IEEE Journal of Microwaves

Received 4 November 2024; revised 29 January 2025; accepted 4 February 2025. Date of publication 7 March 2025; date of current version 14 March 2025.

Digital Object Identifier 10.1109/JMW.2025.3539957

Human Activity Recognition Based on Feature Fusion of Millimeter Wave Radar and Inertial Navigation

JIAJIA SHI [1] (Member, IEEE), YIHAN ZHU[1], JIAQING HE [1], ZHIHUO XU [1] (Senior Member, IEEE), LIU CHU [2] (Senior Member, IEEE), ROBIN BRAUN [3] (Life Senior Member, IEEE), AND QUAN SHI [1] (Member, IEEE)

(Regular Paper)

[1]School of Transportation and Civil Engineering, Nantong University, Nantong 226007, China
[2]School of Physical Science and Technology, ShanghaiTech University, Shanghai 201210, China
[3]School of Electrical and Data Engineering, University of Technology Sydney, Ultimo, NSW 2007, Australia

CORRESPONDING AUTHOR: Quan Shi (e-mail: sq@ntu.edu.cn).

This work was supported in part by the National Natural Science Foundation of China under Grant 62476145, Grant 62471260, Grant 12102203, and Grant 61901235; in part by the Natural Science Foundation of Jiangsu Province under Grant BK20231336; in part by the 6th "333 Talents" Technology Research and Development Talent Foundation of Jiangsu Province Transportation Technology and Achievement Transformation Foundation of Jiangsu Province under Grant 2024G01; in part by the Key Research Project of Nantong under Grant GZ2024001; and in part by Fire and Rescue Bureau Research Program under Grant 2019XFCX31.

ABSTRACT Human activity recognition (HAR) technology is increasingly utilized in domains such as security surveillance, nursing home monitoring, and health assessment. The integration of multi-sensor data improves recognition efficiency and the precision of behavioral analysis by offering a more comprehensive view of human activities. However, challenges arise due to the diversity of data types, dimensions, sampling rates, and environmental disturbances, which complicate feature extraction and data fusion. To address these challenges, we propose a HAR approach that fuses millimeter-wave radar and inertial navigation data using bimodal neural networks. We first design a comprehensive data acquisition framework that integrates both radar and inertial navigation systems, with a focus on ensuring time synchronization. The radar data undergoes range compression, moving target indication (MTI), short-time Fourier transforms (STFT), and wavelet transforms to reduce noise and improve quality and stability. The inertial navigation data is refined through moving average filtering and hysteresis compensation to enhance accuracy and reduce latency. Next, we introduce the Radar-Inertial Navigation Multi-modal Fusion Attention (T-C-RIMFA) model. In this model, a Convolutional Neural Network (CNN) processes the 1D inertial navigation data for feature extraction, while a channel attention mechanism prioritizes features from different convolutional kernels. Simultaneously, a Vision Transformer (ViT) interprets features from radar-derived micro-Doppler images. Experimental results demonstrate significant improvements in HAR tasks, achieving an accuracy of 0.988. This approach effectively leverages the strengths of both sensors, enhancing the accuracy and robustness of HAR systems.

INDEX TERMS Human activity recognition, inertial navigation, micro-Doppler, millimeter wave radar, convolutional neural network, vision transformer.

I. INTRODUCTION

Human Activity Recognition (HAR), the process of recognizing and interpreting human actions, is of paramount importance in numerous practical applications. It can be seamlessly integrated into automated navigation systems [1] to recognize human behaviors, thereby ensuring safe operation. Similarly, HAR enhances surveillance systems [2] by identifying potentially hazardous human activities. Moreover,

HAR underpins a plethora of other domains, such as digital healthcare [3], non-invasive authentication [4], and AR/VR experiences [5], where it plays a crucial role. The comprehension of human activities forms the cornerstone for a multitude of services. For instance, by analyzing an individual's activity log, dietary and fitness recommendations can be tailored based on estimated daily caloric consumption. Furthermore, monitoring falls among elderly individuals enables prompt assistance, effectively mitigating the risk of severe injuries.

In the realm of elderly care and healthcare, HAR offers unique advantages over traditional biometric methods like voice, face, or fingerprint recognition [6]. By leveraging the inherently complex yet stable biological and behavioral patterns of an individual's activities, HAR provides a robust and difficult-to-imitate identification mechanism. Furthermore, its passive nature allows for seamless integration into residents' lives, minimizing disruption and ensuring a natural, unobtrusive monitoring process. This technology can be implemented through a multitude of data acquisition methods [7], including wearable devices [8], ambient sensors, and computer vision-driven cameras. However, each method carries its own set of challenges [9], [10], [11]; wearables might hinder mobility, while vision-based systems may be influenced by lighting conditions or raise privacy concerns. Recognizing these limitations, the integration of multiple sensor modalities and fusion techniques becomes paramount to achieving accurate and reliable HAR in nursing homes. By combining information from diverse sources, these multimodal approaches enhance the system's ability to comprehensively capture the nuances of elderly residents' activities, ultimately improving their quality of life and care.

A. RELATED WORKS

The application of millimeter wave (mmWave) radar in the field of HAR is growing rapidly [12]. Compared with wearable devices and vision-based sensors, mmWave radar provides a non-contact and privacy-protecting human motion sensing technology [13]. mmWave radar maintains excellent performance in bad weather, poor light, and occlusion. In addition, mmWave radar can realize long-range target detection and tracking, which is suitable for large-scale monitoring scenarios, and is not limited by the distance of traditional vision sensors [14]. Therefore, radar-based HAR technology has received much attention as a promising alternative to non-invasive human body recognition.

Doppler effect is an effective form of radar activity representation [15], which is mainly used to measure the overall velocity of the target object, while micro-Doppler effect is a special form of Doppler effect, which is used to describe the frequency change caused by the small movement of the target object [16]. By analyzing the micro-Doppler effect caused by these micro-movements, we can extract the characteristic information related to activity, such as stride frequency, activity period, stride length, etc. Mao et al. [17] proposed a track-tracking and recombination method based on range-Doppler time data for human activity feature estimation. In their simulation, the track tracking accuracy could reach 98.2% to 99.6% . Sharma et al. [18] proposed a walking type recognition system based on mmWave radar, which used flexible analytic wavelet transform and feature sorting technology to effectively classify six walking modes, achieving an overall classification accuracy of 85.5% and 100% in specific cases, with better performance than other comparison methods. Alanazi et al. [19] proposed a novel gait analysis method combining micro-Doppler spectrogram and mmWave radar for bone pose estimation, which used a multi-layer convolutional neural network (CNN) to identify five gait patterns and achieved a high accuracy of 95.7% to 98.8% , proving its application potential in acquiring clinically relevant gait information. While radar sensors are good at range resolution, and accurately distinguishing targets at different radial distances, they are often poor at angular resolution, making it difficult to accurately distinguish targets at the same distance but at different angles. In addition, the micro-Doppler effect mainly captures the radial motion of the target relative to the radar, and it is difficult to accurately capture the non-radial motion components, thus limiting the comprehensive understanding of the dynamic characteristics of the target [20]. In addition, when multiple radar systems are used, crosstalk between signals and multipath effects in complex environments may cause interference, leading to misjudgment [21]. This also limits the accuracy and reliability of target recognition [22], [23].

Inertial navigation technology can effectively address the shortcomings of the aforementioned radar data [24], [25]. Inertial navigation is widely used, including in smartphones, watches, bracelets, smart rings, and other wearable sensors. Mobile phones and wearable sensors can non-invasively monitor activity during walking, running, or even falls and are less intrusive compared to other contact sensors. They do not interfere with the measured object, thereby providing a more accurate reflection of the natural state of gait characteristics. Moreover, the inertial navigation system (INS) can collect these data remotely and analyze them, especially during short-time movements, providing rich information on pedestrian movements [26]. Yang et al. [27] proposed a hybrid network model based on a fast regional convolutional neural network (R-CNN) and gated recurrent unit support vector regression (GRU-SVR), which can build a virtual inertial measurement unit (IMU) when the physical IMU is out of range and maintain the navigation performance of the wearable human-computer interaction system. Experimental results show that this method can maintain the same positioning performance as the trouble-free system under complex gaits. Ding et al. [28] developed a three-node inertial measurement unit system to improve pedestrian positioning accuracy in indoor and outdoor environments without GPS signals. The experimental results show that the system has excellent performance in HAR, step size estimation, and positioning accuracy. Wang et al. [29] developed an indoor multi-motion pattern recognition method based on a multi-node inertial sensor network and long short-term memory artificial neural network and

designed a zero-speed update inertial navigation algorithm suitable for fast motion patterns. Experimental results show that the overall recognition rate of the proposed method is 96.77% , and the error is 1.26% of the total distance. However, long-term use of inertial navigation technology can face challenges from cumulative errors, mainly due to the gradual accumulation of errors in inertial sensors (such as gyroscopes and accelerometers) over time, leading to reduced navigation accuracy. However, by combining it with radar technology, this defect can be effectively remedied.

According to the above description of the characteristics of radar and inertial navigation, we can see that these two sensors have advantages and limitations for HAR. The radar can correct the accumulated errors of the INS, thereby enhancing the system's accuracy and stability in performing HAR. At the same time, inertial navigation can supplement the shortcomings of radar addressing radar's limitations in angular resolution and its inability to accurately capture short-term non-radial movements, providing more comprehensive motion information. Therefore, combining the advantages of radar and inertial navigation can complement each other's limitations and improve the performance and reliability of the HAR system. Moreover, the configuration of fixed millimeter-wave radar and inertial navigation system (INS) attached to moving objects has shown high feasibility and practicality in some specific application scenarios. For example, in the fields of sports training, health monitoring, and smart home monitoring, this configuration can effectively combine the advantages of both, provide accurate motion tracking and analysis, and meet actual needs. Patil et al. [30] proposed an attitude tracking system that integrates 3D light detection, ranging (LiDAR), and IMU sensors, and the results show that the accuracy of both height and position estimation is within the acceptable range of ±3–5 cm. Yu et al. [31] fused data from multiple sensing systems such as IMU sensors, software-defined radio, and radar, and applied neuromorphic computing to perceive and classify human activities, obtaining a confusion matrix classification accuracy of 98.98% . Li et al. [32] used the SFS method to fuse IMU and radar information to create time series data and used SVM and ANN algorithms for classification. The results showed that the method improved the accuracy by about 6% compared with a single data type. It can be seen that the fusion of radar and inertial navigation sensors can improve the effectiveness, reliability, and robustness of the system, improve data reliability, enhance accuracy, expand temporal and spatial coverage, and strengthen the system's real-time performance and information utilization [33].

In the data fusion process, fusion usually occurs at three levels: data, feature, and decision-making. The different modes and dimensions of data often make direct fusion at the data level challenging. At the decision-making level, methods like probability theory synthesize the results of each feature, but these methods have drawbacks such as dependence on model assumptions and sensitivity to noise. Feature-level fusion commonly uses traditional methods like principal component analysis (PCA) and linear discriminant analysis (LDA) or recently emerged neural networks. Compared to traditional methods, neural networks offer significant advantages, as they can automatically learn and extract complex nonlinear features. In the context of HAR, time aggregation is typically achieved through the addition of a temporal network. The model based on a 3-dimensional Convolutional Neural Network (3D CNN) [34] directly extracts spatiotemporal features from the input, which has the disadvantages of requiring a large amount of data, difficulty in network convergence, and high computational cost. The model based on Long Short-Term Memory (LSTM) [35] extracts the pose features from the video input and then uses the three-layer LSTM network to aggregate the time features to generate the final gait features, however, it faces challenges in handling feature changes caused by varying walking speeds. To solve these problems, the Transformer model is proposed for temporal feature aggregation. Li et al. [36] proposed a Transformer ensemble converter model with a time aggregation operation for obtaining ensemble-level spatiotemporal features. Delgado-Santos et al. [37] explored the potential of the Transformer in behavioral biometrics by updating the configuration of the Transformer, thereby improving the biometric authentication performance. Ma et al. [38] apply a Transformer layer to conduct space-time modeling for 4D radar point cloud video, capture walking motion information, and then simulate human walking mode. Although the Transformer retains the features of the time dimension and adopts a multi-head attention mechanism to cope with activity changes caused by speed or pace, compared to CNN models, Transformer models tend to ignore local features and lack features such as shift, scale changes, and hierarchical structures. So, CNN models are still the mainstream method for HAR.

In this regard, some scholars propose to combine the two basic structures of CNN and Transformer, thereby integrating the advantages of both. Conformer [39] fuses local features and global features at different resolutions in an interactive manner. Pyramid Vision Transformer (PVT) [40] uses a progressive pyramid structure to achieve high-resolution output while decreasing the computational load of feature maps. Vision Transformer (ViT) [41] achieves high-resolution output by integrating local and global features at different resolutions in a self-attentional manner. Following this trend of integration, researchers have gradually realized the advantages of integrating multiple models, especially in situations where both spatial and temporal information needs to be considered in HAR. By combining the CNN and Transformer models in the field of HAR, we can better capture the subtle features in walking patterns. This approach can better adapt to different activity changes and improve robustness and accuracy.

B. MOTIVATION AND CONTRIBUTION

Given that traditional HAR methods usually rely on time-consuming and complex data collection processes, are difficult to respond to in real-time, and existing technologies often require interference with the person being detected, or cannot ensure efficiency and reliability in complex environments, this

study aims to achieve efficient human action recognition and ensure the reliability of data collection in actual environments through non-destructive testing technology, fast real-time testing, and relatively convenient data collection processes. At present, no one has performed feature fusion of radar and inertial navigation data to identify gait in the field of HAR. This paper takes advantage of the deep learning model in processing high-dimensional data and effectively integrates two different but complementary data sources, inertial navigation, and radar, to achieve a more comprehensive and accurate understanding of pedestrian behavior. By combining the hybrid structure of CNN and Transformer, we can make full use of its advantages in the global spatiotemporal relationship, and reference channel attention's ability in local feature extraction, which is expected to better process information of different resolutions in HAR tasks and improve the robustness and performance of the system. The main contributions of this study are as follows:

1) A feature fusion model is proposed to deeply fuse bimodal heterogeneous data at the feature level, making full use of the advantages of CNN in extracting features from 1D time series INS data and the superiority of ViT in processing 2D spatial mmWave radar data, thereby improving the accuracy and robustness of HAR and classification.
2) A deep learning method is employed utilizing the T-C-RIMFA network, this method jointly trains and fuses radar and inertial navigation features, mutually weighting them along the channel dimension to form comprehensive feature representations. This approach significantly enhances the model's adaptability to different modal data and improves the performance and stability of the HAR system, especially when processing complex and noisy data scenarios.
3) A channel attention mechanism is introduced to dynamically adjust the weights of different channel features, highlight important features, and suppress redundant information, thereby enhancing the accuracy and robustness of HAR, making the model show stronger adaptability and stability when dealing with different scenarios and noise interference.

II. SYSTEMS AND METHODS

The sample set for this experiment comprises 6 distinct categories of human gait during activity: walking, slow walking (for a slower pace), running, ascending and descending stairs, as well as falling. Each of these categories collected 200 sets of data, and through data augmentation by cropping, the amount of data for each set was expanded fivefold, resulting in a total of 6000 sets of data across all six categories. Fig. 1 shows the system flow chart in this paper, which is mainly divided into (a) data acquisition, (b) data preprocessing, and (c) the construction of the improved network model. The system flow chart covers the whole process from data acquisition to activity classification and recognition. First, the

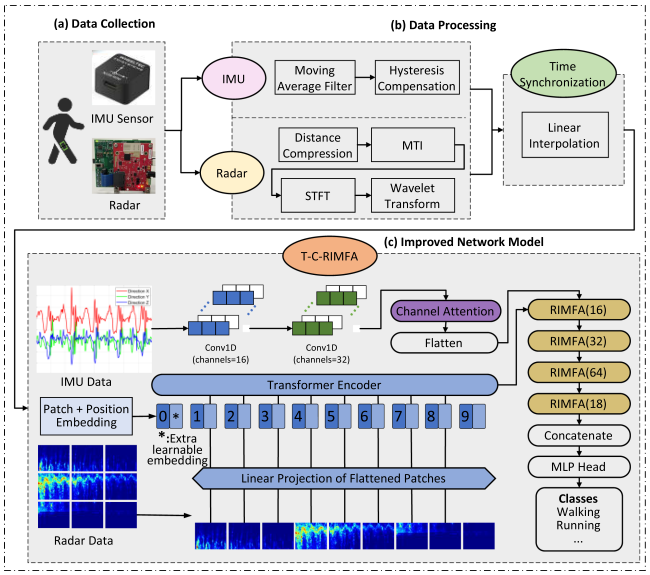


FIGURE 1. System flow chart.

data collection stage is responsible for collecting inertial navigation sensor data and radar data. Then, linear interpolation is employed to effectively align the timestamps of the radar and IMU. Next, the feature fusion phase of inertial navigation and radar data is the core part of the system. At this stage, CNN is used to extract features from inertial navigation sensor data through a 1D convolution layer, and the channel attention mechanism is used to weight features extracted from inertial derivative data from different convolution kernels, while radar data is extracted by ViT to convert radar data into sequences. The Transformer model is also applied to the sequence to capture local patterns and global associations in space. Then, the features of the two data are fused by the proposed T-C-RIMFA, the importance of each feature is dynamically adjusted, and the fused features are finally input into the classifier for activity classification and recognition.

A. DATA COLLECTION

In order to verify the adaptability of the model in different environments, we added two typical indoor and outdoor scenes to the experiment, aiming to reflect the diverse actual application environments. Fig. 2 shows one of the outdoor pedestrian activity data collection systems, where Fig. 2(a) represents the target pedestrian to be measured. Fig. 2(b) shows the inertial navigation system, which uses inertial navigation sensors such as accelerometer, gyroscope, and magnetometer on mobile devices to obtain sensor data such as acceleration, angular speed, and direction of pedestrians. During the data collection process, the IMU sensor was placed on the subject's leg. The leg is the most active part of the human body during gait, providing distinct motion change signals. By placing the IMU sensor on the leg, we can accurately capture dynamic gait-related features, such as stride length and step frequency, which are crucial for behavior recognition tasks. Additionally,

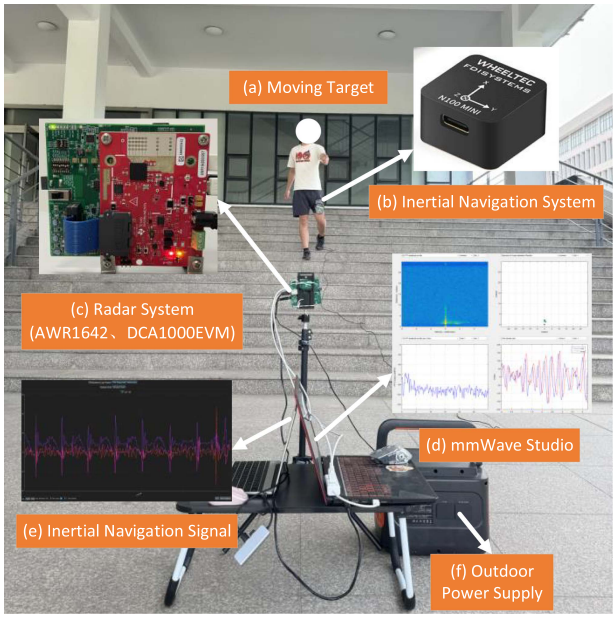


FIGURE 2. Experimental environment diagram.

the movement of the leg is relatively independent, reducing interference from the complex movements of other body parts on the sensor data. IMU sensors are commonly used in wearable devices, and simulating the sensor placement on the leg better aligns with real-world application scenarios, especially in monitoring gait and physical activity, where it has significant advantages. Fig. 2(c) demonstrates the radar system, which mainly transmits information such as pedestrian speed collected by AWR1642 and DCA1000EVM to the computer. In this experiment, pedestrians were within a 1-30 meter range from the radar. Fig. 2(d) represents the data acquisition and processing platform of millimeter-wave radar, which is used to capture and process the raw data acquired from the sensor. Similarly, Fig. 2(e) represents the data acquisition and processing platform for inertial navigation, which is responsible for capturing the raw data obtained from the sensor. Fig. 2(f) represents an outdoor portable power supply that can continuously and stably provide power to the system.

The FMCW mmWave radar, the core component of radar data measurement used in this study, uses the Texas Instruments AWR1642-BOOST radar development board (Red PCB in Fig. 2(c)). It is a high-performance, highly integrated radar development board specifically designed for prototyping and testing a wide range of radar applications, and it comes with a wealth of software tools and development support, including TI's mmWave SDK.

To ensure accurate detection of direction, speed, and angle, we utilized a 77 GHz FMCW radar with a bandwidth of 1798.92 MHz and an RF gain setting of 30 dB. The radar system operates with a sampling rate of 2000, a sample time of 100 μs, and a frame period of 40 ms, ensuring precision and stability in data acquisition. The radar configuration employs a MIMO setup, consisting of 2 transmit antennas and 2 receive antennas, enabling the simultaneous collection of signals from multiple directions and enhancing target detection accuracy. Each data acquisition cycle involves 128 chirps, providing sufficient data for high-precision HAR.

Since the sample size of radar data used in this study is 150 frames/sample, the sample acquisition time for each set of data is 6 seconds. Therefore, it is necessary to use the Texas Instruments DCA1000EVM data acquisition board (Green PCB in Fig. 2(c)) when measuring data. By connecting the DCA1000EVM, the experiment can receive multiple frames of radar data and transmit them to a computer or storage device for subsequent analysis and processing. The acquisition board supports laboratory scene acquisition or mobile scene acquisition and also provides real-time data acquisition and fluidization functions, including preprocessing, filtering, target detection, and tracking. This helps reduce latency in data processing, allowing us to get information about radar performance and target recognition more quickly. In addition, the real-time data stream can also be used for real-time visualization, so that researchers can monitor the performance of radar sensors at any time. This is especially important for studies that require immediate feedback and experimental adjustments.

B. DATA PREPROCESSING

1) RADAR DATA PROCESSING

The acquired radar data is then preprocessed, to extract target-related information from the raw data and eliminate possible interference.

First of all, range compression is one of the key steps in data preprocessing. Through range compression, we can convert the target echo signals at different distances into signals with relatively uniform amplitudes. This processing step helps to improve the dynamic range of the signal so that both long-range and short-range targets can be effectively detected. In this process, the signal usually refers to the IQ signal, that is, the signal composed of in-phase (I) and orthogonal (Q) components. By compressing these IQ signals, we can improve the detection performance of the target, especially in complex environments, and improve the recognition accuracy of various targets. The formula for range compression [42] is as follows:

$$S_{comp}\ (r) = S\,(r) \times e^{-j2\pi f_c \frac{2r}{c}} \tag{1}$$

where $S_{comp}\ (r)$ is the compressed signal; $S(r)$ is the original signal; f_c is the operating frequency of radar; r is the target distance; c is the speed of light; j stands for imaginary units.

Then, using moving target indication (MTI) technology, which is primarily accomplished in the frequency domain, we can distinguish objects from complex backgrounds. MTI compares the signals received at two consecutive time points, but its core mechanism operates in the frequency domain, utilizing filtering to eliminate the echoes of stationary targets and clutter. In this way, MTI highlights the signals of moving targets, enabling their effective detection and tracking amidst

the background noise. The formula for MTI is as follows:

$$H(f) = 1 - e^{-j2\pi f T_r} \tag{2}$$

where $H(f)$ is the filter frequency response.; T_r is the pulse repetition period, which is set to 100 μs, based on experimental settings and empirical values.

Subsequently, we use the short-time Fourier transform (STFT) to decompose the signal into several short-time segments and perform spectrum analysis for each short-time segment, which helps us more accurately understand how the spectrum of the signal changes over time, and thus better distinguish between the target signal and the background noise.

$$X(m, \omega) = \sum_{n=0}^{N-1} x(n)\, w(n-m)\, e^{-j\omega n} \tag{3}$$

where n is a variable in the time domain; $X(m, \omega)$ is the spectrum at time period m and frequency ω; $x(n)$ is the original signal; $w(n-m)$ is a window function; N is the number of sampling points for each time period; Omega is the angular frequency.

At last, the preprocessed signal is subjected to wavelet decomposition via Discrete Wavelet Transform (DWT), resulting in wavelet coefficients at various layers, including low-frequency coefficients and high-frequency coefficients. In this study, we utilize Daubechies wavelets, specifically the Daubechies-4 (Db4) wavelet, due to their excellent performance in signal processing, especially in time-frequency localization and noise suppression. The low-frequency coefficients typically contain the primary information of the signal, while the high-frequency coefficients encompass noise and detailed information. The Daubechies wavelet is particularly effective in radar signal analysis as it allows for multi-scale analysis while preserving the accuracy of feature extraction in compressed data. The formula can be expressed as:

$$W_\psi(j, k) = \sum_n x(n)\, \psi_{j,k}(n) \tag{4}$$

where $x(n)$ represents the original signal, $\psi_{j,k}(n)$ denotes the representation of the wavelet function at scale j and shift k, and $W_\psi(j, k)$ stands for the wavelet coefficients.

Then, thresholding is applied to the high-frequency coefficients at each layer to remove noise. Thresholding can be achieved through either hard thresholding (setting coefficients below the threshold to zero) or soft thresholding (shrinking coefficients below the threshold towards zero). In this experiment, the hard thresholding method is used, and its specific formula is as follows:

$$\widehat{W_\psi}(j, k) = \begin{cases} W_\psi(j, k) & \text{if } \left|W_\psi(j, k)\right| \geq \lambda \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where λ represents the threshold value, which is set to 3.89 mm, based on the specific environmental conditions and the radar system's operational frequency. This value corresponds to the wavelength of the radar signal used in the experiment, and it plays a key role in determining the signal's behavior and its ability to interact with objects of various sizes and materials.

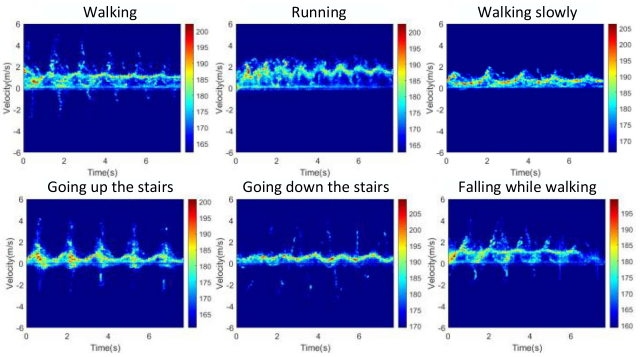


FIGURE 3. Micro-Doppler image of millimeter wave radar.

Using the processed low-frequency coefficients and the high-frequency coefficients after thresholding, wavelet reconstruction is performed to obtain the denoised signal. The specific formula is as follows:

$$x(n) = \sum_{j,k} W_\psi(j, k)\, \tilde{\psi}_{j,k}(n) \tag{6}$$

where $\tilde{\psi}_{j,k}(n)$ denotes the reconstructed wavelet function.

Finally, the above processed mmWave radar data is constructed into a micro-Doppler image with time as the horizontal coordinate and speed as the vertical coordinate as shown in Fig. 3.

2) INERTIAL NAVIGATION DATA PROCESSING

The duration of activity often exceeds the sensor sampling rate and requires a segmentation method to process the signal rather than relying solely on the sample. Splitting data allows a single data point to be associated with a given task [43]. Segmentation is the key step to realizing HAR, which mainly includes manual and automatic segmentation methods. Manual segmentation is done by observers, and automatic segmentation is done by computer algorithms. The best method depends on the specific application, manual segmentation is more accurate, while automatic segmentation is faster and more efficient. In HAR or biological signal data sets, manual segmentation is usually more accurate than automatic segmentation [44], [45], so the inertial data in this paper adopts manual segmentation.

When processing inertial data, the moving average filter is often used to smooth the signal and reduce the high-frequency noise, the specific formula is as follows:

$$y[n] = \frac{1}{N} \sum_{k=0}^{N-1} x[n-k] \tag{7}$$

where k is the variable used to index the data points in the filter window; $x[n]$ is the input signal; $y[n]$ is the output signal; N is the window length of the filter.

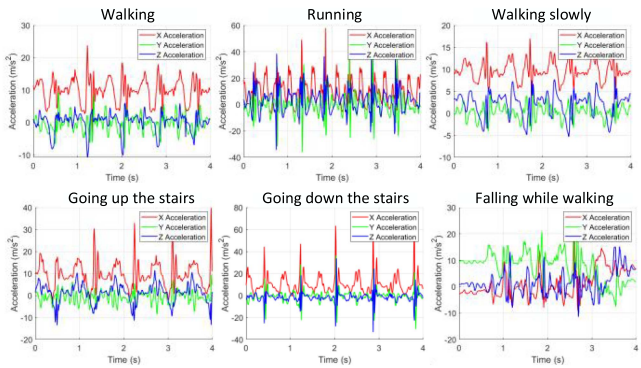


FIGURE 4. Inertial navigation acceleration curve image.

However, the moving average filter may introduce a time delay, and in order to reduce the time delay and better preserve the dynamic characteristics of the signal, the delay compensation method can be combined. Delay compensation reduces delay by adjusting the weight between the input signal at the current time and the output of the filter at the previous time. Assuming that the current time is t, the delay compensation can be achieved by the following formula:

$$y[t] = \alpha \cdot x[t] + (1-\alpha) \cdot y[t-1] \tag{8}$$

where $x[t]$ is the input signal of the current moment; $y[t-1]$ is the filter output at the previous time; $y[t]$ is the filter output at the current time; α is the compensation coefficient, which controls the weight between the input signal at the current time and the output of the filter at the previous time (the compensation coefficient α is usually between 0 and 1, indicating the weight of the input signal at the current time, the closer to 1, the greater the influence of the input signal at the current time, the smaller the lag effect; The closer to 0, the greater the influence of historical data, the more obvious the lag effect). In this experiment, α is set to 0.8 based on the trade-off between smoothing the input signal and minimizing the delay effect. A value of 0.8 provides a good balance, where the influence of the current input signal is significant enough to reduce lag, but still allows enough weight to be given to historical data to maintain stability in the signal processing. This choice of α is based on empirical observations from previous experiments, where it was found to provide optimal performance in terms of both responsiveness and accuracy.

Finally, the above-processed data is constructed into the inertial navigation acceleration curve image with time as the horizontal coordinate and acceleration as the vertical coordinate as shown in Fig. 4.

C. TIME SYNCHRONIZATION

In time series data processing, particularly in applications involving multi-sensor fusion, time synchronization is a crucial step to ensure data consistency and accuracy. For radar and inertial navigation systems, as they may have different sampling rates, timestamp accuracies, and system delays, effective time synchronization becomes imperative.

In order to perform time synchronization, it is first necessary to establish a unified time reference. In this study, we have chosen Global Positioning System (GPS) time as the reference due to its high precision and widespread accessibility. The timestamps of the radar and Inertial Measurement Unit (IMU) systems are converted to the GPS time frame, and the specific conversion formula is as follows:

$$t^{ref}_{radar,i} = t_{radar,i} + \Delta t_{radar} \tag{9}$$

$$t^{ref}_{ins,j} = t_{ins,j} + \Delta t_{ins} \tag{10}$$

where $t_{radar,i}$ and $t_{radar,j}$ are the original timestamps of the radar and IMU respectively, Δt_{radar} and Δt_{ins} are the offsets from their respective times to GPS time, and $t^{ref}_{radar,i}$ and $t^{ref}_{ins,j}$ are the converted timestamps.

Given the inconsistent sampling rates between the radar and IMU in this study, it is necessary to perform sampling rate matching to ensure a one-to-one correspondence between data points. This is typically achieved by adopting linear interpolation, which interpolates the IMU data to match the sampling rate of the radar data. For each radar timestamp $t^{ref}_{radar,i}$, the nearest two IMU timestamps $t^{ref}_{ins,j}$ and $t^{ref}_{ins,j+1}$, along with their corresponding data values $d_{\text{ins},j}$and $d_{\text{ins},j+1}$, are identified. Then, the following formula is used to perform the interpolation:

$$d^{interp}_{ins,i} = d_{ins,j} + \frac{d_{ins,j+1} - d_{ins,j}}{t^{ref}_{ins,j+1} - t^{ref}_{ins,j}} \cdot \left(t^{ref}_{radar,i} - t^{ref}_{ins,j}\right) \tag{11}$$

After completing the aforementioned steps, the timestamps of the radar and IMU data have been aligned under a unified time reference, and the IMU data has been interpolated to match the radar's sampling rate. At this juncture, the converted and interpolated timestamps can be directly utilized for subsequent data fusion and analysis.

III. BIMODAL FEATURE FUSION

In this experiment, only micro-Doppler features and IMU acceleration data were selected because these two types of features are effective enough in reflecting human motion patterns, especially gait recognition and behavior analysis. Although millimeter-wave radar and inertial navigation systems provide rich feature information, the scale differences of different features may lead to mismatches during data fusion, thus affecting model performance. Experimental results show that directly fusing all features does not significantly improve model performance, but may increase noise and errors. Therefore, we choose to simplify the features to avoid the excessive complexity of the model and ensure efficient performance.

A. TRANSFORMER BASED RADAR FEATURE EXTRACTION

In recent years, ViT has developed rapidly as an advanced image classification technology based on a self-attention mechanism. Using the Transformer architecture, ViT has

demonstrated excellent performance when performing computer vision tasks. From the original ViT-Ti [46], ViT-S [47], ViT-B [48], ViT-L [48], and other variants to today's Swin-T [49], MaxViT-B [50] and the latest ViT-G /14 [51], the ViT family continues to grow. These variants provide flexible options for different sizes and types of tasks, enabling Vits to adapt to a wide range of complex image processing needs. Here are the specific steps in the ViT model.

First, ViT divides the image into fixed-size image blocks (patches) and converts each image block into a vector that is used as a token for the input sequence.

$$\boldsymbol{X}_{\boldsymbol{patch}} = patch_{embedding(\boldsymbol{X})} \tag{12}$$

where $\boldsymbol{X}$ is the input image; $\boldsymbol{X}_{\boldsymbol{patch}}$ is a vector representation of an image block.

ViT then uses the encoder part of the Transformer to process the image sequences. The encoder consists of multiple Transformer encoder layers. The transformer's self-attention mechanism is used to calculate the attention weight of each token in the sequence. The specific formula is as follows:

$$Attention\ (\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\boldsymbol{T}}}{\sqrt{d_k}}\right)V \tag{13}$$

where $\boldsymbol{Q}$ is the query matrix; $\boldsymbol{K}$ is the bond matrix; $\boldsymbol{V}$ is a matrix of values; d_k is the dimension of the attention head.

Typically, there is a fully connected feedforward neural network behind each self-attention layer, with the following formula:

$$FFN\,(x) = ReLU\,(x\boldsymbol{W}_1 + b_1)\,\boldsymbol{W}_2 + b_2 \tag{14}$$

where $\boldsymbol{W}_1$ and b_1 are the weights and biases of the first linear layer; $\boldsymbol{W}_2$ and b_2 are the weights and biases of the second linear layer.

Finally, after the self-attention and feedforward neural network of each Transformer encoder layer, layer normalization is typically applied.

$$LayerNorm\ (x) = \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} \tag{15}$$

where μ is the average; σ is the standard deviation; ϵ is a small number that is used for stable computation.

B. CNN BASED INERTIAL NAVIGATION FEATURE EXTRACTION

Convolutional Neural Networks (CNNs) have extensive use in behavioral classification. CNN can extract high-level features from the original input data. For behavior classification, these features can be motion, shape, color, and other information in images, videos, or time series data. Through operations such as the convolutional layer and pooling layer, CNN can extract more abstract and semantically rich feature representations layer by layer. For time series data, such as inertial navigation sensor data used in this study, CNN can extract features from time dimensions through the 1D convolution layer. This processing allows CNN to directly process data that is continuous in time, and to capture local patterns and global associations in time. In the behavior classification task, the process of using CNN to extract features from time series data can be realized by the following steps.

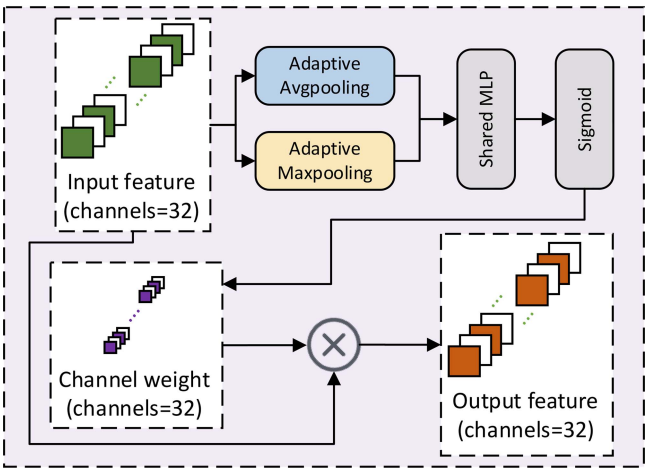


FIGURE 5. Channel attention block.

The first is to extract features in time dimension using a 1D convolution operation. Assuming that the input data is a time series of length L, the size of the convolution kernel is K, and the length of the output feature graph is L', then the convolution operation can be expressed as:

$$\boldsymbol{X}_{\boldsymbol{conv}}\ [i] = f\left(\sum_{j=0}^{K-1} \boldsymbol{W}\left[j\right] \cdot \boldsymbol{X}\left[i+j\right] + b\right) \tag{16}$$

where $\boldsymbol{X}$ is the input time series data; $\boldsymbol{X}_{\mathbf{conv}}$ is the feature graph after convolution; $\boldsymbol{W}$ is the weight of the convolution kernel; b is the offset term; f is the activation function, such as ReLU.

Then there is the use of pooling layers to reduce the size of the feature map while preserving key information. Common pooling operations include maximum pooling and average pooling. Assuming the window size of the pooling operation is P and the length of the output feature graph is L', the pooling operation can be expressed as:

$$\boldsymbol{X}_{\boldsymbol{pool}}\ [i] = \text{Pooling}\,(\boldsymbol{X}_{\boldsymbol{conv}}\ [i : i + P]) \tag{17}$$

where X_{pool} is the feature map after pooling; $Pooling$ is a pooling operation, which can be maximum pooling or average pooling.

C. CHANNEL ATTENTION MECHANISM

Channel attention block is a structure used to enhance a deep learning model's focus on correlations between channels. Its design is inspired by the interaction between channels and the fact that the importance of different channels may vary in different tasks. The main purpose of this structure is to improve the performance and expressiveness of the network by weighting the feature graph of each channel, and its network structure is shown in Fig. 5.

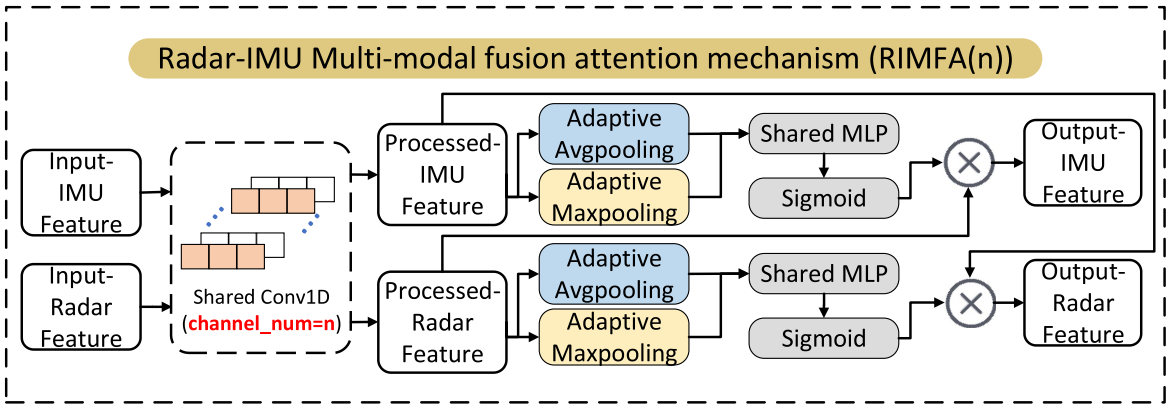


FIGURE 6. Radar-IMU multi-modal fusion attention mechanism.

In the channel attention block, two kinds of pooling operations are first performed on the input feature graph: average pooling and maximum pooling. These two pooling operations capture global information about each channel in the feature graph.

$$AvgPool\left(\boldsymbol{X}\right) = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \boldsymbol{X}_{c,h,w} \tag{18}$$

where H and W are the height and width of the input feature map respectively; $\boldsymbol{X}_{c,h,w}$ represents the pixel values at channel c, height h, and width w in the input feature graph $\boldsymbol{X}$.

$$MaxPool\left(\boldsymbol{X}\right) = \max_{h=1}^{H} \left(\max_{w=1}^{W} (\boldsymbol{X}_{c,h,w}) \right) \tag{19}$$

where max indicates the operation of taking the maximum value. $\boldsymbol{X}_{c,h,w}$ represents the pixel values at channel c, height h, and width w in the input feature graph $\boldsymbol{X}$.

Next, we concatenate the results of average pooling and maximum pooling and calculate the attention weight for each channel through a fully connected layer and activation function. These weights will guide the model to allocate more or less attention to a particular channel. Specifically, we can express the attention weight calculation as:

$$\boldsymbol{W} = \sigma\left(FC\left(\left[AvgPool\left(\boldsymbol{X}\right), MaxPool\left(\boldsymbol{X}\right)\right]\right)\right) \tag{20}$$

where, σ is the activation function, and in our experiment, we used the Sigmoid function. And FC is the full connection layer.

Finally, the channel attention weight $\boldsymbol{W}$ is element-wise multiplied by the original input feature map $\boldsymbol{X}$ to obtain a weighted feature map. This operation can enable the model to better focus on the features that are important for the current task.

$$\boldsymbol{Y} = \boldsymbol{W} \odot \boldsymbol{X} \tag{21}$$

Through these steps, the channel attention block can dynamically adjust the importance of each channel, thereby increasing the model's focus on the correlation between channels and enhancing the model's performance and expressiveness on specific tasks.

D. RADAR-IMU MULTI-MODAL FUSION ATTENTION MECHANISM

To fuse the radar and inertial navigation features, a Radar-inertial navigation Multi-modal Fusion Attention mechanism (RIMFA) is introduced, where radar and inertial navigation features are mutually weighted along the channel dimension. The RIMFA module is shown in Fig. 6. First, Firstly, the radar and inertial navigation features are flattened to obtain one-dimensional features:

$$\begin{aligned} &\boldsymbol{Feature}_{\boldsymbol{Radar}}^{\boldsymbol{Flatten}},\ \boldsymbol{Feature}_{\boldsymbol{IMU}}^{\boldsymbol{Flatten}} \\ &= Flatten\left(\boldsymbol{Feature}_{\boldsymbol{Radar}},\ \boldsymbol{Feature}_{\boldsymbol{IMU}}\right) \end{aligned} \tag{22}$$

After flatten, the Radar and inertial navigation features are input into the shared Conv1D module to obtain the process-radar Feature $\boldsymbol{Feature}'_{\boldsymbol{Radar}}$ and the process-IMU Feature $\boldsymbol{Feature}'_{\boldsymbol{IMU}}$.

$$\begin{aligned} &\boldsymbol{Feature}'_{\boldsymbol{Radar}},\ \boldsymbol{Feature}'_{\boldsymbol{IMU}} \\ &= Conv1d_n\left(\boldsymbol{Feature}_{\boldsymbol{Radar}}^{\boldsymbol{Flatten}},\ \boldsymbol{Feature}_{\boldsymbol{IMU}}^{\boldsymbol{Flatten}}\right) \end{aligned} \tag{23}$$

where n is the number of channels. Next, the $\boldsymbol{Feature}'_{\boldsymbol{Radar}}$ and $\boldsymbol{Feature}'_{\boldsymbol{IMU}}$ are fed into a Channel attention block to generate the Radar weight $\boldsymbol{W}_{\boldsymbol{Radar}}$ and inertial navigation weight $\boldsymbol{W}_{\boldsymbol{IMU}}$ for each channel.

$$\begin{aligned} &\boldsymbol{W}_{\boldsymbol{Radar}},\ \boldsymbol{W}_{\boldsymbol{IMU}} \\ &= Channel\ Attention\left(\boldsymbol{Feature}'_{\boldsymbol{Radar}},\ \boldsymbol{Feature}'_{\boldsymbol{IMU}}\right) \end{aligned} \tag{24}$$

Finally, the $\boldsymbol{Feature}'_{\boldsymbol{Radar}}$ is multiplied by the $\boldsymbol{W}_{\boldsymbol{IMU}}$, and the $\boldsymbol{Feature}'_{\boldsymbol{IMU}}$ is multiplied by the $\boldsymbol{W}_{\boldsymbol{Radar}}$ to obtain the Output-Radar Feature and Output-IMU Feature.

$$\boldsymbol{Feature}_{\boldsymbol{Radar}}^{\boldsymbol{output}} = \boldsymbol{Feature}'_{\boldsymbol{Radar}} \odot \boldsymbol{W}_{\boldsymbol{IMU}} \tag{25}$$

$$\boldsymbol{Feature}_{\boldsymbol{IMU}}^{\boldsymbol{output}} = \boldsymbol{Feature}'_{\boldsymbol{IMU}} \odot \boldsymbol{W}_{\boldsymbol{Radar}} \tag{26}$$

E. IMPROVED T-C-RIMFA NETWORK MODEL

To fully combine the features extracted from the radar sensor and the inertial navigation sensor, we design a method to fuse the two data features. First, feature extraction of inertial navigation sensor data is carried out through a 1D convolution layer, which enables the model to directly process time-continuous data and capture local patterns and global associations in time:

$$\boldsymbol{feature_{IMU}} = ConvBlock\left(\boldsymbol{Data_{IMU}}\right) \tag{27}$$

At this stage, we use a 1D convolutional network to process inertial navigation sensor data. The convolutional network can effectively extract local features from time series data, and gradually extract higher-level feature representations by stacking multiple convolutional layers and pooling layers.

Next, the channel attention mechanism is used to weigh the feature channels extracted from different convolution kernels, so that the network can focus on relatively more important features:

$$\boldsymbol{feature_{CAIMU}} = ChannelAttention\left(\boldsymbol{feature_{IMU}}\right) \tag{28}$$

At the same time, we use ViT for feature extraction of radar data. First, the radar data is converted into a sequence, and then the Transformer model is applied to the sequence to capture spatial local patterns and global associations of the radar data:

$$\boldsymbol{feature_{Radar}} = ViTBlock\left(\boldsymbol{Data_{Radar}}\right) \tag{29}$$

The ViT module first divides the input radar data into several patches of fixed size and projects each patch into a high-dimensional feature space to form a sequence input. These sequences are then processed by multi-layer Transformer encoders to extract global and local features from the radar data.

Then, the extracted inertial navigation feature and radar feature are flattened and fed into the RIMFA module.:

$$\begin{aligned}&\boldsymbol{feature_{CAIMU-RIMFA}}, \boldsymbol{feature_{Radar-RIMFA}}\\ &= RIMFA\left(\boldsymbol{feature_{CAIMU}}, \boldsymbol{feature_{Radar}}\right)\end{aligned} \tag{30}$$

After repeated experimental validation, it has been confirmed that in the RIMFA module, as the network depth increases, the gradual increase in the number of channels means that each channel carries more information, requiring a more refined channel attention mechanism. Therefore, for the RIMFA module, the channel numbers are set to 16, 32, 64, and 128. After each RIMFA module layer, a 4×4 MaxPooling operation is performed, allowing each layer to extract richer and more complex features, thereby enhancing the model's feature representation ability. The use of 16, 32, 64, and 128 channels allows the attention mechanism to weigh and fuse radar and inertial navigation features across different layers. This ensures that the network can fully utilize channel attention at various scales to achieve effective feature fusion.

Then, concatenate the features after passing through the RIMFA module::

$$\begin{aligned}&\boldsymbol{feature_{IMU+Radar}} = Concatenate\\ &\left(\boldsymbol{feature_{CAIMU-RIMFA}}, \boldsymbol{feature_{Radar-RIMFA}}\right)\end{aligned} \tag{31}$$

In this phase, we flatten the features extracted from the inertial navigation sensor. Then, it is concatenated with the radar features to form a comprehensive feature representation. This comprehensive feature contains complementary information from the two sensors, providing richer input for classification tasks.

Finally, the fused features are input into the multi-layer perceptron (MLP) classifier for activity classification and recognition:

$$\boldsymbol{Output} = MLP\left(\boldsymbol{feature_{IMU+Radar}}\right) \tag{32}$$

To prevent overfitting, we have added Dropout processing to the MLP. During training, Dropout can randomly drop a subset of neurons, thereby reducing the model's dependence on some specific neurons and improving the model's ability to generalize.

The designed network structure is shown in Fig. 7. By synthesizing multi-modal data and simultaneously processing the data of the radar sensor and inertial navigation sensor, the characteristic information of the two sensors is fully utilized. Radar data provides spatial local patterns and global association information, while inertial data provides time-continuous local patterns. The combination of the two helps to improve the accuracy and robustness of HAR.

Compared to traditional CNN or Transformer networks that only process IMU or radar, the fusion design enables more comprehensive use of information from multi-modal data. This comprehensive utilization can improve the generalization ability and robustness of the model so that the model has better adaptability and accuracy to the activity data under different environments and conditions.

The fusion design improves the utilization rate of data, fully taps the correlation between different sensor data, and makes the model understand the activity data more comprehensively.

At the same time, the fusion design also enhances the task adaptability of the model, which can better adapt to different HAR tasks and data situations, and has stronger versatility and adaptability.

In addition, the introduction of the channel attention mechanism enables the network model to adjust the weights of feature channels extracted from different convolutional kernels adaptively, so that the network can pay more attention to relatively important features, suppress irrelevant or unimportant features, and improve feature expression and classification performance.

Algorithm 1 is used for HAR based on radar and inertial navigation multi-sensor fusion, which establishes the feasibility of the whole framework theoretically. First, separate feature extraction for each data type is carried out, and then feature level fusion is carried out to integrate features from

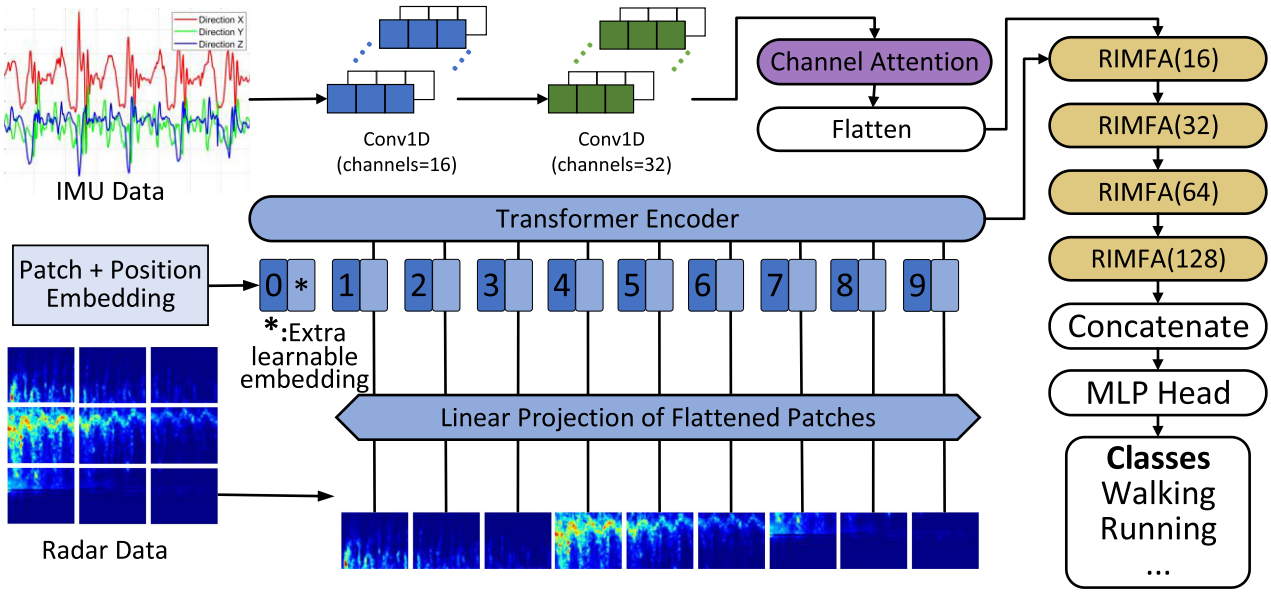


FIGURE 7. T-C-RIMFA network structure diagram.

different signal sources, and the channel attention mechanism is combined to dynamically adjust the importance of each feature. This step ensures that the features extracted by different sensors can effectively complement and enhance each other, and improves the ability of the model to characterize activity. Finally, the fused features are input into the T-C-RIMFA network for activity classification and recognition.

IV. RESULTS AND DISCUSSION

A. METHOD TRAINING AND SETUP

First, in order to realistically simulate the situation in the real environment, we introduce a certain level of ambient noise. Then, before the network model training, the processed data set is divided into a training set, verification set, and test set according to the ratio of 5:3:2. In the stage of model training, parameter optimization, and model training are carried out using the training set. Then, the generalization ability of the model is verified by a validation set and hyperparameters are adjusted to ensure the robustness of the model on different data. Finally, a test set is used to conduct a final evaluation of the model to verify its reliability and generalization performance on real data. This division and evaluation help ensure that our findings are credible and generalizable.

In the model training process, we used the Adam optimizer to effectively train and adjust the model, and the selected learning rate was 1e-5. In order to maintain training speed and control memory consumption, set the batch size to 16. In order to ensure that the model can fully converge and have the ability to generalize, we conducted 300 cycles of training, and this number was obtained through repeated experiments and verification. For the choice of loss function, we use the cross entropy loss function:

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} log\, p_{i,c} \tag{33}$$

where, N is the number of samples; C is the number of categories; $y_{i,c}$ is the value of class C in the actual label of sample i ($y_{i,c} = 1$ if the sample belongs to class C, otherwise $y_{i,c} = 0$); $p_{i,c}$ is the probability that the sample i predicted by the model belongs to class c.

B. COMPARISON OF COMPUTATIONAL RESOURCES

In order to compare the amounts of resources required by the constructed model and the traditional model in operation, the amount of resources required by each model was calculated, and the results are shown in Table 1.

Within this context, total params represent the number of all learnable parameters in the model. Pass size indicates the memory usage during forward propagation. Param size Indicates the memory usage of model parameters. Total size indicates the total memory usage of the model, including model parameters and memory usage during forward propagation.

It can be seen from Table 1 that compared with other models, T-C-RIMFA has fewer Total parameters and a smaller Pass size in the process of forward propagation. This gives T-C-RIMFA an advantage in terms of model size and computing resource requirements and can be run and deployed more efficiently to maintain a certain performance.

C. ACCURACY COMPARISON

As a confusion matrix to evaluate the performance of the algorithm, it can intuitively count the inference error and inference correct value of the classification model. As shown

Algorithm 1: T-C-RIMFA for HAR.

Input: X_R Radar data, X_I IMU data
Output: HAR
/∗**Step1**: Block the radar image ∗/
$in_{channels} = 1,\ patch_{size} = 16, emb_{size} = 256$
$projection = Conv2d(in_{channels},\ emb_{size}, kernel_{size} = patch_{size},\ stride = patch_{size})$
$X_{R\_pro} \leftarrow projection(X_R)$;
$X_{R\_per} \leftarrow permute(X_{R\_pro},\ (0,\ 2,\ 3,\ 1))$
$X_{R\_patch} \leftarrow view(X_{R_{per}},\ (X_{R_{per}}.size(0),\ -1,\ X_{R\_per}.size(-1)))$
∗**Step2**: Encode using Transformer Encoder ∗/
$num_{layers} = 4, emb_{size} = 256, num_{heads} = 8, hidden_{size} = 512, dropout = 0.1$;
$layers = []$;
/∗Create each layer of Transformer Encoder ∗/
for i from 1 to num_{layers}:
  $layer_i = Transformer\ Encoder\ Layer(d_{model} = emb_{size},\ nhead = num_{heads},\ dim_{feedforward} = hidden_{size},\ dropout = dropout)$
  $layers.append(layer_i)$
for $layer$ in $layers$: $X_{R\_out} \leftarrow$ layer($X_{R_{patch}}$)
∗**Step3**: IMU data feature extraction ∗/
$X_{I\_dropout1} \leftarrow Dropout(MaxPool1d(ReLU(Conv1d(X_I))))$
$X_{I\_dropout2} \leftarrow Dropout(MaxPool1d(ReLU(Conv1d(X_{I\_dropout1}))))$
$X_{I\_CA} \leftarrow ChannelAttention(X_{I\_dropout2})$
$X_{I\_out} \leftarrow Flatten(X_{I\_dropout2})$
∗**Step4**: Radar and IMU feature fusion and classification ∗/
$X_{R\_out\_RIMFA},\ X_{I\_out\_RIMFA} \leftarrow RIMFA(X_{R\_out},\ X_{I\_out})$
$features_{combined} \leftarrow Concatenate(X_{R\_out\_RIMFA},\ X_{I\_out\_RIMFA})$
$features_{classification} \leftarrow Linear(features_{combined})$
$Output \leftarrow SoftMax(features_{classification})$

TABLE 1. Amount of Resources Required by Each Model

	Model	Total params	Pass size (MB)	Param size (MB)	Total size (MB)
Radar-single	LeNet[52]	1.98e6	1.31	7.92	9.3
	GoogleNet[53]	5.97e6	15.12	22.77	37.95
	AlexNet[54]	5.70e7	2.76	217.47	220.3
	VGG16[55]	6.51e7	71.5	248.24	319.8
	ViT[56]	2.18e6	2.75	4.49	7.31
IMU-single	CNN_CA[57]	5.59e4	0.02	0.22	0.24
	CNN[58]	5.57e4	0.02	0.22	0.24
	LSTM[59]	2.12e7	0.02	84.99	85.01
	GRU[60]	9.64e6	0.02	38.56	38.58
Radar+IMU	ViT+GRU	1.49e7	4.76	54.67	59.49
	ViT+LSTM	2.50e7	2.78	95.79	98.63
	ViT+CNN	2.23e6	2.77	4.71	7.55
	ViT+CNN_CA	2.23e6	2.77	4.71	7.55
	T-C-RIMFA	3.43e6	3.03	9.51	12.61

in Fig. 8, where (a) to (e) is the confusion matrix diagram of different networks of radar data, (f) to (i) is the confusion matrix diagram of different networks of inertial data, and (j) to (n) is the confusion matrix diagram of different networks of radar and inertial data fusion.

In radar single-modal networks, the accuracy rates of various models reflect their capabilities in processing radar data. Like earlier convolutional neural network models, LeNet and GoogleNet, while demonstrating a certain degree of effectiveness, may be limited by their relatively simple network structures in capturing complex radar image features. AlexNet and VGG16 improve performance by deepening the network structure, with AlexNet achieving an accuracy rate of 0.913, surpassing LeNet and GoogleNet, indicating its greater suitability for processing radar data. However, ViT, a model based on the Transformer architecture, performs most outstandingly among radar single-modal networks, with an accuracy rate as high as 0.921. This demonstrates ViT's advantage in processing global information, especially in capturing key features in radar images.

Compared to radar single-modal networks, the accuracy rates of IMU single-modal networks are generally lower. CNN_CA and CNN, while demonstrating a certain degree of effectiveness, have limited performance. LSTM and GRU, as models for processing time-series data, do not show significant advantages when processing IMU data, with accuracy rates far below those of CNN-based models. However, when radar and IMU data are fused, the accuracy rates of the models are significantly improved. Fusion models combining ViT with GRU and LSTM exhibit good performance, with accuracy rates exceeding those of single radar or IMU models. This indicates that fusing different modal data can bring performance improvements. Among them, the fusion models of ViT+CNN and ViT+CNN_CA perform particularly well, with accuracy rates reaching 0.942 and 0.945, respectively. This further proves that combining ViT with CNN can fully utilize the advantages of both, especially in extracting local and global features.

Among all models, T-C-RIMFA has the highest accuracy rate, reaching 0.988. This fully demonstrates T-C-RIMFA's exceptional capability in processing fused data.

D. DIFFERENT CLASSIFICATION PERFORMANCE COMPARISON

Based on the above research, the performance measures of the data confusion matrix such as accuracy, accuracy, recall rate, and F1 value were compared to measure the classification effect of different models.

Accuracy measures the ratio of the number of samples correctly classified by the model to the total number of samples, and its expression is:

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{34}$$

Among them, TP (True Positives) represents the number of True cases, TN (True Negatives) represents the number of

FIGURE 8. Confusion matrix graph.

true negative cases, FP (False Positives) represents the number of False positive cases, FN (False Negatives) represents the number of false negative cases.

Precision measures how many of the samples predicted by the model to be positive examples are true examples, expressed as:

$$precision = \frac{TP}{TP + FP} \tag{35}$$

Recall measures the proportion of real cases predicted by the model, and its expression is:

$$recall = \frac{TP}{TP + FN} \tag{36}$$

F1 Score is the harmonic average of accuracy and recall rate, which is used to comprehensively measure the classification performance of the model, and its expression is as follows:

$$F1 = \frac{2 * precision * recall}{precision + recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{37}$$

According to the results of various classification indicators shown in Fig. 9, the T-C-RIMFA algorithm has excellent performance in all aspects of classification accuracy, which is maintained at about 98.8% . The results show that with the increase in test times, the accuracy of HAR using the improved T-C-RIMFA algorithm is significantly improved. This shows that even in the face of noise interference or attitude

TABLE 2. Accuracy Comparison

	Model	Accuracy
Radar-single	LeNet[51]	0.880
	GoogleNet[52]	0.825
	AlexNet[53]	0.913
	VGG16[54]	0.907
	ViT[55]	0.921
IMU-single	CNN_CA[56]	0.771
	CNN[57]	0.767
	LSTM[58]	0.514
	GRU[59]	0.641
Radar+IMU	ViT+GRU	0.895
	ViT+LSTM	0.905
	ViT+CNN	0.942
	ViT+CNN_CA	0.945
	T-C-RIMFA	**0.988**

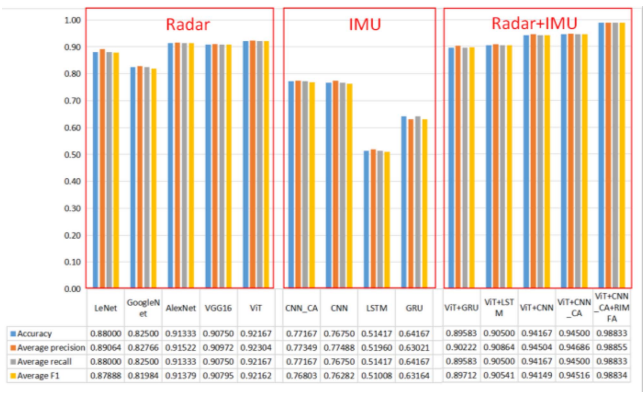


FIGURE 9. Results of different classification indicators.

TABLE 3. Ablation Experiment

	ViT	CNN	CA	RIMFA	Accuracy
a	√	—	—	—	0.92167
b	—	√	—	—	0.76750
c	—	√	√	—	0.77167
d	√	√	√	—	0.94500
e	√	√	√	√	**0.98833**

changes, the algorithm can better cope with complex dynamic environments and changes, and maintain a high recognition accuracy. Therefore, in order to better realize the practical application, we need to further expand the database.

E. ABLATION EXPERIMENT

On the basis of the above research, we designed a series of classification ablation experiments in order to further explore the model's learning of different features and its generalization ability in classification tasks. As shown in the ablation experiment in the table below, we observed changes in the model's performance by gradually adjusting specific features or information in the model's input to reveal how dependent the model is on various features.

In Table 3, various models are evaluated based on their feature extraction methodologies and accuracy rates. Model (a) employs only the ViT for feature extraction, relying solely on radar data, resulting in an accuracy of 0.92167. Model (b) utilizes a CNN for feature extraction, focusing exclusively on inertial navigation sensor data, which yields an accuracy of 0.76750. In Model (c), the CNN is enhanced with a CA mechanism while still using only inertial navigation sensor data, achieving an accuracy of 0.77167. Model (d) integrates both ViT and CNN for feature extraction along with the CA mechanism, utilizing both inertial navigation sensor data and radar data, and achieves a higher accuracy of 0.94500. Finally, model (e) builds on Model (d) by introducing the RIMFA mechanism, resulting in a significant increase in accuracy to 0.98833.

These findings underscore the advantages of combining multiple data sources and advanced feature extraction techniques to enhance model performance.

V. CONCLUSION

In this study, we propose an HAR method, which aims to maintain the accuracy and stability of sensor data under severe weather and accumulated errors. Our method combines the data of millimeter wave radar and inertial navigation sensor, uses ViT to extract the radar data, uses CNN to extract the inertial derivative data, and uses the channel attention mechanism to weight the features extracted from the inertial derivative data from different convolution kernels. The experimental results show that our model T-C-RIMFA can effectively distinguish five activities, including walking, running, slow walking, up and down stairs, as well as falling, with an overall accuracy of 98.8% . Compared with the traditional network model, our method shows a better classification effect.

However, we also realize that in practical applications, especially in multi-person interactions or dynamic environments, multi-target detection and tracking is an issue that cannot be ignored. Although in this study, we mainly focus on feature fusion rather than multi-target tracking, we believe that the challenges of single sensors in multi-target detection can be effectively alleviated through multi-sensor fusion (for example, combining radar and IMU). Future research can explore the combination of angle information (RDA spatial processing) and Kalman filtering technology on this basis to achieve more accurate and stable multi-target detection and trajectory tracking, so as to further improve the performance and practical application value of the system [61], [62].

In the future, our HAR method can be further developed and applied in many aspects. First of all, we can explore the application of this method in practical scenarios, such as fitness monitoring, medical diagnosis, security monitoring, and other fields, to achieve real-time monitoring and analysis of individual activity status. Moreover, we can integrate radar range-Doppler data or point cloud data to enhance spatial resolution and target positioning capabilities and improve the accuracy of activity recognition in more complex and dynamic environments. Secondly, we can consider expanding our HAR technology to gesture recognition, pose recognition, and other biometric recognition technologies to achieve a more comprehensive and diversified body behavior recognition system and provide a more intelligent and convenient solution for human-computer interaction, intelligent security, and other fields. In addition, we can further improve the accuracy and robustness of the model by improving the algorithm and data set to adapt to the application requirements in different environments and scenarios. Ultimately, we aim to build a comprehensive and efficient body behavior recognition system to provide more effective solutions to problems in urban traffic management, intelligent health monitoring, and other fields.

REFERENCES

[1] M. Lu, Y. Hu, and X. Lu, "Driver action recognition using deformable and dilated faster R-CNN with optimized region proposals," *Appl. Intell.*, vol. 50, no. 4, pp. 1100–1111, Apr. 2020, doi: 10.1007/s10489-019-01603-4.

[2] R. M. Raval, H. B. Prajapati, and V. K. Dabhi, "Survey and analysis of human activity recognition in surveillance videos," *Intell. Decis. Technol.*, vol. 13, no. 2, pp. 271–294, 2019, doi: 10.3233/IDT-170035.

[3] S.-H. Hsieh, Y.-J. Tsay, Y.-W. Chen, Y.-Y. Huang, and Y.-X. Yu, "Integrating FMCW radar and RGBD sensor for vital sign detection," in *Proc. IEEE 5th Eurasia Conf. Biomed. Eng., Healthcare Sustainability*, 2023, pp. 175–177, doi: 10.1109/ECBIOS57802.2023.10218461.

[4] S. M. M. Islam, O. Boric-Lubecke, and V. M. Lubecke, "Identity authentication in two-subject environments using microwave Doppler radar and machine learning classifiers," *IEEE Trans. Microw. Theory Techn.*, vol. 70, no. 11, pp. 5063–5076, Nov. 2022, doi: 10.1109/TMTT.2022.3197413.

[5] Y. Wang, "Multi-sensor fusion tracking algorithm based on augmented reality system," *IEEE Sensors J.*, vol. 21, no. 22, pp. 25010–25017, Nov. 2021, doi: 10.1109/JSEN.2020.3034139.

[6] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, D. Zhang, "Biometrics recognition using deep learning: A survey," *Artif. Intell. Rev.*, vol. 56, no. 8, pp. 8647–8695, Aug. 2023, doi: 10.1007/s10462-022-10237-x.

[7] M. H. Khan, M. S. Farid, and M. Grzegorzek, "Vision-based approaches towards person identification using gait," *Comput. Sci. Rev.*, vol. 42, Nov. 2021, Art. no. 100432, doi: 10.1016/j.cosrev.2021.100432.

[8] G. Yang, W. Tan, H. Jin, T. Zhao, and L. Tu, "Review wearable sensing system for gait recognition," *Cluster Comput.*, vol. 22, no. S2, pp. 3021–3029, Mar. 2019, doi: 10.1007/s10586-018-1830-y.

[9] A. A. Aguileta, R. F. Brena, O. Mayora, E. Molino-Minero-Re, and L. A. Trejo, "Multi-sensor fusion for activity recognition—A survey," *Sensors*, vol. 19, no. 17, Sep. 2019, Art. no. 3808, doi: 10.3390/s19173808.

[10] S. Qiu et al., "Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges," *Inf. Fusion*, vol. 80, pp. 241–265, Apr. 2022, doi: 10.1016/j.inffus.2021.11.006.

[11] H. F. Nweke, Y. W. Teh, G. Mujtaba, U. R. Alo, and M. A. Al-garadi, "Multi-sensor fusion based on multiple classifier systems for human activity identification," *Hum.-Centric Comput. Inf. Sci.*, vol. 9, no. 1, pp. 1–44, Dec. 2019, doi: 10.1186/s13673-019-0194-5.

[12] A. Venon, Y. Dupuis, P. Vasseur, and P. Merriaux, "Millimeter wave FMCW RADARs for perception, recognition and localization in automotive applications: A survey," *IEEE Trans. Intell. Veh.*, vol. 7, no. 3, pp. 533–555, Sep. 2022, doi: 10.1109/TIV.2022.3167733.

[13] Y. Zhao, A. Yarovoy, and F. Fioranelli, "Angle-insensitive Human motion and posture recognition based on 4D imaging radar and deep learning classifiers," *IEEE Sensors J.*, vol. 22, no. 12, pp. 12173–12182, Jun. 2022, doi: 10.1109/JSEN.2022.3175618.

[14] J. A. Nanzer, "A review of microwave wireless techniques for human presence detection and classification," *IEEE Trans. Microw. Theory Techn.*, vol. 65, no. 5, pp. 1780–1794, May 2017, doi: 10.1109/TMTT.2017.2650909.

[15] K. Shioiri and K. Saho, "Exploration of effective time-velocity distribution for Doppler-radar-based personal gait identification using deep learning," *Sensors*, vol. 23, no. 2, p. 604, Jan. 2023, doi: 10.3390/s23020604.

[16] Y. Yang, Y. Ge, B. Li, Q. Wang, Y. Lang, and K. Li, "Multiscenario open-set gait recognition based on radar micro-Doppler signatures," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 2519813, doi: 10.1109/TIM.2022.3214271.

[17] T. Mao, Y. Zhang, K. Zhu, T. Wang, and H. Sun, "Estimation of human gait features by trajectory tracking and recombination using radar range-Doppler-time data," *Int. Eng. Technol. Radar Sonar Navigation*, vol. 17, no. 2, pp. 236–246, Feb. 2023, doi: 10.1049/rsn2.12336.

[18] R. R. Sharma, G. Aravind, and R. Dubey, "Radar based automated system for people walk identification using correlation information and flexible analytic wavelet transform," *Appl. Intell.*, vol. 53, no. 24, pp. 30746–30756, Dec. 2023, doi: 10.1007/s10489-023-05159-2.

[19] M. A. Alanazi et al., "Towards a low-cost solution for gait analysis using millimeter wave sensor and machine learning," *Sensors*, vol. 22, no. 15, Jul. 2022, Art. no. 5470, doi: 10.3390/s22155470.

[20] F. Wang, P. Wang, X. Zhang, H. Li, and B. Himed, "An overview of parametric modeling and methods for radar target detection with limited data," *IEEE Access*, vol. 9, pp. 60459–60469, 2021, doi: 10.1109/ACCESS.2021.3074063.

[21] Y. Lang, Q. Wang, Y. Yang, C. Hou, Y. He, and J. Xu, "Person identification with limited training data using radar micro-Doppler signatures," *Microw. Opt. Technol. Lett.*, vol. 62, no. 3, pp. 1060–1068, Mar. 2020, doi: 10.1002/mop.32125.

[22] X. Qiao, Y. Feng, T. Shan, and R. Tao, "Person identification with low training sample based on micro-Doppler signatures separation," *IEEE Sensors J.*, vol. 22, no. 9, pp. 8846–8857, May 2022, doi: 10.1109/JSEN.2022.3162590.

[23] Y. Yang, Y. Zhang, H. Ji, B. Li, and C. Song, "Radar-based Human activity recognition under the limited measurement data support using domain translation," *IEEE Signal Process. Lett.*, vol. 29, pp. 1993–1997, 2022, doi: 10.1109/LSP.2022.3207948.

[24] Y. Meng, W. Wang, H. Han, and M. Zhang, "A vision/radar/INS integrated guidance method for shipboard landing," *IEEE Trans. Ind. Electron.*, vol. 66, no. 11, pp. 8803–8810, Nov. 2019, doi: 10.1109/TIE.2019.2891465.

[25] K. Qiu, T. Qin, J. Pan, S. Liu, and S. Shen, "Real-time temporal and rotational calibration of heterogeneous sensors using motion correlation analysis," *IEEE Trans. Robot.*, vol. 37, no. 2, pp. 587–602, Apr. 2021, doi: 10.1109/TRO.2020.3033698.

[26] N. Abhayasinghe and I. Murray, "Human gait modeling, prediction and classification for level walking using harmonic models derived from a single thigh-mounted IMU," *Sensors*, vol. 22, no. 6, Mar. 2022, Art. no. 2164, doi: 10.3390/s22062164.

[27] S. Yang et al., "Robust navigation method for wearable Human–Machine interaction system based on deep learning," *IEEE Sensors J.*, vol. 20, no. 24, pp. 14950–14957, Dec. 2020, doi: 10.1109/JSEN.2020.3010367.

[28] Y. Ding, Z. Xiong, W. Li, Z. Cao, and Z. Wang, "Pedestrian navigation system with trinal-IMUs for drastic motions," *Sensors*, vol. 20, no. 19, Sep. 2020, Art. no. 5570, doi: 10.3390/s20195570.

[29] Z. Wang, Z. Xiong, L. Xing, Y. Ding, and Y. Sun, "A method for autonomous multi-motion modes recognition and navigation optimization for indoor pedestrian," *Sensors*, vol. 22, no. 13, Jul. 2022, Art. no. 5022, doi: 10.3390/s22135022.

[30] A. K. Patil, A. Balasubramanyam, J. Y. Ryu, P. K. B N, B. Chakravarthi, and Y. H. Chai, "Fusion of multiple lidars and inertial sensors for the real-time pose tracking of human motion," *Sensors*, vol. 20, no. 18, Sep. 2020, Art. no. 5342, doi: 10.3390/s20185342.

[31] Z. Yu et al., "An intelligent implementation of multi-sensing data fusion with neuromorphic computing for human activity recognition," *IEEE Internet Things J.*, vol. 10, no. 2, pp. 1124–1133, Jan. 2023, doi: 10.1109/JIOT.2022.3204581.

[32] H. Li, A. Shrestha, H. Heidari, J. L. Kernec, and F. Fioranelli, "Magnetic and radar sensing for multimodal remote health monitoring," *IEEE Sensors J.*, vol. 19, no. 20, pp. 8979–8989, Oct. 2019, doi: 10.1109/JSEN.2018.2872894.

[33] P. Zhang et al., "Multi-source information fusion based on rough set theory: A review," *Inf. Fusion*, vol. 68, pp. 85–117, Apr. 2021, doi: 10.1016/j.inffus.2020.11.004.

[34] A. S. M. H. Bari and M. L. Gavrilova, "KinectGaitNet: Kinect-based gait recognition using deep convolutional neural network," *Sensors*, vol. 22, no. 7, Mar. 2022, Art. no. 2631, doi: 10.3390/s22072631.

[35] M. Rashmi and R. M. R. Guddeti, "Human identification system using 3D skeleton-based gait features and LSTM model," *J. Vis. Commun. Image Representation*, vol. 82, Jan. 2022, Art. no. 103416, doi: 10.1016/j.jvcir.2021.103416.

[36] G. Li, L. Guo, R. Zhang, J. Qian, and S. Gao, "TransGait: Multimodal-based gait recognition with set transformer," *Appl. Intell.*, vol. 53, no. 2, pp. 1535–1547, Jan. 2023, doi: 10.1007/s10489-022-03543-y.

[37] P. Delgado-Santos, R. Tolosana, R. Guest, F. Deravi, and R. Vera-Rodriguez, "Exploring transformers for behavioural biometrics: A case study in gait recognition," *Pattern Recognit.*, vol. 143, Nov. 2023, Art. no. 109798, doi: 10.1016/j.patcog.2023.109798.

[38] C. Ma and Z. Liu, "A novel spatial–Temporal network for gait recognition using millimeter-wave radar point cloud videos," *Electronics*, vol. 12, no. 23, Nov. 2023, Art. no. 4785, doi: 10.3390/electronics12234785.

[39] Y.-W. Kim, W.-H. Cho, K.-S. Kim, and S. Lee, "Inertial-measurement-unit-based novel human activity recognition algorithm using conformer," *Sensors*, vol. 22, no. 10, May 2022, Art. no. 3932, doi: 10.3390/s22103932.

[40] Y. Zhou, X. Jiang, G. Xu, X. Yang, X. Liu, and Z. Li, "PVT-SAR: An arbitrarily oriented SAR ship detector with pyramid vision transformer," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 291–305, 2023, doi: 10.1109/JSTARS.2022.3221784.

[41] J. N. Mogan, C. P. Lee, K. M. Lim, M. Ali, and A. Alqahtani, "Gait-CNN-ViT: Multi-model gait recognition with convolutional neural networks and vision transformer," *Sensors*, vol. 23, no. 8, Apr. 2023, Art. no. 3809, doi: 10.3390/s23083809.

[42] Y. Zheng, Y. Yang, and W. Chen, "A novel range compression algorithm for resolution enhancement in GNSS-SARs," *Sensors*, vol. 17, no. 7, Jun. 2017, Art. no. 1496, doi: 10.3390/s17071496.

[43] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 3992–4003, doi: 10.1109/ICCV51070.2023.00371.

[44] H. Liu, Y. Hartmann, and T. Schultz, "A practical wearable sensor-based Human activity recognition research pipeline," in *Proc. 15th Int. Joint Conf. Biomed. Eng. Syst. Technol.*, 2022, pp. 847–856, doi: 10.5220/0010937000003123.

[45] G. Sarapata et al., "Video-based activity recognition for automated motor assessment of Parkinson's disease," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 10, pp. 5032–5041, Oct. 2023, doi: 10.1109/JBHI.2023.3298530.

[46] Y. Tang et al., "Patch slimming for efficient vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12155–12164, doi: 10.1109/CVPR52688.2022.01185.

[47] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 7242–7252, doi: 10.1109/ICCV48922.2021.00717.

[48] X. Dong et al., "CLIP itself is a strong fine-tuner: Achieving 85.7% and 88.0% top-1 accuracy with ViT-B and ViT-L on ImageNet," 2022, *arXiv:2212.06138*.
[49] X. Li, S. Chen, S. Zhang, Y. Zhu, Z. Xiao, and X. Wang, "Advancing IR-UWB radar Human activity recognition with Swin transformers and supervised contrastive learning," *IEEE Internet Things J.*, vol. 11, no. 7, pp. 11750–11766, Apr. 2024, doi: 10.1109/JIOT.2023.3330996.
[50] H. Touvron, M. Cord, and H. Jégou, "DeiT III: Revenge of the ViT," in *Comput. Vis. – ECCV*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., Lecture Notes in Computer Science, vol. 13684, Cham: Springer, 2022, doi: 10.1007/978-3-031-20053-3_30.
[51] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1204–1213, doi: 10.1109/CVPR52688.2022.01179.
[52] M. Zhao et al., "Robust and breathable all-textile gait analysis platform based on LeNet convolutional neural networks and embroidery technique," *Sensors Actuators A, Phys.*, vol. 360, Oct. 2023, Art. no. 114549, doi: 10.1016/j.sna.2023.114549.
[53] L. Yang et al., "GoogLeNet based on residual network and attention mechanism identification of rice leaf diseases," *Comput. Electron. Agriculture*, vol. 204, Jan. 2023, Art. no. 107543, doi: 10.1016/j.compag.2022.107543.
[54] A. Nainwal, G. Sharma, V. Kansal, S. Bhatla, and B. Pant, "Comparative study of VGG-13, AlexNet, MobileNet and modified-DarkCovidNet for chest X-ray classification," in *Proc. IEEE 10th Int. Conf. Comput. Sustain. Glob. Develop.*, 2023, pp. 413–417.
[55] I. Ali, I. Junaid, and S. Ari, "VGG-16 based gait recognition using skeleton features," in *Proc. IEEE Int. Conf. Device Intell., Comput. Commun. Technol.*, 2023, pp. 597–601, doi: 10.1109/DICCT56244.2023.10110160.
[56] J. N. Mogan, C. P. Lee, K. M. Lim, and K. S. Muthu, "Gait-ViT: Gait recognition with vision transformer," *Sensors*, vol. 22, no. 19, Sep. 2022, Art. no. 7362, doi: 10.3390/s22197362.
[57] I. Junaid, I. Ali, N. P. Sharma, and S. Ari, "Gait identification using deep convolutional network and attention technique," in *Proc. IEEE 6th Int. Conf. Condition Assessment Techn. Elect. Syst.*, 2022, pp. 334–338, doi: 10.1109/CATCON56237.2022.10077691.
[58] H. Huang, P. Zhou, Y. Li, and F. Sun, "A lightweight attention-based CNN model for efficient gait recognition with wearable IMU sensors," *Sensors*, vol. 21, no. 8, Apr. 2021, Art. no. 2866, doi: 10.3390/s21082866.
[59] M. Sarshar, S. Polturi, and L. Schega, "Gait phase estimation by using LSTM in IMU-based gait analysis—Proof of concept," *Sensors*, vol. 21, no. 17, Aug. 2021, Art. no. 5749, doi: 10.3390/s21175749.
[60] Y.-W. Kim, K.-L. Joa, H.-Y. Jeong, and S. Lee, "Wearable IMU-based human activity recognition algorithm for clinical balance assessment using 1D-CNN and GRU ensemble model," *Sensors*, vol. 21, no. 22, Nov. 2021, Art. no. 7628, doi: 10.3390/s21227628.
[61] Z. Zhang, K. Fu, X. Sun, and W. Ren, "Multiple target tracking based on multiple hypotheses tracking and modified ensemble Kalman filter in multi-sensor fusion," *Sensors*, vol. 19, no. 14, Jul. 2019, Art. no. 3118, doi: 10.3390/s19143118.
[62] D. Gaglione, P. Braca, G. Soldi, F. Meyer, F. Hlawatsch, and M. Z. Win, "Fusion of sensor measurements and target-provided information in multitarget tracking," *IEEE Trans. Signal Process.*, vol. 70, pp. 322–336, 2022, doi: 10.1109/TSP.2021.3132232.

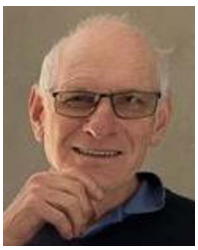
JIAJIA SHI (Member, IEEE) received the B.Sc. and M.E. degrees from Central South University, Changsha, China, in 2007 and 2010, respectively, and the Ph.D. degree from the University of Technology at Sydney, Sydney, NSW, Australia, in 2015. He is currently an Associate Professor with the School of Transportation and Civil Engineering, Nantong University, Nantong, China. His research interests include signal processing.

YIHAN ZHU received the B.S. degree in 2022 from the School of Transportation and Civil Engineering, Nantong University, Nantong, China, where she is currently working toward the M.S. degree. Her current research interests include radar signal processing and deep learning.

JIAQING HE received the B.S. degree from the School of Computer and Software Engineering, Anhui Institute of Information Technology, Wuhu, China, in 2021. He is currently working toward the M.S. degree from the School of Transportation and Civil Engineering, Nantong University, Nantong, China. His research interests include radar signal processing and human behavior recognition.

ZHIHUO XU (Senior Member, IEEE) received the Ph.D. degree in communication and information system from the University of Chinese Academy of Sciences, Beijing, China, in 2016. He founded Radar Remote Sensing Group, Nantong University, China, in 2016. From 2017 to 2018, he was an Academic Visitor with the University of Birmingham, Birmingham, U.K. His research interests include radar system design, radar signal, and image processing.

LIU CHU (Senior Member, IEEE) received the B.E. degree in material science and engineering and the M.E. degree in mechanics from Dalian Maritime University, Dalian, China, in 2010 and 2012, respectively, and the Ph.D. degree in mechanics from the Institut National des Sciences Appliquées de Rouen (INSA Rouen), Rouen, France, in 2017. She is currently a Research Associate with the School of Physical Science and Technology, ShanghaiTech University, Shanghai, China. Her research interests include artificial material microstructure optimization.



ROBIN BRAUN (Life Senior Member, IEEE) received the B.Sc. degree (Hons.) from Brighton University, Brighton, U.K., in 1980, and the M.Sc. (Eng.) and Ph.D. degrees from the University of Cape Town, Cape Town, South Africa, in 1982 and 1986, respectively. He started his academic career with the University of Cape Town, in 1986. In 1998, he moved to the University of Technology Sydney, Sydney, NSW, Australia, where he occupied the Chair of Telecommunications Engineering. His recent work has been in network protocols and the management of complex next-generation networks.

QUAN SHI (Member, IEEE) received the M.S. and Ph.D. degrees in management information systems from the University of Shanghai for Science and Technology, Shanghai, China, in 2005 and 2011, respectively. He is currently a Professor with the School of Transportation and Civil Engineering, Nantong University, Nantong, China. His research interests include the development of signal and image processing and big data techniques.