electronics

MDPI

Article

Wi-Fi-Based Location-Independent Human Activity Recognition with Attention Mechanism Enhanced Method

Xue Ding [1], Ting Jiang [1], Yi Zhong [2,*], Sheng Wu [1], Jianfei Yang [3] and Jie Zeng [4]

[1] School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China; dxue@bupt.edu.cn (X.D.); tjiang@bupt.edu.cn (T.J.); thuraya@bupt.edu.cn (S.W.)
[2] School of Information and Electronics, Beijing Institute of Technology, Bejing 100081, China
[3] School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore 639798, Singapore; yang0478@e.ntu.edu.sg
[4] Department of Electronic Engineering, Tsinghua University, Beijing 100084, China; zengjie@mail.tsinghua.edu.cn
* Correspondence: yi.zhong@bit.edu.cn

Abstract: Wi-Fi-based human activity recognition is emerging as a crucial supporting technology for various applications. Although great success has been achieved for location-dependent recognition tasks, it depends on adequate data collection, which is particularly laborious and time-consuming, being impractical for actual application scenarios. Therefore, mitigating the adverse impact on performance due to location variations with the restricted data samples is still a challenging issue. In this paper, we provide a location-independent human activity recognition approach. Specifically, aiming to adapt the model well across locations with quite limited samples, we propose a Channel–Time–Subcarrier Attention Mechanism (CTS-AM) enhanced few-shot learning method that fulfills the feature representation and recognition tasks. Consequently, the generalization capability of the model is significantly improved. Extensive experiments show that more than 90% average accuracy for location-independent human activity recognition can be achieved when very few samples are available.

Keywords: human activity recognition; Wi-Fi sensing; few-shot learning; location-independent; Channel–Time–Subcarrier Attention Mechanism (CTS-AM)

Citation: Ding, X.; Jiang, T.; Zhong, Y.; Wu, S.; Yang, J.; Zeng, J. Wi-Fi-Based Location-Independent Human Activity Recognition with Attention Mechanism Enhanced Method. *Electronics* **2022**, *11*, 642. https://doi.org/10.3390/electronics11040642

Academic Editors: Syed Aziz Shah, Qammer Hussain Abbasi, Jawad Ahmad and Muhammad Ali Imran

Received: 21 January 2022
Accepted: 17 February 2022
Published: 18 February 2022

1. Introduction

Human activity recognition (HAR) is an indispensable technique that has been widely used in many applications, such as personalized home automation, health surveillance, security and protection, and entertainment [1,2]. Perhaps the most well-known approaches involving human activity recognition are the wearable-devices-based methods [3,4] and the camera-based methods [5,6]. Although both techniques can effectively classify diverse human activities with a low false-alarm rate, they also expose certain shortcomings. For instance, people have to carry the motion sensors whenever and wherever using the wearable-devices-based method to identify the activity, which is inconvenient even if these sensors are harmless to human health. Although the camera-based approach has a definite advantage in perceived performance, it faces tough challenges in privacy protection, obstruction blocking, darkness conditions, etc.

Device-free sensing (DFS) technology effectively overcomes the above shortcomings by only utilizing radio frequency (RF) signals for sensing without being aided by additional devices carried by people [7–10]. There have been extensive studies on DFS leveraging various RF signals, including Frequency-Modulated Continuous Wave (FMCW) [11], MilliMeter-Wave (MMW) [12], Ultra-Wide Band (UWB) [13–16], Wi-Fi [17–20], etc. Among them, since Wi-Fi networks are nearly ubiquitous in both indoor and outdoor environments,

they could play an irreplaceable role in wireless intelligent sensing. Consequently, the study of Wi-Fi-based HAR has increased dramatically in recent years [21–24].

A majority of the existing human behavior recognition systems pay attention to performance improvement in a single position. To promote the industrialization application of this field, research on cross-location sensing systems is emerging. Since the multipath propagation of the RF signal is affected by the location of obstacles, the same human activity at different positions would result in distinct signals patterns, which will severely harm the generalization ability of the model across diverse locations. This challenging problem can be described as the domain shift of different spots.

There are some solutions to solve the issue. "Location-dependent recognition" represents the identification problems requiring sufficient training data at all locations. Ref. [25] relies on plentiful labeled activity samples for each location to obtain precise accuracy. Although sufficient data samples will reduce the domain shift differences and achieve satisfactory accuracy, collecting and annotating plenty of data at each position is labor-intensive with a poor user experience. In order to alleviate the above problems, "Location-independent recognition", which provides only very few samples to be trained at unseen locations within the range of perception in the single environment, is proposed. Ref. [26] separates the location-related background information from the activity signal, which aims to decrease the number of training locations and samples. Nevertheless, the recognition accuracy is modest and still needs to be improved. As can be seen, in the case of insufficient training data samples, the representation of the model is prone to overfitting, resulting in the performance degradation of the model trained in some locations when testing in other locations.

Based on existing solid foundations, it still needs to be further investigated how to mitigate the domain shift to achieve location-independent sensing efficiently with as few training samples as possible. Specifically, some fundamental issues need to be tackled. The features learned from one location would be extremely difficult to transfer to the other unseen positions. Since most of the feature extraction methods, especially the deep learning methods [27], require a training set and testing set to satisfy the independent and identical distribution (IID), it is hard to work well in our case. Therefore, we must extract the features in a more distinguished way, enabling the model to be sufficiently distinctive among different human activities and robust enough against the location variation. Moreover, ensuring the transferable capability of the sensing model with inadequate samples needs to be explored in detail.

To overcome the challenges mentioned above, we provide a location-independent human activity recognition system called LI-HAR. In this system, a Channel–Time–Subcarrier Attention Mechanism (CTS-AM) enhanced Convolutional Neural Network (CNN) is designed for feature representation, so that distinctive features without location-dependent factors will gain more attention. In addition, to improve the transferable ability of the model when only very few training samples from unseen locations are available, metric-based few-shot learning is utilized to enable the model to generalize among different positions. To demonstrate the performance of the proposed method in terms of accuracy and robustness, we collect data samples, including four prescribed activities performed at 24 different locations in an office environment.

The main contributions of this paper are listed as the following three folds:

- We elaborately analyze the influence of location change on Wi-Fi signal distribution of different activities and observe that the existing attempts cannot realize location-independent human activity recognition well.
- To mitigate the adverse effects of location variations in the case of insufficient training samples, we design a system named LI-HAR by leveraging a few-shot learning approach based on a prototypical network improved via a Channel–Time–Subcarrier–Channel Attention Block (CTSC-AB). Unlike previous attempts, our proposed method can simultaneously extract distinguished representations for different activities and

identify them free from the limitation of locations using only very few samples from new positions.
- We built a dataset to evaluate the performance of the proposed approach. Comprehensive experiments demonstrate that the LI-HAR system can promisingly address the challenges presented.

We organize the rest of this paper as follows: Some preliminaries of Wi-Fi sensing are presented in Section 2. Section 3 provides a system overview and a detailed design. In Section 4, the performance evaluation is conducted. We conclude the paper in Section 5.

2. Preliminary and Motivation

For Wi-Fi-based DFS technology, the transmitted signals are affected by objects such as the dynamic human activities in this paper, leading to the superposition of multipath signals in the receiver, which can be utilized for recognition. The influenced communication link between the transmitter (TX) and the receiver (RX) can be depicted by Channel State Information (CSI), which is fine-grained compared with the received signal strength indicator (RSSI).

In the IEEE Wireless Local Area Network (WLAN) standards (such as 802.11 a/g/b/n/ac/ax), which are supported by most of the Wi-Fi routers, 802.11n and later versions support both Multiple-Input Multiple-Output (MIMO) and Orthogonal Frequency Division Multiplexing (OFDM) technologies, which could provide a higher data rate. Thus, CSI can be extracted from multiple channels.

We take advantage of the CSI of the links to sense human activity. We denote $\boldsymbol{y}$ and $\boldsymbol{x}$ as the received signal and transmitting signal, respectively. Their relationship can be expressed as

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n} \tag{1}$$

where $\boldsymbol{H}$ is the CSI channel matrix and $\boldsymbol{n}$ is the noise vector. In addition, $\boldsymbol{H}$ is a complex matrix including amplitude and phase. Specifically, when the transceiver is equipped with three antennas, the CSI measurement at frame t can be described as

$$H(t) = H_{ij}^{s}(t) = \begin{bmatrix} H_{11}^{s}(t) & H_{12}^{s}(t) & H_{13}^{s}(t) \\ H_{21}^{s}(t) & H_{22}^{s}(t) & H_{23}^{s}(t) \\ H_{31}^{s}(t) & H_{32}^{s}(t) & H_{33}^{s}(t) \end{bmatrix} \tag{2}$$

where $H_{ij}^{s}(t)$ indicates the s-th subcarrier of CSI between the i-th transmitting antenna and j-th receiving antenna at each frame t.

2.1. Experiment Setup

To build a human activity recognition dataset and evaluate the proposed method, we collected the data in a cluttered office. The data collection experimental scene and transceiver device are demonstrated in Figure 1. The room size was approximately 6 m $\times$ 8 m. Intel 5300 and Qualcomm Atheros Wi-Fi cards are the most widely used cards for CSI measurements. The open-source CSI Tool, including 802.11n CSI Tool [28] and the Atheros CSI Tool [29], enabled CSI to be exported from commodity wireless Network Interface Cards (NICs). In this paper, an Intel 5300 NIC and Linux 802.11n CSI Tool were utilized to collect the raw CSI measurement. Both the transmitter and receiver work within 802.11n wireless protocol in monitor mode and operate on channel 64 in the 5 GHz frequency band. The bandwidth is 20 MHz. Since the surrounding Wi-Fi devices operate at 2.4 GHz, they only have a tiny impact on the sensing signal.

In this paper, we collected the activity data at 24 distinct positions. The specified location layout is shown in Figure 2. The distance between transmit and receive antennas was 4 m. The adjacent data acquisition positions were approximately 0.6 m apart. Four typical activities were predefined, as demonstrated in Table 1. It is also worthwhile to mention that only a single person was allowed to perform the activity for each case. To obtain enough data samples in the experiment, we collected at least 50 samples for each

activity at each spot. The sampling rate was set as 200 frames per second. Considering each activity lasts for 3.5∼4 s corresponding to 700∼800 frames, we intercepted 750 frames as an activity sample. Since the antenna number of TX and RX are both three and 30 groups of subcarriers can be obtained from each pair of transceiver antennas, the total number of subcarriers is $3 \times 3 \times 30 = 270$.

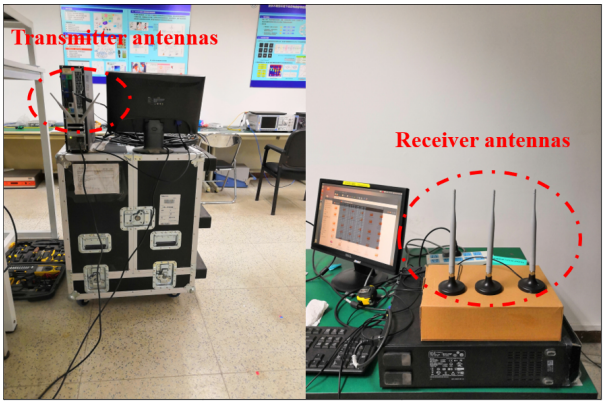

Figure 1. Data collection experimental scene and transceiver device.

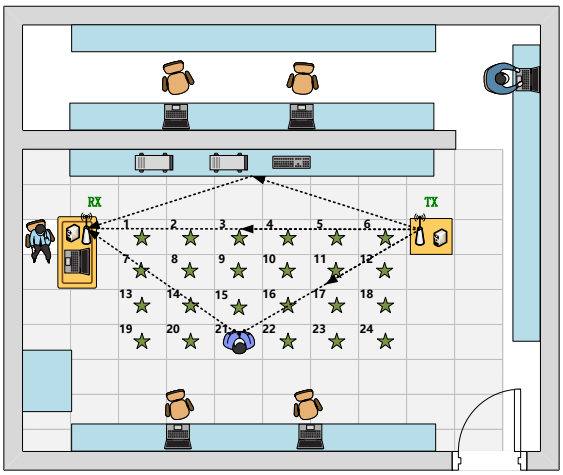

Figure 2. The layout of data collection locations.

Table 1. Predefined activities.

Mark	Activity
O	Drawing a circle with right hand
X	Drawing a cross with right hand
PO	Pushing and opening with two arms
UP	Lifting up and laying down two arms

2.2. Problem Analysis

In this section, we investigate the influence of different human activities at various locations on the transmission of Wi-Fi signals in multiple ways. As demonstrated in Figure 3, the CSI amplitudes of the received signals measured with four different human actions at a fixed location are given. Each subgraph of Figure 3 shows the regular attributes reflected in the different samples of the same action. Moreover, we can observe that each human activity yields diverse feature patterns in the received signals. This is the key to the realization of human activity recognition.

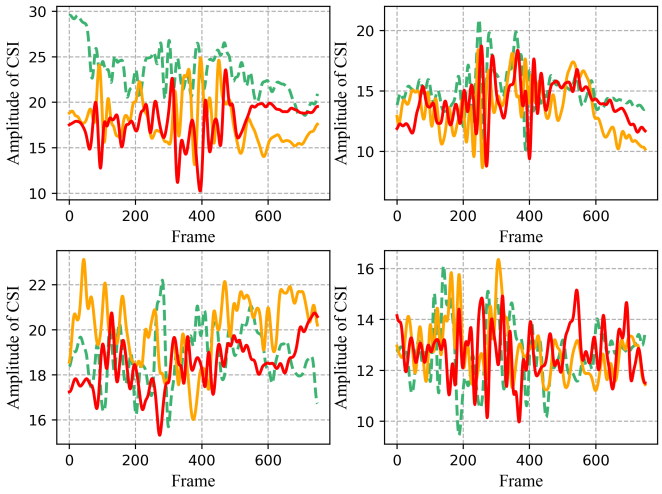

Figure 3. CSI amplitude of four distinct human activities at the identical position. The three curves in each subgraph represent three samples for the same action. The horizontal axis denotes the frame; the ordinate indicates amplitude of CSI.

Inspired by [26], to further demonstrate the location-dependent challenges of Wi-Fi-based HAR, we leverage the low rank and sparse decomposition (LRSD) algorithm to separate the original signal into the low-rank and the sparse part, which describe the background and the activity, respectively. Figure 4 displays the sparse component of CSI for the identical activity at four different positions. As can be observed, although the location-related information has been removed to a certain extent, the same action in each location shows distinct features. Inevitably, the low-rank part contains information related to activities, while the sparse part also consists of certain location-related details. Therefore, further signal processing is still required to extract action-related information.

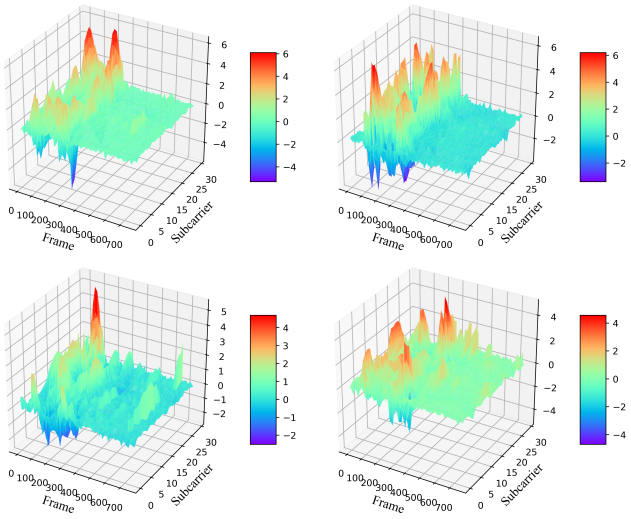

Figure 4. The sparse component of CSI for the identical activity at four different positions.

Consequently, although it is relatively easy to classify the activities at a single location, it may not be possible to ensure good recognition accuracy for location-independent sensing unless some location-invariant features can be extracted. For this reason, an attention mechanism enhanced approach was selected for feature representation since it can concentrate on the features that can generalize to different locations [30].

For a more comprehensive analysis of the problem, the maximum mean difference (MMD) metric was utilized to investigate the distribution difference for data samples collected at the same or different locations. We compared the distributions between two

datasets with MMD through a kernel two-sample test [31], assuming that X, Y are two sample sets of the same activity from two locations. The empirical estimate of MMD distance between them is as follows:

$$MMD^2(X,Y) = \left\| \frac{1}{n_1}\sum_{i=1}^{n_1}\phi(x_i) - \frac{1}{n_2}\sum_{j=1}^{n_2}\phi(y_j) \right\|_{\mathcal{H}}^2 \quad (3)$$

where $x_i \in X$ and $y_i \in Y$ are the randomly selected sample; n_1 and n_2 are the number of X and Y, respectively. ϕ denotes the kernel function that maps the original data to a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$. Empirically, Gaussian kernel is utilized in this paper.

Using the same action at 12 distinct locations, we depict the calculated MMD results for all cases in Figure 5. As illustrated, the MMD of data samples is usually smaller at the same location while becoming larger at different locations. Meanwhile, it is also clearly seen that activity samples at an adjacent area, as shown in Figure 2, have a closer distribution.

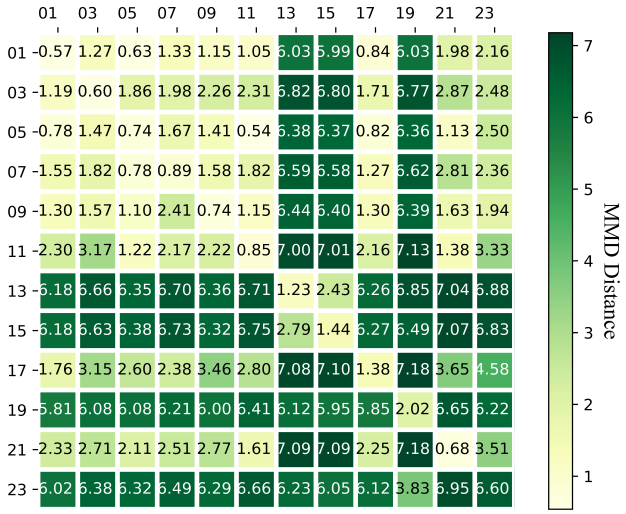

Figure 5. MMD distance of data samples at the same or different locations.

In the light of the analyses above, both the human activities and the location variations can influence signal transmission. Therefore, it is challenging to realize location-independent sensing, especially when inadequate training samples are available. Aiming to drive wireless sensing technology from academic research to industrial application, we urgently need to propose a method to alleviate the issue of domain shifts caused by positional differences, which force the models to fail to generalize between different locations. Considering the case of insufficient samples, we propose a model inspired by few-shot learning [32,33] to achieve LI-HAR in this paper.

3. Materials and Methods

In this part, we first introduce the system overview of LI-HAR. Then, the attention mechanism enhanced characteristic extraction method is proposed. Finally, an improved prototypical network for a few-shot, learning-based HAR method is presented and analyzed.

3.1. System Overview

As shown in Figure 6, the architecture of LI-HAR includes data collection, data preprocessing, feature representation, and model training/testing. Firstly, raw CSI samples were obtained from the Wi-Fi communication system. Then, we computed the amplitude of CSI and denoised the signal with a Butterworth low-pass filter. Moreover, data segmentation was performed to separate the data into multiple samples—the size of which is the frame $\times$ subcarrier, indicating the time an activity lasts for multiplied by the number of

subcarriers. In the third step, the data were transformed to high-dimensional embedding vectors via CNN; then, the attention block was applied to enhance discriminative feature learning. In the final step, aiming to achieve location-independent sensing with a minimum number of samples, the human activity sensing approach with a few-shot learning method was proposed. Next, we present the detailed implementation of the last two steps.

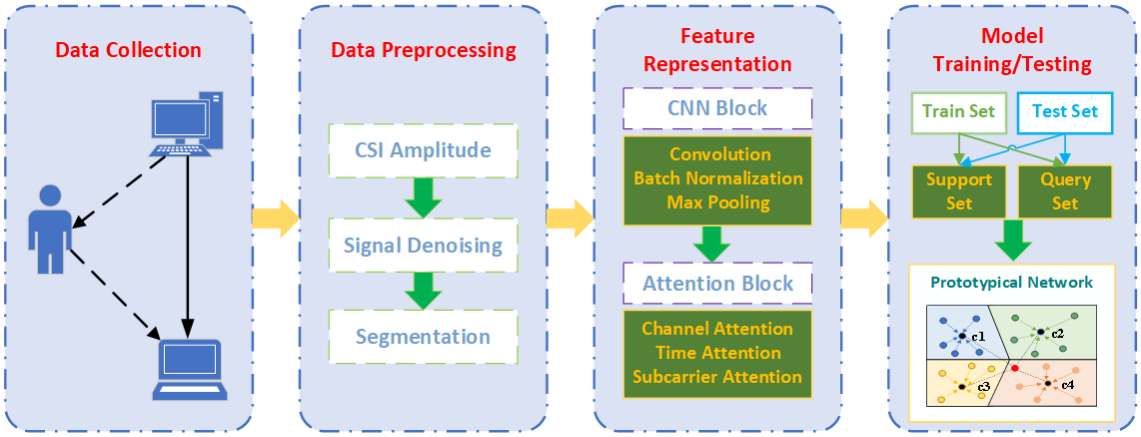

Figure 6. The system architecture for LI-HAR.

3.2. Channel–Time–Subcarrier Attention Mechanism-Based Feature Representation

The deep learning method attracts extensive attention for its powerful capability in feature representation. In terms of feature extraction for two-dimensional data, CNN has been proved to process unparalleled advantages. It not only plays an important role in computer vision but is also widely involved in DFS. However, for the case of location-independent sensing in this paper, apart from the deep features captured by CNN, we expect to gain more generalized characteristics correlated to the activities rather than the locations. This means that the model should place a greater emphasis on the common attributes across different positions.

Each convolution kernel of CNN extracts different types of features and has a distinctive generalization ability. Each channel of the feature maps produced by different kernels represents the characteristics that the kernels learn from different subspaces. Each channel has a distinct significance for various specific tasks; for instance, some channels capture similar features among different positions of the same action. Therefore, by providing such channels with a higher proportion, the output features will be more discriminative for the activities, even at different locations. In addition, the critical ingredient of the temporal sequence can be hardly observed by CNN. As far as the procedure of action is concerned, it includes raising hands or arms; performing some activities; and finally, putting down hands or arms. Intuitively, we are more concerned with the movement in the middle of the process in comparison with lifting and lowering. Therefore, information on the time axis has different degrees of importance. More importantly, due to the influence of frequency selective fading caused by the multipath effect, each subcarrier carries diverse information related to activities and locations. Some subcarriers may be more affected by actions while others may be susceptible to the environment. Further, the distinction and correlation of different subcarriers would be problematic to capture. Accordingly, the interchannel, intertime, and inter-subcarrier relationships should be exploited to produce different weight distributions.

To obtain the discriminative features suitable for different activities irrespective of locations, we proposed a Channel–Time–Subcarrier Attention Mechanism (CTS-AM) improved CNN network in this paper. Specifically, we orderly arrange channel, time–subcarrier, and channel attention modules to form Channel–Time–Subcarrier–Channel Attention Block (CTSC-AB). Figure 7 demonstrates the structure of the channel attention module and the time–subcarrier attention module. Then, we are devoted to learning the attention maps

(weight distributions), which can indicate the significance of different parts in the feature maps and sequentially infer the information that should be emphasized or suppressed. Finally, the attention maps are multiplied by the input feature maps to refine the feature adaptively. Here, we design a CNN network as the feature extractor. It consists of five blocks, each of which possesses a convolutional layer, a batch normalization layer, and a max-pooling layer. Moreover, 64 filters with the kernel size (3,3) are utilized. The activation function is rectified linear unit (ReLU).

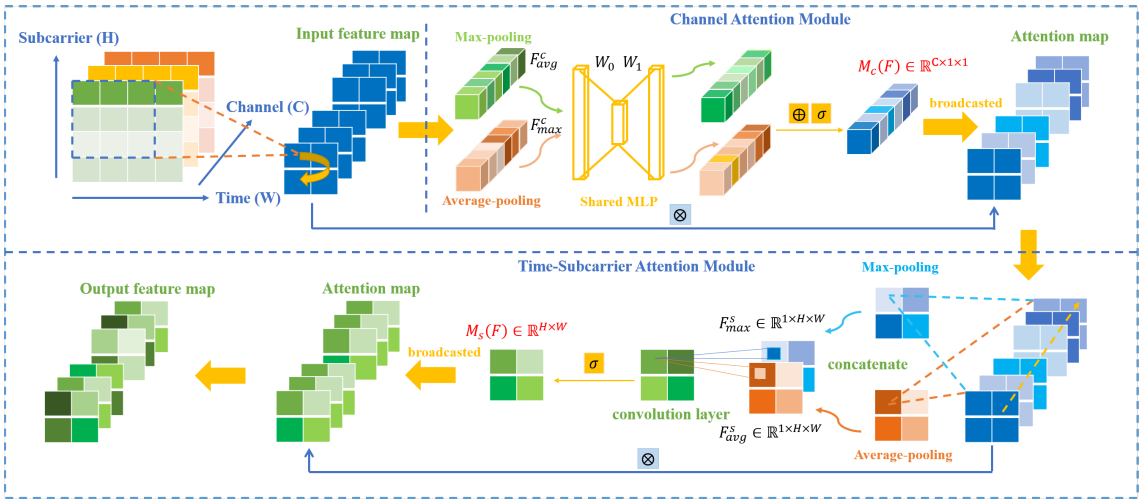

Figure 7. The structure of the channel attention module and the time–subcarrier attention module. σ denotes the sigmoid function. $\otimes$ denotes elementwise multiplication. $\oplus$ denotes elementwise summation.

We assume that the feature map calculated by CNN is $F \in \mathbb{R}^{C \times H \times W}$, where C indicates the number of channels; H and W represent the height (subcarriers) and width (frames/time), respectively. We obtain channel significance $W_{channel1}$ through the channel attention module. Then, we focus on the time–subcarrier dimension through the time–subcarrier attention module to gain the time significance and subcarrier significance $W_{time-subcarrier}$. We append a channel attention module again to enhance the generalization of the time–subcarrier attention block. The whole attention operation can be denoted as

$$F' = W_{channel1} \otimes F \tag{4}$$

$$F'' = W_{time-subcarrier} \otimes F' \tag{5}$$

$$F''' = W_{channel2} \otimes F'' \tag{6}$$

where $\otimes$ indicates elementwise multiplication. During multiplication, the attention values are broadcasted accordingly; channel attention values are broadcasted along the time–subcarrier dimension and vice versa. Then, each attention module illustrated in Figure 7 will be described in detail.

Channel Attention Module. To calculate channel attention weights, we concentrate time–subcarrier knowledge of each feature map through average-pooling and max-pooling operations, respectively. Thereby, two representation including F^c_{avg} and F^c_{max} describing average-pooled features and max-pooled features can be obtained. Then, they are forwarded to a shared network, which consists of a multilayer perceptron (MLP) with one hidden layer to generate a channel attention map $M_c(F) \in \mathbb{R}^{C \times 1 \times 1}$. This indicates the significance of each subspace feature corresponding to each kernel. To decrease the number of parameters, we fixed the hidden activation size to $\mathbb{R}^{C/r \times 1 \times 1}$, where r is the reduction ratio. In this paper, r is set to 8. Then, we combine the output feature vectors from the

shared network by elementwise summation. Specifically, the channel attention can be expressed as

$$\begin{aligned} M_c(F) &= \sigma(M(AvgPool(F)) + M(MaxPool(F))) \\ &= \sigma\left(W_1\left(W_0\left(F^c_{avg}\right)\right) + W_1(W_0(F^c_{max}))\right) \end{aligned} \tag{7}$$

where σ indicates the sigmoid function, $W_0 \in \mathbb{R}^{C/r \times C}$, and $W_1 \in \mathbb{R}^{C \times C/r}$. M in the equation is short for MLP.

Time–Subcarrier Attention Module. To calculate the time–subcarrier attention, the average-pooling and max-pooling along the channel axis is conducted to form two 2D features across the channel: $F^s_{avg} \in \mathbb{R}^{1 \times H \times W}$ and $F^s_{max} \in \mathbb{R}^{1 \times H \times W}$. Then, we concatenate them to build an efficient feature descriptor. Finally, a convolution layer is utilized to produce a time–subcarrier attention map $M_s(F) \in \mathbb{R}^{H \times W}$, which indicates the information to focus on along the time and subcarrier dimension. In summary, the time–subcarrier attention can be denoted as

$$\begin{aligned} M_s(F) &= \sigma\left(f^{7\times7}([AvgPool(F); MaxPool(F)])\right) \\ &= \sigma\left(f^{7\times7}\left(\left[F^s_{avg}; F^s_{max}\right]\right)\right) \end{aligned} \tag{8}$$

where σ denotes the sigmoid function, $f^{7\times7}$ represents a convolution operation with the filter size of 7×7, and $[\cdot;\cdot]$ means concatenation operation.

3.3. Few-Shot, Learning-Based Human Activity Recognition

Inspired by the prototypical network [34], we propose a few-shot, learning-based activity sensing method. Unlike the traditional approach that is applied to solve the issue involving new class learning with very few samples, this paper utilizes it for the perception of activity at the new location when insufficient training samples are available.

We conduct the few-shot learning task by subsampling the positions and the samples to make up a training set S_{Train}, a validation set $S_{Validation}$, and a testing set S_{Test}. Then, the support set $S_{Support}$ and the query set S_{Query} are randomly selected from the three sets. We assume that there are N activities in total. The detailed procedure of class prediction and training loss computation is described in Algorithm 1.

In this paper, Adam is applied to optimize the parameters. The exponential decay rate $\rho 1$ and $\rho 2$ are empirically set as 0.9 and 0.999, and the learning rate is set as 0.001. We train the model for 40 epochs.

Algorithm 1 Pseudocode of class prediction and training loss computation

Input: The number of activity class K, Training set S_{Train}=$\{(x_1, y_1), \ldots, (x_N, y_N)\}$, the feature extractor f_φ, the number of support samples of each class in the training set N_s, the number of query samples of each class in the training set N_Q.

Output: Predicted class label $\hat{y}$, loss J.

for batch in traning set **do**
  for class k in classes $\{1, \ldots, K\}$ **do**
    Choose N_s support samples for the class k to form the support set $S^k_{Support}$.
    Choose N_Q support samples for the class k to form the support set S^k_{Query}.
    $S^k_{Support} \rightarrow f_\varphi\left(S^k_{Support}\right)$
    $S^k_{Query} \rightarrow f_\varphi\left(S^k_{Query}\right)$ // feature embedding
    Calculate the prototype of each class with the support samples

$$C_k = \frac{1}{\left|S^k_{Support}\right|} \sum_{(x_i, y_i) \in S^k_{Support}} f_\varphi(x_i)$$

  end for
  Loss $J = 0$ //Loss initialization
  for query samples $(\bar{x}, \bar{y})$ in S^k_{Query} **do**
    for class k in classes $\{1, \ldots, K\}$ **do**
      Calculate the distance between the query samples and the prototype of each class

$$d_k\left(C_k, f_\varphi(\bar{x})\right) = \left\|C_k - f_\varphi(\bar{x})\right\|^2$$

    end for
    $\hat{y} = \arg\min_k d_k\left(C_k, f_\varphi(\bar{x})\right)$ // Predicted label
    $J \leftarrow J + \frac{1}{K} \log\_soft\max(d_k)$ // Loss update
  end for
end for

4. Performance Evaluation

This section conducts the experiments and evaluates the performance of the proposed LI-HAR system. Firstly, we investigate the feasibility of the recognition method enhanced by the CTS-AM improved few-shot learning. Then, to verify the superiority, a comparison study is performed. At last, the robustness is explored via evaluating the accuracy under different conditions, including the training locations selection strategies, the number of training positions and shots, as well as different signal-to-noise ratio (SNR) levels.

4.1. Feasibility Evaluation

Overall performance. In this section, we elaborate on the system performance under three different experiment settings. Firstly, the training position is identical to the test position, which is a random one of our 24 locations shown in Figure 2. Both the training samples and the testing samples come from a single position; thus, we call it single-location recognition. Secondly, we discuss mixed-location recognition, where the training and testing samples are selected from all 24 locations and we have the same number of samples at each position. Thirdly, to evaluate the performance of location-independent sensing, we choose the samples from part of the locations as the training set and the samples from all 24 locations as the testing set. Specifically, we divided 50 samples for each activity at different positions into the training set, validation set, and testing set at a proportion of 60%, 20%, and 20%, respectively. We present the overall performance with four training locations selected from 24 positions, which follows the principle of equal interval sampling. The dataset allocation in the location-independent HAR evaluation is shown in Table 2.

Table 2. Dataset allocation in the location-independent HAR evaluation.

Number	Activities	Locations	Samples
Training Set	4	4	30
Validation Set	4	24	10
Testing Set	4	24	10

In this section, according to the training strategy of the few-shot learning approach, we randomly select five support points and 15 query points from the training, verification, and testing datasets of each activity, respectively. It is worth noting that the locations of the five support samples and 15 query samples of each activity are random and may be the same or different. More importantly, only data from the training set are utilized to update the model parameters; specifically, the model parameters do not change during the validation and testing steps.

As depicted in Table 3, LI-HAR performs well in the three experiment settings. In particular, the overall accuracy of the single-location and the mixed-location activity recognition can reach 99.39% and 98.52%, respectively. Moreover, each activity can achieve satisfactory accuracy. For location-independent recognition, the accuracy can still attain 91.98%. Specifically, O and PO are easier to distinguish relative to the other two. In conclusion, our LI-HAR system is promising for both location-dependent and location-independent recognition.

Table 3. The recognition accuracy of different activities.

Accuracy (%)	O	X	PO	UP	Average	Overall
Single location	100	100	100	100	100	**99.39**
Mixed locations	100	100	100	93.33	98.33	**98.52**
Location-independent	100	86.67	93.33	86.67	91.67	**91.98**

Feature visualization. To further illustrate the feasibility of our system, we reduce the dimension of the feature representation before and after mapping the embedding space to two dimensions and visualize the output leveraging the T-SNE method [35]. As shown in Figure 8, samples that belong to the same category are clustered together after the process while cluttered and intuitively indistinguishable before the process. To some extent, it demonstrates that the proposed system is effective while applying the few-shot learning strategy based on the prototypical network with the CNN and CTS-AM feature representative method.

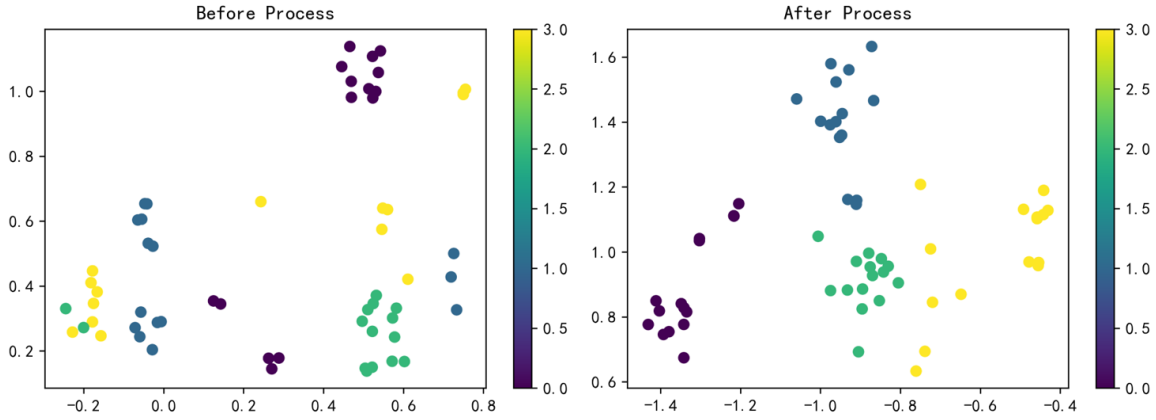

Figure 8. T-SNE of test samples before and after the model learning process.

4.2. Superiority Evaluation

Comparison with different attention mechanisms. We perform a comprehensive study to verify the superiority of the CTS-AM in the prototypical network. In the following part, the term "PN" in the table or figure is shot for the prototypical network. In this section, to distinguish the proposed attention structure from the others, we apply CTSC-AB to denote our method. Besides the CTSC-AB, we have explored the impacts of each part, including channel attention and time–subcarrier attention. Furthermore, we have studied different combinations of them, consisting of channel attention (C-AB), time–subcarrier attention (TS-AB), channel–time–subcarrier attention (CTS-AB), and time–subcarrier–channel attention (TSC-AB).

Table 4 has illustrated the advantages of CTSC-AB with the different number of training locations. We take 4/6 training locations with equal interval sampling from 24 locations as examples. The average accuracy can be improved by 1.86% and 2.39% when there are 4 and 6 training locations, respectively. The results demonstrate that our CTSC-AB improved the prototypical network as the most effective method compared with the others in Table 4. As can be seen, all the attention blocks enhance the results; however, the single block offers only a limited accuracy boost, while the combination of the C-AB and TS-AB can achieve a relatively stable and superior performance improvement.

Table 4. The average accuracy for different attention mechanisms.

Accuracy (%)	4 Training Locations	6 Training Locations
PN	90.12	91.66
PN + C-AB	90.65	92.68
PN + TS-AB	90.16	92.06
PN + CTS-AB	91.21	92.98
PN + TSC-AB	90.67	92.88
PN + CTSC-AB	**91.98**	**94.05**

Comparison with different recognition approaches. To verify the superiority of the proposed method, we explore the other two typical approaches, which are CNN and WiHand [26]. CNN is a classical feature representation method that is the most commonly used in wireless sensing. LRSD-based WiHand aims to remove activity-irrelevant information and outperforms the other location-independent sensing method. In this part, we conduct the comparison study and discuss the recognition accuracy of the three aforementioned experimental settings. The comparison results are shown in Table 5. Note that the selections of the four training locations are the same as the above feasibility evaluation. Thirty subcarriers are used for recognition and 20 features are extracted for WiHand. In Table 5, S L, M L, and L I are short for single location, mixed location, and location-independent, respectively.

Table 5. Comparison study of different recognition methods.

Accuracy (%)	S L	M L	L I
CNN	98.46	94.02	80.24
WiHand	96.15	91.50	80.18
PN	99.07	97.22	90.12
PN + CTSC-AB	**99.39**	**98.52**	**91.98**

We can observe that all three approaches achieve high accuracy in single-location recognition. The results are also promising when recognizing activity with the mixed locations. For location-independent recognition, LI-HAR (PN + CTSC-AB) has an average 91.89% accuracy with only four training locations. It is noted that the few-shot, learning-based method improves the average recognition accuracy by 9.88% and 9.94% compared with CNN and WiHand. When the prototypical network is improved by CTSC-AB, the recognition rate increases by 11.74% and 11.80%, respectively. Although CNN

has an absolute advantage in the case of IID, it fails to identify activity when the distribution varies at different positions. In conclusion, LI-HAR has certain advantages to realize location-independent perception when very few training samples are available.

4.3. Robustness Evaluation

Performance of LI-HAR in terms of different training location selection strategies and different numbers of training locations. In this section, we adopt two training position selection strategies to show the robustness of the proposed method. One is the scheme used in the above experiment, in which the training locations are selected with an equal interval from 24 locations. Taking four training positions as an example, we pick one in every six positions and there are six options. Namely, the four locations of each column in Figure 2 are an option. The other strategy is depicted in Figure 9, in which the training locations satisfy the axial or centroid symmetry. The positions are relatively decentralized rather than spreading in a line parallel to the transceiver. These two position selection strategies could demonstrate the generalization of the system. In addition, we will discuss how the number of training locations influences sensing capability. Specifically, we explore 4/6/8/12/24 positions for training and 24 positions for testing.

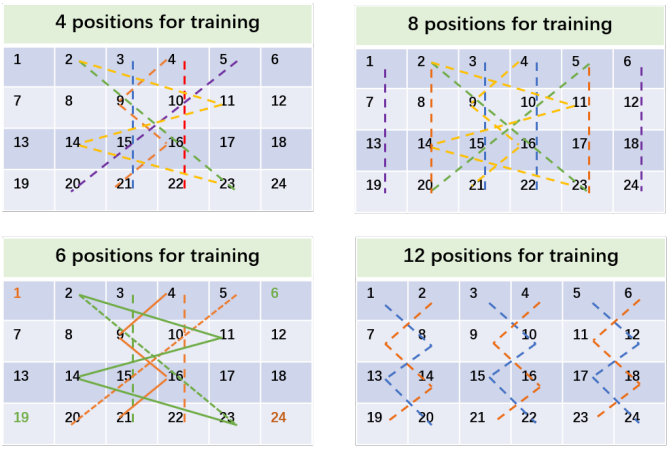

Figure 9. The training location selection strategies. Specifically, for 4/8/12 training locations, the positions where the same colored straight line goes through or the inflection points and the enthesis of the same colored broken lines constitute the training locations. For six training locations, the straight lines or broken lines along with the same colored marked positions form the training locations.

As illustrated in Table 6, when we choose four training positions, the accuracy for the first training position selection strategy is 91.98%, while for the second training position, the selection strategy is 92.90%. The results indicate that the accuracy of strategy 1 is better than that of strategy 2, except for four training locations. This may be due to some off-center combination of positions in Strategy 1 (such as 1, 7, 13, 19 and 6, 12, 18, 24); the distribution of other locations is considerably different. Nevertheless, satisfactory identification results can still be obtained. We show the average recognition accuracy of each four-training-position combination scheme in Table 7. It is also concluded that more training positions lead to higher recognition accuracy.

Table 6. The average recognition accuracy for different training position selection strategies and different numbers of training locations.

Accuracy (%)	4	6	8	12	24
Strategy 1	91.98	94.05	95.04	96.69	98.52
Strategy 2	92.90	93.98	94.61	95.07	98.52

Table 7. The average recognition accuracy for distinct four-training-position combination scheme.

Training Locations	Accuracy (%)
1, 7, 13, 19	89.46
2, 8, 14, 20	91.89
3, 9, 15, 21	93.17
4, 10, 16, 22	93.61
5, 11, 17, 23	92.90
6, 12, 18, 24	90.83
Average	91.98

Performance of LI-HAR for different number of shots. We discuss how the size of support points for testing utilized to calculate the prototypes in CTS-AM improved prototypical network influences the recognition accuracy. We take four training locations and 24 testing locations as an example. The prototypical network for few-shot learning before/after the improvement by CTS-AM is discussed. The identification results with strategy 1 and strategy 2 are listed in Figure 10.

It is noted that although there is only one sample for each action of the unseen locations, the improved prototypical network can still obtain a promising result. Moreover, the accuracy will boost with a larger sample size. This is because the representation capability of the prototype we compute to describe a category relies on the number of support samples.

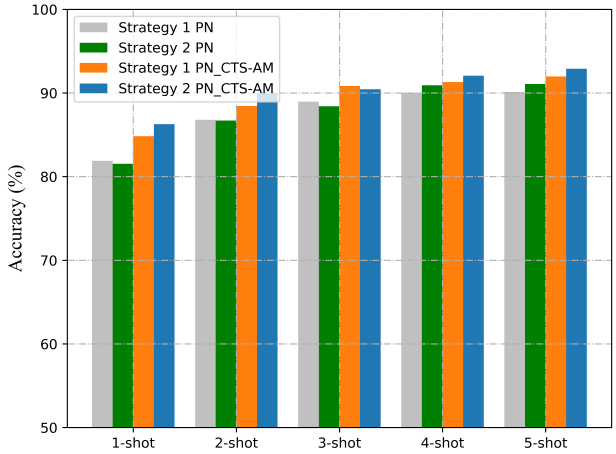

Figure 10. The average recognition accuracy for distinct shots.

Performance of LI-HAR with different SNR levels. To verify the robustness of the system under different noise intensity environments, we add Gaussian white noise with a mean value of 0 and a variance of 1 to the original CSI data, generating signals with different SNR. We apply activities from 4 locations with strategy 1 for training and 24 locations for testing. We discuss 3-shots and 5-shots. The recognition results are illustrated in Figure 11.

It can be observed that the recognition accuracy improves with the increase in SNR. When applying the CTS-AM enhanced prototypical network, the 5-shots outperform 3-shots. In summary, the proposed method can fulfill location-independent activity recognition with very few samples.

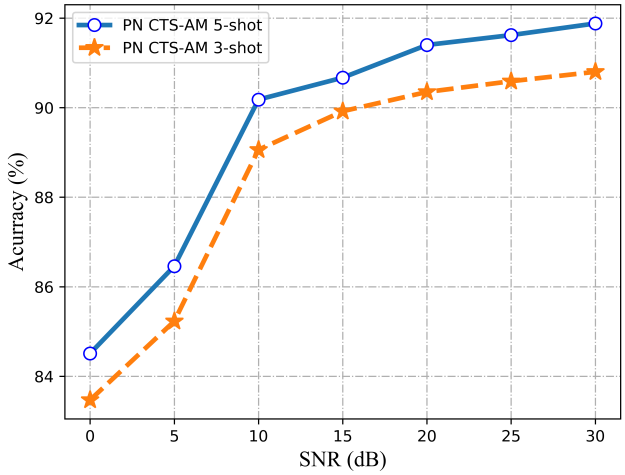

Figure 11. The average recognition accuracy with the variation in SNR.

5. Discussion

Results Discussion: In this paper, we promisingly achieve Wi-Fi-based location-independent human activity recognition with limited training samples at the unseen positions. The performance evaluation involving feasibility and superiority indicates that the proposed method possesses great sensing capability. Especially, the robustness evaluation involving different training location selection strategies, different numbers of training locations and shots, and different SNR levels shows great potential ability in practical application scenarios. In addition, we recorded the training time used for the proposed method. In the case of training 20 epochs, the training process took less than two minutes. In this case, the model can converge well and achieve ideal recognition accuracy. When we train the model for 40 epochs, the training process took less than four minutes. Both can meet the time requirements of offline training and online recognition.

Limitation: Although some progress has been made in location-independent human activity recognition, there are still many challenging issues and limitations that need to be solved. Firstly, since we are preliminarily exploring the feasibility of the proposed method, we only discuss the situation where there is only one active target in the environment. However, in practical application scenarios, there is usually interference from others; thus, how to remove the influence of such interference is also a key issue. In addition, as required in the data collection process of most current wireless perception studies, we will also limit the consistency of motion sample collection, such as the orientation and space of the activity, which will have a great impact on signal transmission.

Future Work: In the future, researchers could gradually remove the restriction and explore a more generalized and robust method to realize the human activity perception independent of various external factors. In addition, more complex application scenarios involving multiple targets and non-line-of-sight situations should be considered. Besides Wi-Fi devices, the multisignal and multiterminal fusion methods will also provide a broader idea for the development of high-precision intelligent sensing.

6. Conclusions

In this paper, we propose a novel location-independent human activity recognition system named LI-HAR. The system owns the capability of transferring the knowledge acquired from some locations to others. Technically, the proposed few-shot learning recognition approach is based on the CTS-AM improved prototypical network, which can learn the feature representation at all locations with only very few samples. The method concentrates on the common characteristics of distinct positions and extracts discriminable characteristics of each action. We built a comprehensive dataset for evaluation. The experiment results demonstrate that the method can attain an average accuracy of more than 90%, with four locations for training and 24 locations for testing, given only five samples

for each activity. Consequently, it concludes that the proposed method is achievable for location-independent human activity recognition.

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

Funding: This work is supported by the National Natural Sciences Foundation of China (No. 62071061), and the BUPT Excellent Ph.D. Students Foundation (No. CX2019110), and Beijing Institute of Technology Research Fund Program for Young Scholars.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ma, Y.; Zhou, G.; Wang, S. WiFi sensing with channel state information: A survey. *ACM Comput. Surv.* **2019**, *52*, 1–36. [CrossRef]
2. Wang, D.; Yang, J.; Cui, W.; Xie, L.; Sun, S. Multimodal CSI-based Human Activity Recognition using GANs. *IEEE Internet Things J.* **2021**, *8* , 17345–17355. [CrossRef]
3. Bianchi, V.; Bassoli, M.; Lombardo, G.; Fornacciari, P.; Mordonini, M.; De Munari, I. IoT Wearable Sensor and Deep Learning: An Integrated Approach for Personalized Human Activity Recognition in a Smart Home Environment. *IEEE Internet Things J.* **2019**, *6*, 8553–8562. [CrossRef]
4. Wang, A.; Zhao, S.; Zheng, C.; Chen, H.; Liu, L.; Chen, G. HierHAR: Sensor-Based Data-Driven Hierarchical Human Activity Recognition. *IEEE Sens. J.* **2021**, *21*, 3353–3365. [CrossRef]
5. Hao, X.; Li, J.; Guo, Y.; Jiang, T.; Yu, M. Hypergraph Neural Network for Skeleton-Based Action Recognition. *IEEE Trans. Image Process.* **2021**, *30*, 2263–2275. [CrossRef] [PubMed]
6. Zhang, H.B.; Zhang, Y.X.; Zhong, B.; Lei, Q.; Yang, L.; Du, J.X.; Chen, D.S. A comprehensive survey of vision-based human action recognition methods. *Sensors* **2019**, *19*, 1005. [CrossRef]
7. Liu, J.; Liu, H.; Chen, Y.; Wang, Y.; Wang, C. Wireless Sensing for Human Activity: A Survey. *IEEE Commun. Surv. Tutor.* **2020**, *22*, 1629–1645. [CrossRef]
8. Huang, X.; Dai, M. Indoor Device-Free Activity Recognition Based on Radio Signal. *IEEE Trans. Veh. Technol.* **2016**, *66*, 5316–5329. [CrossRef]
9. Virmani, A.; Shahzad, M. Position and orientation agnostic gesture recognition using wifi. In Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services, MobiSys, Niagara Falls, NY, USA, 19–23 June 2017; pp. 252–264.
10. Zhang, R.; Jing, X.; Wu, S.; Jiang, C.; Yu, F.R. Device-Free Wireless Sensing for Human Detection: The Deep Learning Perspective. *IEEE Internet Things J.* **2020**, *8*, 2517–2539. [CrossRef]
11. Ding, C.; Hong, H.; Zou, Y.; Chu, H.; Zhu, X.; Fioranelli, F.; Le Kernec, J.; Li, C. Continuous human motion recognition with a dynamic range-Doppler trajectory method based on FMCW radar. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 6821–6831. [CrossRef]
12. Wang, Y.; Liu, H.; Cui, K.; Zhou, A.; Li, W.; Ma, H. m-Activity: Accurate and Real-Time Human Activity Recognition via Millimeter Wave Radar. In Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 8298–8302.
13. Zhong, Y.; Yang, Y.; Zhu, X.; Dutkiewicz, E.; Zhou, Z.; Jiang, T. Device-free sensing for personnel detection in a foliage environment. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 921–925. [CrossRef]
14. Huang, Y.; Zhong, Y.; Wu, Q.; Dutkiewicz, E.; Jiang, T. Cost-Effective Foliage Penetration Human Detection under Severe Weather Conditions based on Auto-Encoder/Decoder Neural Network. *IEEE Internet Things J.* **2018**, *6*, 6190–6200. [CrossRef]
15. Zhong, Y.; Yang, Y.; Zhu, X.; Huang, Y.; Dutkiewicz, E.; Zhou, Z.; Jiang, T. Impact of Seasonal Variations on Foliage Penetration Experiment: A WSN-Based Device-Free Sensing Approach. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 5035–5045. [CrossRef]
16. Zhong, Y.; Bi, T.; Wang, J.; Wu, S.; Huang, Y. Low data regimes in extreme climates: Foliage penetration personnel detection using a wireless network-based device-free sensing approach. *Ad Hoc Netw.* **2021**, *114*, 102438. [CrossRef]
17. Yousefi, S.; Narui, H.; Dayal, S.; Ermon, S.; Valaee, S. A survey on behavior recognition using WiFi channel state information. *IEEE Commun. Mag.* **2017**, *55*, 98–104. [CrossRef]

18. Zhou, S.; Zhang, W.; Peng, D.; Liu, Y.; Liao, X.; Jiang, H. Adversarial WiFi Sensing for Privacy Preservation of Human Behaviors. *IEEE Commun. Lett.* **2020**, *24*, 259–263. [CrossRef]
19. Huang, J.; Liu, B.; Chen, C.; Jin, H.; Liu, Z.; Zhang, C.; Yu, N. Towards Anti-Interference Human Activity Recognition Based on WiFi Subcarrier Correlation Selection. *IEEE Trans. Veh. Technol.* **2020**, *69*, 6739–6754. [CrossRef]
20. Tang, Z.; Liu, Q.; Wu, M.; Chen, W.; Huang, J. WiFi CSI gesture recognition based on parallel LSTM-FCN deep space-time neural network. *China Commun.* **2021**, *18*, 205–215. [CrossRef]
21. Wu, C.; Yang, Z.; Zhou, Z.; Liu, X.; Liu, Y.; Cao, J. Non-Invasive Detection of Moving and Stationary Human with WiFi. *IEEE J. Sel. Areas Commun.* **2015**, *33*, 2329–2342. [CrossRef]
22. Wang, H.; Zhang, D.; Wang, Y.; Ma, J.; Wang, Y.; Li, S. RT-Fall: A real-time and contactless fall detection system with commodity WiFi devices. *IEEE Trans. Mob. Comput.* **2016**, *16*, 511–526. [CrossRef]
23. Wang, W.; Liu, A.X.; Shahzad, M.; Ling, K.; Lu, S. Device-free human activity recognition using commercial WiFi devices. *IEEE J. Sel. Areas Commun.* **2017**, *35*, 1118–1131. [CrossRef]
24. 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]
25. Zhong, Y.; Wang, J.; Wu, S.; Jiang, T.; Wu, Q. Multi-Location Human Activity Recognition via MIMO-OFDM Based Wireless Networks: An IoT-Inspired Device-Free Sensing Approach. *IEEE Internet Things J.* **2020**, *8*, 15148–15159. [CrossRef]
26. Lu, Y.; Lv, S.; Wang, X. Towards Location Independent Gesture Recognition with Commodity WiFi Devices. *Electronics* **2019**, *8*, 1069. [CrossRef]
27. Huang, Y.; Wu, Q.; Xu, J.; Zhong, Y.; Zhang, Z. Unsupervised Domain Adaptation with Background Shift Mitigating for Person Re-Identification. *Int. J. Comput. Vis.* **2021**, *129*, 2244–2263. [CrossRef]
28. Halperin, D.; Hu, W.; Sheth, A.; Wetherall, D. Tool release: Gathering 802.11n traces with channel state information. *ACM SIGCOMM Comput. Commun. Rev.* **2011**, *41*, 53. [CrossRef]
29. Xie, Y.; Li, Z.; Li, M. Precise Power Delay Profiling with Commodity Wi-Fi. *IEEE Trans. Mob. Comput.* **2019**, *18*, 1342–1355. [CrossRef]
30. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. *CBAM: Convolutional Block Attention Module*; Springer: Cham, Switzerland, 2018.
31. Gretton, A.; Borgwardt, K.M.; Rasch, M.J.; Schölkopf, B.; Smola, A. A kernel two-sample test. *J. Mach. Learn. Res.* **2012**, *13*, 723–773.
32. Lake, B.; Salakhutdinov, R.; Gross, J.; Tenenbaum, J. One shot learning of simple visual concepts. In Proceedings of the Annual Meeting of the Cognitive Science Society, Boston, MA, USA, 20-23 July, 2011, Volume 33.
33. 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.
34. Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for few-shot learning. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4077–4087.
35. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.