MDPI

Article

Wi-Fi-Based Location-Independent Human Activity Recognition via Meta Learning

Xue Ding [1], **Ting Jiang** [1], **Yi Zhong** [2,*], **Yan Huang** [3] and **Zhiwei Li** [1]

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.); lzw_fire@bupt.edu.cn (Z.L.)
2   School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China
3   Global Big Data Technologies Centre (GBDTC), School of Electrical and Data Engineering, University of Technology Sydney, Sydney, NSW 2007, Australia; Yan.Huang-3@student.uts.edu.au
*   Correspondence: yi.zhong@bit.edu.cn

Abstract: Wi-Fi-based device-free human activity recognition has recently become a vital underpinning for various emerging applications, ranging from the Internet of Things (IoT) to Human–Computer Interaction (HCI). Although this technology has been successfully demonstrated for location-dependent sensing, it relies on sufficient data samples for large-scale sensing, which is enormously labor-intensive and time-consuming. However, in real-world applications, location-independent sensing is crucial and indispensable. Therefore, how to alleviate adverse effects on recognition accuracy caused by location variations with the limited dataset is still an open question. To address this concern, we present a location-independent human activity recognition system based on Wi-Fi named WiLiMetaSensing. Specifically, we first leverage a Convolutional Neural Network and Long Short-Term Memory (CNN-LSTM) feature representation method to focus on location-independent characteristics. Then, in order to well transfer the model across different positions with limited data samples, a metric learning-based activity recognition method is proposed. Consequently, not only the generalization ability but also the transferable capability of the model would be significantly promoted. To fully validate the feasibility of the presented approach, extensive experiments have been conducted in an office with 24 testing locations. The evaluation results demonstrate that our method can achieve more than 90% in location-independent human activity recognition accuracy. More importantly, it can adapt well to the data samples with a small number of subcarriers and a low sampling rate.

Keywords: Wi-Fi sensing; human activity recognition; location-independent; meta learning; metric learning; few-shot learning

1. Introduction

Human Activity Recognition (HAR) has been considered as an indispensable technology in many Human–Computer Interaction (HCI) applications, such as smart home, health care, security surveillance, virtual reality, and location-based services (LBS) [1,2]. Traditional human activity sensing approaches are the wearable sensor-based methods [3,4] and the camera (vision)-based methods [5,6]. While promising and widely used, these device-based approaches suffer from respective drawbacks, making them fail to be suitable for all the application scenarios. For instance, the wearable sensor-based method works only if the users are carrying the sensors, such as smartphones, smart shoes, or smartwatches with built-in inertial measurement units (IMUs), including gyroscope, accelerometer, magnetometer, etc. However, it is inconvenient for constant use. In addition, although the camera (vision)-based method could potentially achieve satisfactory accuracy, it is limited by certain shortcomings, such as privacy leakage, line-of-sight (LOS) and light conditions, etc. Moreover, both methods require dedicated devices, which are high cost. In addition, the durability of the devices is another critical factor that should be considered.

Recently, Wi-Fi-based human activity recognition has attracted extensive attention in both academia and industry, becoming one of the most popular device-free sensing (DFS) technologies [7,8]. Compared with the other wireless signals, such as Frequency Modulated Continuous Wave (FMCW) [9,10], millimeter-wave (MMW) [11,12], and Ultra Wide Band (UWB) [13–17], Wi-Fi possesses the most prominent and potential advantage, which is that it is ubiquitous in people's daily lives. Leveraging the commercial off-the-shelf (COTS) devices, Wi-Fi-based human activity recognition obviates the need for additional specialized hardware. Beyond this, it also has the same merits as other wireless signals, including the capability to operate in darkness and non-line-of-sight (NLOS) situations while providing better protection of users' privacy in the meantime. As a result, research on Wi-Fi-based human activity recognition has proliferated rapidly over the past decade [18–21].

Previous attempts involving Wi-Fi-based sensing yielded great achievements, such as E-eyes [22], CARM [23], etc. However, the major challenge referring to the generalization performance of the approaches and systems has not been fully explored and solved. For instance, when deployed in a room, the system must work well in each location rather than a specified location. Location-independent sensing is one of the most necessary generalization capabilities. It can also be regarded as the ability of a method to transfer among different locations. Note that this is a crucial factor to determine whether the technology can be commercialized. According to the principle of wireless perception, it is not difficult to find that human activities in different locations have different effects on signal transmission. Specifically, activities conducted by people in different locations will change the path of wireless signal propagation in different ways, leading to diverse multipath superimposed signals at the receiver. It is worth noting that these signals have different data distributions, which can be treated as different domains. Hence, it is clear that the human activity recognition model trained in a specific domain will not work well in the other domains. The most obvious solution is to provide abundant data for each domain to learn the characteristics of activities in the different domains. However, it is labor-intensive, time-consuming, and with poor user experience to obtain a large amount of data in practical applications. Therefore, how to utilize as few samples as possible to solve the problem of location-independent perception to achieve outstanding generalization performance is desired.

Some solutions have been proposed to solve the above problems, and remarkable progress has been made, which lays a solid foundation for realizing location-independent sensing with good generalization ability. The solutions fall into the following four categories: (1) Generate virtual data samples for each location [24], (2) Separate the activity signal from the background [25,26], (3) Extract domain-independent features [27], and (4) Domain adaptation and transfer learning. Some approaches involving other domains (such as environment, orientation, and person) can also be grouped into these four categories. However, they pay less attention to location-independent sensing [28–31]. Although the above methods promote the process of device-free human activity recognition from academic research to industrial application, there are still some limitations. WiAG [24] requires the user to hold a smartphone in hand for one of the training sample collections in order to estimate the parameters. Widar 3.0 [27] is limited by link numbers and complex parameter estimation methods. FALAR [25] benefits from its development of a new OpenWrt firmware which can get fine-grained Channel State Information (CSI) of all the 114 subcarriers, improving data resolution. Similarly, high transmission rates of the perception signal (such as 2500 packets/s in Lu et al. [26]) can also boost the resolution. As the author described by Zhou et al. [30], a low sampling rate may miss some key information, which accounts for the deterioration in the system performance. However, using shorter packets helps reduce latency and has less impacts on communication. The detailed discussions about the effect of different sampling rates on the sensing accuracy can be found in the evaluation in [27,30]. In summary, a location-independent method that can adapt to data

with a small number of antennas and subcarriers as well as a small data transmission rate is required.

This work aims to realize device-free location-independent human activity recognition using as few samples as possible. It means that the model trained with the source domain data samples can perform well on the target domain with only very few data samples. We describe our task as a few-shot learning problem, improving the performance of the model in the unseen domain when its amount of available data are relatively small [32]. The task is also consistent with meta learning, whose core idea is learning to learn [33]. They have been successfully applied in a variety of fields to solve classification tasks. Inspired by the typical meta learning approach matching network, we apply the learning method obtained from the source domain to the target domain by means of metric learning [34,35]. Assuming that, although there is no stable feature that can describe a class of actions well, we can still identify its category through maximizing the inter-class variations and minimizing intra-class differences. Judging the category of a sample by calculating the distance can be regarded as a learning method. To realize location-independent sensing, we expect to learn not only the discriminative features representation specific to our task but also the distance function and metric relationships that can infer the label with a confident margin.

In this paper, we first comprehensively and visually investigate the effects of the same activity at different locations on wireless signal transmission. We also analyze the signal received in different antennas and subcarriers with different sampling rates. Moreover, we discuss how different locations affect signal transmission without any other variable influence factors by utilizing data collected from the anechoic chamber. Then, we propose a device-free location-independent human activity recognition system named WiLiMetaSensing, which is based on meta learning to enable few-shot learning sensing. Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) are introduced for feature representation. Unlike the traditional feature extraction process for Wi-Fi signal based on LSTM, in this paper, the memory capacity of LSTM is utilized to retain the valuable information of the samples from all the activities. In addition, an attention mechanism-based metric learning method is used to learn the metric relations of the activity with the same or different categories. Finally, extensive experiments are conducted to explore the recognition performance of the proposed system. The evaluation refers to the property involving single location, mixed locations, and location-independent sensing. Unlike existing evaluations, we reduce the sampling rate, the number of subcarriers, and antennas. Experiments show that WiLiMetaSensing achieves satisfying results with robust performance in a variety of situations.

2. Preliminary

2.1. Channel State Information

The Wi-Fi-based wireless sensing principle is leveraging the influence of perceptual targets on the transmitted signal for recognition. During the transmission from the transmitter (TX) to the receiver (RX), the wireless signal would be refracted, reflected, and scattered when encountering obstacles and objects (dynamic or static), which results in the superposition of multipath signals at the receiver. In a Multiple Input Multiple Output (MIMO) and Orthogonal Frequency Division Multiplexing (OFDM)-based Wi-Fi communication system, this process can be described by fine-grained CSI. In recent years, the physical layer information of some commercial off-the-shelf (COTS) Wi-Fi devices has gradually become available, making it possible to obtain CSI directly [36]. Compared with coarse-grained Received Signal Strength Indicator (RSSI), CSI provides richer channel characteristics.

Letting y and x respectively denote the received signal and transmitting signal, the relation between y and x can be modeled as:

$$y = Hx + N \tag{1}$$

where H is the channel matrix, and N is the noise vector. H completely describes the characteristics of the channel. The process of calculating the channel matrix is called

channel estimation. *H* can be represented in either channel frequency response (CFR) in frequency domain or channel impulse response (CIR) in time domain. The former is given by

$$\left\|H_{ij}(f_k)\right\|e^{j\angle H_{ij}(f_k)}, k \in [1, N_S], i \in [1, N_t], j \in [1, N_r] \tag{2}$$

where $H_{ij}(f_k)$ is a complex number, which denotes the CSI corresponding to the subcarrier k whose carrier frequency is f_k. $\left\|H_{ij}(f_k)\right\|$ and $\angle H_{ij}(f_k)$ denote amplitude and phase, respectively. i and j are the index of TX and RX antennas, respectively. N_t and N_r stand for the number of antennas at the TX and RX, respectively. N_s represents the number of subcarriers for each pair of transceiver antennas.

2.2. Data Acquisition

To thoroughly analyze the challenges in Wi-Fi-based human activity recognition and evaluate the performance of the method proposed in this paper, we built a dataset in an office environment. The data collection scene is shown in Figure 1. Specifically, Linux 802.11n CSI Tool based on Intel 5300 Network Interface Card (NIC) is leveraged to acquire the raw CSI data [36]. The TX and RX work on 802.11n and operate on a 5 GHz frequency band with a bandwidth of 20 MHz. They are both equipped with three antennas. In addition, 30 subcarriers from each TX-RX pair can be obtained. Thus, there are $3 \times 3 \times 30$ subcarriers in total. We can only use the signal collected from part of the antennas. The data transmission rate is 200 frames/s. We can also subsample the signal measurement to verify the performance of the system at low sampling rates.

Figure 1. Data collection experimental scene in the office.

Table 1 shows the predefined four activities conducted by six volunteers (five males and one female), whose ages range from 23 to 30. We collected the data in a cluttered office environment with lots of tables, chairs, and experimental facilities. The room size is approximately 6 m × 8 m. The distance between the antennas of the TX and the RX is 4 m, and the antennas were both fixed at 1.2 m above the floor. The samples are collected at 24 different locations within a region between the transceivers. The specified location layout is given in Figure 2. The distance between adjacent positions is approximately 0.6 m. We collect 50 samples for each activity at each location for each person. Since the initial sampling rate is 200 frames/s, and the actual duration of the actions is 3.5∼4 s, namely 700∼800 frames, we take 750 frames as a sample.

Table 1. Predefined activities.

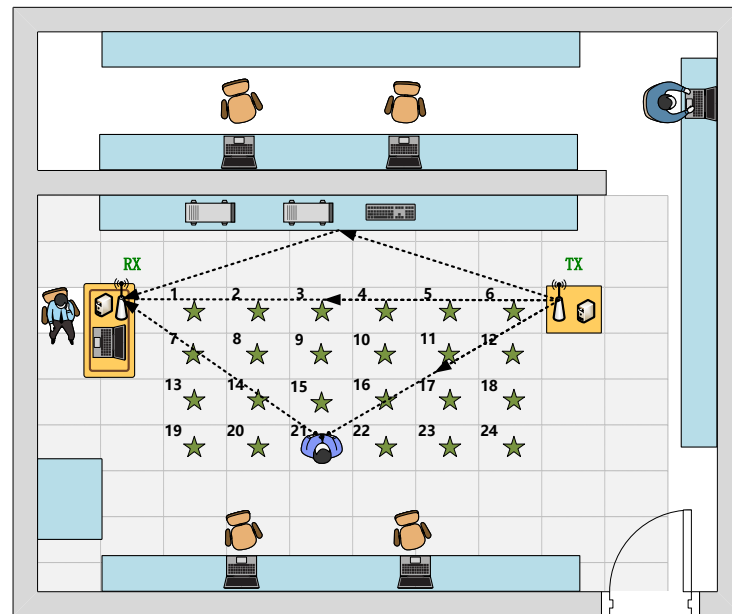| Mark	Activity
O	Draw a circle with right hand
X	Draw a cross with right hand
PO	Lift up and lay down two arms
UP	Push and open with two arms

Figure 2. The layout of data collection locations.

To further demonstrate the influence of activities at different locations on the transmitted signal, we also conducted some experiments in a half-wave anechoic chamber. It is a six-sided box with a shielded design, covering the electromagnetic wave absorbing material inside except for the floor. It simulates an ideal open field situation, in which the site has an infinitely large, well conductive ground plane. In a semi-anechoic chamber, since the ground is not covered with absorbent material, the reflected path will be existing, so that the signal received by the receiving antenna will be the sum of the direct and reflected path signals. More importantly, without the influence of the environment and the other same frequency wireless interference, the same activity conducted by the same person at different locations can effectively reflect the characteristics affected by the locations. The data collection scene is shown in Figure 3. Four activities in Table 1 are conducted by one person. The distance between the antennas of the TX and the RX is 3 m, and the antennas were both fixed at 1.1 m above the floor. The samples are collected at five different locations whose coordinates are (0, 0), (0.6 m, 0.6 m), (0.6 m, −0.6 m), (−0.6 m, −0.6 m), (−0.6 m, 0.6 m). (0,0) is the midpoint of the line between the TX and the RX.

2.3. Problem Analysis

To illustrate the issues and challenges of location-independent human activity recognition using Wi-Fi signals, we comprehensively analyze the CSI measurements involving different human activities at distinct locations collected in the office and anechoic chamber.

As shown in Figure 4, at a fixed location in both two environments, CSI amplitudes of the received signal for four different activities own different waveforms, leading to diverse characteristic patterns. Furthermore, it can be observed that the two different samples of the same activity seem to have a very similar variation tendency. These are the fundamentals of wireless sensing.
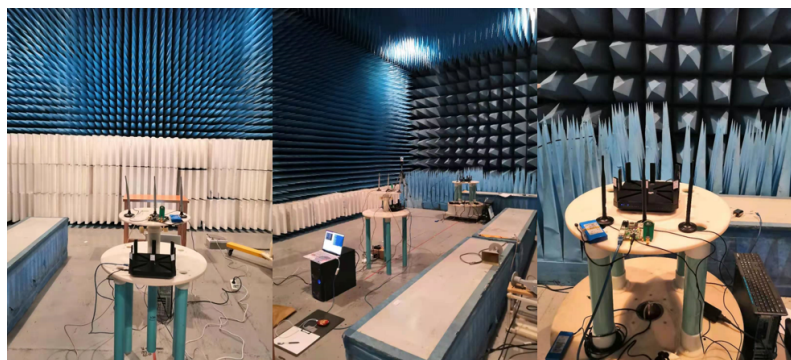
Figure 3. Data collection experimental scene in the anechoic chamber.

As illustrated in Figure 5, the measured signals possess varying CSI amplitudes for the same activity at different locations. Particularly in the anechoic chamber, other variables were eliminated as far as possible except for the locations, which more clearly reflects the influence of different positions on the signal transmission. As can be seen, although it is relatively easy to identify the categories of human activities by translating the CSI patterns at a single location, it may not be possible to ensure good classification accuracy for location-independent sensing. A practicable solution is to minimize the distance of the same activity in different locations, while maximizing the distance between different actions, and apply this learned metric relationship to the target domain. For this reason, a metric learning-based approach is selected for location-independent human activity recognition.

In order to further explore the influence of activities on signal transmission, we illustrate the distinction of the signal between the empty environment and the activity-influenced environment. The three-dimensional maps of the signal are shown in Figure 6, which indicate that the fluctuation of the signal in an empty environment and activity-influenced environment. Each point on the stereogram represents the amplitude of signal corresponding to the frame and subcarrier. From the figure, we can see a higher level of chaos in the three-dimensional waveform of the activity-influenced environment than the empty environment.
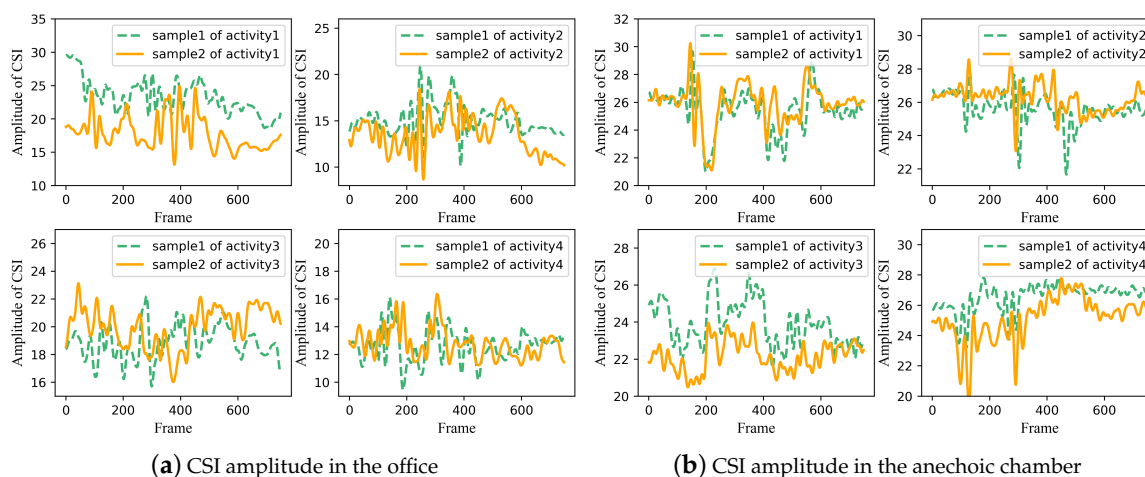

(**a**) CSI amplitude in the office

(**b**) CSI amplitude in the anechoic chamber

Figure 4. CSI amplitude of four different activities at the same location in two experimental scenes. (**a**) CSI amplitude in the office. (**b**) CSI amplitude in the anechoic chamber. Two curves in each subgraph are two samples for the same activity. The horizontal axis of each subgraph represents the frame, the ordinate of each subgraph indicates amplitude of CSI.
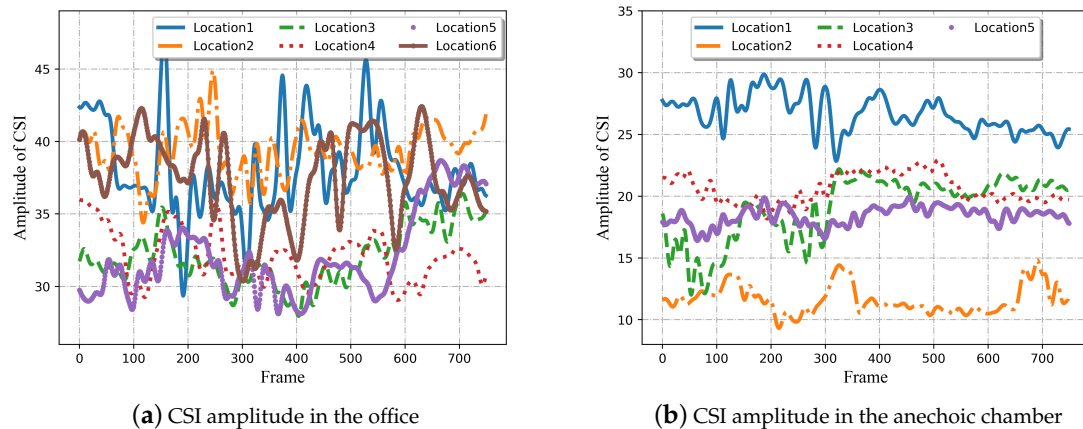
(**a**) CSI amplitude in the office

(**b**) CSI amplitude in the anechoic chamber

Figure 5. CSI amplitude of the same activity at different locations in two experimental scenes. (**a**) CSI amplitude in the office. (**b**) CSI amplitude in the anechoic chamber. Each curve in each subgraph represents an activity sample at one location.
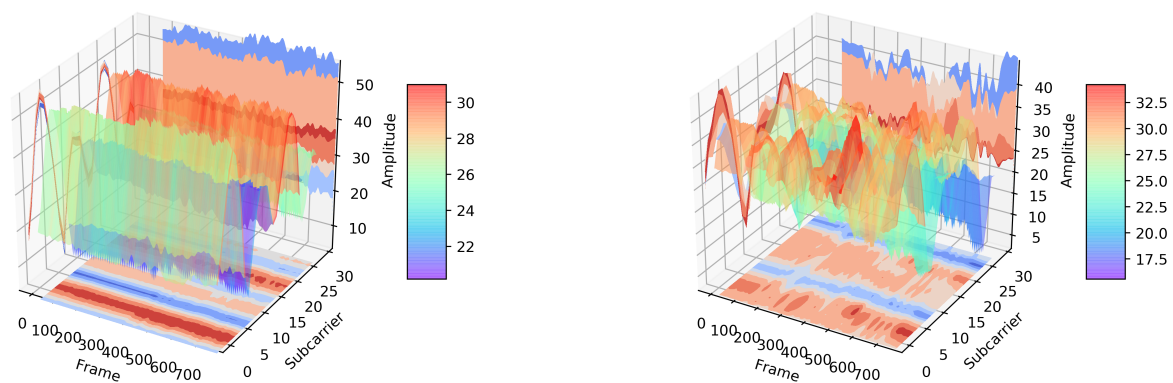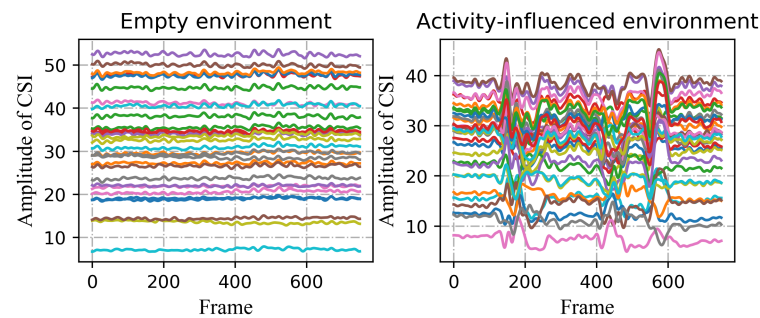

(**a**) Three-dimensional map of signal in empty environment.

(**b**) Three-dimensional map of signal in activity-influenced environment

Figure 6. Three-dimensional map of the signal in empty environment and activity-influenced environment. (**a**) Three-dimensional map of signal in empty environment. (**b**) Three-dimensional map of signal in activity-influenced environment. The three coordinate axes are frame-axis, subcarrier-axis, and amplitude-axis, respectively. The three-dimensional waveform can be mapped to three planes, including the planes parallel to the subcarrier-axis and frame-axis, and perpendicular to the amplitude-axis.

To demonstrate the difference more clearly, Figure 7 shows the two-dimensional maps corresponding to the two vertical planes in Figure 6. As can be seen in Figure 7a, compared with the activity-influenced environment, the amplitude of each subcarrier is almost constant in the empty environment. In other words, the signal waveform changes smoothly with time when there is no human activity interference, while it changes obviously when the signal transmission is affected by human activity. In addition, the activity has a great influence on some subcarriers and a relatively small influence on others.

In Figure 7b, we name the curves channel waveforms, which could reflect the channel state to some extent, revealing the states of each subcarrier. The curve will change with the influence of the activity and the surrounding environment, such as other signal sources, interior layout, and furnishings, especially obstacles on the line-of-sight path. In the left figure, the amplitude of each subcarrier is almost unchanged within 3.5 s, while, in the right figure, the amplitude of each subcarrier varies to different degrees. In each environment, there is a basic channel waveform describing the channel situation (shown as the subgraph on the left of Figure 7b). After being affected by human activity, the curve generates an

additional perturbation based on the basic waveform (shown as the subgraph on the right side of Figure 7b). The thickness of the whole curve represents the fluctuation degree of CSI subcarriers, which shows the extent to which human activity and the surrounding environment affect the transmission of signals. Therefore, we should pay more attention to the added activity-related changes. Deep learning methods can be used to extract action-specified characteristics.

(**a**) Signal waveforms in empty environment and activity-influenced environment.
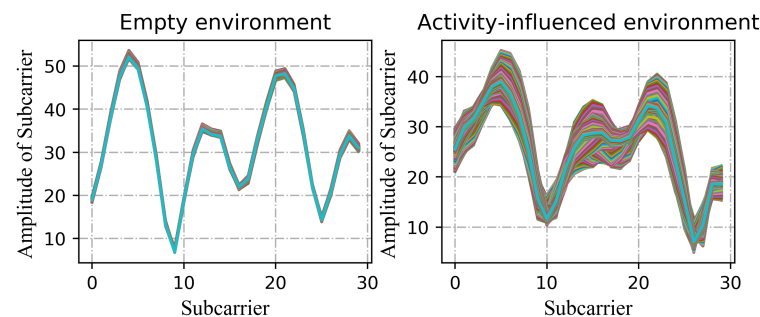
(**b**) Channel waveforms in empty environment and activity-influenced environment.

Figure 7. Two-dimensional map of signal in empty environment and activity-influenced environment. (**a**) The horizontal axis represents the frame/packet, the ordinate indicates the amplitude of CSI. Each curve in the figure represents one of the 30 subcarriers; (**b**) the horizontal axis represents the subcarrier index, the ordinate indicates the amplitude of the subcarriers. Each curve in the figure represents one of the 750 curves, which illustrate the amplitude change of each subcarrier within 3.5 s (The sampling rate is 200 frames/s).

We also investigate the CSI measurements in different TX-RX antenna pairs. As shown in Figure 8, we can see it intuitively, the three subgraphs of each row vary largely, and the three subgraphs of each column are similar in amplitude changes, with a horizontal shift, which can be explained by the phase shift caused by the delay of different transmitting antennas arriving at the receiving antenna. Therefore, the information carried by $1 \times 3 \times 30$ subcarriers from one transmit antenna and three receive antennas is enough for a sample description. Although more subcarriers cover richer information, it is more desirable to extract sufficient activity characteristics from only one transceiver antenna pair, which can effectively reduce computing costs and obviate the need for the number of antennas. In this paper, we hope that the proposed method can be applied to data samples with a small number of antennas and subcarriers.

In this part, we study the signal affected by human activity with different sampling rates. As shown in Figure 9, as the sampling rate decreases, the signal becomes smoother. It may remove some of the noise, but, more importantly, it will lose some of the details referring to the activity. In this paper, while realizing the location-independent human activity recognition, we try our best to ensure the sensing performance of the data samples with a small sampling rate.
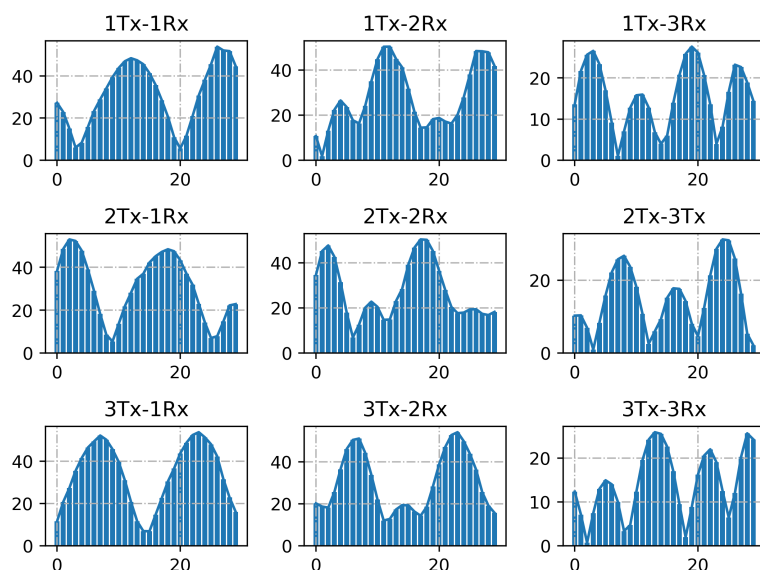
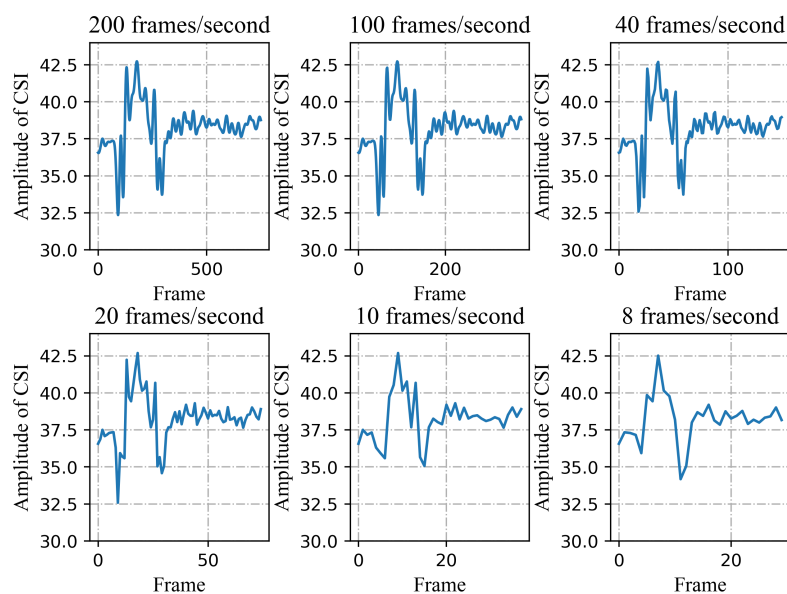**Figure 8.** Amplitude of subcarrier of nine TX-RX antenna pairs.

Figure 9. CSI Amplitude of the human activity with different sampling rates.

3. WiLiMetaSensing

In this section, we provide a detailed introduction to the proposed WiLiMetaSensing system. We first present the system overview. Then, a CNN-LSTM-based feature representation method is described. Finally, an attention mechanism enhanced metric learning-based human activity recognition method is presented.

3.1. System Overview

The workflow of the location-independent human activity recognition system WiLiMetaSensing is shown in Figure 10, which mainly consists of four parts, including data collection, data preprocessing, feature representation, and model training/testing. In the data collection phase, we collect the raw CSI measurements, which describe the changes in the environment. In the data preprocessing step, the amplitude is calculated by the raw complex CSI. Due to the noisy raw data, a 5-order lowpass Butterworth filter is utilized for denoising. Beyond that, the collected data are divided into samples with the size of time × subcarrier, which indicates the number of frames corresponding to an activity multiplied by

the number of subcarriers. Then, we map the data samples to high dimensional embedding space to fulfill the feature representation through CNN and LSTM. Finally, in order to achieve location-independent perception with as few samples as possible, regarding a few-shot learning problem, the human activity perceptive method based on metric learning is proposed. Subsequently, we will introduce the system in detail.
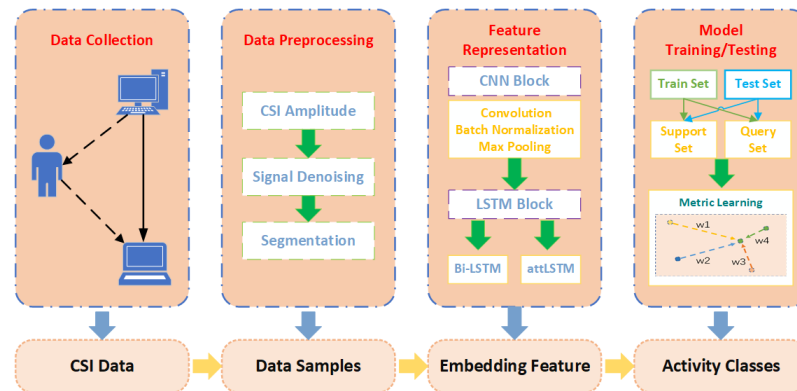

Figure 10. The workflow of WiLiMetaSensing.

3.2. CNN-LSTM-Based Feature Representation

In this section, in order to extract activity-specified and location-independent features from input samples for few-shot learning, deep learning methods, including CNN and LSTM, are introduced for feature representation shown as Figure 11. Following the learning strategy of meta learning, the data samples are divided into two parts, including the support set and the query set with the same data selection strategy, which will be presented in detail in the next section.

We use x_i and $\hat{x}$ to denote the samples from the above two sets. $S = \{x_i\}, i \in 1, \ldots, n \times k$ indicates the support set which is made up of samples from n categories, and k samples for each class. $g(x_i, S)$ and $f(\hat{x}, S)$ are modeled to achieve feature representation of x_i and $\hat{x}$ fully conditioned on the support set, respectively.

The feature embedding function $g(x_i, S)$ for each sample x_i can be expressed as:

$$g'(x_i) = CNN(x_i) \tag{3}$$

$$\overrightarrow{h_i}, \overrightarrow{c_i} = \overrightarrow{LSTM}(g'(x_i), \overrightarrow{h}_{i-1}, \overrightarrow{c}_{i-1}) \tag{4}$$

$$\overleftarrow{h_i}, \overleftarrow{c_i} = \overleftarrow{LSTM}(g'(x_i), \overleftarrow{h}_{i-1}, \overleftarrow{c}_{i-1}) \tag{5}$$

$$g(x_i, S) = \overrightarrow{h_i} \oplus \overleftarrow{h_i} + g'(x_i) \tag{6}$$

The samples are first mapped to high-dimensional embedding space through CNN to capture the feature in subcarrier and time dimensions. Specifically, the embedding model is made up of a cascade of blocks, each including a convolutional layer, a batch normalization layer, and a MaxPooling layer, followed by a fully-connected layer. The activation function is a rectified linear unit (ReLU).

The samples embedded by CNN form a sequence, which serves as the input of bidirectional long short-term memory (Bi-LSTM). It consists of a forward propagation LSTM and a backward propagation LSTM. The basic structure of LSTM is shown in Figure 12, which consists of three control gates, including an input gate i_t, a forget gate f_t, an output gate o_t. In addition, a memory cell c_t and a hidden unit h_t are also significant components. With the current input x_t, the hidden state h_{t-1}, and cell state c_{t-1} at time $t - 1$, the LSTM parameters at timestep t can be calculated as follows:

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \tag{7}$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{8}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{9}$$

$$\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_C) \tag{10}$$

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \tag{11}$$

$$h_t = o_t \times \tanh(C_t) \tag{12}$$

where W_f, W_i, W_o are the weight and b_f, b_i, b_o are the bias of the three gates. σ and tanh denote sigmoid and hyperbolic tangent activation functions, respectively. $\times$ stands for the element-wise multiplication.
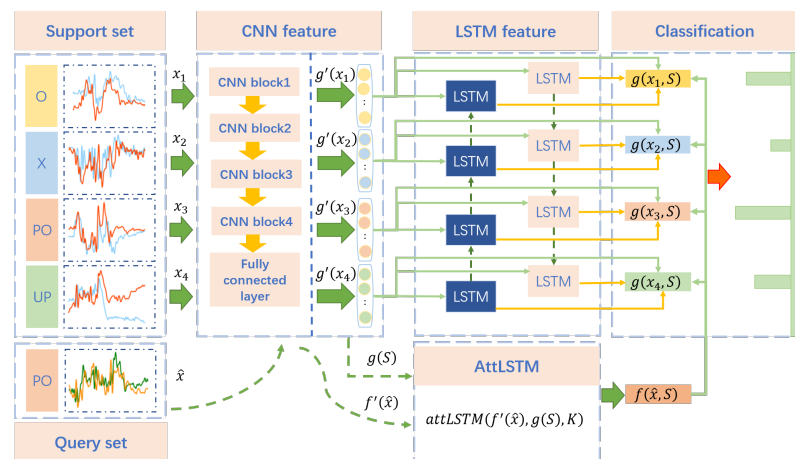

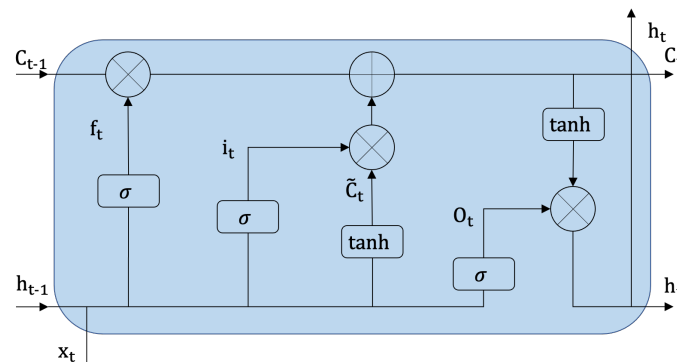**Figure 11.** The architectures of human activity recognition method.

Figure 12. The structure of the LSTM cell.

Through the forget gate, the previous memory cell can be selectively forgotten. The input gate controls the current input, while the output gate determines how the memory unit is converted to a hidden unit. However, the LSTM network processes the sequential data in one direction resulting in only partial categories of features that can be utilized. Therefore, the Bi-LSTM is leveraged to merge the information from two directions of the sequence. The final hidden vector of the Bi-LSTM at the $t - th$ moment can be expressed as:

$$h_t = \vec{h}_t \oplus \overleftarrow{h}_t \tag{13}$$

where $\oplus$ is the concatenation operation, $\vec{h}_t$ and $\overleftarrow{h}_t$ are the outputs (hidden vector) of the forward LSTM and the backward LSTM, respectively.

Through the above CNN-LSTM feature representation, we aim to leverage the common characteristics of different activities to calibrate the high-dimensional embedding of each sample. In other words, in the feature representation of each class sample, the information of other class samples can be used. As we all know, the received CSI measurements

contain not only dynamic activity information but also static environment information and varying location information. Therefore, there are some common features about the background for different samples in each category. We hope that the model can learn and memorize the common characteristics of different types of activities, as well as the distinct information of different categories. The distinct information can be utilized to increase the distance of inter-class, and reduce the distance of intra-class.

The embedding function $f(\hat{x}, S)$ for a query sample $\hat{x}$ is defined as follows:

$$f(\hat{x}, S) = attLSTM(f'(\hat{x}), g(S), K) \tag{14}$$

where f' is a neural network, the same as g'. K denotes the number of "processing" steps following work from Vinyals et al. [37]. $g(S)$ represents the embedding function g applied to each element x_i from the set S. Thus, the state after k processing steps is as follows:

$$\hat{h}_k, c_k = LSTM(f'(\hat{x}), [h_{k-1}, r_{k-1}], c_{k-1}) \tag{15}$$

$$h_k = \hat{h}_k + f'(\hat{x}) \tag{16}$$

$$r_{k-1} = \sum_{i=1}^{|S|} a(h_{k-1}, g(x_i)) g(x_i) \tag{17}$$

$$a(h_{k-1}, g(x_i)) = soft \max\left(h_{k-1}^T g(x_i)\right) \tag{18}$$

Noting that the $LSTM(x, h, c)$ in both g and f follows the same LSTM implementation defined by Sutskever et al. [38].

3.3. Metric Learning-Based Human Activity Recognition

Our location-independent activity recognition task can be described as a few-shot learning problem and a meta learning task. Meta learning trains the model from a large number of tasks and learns faster on new tasks with a small amount of data. Unlike the traditional meta learning and few-shot learning methods, which apply the model learned from some classes (source domain) to the other new classes (target domain) with very few samples from the new classes, our work is intended to utilize the model to the data with the same label, but with different data distribution.

Meta learning includes training process and testing process, which is called meta-training and meta-testing. In our task, samples in part of locations are selected as the source domain data, while samples from other locations are the target domain data. Both the source domain data and the target domain data are classified into the support set and query set with the same data set selection strategy.

Assuming that there is a source domain sample set S with n classes, and a target domain set T with the same n classes. We randomly select support sets $S' = \{(x_i, y_i)\}_{i=1}^{n \times m}$ and $T' = \{(x_i, y_i)\}_{i=1}^{n \times k}$, query sets $S'' = \{(\hat{x}, \hat{y})\}_{i=1}^{n \times l}$ and $T'' = \{(\hat{x}, \hat{y})\}_{i=1}^{n \times t}$ from S and T datasets. m and l, k, and t are the number of samples picked from each class of source domain and target domain, respectively. This is the so-called k-shot learning. More precisely, leveraging the support set S' from the source domain, we learn a function which can map test samples $\hat{x}$ from S'' to a probability distribution $P(\hat{y}|\hat{x}, S')$ over outputs $\hat{y}$. P is a probability distribution parameterized by a CNN-LSTM feature representation neural network and a classifier. In the target domain, when a new support set T' is given, we can simply use the function P to make a prediction $P(\hat{y}|\hat{x}, T')$ for each test sample $\hat{x}$ from T''. In short, we predict the label $\hat{y}$ for the unseen sample $\hat{x}$ and a support set S' can be expressed as:

$$\hat{y} = \arg \max_y P(y|\hat{x}, S') \tag{19}$$

A simple method to predict $\hat{y}$ is calculating a linear combination of the labels in the support set as follows:

$$\hat{y} = \sum_{i=1}^{N} a(\hat{x}, x_i) y_i \tag{20}$$

where a is an attention mechanism which is shown as:

$$a(\hat{x}, x_i) = \frac{e^{c(f(\hat{x}), g(x_i))}}{\sum_{j=1}^{N} e^{c(f(\hat{x}), g(x_j))}} \tag{21}$$

It is softmax over the cosine similarity c of the embedding functions f and g, which are the feature representation neural network. In addition, the cosine similarity is calculated as:

$$c(f(\cdot), g(\cdot)) = \cos(f(\cdot), g(\cdot)) = \frac{f(\cdot) \cdot g(\cdot)}{\|f(\cdot)\| \|g(\cdot)\|} \tag{22}$$

The training procedure is an episode-based training, which is a form of meta-learning, learning to learn from a given support set to minimize a loss over a batch. More specifically, we define a task T as a distribution over possible label sets L (four activities in our experiment). To form an "episode" to compute gradients and update our model, we first sample L from T (e.g., L could be the label set X, O, PO, UP). We then use L to sample the support set S and a batch B (i.e., both S and B are labelled examples of X, O, PO, UP). The network is then trained to minimize the error predicting the labels in the batch B conditioned on the support set S. More precisely, the training objective is as follows:

$$\theta = \arg\max_{\theta} E_{L \sim T} \left[E_{S \sim L, B \sim L} \left[\sum_{(x,y) \in B} \log P_{\theta}(y|x, S) \right] \right] \tag{23}$$

where θ represents the parameters of the embedding function f and g.

4. Evaluation

In this section, we evaluate the performance of the proposed WiLiMetaSensing system through extensive experiments. The evaluation contains the following three parts. Firstly, we explore the feasibility and effectiveness of our system. Then, we investigate the system Modules. Finally, the robustness of the system is discussed by demonstrating the influence of different data samples.

4.1. Experiment Setup

We first evaluate the performance of our sensing method in a traditional way, including the single location sensing and the mixed locations sensing. In addition, we validate the effectiveness of location-independent sensing. There are 50 samples for each activity at each location for each person, 60% of which are randomly selected as the training set, 20% as the validation set, and the rest as the testing set. For single location sensing, we train and test at the same location. For mixed locations sensing, we apply the activities of all the locations for training and testing. For location-independent sensing, we show the overall average accuracy with four locations for training and 24 locations for testing. In this section, we show the overall accuracy for one-shot learning using the samples with 200 frames/s sampling rate, which lasts for 3.5 s, and 90 subcarriers. According to the training strategy of meta learning method, when we test for k-shot learning, we set the number of samples in each category of the support set as k for the testing sets. We set the support set of training and validation sets the same as the testing sets.

Specifically, the CNN embedding module consists of four CNN blocks, each including a convolutional layer, a batch normalization layer, and a 2×2 max-pooling layer, followed by a fully-connected layer with 64 neurons. In addition, 64 filters with the kernel size 3×3 are used. In the Bi-LSTM embedding module, the number of hidden units is $n * k$, which

is the number of activities multiplied by the *k*-shot. The input size of Bi-LSTM is decided by the dimension of a fully-connected layer which is 64. The number of hidden layers is 1. Hidden size (the dimension of the hidden layer) is 32, while, in attLSTM, it is 64. We minimize the cross-entropy loss function with Adam to optimize the model parameters. The exponential decay rate $\rho1$ and $\rho2$ are empirically set as 0.9 and 0.999. The learning rate is set as 0.0001. The total number of training iterations is 300. The batch size is set as 16. Unless otherwise specified, the following evaluations follow the above settings.

4.2. Overall Performance

Table 2 illustrates the recognition average accuracy of our method compared with the traditional deep learning method CNN and WiHand [25]. WiHand is based on the low rank and sparse decomposition (LRSD) algorithm and extracts the histogram of the gesture CSI profile as the features, which outperforms the other location-independent approach. It can be seen that our system outperforms these two methods in both location-dependent sensing and location-independent sensing. All the methods can recognize with high accuracy for single location sensing and mixed locations sensing. For the location-independent sensing, WiLiMetaSensing can also obtain an average 91.11% recognition accuracy, which is about 7% higher than CNN, and about 9% higher than WiHand. Specifically, the confusion matrix of a test for our location-independent human activity recognition method is shown in Figure 13 with a 91.41% accuracy. We can see that all of the activities can be recognized with high accuracy. Note that Table 2 shows the optimal recognition accuracy of WiHand with 30 subcarriers and 20 features. We analyze the reason why WiHand did not perform as well as the original dataset, including (1) The nine data collection locations of WiHand are relatively close to the TX and RX, while our 24 locations have a wider coverage. (2) The sampling rate of WiHand is 2500 packets/s, which is much larger than our 200 packets/s. (3) WiHand could extract CSI streams of all 56 subcarriers from the customized drivers, while ours is 30 subcarriers. A higher sampling rate and more subcarriers may provide richer fine-grained information. After the matrix decomposition, more activity-related information will be preserved.

Table 2. The recognition accuracy for single location sensing, mixed locations sensing, and location-independent sensing.

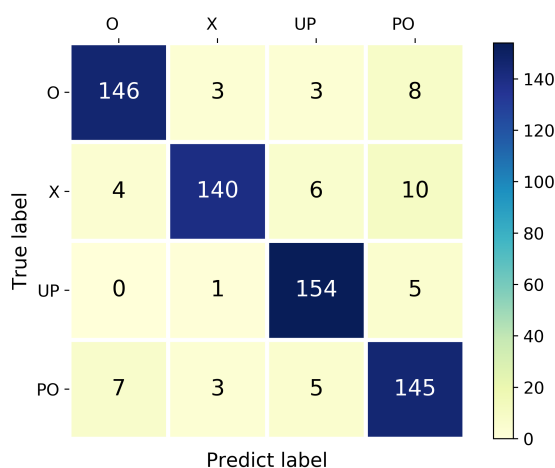| Accuracy (%)	WiLiMetaSensing	CNN	WiHand [26]
Single location	99.13	99.00	96.15
Mixed locations	98.36	95.53	91.50
Location-independent	91.11	84.02	82.20

Figure 13. The confusion matrix of location-independent human activity recognition.

4.3. Module Study

Comparison with different feature representation modules. In this section, we explore the effect of the embedding module g (Bi-LSTM) and f (attLSTM) for the samples from the support set and the query set. We test for one-shot learning using the samples with 90 subcarriers. From Table 3, we can see that both modules enhance the performance of the method. Leveraging all the activity samples from the support set, common features can be obtained to adjust the feature representation, so as to pay more attention to the location-independent features. The embedding module for the query set enables the sample in the source domain to effectively calibrate the feature representation of the sample in the target domain.

Table 3. The recognition accuracy with different embedding modules.

Embedding Modules	Accuracy (%)
WiLiMetaSensing with Bi-LSTM with attLSTM	91.11
WiLiMetaSensing with Bi-LSTM without attLSTM	90.25
WiLiMetaSensing without Bi-LSTM without attLSTM	88.73

4.4. Robustness Evaluation

Performance of location-independent sensing in terms of different number of training locations. The activity samples of each position have different data distributions. The further the distance of the locations, the higher the probability of a broader distribution distance will be. Therefore, when it comes to the samples collected for training the models, we hope the positions of the training samples become more decentralized. We adopt a fixed training position selection strategy, in which the positions should be distributed as far as possible in the entire space, instead of clustering together in a line parallel to the transceiver. We choose 4/6/8/12/24 locations for training and 24 locations for testing. The selections of fixed 4/6/8/12 training positions are depicted in Figure 14. 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 samples. For six training locations, the straight lines or broken lines together with the same colored marked locations form the training pairs.
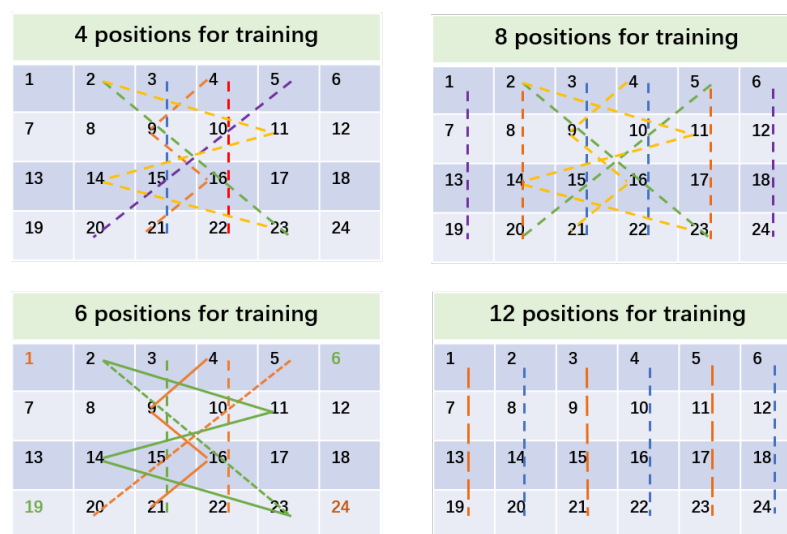

Figure 14. The layout of the training locations.

As demonstrated in Table 4, when we pick four training locations and 1-shot, the accuracy is 91.11%. When eight training locations and 1-shot are selected, the accuracy is 92.66%. The results indicate that the more training positions there are, the higher accuracy the recognition obtains.

Table 4. The average recognition accuracy for different numbers of training locations.

Number of Training Locations	4	6	8	12	24
Accuracy (%)	91.11	92.23	94.98	96.00	98.36

Performance of location-independent sensing for samples with different numbers of subcarriers. We explore one-shot human activity recognition with different numbers of subcarriers. As illustrated in Table 5, the recognition accuracy reduces with the decrease of the number of subcarriers. However, it still maintains an acceptable recognition rate when there are only 30 subcarriers from one pair of antenna.

Table 5. The accuracy for different number of subcarriers with four training locations.

Training Locations	90 Subcarriers	60 Subcarriers	30 Subcarriers
3,8,15,20	92.63	91.75	86.75
1,10,13,22	91.25	88.50	87.50
2,8,14,20	91.75	90.75	87.75
3,9,15,21	89.13	89.00	85.13
1,8,15,22	91.25	89.50	85.75
4,9,14,19	90.63	90.13	85.00
Average accuracy (%)	91.11	89.94	86.31

Performance of location-independent sensing for different TX-RX antenna pairs. We investigate the recognition accuracy with 30 subcarriers from different TX-RX antennas. As shown in Table 6, different antenna pairs have similar recognition effects. The difference reflects that different antenna pairs contain more or less diverse information. Therefore, 90 subcarriers which integrate these features can obtain superior results. Note that, in Table 6, iTX-jRX represents CSI data from i-th TX and j-th RX.

Table 6. The accuracy for different TX-RX antenna pairs.

TX-RX	1TX-1RX	1TX-2RX	1TX-3RX
Accuracy (%)	86.31	87.00	85.60

Performance of location-independent sensing for different number of shots. We explore the number of samples in support set for testing. As examples, we also select four locations for training and 24 locations for testing. The samples with 90 subcarriers are used. The identification results are listed in Table 7. It is noted that all the average accuracy is above 90%, and the accuracy will increase with the growth of the sample size.

Table 7. The accuracy for different number of shots with four training locations.

Number of Shots	1-Shot	2-Shot	3-Shot
Accuracy (%)	91.11	92.25	93.21

Performance of location-independent sensing for samples with different sampling rates. We collect CSI measurements at the initial transmission rate of 200 packets/s, and down-sample the 750 CSI series to 375, 250, 150, 75. The one-shot results with different sampling rates are shown in Figure 15. As can be seen, when he sampling rate decrease to 20 frames/s, the method can still obtain satisfying accuracy.
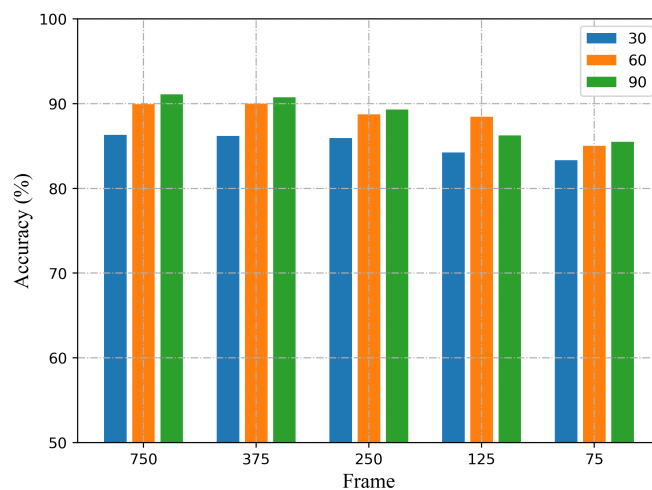
Figure 15. The recognition accuracy with different number of subcarriers and sampling rates.

5. Limitations and Future Work

Although the proposed WiLiMetaSensing system realizes location-independent sensing with very few samples, there remain many challenges to be overcome. First of all, there are some strict restrictions in the data collection process. For example, volunteers are required to perform the same activity facing nearly the same direction in the same room. Consequently, except for the types of activity and location variations, other factors that would affect the transmission of the signals are not seriously taken into account, such as the status of a person (e.g., pose and direction) and the environmental variations (i.e., altering the room or the locations of surrounding objects). In addition, as signals can be easily blocked, reflected, or scattered by different targets, the existence of other people would also result in different signal patterns. However, the impact of interference from the other person on the classification accuracy was either not considered. As a result, in future work, we will further explore the generalized and robust human activity recognition method with an adequate account of these aforementioned factors. Only in this way can human activity recognition technology develop from academic research to industrial application.

6. Conclusions

In this paper, we present a novel human activity recognition system, named WiLiMetaSensing. It realizes location-independent sensing with very few samples in the Wi-Fi environment. Inspired by the idea of meta learning, we endow the system with the ability that can utilize the knowledge acquired from one location for others. Technically, we propose a CNN-LSTM feature representation and metric learning-based human activity recognition system. The model focuses on the common characteristics of different locations and extracts discriminative features for different activities. The performance evaluation is conducted on the comprehensive dataset we build. It demonstrates that the WiLiMetaSensing system can achieve an average accuracy of 91.11%, with four locations for training, given only one sample for other testing locations. More importantly, it can well adapt to the data samples with a small number of subcarriers and a low sampling rate. Therefore, we can firmly conclude that the presented approach is feasible and robust for location-independent sensing.

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

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. Kim, E.; Helal, S.; Cook, D. Human activity recognition and pattern discovery. *IEEE Pervasive Comput.* **2009**, *9*, 48–53. [CrossRef]
3. Lara, O.D.; Labrador, M.A. A survey on human activity recognition using wearable sensors. *IEEE Commun. Surv. Tutor.* **2012**, *15*, 1192–1209. [CrossRef]
4. Wang, J.; Chen, Y.; Hao, S.; Peng, X.; Hu, L. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognit. Lett.* **2019**, *119*, 3–11. [CrossRef]
5. D'Sa, A.G.; Prasad, B. A survey on vision based activity recognition, its applications and challenges. In Proceedings of the 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), Majitar, Sikkim, 25–28 February 2019; pp. 1–8.
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. 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**. [CrossRef]
8. Liu, J.; Teng, G.; Hong, F. Human Activity Sensing with Wireless Signals: A Survey. *Sensors* **2020**, *20*, 1210. [CrossRef]
9. 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]
10. Adib, F.; Kabelac, Z.; Katabi, D.; Miller, R.C. 3D tracking via body radio reflections. In Proceedings of the 11th USENIX Symposium on Networked Systems Design and Implementation (NSDI 14), Seattle, WA, USA, 2–4 April 2014; pp. 317–329.
11. Yang, Z.; Pathak, P.H.; Zeng, Y.; Liran, X.; Mohapatra, P. Monitoring vital signs using millimeter wave. In Proceedings of the 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing, Paderborn, Germany, 4–8 July 2016; pp. 211–220.
12. Lien, J.; Gillian, N.; Karagozler, M.E.; Amihood, P.; Schwesig, C.; Olson, E.; Raja, H.; Poupyrev, I. Soli: Ubiquitous gesture sensing with millimeter wave radar. *ACM Trans. Graph.* **2016**, *35*, 1–19. [CrossRef]
13. Zhong, Y.; Dutkiewicz, E.; Yang, Y.; Zhu, X.; Zhou, Z.; Jiang, T. Internet of mission-critical things: Human and animal classification—A device-free sensing approach. *IEEE Internet Things J.* **2017**, *5*, 3369–3377. [CrossRef]
14. 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]
15. 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]
16. 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]
17. Zhong, Y.; Bi, T.; Wang, J.; Wu, S.; Jiang, T.; Huang, Y. Low data regimes in extreme climates: Foliage penetration personnel detection using a wireless network-based device-free sensing approach. *Ad Hoc Networks* **2021**, *114*, 102438. [CrossRef]
18. 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]
19. Shi, Z.; Zhang, J.A.; Xu, Y.D.R.; Cheng, Q. Environment-Robust Device-free Human Activity Recognition with Channel-State-Information Enhancement and One-Shot Learning. *IEEE Trans. Mob. Comput.* **2020**. [CrossRef]
20. 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**. [CrossRef]
21. Yang, J.; Zou, H.; Zhou, Y.; Xie, L. Learning gestures from wifi: A siamese recurrent convolutional architecture. *IEEE Internet Things J.* **2019**, *6*, 10763–10772. [CrossRef]
22. Wang, Y.; Liu, J.; Chen, Y.; Gruteser, M.; Yang, J.; Liu, H. E-eyes: Device-free location-oriented activity identification using fine-grained wifi signatures. In Proceedings of the 20th Annual International Conference on Mobile Computing and Networking, Seattle, WA, USA, 2–4 April 2014; pp. 617–628.
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. 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, Niagara Falls, NY, USA, 19–23 June 2017; pp. 252–264.

25. Yang, J.; Zou, H.; Jiang, H.; Xie, L. Fine-grained adaptive location-independent activity recognition using commodity WiFi. In Proceedings of the 2018 IEEE Wireless Communications and Networking Conference (WCNC), Barcelona, Spain, 15–18 April 2018; pp. 1–6.

26. Lu, Y.; Lv, S.; Wang, X. Towards Location Independent Gesture Recognition with Commodity WiFi Devices. *Electronics* **2019**, *8*, 1069. [CrossRef]

27. Yue, Z.; Yi, Z.; Kun, Q.; Guidong, Z.; Yunhao, L.; Chenshu, W.; Zheng, Y. Zero-Effort Cross-Domain Gesture Recognition with Wi-Fi. In Proceedings of the 17th Annual International Conference on Mobile Systems, Applications and Services (MobiSys '19), Seoul, Korea, 17–21 June 2019.

28. Jie, Z.; Zhanyong, T.; Meng, L.; Dingyi, F.; Petteri, N.; Wang, Z. Crosssense: Towards cross-site and large-scale wifi sensing. In Proceedings of the 24th Annual International Conference on Mobile Computing and Networking (MobiSys '18), New Delhi, India, 29 October–November 2018; pp. 305–320.

29. Wu, X.; Chu, Z.; Yang, P.; Xiang, C.; Zheng, X.; Huang, W. TW-See: Human Activity Recognition Through the Wall with Commodity Wi-Fi Devices. *IEEE Trans. Veh. Technol.* **2018**. [CrossRef]

30. Zhou, Q.; Xing, J.; Yang, Q. Device-free occupant activity recognition in smart offices using intrinsic Wi-Fi components. *Build. Environ.* **2020**, *172*, 106737. [CrossRef]

31. Jiang, W.; Miao, C.; Ma, F.; Yao, S.; Wang, Y.; Yuan, Y.; Xue, H.; Song, C.; Ma, X.; Koutsonikolas, D.; et al. Towards environment independent device free human activity recognition. In Proceedings of the 24th Annual International Conference on Mobile Computing and Networking (MobiSys '18), New Delhi, India, 29 October–2 November 2018; pp. 289–304.

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. Vilalta, R.; Drissi, Y. A Perspective View and Survey of Meta-Learning. *Artif. Intell. Rev.* **2002**, *18*, 77–95. [CrossRef]

34. Vinyals, O.; Blundell, C.; Lillicrap, T.; Kavukcuoglu, K.; Wierstra, D. Matching Networks for One Shot Learning. *arXiv* **2016**, arXiv:1606.04080.

35. Bellet, A.; Habrard, A.; Sebban, M. A Survey on Metric Learning for Feature Vectors and Structured Data. *arXiv* **2013**, arxiv:1306.6709.

36. Halperin, D.; Hu, W.; Sheth, A.; Wetherall, D. Predictable 802.11 packet delivery from wireless channel measurements. *ACM Sigcomm Comput. Commun. Rev.* **2010**, *40*, 159–170. [CrossRef]

37. Vinyals, O.; Bengio, S.; Kudlur, M. Order Matters: Sequence to Sequence for Sets. *arXiv* **2016**, arXiv:1511.06391.

38. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to Sequence Learning with Neural Networks. *Adv. Neural Inf. Process. Syst.* **2014**. [CrossRef]