



## Article

# Transfer-Learning-Based Human Activity Recognition Using Antenna Array

Kun Ye <sup>1,2,3</sup>, Sheng Wu <sup>1,2,3</sup>, Yongbin Cai <sup>1,2,3</sup>, Lang Zhou <sup>1,3,4</sup>, Lijun Xiao <sup>1,2,3</sup>, Xuebo Zhang <sup>3,5,\*</sup> , Zheng Zheng <sup>1,2,3</sup> and Jiaqing Lin <sup>1,2,3</sup>

<sup>1</sup> Shenzhen Research Institute, Xiamen University, Shenzhen 518000, China; yekun@stu.xmu.edu.cn (K.Y.); wusheng@stu.xmu.edu.cn (S.W.); caiyongbin@stu.xmu.edu.cn (Y.C.); zhoulang@stu.xmu.edu.cn (L.Z.); 23320191153346@stu.xmu.edu.cn (L.X.); 23320211154270@stu.xmu.edu.cn (Z.Z.); jqite@xmu.edu.cn (J.L.)

<sup>2</sup> School of Informatics, Xiamen University, Xiamen 361005, China

<sup>3</sup> Key Laboratory of Southeast Coast Marine Information Intelligent Perception and Application, Ministry of Natural Resources, Xiamen 363000, China

<sup>4</sup> School of Electronic Science and Engineering (National Model Microelectronics College), Xiamen University, Xiamen 361005, China

<sup>5</sup> Whale Wave Technology Inc., Kunming 650200, China

\* Correspondence: xby\_zhang@nwnu.edu.cn

**Abstract:** Due to its low cost and privacy protection, Channel-State-Information (CSI)-based activity detection has gained interest recently. However, to achieve high accuracy, which is challenging in practice, a significant number of training samples are required. To address the issues of the small sample size and cross-scenario in neural network training, this paper proposes a WiFi human activity-recognition system based on transfer learning using an antenna array: Wi-AR. First, the Intel5300 network card collects CSI signal measurements through an antenna array and processes them with a low-pass filter to reduce noise. Then, a threshold-based sliding window method is applied to extract the signal of independent activities, which is further transformed into time–frequency diagrams. Finally, the produced diagrams are used as input to a pretrained ResNet18 to recognize human activities. The proposed Wi-AR was evaluated using a dataset collected in three different room layouts. The testing results showed that the suggested Wi-AR recognizes human activities with a consistent accuracy of about 94%, outperforming the other conventional convolutional neural network approach.

**Keywords:** activity recognition; WiFi sensing; transfer learning; CSI; ResNet18



**Citation:** Ye, K.; Wu, S.; Cai, Y.; Zhou, L.; Xiao, L.; Zhang, X.; Zheng, Z.; Lin, J. Transfer-Learning-Based Human Activity Recognition Using Antenna Array. *Remote Sens.* **2024**, *16*, 845. <https://doi.org/10.3390/rs16050845>

Academic Editors: Prasad S. Thenkabail, Gerardo Di Martino, Jiahua Zhu, Xinbo Li, Shengchun Piao, Junyuan Guo, Wei Guo, Xiaotao Huang and Jianguo Liu

Received: 15 December 2023

Revised: 15 February 2024

Accepted: 21 February 2024

Published: 28 February 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

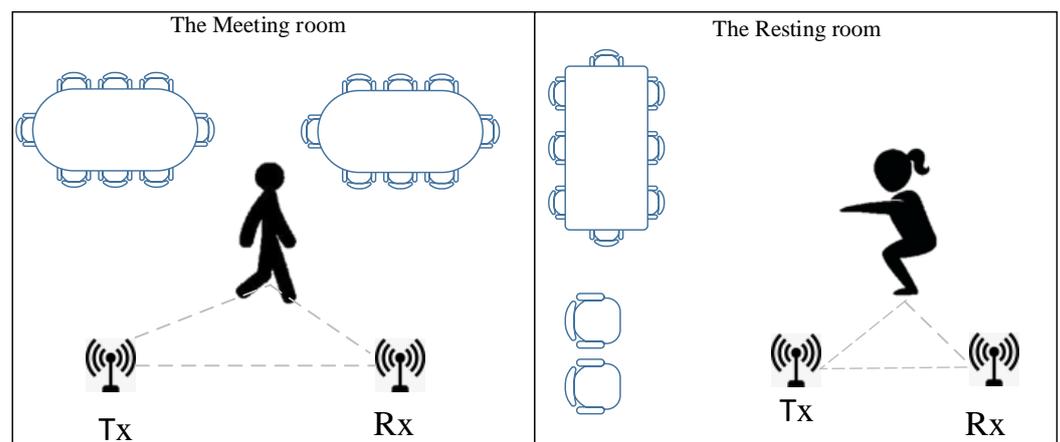
## 1. Introduction

A variety of scenarios have drawn intense attention to human activity recognition (HAR), including health monitoring [1], smart homes [2], and fall detection [3]. In general, traditional HAR systems are based on wearable devices [4–6] or cameras [7,8]. However, in camera-based systems, users run the risk of compromising their privacy.

Due to the low cost and low equipment requirement of Received Signal Strength Indication (RSSI) and Channel State Information (CSI), the RSSI and CSI from commercial WiFi antenna array devices have become widely used in activity recognition. For example, PAWS [9] and WiFinger [10] are RSSI-based [11] methods, whose recognition accuracy is relatively low due to the limited perception performance of the RSSI. WiFall [12], CARM [13], and TensorBeat [14] are CSI-based methods that have higher accuracy and data resolution than the RSSI-based methods. The CSI-based works have been used widely in WiFi sensing, such as gestures [15], gait [16], and breath rate [17], which is also considered in the activity-recognition model proposed in this paper.

Due to the ubiquity of WiFi signals, CSI-based activity recognition utilizes only the wireless communication function and does not require any physical sensors, which provides a great improvement in the security and protection performance of privacy. CSI-based

activity-recognition schemes are often composed of four steps, i.e., data processing, activity classification, feature extraction, and activity detection [12,18]. The corresponding works use primitive signal features, which carry much information caused by human activity and the environment changing with different layouts. At the same time, the CSI extracted by the antenna array is affected by environmental changes, which results in different impacts on the wireless link by human behavior in different scenarios, shown in Figure 1, where two people are performing walking and squatting in two different rooms. It is called the cross-scene problem. The cross-scene problem refers to the ability of the model to generalize across different environments or scenarios and to handle the transition from one scenario to another, which also means that the system needs to be adaptable to recognize activities in multiple scenarios. The multipath channel caused by a specific activity varies with changing environment deployment in different scenes [19,20]. Current works have applied machine learning to solve this problem [21–24]. A hybrid image dataset (ADORESet) has been proposed, which combines real and synthetic images to improve object recognition in robotics, bridging the gap between real and simulation environments [25]. The possibility of connecting object visual recognition with physical attributes such as weight and center of gravity has been explored to improve object manipulation performance via deep neural networks [26]. However, much of this work relies on many training samples to improve accuracy, which is unrealistic when collecting data in reality. The experimental environments are all single scenes, which cannot verify the generalization ability of the models. Therefore, more flexible methods need to be developed for CSI-based human activity recognition with small samples and across scenes.



**Figure 1.** Cross-scene activity recognition.

Addressing the problem of recognizing human activity in cross-scenario and small sample environments, this paper proposes a transfer-learning-based activity-recognition system using an antenna array: Wi-AR. The proposed structure uses the pretrained network to reduce the system's computational complexity instead of training it from scratch [27], which avoids the problem of overfitting. In the method we propose, the original CSI data collected through the antenna array are first processed by a low-pass filter for noise reduction. The purpose of the threshold-based sliding window technique is to determine the beginning and conclusion of activity in a protracted signal. We can then extract the valid segment of activity from the CSI data. Time–frequency diagrams are then created by applying the short-time Fourier transform (STFT) on the four segmented datasets. Finally, the time–frequency diagrams are fed into the pretrained ResNet18 for identification and classification. Based on the simulation results, it is possible to reach 94.2% precision with the proposed Wi-AR system, which is superior to other convolutional neural network (CNN) models. This paper contributes the following:

- (1) This paper proposes a low-cost, non-intrusive human activity-recognition system called Wi-AR, which uses antenna arrays to detect WiFi signals without the need for any devices.
- (2) An activity feature extraction algorithm is proposed to perform the feature segmentation of different activities to detect start and end moments in noisy environments. Using a threshold-based sliding window approach, activity periods can be extracted from CSI data more efficiently.
- (3) The transfer learning strategy employs the fine-tuning of the CNN for training small samples in a changing environment, which improves the accuracy and robustness of activity recognition and avoids overfitting during the training process.

This paper's remaining sections are arranged as follows: Section 2 presents the proposed scheme and the preliminaries. The data collection and preprocessing are investigated in Section 3. In Section 4, the activity-recognition model is proposed. Section 5 shows the experimental validation results, and the last section concludes the paper.

## 2. Proposed Scheme and Preliminaries

### 2.1. Proposed Scheme

Wi-AR uses WiFi devices to recognize human movement without the need for a device. As shown in Figure 2, CSI is collected and then processed to reduce the noise. There are four steps in Wi-AR, among which two steps are essential, i.e., activity segmentation and activity recognition. More specifically, the whole Wi-AR system framework can be concluded as follows:

**CSI Collection:** Wi-AR system collects CSI by Intel 5300 network interface card and WiFi devices.

**Data Preprocessing:** In terms of amplitude information, the Butterworth filter is chosen to reduce the noise.

**Activity Segmentation:** Utilizing the processed CSI series, Wi-AR divides the whole CSI series into four segments, which represent four different activities. To determine the beginning and conclusion of each action, Wi-AR uses a sliding window technique based on thresholds.

**Activity Recognition:** The time–frequency diagram generated by STFT is used for classification. For activity recognition, Wi-AR uses a deep convolutional neural network that has been trained using ResNet18.

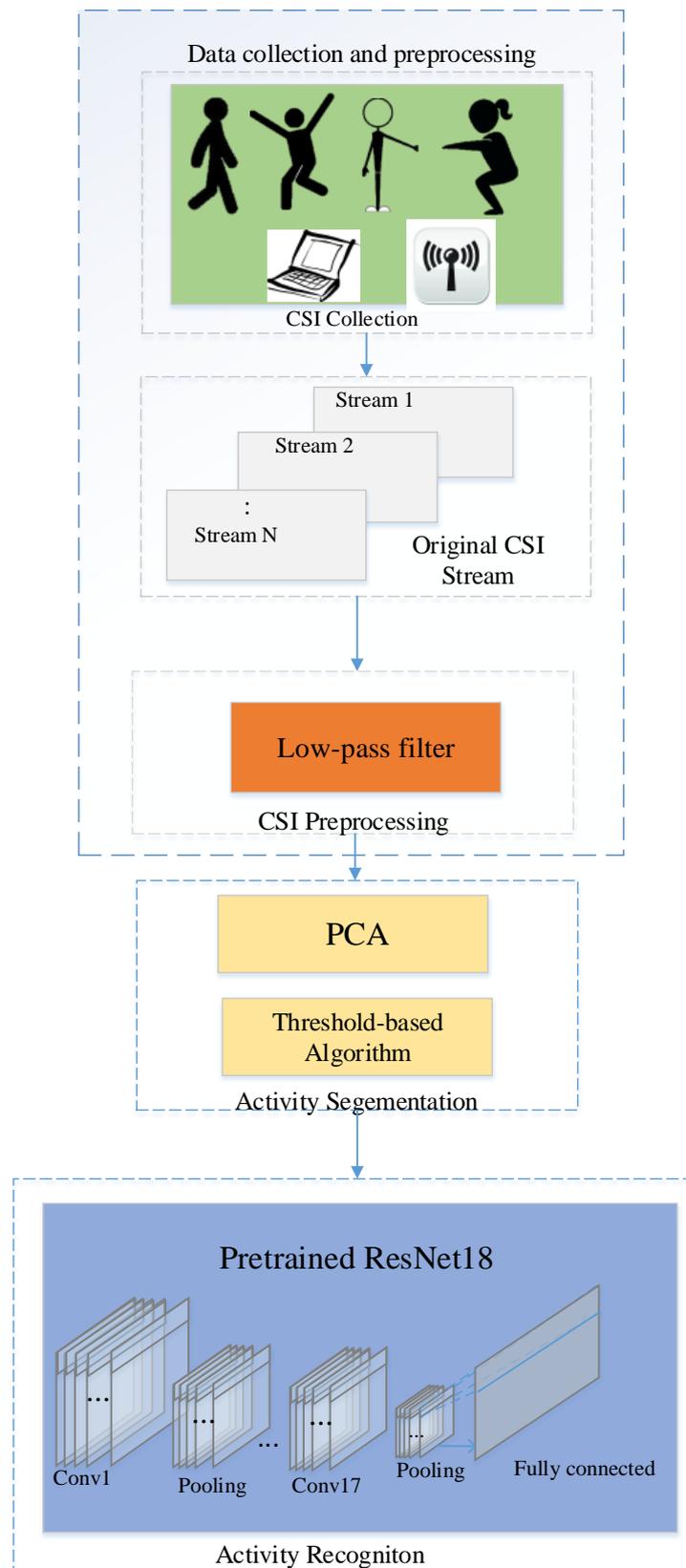
### 2.2. Preliminary

(1) *Channel State Information:* The proposed Wi-AR exploits commercial WiFi devices to obtain CSI. In 802.11n, each multiple-input multiple-output (MIMO) link comprising multiple subcarriers uses orthogonal frequency division multiplexing (OFDM) technology. Each link has a unique channel frequency response caused by the CSI. Due to the 802.11n protocol, the WiFi network has 56 OFDM subcarriers in a 20 MHz band. Using the development tools of the Intel5300 network interface card [28], we can obtain the CSI from 30 subcarriers of the antenna array. Let the number of transmitter and reception antennae be denoted by  $N_t$  and  $N_r$ . If  $X_i$  denotes the transmit signal vector of each packet  $i$  and  $Y_i$  denotes the receive signal vector of each packet  $i$ , the received signal of the network interface card can be represented as:

$$Y_i = H_i X_i + N_i, \quad i \in [1, N] \quad (1)$$

where  $N_i$  is the white Gaussian noise vector,  $H_i$  is the CSI channel matrix, and  $N$  is the total number of received packets. Consequently, for every communication link, the total 30 subcarriers can be obtained,  $H = [H_1, H_2 \dots H_{30}]$  is an expression for the CSI channel matrix, and the total  $N_T \times N_R \times 30$  CSI values are finally obtained. The CSI value for each subcarrier, including amplitude and phase information, is denoted by  $H_i$  in Equation (1)

and can be expressed as  $H_i = \|H_i\|e^{j(2\pi f_i + \theta_i)}$ , where the magnitude frequency and phase of the  $i$ -th subcarrier are represented by  $\|H_i\|$ ,  $f_i$ , and  $\theta_i$ .



**Figure 2.** System overview.

(2) *Reflection of Activity on CSI*: In a WiFi system, the CSI is sensitive to environmental changes, and thus, it can be used to describe the channel frequency response (CFR). Assume  $H(f, t)$  characterizes the CFR. It can be described as follows:

$$H(f, t) = \sum_{k=1}^K \alpha_k(t) e^{-j2\pi f \tau_k(t)}, \quad (2)$$

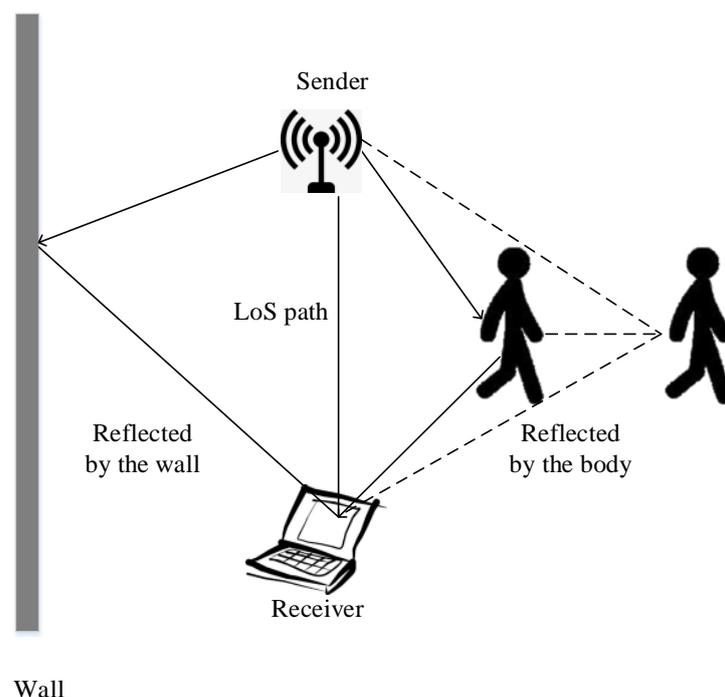
where  $e^{-j2\pi f \tau_k(t)}$  is the phase difference given by carrier frequency shift,  $\alpha_k(t)$  is the channel attenuation, and  $K$  is the total of the multipath numbers. Human activity can impact the WiFi signal on many pathways between the WiFi transmitter and the WiFi receiver. Let  $\lambda$  and  $d(t)$  represent the signal length and the change in the reflection path length, respectively, to better explain the link between human activity and the variations in the WiFi signal. Given that  $f_D = -\frac{1}{\lambda} \cdot \frac{d}{dt} d(t)$  is the frequency shift, we obtain:

$$H(f, t) = e^{-2\pi \Delta f t} (H_s(f) + \sum_{k \in P_d} \alpha_k(t) e^{j2\pi \int_{-\infty}^t f D_k(u) du}), \quad (3)$$

where  $P_d$  denotes the dynamic pathways,  $H_s(f)$  represents the total CFR of the static paths, and  $\Delta f$  is the carrier frequency offset (CFO). Preprocessing can be used to filter out the high-frequency components in static response, as the CFR power fluctuates mostly due to human activity. Because of multipath effects, the value of CFR fluctuates with dynamic components, which may be used to detect human activity. Based on the Friis free space propagation equation [29], as illustrated in Figure 3, the power of receiver is defined as:

$$P_d = \frac{p_t G_t G_r \lambda^2}{(4\pi)^2 (d + 4h + \Delta)^2}, \quad (4)$$

where  $d$  is the distance between the transceiver pair and  $\lambda$  is the signal wavelength. The WiFi transmitter and reception powers are denoted by the variables  $P_t$  and  $P_r$ , respectively. The transmitter and receiver gains are  $G_t$  and  $G_r$ , while the vertical distance is indicated by  $h$ . The reflection path's length is  $\Delta$ . The receiving power varies with the distance between the transceiver pair, as shown by Equation (4). As a result, the shift in CSI may be used to identify human activities.



**Figure 3.** WiFi signal reflection scenario.

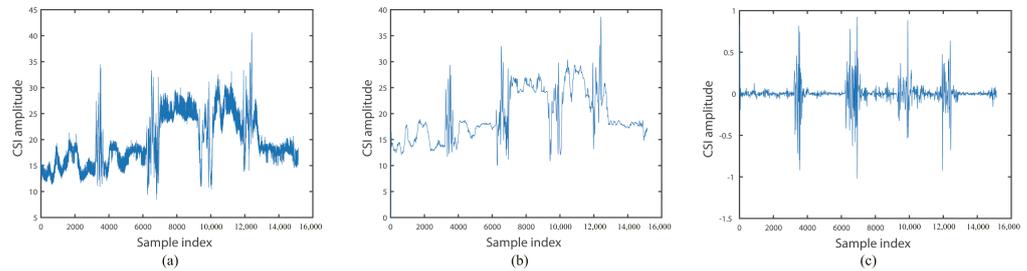
### 3. Data Collection and Preprocessing

#### 3.1. CSI Collection and Denoising

The proposed Wi-AR collects the CSI with  $N_t$  transmitting antennae,  $N_r$  receiving antennae, and 30 CSI subcarriers reported by Intel5300 network interface card, and it can obtain  $30 \times N_t \times N_r$  CSI streams for all communication links. First, a Butterworth low-pass filter is used to reduce noise and remove the high frequency. In our experiments, we designed the Butterworth filter as a low-pass filter, in which case the noise is usually considered to be a high-frequency signal, while the human activity signal that we wish to preserve is a low-frequency signal. Specifically, the frequency response of the Butterworth filter is smooth without abrupt jumps, which gives it an advantage in processing human activity signals. Its main advantage is that it has the flattest frequency response within the pass bandwidth, which means that it processes all frequencies consistently within this range. The original CSI stream and the denoised CSI stream are depicted in Figures 4a and 4b, respectively. It can be observed that CSI generated by human activity is covered by noise. Since the Butterworth filter can maximize the passband flatness of the filter and reduce the high-frequency noise, it is exploited before activity segmentation. The expression for the Butterworth low-pass filter is

$$|\mathbb{H}(\omega)|^2 = \frac{1}{1 + \left(\frac{\omega}{\omega_c}\right)^{2n}} = \frac{1}{1 + \epsilon^2 \left(\frac{\omega}{\omega_p}\right)^{2n}}, \quad (5)$$

where the filter order is denoted by  $n$ , the cut-off frequency is  $\omega_c$ , the passband edge frequency is  $\omega_p$ , and the value of  $|\mathbb{H}(\omega)|^2$  at the passband edge is  $1 + \epsilon^2$ .



**Figure 4.** CSI signal preprocessing. (a) The original CSI signal. (b) The low-pass filtering signal. (c) The first-order difference signal.

While the high-frequency noise can be successfully reduced by the signal following the Butterworth low-pass filter, it cannot reflect the characteristic change in the signal from the waveform. To fully reflect the amplitude information of the low-pass filtering signal, it is necessary to process the first-order difference of the signal. Figure 4b,c represents the filtered CSI signal and the first-order difference CSI signal, respectively. It can be seen that the signal after the first-order difference can reflect the signal characteristics of the four behaviors, providing favorable conditions for the activity segmentation. The definition of the first-order difference can be expressed as  $y(m) = x(m) - x(m - 1)$ , where  $x(m)$  represents the CSI value corresponding to the  $m$ -th sample index.

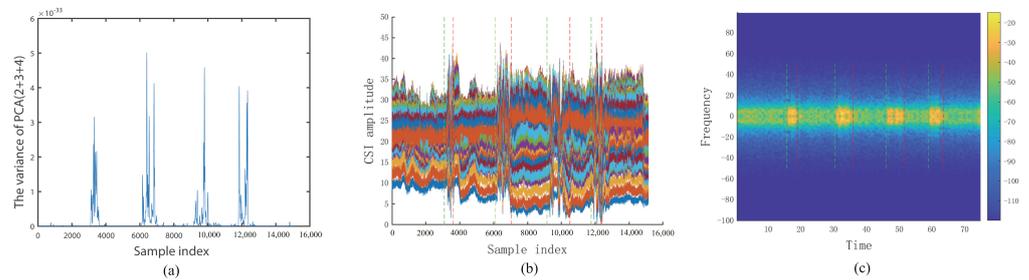
#### 3.2. Activity Segmentation Based on Domain Adaptation

The purpose of activity segmentation is to truncate the start and end moments of human activity from a long signal to extract the complete signal containing the whole behavior. In this paper, to improve the robustness of the segmentation algorithm according to different room layouts, a threshold-based sliding window method is adopted, shown in Algorithm 1. First, principal component analysis (PCA) is used to extract features, and then several components of PCA are chosen to calculate variance. The first PCA component is not used as it contains very little useful information [13]. Second, the moving variance of the total of the aforementioned primary components is computed using a

sliding window. PCA can transform high-dimensional data into low-dimensional data while retaining as much variation information as possible from the original data. This allows us to reduce computational complexity while still retaining the main features of the data. At the same time, PCA can effectively remove noise by retaining the main components of the data. Moving variance can adapt to changes in the data in real time, and due to the adjustability of the window size, this makes it very flexible when dealing with different datasets. The variance of PCA components is shown in Figure 5a. Moving variance depicts the difference of packets reflected by the activity. The variance of PCA is defined as follows:

$$\sigma_i^2 = \frac{\sum_1^m (x_{i+j-1} - \bar{x})^2}{m}, \quad (i = 1, 2, \dots, n - m) \quad (6)$$

where  $\bar{x}$  is the mean value of samples and  $m$  represents step size. Then, the median of the variance is calculated by the sliding window, which also reflects the changing trend. The threshold is calculated according to the data in the priority stationary environment. In general, One-tenth of the maximum value of the data at static time is used as the standard. Large threshold standards are chosen to avoid the effects of static mutations, and compensation needs to be made before the beginning and end of the behavior. For this purpose, the compensation number is set to half of the sampling rate, i.e., 100 sampling points. Therefore, the real start and the end are  $sta - 100$  and  $fin + 100$ . In this way, we finally obtain the result of activity segmentation in the original CSI amplitude, shown in Figure 5b. The figure illustrates that the green dotted line and the red dotted line, respectively, represent the beginning and the end of the activity, and the segmented signal contains complete activities. To be suitable for different scenes, different thresholds are selected. The activity extraction results in a time–frequency domain are shown in Figure 5c. The relevant content of the time–frequency domain diagram will be introduced in Section 3.3.



**Figure 5.** Activity segmentation. (a) The variance of principal component sum. (b) Activity segmentation on original CSI. (c) Time–frequency feature segmentation diagram.

---

#### Algorithm 1 Activity segmentation algorithm.

---

**input:** The amplitude  $\alpha(f, t)$ ;  
 The length of variance window and stride  $w_1, s_1$ ;  
 The length of the median window  $w_2, s_2$ ;  
 The Minimum interval between two actions;  
 The variation between the maximum and minimum values of the stationary environment;

**output:** The start and finish time index  $sta, fin$ ;

- 1:  $[score] = pca(\alpha)$
- 2:  $pca(a) = score(:, 2) + score(:, 3) + score(:, 4)$
- 3:  $n = 1$ ;
- 4: **for**  $i = 1 : w_1 : length(pcadata) - w_1$  **do**
- 5:      $pcavar(n) = var(pcadata(i : i - 1 + w_1))$ ;
- 6:      $n = n + 1$ ;
- 7: **end**
- 8:  $sta = [], fin = []$ ;

---

**Algorithm 1 Cont.**


---

```

9:  $n = 1$ ;
10: for  $ii = 1 : s_2 : \text{length}(pcavar) - s_2$  do
11:    $index(n) = \text{median}(pcavar(ii : ii + s_2))$ ;
12:    $temp = 1 + (ii - 1) \times w_1$ ;
13:   if [ $\text{length}(sta) > \text{length}(fin)$ ] and
       [ $index(n) > \text{threshold}$ ] then
14:      $sta = [sta, temp]$ 
15:     if  $\text{length}(sta) > \text{length}(fin)$  and
          $index(n) < \text{threshold}$  then
16:       if  $temp - sta(\text{end}) \geq \text{minint}$  then
17:          $fin = [fin, temp]$ ;
18:          $n = n + 1$ 
19:       end
20:     end
21:   end
22: end for
23:  $fin = fin + 100$ ;
24:  $sta = sta - 100$ ;

```

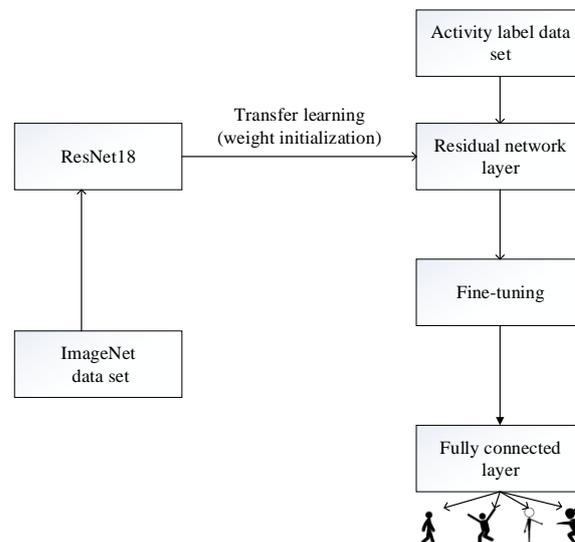
---

**3.3. STFT Transform**

Wi-AR converts the waveforms to time–frequency diagrams using the STFT to extract the signal’s combined time–frequency properties. Considering the trade-off of the frequency-time resolution, Wi-AR sets the sliding window step size of 256 samples in this paper.

**4. Activity Recognition Model**

The proposed Wi-AR adopts the ResNet18 trained by ImageNet as a classification network combined with transfer learning. To avoid losing the source weights, the classifier is first trained using the initial parameters to train all the network’s weights at a low learning rate. Then, the last fully connected layer is modified to suit the target dataset. More specifically, the learned features and weights in the pretraining process are transferred to the recognition network of human activity. After that, the time-frequency diagrams of the CSI are input into the pretrained ResNet18 to train the recognition model. In Figure 6, the layer with a complete connection number is finally substituted with the categories of human activity.



**Figure 6.** Model flow chart.

#### 4.1. ResNet18

In deep learning, the complexity of traditional CNN increases with layers. It is suggested to use a deep residual network (ResNet) to address this situation [30]. Compared with traditional CNN, ResNet is easier to train and has a faster convergence since the whole model only needs to pay attention to the difference between input and output. Moreover, the increase in depth does not increase the amount of computation but increases the accuracy and the efficiency of network training. ResNet combines a deep convolutional neural network with a specially designed residual structure, which can achieve an intense network [31]. To obtain the deep features in the image, which can have a better characterization of the human activity, and considering the depth and calculation of the network, the ResNet18 is finally chosen as the activity classification model, which includes 17 convolutional layers and one full connection layer. The learning process is simplified because ResNet mainly learns residual rather than the complete output. The convolution operation computation is described as

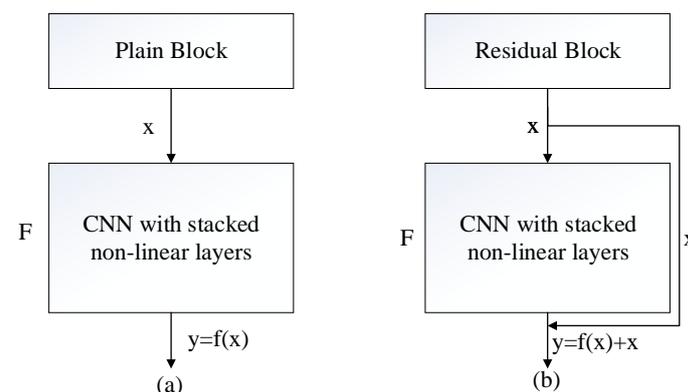
$$X_j^L = f\left(\sum_{i \in M_j} X_i^L \times K_{i,j}^L + b_j^L\right), \quad (7)$$

where  $M_j$  is the input feature map,  $L$  denotes the number of layers in the neural network, and  $K_{i,j}^L$  suggests the convolution kernel. The activation function is  $f$ , and the unique offset  $b$  is output for each layer of the feature graph.

To extract more features, the number of convolutional layers increases. However, with an increasing number of convolutional layers, there is a risk of gradient dispersion and gradient explosion. The residual unit in ResNet18 can effectively solve the above problem. The core idea is to divide network output into two parts: identity mapping and residual mapping. The definition of the residual unit is

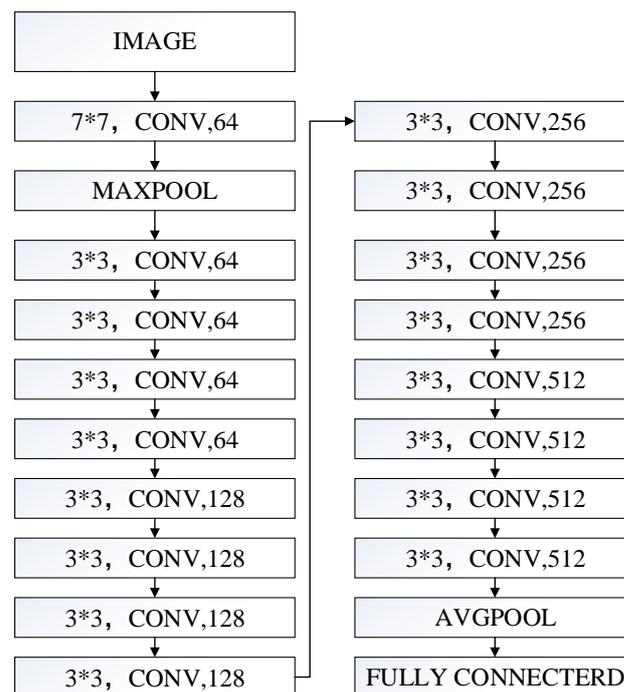
$$X_{k+1} = f(F(X_k, W_k) + h(X_k)), \quad (8)$$

where  $X_k$  and  $X_{k+1}$  stand for the input and output of the  $k$ -th residual unit, respectively.  $F(X_k, W_k)$  is the residual mapping that must be learned. The activation function is denoted by  $f$  and the convolution kernel by  $W$ . Figure 7a displays the CNN learning block with less stacked non-linear layers through a direct mapping  $x \rightarrow y$  representing  $F(x)$  and  $x$  as stacked non-linear layers and the identity function, respectively. Figure 7b shows the identity mapping through the residual function  $F(x)$ , where  $y = F(x) + x$ , as proposed in [32].



**Figure 7.** (a) Direct mapping in plain CNN (b) Identity mapping in ResNet.

In ResNet18, both the max-pooling and the average-pooling techniques are used. Reducing training parameters in the network is the aim of the max-pooling layer, which comes after the convolutional layer [33]. The last completely linked layer, which consists of four nodes symbolizing four distinct activities, is added to receive the output of the human activity. Figure 8 illustrates this process.



**Figure 8.** ResNet18 network structure.

#### 4.2. Transfer Learning

One common deep learning technique that is frequently applied to tiny training samples is transfer learning. To accomplish the migration, model-based transfer learning is used [34], which looks for pretrained parameter weights that the network's bottom layer may share. After obtaining the pretrained model with the ImageNet dataset, we replace the random initialization model parameters with the new parameters except for the last fully connected layer. Since each CNN in the ImageNet dataset is trained using 1000 classes, and our activities include four classes, fine-tuning is used to make sure that the network weights will not change too quickly and will fit our data without altering the original model. Therefore, a small learning rate is set to avoid gradient vanishing.

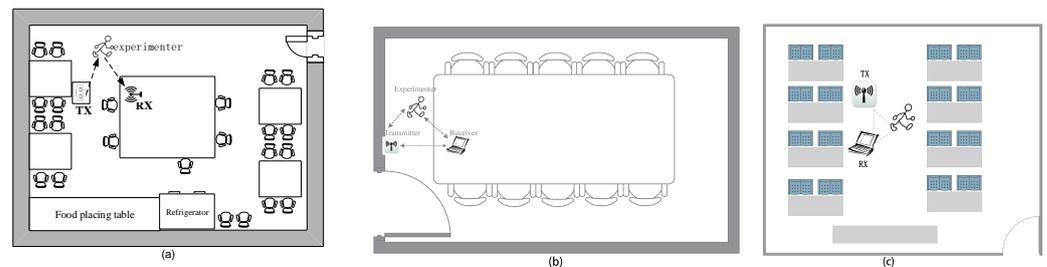
### 5. Experimental Evaluation

To show the model's capacity for generalization, the performance of the suggested Wi-AR is first assessed on a dataset that the user has collected, and then it is tested on additional datasets. We also contrast our model's accuracy with that of other CNN models in this section.

#### 5.1. Experiment Setup

In the experiment, we collected CSI from various rooms using a computer equipped with an Intel 5300 network interface card and a commercial WiFi router. The router is equipped with three antennae, and the reason for using three antennae is mainly related to the working principle of the MIMO (Multiple-Input Multiple-Output) system. In a MIMO system, multiple antennae can send and receive multiple streams of data at the same time, therefore increasing the transmission rate and reliability of the system. Specifically, three antennae can form three independent antenna links, each of which can receive an independent data stream. These three links form 90 subcarriers, i.e., for each timestamped data, they are composed of 90 subcarriers. In this way, by analyzing the CSI of each subcarrier, we can obtain more detailed and accurate information about the channel state and thus better understand and utilize the wireless channel. In addition, multiple antennae can provide more spatial diversity and spatial multiplexing, thus improving the capacity and anti-interference capability of the system. Spatial diversity improves the reception

quality of the signal, while spatial multiplexing increases the system's data transmission rate. The router is equipped with three antennae, forming an antenna array. The room is furnished with tables and chairs, as shown in Figure 9. The data gathered in the meeting and rest areas serve as the training set for the experiment, while the data gathered in three separate rooms serves as the testing set. A total of 120 pieces of data were randomly chosen from the three scenarios, and 240 samples were chosen from each of the first two scenarios to represent the model's generalization. As a result, there are 480 samples in the training set and 120 samples in the testing set. Three antennae are installed on the WiFi transmitter's receiving side and two on the transmitter side. The sampling rate is 200 packets/s, and 30 subcarriers from each transceiver pair are obtained. At the same time, three volunteers with different body shapes are asked to perform four kinds of activities. Four behaviors contain two coarse-grained activities (jumping and walking) and two relatively fine-grained activities (squatting and leg lifting). The volunteers are asked to perform each activity one by one and keep them for five seconds. Each task has a total of 150 samples split into a test set and a training set. If the generated time–frequency diagrams are not normalized, the ranges of feature value distribution will vary differently. To avoid such a problem, Wi-AR first normalizes all time–frequency diagrams and further resizes them to suit the pretrained model. Lastly, the amount of human activity changes the final completely connected layer.



**Figure 9.** Different room layout for data collection. (a) The rest-room layout. (b) The meeting room layout. (c) The class-room layout.

Before the experiment, several initial parameters need to be defined, which are listed in Table 1. The batch size is set to 16, and the iteration value is 30. In this research, a reduced starting learning rate of 0.001 is chosen to prevent overfitting.

**Table 1.** Training parameters.

| Parameter             | Value     |
|-----------------------|-----------|
| Image size            | 224 × 224 |
| Epoch                 | 30        |
| Batch size            | 16        |
| Initial learning rate | 0.001     |
| Lr-function           | StepLR    |

### 5.2. Experimental Validation

The proposed Wi-AR produces time-frequency diagrams by STFT. Figure 10 displays the time–frequency diagrams for four different types of activities.

The scene's computers are equipped with an NVIDIA GeForce GTX 1660S GPU, an Intel 10500 CPU, and 32 GB of RAM. Then, use the cross-entropy loss function. If the predicted value is the same as the true value, it approaches 0. If the predicted value is different from the true value, the cross-entropy loss function will become very large. Training and testing curves are, respectively, recorded in Figures 11 and 12, which represent the value changes in accuracy and loss in the training and testing process. Both can converge after 20 iterations. In addition, a detailed evaluation of the classification result

is conducted using three types of metrics. The metrics are defined as follows: (1) Precision is defined as  $\frac{TP}{TP + FP}$ , where the ratios of correctly marked activities ( $TP$ ) and erroneously marked activities ( $FP$ ) are expressed. (2)  $\frac{TP}{TP + FN}$  is the definition of recall, where  $FN$  represents erroneously promoted negative samples. (3) The formula for the F1-score (F1) is  $F1 = \frac{2 \times PR \times RE}{PR + RE}$ . Table 2 tabulates the testing data and shows the high precision and F1-score. Figure 13a shows the confusion matrix whose rows represent the predicted activity and columns refer to the actual activity. We find that the accuracy of leg lifting and squatting can achieve 100% due to their obvious features, and the average accuracy is 94.2%. Walking and jumping have similar movements, so the accuracy of recognition is not as good as it is for the other two activities. The dispersion of various activity data is displayed using the ResNet18 model under T-SNE visualization in Figure 13b. It can be intuitively seen that the model can distinguish different activities well.

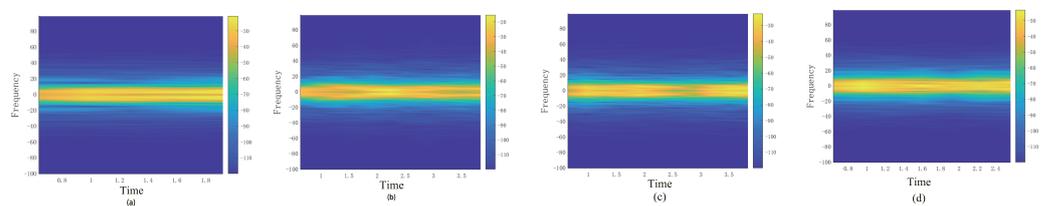


Figure 10. The time-frequency diagrams of four kinds of activities. (a) Jumping. (b) Walking. (c) Squatting. (d) Leg lifting.

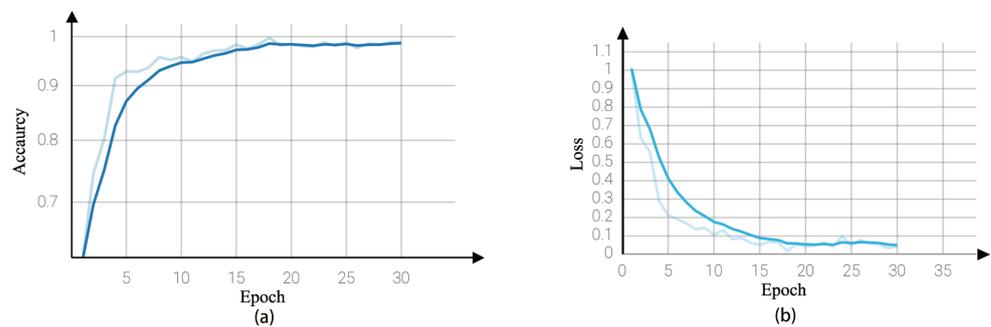


Figure 11. The training g accuracy and loss. (a) The accuracy of training. (b) The training accuracy and loss.

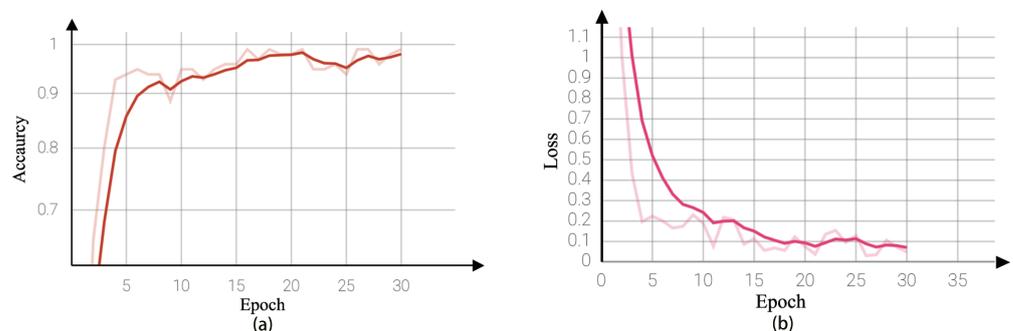
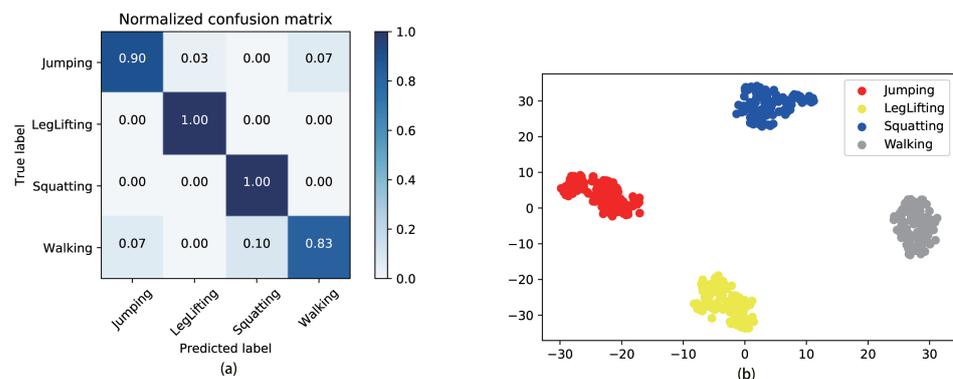


Figure 12. The testing accuracy and loss. (a) The accuracy of testing. (b) The loss of testing.

**Table 2.** Testing Results of the ResNet18.

| Human Activity | Precision | Recall | F1-Score |
|----------------|-----------|--------|----------|
| Jumping        | 0.90      | 0.87   | 0.93     |
| Leg Lifting    | 1.00      | 0.97   | 0.98     |
| Squatting      | 1.00      | 1.00   | 0.97     |
| Walking        | 0.83      | 0.93   | 0.92     |

**Figure 13.** (a) The confusion matrix of human activity recognition. (b) T-SNE visualization after the ResNet18 model for four kinds of activities.

To verify the generalization ability of the recognition model, the proposed Wi-AR is also tested on the dataset collected by the team of Chunjing Xiao [35] for more evaluation. The differences are that their sampling rate is 50 packets/s, and the WiFi router has only one antenna. Ten different types of activities make up the dataset, five of which are fine-grained and five of which are coarse-grained. The experiment is conducted using two relatively fine-grained activities (hand swing and drawing O) and two coarse-grained activities (running and squatting). The result of testing is tabulated in Table 3, from which we can also see high accuracy, achieving more than 94% accuracy among each activity. And since the drawing circle has unique movement characteristics, the recognition accuracy achieves 100% for the given dataset. As a result, the validation results show that the recognition model has the capacity for generalization.

**Table 3.** Testing results of the ResNet18 on another dataset.

| Human Activity | Precision | Recall | F1-Score |
|----------------|-----------|--------|----------|
| Drawing O      | 1.00      | 1.00   | 1.00     |
| Hand Swing     | 0.96      | 0.87   | 0.91     |
| Running        | 0.94      | 0.97   | 0.95     |
| Squatting      | 0.94      | 1.00   | 0.97     |

To compare the accuracy of the pretrained networks, we choose different CNN models, such as Alexnet, VGG11, and ResNet34. Meanwhile, some classical machine learning algorithms, such as decision trees, random forests, SVMs, etc., are also used as comparative tests. The results of the CNN classifiers in terms of accuracy and time consumed are shown in Table 4 to show the performance of the different CNN networks in comparison with the model in this paper. The accuracy of classical machine learning algorithms is shown in Table 5. The experimental results show that the recognition accuracy of classical machine learning algorithms is generally low because the action features carried by CSI signals are significantly reduced after going through the wall, and ordinary machine learning algorithms cannot accurately classify them, and more complex deep networks are needed to extract their features. Among the CNN classifiers, the accuracy of ResNet18 is 0.8% higher than that of VGG19, but the time consumed is about 25% of that of

VGG19. In addition, ResNet18 shows higher accuracy and less time consumption than other ResNet models.

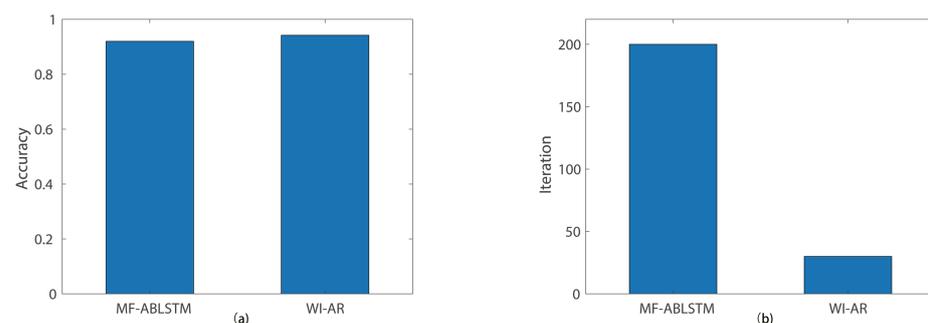
**Table 4.** The accuracy of each CNN classification.

| CNN Models | Accuracy Rate | Time Consumption |
|------------|---------------|------------------|
| AlexNet    | 85.00%        | 4 m 22 s         |
| Vgg11      | 91.67%        | 7 m 33 s         |
| Vgg13      | 92.50%        | 9 m 16 s         |
| Vgg16      | 92.67%        | 10 m 45 s        |
| Vgg19      | 93.33%        | 13 m 10 s        |
| ResNet18   | 94.17%        | 3 m 29 s         |
| ResNet34   | 90.00%        | 5 m 23 s         |
| ResNet50   | 84.17%        | 6 m 43 s         |
| ResNet101  | 86.67%        | 8 m 17 s         |
| ResNet152  | 92.50%        | 10 m 23 s        |

**Table 5.** Accuracy of classical machine learning algorithms

| Machine Learning Algorithms | Accuracy Rate |
|-----------------------------|---------------|
| Naive Bayesian              | 44%           |
| KNN                         | 58%           |
| Decision tree               | 65%           |
| SVM                         | 74%           |
| Proposed                    | 94%           |

Moreover, MF-ABLSTM [36] leverages attention-based long short-term memory neural network and time–frequency domain features for small CSI sample-based activity recognition, achieving 92% with 490 training and testing samples after 200 iterations. Due to the proposed method being able to train very deep neural networks, it avoids the problem of gradient vanishing and improves the model’s expressive power and performance. It uses residual connections to preserve the original features, making the learning of the network smoother and more stable, further improving the accuracy and generalization ability of the model. During training, gradient vanishing and exploding problems can be avoided, accelerating network convergence. Therefore, Wi-AR achieves 94.2% with 600 samples after 30 iterations. The results of the comparison are, respectively, shown in Figure 14a,b, which demonstrate that our proposed Wi-AR scheme achieves higher accuracy with fewer iterations for small sample-based activity recognition.



**Figure 14.** (a) The comparison of activity-recognition accuracy. (b) The comparison of training iterations.

## 6. Conclusions

In response to the small sample size and cross-scenario issues in activity recognition, this paper proposes the Wi-AR human activity-recognition system, which is based on channel state data and antenna arrays. Wi-AR collects CSI data from an array of antennae for a sequence of four kinds of activities, preprocesses the collected CSI signal, transforms it into time–frequency diagrams, and marks samples for supervised machine learning. The experimental results show that this method based on transfer learning can achieve 94% accuracy with a small number of samples. We can see that Wi-AR is relevant in single-person multi-scene environments. In future work, we will consider more realistic multi-user human activity-recognition scene recognition. Meanwhile, it is also a challenging problem to do effective differentiation for some similar actions. For the problem of difficult label annotation of sensing data, semi-supervised learning is also an effective solution to deal with this difficulty, which is also the focus of our future work.

**Author Contributions:** Conceptualization, K.Y., L.X. and X.Z.; methodology, K.Y., L.X. and X.Z.; software, K.Y.; validation, K.Y., Y.C. and Z.Z.; formal analysis, K.Y.; investigation, K.Y.; resources, K.Y.; data curation, Y.C.; writing—original draft preparation, K.Y. and L.X.; writing—review and editing, S.W. and L.Z.; visualization, K.Y.; supervision, S.W.; project administration, J.L.; funding acquisition, J.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Key Program of Marine Economy Development Special Foundation of the Department of Natural Resources of Guangdong Province (GDNRC [2022]19) and the Natural Resources Science and Technology Innovation Project of Fujian KY-080000-04-2022-025.

**Data Availability Statement:** The data presented in this study are available on reasonable request from the corresponding author.

**Conflicts of Interest:** Author Xuebo Zhang was employed by the company Whale Wave Technology Inc. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References

1. Wang, X.; Yang, C.; Mao, S. PhaseBeat: Exploiting CSI phase data for vital sign monitoring with commodity WiFi devices. In Proceedings of the 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS), Atlanta, GA, USA, 5–8 June 2017; pp. 1230–1239.
2. Zou, H.; Zhou, Y.; Yang, J.; Jiang, H.; Xie, L.; Spanos, C.J. WiFi-enabled device-free gesture recognition for smart home automation. In Proceedings of the 2018 IEEE 14th International Conference on Control and Automation (ICCA), Anchorage, AK, USA, 12–15 June 2018; pp. 476–481.
3. Palipana, S.; Rojas, D.; Agrawal, P.; Pesch, D. FallDeFi: Ubiquitous fall detection using commodity WiFi devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2018**, *1*, 155. [\[CrossRef\]](#)
4. Yatani, K.; Truong, K.N. Bodyscope: A wearable acoustic sensor for activity recognition. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing, Pittsburgh, PA, USA, 5–8 September 2012; pp. 341–350.
5. Fortino, G.; Galzarano, S.; Gravina, R.; Li, W. A framework for collaborative computing and multi-sensor data fusion in body sensor networks. *Inf. Fusion* **2015**, *22*, 50–70. [\[CrossRef\]](#)
6. Ghasemzadeh, H.; Panuccio, P.; Trovato, S.; Fortino, G.; Jafari, R. Power-aware activity monitoring using distributed wearable sensors. *IEEE Trans. Hum. Mach. Syst.* **2014**, *44*, 537–544. [\[CrossRef\]](#)
7. Bodor, R.; Jackson, B.; Papanikolopoulos, N. Vision-based human tracking and activity recognition. In Proceedings of the 11th Mediterranean Conference on Control and Automation, Rhodes, Greece, 18–20 June 2003; Volume 1.
8. De Sanctis, M.; Cianca, E.; Di Domenico, S.; Provenziani, D.; Bianchi, G.; Ruggieri, M. Wibecam: Device free human activity recognition through WiFi beacon-enabled camera. In Proceedings of the 2nd Workshop on Workshop on Physical Analytics, Florence, Italy, 22 May 2015; pp. 7–12.
9. Gu, Y.; Ren, F.; Li, J. Paws: Passive human activity recognition based on WiFi ambient signals. *IEEE Internet Things J.* **2015**, *3*, 796–805. [\[CrossRef\]](#)
10. Tan, S.; Yang, J. WiFinger: Leveraging commodity WiFi for fine-grained finger gesture recognition. In Proceedings of the 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing, Paderborn, Germany, 5–8 July 2016; pp. 201–210.
11. Yang, Z.; Zhou, Z.; Liu, Y. From RSSI to CSI: Indoor localization via channel response. *ACM Comput. Surv.* **2013**, *46*, 25. [\[CrossRef\]](#)
12. Chen, Z.; Zhang, L.; Jiang, C.; Cao, Z.; Cui, W. WiFi CSI based passive human activity recognition using attention based BLSTM. *IEEE Trans. Mob. Comput.* **2018**, *18*, 2714–2724. [\[CrossRef\]](#)

13. Wang, W.; Liu, A.X.; Shahzad, M.; Ling, K.; Lu, S. Understanding and modeling of WiFi signal based human activity recognition. In Proceedings of the 21st Annual International Conference on Mobile Computing and Networking, Paris, France, 7–11 September 2015; pp. 65–76.
14. Wang, X.; Yang, C.; Mao, S. TensorBeat: Tensor decomposition for monitoring multiperson breathing beats with commodity WiFi. *ACM Trans. Intell. Syst. Technol.* **2017**, *9*, 8. [[CrossRef](#)]
15. Feng, C.; Arshad, S.; Yu, R.; Liu, Y. Evaluation and improvement of activity detection systems with recurrent neural network. In Proceedings of the 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, USA, 20–24 May 2018; pp. 1–6.
16. Abdelnasser, H.; Youssef, M.; Harras, K.A. WiGest: A ubiquitous WiFi-based gesture recognition system. In Proceedings of the 2015 IEEE Conference on Computer Communications (INFOCOM), Hong Kong, China, 26 April–1 May 2015; pp. 1472–1480.
17. Liu, X.; Cao, J.; Tang, S.; Wen, J. Wi-Sleep: Contactless sleep monitoring via WiFi signals. In Proceedings of the 2014 IEEE Real-Time Systems Symposium, Rome, Italy, 2–5 December 2014; pp. 346–355.
18. Duan, S.; Yu, T.; He, J. WiDriver: Driver activity recognition system based on WiFi CSI. *Int. J. Wirel. Inf. Netw.* **2018**, *25*, 146–156. [[CrossRef](#)]
19. Zhang, Y.; Wang, X.; Wang, Y.; Chen, H. Human Activity Recognition Across Scenes and Categories Based on CSI. *IEEE Trans. Mob. Comput.* **2020**, *21*, 2411–2420. [[CrossRef](#)]
20. Wang, J.; Chen, Y.; Hu, L.; Peng, X.; Philip, S.Y. Stratified transfer learning for cross-domain activity recognition. In Proceedings of the 2018 IEEE International Conference on Pervasive Computing and Communications (PerCom), Athens, Greece, 19–23 March 2018; pp. 1–10.
21. Zheng, V.W.; Hu, D.H.; Yang, Q. Cross-domain activity recognition. In Proceedings of the 11th International Conference on Ubiquitous Computing, Orlando, FL, USA, 30 September–3 October 2009; pp. 61–70.
22. Hu, D.H.; Zheng, V.W.; Yang, Q. Cross-domain activity recognition via transfer learning. *Pervasive Mob. Comput.* **2011**, *7*, 344–358. [[CrossRef](#)]
23. Wang, J.; Zhang, L.; Gao, Q.; Pan, M.; Wang, H. Device-free wireless sensing in complex scenarios using spatial structural information. *IEEE Trans. Wirel. Commun.* **2018**, *17*, 2432–2442. [[CrossRef](#)]
24. Ding, X.; Jiang, T.; Zhong, Y.; Wu, S.; Yang, J.; Xue, W. Improving WiFi-based Human Activity Recognition with Adaptive Initial State via One-shot Learning. In Proceedings of the 2021 IEEE Wireless Communications and Networking Conference (WCNC), Nanjing, China, 29 March–1 April 2021; pp. 1–6.
25. Bayraktar, E.; Yigit, C.B.; Boyraz, P. A hybrid image dataset toward bridging the gap between real and simulation environments for robotics: Annotated desktop objects real and synthetic images dataset: ADORESet. *Mach. Vis. Appl.* **2019**, *30*, 23–40. [[CrossRef](#)]
26. Bayraktar, E.; Yigit, C.B.; Boyraz, P. Object manipulation with a variable-stiffness robotic mechanism using deep neural networks for visual semantics and load estimation. *Neural Comput. Appl.* **2020**, *32*, 9029–9045. [[CrossRef](#)]
27. Liang, H.; Fu, W.; Yi, F. A survey of recent advances in transfer learning. In Proceedings of the 2019 IEEE 19th International Conference on Communication Technology (ICCT), Xi'an, China, 16–19 October 2019; pp. 1516–1523.
28. Halperin, D.; Hu, W.; Sheth, A.; Wetherall, D. Tool release: Gathering 802.11 n traces with channel state information. *ACM SIGCOMM Comput. Commun. Rev.* **2011**, *41*, 53–53. [[CrossRef](#)]
29. Rappaport, T.S. *Wireless Communications—Principles and Practice*, (The Book End). *Microw. J.* **2002**, *45*, 128–129.
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1026–1034.
31. Li, Y.; Ding, Z.; Zhang, C.; Wang, Y.; Chen, J. SAR ship detection based on resnet and transfer learning. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 1188–1191.
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
33. Song, X.; Chen, K.; Cao, Z. ResNet-based Image Classification of Railway Shelling Defect. In Proceedings of the 2020 39th Chinese Control Conference (CCC), Shenyang, China, 27–29 July 2020; pp. 6589–6593.
34. Zhu, Y.; Chen, Y.; Lu, Z.; Pan, S.J.; Xue, G.R.; Yu, Y.; Yang, Q. Heterogeneous transfer learning for image classification. In Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 7–11 August 2011.
35. Xiao, C.; Lei, Y.; Ma, Y.; Zhou, F.; Qin, Z. DeepSeg: Deep-Learning-Based Activity Segmentation Framework for Activity Recognition Using WiFi. *IEEE Internet Things J.* **2020**, *8*, 5669–5681. [[CrossRef](#)]
36. Tian, Y.; Li, S.; Chen, C.; Zhang, Q.; Zhuang, C.; Ding, X. Small CSI Samples-Based Activity Recognition: A Deep Learning Approach Using Multidimensional Features. *Secur. Commun. Netw.* **2021**, *2021*, 5632298. [[CrossRef](#)]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.