental results. The fifth part makes a brief summary of this paper.

2 Structure

2.1 Attention Mechanism and Width

The human attention mechanism is a means and method for the human brain to quickly delete and select high-value from the large amount of information already obtained in the mind. This method is introduced into network learning in deep learning [11]. This attention mechanism is widely used in various types of deep learning tasks such as natural language processing, image classification and speech recognition, and has achieved remarkable results. The paper [12] indicates that through a series of convolutional layers and nonlinear activation functions combined with downsampling, the CNN network can capture the hierarchical mode of the picture and obtain a global view. At the same time, studies have shown that the generated features can be enhanced by integrating the learning mechanism into the CNN network, which helps to capture the spatial correlation between the features. Hu et al. [12] introduced a new structural unit, called squeeze and excitation (SE) block, when studying the relationship between the network design on the channel, and its purpose is to explicitly model the network convolution feature the interdependence between channels improves the quality of the representation produced by the network. The network proposes a mechanism that allows the network to perform feature recalibration. Through this mechanism, the network can learn to use global information to selectively emphasize information features and suppress information that is less relevant to the original information. When using the attention mechanism, considering that the average value is only related to first-order information, it will cause a lot of information to be lost when using the global average at the end of the weighted neural network for image representation [13]. So we quote an incentive module (rSE-block) in the paper to make up for a lot of information lost in SEblock due to global average pooling. We combine the attention mechanism with the residual network. By adding the attention mechanism, we use this mechanism to extract the relevant information of the CSI signal that is ignored by the network, and then finally add the acquired information to the network to improve the network Generalization ability of CSI signal.

Under certain circumstances, the accuracy of the network will increase with the depth of the network, but when the depth of the network reaches a certain level, if the accuracy of the network is to be further improved, the network depth must be doubled. A method of increasing the width of the network is proposed to improve the overall performance of the network [14]. A wide network can enable each layer to learn more abundant features, so as to increase the width of the network and reduce the number of network layers under certain circumstances, and also improve the accuracy of the network. Some scholars have proved that increasing the width of the network can improve the performance of classification [15–18]. In order to ensure that the accuracy of the existing network is improved while the number of network layers is basically unchanged, we use CNN as the feature extraction layer at the beginning of the data input. Knowing that the optimal network width of the input layer of RestNet is 64, so our input here increases the network width to 64 layers and doubles the span. By increasing the network width layer by layer, the existing network has a limited number of layers. Based on each layer, better feature extraction performance is obtained.

2.2 Threshold

Thresholds are divided into soft thresholds and hard thresholds. The hard threshold is a threshold that cannot be exceeded by the detected data. The soft threshold means a threshold that specifies the range of variation of the detected data. In the process of data analysis, if the data is truncated at the change, that is, a certain segment of data is all set to 0, a sudden change will occur at the data waveform. The use of soft threshold is a method often used in current data analysis and data processing. Among them, the concept of soft threshold was first proposed in the papers [19,20], taking the soft threshold expression of paper [19] as an example threshold expression method, the soft threshold expression is:

$$\eta_s(\omega,\lambda) = \mathrm{sgn}(\omega)(|\omega| - \lambda)_+ \tag{1}$$

In the expression (1), ω is a variable, λ is a threshold (non-negative value), sgn is a symbolic function, and the sgn $(|\omega|\text{-}\lambda)+$ means that when $(|\omega|\text{-}\lambda)>0$ is equal to $|\omega|\text{-}\lambda$, when $(|\omega|\text{-}\lambda)<0$, it is equal to 0. According to the characteristics of the symbolic function, the expression here can be rewritten as:

$$\eta_s(\omega,\lambda) = \begin{cases} \omega+\lambda, & \omega \le -\lambda \\ 0, & |\omega| \le \lambda \\ \omega-\lambda, & \omega \ge \lambda \end{cases} \tag{2}$$

Literature [21] also uses another expression of threshold. The specific details of the expression are:

$$\mathrm{soft}(u,a) \equiv \mathrm{sign}(u)\max\{|u|-a,0\} \tag{3}$$

The function of max|u|-a,0 is the same as that of $(|w|-\lambda)+$ in formula (1). When $(|u|-a) > 0$, $\max|u|-a,0 = (|u|-a)$, and when $(|u|-a) < 0$, $\max\ |u|-a,\ 0 = 0$. The expression of the threshold formula in the papers [22] and [23] is:

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

where is the threshold. We can find that the expressions of these soft thresholds have the same meaning, whether sgn(x) or sign(x) is a symbolic function, where the result is 1 when $x > 0$, and when $x < 0$ The result is -1. The meanings of these expressions about soft thresholds are the same, of course (4) expressions are easier to understand. Since CSI data is sensitive to changes in the surrounding environment, and the data is related to the front and back, the research direction of this paper chooses to use the performance of soft threshold to further analyze the data. Among them, the use of soft threshold can make the data near zero to be set, that is, to filter out the components of irrelevant signals. When the data is less than 0, it is not all set to 0, but the basic characteristics of the negative data are retained, ensuring the maximum data characterize.

3 Improved Residual Network Model

The soft threshold method has been applied in many documents. This article introduces the concept of soft threshold to further optimize the network. In this section we will describe this paper network knowledge and detailed explanation of the specific implementation steps of our network.

3.1 Feature Extraction

The CSI signal we collect is a series of actions. The correlation between the beginning and the end of the signal is similar to the correlation between pixels that are far apart in the picture. There is a strong correlation between the adjacent signals. To some extent, there is a similar relationship between our CSI signal and the picture, so here we use a convolutional network similar to the picture to extract the characteristics of the CSI signal. Finally, through network processing, the information of the entire CSI sequence is obtained through these characteristic signals.

In order to fully extract the information of the input CSI data, we improved the collection of the network signal on the basis of the residual network, and increased the network's feature extraction ability of the input CSI signal by adding the convolution layer of the input signal. We found that the performance of the network is improved with the increase of feature extraction layers by fixing other network structures.

Since the number of sub-carriers of each antenna of the CSI signal is 30, we set the input dimension to 30 in the early stage of signal input, the size of the convolution kernel in the first layer of convolution is set to 7, the step size is set to 2, and the padding is 3. No offset, the output dimension is set to 128. In order to match the signal features output by the final feature layer with the network input features, the step size of the remaining three convolutional layers is set to 1, the size and padding of the convolution kernel are not changed, and no bias is used. The input dimension of the second convolutional layer is the output dimension of the first convolutional layer, and the output dimension of the second layer is set to 256. Similarly, in order to ensure accurate transmission of network dimensions, we set the input dimension of the next layer of convolutional layer to be the output dimension of the previous layer of convolutional layer. The output dimension of the third layer is set to 512, and at the bottom we set the dimension of the feature output layer back to 128. In order to ensure fast data learning and strong generalization ability, we use BN to normalize the data, use the nonlinear function sigmoid as the activation function, and finally use the output data as the feature input of the final network.

3.2 Residual Block

In depth learning, the effect of increasing training with the number of layers of the network will be better and better. But as the depth of the layer is deeper, it is inevitable that there

will be problems such as gradient disappearance and gradient explosion [24,25]. This will cause the performance of the network to become worse and worse as the number of network layers increases. However, Standardization [25] and intermediate layer standardization [26] solve this problem to a large extent. It makes the network with dozens of layers start to converge to the stochastic gradient descent with directional propagation. When deeper networks can begin to converge, as the depth of the network increases, the accuracy reaches saturation and then rapidly degrades. In the paper [27], a deep residual learning framework is introduced to solve this type of degradation problem.

In the process of building the network, we used Batch Normalization (BN) as the normalization of the data, and the activation function was Relu. It is an optimization training method proposed by Google [26]. The BN algorithm is generally placed before the activation function. Normalize the data before entering the activation function, which will help the outcome of the input data offset and increase the impact. BN helps speed up training and increase learning rate. A fast learning rate can be generated even when the learning rate is small. In addition, after the BN algorithm, the training data set can also be summarized to prevent the training from shifting. The expression of BN is shown below:

$$\mu = \frac{1}{N_{\text{batch}}} \sum_{n=1}^{N_{\text{batth}}} x_n \tag{5}$$

$$\sigma^2 = \frac{1}{N_{\text{batch}}} \sum_{n=1}^{N_{\text{batch}}} (x_n - \mu)^2 \tag{6}$$

$$\hat{x}_n = \frac{x_n - \mu}{\sqrt{\sigma^2 + \varepsilon}}, \quad y_n = \gamma \hat{x}_n + \beta \tag{7}$$

where x_n and y_n respectively represent the input and output features of the nth observation in the smallest batch. γ and β are the trainable parameters of the two measured and transformed distributions, and ϵ is a constant close to zero.

Activation function we use the most popular Relu function. As shown in Fig. 1, Relu is a piecewise linear function. When x > 0, the gradient is always only 1, there is no gradient consumption, and the convergence speed of the network will also increase. When x < 0, the negative value becomes 0, and the integrity remains unchanged. The more neurons that are 0 after the training is completed, the greater the sparsity, the more representative the extracted features and the stronger the generalization ability. This is a one-sided suppression operation, which can make neurons in the neural network have coefficient activation in order to better mine the phase characteristics of the data for our model and better fit the data.

$$\text{ReL } U(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases} \tag{8}$$

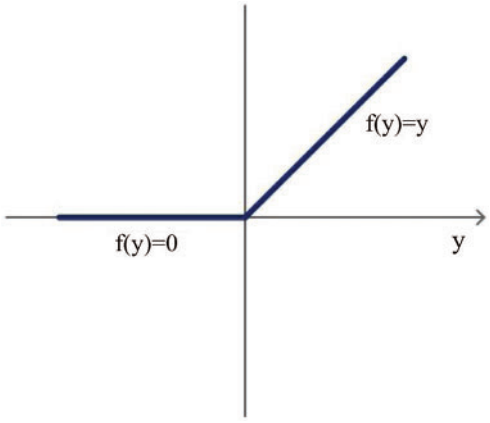

Figure 1: Relue function

Compared with ordinary convolutional networks, residual networks use cross-layer identity connections to reduce the difficulty of convolutional neural networks. ResNet [27] has been proven effective in multiple tasks, such as the paper [28] for target detection and [27] for image classification. The traditional residual module is shown in Fig. 2 below. The module is divided into two channels by input x into the network. After input x passes through the weight layer, the output data function is activated by BN and activation function Relu. Finally, the result is sent to the weight layer, and the final output function F(x) is added to the x of the shortcut path, and the result of the addition is sent to the Relu activation function. The final result is the output of the module.

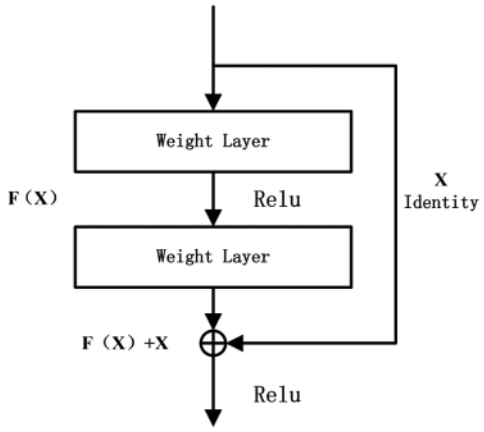

Figure 2: Traditional residual network model

The standard residual network implementation is used to process 2D data input, such as two-dimensional image input. Since most of the inputs of CSI signals are one-dimensional inputs, in order to adapt to our research direction and signal characteristics, we change the network structure to one-dimensional in order to adapt to the CSI signal. Fig. 3 shows the 1D residual module, where x is the input data of the module, and we use y to represent the final output of the module. In the one-dimensional residual module, we change the convolution function and BN into one-dimensional style. Since x has undergone two layers of convolution, the size of the

output f(x) and x may be different. The 1D convolution and BN1D in the shortcut are to ensure that the size of f(x) is the same as the size of x.

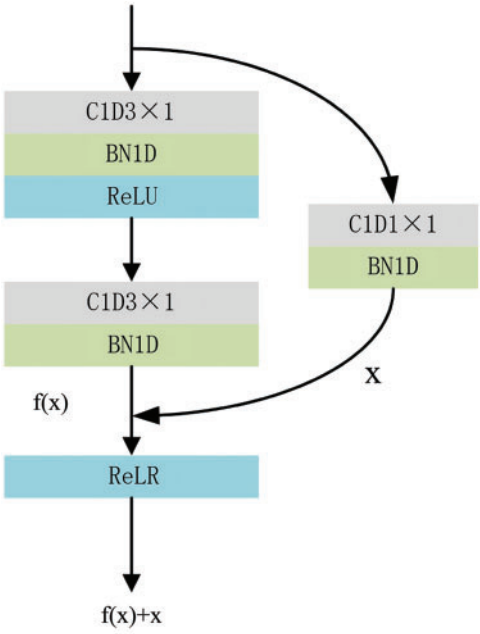

Figure 3: One-dimensional residual network model

3.3 Network Structure

In this part, we introduce the concept of soft threshold and attention mechanism into the deep learning method, mainly to improve the residual module of the residual network, and then embed the entire network to improve the overall performance of the network. We increase the width of the network without changing the number of network layers, and at the same time introduce an attention mechanism to reduce the components that are not related to the signal through the network self-learning threshold. In order to facilitate the introduction of the network, we named the improved module AWTR in this paper. This is the abbreviation of Attention-Wild-Threshold-RestNet, and the network composed of this module is called AWTR network.

3.3.1 Model 1-Residual Network Based on Soft Threshold

The AWTR network is an improvement of the ResNet network by using the attention mechanism and the concept of soft threshold. We use the attention mechanism [12] to introduce multi-channel learning features to learn the input CSI data, and use the laziness between channels of the branch network convolution feature to improve the quality of network results. The attention mechanism can make the network use the global information to learn the input data, and at the same time, the soft threshold embedding makes the network weaken or even filter out the information weakly related to the main information in the data.

The following Fig. 5 is the threshold parameter extraction network model in the reference paper [22], named Threshold Moldel. When the threshold network is data input, we use the output of the two-layer convolution of the original residual module as the input of the threshold network. In order to use the network to set a suitable threshold, the output data after the two-

layer convolution will continue to be introduced into the branch flow GAP operation to obtain 1D vector. Its function is to summarize the overall information of the input data, and then propagate the vector into the two-layer fully connected layer network. After a series of nonlinear and convolution operations, due to the nonlinear characteristics of the relu function and the characteristic curve The characteristic curve of the soft threshold function is very similar, so we use Relu as the intermediate activation function in the network. A sigmoid function is used at the end of the network to scale the parameters to (0,1). The range can be expressed as:

$$\alpha = \frac{1}{1+e^{-z}} \tag{9}$$

where z is the output of the two fully connected layers in the branched network, and is the scaling parameter of the network settings we finally obtained. In order to ensure the original characteristics of the data, we keep the positive and negative signs corresponding to the threshold unchanged without changing the data structure. We set the feature information that is considered unimportant within the threshold to 0, and the positive and negative signs outside the threshold Both the value and the negative value are adjusted according to the threshold, so as to ensure that both the positive data and the negative data in the data retain their characteristics, and the data is adjusted. Fig. 4 is a schematic diagram of the threshold processing process.

From the figure we can see that the slope of the output of the input data is 1 or 0, which has a considerable inhibitory effect on preventing the gradient from disappearing and exploding in the network. The derivative expression is:

$$\frac{\partial y}{\partial x} = \begin{cases} 1 & x > \tau \\ 0 & -\tau \le x \le \tau \\ 1 & x < -\tau \end{cases} \tag{10}$$

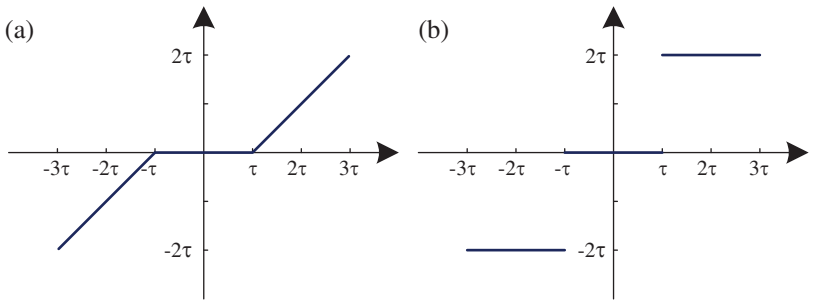

Figure 4: Threshold diagram: (a) shows the output slope of the input data; (b) shows the value of the slope

The first two convolution modules after we input the data into the original network [29] are called Residual Block. The branch network that sets and adjusts the threshold of the network is represented by multiple Threshold Models, and finally the non-linear function Sigmoid is used as output. We regard Fig. 5 as an improved module of the residual module. The difference is that the module can automatically learn through the network and set the threshold for the adjustment of data parameters, avoiding unnecessary trouble caused by manual setting. Parameters can be improved layer by layer according to the addition of modules. We superimpose the SE module

in the original residual module and the threshold module, and adjust the network layer by layer through the branched attention mechanism and threshold.

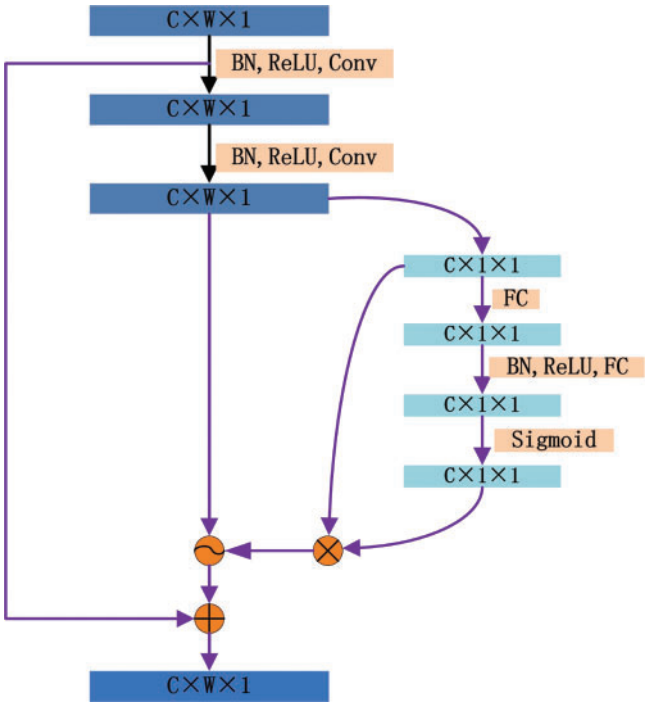

Figure 5: Unilateral residual network model

3.3.2 Model 2-Widened Soft Threshold Network 1

The paper [14] proves that widening the width of the network under certain conditions can help improve the performance of the network. In order to enable the network to learn the characteristic information of the signal more comprehensively, we further improve the network framework on the basis of the network framework 1. We input the input signal into two AWRSN modules at the same time. Figs. 3–8 show the improved network model. We regard the two convolutional layers through which the data first came in as general Residual modules. The side module with soft threshold is composed of Threshold Model. The signal feature is extracted and analyzed through the dual-channel module, and finally passes through a sigmoid function. And then take the weighted average of the output signal, that is, the output:

$$y = \frac{x1 + x2}{2} \tag{11}$$

For the convenience of recording, we call this module AWTR-2 Block.

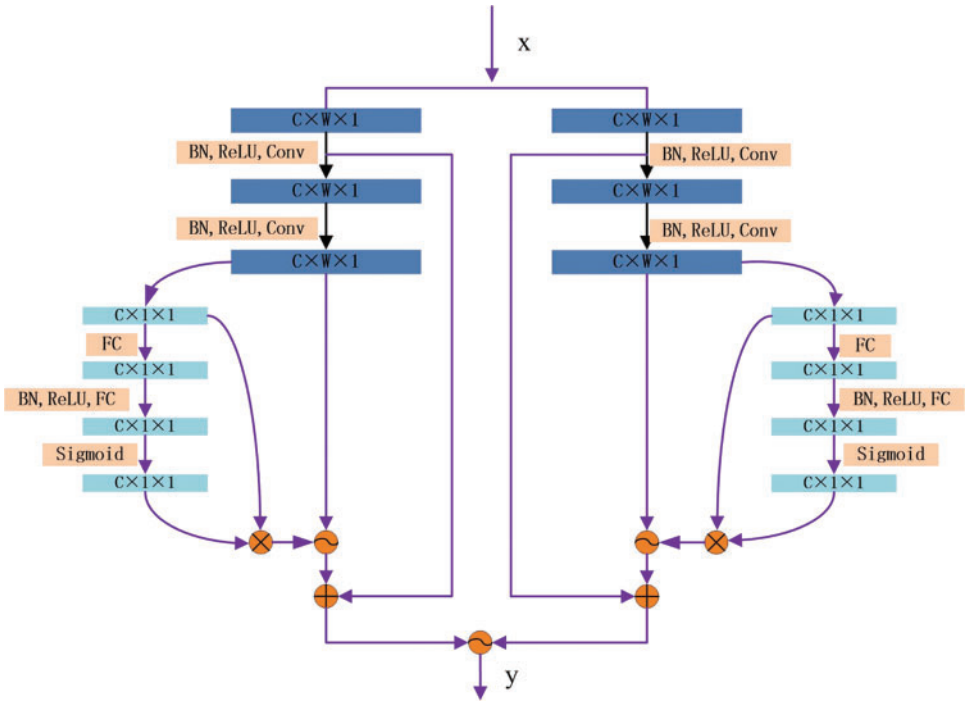

Figure 6: Bilateral residual network model

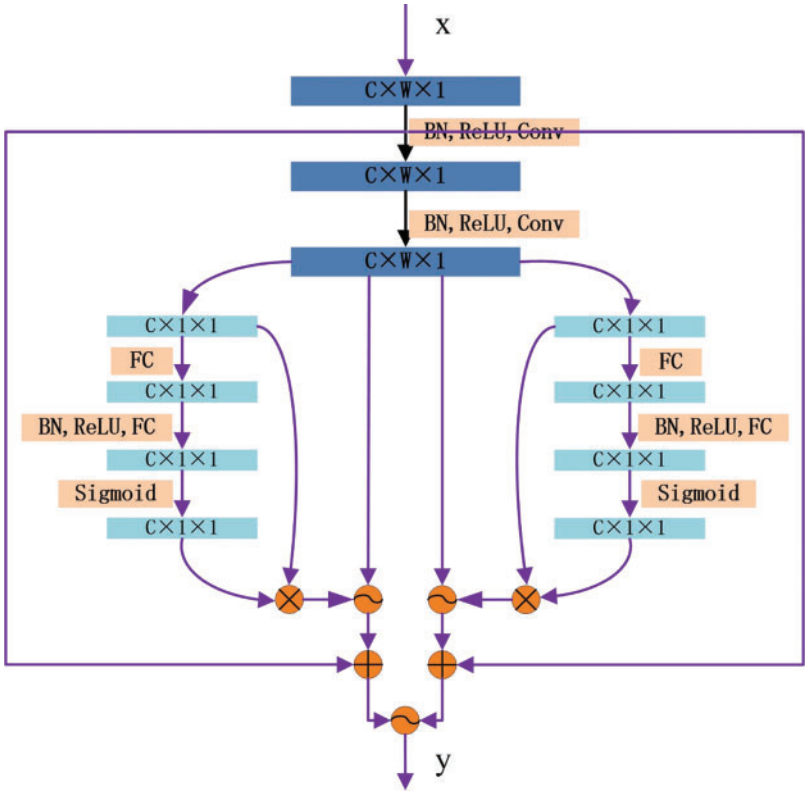

Figure 7: Bilateral residual network model: This is an improvement of the bilateral residual network model, which is used to compare the experimental results

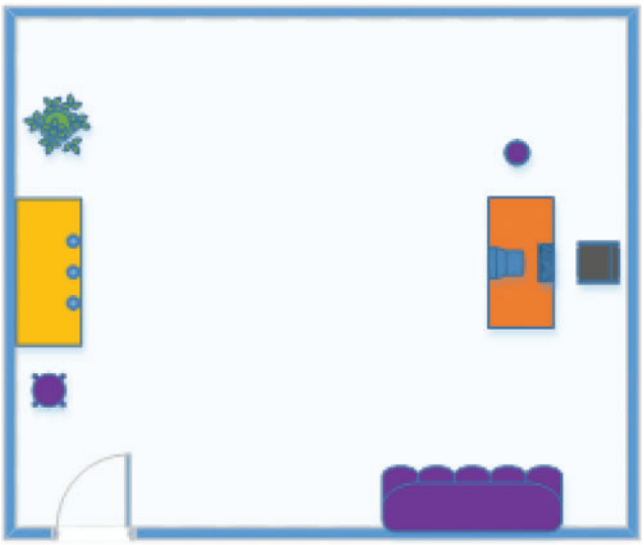

Figure 8: Experimental environment 1: Office environment

3.3.3 Model 3-Widened Soft Threshold Network 2

The third network structure we designed is as shown in the figure. After the data passes through the residual module, the output information is passed through two threshold modules to extract dual thresholds, and then use the extracted thresholds as network data, and finally combine residual features. The parameter items passed down by the short-circuit path are combined. The two parameters extract data separately. With the continuous superposition of the number of layers, the new parameters are constantly updated. Finally, we have verified that the effect of taking the average of the outputs of the two modules is better than the sum. In this paper, we named this module AWRSN-3 Block.

4 Collection and Analysis of Experimental Data

In this part, we first introduced the test equipment and briefly introduced the on-site data collection. Secondly, we briefly discussed the problems and correction methods in the experiment process. Finally, we proved the effectiveness of our network and method through a series of comparative data. In addition, our data set is still being supplemented and improved.

4.1 Experimental Equipment

Our data collection tool is to use CSI Tool, and the data collection mode is Moniter mode. Use an ASUS notebook that has replaced the Intel 5300 network card as the signal transmitter, and a desktop computer with the built-in network card replaced with a 5300 network card as the signal receiver. The transmitting end adopts the built-in antenna of the notebook computer, set to 1 antenna (Tx) to transmit data, the receiving end (Rx) adopts the external extension antenna of the desktop computer, and 3 external antennas are used as the receiving end. According to theory and practical experience, when the antenna is selected, when the length of the antenna is 1/4 of the radio wavelength, the transmission and reception efficiency of the antenna is the highest. Each receiving antenna is placed at the same horizontal spacing and half-wavelength distance. The wavelength formula is: $\lambda = u/f$, u is the wave speed, f is the evaluation rate, and λ is the

wavelength. In the experiment, we set the sampling frequency to 1 KHz, each action requires volunteers to complete within 3 s, and a total of 3,000 data packets are collected for each action.

In this data collection process, we collected motion data of two different indoor environments. Fig. 8 shows the first experiment scene. The experiment scene is a meeting room of 8 m $*$ 10 m square meters. The transmitting end and the receiving end are on a table about 1m away from the ground, and the horizontal distance is 2 m. Volunteers are required to complete the action in the middle of the signal transmitting and receiving end within a specified time. After the action is completed, they remain still and wait for the completion of the data collection and proceed directly to the next set of data collection. In order to ensure the reproducibility of the experimental results, we changed the location of the experimental data collection, and selected a meeting room with a size of 15 m $*$ 20 m as the second location for sample collection, including several tables, as shown in Fig. 9 for the experiment data of two acquisition scenes.

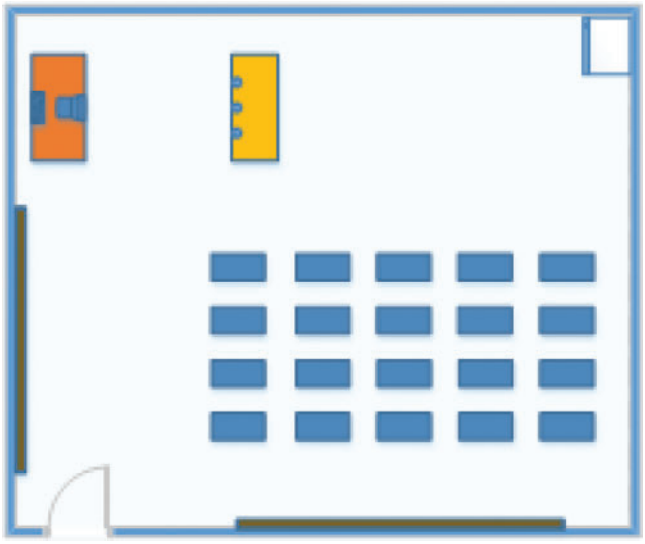

Figure 9: Experimental environment 2: Meeting room environment

4.2 Data Collection

Considering that there are too many users in the 4 GHz frequency band, and there are a large number of other equipment frequency band conflicts, data collection in the 4G frequency band may cause data loss. After that, we also confirmed that this situation does exist, so we abandon the 4G working frequency band and jointly use the working frequency of the antenna. We choose the working frequency band of the collected signal as 5.805 GHz and the channel bandwidth as 20 MHZ. This not only ensures that the working frequency band used for signal acquisition avoids the crowded 4G frequency band, but also conforms to the antenna working frequency range. This setting does improve the channel quality and the data loss situation is improved to the greatest extent.

4.2.1 CSI Data

We installed the CSI Tool tool on a laptop and a desktop computer that was replaced with an Intel 5300 network card, using a laptop antenna as the signal transmitter, and the desktop computer using three external extension antennas as the signal receiving end. Therefore, The

collected CSI data is a $1*3$ channel state information stream. Since there are 30 sub-carriers in the received data per day, the received CSI data can be expressed by the following Eq. (12):

$$H = \begin{bmatrix} H_{1,1} & H_{1,2} & \dots & H_{1,30} \\ H_{2,1} & H_{2,2} & \dots & H_{2,30} \\ H_{3,1} & H_{3,2} & \dots & H_{3,30} \end{bmatrix} \tag{12}$$

H is the collected CSI data, and each element in the CSI matrix can be expressed as:

$$H_{i,j} = \left\| H_{i,j}\left(f_k\right) \right\| e^{j\angle H_{i,j}(f_k)} \tag{13}$$

where the $\left\| H_{i,j}\left(f_k\right) \right\|$ and $\angle H_{i,j}\left(f_k\right)$ represent the amplitude and phase of the CSI signal, i and j represent the number of the quepse antenna and subcarriers. In the experiment, we can collect the amplitude attenuation, phase shift and delay of each antenna on each carrier as the research object of CSI value. In this paper, we use the amplitude as the reference signal to carry out the experiment.

4.2.2 Data Set 1

In order to simulate the scene of education-related actions, taking into account the on-site situation of the equipment, the experiment is conducted indoor. The transmitting end and the receiving end are 0.8 m away from the ground, and the horizontal distance is 2 m. Volunteers are in the middle, with their faces facing perpendicular to the line between the signal transmitter and receiver. We use a desktop computer with Intel 5300 network card as the data receiving end, and a laptop as the transmitting end. The CSI fluctuation frequency caused by traditional indoor human activities does not exceed 300 Hz. In order to ensure the integrity of the collected data, according to Nyquist sampling theorem, when the sampling frequency is greater than 2 times the highest frequency in the signal, the signal after sampling can completely retain the original signal, so we set the sampling frequency to 1,000 Hz. During the sampling process, we set the total transmission time to 3 s, and the entire data reception generates a total of 3,000 data packets. Since there is one antenna at the transmitting end and three antennas at the receiving end, each antenna has a total of 30 sub-carriers, so the final received each action CSI data is a $1*3$ channel state information stream.

This experiment collects six groups of actions. In order to simulate the authenticity of the classroom, we ask the volunteers to complete these actions while sitting down. These actions are the common actions of students in class. The actions are standing up, sitting down, listening, clapping, raising hands, bowing.

Fig. 10 shows the action diagram,we stipulate that the actions include raising hands, sitting down, listening, clapping, standing up, and lowering the head. The data in this order are labeled A1 to A6. In order to ensure the consistency of the actions, we have done certain exercise normative training for the volunteers, the action norms are as follows?

- Stand up: volunteers are required to stand up from a sitting state and keep their bodies free from shaking after standing up;
- Sit down: When sitting down, the volunteers are required to not shake their bodies before and after sitting down, and keep their bodies still after sitting down;
- Listen: The listening action requires the volunteer to sit upright and maintain the listening action;

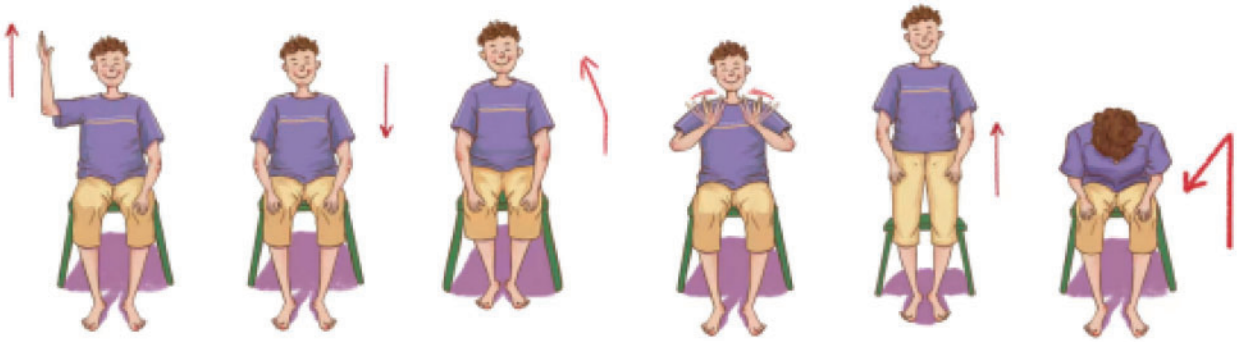

Figure 10: Experimental action view: The actions from left to right are hands up, sitting down, sitting up, clapping, standing up and bowing

- Applause: Volunteers are required to sit upright. Except for the arms, there is no action of the body. The arms are slightly bent forward and clap once, then put down to keep the body still;
- Raise hand: Raising hands requires volunteers to sit upright with one hand and raise their arms over the head. The arms may be slightly bent;
- Head down: Ask the volunteers to lower their heads and keep their heads down.

A total of six volunteers were convened in this experiment, and each volunteer performed each action 30 times. Data were collected in experiment scene 1 and experiment scene 2. Each experiment scene collected 1080 CSI data, totaling 2160 CSI data.

4.2.3 Date Set 2

In this experiment, six groups of actions were collected. During the experiment, except for the sitting action, there was an obstruction between the transmitter and the receiver when the data was collected. The rest of the action supporters remain standing, and there are no obstructions in the signal area. Respectively enable volunteers to complete specific actions within a specified time. These actions are raising hands, waving hands, applauding, bowing, walking and sitting down.

Fig. 11 shows the action diagram,we stipulate that the actions are raising hands, waving hands, applauding, bending over, walking, and sitting down, which are marked as A1 to A6 in this order. In order to ensure a certain standardization of the action data, we have conducted certain standardized training for volunteers. Is the specification requirement of the action:

- Raise hand: The volunteer keeps standing, the hand exceeds the top of the head during the hand-raising action, the arm may be slightly bent, and the body remains still for him;
- Wave: Waving action arms slightly bend upwards, palms over the head, wave with one hand once and then lower;
- Bow: The lower limbs and arms do not move when bending over, and the upper body leans forward between 45 and 50 degrees;
- Applause: Applause action design action in order to better distinguish, we asked the volunteers to straighten their arms, high-five forward, and then put them down;
- Walk: Walking is centered on the horizontal connection line between the signal transmitter and the receiver, and the distance between the beginning and the end of the walk is exactly the same. A total of three steps are taken, and the body remains still after the end;

Figure 11: Experimental action view: The actions from left to right are raising hands, waving hands, bending over, clapping hands, walking, and sitting down

- Sit down: When sitting down, keep the upper body still, and keep the body still without tilting back and forth after sitting normally.

The six action waveforms in the figure are the CSI data of A1–A6 actions in order. Depending on the action, the waveform of each action also shows different status. A total of six volunteers were recruited for this group of experiments, and each volunteer repeated each action 30 times. Finally, a total of 1080 CSI data were collected.

4.3 Data Preprocessing

In this paper, we use the amplitude signal of CSI as the analysis signal and ignore the phase signal. Due to the rigid conditions of the test environment and equipment, a few packets are lost during the data collection process. We first use interpolation to complete the missing data to ensure the integrity of the collected data. Since the CSI signal is extremely sensitive to environmental changes, outliers are likely to be generated in the process of collecting the signal. From a scientific point of view, it is necessary to calculate its standard deviation. Three times the standard deviation is the limit. Anything out of the range will be eliminated as abnormal data. We use filloutliers to eliminate outliers in the collected data. Due to the uncontrollable factors of the equipment's own hardware and environment, and the signal will generate a lot of high-frequency noise during the transmission process, the collected data will be doped with a lot of noise. For this reason, we use the most common Butterworth filter for the signal perform filtering. Fig. 12 shows the original and filtered waveforms of thirty sub-carriers of an antenna. It can be seen that the filtered signal filters out a large amount of high-frequency noise, but retains the original data trend of the signal.

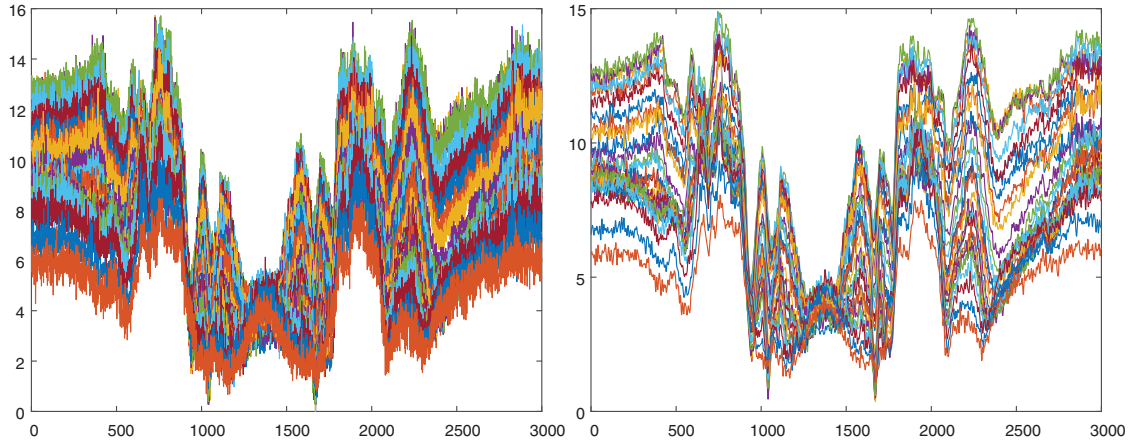

Figure 12: The original and filtered waveforms of 30 sub-carriers of an antenna: The left picture is a waveform diagram of 30 sub-carriers without filtering. The picture on the right is the waveform after filtering

Fig. 13 is a subcarrier waveform diagram of the collection action. We select the CSI data collected by an antenna and show the original signal of one action and the signal after data processing. It can be seen from the figure that the collected original signal waveform is mixed with a lot of noise signals, which makes the original waveform signal more complicated, and the signal after filtering becomes smoother.

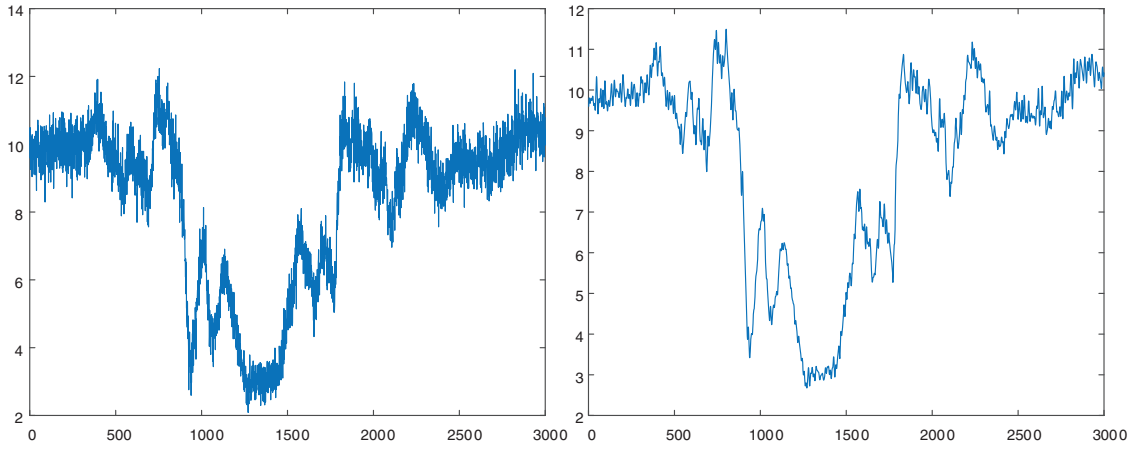

Figure 13: The original waveform and filtered waveform of a subcarrier: The picture on the left is a waveform diagram of a single sub-carrier without filtering. The picture on the right is the waveform after filtering

5 Experiment and Model Evaluation

5.1 Experiment Setup

In this section, we use the model to analyze the data sets collected in different environments, and verify the reliability of the model through the test result data. In the experiment, we use

the model's data in different experimental scenarios to analyze various aspects. Each environment recruits 6 volunteers, each of whom does six actions. The specific actions have been introduced before. Each action is repeated 30 times, the number of data collected in each environment is 1080, and a total of 2160 sets of actions are collected in the two experimental environments. After that, to verify the experimental guess, an additional 1080 sets of data are collected for verification, and finally a total of 3240 sets of data are collected data. We divide each group of data equally by 8:1:1, and divide the data into training set, test set and validation set.

5.2 The Same Action Assessment of Standing and Sitting State

In order to verify the effect of sitting and standing actions on the CSI action recognition effect, we selected the action data of the experimental area without obstructions and placed the chair as an obstruction in the experimental area, so that the experimenter was in the sitting state data that completes the same action. As shown in Fig. 14, we recorded the data of actions completed in the standing state as Case 1, and the actions completed in the sitting state as Case 2. We used three model networks to analyze the data set. It can be seen from the figure that regardless of the data of the model, the recognition accuracy of the action completed in the standing state is higher than that of the action completed in the sitting state. Considering the fine-grained signal characteristics of CSI, the receiver will receive the subtle action states of the human body when the human body acts in the standing state, so as to identify more accurately. However, in the sitting state, the receiving data not only reduces the characteristic data, but also increases the multipath effect of the signal, which invisibly increases the difficulty of the data. We concluded that subjects who were alone in the signal region had a higher recognition rate for receiving the signal than those who were sitting.

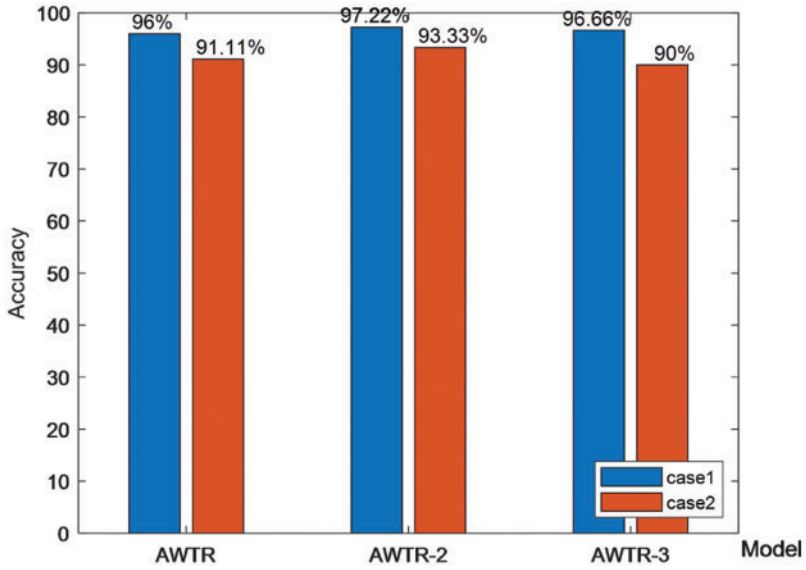

Figure 14: Comparison of the same action data between standing and sitting states

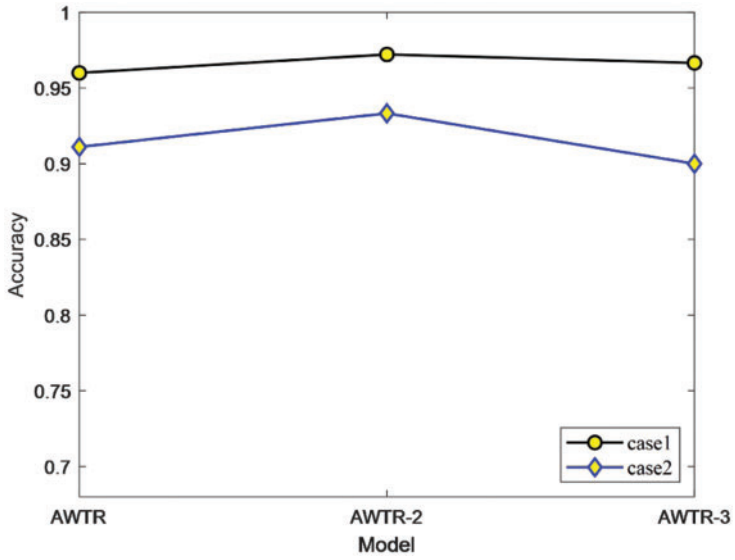

Figure 15: Comparison of the same action data between standing and sitting states

5.3 Experimental Results in Different Experimental Environments

In order to evaluate the influence of environment on CSI signal detection, we conducted experiments using the same action data collected in different environments. During the experiment, the placement of the experimental equipment and the experimental conditions were almost restored. The volunteers were still the same group of people. The only difference was the different environment. In order to highlight the influence of the experimental environment on the final precision, we tried our best to ensure the controllability of the data. The content of the actions in each environment was the same, and the placement of equipment and equipment was extremely restored.

As shown in Figs. 16 and 17, the data result we collected in Scene 1 and the data result of Scene 2 are denoised as Scene 1. In the second experiment environment, due to the complexity of the environment and the multipath effect, the accuracy of CSI recognition in the laboratory with complex environment is slightly reduced. It can be concluded from the experimental results that the environment is still very important for the recognition of CSI signals. Even when the equipment and data are highly restored, when the data collection environment is more complex, it is the accuracy of recognition under the same equipment model. It will be more difficult.

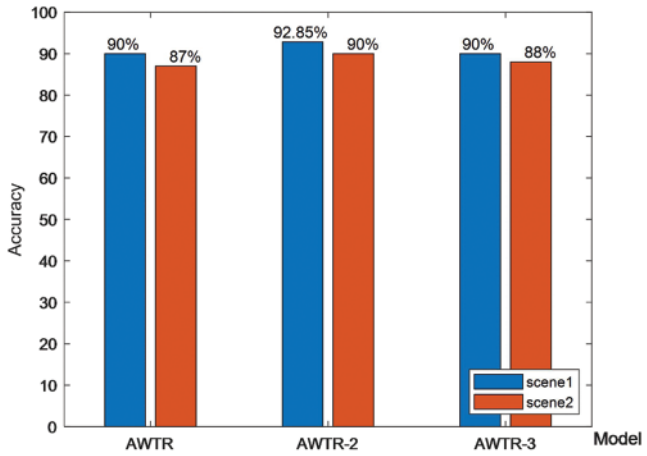

Figure 16: Data comparison of different experimental environments

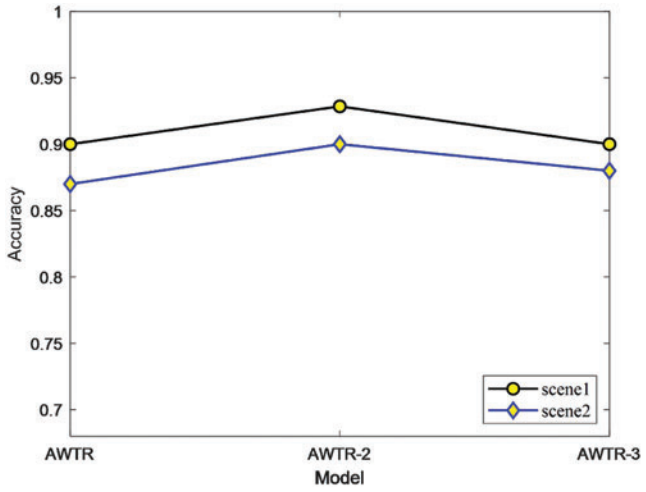

Figure 17: Data comparison of different experimental environments

5.4 Data Model Optimization

In order to make the results of the experiment more convincing, we collected data in experiment Scene 1 and experiment Scene 2 respectively at the beginning of the experiment, and the placement of equipment during the data collection process of each experiment scene was extremely restored. At the same time, it can be understood from Figs. 18 and 19 that the experimental actions in different states during the experiment also have a great correlation to the final recognition accuracy. The final accuracy of the experimental data collected when the volunteer completes the action in the standing state is better than the voluntary The action is completed by the person in the sitting state. In order to make the experiment effect better, we collected an additional set of data in experiment Scene 1. This time the data requires volunteers to complete the required actions while standing, which is the data of data set 2. To this end, we combine three network models for analysis.

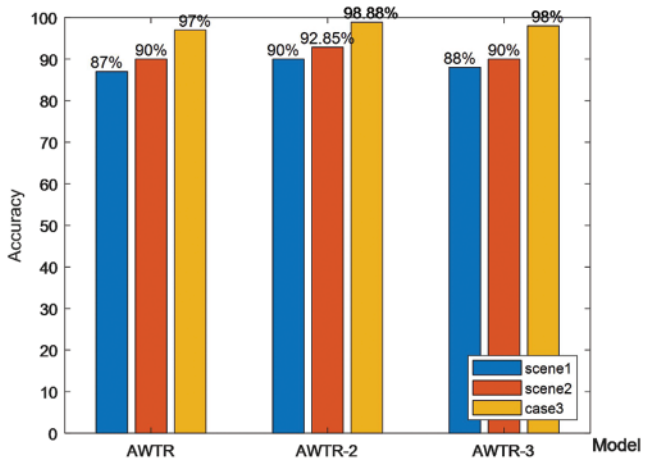

Figure 18: Data comparison of three network frameworks in different environments

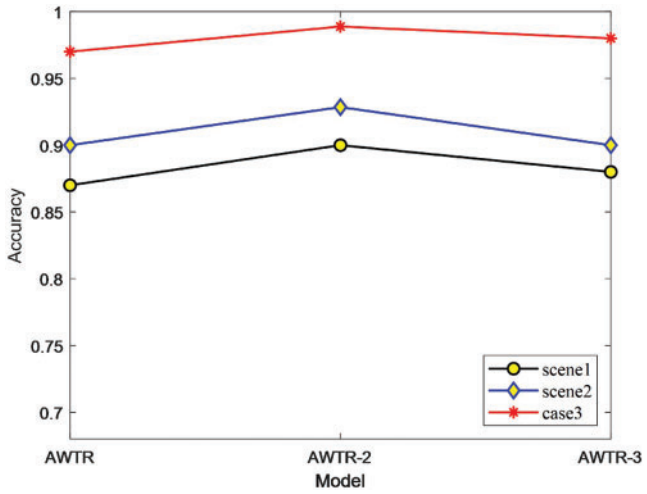

Figure 19: Data comparison of three network frameworks in different environments

According to the data of previous experiments, we know that the complexity of the experimental environment and the presence or absence of obstructions in the experimental area will have a certain impact on the final accuracy of the experiment. The data in the sitting state of environment one is recorded as Scene 1, the data of the completed action in the sitting state of environment two is recorded as Scene 2, and the action completed in the standing state of environment one is recorded as Case 3. It can be seen from the figure that when there are no other obstructions in the signal area and the volunteers complete the action data in a standing state, the effect is best, which also verifies the conclusions of our previous experiments.

5.5 Model Evaluation

The model we designed is to introduce attention mechanism and threshold on the basis of residual network. Through the attention mechanism, the network can learn the characteristics of input data and transmission data better. Meanwhile, the threshold value of network self-learning is used to weaken the part irrelevant to the signal layer by layer. Considering that increasing the width of the network helps to improve the accuracy of the network when the number of network layers is certain, we improve two network structures based on the AWTR module, and compare the performance of the three network models respectively.

In order to evaluate the performance of the model, we controlled the experimental variables as much as possible. For this reason, we combined the network model and the data collected under different environments to carry out the analysis. In Fig. 20, Res (Scene 1) represents the experimental data of the original residual network in Scene 1. We selected the residual networks of layers 18, 34 and 50 to analyze the data collected in Scene 1 and Scene 2. AWTR (Scene 1) and AWTR (Scene 2) respectively represent the data of our network model on two data sets. For the AWTR model, the abscissa number represents the three models, respectively. According to Fig. 20, combined with Figs. 14 and 15, it can be concluded that AWTR-2 model has the best performance in terms of accuracy no matter which experimental environment it is in. The highest accuracy of the residual network was 91.11 percent in the two experimental environments, while the highest accuracy of the AWTR-2 model was 92.85 percent. Even in the same environment, the effect of the residual network under the AWTR model could achieve a good effect.

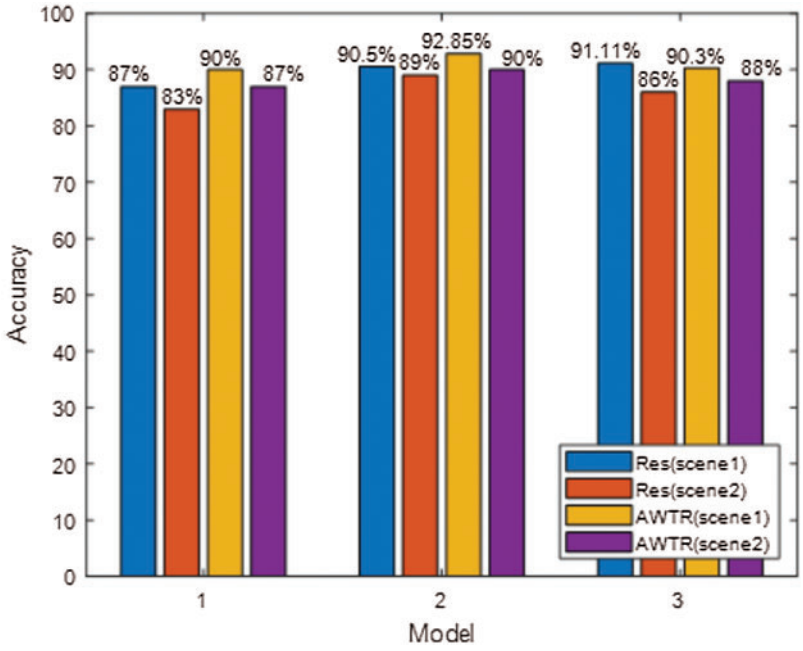

Figure 20: Evaluate the network model based on the data results in different environments

Fig. 21 shows the confusion matrix of the data set when the model AWTR-2 is unobstructed in Scene 1. It can be seen from the figure that the model recognizes 100 percent of the actions of raising hands, bending over, walking and sitting, and the recognition of waving and applauding are 96 percent and 98 percent, respectively.

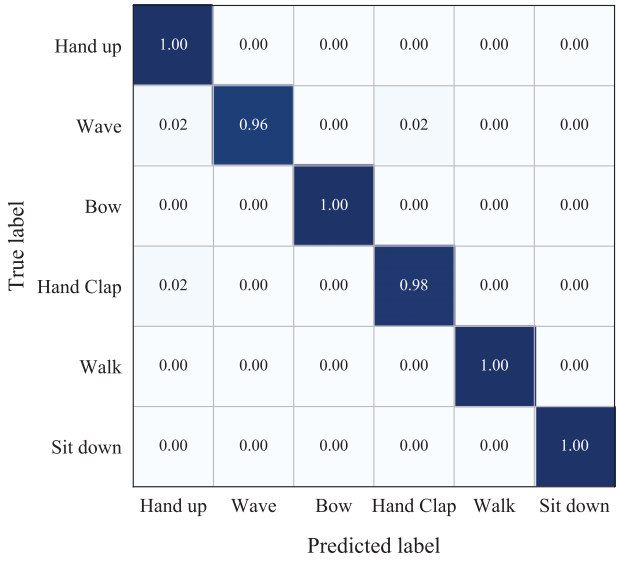

Figure 21: Confusion matrix of true label and predicted label

Accuracy recall rate is introduced to evaluate the results of the model. The accuracy calculation formula is as follows:

$$\text{precision} = \frac{TP}{TP + FP} \tag{14}$$

TP and FP respectively represent true positive and false positive, and the calculation expression of recall rate is as follows:

$$\text{recall} = \frac{TP}{TP + FN} \tag{15}$$

FN is a false negative. Table 1 shows the accuracy, precision, and recall rates for each action.

Table 1: Experimental results of AWTR-2 network

	Accuracy	Precision	Recall
Hand Up	1	0.96	1
Wave	0.95	0.96	0.96
Bow	1	1	1
Hand Clap	0.98	0.98	0.98
Walk	1	1	1
Sit Down	1	1	1

6 Conclusions

In this paper, an improved deep residual network model combined with CSI signals is used to identify human actions. We build a new model by setting the threshold module to further improve the accuracy of the network. We introduce threshold method to optimize the network through attention mechanism, and expand the network structure on the basis of the original network to improve the performance of the network. We collected three sets of data on students' movements in two environments, with a total of 3,240 data collected. In addition, our data is consistently in a phase of improvement. We use the data set to verify the impact of the environment and the experimental site conditions on the final accuracy rate, and finally use the conclusions drawn to get a new verification on the new data set, which confirms the accuracy of our verification. Our network is improved on the basis of the ResNet network. Finally, the improved network is compared with the original network. The experimental results show that the accuracy of our improved network is higher. Because the indoor action is small and difficult to distinguish, it is more difficult, so our method is more practical.

Funding Statement: This work was supported by Innovation Capability Support Program of Shaanxi (Program No. 2018TD-016) and Key Research and Development Program of Shaanxi (Program No. 2019ZDLSF02-09-02).

Conflicts of Interest: We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

References

1. Alsaade, F. W., Theyazn, H. H., Aldhyani M. H. (2021). Developing a recognition system for classifying COVID-19 using a convolutional neural network algorithm. *Computers, Materials & Continua, 68(1),* 805–819. DOI 10.32604/cmc.2021.016264.
2. Gavrilyuk, K., Sanford, R., Javan, M., Snoek, C. G. M. (2020). Actor-transformers for group activity recognition. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, WA, USA.
3. Li, Y., Ji, B., Shi, X., Zhang, J., Kang, B. et al. (2020). Tea: Temporal excitation and aggregation for action recognition. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, WA, USA.
4. Zhang, J., Shao, J., Cao, R., Gao, L., Xu, X. et al. (2020). Action-centric relation transformer network for video question answering. *IEEE Transactions on Circuits and Systems for Video Technology*. DOI 10.1109/TCSVT.2020.3048440.
5. Aggarwal, J. K., Ryoo, M. S. (2011). Human activity analysis: A review. *ACM Computing Surveys, 43(3),* 1–43. DOI 10.1145/1922649.1922653.
6. Lien, J., Gillian, N., Karagozler, M. E., Amihood, P., Schwesig, C. et al. (2016). Soli: Ubiquitous gesture sensing with millimeter wave radar. *ACM Transactions on Graphics, 35(4),* 142.1–142.19. DOI 10.1145/2897824.2925953.
7. Wang, W., Liu, A. X., Shahzad, M., Ling, K., Lu, S. (2015). Understanding and modeling of wifi signal based human activity recognition. *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, New York, NY, USA: Association for Computing Machinery. http://arxiv.org/abs/1406.6247.1406.6247.
8. Halperirr, D., Hu, W., Sheth, A., Wetherall, D. (2011). Tool release: Gathering 802.11n traces with channel state information. *ACM Sigcomm Computer Communication Review, 41(1),* 53–53. DOI 10.1145/1925861.1925870.
9. Mnih, V., Heess, N., Graves, A., Kavukcuoglu, K. (2014). Recurrent models of visual attention. *Advances in Neural Information Processing Systems, 3,* 2204–2212.
10. Mohammed, A. Q., Li, F. (2016). Wiger: Wifi-based gesture recognition system. *International Journal of Geo-Information, 5(6),* 92. DOI 10.3390/ijgi5060092.
11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L. et al. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NY, USA: Curran Associates Inc.
12. Hu, J., Shen, L., Sun, G. (2018). Squeeze-and-excitation networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. DOI 10.1109/CVPR.2018.00745.
13. Jie, W., Xiao, Z., Gao, Q., Hao, Y., Wang, H. (2017). Device-free wireless localization and activity recognition: A deep learning approach. *IEEE Transactions on Vehicular Technology, 66(7),* 6258–6267. DOI 10.1109/TVT.2016.2635161.
14. Zagoruyko, S., Komodakis, N. (2016). Wide residual networks. *Proceedings of the British Machine Vision Conference*, UK: BMVA Press. DOI 10.5244/C.30.87.
15. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. et al. (2015). Going deeper with convolutions. *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA. DOI 10.1109/CVPR.2015.7298594.
16. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *Computer Vision and Pattern Recognition,* 2818–2826. DOI 10.1109/CVPR.2016.308.
17. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, San Francisco, California, USA: AAAI Press.
18. Roy, S. K., Dubey, S. R., Chatterjee, S., Baran Chaudhuri, B. (2020). Fusenet: Fused squeeze-and-excitation network for spectral-spatial hyperspectral image classification. *IET Image Processing, 14(8),* 1653–1661. DOI 10.1049/iet-ipr.2019.1462.
19. Johnstone, D. I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika, 81(3),* 425–455. DOI 10.1093/biomet/81.3.425.

20. Donoho, D. L. (2002). De-noising by soft-thresholding. *IEEE Transactions on Information Theory, 41(3),* 613–627. DOI 10.1109/18.382009.
21. Wright, S. J., Nowak, R. D., Figueiredo, M. A. T. (2008). Sparse reconstruction by separable approximation. *IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, NV, USA. DOI 10.1109/ICASSP.2008.4518374.
22. Zhao, M., Zhong, S., Fu, X., Tang, B., Pecht, M. (2020). Deep residual shrinkage networks for fault diagnosis. *IEEE Transactions on Industrial Informatics, 16(7),* 4681–4690. DOI 10.1109/TII.9424.
23. Isogawa, K., Ida, T., Shiodera, T., Takeguchi, T. (2018). Deep shrinkage convolutional neural network for adaptive noise reduction. *IEEE Signal Processing Letters, 25(2),* 224–228. DOI 10.1109/LSP.2017.2782270.
24. Bengio, Y., Simard, P., Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks, 5(2),* 157–166. DOI 10.1109/TNN.72.
25. Glorot, X., Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of Machine Learning Research*, vol. 9. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Chia Laguna Resort, Sardinia, Italy: PMLR.
26. Ioffe, S., Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Proceedings of Machine Learning Research*, vol. 37. *Proceedings of the 32nd International Conference on Machine Learning*. Lille, France: PMLR.
27. He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*.
28. Ren, S., He, K., Girshick, R., Sun, J. (2017). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6),* 1137–1149. DOI 10.1109/TPAMI.2016.2577031.
29. He, K., Zhang, X., Ren, S., Sun, J. (2016). Identity mappings in deep residual networks. *European Conference on Computer Vision*, vol. 9908, pp. 630–645. USA: Springer International Publishing. DOI 10.1007/978-3-319-46493-0.