Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.063295

ARTICLE

Through-Wall Multihuman Activity Recognition Based on MIMO Radar

Changlong Wang[1], Jiawei Jiang[1], Chong Han[1,2,*], Hengyi Ren[3], Lijuan Sun[1,2] and Jian Guo[1,2]

[1]College of Computer, Nanjing University of Posts and Telecommunications, Nanjing, 210003, China
[2]Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing University of Posts and Telecommunications, Nanjing, 210023, China
[3]College of Information Science and Technology, Nanjing Forestry University, Nanjing, 210000, China

*Corresponding Author: Chong Han. Email: hc@njupt.edu.cn

Received: 10 January 2025; Accepted: 27 February 2025; Published: 19 May 2025

ABSTRACT: Existing through-wall human activity recognition methods often rely on Doppler information or reflective signal characteristics of the human body. However, static individuals, lacking prominent motion features, do not generate Doppler information. Moreover, radar signals experience significant attenuation due to absorption and scattering effects as they penetrate walls, limiting recognition performance. To address these challenges, this study proposes a novel through-wall human activity recognition method based on MIMO radar. Utilizing a MIMO radar operating at 1–2 GHz, we capture activity data of individuals through walls and process it into range-angle maps to represent activity features. To tackle the issue of minimal variation in reflection areas caused by static individuals, a multi-scale activity feature extraction module is designed, capable of extracting effective features from radar signals across multiple scales. Simultaneously, a temporal attention mechanism is employed to extract keyframe information from sequential signals, focusing on critical moments of activity. Furthermore, this study introduces an activity recognition network based on a Deformable Transformer, which efficiently extracts both global and local features from radar signals, delivering precise human posture and activity sequences. In experimental scenarios involving 24 cm-thick brick walls, the proposed method achieves an impressive 97.1% accuracy in activity recognition classification.

KEYWORDS: MIMO radar; human activity; Transformer; through-wall

1 Introduction

In recent years, with the rapid advancement of technology, Human Activity Recognition (HAR) [1–3] has found extensive applications in smart homes [4,5], health monitoring [6,7], human-computer interaction [8], and security surveillance [9]. However, traditional activity recognition methods relying on visual sensors are often susceptible to external environmental factors such as lighting and weather, making it challenging to maintain stable performance in complex and dynamic real-world scenarios. In addition, the data collected by visual sensors frequently involve personal privacy, posing significant risks of privacy breaches. These limitations have, to some extent, constrained the breadth and depth of their practical applications.

Compared to visual sensors, certain radio frequency sensors, such as WiFi [10,11] and radar [12], can achieve activity recognition by analyzing received reflected echoes. However, WiFi devices suffer from relatively low localization accuracy and are highly susceptible to environmental changes. While millimeter-wave radar offers higher localization precision and stronger resistance to environmental interference, its

Copyright © 2025 The Authors. Published by Tech Science Press.
This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

high-frequency characteristics result in relatively weaker penetration capability. In contrast, multiple-input-multiple-output (MIMO) radar achieves localization accuracy comparable to millimeter-wave radar while offering superior penetration performance, making it better suited for activity recognition in complex environments. WiFi thermal imaging offers good real-time performance, typically completing data acquisition and processing in a relatively short time. Since it does not require additional sensors, the data collection process is relatively simple, and the processing speed is fast, making it suitable for scenarios requiring quick responses. However, WiFi has a lower resolution and is more suited for detecting the general position and basic posture of humans. In contrast, MIMO radar technology usually requires more complex algorithms for signal processing and data analysis to generate thermal maps and behavioral analysis results, resulting in longer processing times. Nevertheless, MIMO radar can effectively identify target behaviors in complex environments, such as obstructed spaces or between different rooms, making it particularly suitable for through-wall detection and micro-motion behavior detection.

In current mainstream radar-based human activity recognition methods [13,14], most rely on extracting micro-Doppler features and inputting these features into neural network models to predict target activity. However, this approach has certain limitations, primarily in its inability to associate predicted activities with the specific locations of individuals in the scene. Even with the use of Range-Doppler images [15], it remains difficult to effectively separate different targets when multiple individuals are located at the same distance. Furthermore, when recognizing activities of multiple targets, it typically requires separating the reflection areas of the targets first, and then performing individual activity recognition for each target, which inadvertently increases both the time cost and complexity of the recognition process. More challenging still, for stationary individuals, traditional methods struggle to detect and recognize activities, as they fail to produce significant changes in radar Doppler signals.

Through-wall human activity recognition technology holds significant practical value, particularly in search and rescue and security monitoring. By penetrating walls or obstacles to monitor human activities in real time, it not only enhances rescue efficiency in complex environments but also provides precise behavioral alerts for security protection. In disaster rescue operations, this technology enables the assessment of trapped individuals' conditions without direct contact, greatly improving the timeliness and accuracy of rescue efforts. In radar-based human activity recognition tasks, especially in through-wall scenarios, the radar signal reflections triggered by human movements are often weak and complex due to the obstruction caused by walls and the surrounding environment. In most cases, these signal variations are not sufficiently prominent, making traditional methods that rely on manual signal analysis and feature extraction ineffective in handling dynamic scenes and actions, thus failing to achieve satisfactory recognition results. However, the rapid development of deep learning technologies [16–19] in recent years has provided a new breakthrough for radar signal processing. Deep learning [20–22], with its powerful feature extraction capabilities, can capture complex contextual information from radar heatmaps, enabling automatic recognition and classification of various activities. However, recognizing subtle motion changes and stationary activities remains a significant challenge due to the extremely weak variations in the radar signal they induce. Therefore, it is essential to further optimize network structures and design feature extraction strategies specifically tailored for radar heatmaps to enhance the ability to distinguish micro-movements and stationary activities.

This study introduces a through-wall multihuman activity recognition system based on MIMO radar, named TW-MHAR. The system employs a MIMO frequency-modulated continuous wave radar operating in the 1–2 GHz band to capture reflection signals of human activities. This frequency band offers excellent penetration capabilities, allowing the radar to traverse common obstacles such as wood and bricks and generate high-precision radar reflection heatmaps, providing reliable data support for activity recognition in through-wall scenarios. To address the challenge of weak signals caused by certain activities, a novel

radar-based human activity recognition network, RHACNet, is proposed. RHACNet begins by extracting spatial features from each sequential radar signal to fully capture multi-temporal, multi-scale feature representations, effectively capturing the intricate details of complex activities. Subsequently, the network employs a temporal attention mechanism to integrate features across the temporal dimension, focusing on dynamic changes in key temporal features to enhance sensitivity to behavioral pattern variations. Moreover, an end-to-end multihuman activity classification module is incorporated, strengthening the system's ability to handle complex scenarios and enabling efficient classification of diverse human activities.

The main contributions of this paper are as follows:

- We propose TW-MHAR, a through-wall multihuman activity recognition system based on MIMO radar.
- We introduce RHACNet, designed to identify and distinguish signal variations of diverse micro-movements and even stationary activities across multiple temporal and spatial scales.
- We conduct comprehensive experiments on a multihuman activity recognition dataset collected in through-wall scenarios, demonstrating the effectiveness and advantages of the proposed method.

The remainder of this paper is organized as follows: Section 2 provides a review of related studies on human activity recognition using various sensors. Section 3 details the structure of the radar device and the signal processing methods employed in this study. Section 4 proposes the architecture of a real-time through-wall multihuman activity recognition system based on MIMO radar. Section 5 describes the experimental setups and the dataset used. Section 6 presents the experimental results and performance analysis. Finally, Section 7 concludes the paper with a summary and future outlook.

2 Related Work

This section will explore the application methods of various wireless sensors in activity recognition, including WiFi, RFID, and radar.

WiFi-based human activity recognition methods primarily infer human activities by analyzing WiFi Channel State Information (CSI) or Received Signal Strength Indicator (RSSI). Luo et al. [23] introduced the Transformer model into WiFi sensing applications and designed five HAR Vision Transformer architectures based on WiFi CSI. Jiao et al. [24] proposed a novel framework based on Gramian Angular Fields (GAFs), incorporating two modules: Gramian Angular Summation Fields (GASF) and Gramian Angular Difference Fields (GADF). This framework effectively extracts information from CSI and converts it into CSI-GAF images, followed by convolutional neural networks for activity prediction. Sheng et al. [25] presented a Cross-Domain Sensing Framework comprising a Nearest Neighbor-based Domain Selector (NNDS) and a Fine-to-Coarse Granular Transformer Network (FCGTN). NNDS evaluates the similarity between the source and target domains by measuring local and global feature distributions. Experiments demonstrate that this method achieves an activity recognition rate of 89.8% in cross-domain scenarios. Yadav et al. [26] proposed a universal framework for human activity recognition based on CSI. This framework utilizes two hybrid strategies for data augmentation, which are then fed into an enhanced InceptionTime network architecture. The approach achieved accuracies of 98.20%, 98%, and 95.42% on the ARIL, StanWiFi, and SignFi datasets, respectively.

RFID-based human activity recognition methods primarily infer behavioral patterns by analyzing the dynamic variations of RFID tags associated with the human body. Zhao et al. [27] proposed a non-wearable RFID-based human motion recognition approach that integrates phase and RSSI data to enhance data diversity. They developed a combined processing method to effectively eliminate device-induced thermal noise and reduce environmental interference, while employing a spatio-temporal graph convolutional neural network to construct an efficient classification model for human motion signals, achieving an overall

recognition accuracy of 92.8%. Qiu et al. [28] introduced LD-Recognition, an RFID-based classroom motion recognition system. This system utilizes a multi-channel attention graph convolutional neural network to deeply analyze motion-related phase and signal strength, achieving a recognition accuracy as high as 96.9%.

Radar-based human activity recognition methods identify activities by transmitting signals and analyzing echoes reflected from the human body. Song et al. [29] proposed a framework for reconstructing human poses and classifying activities using 4D radar imaging. This framework employs ultra-wideband MIMO radar as the detection sensor to capture 4D imaging data, including range, azimuth, elevation, and time. A 3D convolutional neural network-based pose reconstruction model is used to generate 3D human poses, while a dual-branch network leveraging multi-frame 3D poses and 4D radar images classifies activities, achieving a recognition accuracy of 94.75%. Froehlich et al. [30] addressed the challenge of measuring lateral velocity, which is unattainable with a single radar, by employing a radar network comprising two spatially orthogonal millimeter-wave MIMO radars. They designed a radar activity recognition network combining CNN and LSTM for the dual radar data, achieving an average recognition rate of 74.44% across four activities. Zhu et al. introduced a hybrid classifier combining convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for spatiotemporal pattern extraction. They first applied 2D CNNs to radar data to extract spatial features from spectrograms. Subsequently, bidirectional gated recurrent units (Bi-GRU) were used to capture both long-term and short-term temporal dependencies in the feature maps generated by 2D CNN, achieving a classification accuracy of approximately 90.8% across nine categories of human activity. Yu et al. [31] proposed a novel non-intrusive human activity recognition system using millimeter-wave radar. The system first converts millimeter-wave signals into point clouds and employs an improved voxelization method to account for the spatiotemporal point clouds in the physical environment. A dual-view convolutional neural network is then used to learn human activities from the sparse data enriched by the radar's rotational symmetry. The system achieved accuracies of 97.61% and 98% in fall detection and activity classification tasks, respectively, across datasets with seven different activities.

3 Radar Architecture and Signal Processing

The radar system used in this study employs a 4-transmitter, 16-receiver antenna array MIMO radar [32] with a center frequency of 1.5 GHz, a bandwidth of 1 GHz, and a sweep period of 0.05 s. The transmitted signal from the i-th transmitting antenna of the radar at the k-th moment is represented as shown in (1):

$$s_{k,i}(t) = \sqrt{2P_{k,i}}e^{2j\pi f_c + \frac{\pi j k t^2}{T_p}} \tag{1}$$

The transmitted power is denoted as $P_{k,i}$, where f_c is the radar's center frequency, k represents the chirp rate of the linear frequency-modulated pulse, and T_p is the sweep period of the linear frequency-modulated pulse. After penetrating the wall and reflecting off the human body, the received echo signal can be expressed as shown in (2):

$$I_{m,n}(t) = A(t)e^{2j\pi \frac{S_{m,n}}{\lambda} t} \tag{2}$$

The term $A(t)$ represents the amplitude at time t, λ denotes the wavelength corresponding to the radar's center frequency, and $S_{m,n}$ represents the total path distance traveled by the chirped signal transmitted from the m-th transmitting antenna, reflected, and received by the n-th receiving antenna. Subsequently, the sampled signal from a transmit-receive antenna pair is combined and processed using a three-dimensional

Fourier transform to compute the final 3D power matrix. The specific calculation is shown in (3):

$$p(\varphi,\theta,\gamma) = \left| \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{t=1}^{T} I_{m,n}(t) e^{j2\pi\frac{kr}{c}t} e^{j\frac{2\pi}{\lambda}sin\theta(ns_n cos\theta + ms_m sind\varphi)} \right| \tag{3}$$

where variables φ, θ, and γ represent the elevation angle, azimuth angle, and range, respectively. The parameter k denotes the chirp slope, c is the signal propagation speed, s_m and s_n indicate the spacing of the transmitting and receiving antennas, respectively. These parameters are critical in determining the spatial and temporal resolution of the radar system, as well as in constructing the accurate three-dimensional power matrix through the signal processing pipeline.

The power matrix P encapsulates the reflection signals of all objects in the scene. To isolate human reflection signals, a background subtraction method is employed. Specifically, multiple frames of scene reflection signals are continuously captured and averaged to generate background data. Subsequent frames are then subtracted from this averaged background, producing what is referred to as radar static heatmaps. To further extract motion information of humans in the scene, a Doppler static removal method is introduced. By calculating the Doppler velocity of spatial points across multiple frames and filtering out zero-velocity components, the motion velocity information of the target is extracted, resulting in radar dynamic heatmaps. Next, horizontal and vertical projections are performed on both types of radar heatmaps, generating radar static horizontal/vertical heatmaps and radar dynamic horizontal/vertical heatmaps. This processing approach not only significantly reduces the parameter count required for network training but also retains critical information from the scene, facilitating efficient learning and inference by the model.

4 Methods

This paper proposes a system for through-wall multihuman activity recognition, named TW-MHAR, with its framework illustrated in Fig. 1. The system primarily consists of a Radar Human Activity Recognition Network (RHACNet) and a visual motion capture system. The radar and visual systems are connected to a terminal via a synchronization interface to enable collaborative and synchronized data analysis. In the system architecture, the radar device is positioned behind the wall to capture through-wall signals reflected by human bodies, while the visual motion capture system is deployed within the wall to accurately record 3D human pose sequences. Each captured pose sequence is manually annotated with corresponding activity labels, which are subsequently used as supervised training signals for the RHACNet network. RHACNet forms the core of this system and comprises two main components: the radar feature extraction module and the human activity decoding module. The radar feature extraction module is responsible for extracting deep features related to human poses and activities from the spatial and temporal dimensions of sequential radar data. By constructing multi-scale and multi-dimensional feature representations, this module effectively mines valuable information from radar signals. The human activity decoding module takes the abstract features output by the radar feature extraction module and decodes them into corresponding human pose and activity sequences, enabling precise activity recognition. Through this design, TW-MHAR achieves efficient transformation from radar signals to activity semantics while leveraging the precise pose data provided by the visual motion capture system to supply high-quality supervision for radar network training. The following sections will delve into the detailed designs of the radar feature extraction module and the human activity decoding module. Micro-movements are often misclassified as noise or background interference, especially at low resolutions. However, by integrating multi-scale features, the system can model signals across multiple scales simultaneously, enhancing sensitivity to these subtle changes. During through-wall detection, different sensors or radar systems may introduce variations in perspective or resolution due to environmental and hardware differences. This makes single-scale feature extraction insufficient to

accommodate all scenarios effectively. By leveraging multi-scale feature fusion, the model can adapt to varying resolutions and environmental changes, improving robustness in dynamic settings.

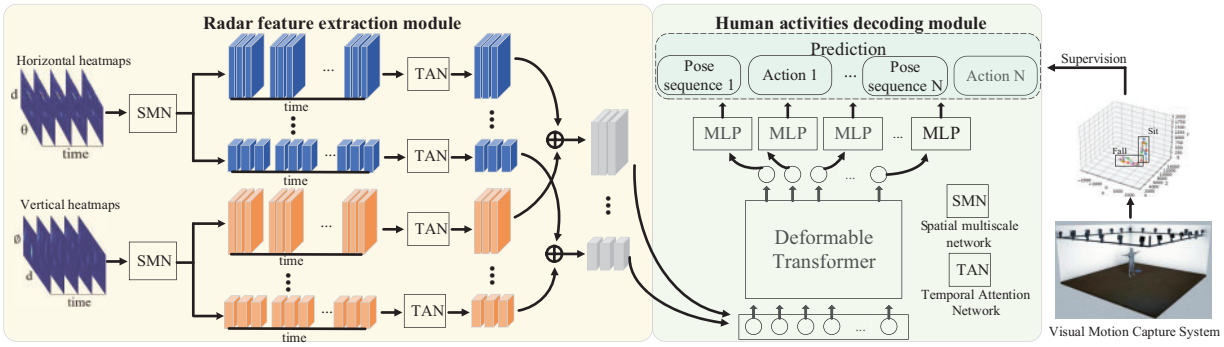

Figure 1: The framework of the through-wall multihuman activity recognition system. The system framework begins by extracting features from two types of radar heatmaps through the radar feature extraction module. Subsequently, the human activity decoding module processes these features to derive human poses and activities. The labeling component utilizes precise annotations provided by the vision motion capture system for supervised training

4.1 Radar Feature Extraction Module

In this network, the input consists of sequential horizontal radar heatmaps $I_h = \{x_1, x_2, \ldots, x_N\}$ and vertical radar heatmaps $I_v = \{y_1, y_2, \ldots, y_N\}$, where N represents the length of the temporal frames of the radar. First, we employ a Spatial Multi-scale Network (SMN) to extract multi-scale features from each frame of the radar heatmaps. The SMN is composed of multiple 2D convolution layers, pooling layers, normalization layers, and stacked residual structures, designed to extract spatial features at various scales from the raw heatmaps. By applying SMN to each frame of the horizontal and vertical radar heatmaps, we obtain multi-scale features denoted as $f_h = \left\{\left\{x_1^{1'}, x_1^{2'}, \ldots, x_1^{M'}\right\}, \left\{x_2^{1'}, x_2^{2'}, \ldots, x_2^{M'}\right\}, \ldots, \left\{x_N^{1'}, x_N^{2'}, \ldots, x_N^{M'}\right\}\right\}$ and $f_v = \left\{\left\{y_1^{1'}, y_1^{2'}, \ldots, y_1^{M'}\right\}, \left\{y_2^{1'}, y_2^{2'}, \ldots, y_2^{M'}\right\}, \ldots, \left\{y_N^{1'}, y_N^{2'}, \ldots, y_N^{M'}\right\}\right\}$ where M represents the number of scales in feature extraction. Next, we use a Temporal Attention Network (TAN) to aggregate temporal features from the sequential radar heatmaps. TAN employs multiple stacked feedforward neural networks to compute temporal attention weights for each radar frame. The output of each feedforward network represents the feature weight at the current time step, which is normalized using a softmax function to ensure consistency in the weighted sum of features across time steps. Through this process, we perform weighted summation of features across time steps, obtaining feature vectors representing multi-scale temporal context information, denoted as $f_h' = \left\{\left\{k_1^1 x_1^{1'}, k_2^1 x_2^{1'}, \ldots, k_N^1 x_N^{1'}\right\}, \left\{k_1^2 x_1^{2'}, k_2^2 x_2^{2'}, \ldots, k_N^2 x_N^{2'}\right\}, \ldots, \left\{k_1^M x_1^{M'}, k_2^M x_2^{M'}, \ldots, k_N^M x_N^{M'}\right\}\right\}$ and $f_v' = \left\{\left\{q_1^1 y_1^{1'}, q_2^1 y_2^{1'}, \ldots, q_N^1 y_N^{1'}\right\}, \left\{q_1^2 y_1^{2'}, q_2^2 y_2^{2'}, \ldots, q_N^2 y_N^{2'}\right\}, \ldots, \left\{q_1^M y_1^{M'}, q_2^M y_2^{M'}, \ldots, q_N^M y_N^{M'}\right\}\right\}$, where k and q denote the temporal attention weights in the horizontal and vertical radar heatmaps, respectively. Finally, to integrate information from both radar heatmaps, we perform element-wise addition of intermediate features at the same scale from the horizontal and vertical heatmaps. This operation facilitates the fusion of multi-scale spatial and temporal information from different perspectives, resulting in a richer and more accurate representation. The final output feature $f = f_h' + f_v' = \left\{z^1, z^2, \ldots, z^M\right\}$ of the network encapsulates comprehensive spatiotemporal contextual information for subsequent task processing. This approach effectively extracts and integrates multi-scale spatial and temporal features from sequential radar heatmaps, enhancing the network's representational capacity for handling complex scenarios. By improving

feature extraction precision and capturing variations in targets across different time steps and spatial scales, this method significantly boosts recognition performance in challenging environments.

4.2 Human Activity Decoding Module

In traditional Transformer feature mapping methods, querying every position in the feature map leads to extremely high computational costs, especially when dealing with high-dimensional feature maps. To address this issue and improve efficiency, we introduce the Deformable Transformer [33], which replaces global queries with deformable convolutions. This approach enables the attention module to focus only on a few sampled points in the local regions of the feature map. Such localized querying significantly reduces computational overhead while allowing the network to concentrate on key local features. For the output features f from the Radar Feature Extraction Module, we first flatten the width and height dimensions at each scale. For each query element q, the deformable attention module is defined as:

$$DAM(z_q, p_q, f) = \sum_{z=1}^{M} W_z[\sum_{l=1}^{L}\sum_{k=1}^{K} A_{mlqk} W'_z f(\phi(p_1 + \Delta p_{mlqk}))] \tag{4}$$

where p_q represents the query element at the reference point q, z_q denotes the feature value at point q, m refers to the index of the multi-head attention head, l indicates the scale level of the multi-scale features, and k is the index of the sampling point. A_{mlqk} and Δp_{mlqk} correspond to the attention weight and sampling point offset for the k-th sampling point of the m-th attention head at the l-th scale, respectively. Through the Deformable Attention Module, the sampling points for the query elements are no longer fixed but dynamically adjusted based on the offsets Δp_{mlqk}. This allows the module to focus on regions of the feature map with higher local relevance. After processing through multiple stacked multi-head, multi-scale attention modules, the resulting features are passed into a Feed-Forward Network (FFN) for further refinement and enhancement. In the FFN, the features are processed through a multi-layer perceptron (MLP), ultimately outputting multiple sets of human pose candidate sequences $P \in \mathbb{R}^{n\times p\times 3}$ and human activity sequences $A \in \mathbb{R}^{n\times(c+1)}$, where n represents the number of predicted candidate targets, p denotes the number of predicted keypoints, c indicates the number of predicted activity classes. A prediction of $c+1$ for an activity implies that the candidate target corresponds to the background. This approach allows the Deformable Transformer to not only significantly improve processing efficiency but also enhance the model's adaptability and representational power by dynamically adjusting its attention to relevant regions. In tasks such as human pose recognition and activity prediction, it excels at capturing local features and dynamic variations of the targets with greater precision.

The final output of the network encompasses two tasks: pose regression and activity prediction. For the pose regression task, we employ L1Loss to calculate the regression loss, aiming to minimize the absolute error between the predicted and ground truth keypoint positions. For the activity prediction task, we use cross-entropy loss to calculate the classification loss, targeting the minimization of the difference between the predicted and ground truth activity classes. Assuming the human pose labelsP' are and the activity labelsA' are, the loss functions are computed as follows:

$$\zeta = \frac{1}{K}\sum_{i=0}^{K}[-A' log A - (1-A) log(1-A) + |p - p'|] \tag{5}$$

5 Experimental Setup

The experimental setup includes a 24 cm brick wall, with the radar positioned outside the wall and a ring of motion capture devices placed inside to capture precise human pose and activity data. The experimental area spans a 6 × 15 meter range behind the wall, where participants freely perform six types of activities,

including walking, sitting, and lying down. At most, three individuals are present in the scene simultaneously. The radar and motion capture devices are synchronized and connected to a terminal, collecting data at a rate of 20 frames per second. A total of 1,400,000 frames of data were recorded, which were split into training and testing sets in a 4:1 ratio. Some of the data is shown in Fig. 2.

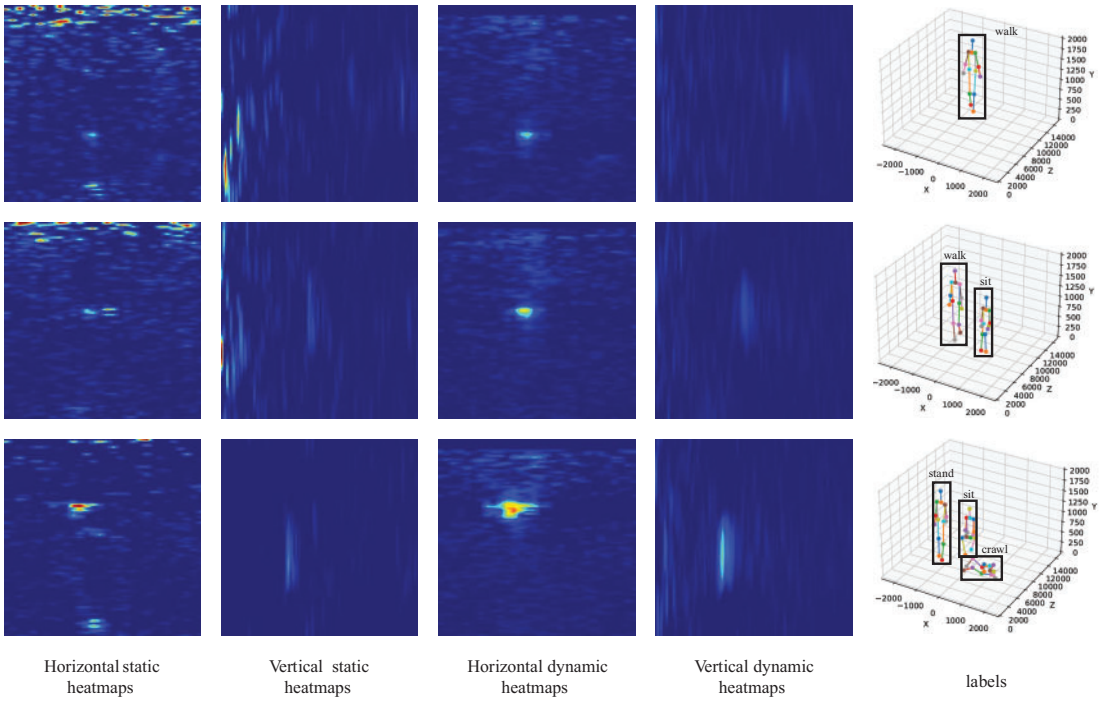

Figure 2: Data captured by radar and labels captured by motion capture devices in scenes with 1 to 3 people

6 Results

6.1 Performance Comparison with Existing Methods

The performance comparison results of the algorithm proposed in this paper and the existing methods are shown in Table 1. Zhao et al. [34] proposed a method using single-frame raw IQ radar data as input, where the model extracts features from individual frames to recognize human activities. In their experiments, the method achieved an accuracy of 94.7%, precision of 95.0%, recall of 94.7%, and an F1 score of 94.8%. Although it performed well in terms of precision, its limited use of temporal information in radar signals constrained its ability to recognize subtle motions effectively. Wang et al. [35] utilized a sequence of 32 consecutive range-angle maps as input, modeling dynamic activity changes by incorporating the temporal dimension. Their method achieved an accuracy of 96.0%, precision of 94.4%, recall of 94.9%, and an F1 score of 94.7%. Compared to Zhao et al., Wang et al.'s approach significantly improved accuracy and recall, demonstrating better capability in modeling the temporal dynamics of activities. The method proposed in this paper further enhances the recognition of subtle motions by comprehensively considering both temporal and spatial information. Specifically, our approach introduces a multi-scale feature extraction strategy to improve the model's sensitivity to fine signal variations, excelling in recognizing micro-movements or continuous actions. Additionally, by integrating attention mechanisms, our model effectively captures the temporal continuity of signals, enhancing the robustness of activity recognition. In experiments, our method achieved an accuracy of 97.1%, precision of 94.7%, recall of 95.9%, and an F1 score of 95.2%. Compared to existing

methods, the proposed approach demonstrates significant advantages in accuracy and recall, highlighting its superior adaptability in capturing subtle motions and handling complex dynamic backgrounds.

Table 1: Performance comparison with existing advanced algorithms

Methods	Accuracy	Precision	Recall	F1 score
Zhao et al. [34]	94.7%	95.0%	94.7%	94.8%
Wang et al. [35]	96.0%	94.4%	94.9%	94.7%
Ours	97.1%	94.7%	95.9%	95.2%

Fig. 3 illustrates the confusion matrix of the proposed method. Despite the attenuation of through-wall radar signals after penetrating the wall, the vast majority of samples are correctly classified. Dynamic activities such as "walking" and "crawling" demonstrate stronger recognition capabilities due to their pronounced temporal and spatial variations in radar signals. In contrast, static activities like "sitting" and "standing" exhibit slightly higher misclassification rates. However, by applying multi-scale features and temporal attention mechanisms, the model effectively distinguishes these subtle signal characteristics. Fig. 4 presents the Principal Component Analysis (PCA) [36] visualization of various activities. This visualization highlights the clustering of different activity classes in the feature space, demonstrating the model's ability to extract discriminative features for robust activity recognition. Table 2 presents the computational complexity and runtime of RHACNet and its submodules, Spatial Multiscale Network (SMN) and Temporal Attention Network (TAN). RHACNet, the entire network, has the highest FLOPs at 91.07 G and 38.32 M parameters, requiring 0.13 s for execution. SMN, a submodule of RHACNet, is more efficient with 8.20 G FLOPs and 25.56 M parameters, taking only 0.03 s. TAN, another submodule, requires 30.43 G FLOPs and 6.8 M parameters but is the most computationally efficient, with a runtime of just 0.01 s.

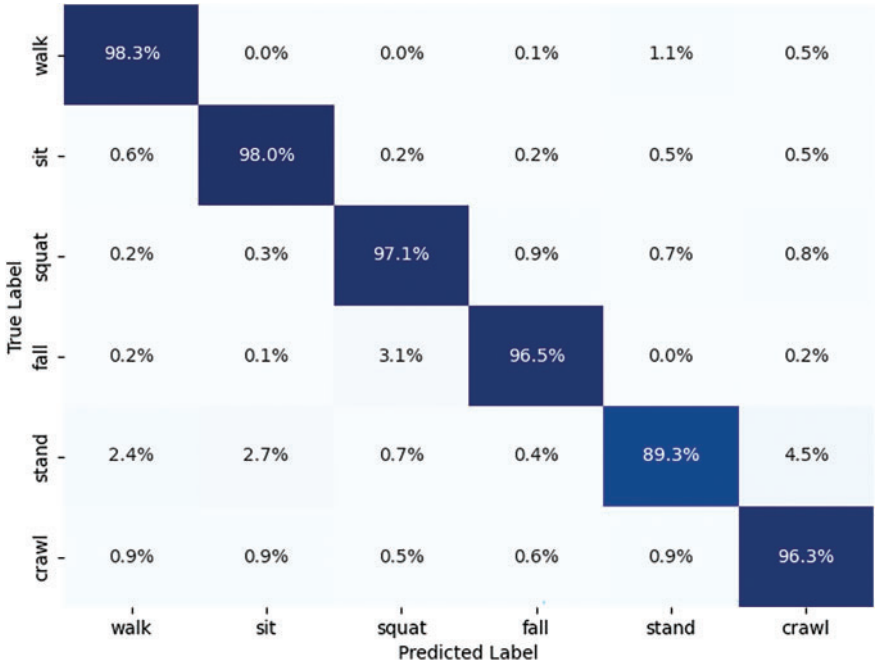

Figure 3: Confusion matrix of classification results for different human activities

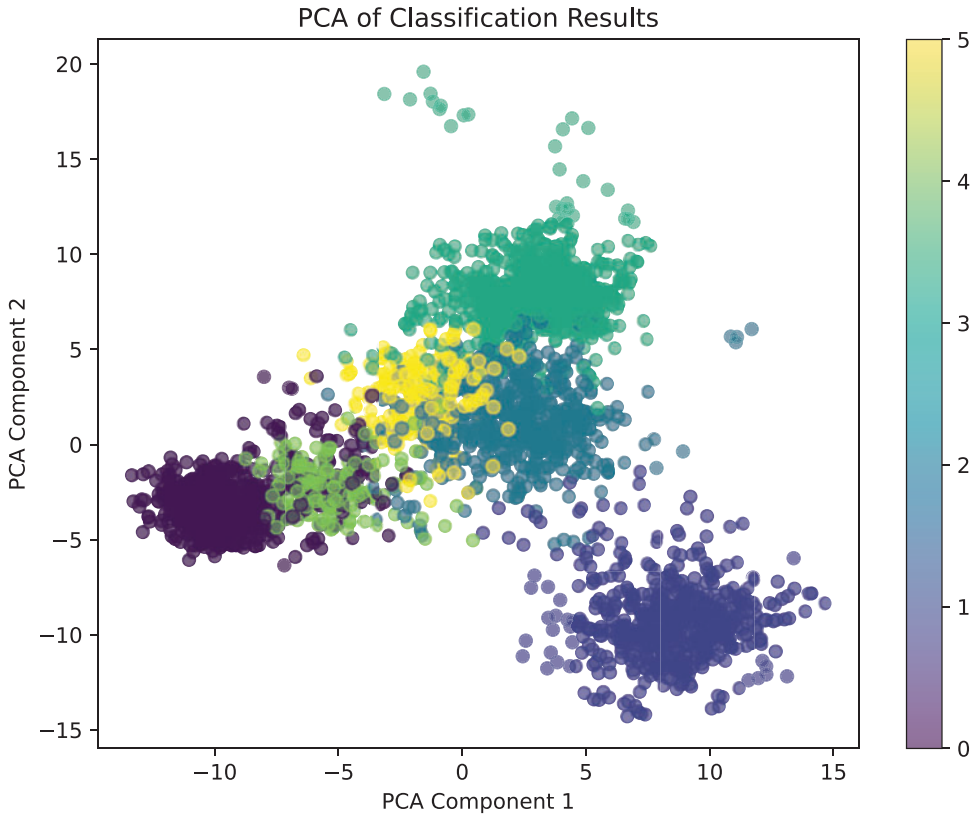

Figure 4: PCA visualization of different human activities

Table 2: Cost of network modules

Modules	FLOPs (G)	Params (M)	Times (s)
RHACNet	91.07	38.32	0.13
SMN	8.20	25.56	0.03
TAN	30.43	6.8	0.01

6.2 Impact of Trainning Parameters

Fig. 5 illustrates the impact of the number of temporal frames on the final recognition accuracy. It can be observed that the recognition rate exhibits a gradual upward trend as the number of frames increases. When the frame count reaches 32, the recognition accuracy peaks, after which it shows a slight decline. This phenomenon occurs because the increase in temporal frames provides richer sequential information. With fewer frames, the dynamic features captured by the model are limited, resulting in relatively lower recognition accuracy. As the number of frames increases, more temporal information is introduced, enhancing the network's ability to recognize activity patterns, and thereby improving the recognition rate. However, when the number of temporal frames exceeds a certain threshold, the additional sequential information may become redundant. This redundancy contributes little to further improving recognition performance and may even lead to increased focus on irrelevant features, reducing the model's feature extraction effectiveness.

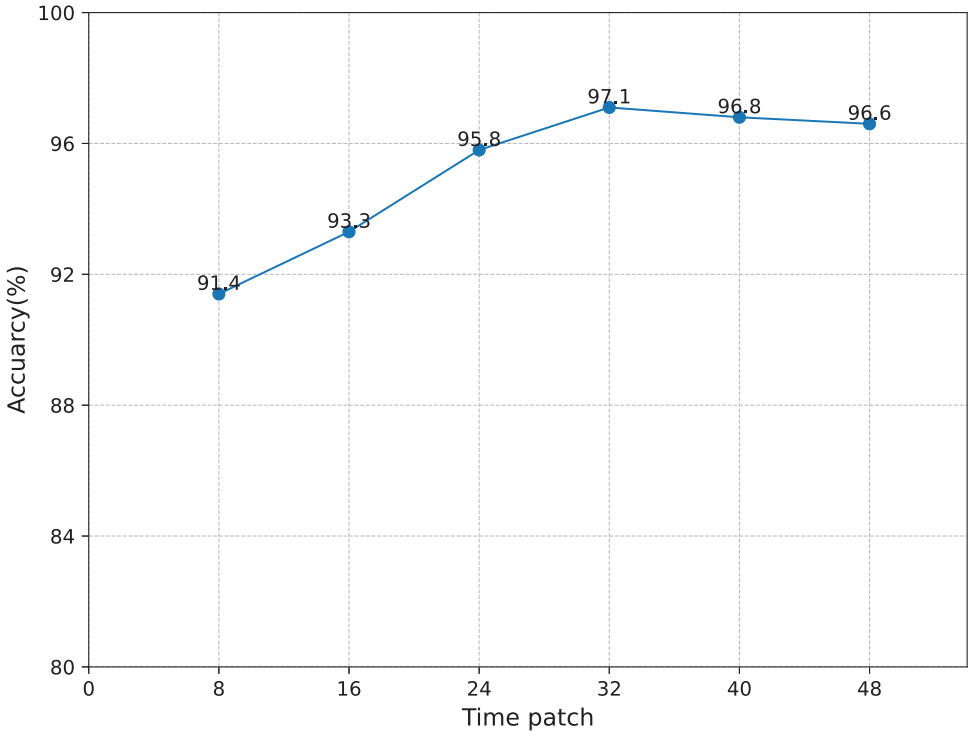

Figure 5: The influence of input radar time frame length on recognition rate

Fig. 6 illustrates the impact of the number of predicted candidate targets on recognition accuracy. It can be observed that as the number of candidate targets increases, the recognition rate gradually improves. However, once the number of candidates exceeds a certain threshold, the accuracy begins to decline slowly. This phenomenon can be attributed to the dual effects of resolution and sparsity influenced by the number of candidate targets. When the number of predicted candidates is small, the model lacks sufficient resolution, causing multiple radar reflection regions to be incorrectly aggregated into the same candidate target location. This aggregation effect significantly reduces recognition accuracy, particularly in multi-target or complex scenarios. As the number of candidate targets increases, the model's resolution improves, allowing more precise separation and identification of individual target regions, thereby enhancing overall recognition performance. However, when the number of predicted candidates exceeds a reasonable range, the model encounters sparsity issues in practical applications. In such cases, candidate positions become overly dense relative to the actual number of targets, leading to the allocation of substantial resources to invalid candidate regions, ultimately hindering the model's effectiveness.

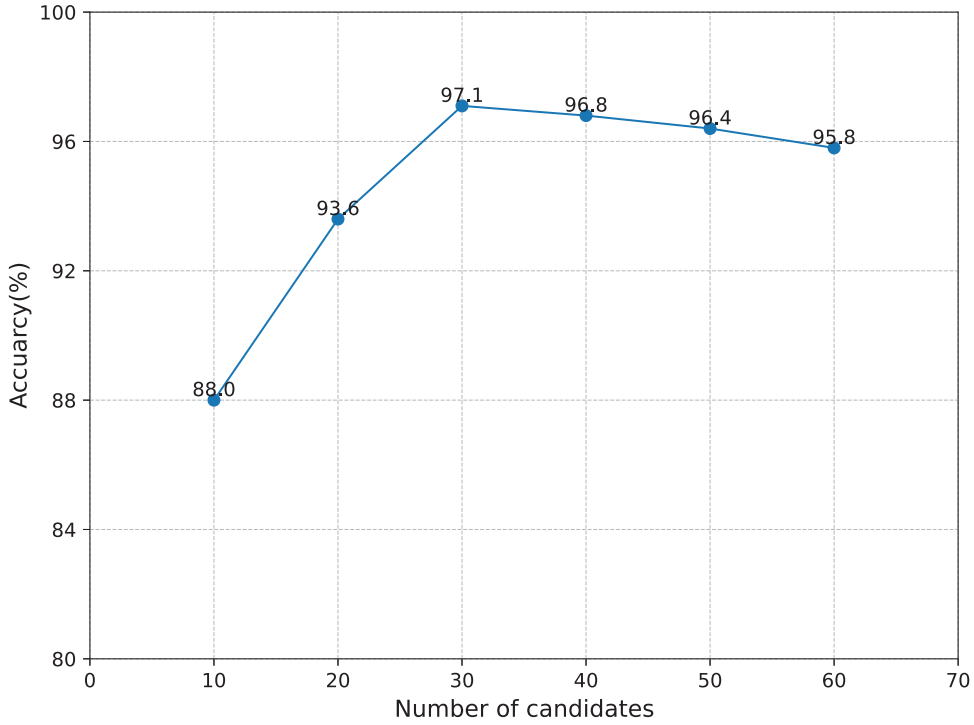

Figure 6: The impact of the number of prediction candidates on the recognition rate

7 Conclusion

This paper presents a through-wall human activity recognition method based on MIMO radar. By extracting behavioral features from multi-scale radar signals and incorporating temporal attention mechanisms to capture keyframes in sequential signals, the proposed method effectively identifies human activities in through-wall environments. Leveraging the Deformable Transformer module, it efficiently extracts both global and local features, directly outputting human pose and activity sequences. In experimental scenarios involving a 24 cm thick brick wall, the proposed method achieves an activity recognition accuracy of 97.1%, significantly outperforming existing approaches and demonstrating its performance advantages in complex scenarios.

Although the proposed method demonstrates outstanding performance in activity recognition tasks, there are still areas for optimization. Future work will focus on enhancing the predictive upper limit of the network: due to the inherent resolution limitations of radar signals, the ability to predict multiple targets may be constrained. To address this, we plan to redesign the sampling and imaging methods of radar signals to improve the system's resolution for multi-target activities, thereby further increasing the predictive upper limit of the system. Another focus is improving the recognition of static activities: for prolonged static human activities, the lack of prominent Doppler features often causes radar reflection signals to be overwhelmed by background noise, leading to misclassification as background information. To tackle this challenge, we aim to design more refined feature extraction modules to enhance the network's sensitivity to static target signals, enabling it to accurately distinguish static human signals from background noise.

Acknowledgement: The authors are grateful to all the editors and anonymous reviewers for their comments and suggestions.

Funding Statement: This research was supported by National Natural Science Foundation of China (No. 62272242) and Postgraduate Research & Practice Innovation Program of Jiangsu Province (Nos. KYCX21_0800, KYCX23_1082).

Author Contributions: Writing—original draft preparation, Changlong Wang; supervision, Chong Han; resources, Jiawei Jiang and Hengyi Ren; formal analysis, Lijuan Sun and Jian Guo. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Wang X, Wu Z, Jiang B, Bao Z, Zhu L, Li G, et al. HARDVS: revisiting human activity recognition with dynamic vision sensors. Proc AAAI Conf Artif Intell. 2024;38(6):5615–23. doi:10.1609/aaai.v38i6.28372.
2. Raj MS, George SN, Raja K. Leveraging spatio-temporal features using graph neural networks for human activity recognition. Pattern Recognit. 2024;150(3):110301. doi:10.1016/j.patcog.2024.110301.
3. Islam MM, Nooruddin S, Karray F, Muhammad G. Multi-level feature fusion for multimodal human activity recognition in Internet of Healthcare Things. Inf Fusion. 2023;94(12):17–31. doi:10.1016/j.inffus.2023.01.015.
4. Huda NU, Ahmed I, Adnan M, Ali M, Naeem F. Experts and intelligent systems for smart homes' transformation to sustainable smart cities: a comprehensive review. Expert Syst Appl. 2024;238(9):122380. doi:10.1016/j.eswa.2023.122380.
5. Diraco G, Rescio G, Siciliano P, Leone A. Review on human action recognition in smart living: sensing technology, multimodality, real-time processing, interoperability, and resource-constrained processing. Sensors. 2023;23(11):5281. doi:10.3390/s23115281.
6. Kulkarni MB, Rajagopal S, Prieto-Simón B, Pogue BW. Recent advances in smart wearable sensors for continuous human health monitoring. Talanta. 2024;272:125817. doi:10.1016/j.talanta.2024.125817.
7. Chen J, Wang W, Fang B, Liu Y, Yu K, Leung VC, et al. Digital twin empowered wireless healthcare monitoring for smart home. IEEE J Sel Areas Commun. 2023;41(11):3662–76. doi:10.1109/JSAC.2023.3310097.
8. Lyu Z. State-of-the-art human-computer-interaction in metaverse. Int J Hum-Comput Interact. 2024;40(21):6690–708. doi:10.1080/10447318.2023.2248833.
9. Kazeminajafabadi A, Imani M. Optimal joint defense and monitoring for networks security under uncertainty: a POMDP-based approach. IET Inf Secur. 2024;2024(1):7966713. doi:10.1049/2024/7966713.
10. Sheng B, Li J, Gui L, Guo Z, Xiao F. LiteWiSys: a Lightweight system for WiFi-based dual-task action perception. ACM Trans Sensor Netw. 2024;20(4):1–19. doi:10.1145/3632177.
11. Moghaddam MG, Shirehjini AAN, Shirmohammadi S. A WiFi-based method for recognizing fine-grained multiple-subject human activities. IEEE Trans Instrum Meas. 2023;72:1–13. doi:10.1109/TIM.2023.3289547.
12. Sørensen KA, Kusk A, Heiselberg P, Heiselberg H. Finding ground-based radars in SAR images: localizing radio frequency interference using unsupervised deep learning. IEEE Trans Geosci Remote Sensing. 2023;61:1–15.
13. Li X, Zhang S, Chen S, Xiao Z. Through-wall multi-person action recognition using enhanced YOLOv5 and IR-UWB radar. IEEE Sens J. 2025;25(3):5711–22.
14. Yin W, Shi LF, Shi Y. Indoor human action recognition based on millimeter-wave radar micro-Doppler signature. Measurement. 2024;235:114939. doi:10.1016/j.measurement.2024.114939.
15. Acar YE, Ucar K, Saritas I, Yaldiz E. Classification of human target movements behind walls using multi-channel range-doppler images. Multimed Tools Appl. 2024;83(18):56021–38. doi:10.1007/s11042-023-17759-8.

16. Geng H, Hou Z, Liang J, Li X, Zhou X, Xu A. Motion focus global-local network: combining attention mechanism with micro action features for cow behavior recognition. Comput Electron Agric. 2024;226(4):109399. doi:10.1016/j.compag.2024.109399.
17. Li C, Huang Q, Mao Y, Li X, Wu J. Multi-granular spatial-temporal synchronous graph convolutional network for robust action recognition. Expert Syst Appl. 2024;257(10):124980. doi:10.1016/j.eswa.2024.124980.
18. Ren H, Sun L, Fan X, Cao Y, Ye Q. IIS-FVIQA: finger vein image quality assessment with intra-class and inter-class similarity. Pattern Recognit. 2025;158(1):111056. doi:10.1016/j.patcog.2024.111056.
19. Ren H, Sun L, Ren J, Cao Y. FV-DDC: a novel finger-vein recognition model with deformation detection and correction. Biomed Signal Process Control. 2025;100(2):107098. doi:10.1016/j.bspc.2024.107098.
20. Xue Q, Zhang X, Zhang Y, Hekmatmanesh A, Wu H, Song Y, et al. Non-contact rPPG-based human status assessment via feature fusion embedding anti-aliasing in industry. Comput Ind. 2025;165(4):104227. doi:10.1016/j.compind.2024.104227.
21. Yu X, Liang X, Zhou Z, Zhang B. Multi-task learning for hand heat trace time estimation and identity recognition. Expert Syst Appl. 2024;255(4):124551. doi:10.1016/j.eswa.2024.124551.
22. Yu X, Liang X, Zhou Z, Zhang B, Xue H. Deep soft threshold feature separation network for infrared handprint identity recognition and time estimation. Infrared Phys Technol. 2024;138(4):105223. doi:10.1016/j.infrared.2024.105223.
23. Luo F, Khan S, Jiang B, Wu K. Vision transformers for human activity recognition using WiFi channel state information. IEEE Internet Things J. 2024;11(17):28111–22. doi:10.1109/JIOT.2024.3375337.
24. Jiao W, Zhang C. An efficient human activity recognition system using wifi channel state information. IEEE Syst J. 2023;17(4):6687–90. doi:10.1109/JSYST.2023.3293482.
25. Sheng B, Han R, Cai H, Xiao F, Gui L, Guo Z. CDFi: cross-domain action recognition using WiFi signals. IEEE Trans Mob Comput. 2024;23(8):8463–77. doi:10.1109/TMC.2023.3348939.
26. Yadav SK, Sai S, Gundewar A, Rathore H, Tiwari K, Pandey HM, et al. CSITime: privacy-preserving human activity recognition using WiFi channel state information. Neural Netw. 2022;146(11):11–21. doi:10.1016/j.neunet.2021.11.011.
27. Zhao C, Wang L, Xiong F, Chen S, Su J, Xu H. RFID-based human action recognition through spatiotemporal graph convolutional neural network. IEEE Internet Things J. 2023;10(22):19898–912. doi:10.1109/JIOT.2023.3282680.
28. Qiu Q, Wang T, Chen F, Wang C. LD-recognition: classroom action recognition based on passive RFID. IEEE Trans Comput Soc Syst. 2023;11(1):1182–91. doi:10.1109/TCSS.2023.3234423.
29. Song Y, Dai Y, Jin T, Song Y. Dual-task human activity sensing for pose reconstruction and action recognition using 4-D imaging radar. IEEE Sens J. 2023. doi:10.1109/JSEN.2023.3308788.
30. Froehlich AC, Mejdani D, Engel L, Braeunig J, Kammel C, Vossiek M, et al. A millimeter-wave MIMO radar network for human activity recognition and fall detection. In: 2024 IEEE Radar Conference (RadarConf24); 2024; Piscataway, NJ, USA: IEEE. p. 1–5.
31. Yu C, Xu Z, Yan K, Chien YR, Fang SH, Wu HC. Noninvasive human activity recognition using millimeter-wave radar. IEEE Syst J. 2022;16(2):3036–47. doi:10.1109/JSYST.2022.3140546.
32. Wang C, Han C, Gao X, Ren H, Sun L, Guo J. RTMP-ID: real-time through-wall multi-person identification based on MIMO radar. IEEE Internet Things J. 2024. doi:10.1109/JIOT.2024.3510941.
33. Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable DETR: deformable transformers for end-to-end object detection. arXiv:2010.04159. 2020.
34. Zhao R, Ma X, Liu X, Liu J. An end-to-end network for continuous human motion recognition via radar radios. IEEE Sensors J. 2020;21(5):6487–96. doi:10.1109/JSEN.2020.3040865.
35. Wang C, Zhu D, Sun L, Han C, Guo J. Real-Time through-wall multihuman localization and behavior recognition based on MIMO radar. IEEE Trans Geosci Remote Sensi. 2023;61:1–12. doi:10.1109/TGRS.2023.3335484.
36. Yang J, Zhang D, Frangi AF, Yang JY. Two-dimensional PCA: a new approach to appearance-based face representation and recognition. IEEE Trans Pattern Anal Mach Intell. 2004;26(1):131–7. doi:10.1109/TPAMI.2004.1261097.