applied sciences

MDPI

Article

Location Adaptive Motion Recognition Based on Wi-Fi Feature Enhancement

Wei Shi [1], Meichen Duan [1], Hui He [1,*], Liangliang Lin [1], Chen Yang [1], Chenhao Li [1,2] and Jizhong Zhao [1]

[1] Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China
[2] Ant Rongxin (Chengdu) Network Technology Co., Ltd., Chengdu 610040, China
* Correspondence: huihe@xjtu.edu.cn

Abstract: Action recognition is essential in security monitoring, home care, and behavior analysis. Traditional solutions usually leverage particular devices, such as smart watches, infrared/visible cameras, etc. These methods may narrow the application areas due to the risk of privacy leakage, high equipment cost, and over/under-exposure. Using wireless signals for motion recognition can effectively avoid the above problems. However, the motion recognition technology based on Wi-Fi signals currently has some defects, such as low resolution caused by narrow signal bandwidth, poor environmental adaptability caused by the multi-path effect, etc., which make it hard for commercial applications. To solve the above problems, we first propose and implement a position adaptive motion recognition method based on Wi-Fi feature enhancement, which is composed of an enhanced Wi-Fi features module and an enhanced convolution Transformer network. Meanwhile, we improve the generalization ability in the signal processing stage to avoid building an extremely complex model and reduce the demand for system hardware. To verify the generalization of the method, we implement real-world experiments using 9300 network cards and the PicoScenes software platform for data acquisition and processing. By contrast with the baseline method using original channel state information(CSI) data, the average accuracy of our algorithm is improved by 14% in different positions and over 16% in different orientations. Meanwhile, our method has best performance with an accuracy of 90.33% compared with the existing models on public datasets WiAR and WiDAR.

Keywords: human motion recognition; channel state information; multi-signal classification algorithm; wireless perception

Citation: Shi, W.; Duan, M.; He, H.; Lin, L.; Yang, C.; Li, C.; Zhao, J. Location Adaptive Motion Recognition Based on Wi-Fi Feature Enhancement. *Appl. Sci.* **2023**, *13*, 1320. https://doi.org/10.3390/app13031320

Academic Editor: Eui-Nam Huh

Received: 19 December 2022
Revised: 13 January 2023
Accepted: 16 January 2023
Published: 18 January 2023

1. Introduction

Due to the popular concept of artificial intelligence (AI) technology and the rapid development of intelligent equipment and the Internet of Things (IoT), indoor human action recognition technology is widely used in many fields, such as human–computer interaction to improve production and life efficiency [1–3], nursing [4], and monitoring [5,6].

Human action recognition techniques can be broadly divided into two categories. One of them is bound motion recognition, which requires individuals to carry radio frequency identification (RFID) tags [7,8], sensor devices [9], or other special devices [10–12]. These methods indeed have high accuracy, while they cannot be applied to some confidential scenes that prohibit carrying devices. The other is unbound motion recognition, which has certain universality, such as using wireless signals to recognize human actions. Due to the existence of communication modules in intelligent equipment, wireless signals have the advantages of wide distribution, high signal strength in indoor environments, excellent penetration, and avoid the risk of personal privacy disclosure. Wi-Fi signal is the most widely distributed wireless signal [13–15]. According to the changes of CSI or other data in the received Wi-Fi signal, the human body and its actions in the environment can be analyzed and identified [15–17].

There are also many techniques on action recognition based on Wi-Fi. The Multiple Signal Classification [18] (MUSIC) algorithm is mainly used to estimate the incidence angle

of source signals in space and traverses the spectral function values corresponding to the incidence angles. Then, the maximum value of them can be used to deduce the source position. However, the MUSIC algorithm is mainly used in the field of radar direction finding, which requires a huge antenna array for accurate estimation. For common Wi-Fi devices, the number of antennas and bandwidth are limited, and the estimation accuracy will be greatly reduced when applied to human body location scenes due to the multi-path effect. Moreover, among all the parameters of the signal affected by human actions (Time of Flight, Angle of Arrival, Doppler Frequency Shift (DFS) and Received Signal Strength Information (RSSI)), Doppler Frequency Shift reflects the human action speed information best [19]. When the human body is running, walking, or any other actions, the movement distance is usually more than 2m, and DFS changes significantly. Nevertheless, for the actions with small displacement such as standing up and sitting down, DFS changes slightly, which makes it difficult to estimate the velocity distribution of such actions.

From above considerations, a position-adaptive motion recognition method based on Wi-Fi feature enhancement is proposed to solve the multi-path effect under constrained conditions and improve the generalization ability of Wi-Fi motion recognition. Taking the human body as the reference, the three-dimensional velocity distribution independent of the human body's position is extracted from the Doppler frequency shift as the feature to enhance its generalization performance of the position. In the action classification stage, a dynamic convolution transformer network is designed to realize action classification, which is used to improve the ability to extract local features and classify different actions. Ultimately, we implement improving motion classification performance through Wi-Fi signal with good generalization ability and high accuracy.

Our main contributions can be summarized as follows:

(1) Based on the MUSIC algorithm, a novel method adding time of flight (ToF) and the offset parameters is proposed to improve the positioning accuracy to obtain an accurate human position. We design a dynamic signal amplification method referring to the Fresnel model to strengthen the influence of human motion to obtain DFS. By changing the frame of reference, we establish the corresponding relationship between DFS and speed component and extract the speed distribution independent of position as the motion characteristics. Overall, we improve the generalization ability in the signal processing stage to avoid building an extremely complex model and reduce the training cost and the demand for system hardware.

(2) In this paper, we propose a dynamic convolution transformer network. Based on the Transformer model, the human action speed distribution is taken as the input to extract the features. In addition, Gaussian range coding is introduced to retain the timing information and reduce the feature differences caused by individual factors. The adaptive capture improves the local feature extraction ability and the classification effect of different actions, and it realizes action recognition with good generalization and high accuracy.

(3) We first propose and implement position adaptive motion recognition based on Wi-Fi feature enhancement, and it has excellent generalization ability in different scenarios, positions, and orientations. We also conduct experiments on WiAR and WiDAR datasets, and compare our method with Widar 3.0, EI, and CARM. Experimental results show that our method outperforms the compared methods in terms of efficiency and accuracy.

The rest of the paper is organized as follows. We first introduce the related work about the main action recognition methods. Then Section 3 introduces the proposed method to enhance Wi-Fi features. The following section provides the enhanced convolution Transformer network to classify actions. Finally, the experiments and conclusions are presented.

2. Related Work

Currently, Internet of Things (IoT) has been applied to sensors, devices and software to enhance the performance. One of the most important components for IoT working is protocols and standards, and they can be divided into data protocols (WebSocket, Hyper Text Transfer Protocol and Direct Digital Frequency Synthesis) and network protocols

(Wi-Fi, Bluetooth and ZigBee). Given that IoT system is vulnerable to attacks, some latest research [20,21] focuses on the security measures. Because of the limited capabilities of 5G, some research has put forward new advancement in next generation mobile wireless communication (6G), especially connected Intelligence [22,23]. Furthermore, indoor human motion recognition has become a popular field in recent research [24] combing IoT, smart devices, and AI. Meanwhile, with the coverage of optical cables and the popularity of indoor routers, Wi-Fi signals widely exist indoors. Furthermore, water accounts for about 70% of the human body. Thus, the human body has good reflectivity to indoor Wi-Fi signals, and action recognition can be realized by distinguishing the varying signals and corresponding them to human actions [25]. According to the different types of information collected, recognition technology using Wi-Fi signals can be divided into two categories: motion recognition based on RSSI and motion recognition based on CSI.

Because of the high sensitivity to moving objects, some research works use RSSI information to image different moving objects [26] and RSSI for human motion recognition [27,28]. RSSI will no longer decrease monotonically with the increase in propagation distance due to small-scale shadow fading caused by signal multi-path propagation in an indoor environment, which limits the accuracy of measurement. CSI contains amplitude and phase information, so it has higher sensitivity to individual motion. At present, many achievements have been made in motion recognition based on CSI in Wi-Fi signals, such as e-eyes [29], CARM [19], WiGest [30], and WIMU [31]. These wireless motion recognition studies extract statistical features (such as histogram of signal amplitude [29]) or physical features (such as power distribution of Doppler shift [19]) from the CSI of Wi-Fi signal and associate the features with human motion. However, due to the lack of spatial resolution, the wireless signal also carries adverse environmental information irrelevant to the action, which limits the effect of recognition.

Therefore, researchers began optimizing the recognition model's cross-domain generalization ability, such as using transfer learning and confrontational learning [32–34]. However, these methods need to add new data sets to the recognition model every time, which greatly increases the training workload. With the rise of deep learning, many attempts have been made to improve the generalization performance of Wi-Fi action recognition. For example, Widar 3.0 [35] mine the characteristics extracted by CSI in spatial dimension and time dimension, respectively, by using convolutional neural network (CNN) and recurrent neural network (RNN) to distinguish common human–computer interaction actions. STFNets [36] proposed a new neural network construction module: a short-time Fourier neural network. It directly learns the characteristics in the frequency domain of various sensor inputs. EI [32] designed a more complex network structure, defined a new loss function, and directly used the new model to learn the common expression of signal features in different environments based on making full use of unlabeled data. Therefore, it is necessary to use deep learning to improve generalization ability while avoiding more complex network structures from slowing down or even hindering training and consuming training data excessively [37].

3. Enhanced Wi-Fi Features Based on Human Body Speed Distribution

In this section, we present an improved MUSIC algorithm to enhance the resolution ratio of arrival angle by introducing time of flight, then solve the sampling time offset and sampling frequency offset caused by the time of flight. Given that some human actions have too little influence on signals to estimate DFS, a dynamic signal amplification method based on the Fresnel Zone model is proposed, which establishes the corresponding relationship between DFS and the human motion velocity distribution in the human reference frame.

3.1. Human Body Location Based on Improved Multi-Signal Classification Algorithm

Wi-Fi technology uses several different subcarriers to transmit signals, which are modulated on different subcarriers to achieve the purpose of parallel transmission. The subcarriers are superimposed during propagation and accepted by the receiver antenna. The key of

the MUSIC algorithm is to calculate the phase shift caused by the distance among antennas to estimate AOA (angle of arrival). However, this method requires many antennas, and the estimation accuracy will be greatly reduced when applied to human body location scenes due to the multi-path effect. To make phase differences more obvious and improve the resolution of AOA, we introduce ToF to the MUSIC algorithm. The signals from L signal resources are accepted by receiving array consisting of M antennas. For one signal, different subcarriers arrive at one receiving antenna with the same incident angle φ. Because of the different frequencies, the generated original phase shift formula is as follows

$$\Omega = \frac{2\pi(\mathrm{L}-1)\mathrm{d}(f_1 - f_2)\cos(\varphi)}{c} \tag{1}$$

where f_1 and f_2 are the frequency of two subcarriers and c is the speed of light. Due to Wi-Fi's narrow bandwidth signal, the subcarriers' frequency is close. Since light travels very fast, the phase difference is not significant enough to be involved. After adding ToF, the phase shift formula becomes

$$\Omega = 2\pi(f_1 - f_2)\tau \tag{2}$$

where τ is the time of flight. Through importing a new parameter, the distinction of the phase shift is strengthened. Furthermore, the spatial spectral function is expressed as:

$$P_{mu}(\varphi, \theta, \tau) = \frac{1}{a^H(\varphi, \theta, \tau)UU^H a(\varphi, \theta, \tau)} \tag{3}$$

According to the spatial spectrum function, traversing all azimuth φ, pitch θ, and ToF τ at certain intervals, we can find out the number of maxima and the corresponding azimuth and pitch and then deduce the position of the human body in reverse. By the signal's angle of arrival, the source azimuth can be uniquely determined.

The introduction of ToF will lead to the time offset in random sampling, and we leverage the multiple linear regression method to solve the phase noise. STO (Sampling Time Offset) and SFO (Sampling Frequency Offset) equivalently influence the transmitting or receiving antennas. The extra delay is constant in the transmission path of different antennas with the same sampling, but inconsistent phase errors are generated in different sampling packets. The offset introduced by SFO to all paths appears as a linear frequency term. Namely, STO offset τ_{sto} leads to $2\pi\Delta f(k-1)\tau_{sto}$ phase offset in the k-th subcarrier, and Δf is the frequency interval between carriers.

For each CSI measurement, the offset is removed by removing the linear fitting of the spread phase shift on all antenna subcarriers. Assuming that $\psi(i, j, k)$ is the CSI phase of the k-th subcarrier transmitted by the j-th transmitting antenna and received by the i-th receiving antenna, the best linear fitting can be obtained as follows:

$$\hat{\tau}_{\mathrm{sto}} = \arg\min_{\beta_1} \sum_{N,M}^{i,j=1} \sum_{K}^{k=1} (\psi(i, j, k) + 2\pi\Delta f(k-1)\beta_1 + \beta_2)^2 \tag{4}$$

where β_1 is the common slope of receiving phase response antennas and β_2 is offset. The corrected CSI phase is

$$\hat{\psi}(i, j, k) = \psi(i, j, k) - 2\pi\Delta f(k-1)\hat{\tau}_{sto} \tag{5}$$

Finally, the AoA of the reflected signal can be obtained from CSI information, and the position information of the human body can be obtained from the arrival angle of the human body signal [38–40].

3.2. DFS Estimation Based on Dynamic Signal Amplification Method

To obtain DFS caused by the movement of the human body between a pair of transmitting and receiving antennas [40], we propose a method to amplify the dynamic signal changes caused by the human body and extract the required DFS from it. According to the Fresnel zone model, the larger the motion amplitude, the greater the signal change, and the easier it is to detect the dynamic path. The signal change is a segment of a sine wave. When the path length change caused by action is small, its phase change is not easily observed. However, the same displacement at different starting positions will cause different signal changes.

Therefore, to change the phase difference to amplify the signal, the phase of the static vector can be changed by adding a virtual path to change the phase difference between the dynamic vector and the static vector. The specific process is as follows:

(A) Search for phase shift. As shown in Figure 1, H_s,H_d, H_v and H_s' are static variable, dynamic variable, virtual variable, and static synthetic variable, respectively, and H_s' is the final variable to be updated. The dynamic and static vector phase difference $\Delta\theta_{sd}$ can not be obtained directly from the original signal but can be traversed between $0-2\pi$, and the phase shift with the strongest sensing ability is the best one.

(B) Calculate multiple radial quantities. Estimate the static variables approximately by the signal average value periodically. In Figure 2, H_{v1} and H_{v2} are two virtual paths with different lengths; they and H_s are combined into different static paths, H_{s1}' and H_{s2}', respectively. It can be observed that although H_{s1}' and H_{s2}' have different lengths, they have the same phase shift, which ensures the same degree of improvement in sensing performance. To simplify the problem, we can set $|H_s'|$to a value equal to $|H_s|$, which will not affect the correctness of the result. To date, according to the known parameters, the amplitude of the synthetic static path can be constructed:

$$|H_v| = \sqrt{|H_s|^2 + |H_s'|^2 - 2|H_s||H_s'|\cos\omega} \tag{6}$$

After the amplitude of H_v is obtained, $\delta = \arcsin\frac{\sin\omega|H_{s'}|}{|H_v|}$ can be obtained according to $\frac{|H_v|}{\sin\omega} = \frac{|H_s'|}{\sin\delta}$. The phase of virtual path H_v is $\theta_v = \theta_s + \delta - \pi$, and θ_s is the phase of static path vector. According to the given ω, the virtual static vector can finally be obtained as follows:

$$H_v = |H_v|e^{j\theta_v} \tag{7}$$

(C) Add a composite path. After H_v is obtained, this path is created by MATLAB and added to the original signal. For the original signal of $S_0 = (CSI_1, CSI_2, \ldots, CSI_P))$ with p-CSI samples, the new signal after adding virtual path is $S = (CSI_1 + H_v, CSI_2 + H_v, \ldots, CSI_P + H_v)$.

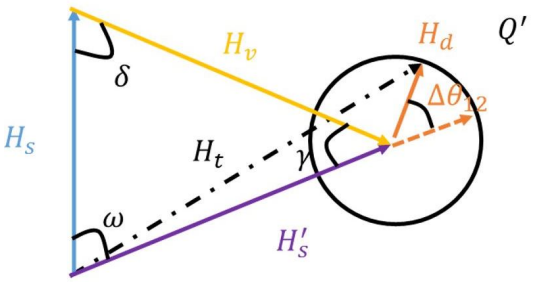


Figure 1. Diagram without virtual diagram.

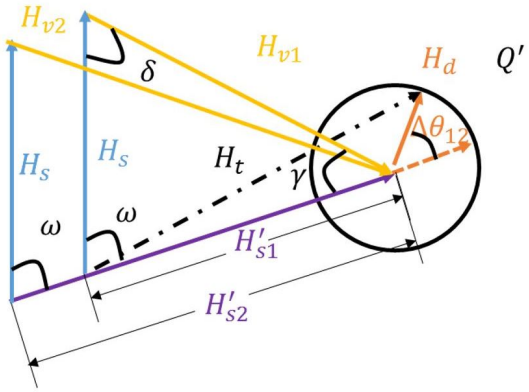


Figure 2. Diagram with virtual diagram.

To date, a new signal with strong dynamic perception ability has been obtained. The DFS caused by the dynamic path can be extracted by time-frequency analysis of this signal, and its symbol is $f_D(t)$, which is related to the change of signal propagation path length $d(t)$:

$$f_D(t) = -\frac{1}{\lambda}\frac{d}{dt}d(t) \tag{8}$$

where λ is the signal wavelength. The channel response of the dynamic path can be expressed by DFS:

$$H(t) = H_s(t) + H_d(t) = H_s(t) + \sum_{\ell \in L_d} \alpha_\ell(t) e^{-j2\pi \int_{-\infty}^{f} f_{D_l}(u)\mathrm{du}} \tag{9}$$

where $H(t)$ is the resultant signal accepted at the receiver, which is a linear superposition of $H_s(t)$ and $H_d(t)$. $\alpha(t)$ is the original phase shift and amplitude in complex form.

In a short time, α and $f_D$can be considered to be constant, as shown in the following formula:

$$\mathcal{H}(t) \approx H_s(t) + \sum_{\ell \in L_d} \alpha_\ell(t) B\left(f_{D_\ell}(t)\right) \tag{10}$$

where B is the window of dividing signal segments, and L_d is the number of dynamic paths. By applying a short-time Fourier transformation to the new CSI, we can obtain the DFS caused by human actions.

3.3. Motion Velocity Distribution Model Based on Human Frame of Reference

When a human body makes an action, its body parts (for example, two hands, two arms, and the trunk) move at different speeds, and DFS is highly correlated with the orientation of the human body position. The movement speed distribution obtained from DFS is also related to human body position. However, in the human frame of reference, the segmentation of the moving body is only associated with the type of action and the distribution of movement speed, which is irrelevant to the position. Thus, we must transform to the human frame of reference to reduce parameters.

It is necessary to estimate the target's position information in the Z-axis direction and two-dimensional plane simultaneously to establish the human body's three-dimensional coordinate system. The local coordinate system centered on the human body takes the front-facing direction of the human body as the X-axis, and the other direction parallels the ground and is perpendicular to the X-axis as the Y-axis. To include the actual height of most people, we set the altitude datum of the Z-axis to 1.3 m relative to the ground and obtain the three-dimensional coordinates of the transmitting end and receiving end of the i-th channel in the human coordinate system:

$$l_t^{(i)} = \left(x_t^{(i)}, y_t^{(i)}, z_t^{(i)}\right) \tag{11}$$

$$l_r^{(i)} = \left(x_r^{(i)}, y_r^{(i)}, z_r^{(i)}\right) \tag{12}$$

Velocity component $\vec{v} = (v_x, v_y, v_z)$ caused by human action contributes to the frequency component. We label it as $f^{(\mathrm{i})}(\vec{v})$, and the following is the corresponding relation for the i-th channel:

$$f^{(\mathrm{i})}(\vec{v}) = a_x^{(\mathrm{i})} v_x + a_y^{(\mathrm{i})} v_y + a_z^{(\mathrm{i})} v_z \tag{13}$$

where $a_x^{(\mathrm{i})}$,$a_y^{(\mathrm{i})}$,$a_z^{(\mathrm{i})}$ are the coefficients determined by the three-dimensional position information of the transmitting end and the receiving end:

$$a_x^{(i)} = \frac{1}{\lambda}\left(\frac{x_t^{(i)}}{\left\|l_t^{(i)}\right\|_2} + \frac{x_r^{(i)}}{\left\|l_r^{(i)}\right\|_2}\right) \tag{14}$$

$$a_y^{(i)} = \frac{1}{\lambda}\left(\frac{y_t^{(i)}}{\left\|l_t^{(i)}\right\|_2} + \frac{y_r^{(i)}}{\left\|l_r^{(i)}\right\|_2}\right) \tag{15}$$

$$a_z^{(i)} = \frac{1}{\lambda}\left(\frac{z_t^{(i)}}{\left\|l_t^{(i)}\right\|_2} + \frac{z_r^{(i)}}{\left\|l_r^{(i)}\right\|_2}\right) \tag{16}$$

According to Equations (14)–(16), the coefficients $a_x^{(i)}$,$a_y^{(i)}$,$a_z^{(i)}$ between human body's velocity and frequency are obtained. Due to the limitation of human body structure (such as joints), the number of reflected multi-path signals is limited. The possible velocity components caused by human motion in the X and Y axes are generally no more than 8, and that in the Z axis is generally no more than 5. Thus, we traverse all possible $\vec{v}$ in a certain speed range according to the set step size and multiply it by corresponding a_x,a_y,a_z. The possible frequency component generated by the possible speed component of the human body is obtained. If the frequency component calculated is between the maximum and minimum sampling frequencies, the corresponding position of the allocation matrix is 1; otherwise, it is 0, and the initial allocation matrix in i-th link $A_{F*PN^2}^{(i)}$ is obtained. F represents the number of samples, namely the number of peak values in the DFS profile or the number of actual velocity components. N is the number of possible velocity components in X or Y axes [35], and P is that in Z axes.

$$A_{j,k}^{(i)} = \begin{cases} 1 & f_{min} < f^{(i)}(\vec{v}_k) < f_{max} \\ 0 & others \end{cases} \tag{17}$$

where f_{min} and f_{max} are the minimum and maximum frequency sampling frequency, respectively, $f^{(i)}(\vec{v}_k)$ is possible velocity components correspond to frequency components.

The DFS obtained in Section 3.2 is segmented according to the set time t, and the average value is taken as the actual value of the segment.

$$\min_V \sum_M \left| EMD\left(A^{(i)}V, D^i\right)\right| + \eta\|V\| \tag{18}$$

where M is the number of Wi-Fi transceiver links. The sparseness of the number of velocity components is defined by the term $\eta\|V\|$, where η is the sparseness coefficient, and $\|V\|$ is the number of non-zero velocity components. To alleviate the approximate error caused by the quantification of velocity distribution, EMD (The Earth Mover's Distance) [41] is used to characterize the difference between the two distributions. The difference between the actual and the estimated Doppler spectrum calculated by the allocation matrix A is measured by the Formula (18). We can obtain the corresponding human motion speed distribution by iteratively reducing these two distributions' differences.

4. Enhanced Convolution Transformer Network

In this section, the multi-scale convolution module is added to the Transformer model to enhance its local feature extraction ability further and improve the classification performance of different actions. In addition, Gaussian range coding is introduced to retain the time sequence information of data, which reduces the difference caused by individual factors in the same action. Finally, human motion recognition is realized.

4.1. Model Overview

The human motion distribution data extracted from CSI is used as the input of the convolution enhanced Transformer model, as shown in Figure 3. The main target is to classify. Thus, we only use the Encoder component in the Transformer model. Firstly, the input data is processed by the multi-head self-attention module [42] after the location code. A weighted sum of the values is output, where the weight assigned to each value is computed by the dot product of the query with the corresponding key.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{19}$$

where $Q \in R^{L\times d_k}$, $K \in R^{L\times d_k}$, $V \in R^{L\times d_V}$, d_k, d_v are queries, keys, values, the dimension of Q and K, and the dimension of V, respectively. Q, K, and V are obtained by linear variation of the input data.

$$Q = XW_Q \tag{20}$$

$$K = XW_K \tag{21}$$

$$V = XW_V \tag{22}$$

where $W_Q \in R^{d_(in)\times d_K}, W_K \in R^{d_(in)\times d_k}, W_w \in R^{d_(in)\times d_v}$ are learnable weight parameters. $d_(in)$ is the dimension of the input vector. To concentrate information from different subspaces, the queries, keys, and values use different linear variations methods to change h times (h-head). The output of the projection is wired and projected again to produce the final output

$$\text{MultiHead}(Q, K, V) = [\text{head}_1, \ldots, \text{head}_h]W_O \tag{23}$$

head $_i$ = Attention$\left(XW_Q^i, XW_K^i, XW_V^i\right), W_O \in R^{hd_m\times d_k}$ is the final projection matrix.. Multi-head self-attention module makes the Transformer learn characteristic from different subspace. In order to facilitate the residual connection, $d_in = h * d_v$. After residual connection and normalization, this structure can make the backward and forward propagation of information smoother and alleviate the gradient explosion problem.

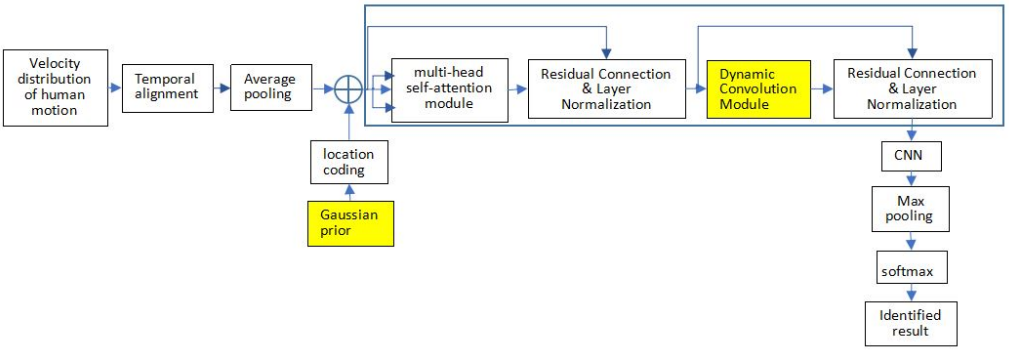


Figure 3. Convolutional Feature Augmentation Transformer Model Architecture.

Figure 3 shows the entire network structure. Furthermore, the main part is the Transformer. Different from the usual Transformer, we use a dynamic convolution module

instead of PFFN to capture feature information at different scales and strengthen the model's performance. Meanwhile, we use Gaussian Prior function to enhance the ability to express human actions' temporal features and retain the data sequence. Finally, extract the corresponding action features for action classification.

4.2. Dynamic Convolution Module

In order to select appropriate convolution parameters for different action inputs, we add a convolution module with adaptive size and quantity to the transformer based on the idea of dynamic convolution so that the convolution size and the number of convolutions can vary with the input. Setting two parallel convolutions for calculating these two parameters will be more complex. Thus, we use the same set of parallel convolutions. The size is 1×1, and the number is K. As shown in Figure 4, we perform mean pooling on the input feature matrix, and then go through two layers of full connection and ReLu unit, respectively, use the softmax function to calculate the weight of the convolution kernel size, use the sigmoid function to calculate the weight of the number of convolution kernels, and train them jointly to calculate the final dynamic convolution parameters, and convolve with the input features to obtain local features. The input and output of the dynamic convolution formally are:

$$P = \text{ReLU}(\text{ Dropout }(BN(\text{Conv}(\tilde{W}(X); X)))) \tag{24}$$

In the formula: $\tilde{W}$ is the weight function of the convolution kernel parameter, which is a function of the input X

$$\tilde{W}(X) = \sum_{k=1}^{K} \pi_k(X)\tilde{W}_k \tag{25}$$

In the formula: π is the weight parameter of a certain parameter value, and k is the kth parallel convolution kernel.

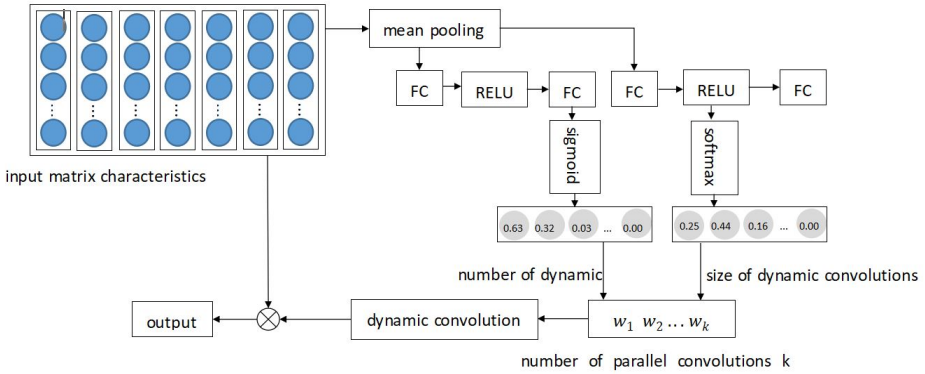


Figure 4. Dynamic CNN structure.

4.3. Representation Enhancement Based on Gaussian Prior

Although the Transformer model uses the attention mechanism to avoid the defects of network structures such as CNN and RNN, the attention mechanism of the Transformer model does not include the positional relationship between data units, namely, it can not focus on intermediate state actions in a complete action because they are the same. So, the features are not distinguishable completely in the Transformer module.

The attention mechanism can calculate the semantic correlation between different units but ignore the influence of distance. Thus, we use Gaussian distribution with the variance of $\sigma^2 = 1/(2\pi)$ and the probability density of $\phi(d) = e^{-\pi d^2}$ to increase the importance of close cells and reduce the weight of long-distance units. At the same time, Equation (26) transforms the Gaussian distribution into a bias term, saving additional multiplications.

$$\begin{aligned}\tilde{x}_i &= \sum_j \frac{\phi(d_{i,j})\,\mathrm{comp}_{i,j}}{Z_1} x_j = \sum_j \frac{e^{-d_{i,j}^2} \cdot e^{(x_i \cdot x_j)}}{Z_2} x_j \\ &= \sum_j \frac{e^{-d_{i,j}^2 + (x_i \cdot x_j)}}{Z_2} x_j \\ &= \sum_j \mathrm{Softmax}\left(-d_{i,j}^2 + (x_i \cdot x_j)\right) x_j\end{aligned} \tag{26}$$

where x_i represents the action speed distribution with position i in the action input x, d is the distance between units, and $Z_1 = \sum_k \phi(d_{i,k})\,\mathrm{comp}_{i,k}$ and $Z_2 = \sum_k e^{-d_{i,k}^2 + (x_i \cdot x_k)}$ are the normalization factors. We calculate the compatibility function according to the softmax of the point multiplication to obtain the correlation between the distribution, namely $\mathrm{Com}_{i,j} = \mathrm{Softmax}(x_i \cdot x_j)$.

5. Experiments and Analysis

In this section, we first provide in-depth ablation studies to analyze our method and then conduct comparison experiments among different models on common data sets, illustrating the whole algorithm's effectiveness and superiority.

5.1. Experiment Settings

Select two 9300 network interface controllers (NICs) as the transmitter and receiver of the signal and the CSI data is collected on the PicoScenes platform. The height of the equipment at the transmitting and receiving end is 1.3 m, located in the middle of the human body to realize three transmitting and six receiving points. The experimenter collected movement at different positions and orientations in the open room, office, apartment room, and corresponding aisle areas, including walking, running, etc. Ten acquisitions were carried out for different actions, each lasting for 5 min. The signal transmission frequency was set to 1000 Hz, the transmission energy was set to 30 dBm, and the subcarrier bandwidth was set to 20 MBps.

When training the Transformer network, the appropriate parameters of the network structure are determined through parameter debugging. The average pool step is set to 4. The number of Gaussian distributions k = 10. The number of Transformer layers is set to h = 5, the input and output dimension is set to 90, and the number of heads in the multi-head attention mechanism is set to 9. The number of parallel convolutions in dynamic convolution structure is 6, and dropout is set to 0.1. There are two convolution types in the prediction layer. The size is 10, 40, the number of each type W is 128, and the dropout of this layer is set to 0.5.

5.2. Ablation Analysis of Our Method

We conduct ablation studies to verify the importance of the methods in Sections 3 and 4, including the importance of correcting errors caused by noise, the effect of motion distribution accuracy, and the contribution of Convolution module and Gaussian prior.

Importance of correcting errors caused by noise: To enhance the phase difference of the MUSIC algorithm, we introduce time-of-flight, but the resulting STO and SFO at the same time introduce phase errors in the received signals of different packets. As shown in Figure 5, the left side is the CSI map of the original sampling, and the right side is the result after using a multivariate linear combination. It can be seen that the phase offset has been aligned.

Effect of motion distribution estimation: The human action velocity distribution is used as the input of the network, so the accuracy of the action velocity distribution estimation directly affects the effect of action recognition, which in turn depends on the accuracy of the human body position estimation. To verify the effect of Section 3, we add joint parameters and human body dynamic signal amplification steps for comparison experiments.

As shown in Figure 6, it can be seen that the accuracy rate has been significantly improved after introducing our method.

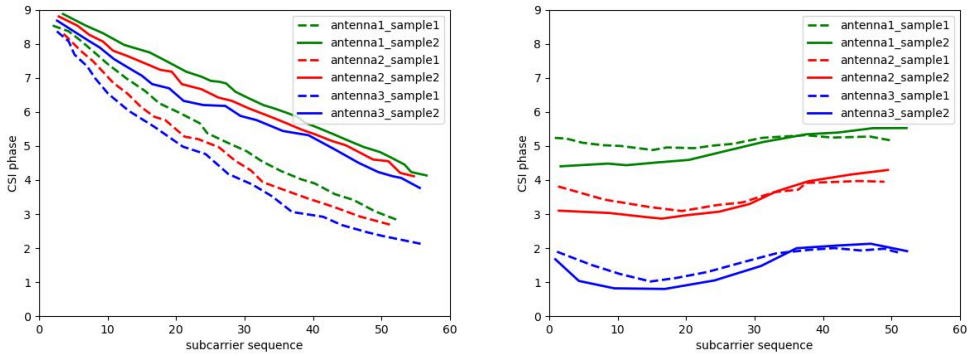


Figure 5. The correction of phase noise. (**a**) Original CSI phase. (**b**) Corrected CSI phase.

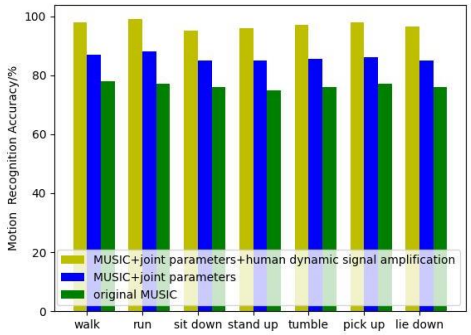


Figure 6. Comparison diagram of action speed distribution effect.

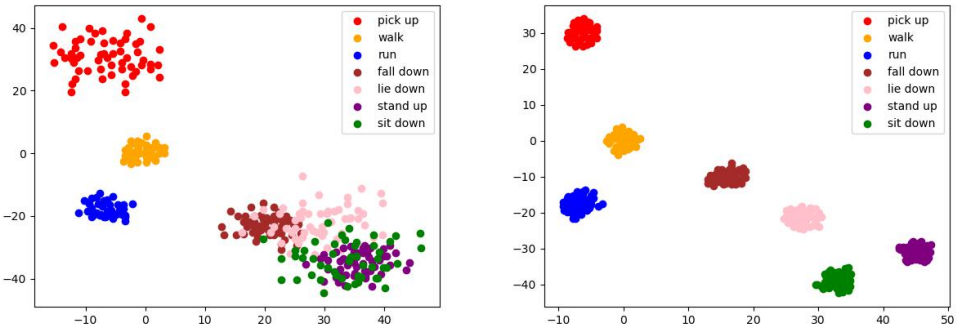
Contribution of dynamic Convolution module and Gaussian prior: By using the weight function to adjust the convolution kernel parameters corresponding to different actions dynamically, the model extracts action features more accurately. To consider the characteristics of time series, Gaussian priors are added. As shown in Figure 7a, the features of similar actions are confused, and the features of the same action are scattered and easily identified as other action features. Figure 7b shows the classification performance of the model using the Convolution module and Gaussian prior. We can find that the feature difference between different actions increases, and at the same time the action features that are opposite in time can be better distinguished.

5.3. Comparison with Other Methods

In order to better confirm the improvement of the proposed algorithm, we compare the method in Section 3.1 with the other two positioning models for AoA estimation, and contrast the recognition of our algorithm with other methods based on CSI using public datasets.

Two typical positioning models are used to estimate human body positions from CSI data collected at different locations. Figure 8 shows the cumulative error of AoA estimation for dynamic human bodies by different models. The median errors of SpotFi [43], Widar2.0 [38], and the model in this paper are 10.3°, 5.1°, and 3.7°, respectively. Among them, SpotFi uses the MUSIC algorithm for positioning, and the accuracy of this method greatly reduces when the number of antennas is insufficient. Widar2.0 uses the traditional maximum likelihood and the extended SAGE algorithm. SAGE algorithm completes parameter decomposition by continuous transformation of index set and decomposes the maximization problem of AoA into the corresponding maximization problems of index set

parameters. This method does not guarantee certain convergence to the optimal value and a local optimum usually results in not reaching the desired estimate.



(**a**) (**b**)

Figure 7. Classification performance. (**a**) Model performance without Convolution module and Gaussian prior. (**b**) Model performance with Convolution module and Gaussian prior.

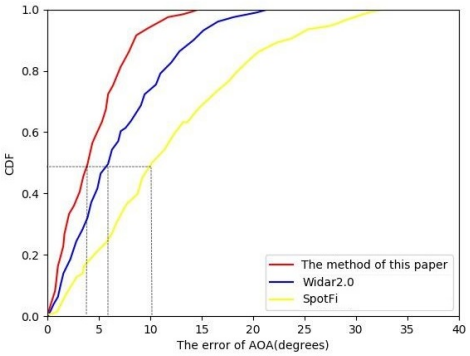


Figure 8. AOA cumulative error of different models.

There are also many action recognition methods based on CSI; however, the corresponding of action sets are not exactly the same. For the sake of fairness, the CSI action data set in the publicly available WiAR [44] and the human activity recognition data set WiDAR [35] from the Tsinghua University team are selected. We selected CARM [19], Widar 3.0 system [35], EI [32], and our model by PRF values (precision, recall, F1-score) as indicators for comparison.

As shown in Tables 1 and 2, CARM uses DFS as the learning feature and adopts the HMM model, so the performance is the worst when human motion data in different environments and different positions are added to the datasets. EI combines the adversarial network and specially trains the loss, and additionally utilizes the features of the unlabeled data in the target domain, and its recognition effect is better, but the adversarial network model is uncontrollable and the training is difficult. Widar 3.0 uses CNN to extract signal spatial features and GRU to extract temporal features. The time normalization of action features must be strictly performed and the training time is long, otherwise the recognition effect cannot reach expectations. Our model transforms the DFS into the three-dimensional velocity distribution of human actions, which improves the efficiency and accuracy of action recognition compared with other recognition models.

Table 1. Comparison of recognition accuracy of all models in WiAR.

	This Paper	Widar 3.0	EI	CARM
F1 Score	90.24	88.79	92.32	70.45
Precision	91.65	88.12	92.43	70.60
Recall/TPR	90.33	89.48	92.21	70.32

Table 2. Comparison of recognition accuracy of all models in WiDAR.

	This Paper	Widar 3.0	EI	CARM
F1 Score	92.98	91.79	91.27	80.41
Precision	90.65	91.12	91.22	80.60
Recall/TPR	91.33	92.48	91.34	80.32

5.4. Generalization Experiment

To verify the generalization of our method, we conduct experiments in different positions, orientations and scenarios, employ 9300 network cards and the PicoScenes software platform for data acquisition and processing, and evaluate the performances of using human velocity distribution compared with that of using original CSI data.

In the area of 2 m × 2 m, the positions of the transmitter and receiver are fixed as shown in Figure 9. The variation interval of the human body position is [u,w,x,y,z], and the variation interval of the human body orientation is [1,2,3,4,5].

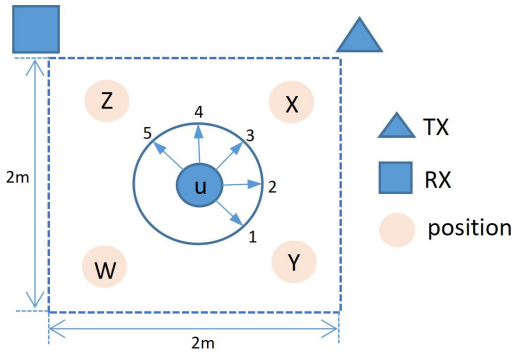


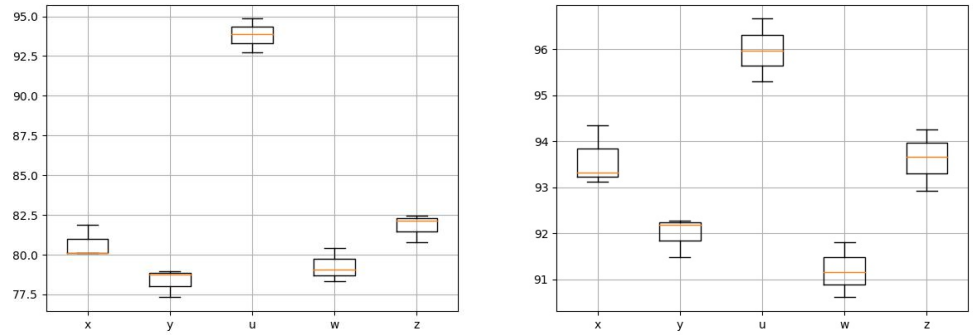
Figure 9. Map of the device and the human body position.

The direction of the human body is toward 4, and the position of the human body is changed in x, y, u, w, and z. Comparing Figure 10a with Figure 10b, we can see that the recognition accuracy of using the human velocity distribution data is much higher than that of using the original CSI data, and the accuracy is even improved by about 14% on average. When the position of the human body changes from the sender to the receiver, the average accuracy of different actions in any position using the human velocity distribution data is above 90%. This indicates that the model has good generalization performance in the location domain.

The different orientation of the human body will also have a certain impact on the recognition. The human position is fixed as u, and the human orientation is varied in orientations 1, 2, 3, 4, and 5 of Figure 9. Figure 11a shows the action recognition results of the original CSI as input under different orientations, and the average accuracy is 68.2%. We can also find that when the orientation changes, the accuracy of the model also changes. It is mainly because the path of the Wi-Fi signal obeys the law of reflection. In some human body orientations, part of the action reflection signals are occluded by the body, which leads to less action influence in the signal received by the receiver than the normal situation and then leads to the reduction in recognition accuracy. Figure 11b illustrates the accuracy of action recognition under different orientations with human

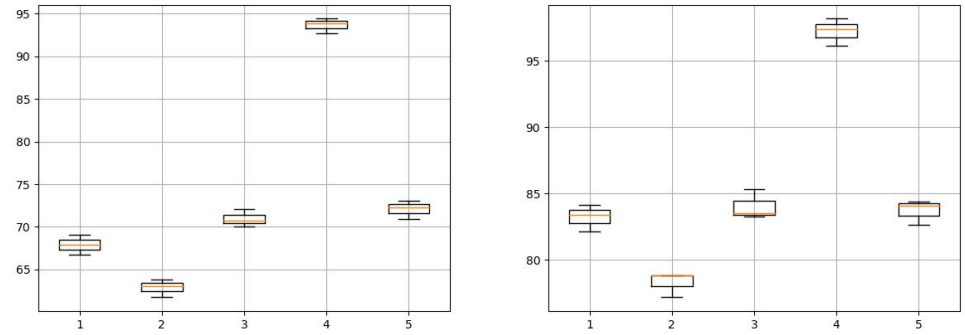
velocity distribution as input. The average accuracy in this case is 84.4%. In Figure 11b, recognition accuracy is improved under different orientations, indicating that the model has certain generalization performance in the orientation domain. However, in the directions which are easily occluded by the body (orientations 1 and 2), the model cannot eliminate the influence caused by the signal reflection law.

We also test the algorithm in different scenarios, as shown in Figure 12, among which the training accuracy rate of the open indoor scene is the highest, and the accuracy rate of the office scene is the lowest. Due to the complex environment, the wavelet transform cannot completely remove the noise, which interferes with the performance of the action recognition, but the accuracy rate is still above 84%.



(**a**) (**b**)

Figure 10. Accuracy of motion recognition in different positions. (**a**) Using CSI data. (**b**) Using human velocity distribution.



(**a**) (**b**)

Figure 11. Accuracy of motion recognition under different orientations. (**a**) Using CSI data. (**b**) Using human velocity distribution.

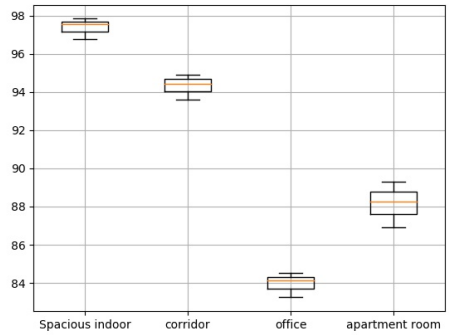


Figure 12. Accuracy of motion recognition in different scenarios.

6. Conclusions

In order to adapt to the common equipment, we propose a human body positioning method based on the improved MUSIC algorithm. In view of the problem that the dynamic path changes caused by some actions are slight, we amplify the dynamic signal changes based on the Fresnel area model and obtain the DFS of human motion. According to the human position and DFS, the motion speed distribution based on the human frame of reference is constructed. Then, we use Transformer to capture the movement characteristics of the input data and add a dynamic convolution module and Gaussian Coding in order to enhance the local feature extraction ability of the model and improve the classification performance of different movements. Our model improves the generalization ability without complex network structures, which has excellent potential for motion recognition in different environments, locations, and orientations. In future work, we will expand the single-user scene into a multi-user scene.

Author Contributions: Conceptualization and methodology, W.S.; software and writing—original draft preparation, M.D. and L.L.; writing—review and editing, H.H. and J.Z.; formal analysis and investigation, L.L. and C.Y.; data curation, C.L.; supervision, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data are contained within the article.

Acknowledgments: We acknowledge material support given by Chenwei Cui.

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
MUSIC	Multiple signal classification
ToF	Time of flight
DFS	Doppler frequency shift
RSSI	Received signal strength information
CSI	Channel state information
CNN	Convolutional neural network
RNN	Recurrent neural network
AOA	Angle of arrival
STO	Sampling time offset
SFO	Sampling frequency offset
EMD	The Earth Mover's Distance
RFID	Radio frequency identification

References

1. Peng, F.; Zhang, H. Research on Action Recognition Method of Dance Video Image Based on Human-Computer Interaction. *Sci. Program.* **2021**, *2021*, 8763133. [CrossRef]
2. Lou, M.; Li, J.; Wang, G.; He, G. AR-C3D: Action recognition accelerator for human-computer interaction on FPGA. In Proceedings of the 2019 IEEE International Symposium on Circuits and Systems (ISCAS), Sapporo, Japan, 26–29 May 2019; IEEE: New York, NY, USA, 2019; pp. 1–4.
3. Chiu, W.Y.; Tsai, D.M. ICA-based Action Recognition for Human-computer Interaction in Disturbed Backgrounds. In Proceedings of the GRAPP/IVAPP, Rome, Italy, 24–26 February 2012; pp. 519–522.
4. Zhu, Y.; Lan, T.; Yang, Y.; Robinovitch, S.; Mori, G. Latent Spatio-temporal Models for Action Localization and Recognition in Nursing Home Surveillance Video. In Proceedings of the MVA, Kyoto, Japan, 20–23 May 2013; pp. 463–466.

5. Sun, H.; Chen, Y. Real-Time Elderly Monitoring for Senior Safety by Lightweight Human Action Recognition. In Proceedings of the 2022 IEEE 16th International Symposium on Medical Information and Communication Technology (ISMICT), Lincoln, NE, USA, 2–4 May 2022; IEEE: New York, NY, USA, 2022; pp. 1–6.
6. Adewopo, V.; Elsayed, N.; Anderson, K. Baby Physical Safety Monitoring in Smart Home Using Action Recognition System. *arXiv* **2022**, arXiv:2210.12527.
7. Ma, Q.; Li, X.; Li, G.; Ning, B.; Bai, M.; Wang, X. MRLIHT: Mobile RFID-based localization for indoor human tracking. *Sensors* **2020**, *20*, 1711. [CrossRef] [PubMed]
8. Zhao, J.; Zhou, J.; Yao, Y.; Li, D.a.; Gao, L. RF-motion: A device-free RF-based human motion recognition system. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 1497503. [CrossRef]
9. Lattanzi, E.; Calisti, L.; Freschi, V. Automatic unstructured handwashing recognition using smartwatch to reduce contact transmission of pathogens. *arXiv* **2021**, arXiv:2107.13405.
10. Yang, G.; Tan, W.; Jin, H.; Zhao, T.; Tu, L. Review wearable sensing system for gait recognition. *Clust. Comput.* **2019**, *22*, 3021–3029. [CrossRef]
11. Yang, J.; Li, Q.; Wang, X.; Di, P.; Ding, H.; Bai, Y.; Dong, W.; Zhu, S. Smart wearable monitoring system based on multi-type sensors for motion recognition. *Smart Mater. Struct.* **2021**, *30*, 035017. [CrossRef]
12. Zhuang, W.; Chen, Y.; Su, J.; Wang, B.; Gao, C. Design of human activity recognition algorithms based on a single wearable IMU sensor. *Int. J. Sens. Netw.* **2019**, *30*, 193–206. [CrossRef]
13. Zhao, J.; Liu, L.; Wei, Z.; Zhang, C.; Wang, W.; Fan, Y. R-DEHM: CSI-based robust duration estimation of human motion with WiFi. *Sensors* **2019**, *19*, 1421. [CrossRef]
14. Fei, H.; Xiao, F.; Han, J.; Huang, H.; Sun, L. Multi-variations activity based gaits recognition using commodity WiFi. *IEEE Trans. Veh. Technol.* **2019**, *69*, 2263–2273. [CrossRef]
15. Zhang, J.; Wei, B.; Wu, F.; Dong, L.; Hu, W.; Kanhere, S.S.; Luo, C.; Yu, S.; Cheng, J. Gate-ID: WiFi-based human identification irrespective of walking directions in smart home. *IEEE Internet Things J.* **2020**, *8*, 7610–7624. [CrossRef]
16. Duan, P.; Li, H.; Zhang, B.; Cao, Y.; Wang, E. APFNet: Amplitude-phase fusion network for CSI-based action recognition. *Mob. Netw. Appl.* **2021**, *26*, 2024–2034. [CrossRef]
17. Chen, H.; Zhang, Y.; Yin, Y.; He, F. Human Activity Recognition Based on CSI fragment with Action-value Method. In Proceedings of the 2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML), Hangzhou, China, 25–27 March 2022; IEEE: New York, NY, USA, 2022; pp. 448–455.
18. Schmidt, R. Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas Propag.* **1986**, *34*, 276–280. [CrossRef]
19. 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]
20. Inayat, U.; Zia, M.F.; Mahmood, S.; Khalid, H.M.; Benbouzid, M. Learning-Based Methods for Cyber Attacks Detection in IoT Systems: A Survey on Methods, Analysis, and Future Prospects. *Electronics* **2022**, *11*, 1502. [CrossRef]
21. Inayat, U.; Ali, F.; Khan, H.M.A.; Ali, S.M.; Ilyas, K.; Habib, H. Wireless Sensor Networks: Security, Threats, and Solutions. In Proceedings of the 2021 International Conference on Innovative Computing (ICIC), Lahore, Pakistan, 9–10 November 2021; IEEE: New York, NY, USA, 2021; pp. 1–6.
22. Ali, S.; Sohail, M.; Shah, S.B.H.; Koundal, D.; Hassan, M.A.; Abdollahi, A.; Khan, I.U. New Trends and Advancement in Next Generation Mobile Wireless Communication (6G): A Survey. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 9614520. [CrossRef]
23. Letaief, K.B.; Chen, W.; Shi, Y.; Zhang, J.; Zhang, Y.J.A. The roadmap to 6G: AI empowered wireless networks. *IEEE Commun. Mag.* **2019**, *57*, 84–90. [CrossRef]
24. Zou, Y.; Liu, W.; Wu, K.; Ni, L.M. Wi-Fi radar: Recognizing human behavior with commodity Wi-Fi. *IEEE Commun. Mag.* **2017**, *55*, 105–111. [CrossRef]
25. Wang, F.; Zhou, S.; Panev, S.; Han, J.; Huang, D. Person-in-WiFi: Fine-grained person perception using WiFi. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5452–5461.
26. Wilson, J.; Patwari, N. Radio tomographic imaging with wireless networks. *IEEE Trans. Mob. Comput.* **2010**, *9*, 621–632. [CrossRef]
27. Sigg, S.; Blanke, U.; Tröster, G. The telepathic phone: Frictionless activity recognition from wifi-rssi. In Proceedings of the 2014 IEEE International Conference on Pervasive Computing and Communications (PerCom), Budapest, Hungary, 24–28 March 2014; IEEE: New York, NY, USA, 2014; pp. 148–155.
28. Han, J.; Qian, C.; Yang, P.; Ma, D.; Jiang, Z.; Xi, W.; Zhao, J. GenePrint: Generic and accurate physical-layer identification for UHF RFID tags. *IEEE/ACM Trans. Netw.* **2015**, *24*, 846–858. [CrossRef]
29. 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, Maui, HI, USA, 7–11 September 2014; pp. 617–628.
IEEE J. Sel. Areas Commun. **2017**, *35*, 1118–1131. [CrossRef]
30. Abdelnasser, H.; Youssef, M.; Harras, K.A. Wigest: A ubiquitous wifi-based gesture recognition system. In Proceedings of the 2015 IEEE Conference on Computer Communications (INFOCOM), Kowloon, Hong Kong, 26 April–1 May 2015; IEEE: New York, NY, USA, 2015; pp. 1472–1480.

31. Venkatnarayan, R.H.; Page, G.; Shahzad, M. Multi-user gesture recognition using WiFi. In Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services, Munich, Germany, 10–15 June 2018; pp. 401–413.
32. 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, New Delhi, India, 29 October–2 November 2018; pp. 289–304.
33. Zhang, J.; Tang, Z.; Li, M.; Fang, D.; Nurmi, P.; Wang, Z. CrossSense: Towards cross-site and large-scale WiFi sensing. In Proceedings of the 24th Annual International Conference on Mobile Computing and Networking, New Delhi, India, 29 October–2 November 2018; pp. 305–320.
34. 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.
35. Zheng, Y.; Zhang, Y.; Qian, K.; Zhang, G.; Liu, Y.; Wu, C.; Yang, Z. Zero-effort cross-domain gesture recognition with Wi-Fi. In Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services, Seoul, Republic of Korea, 17–21 June 2019; pp. 313–325.
36. Yao, S.; Piao, A.; Jiang, W.; Zhao, Y.; Shao, H.; Liu, S.; Liu, D.; Li, J.; Wang, T.; Hu, S.; et al. Stfnets: Learning sensing signals from the time-frequency perspective with short-time fourier neural networks. In Proceedings of the World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; pp. 2192–2202.
37. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, Sardinia, Italy, 13–15 May 2010; pp. 249–256.
38. Li, X.; Zhang, D.; Lv, Q.; Xiong, J.; Li, S.; Zhang, Y.; Mei, H. IndoTrack: Device-free indoor human tracking with commodity Wi-Fi. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2017**, *1*, 1–22. [CrossRef]
39. Soltanaghaei, E.; Kalyanaraman, A.; Whitehouse, K. Multipath triangulation: Decimeter-level wifi localization and orientation with a single unaided receiver. In Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services, Munich, Germany, 10–15 June 2018; pp. 376–388.
40. Qian, K.; Wu, C.; Zhang, Y.; Zhang, G.; Yang, Z.; Liu, Y. Widar2. 0: Passive human tracking with a single Wi-Fi link. In Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services, Munich, Germany, 10–15 June 2018; pp. 350–361.
41. Rubner, Y.; Tomasi, C.; Guibas, L.J. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vis.* **2000**, *40*, 99–121. [CrossRef]
42. Li, B.; Cui, W.; Wang, W.; Zhang, L.; Chen, Z.; Wu, M. Two-stream convolution augmented transformer for human activity recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 2–9 February 2021; Volume 35; pp. 286–293.
43. Kotaru, M.; Joshi, K.; Bharadia, D.; Katti, S. Spotfi: Decimeter level localization using wifi. In Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication, London, UK, 17–21 August 2015; pp. 269–282.
44. Guo, L.; Wang, L.; Lin, C.; Liu, J.; Lu, B.; Fang, J.; Liu, Z.; Shan, Z.; Yang, J.; Guo, S. Wiar: A public dataset for wifi-based activity recognition. *IEEE Access* **2019**, *7*, 154935–154945. [CrossRef]