Sparsity-based Human Activity Recognition with PointNet using a Portable FMCW Radar

Ding, Chuanwei; Zhang, Li; Chen, Haoyu; Hong, Hong; Zhu, Xiaohua; Fioranelli, Francesco

DOI
10.1109/JIOT.2023.3235808

Publication date
2023

Document Version
Final published version

Published in
IEEE Internet of Things Journal

Sparsity-Based Human Activity Recognition With PointNet Using a Portable FMCW Radar

Chuanwei Ding, *Student Member, IEEE*, Li Zhang, *Student Member, IEEE*, Haoyu Chen, Hong Hong, *Senior Member, IEEE*, Xiaohua Zhu, *Member, IEEE*, and Francesco Fioranelli, *Senior Member, IEEE*

***Abstract*—Radar-based solutions have attracted great attention in human activity recognition (HAR) for their advantages in accuracy, robustness, and privacy protection. The conventional approaches transform radar signals into feature maps and then directly process them as visual images. While effective, these image-based methods may not be the best solutions in terms of representation efficiency to encode the relevant information for classification. This article proposes a novel HAR method combining sparse theory and PointNet network, with both operations in the time-Doppler (TD) and range-Doppler (RD) domains. First, sparsity-based feature extraction is introduced to use a limited number of sparse solutions to characterize human activities in the form of TD sparse point clouds (TDSP) or dynamic RD sparse point clouds (DRDSP). This new representation is validated by comparing the reconstructed and original signals. Then, PointNet networks are adopted to summarize multidomain features and predict human activity labels by a sparse set of input point clouds. Comprehensive experiments were conducted to demonstrate that the proposed method can yield a higher representation efficiency, classification accuracy, and better generalization capability than existing ones.**

***Index Terms*—Frequency-modulated continuous wave (FMCW) radar, human activity recognition (HAR), PointNet, sparse representation.**

I. Introduction

HUMAN activity recognition (HAR) has drawn significant attention in wide applications, such as public monitoring, disaster rescue, intelligent interaction, and assisted living [1], [2], [3], [4], [5]. Numerous approaches for HAR have been studied [6], [7], [8], [9], [10], [11]. Radar-based solutions may complement conventional wearable and video technologies because of their advantages in accuracy, robustness, and privacy protection [12], [13], [14]. Radar reflections from human subjects performing activities cause modulations in the signal frequencies defined as micro-Doppler. These signatures contain prominent features that are specific to different human activities.

Most of the conventional radar-based HAR works can be considered to be inspired from image-based methods. Their primary idea is to transform radar signals into matrices or feature maps first and then directly process them as images. The typical feature maps in the literature include time-range (TR) maps, time-Doppler (TD) maps, and range-Doppler (RD) maps/frames. After generating these images, subsequent feature extractions generally include two categories: 1) traditional methods and 2) deep learning ones.

Traditional methods apply various algorithms to extract multidomain features from feature maps as input to the classifiers. They consist of envelope features, singular value decomposition (SVD) features, principal component analysis (PCA) features, constant false alarm rate algorithm (CFAR) features, among others. For example, in [15], envelope-based features, including extreme Doppler, torso frequency, and event length, were extracted from the TD map to a support vector machine (SVM). In [16], the SVD algorithm was applied on the TD map to obtain feature vectors and their statistic information, including average, standard deviation, and variance, was used with a Naïve Bayes classifier to distinguish between armed and unarmed people. In [17], a multilinear PCA was introduced to extract features from a radar data cube, a fusion of TR and TD maps. The output principal components were used as extracted features for HAR. In [18] and [19], a CFAR detector was applied on range-velocity maps to get the pixels through a power intensity threshold to construct feature point groups. In [20], a cell-averaging CFAR (CA-CFAR) algorithm was utilized to extract multiple scatterers of the human body from the RD map. The extracted scatterers constitute a set of point clouds to reveal the person's physical shape information. In [21], the high-intensity RD points were extracted in a time sequence. Their statistic values were calculated to indicate the dynamic RD trajectory of human activities. In general, traditional approaches rely on empirical thresholds and are easily influenced by noise resulting in a poor generalization.

Manuscript received 18 November 2022; accepted 6 January 2023. Date of publication 10 January 2023; date of current version 23 May 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFC2005302; in part by the National Natural Science Foundation of China under Grant 61871224 and Grant 62201259; and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20220942 and Grant BK20220940. *(Chuanwei Ding and Li Zhang contributed equally to this work.) (Corresponding author: Hong Hong.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Nanjing Integrated Traditional Chinese and Western Medicine Hospital Human Research Project under Application No. 201812001.

Chuanwei Ding, Li Zhang, Haoyu Chen, Hong Hong, and Xiaohua Zhu are with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: hongnju@njust.edu.cn).

Francesco Fioranelli is with the MS3 Group, Faculty of Electrical Engineering, Mathematics and Computer Science, TU Delft, 2628CD Delft, The Netherlands (e-mail: f.fioranelli@tudelft.nl).

Digital Object Identifier 10.1109/JIOT.2023.3235808

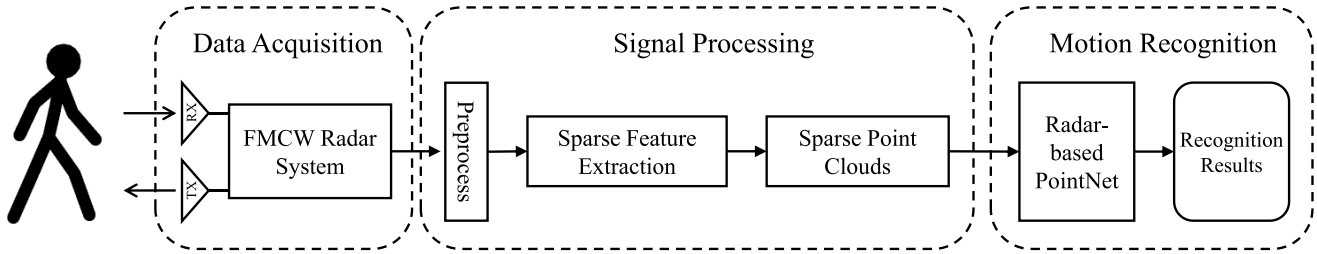

Fig. 1. Framework of the proposed sparsity-based PointNet method.

Deep learning has drawn significant attention in HAR for its advantages in automatic deep features extraction and large data processing [22], [23], [24], [25]. Deep learning methods can directly treat feature maps as images to automatically extract features and classification boundaries by applying hierarchical abstractions and generalization. The popular deep neural networks in HAR can be divided into three categories: 1) deep convolutional neural network (DCNN); 2) recurrent neural network (RNN); and 3) DCNN + RNN. In [26], a DCNN was applied to the TD map for the joint learning of necessary features and classification boundaries without any explicit features. A dual-channel DCNN (DC-DCNN) was proposed in [27] for human gait recognition. Two separate TD spectrums with different temporal resolutions were taken as input to the dual-channel architecture. Moreover, a 3-D convolutional neural network (CNN)-based network was designed to extract the complex features from RD frames and achieved good performance in real-time fall detection [28]. RNN-based methods perform well in processing sequential data streams for considering both current and previous observations [29], [30]. In [31], long short-term memory (LSTM) units were introduced to automatically identify sequential features from TD maps to classify six different types of human motions. In [32], both temporal forward and backward correlated information in the Doppler spectrogram were captured by a bidirectional LSTM (Bi-LSTM) architecture for continuous activity monitoring and classification. Furthermore, in [33], a deep neural network combining DCNN and LSTM was proposed to process TD maps to reach approximately a global accuracy rate of 90% in identifying falling and nonfalling events based on ultrawide band (UWB) radars.

However, image-based processing of radar data may not be the optimal solution in all cases and suffer from poor interpretability or artefacts, such as the side lobes of stronger components that may hide the contribution of weaker ones. Sparse theory can provide a new perspective for radar signal processing in HAR because of its potential for less dependency on empirical parameters and the possibility to compact the relevant information into a small number of points/coefficients. This technique has been studied for instance to extract dynamic hand gesture micro-Doppler features [34]. Moreover, our preliminary work has validated its feasibility in extracting TD features of human activities [35]. However, the study was limited to micro-Doppler features and traditional machine learning classifiers, which may not be the best choice.

In this work, a novel method for HAR combining sparse theory and PointNet network is proposed. First, sparse representation theory was applied to project echo signals in both TD and RD domains. Next, orthogonal matching pursuit (OMP) and its modified version were adopted to achieve corresponding sparse solutions. These sparse solutions could use a limited number of nonzero items to characterize human activities with physical meanings in the form of TD sparse point clouds (TDSP) or dynamic RD sparse point clouds (DRDSP). These representations can be validated by a comparison between the reconstructed signals and the original ones. Then, PointNet architectures were adopted to take the coordinates and values of the point clouds as input and predict their corresponding labels. The results of comprehensive experiments demonstrate that our proposed method obtains significant improvement compared to the conventional image-based ones. The contributions of this study are as follows.

1) To the best of our knowledge, this is the first work to investigate the fusion of sparse theory and PointNet network for HAR, differently from point cloud processing for mm-wave radar data as in [18], [19], and [20].
2) A sparsity-based algorithm is formulated to extract TD and RD features from human activity in the form of sparse point clouds. The advantages of the sparsity-based algorithm are its high representation efficiency, less dependency on empirical parameters, and clear physical meaning.
3) PointNet networks are applied to support the multidomain sparse point clouds as direct input for classification. This architecture overcomes the constraint of the conventional ones which requires transforming point clouds into images, and shows high robustness to small perturbation of input points [36].

The remainder of this article is organized as follows. Section II introduces the theory and proposed algorithm. In Section III, we explain the system, data acquisition, and implement details. Section IV presents analysis and discussion of the key parameters and recognition performance. Section V contains the conclusion.

II. Theory and Proposed Algorithm

The simplified overview of our proposed sparsity-based PointNet method for HAR is illustrated in Fig. 1. In the data acquisition part, the echo signals are recorded by a frequency-modulated continuous wave (FMCW) radar system. In the signal processing part, after preprocessing, sparse feature

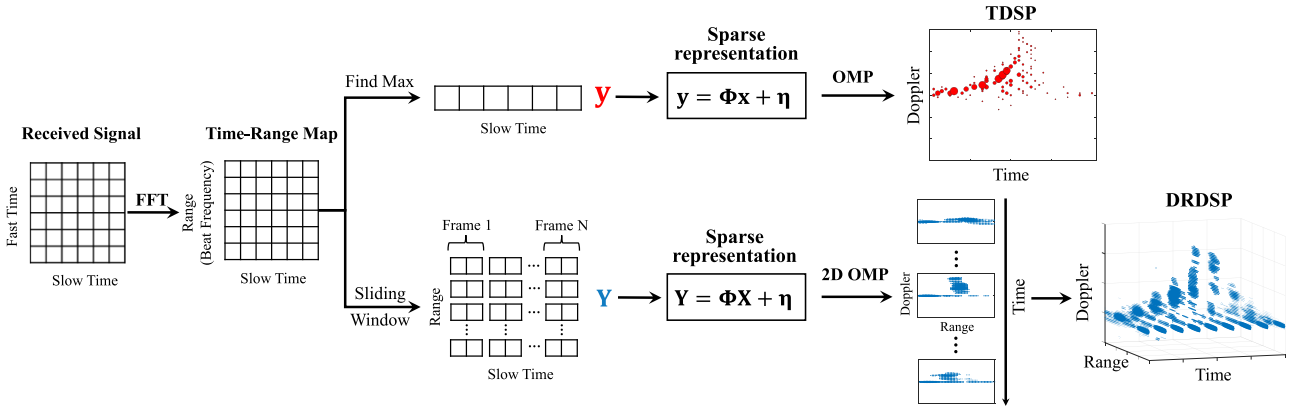

Fig. 2. Flowchart of our proposed sparsity-based feature extraction algorithms in TD (top) and RD (bottom) data domains.

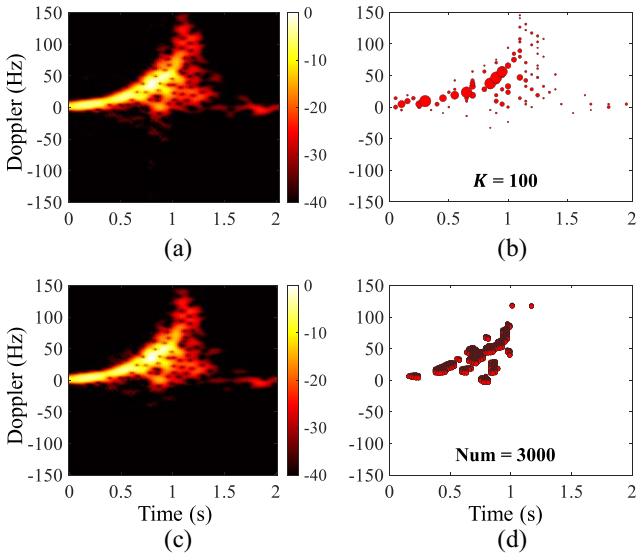

Fig. 3. Feature maps of an example of *fall* motion in TD domain. (a) Conventional TD map with STFT. (b) TDSP. (c) Reconstructed TD map with TDSP. (d) TD point cloud generated by applying CFAR on the conventional TD map.

extraction algorithms in the TD and RD domains are proposed to obtain the corresponding sparse point clouds, respectively. In the activity classification part, the above point cloud features are input to the proposed PointNet network to achieve the final classification results.

A. Sparsity-Based Features in Time-Doppler Domain

The flowchart of the sparsity-based feature extraction in the TD domain is illustrated in Fig. 2 (Top). The received signals can be expressed in a form of the signal matrix consisting of slow time and fast time. In the preprocessing part, a fast Fourier transform (FFT) is performed along the fast time axis to obtain the TR map. After direct current (dc) removal, the range bin with the maximum intensity value is regarded as where human activities happen. This can be denoted as an $N \times 1$ vector $\mathbf{y}$. Traditionally, $\mathbf{y}$ can be transformed into a TD map by short-time Fourier transform (STFT). Fig. 3(a) shows the conventional TD map of an example *fall* motion. It can be noticed that the high-intensity parts only occupy a limited area in the TD map. In other words, the projection of human activity information in the TD domain can be assumed sparse. Inspired by the work in [34], $\mathbf{y}$ can be expressed as a sparse representation

$$\mathbf{y} = \mathbf{\Phi}\boldsymbol{x} + \boldsymbol{\eta} \tag{1}$$

where $\mathbf{\Phi}$ is an $N \times M$ sparse dictionary, $\mathbf{x}$ indicates an $M \times 1$ sparse vector, and $\boldsymbol{\eta}$ represents an $N \times 1$ noise vector. To project the received signals into the TD domain, $\mathbf{\Phi}$ is set as the Gaussian-windowed Fourier basis signal

$$\mathbf{\Phi}[:, m] = [\phi_m(1), \phi_m(2), \ldots, \phi_m(N)]^T \tag{2}$$

where

$$\begin{aligned}\phi_m(n) &\triangleq \phi(n|t_m, f_m) \\ &= \frac{1}{2^{\frac{1}{4}}\sqrt{\sigma}} \exp\left[-\frac{(n-t_m)^2}{\sigma^2}\right] \exp(-j2\pi f_m n)\end{aligned} \tag{3}$$

t_m is the time shifting, f_m is the frequency shift, σ indicates the variance of the Gaussian window, and $n = 1, 2, \ldots, N$.

According to the sparse theory [37], when $K \leq N < M$, the K-sparse approximated vector, $\widehat{\mathbf{x}}$ can be calculated from $\mathbf{y}$ as follows:

$$\widehat{\mathbf{x}} = \arg\min_{x} \|\mathbf{y} - \mathbf{\Phi}\boldsymbol{x}\|_2, \text{s.t.} \|\mathbf{x}\|_0 \leq K \tag{4}$$

where $\|\cdot\|_0$ and $\|\cdot\|_2$ are L_0 and L_2 norm, respectively. K represents the number of nonzero items in $\widehat{x}$. Since human activity is always related to high-intensity parts in the TD domain, the classic OMP algorithm is considered to solve (4) [38]. During each iteration of OMP, the sparse support of $\mathbf{x}$ is calculated at first. Then, the nonzero coefficients are determined by the least square estimator. Therefore, the estimated K-sparse solution can be expressed as follows:

$$\begin{aligned}\widehat{\mathbf{x}} &= \text{OMP}(\mathbf{y}, \mathbf{\Phi}, K) \\ &= (0, \ldots, \widehat{x}, 0, \ldots, \widehat{x}, 0, \ldots, \widehat{x}_k)^T\end{aligned} \tag{5}$$

where $\widehat{x}_k (k = 1, 2, \ldots, K)$ represents nonzero items. In this way, $\mathbf{y}$ can be expressed as follows:

$$\mathbf{y}(n) = \sum_{k=1}^{K} \widehat{x}_k \phi(n|t_k, f_k) + \boldsymbol{\eta}(n). \tag{6}$$

Algorithm 1 Modified 2-D OMP Algorithm

Input: Sparse dictionary $\mathbf{\Phi}$, the received signal matrix $\mathbf{Y}$, sparsity K;
Output: K-Sparse approximated matrix $\widehat{\mathbf{X}}$;
Initialization: Residual $\boldsymbol{\gamma} = \mathbf{Y}$;
for $k = 1$ to K **do**
Step 1: $[i_k, j_k] = \arg\max_{i,j} |\langle \boldsymbol{\gamma}_j, \boldsymbol{\phi}_i \rangle|$
Step 2: $\widehat{x} = \arg\min_{x} \|\boldsymbol{\gamma}_{j_k} - \boldsymbol{\phi}_{i_k} x\|_2$
Step 3: $\boldsymbol{\gamma}_{j_k} = \boldsymbol{\gamma}_{j_k} - \boldsymbol{\phi}_{j_k} \widehat{x}$
Step 4: $\widehat{\mathbf{X}}(i, j) = \widehat{x}$
end

The position coordinates of nonzero items $\widehat{x}_k$ in K-sparse solution $\widehat{\mathbf{x}}$ represent the extracted time and Doppler features. Thus, the TDSP of $\mathbf{y}$ can be denoted as follows:

$$T(\mathbf{y}) = (t_k, f_k, |\widehat{x}_k|), \quad k = 1, 2, \ldots, K \tag{7}$$

where $|\widehat{x}_k|$ represents the power intensity at (t_k, f_k) in TDSP.

The above method is performed on the same fall sample from Fig. 3(a). The sparsity K is set as 100, and the variance σ of the Gaussian window is set to be 32 as a tradeoff between time and Doppler resolutions. The nonzero items in the sparse solution calculated by (6) and (7) can be rearranged to consist of a TDSP, as shown in Fig. 3(b). In detail, red points indicate the 100 nonzero items, i.e., sparse point cloud. Their x-coordinates represent time information, y-coordinates represent Doppler information, and the point sizes indicate power intensity. Therefore, these sparse points can reveal the time, Doppler, and power intensity features of human activity.

Moreover, once obtaining the sparse dictionary $\mathbf{\Phi}$ and the sparse solution $\widehat{\mathbf{x}}$, we can achieve the reconstructed signal y_{rec} as follows:

$$\mathbf{y}_{\text{rec}} = \mathbf{\Phi}\widehat{\mathbf{x}}. \tag{8}$$

Next, STFT is performed on the reconstructed signal to obtain a reconstructed TD map, as illustrated in Fig. 3(c). Compared with Fig. 3(a), the reconstructed signal y_{rec} preserved most TD information and suppress low-intensity noises. This validates that the TDSP can use only 100 feature points, instead of a 1800×100 picture, to characterize human activity in time, Doppler, and power intensity domains with a small information loss.

Furthermore, a 2-D CFAR detector [39] was applied on the STFT-based TD map of Fig. 3(a) as a conventional acquisition method for point cloud features. The CFAR detector depends on empirical parameters which need to be fine-tuned for a good performance. In detail, these parameters include: the train cell size set as 3×3, the guard cell size set as 2×2, and the probability of false alarm set as 0.483. Then, the detected points are shown in Fig. 3(d). Though the number of CFAR-based point clouds was increased to 3000, most of them were overlapping and distributed in a small high-intensity area. These clustered point clouds can hardly describe the complete trend of Doppler changes over time, not to mention the correct representation of TD information.

This phenomenon is assumed to be caused by image-based processing way. When using STFT, the side lobes of stronger components may hide the contribution of weaker components that are not detected by the CFAR or require a large number of points for their representation. On the contrary, in the proposed sparsity-based feature extraction method, OMP is adopted to address this problem. In Algorithm 1 step 3, the update of residue $\boldsymbol{\gamma} = \boldsymbol{\gamma} - \boldsymbol{\phi}_{j_k}\widehat{x}$ removes the "explained signal portion" and takes the "unexplained portion" as the residue at each iteration. This helps ignore the side lobe effects of strong components and avoid repeatedly extracting the same component, thus, improving feature extraction.

B. Sparsity-Based Features in Range-Doppler Domain

The distribution of Doppler information in the range domain is also essential for HAR. The sparsity-based feature extraction in the RD domain is proposed as shown in Fig. 2 (Bottom).

Unlike in the TD domain, the target signal that needs to be represented in sparse form changes from a vector signal in the chosen range bin to an $N \times M$ matrix signal $\mathbf{Y}$ in all range bins. N is the slow time index, and M is the range bin index. Then, matrix $\mathbf{Y}$ can be expressed as a sparse representation in the RD domain

$$\mathbf{Y} = \mathbf{\Phi}\boldsymbol{X} + \boldsymbol{\eta} \tag{9}$$

where $\mathbf{\Phi}$ represents an $N \times N$ sparse dictionary, $\mathbf{X}$ is an $N \times M$ sparse matrix, and $\boldsymbol{\eta}$ indicates an $N \times M$ noise matrix. Similarly, $\mathbf{X}$ can be called K-sparse signal if there are only K nonzero items. To achieve the sparse representation of human activity in the RD domain, the Fourier basis function is adopted as the sparse dictionary $\mathbf{\Phi}$. Then, the nth row, mth column of the dictionary can be expressed as follows:

$$\mathbf{\Phi}(n, m) = e^{-j\frac{2\pi}{N}(n-1)\times(m-1)}. \tag{10}$$

When $K \leq \text{NM}$, the sparse matrix $\mathbf{X}$ can be approximated from $\mathbf{Y}$ as follows:

$$\widehat{\mathbf{X}} = \|\mathbf{Y} - \mathbf{\Phi}\boldsymbol{X}\|_2, \qquad \text{s.t. } \|\mathbf{X}\|_0 \leq K. \tag{11}$$

As shown in Algorithm 1, the OMP algorithm can be modified for 2-D processing to search for the optimal solution along both the range and Doppler axis. Then, (11) can be solved to obtain sparse solutions as follows:

$$\widehat{\mathbf{X}} = \text{OMP}(\mathbf{Y}, \mathbf{\Phi}, K) = \begin{bmatrix} 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \widehat{x}_1 & \cdots & \iddots & \vdots \\ 0 & \widehat{x}_2 & 0 & \cdots & 0 \\ \vdots & \iddots & \cdots & \widehat{x}_k & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \end{bmatrix} \tag{12}$$

where $\widehat{\mathbf{X}}$ is a K-sparse matrix, i.e., the sparse solution, and $\widehat{x}_k$ $(k = 1, 2, \ldots, K)$ are corresponding nonzero items. Similarly, their positions in the matrix indicate the range and Doppler information, and their values denote power intensity. Therefore, the sparse solution matrix $\widehat{\mathbf{X}}$ can be used to form an RD sparse point cloud of human activities from the received signal $\mathbf{Y}$. Furthermore, since human activity is

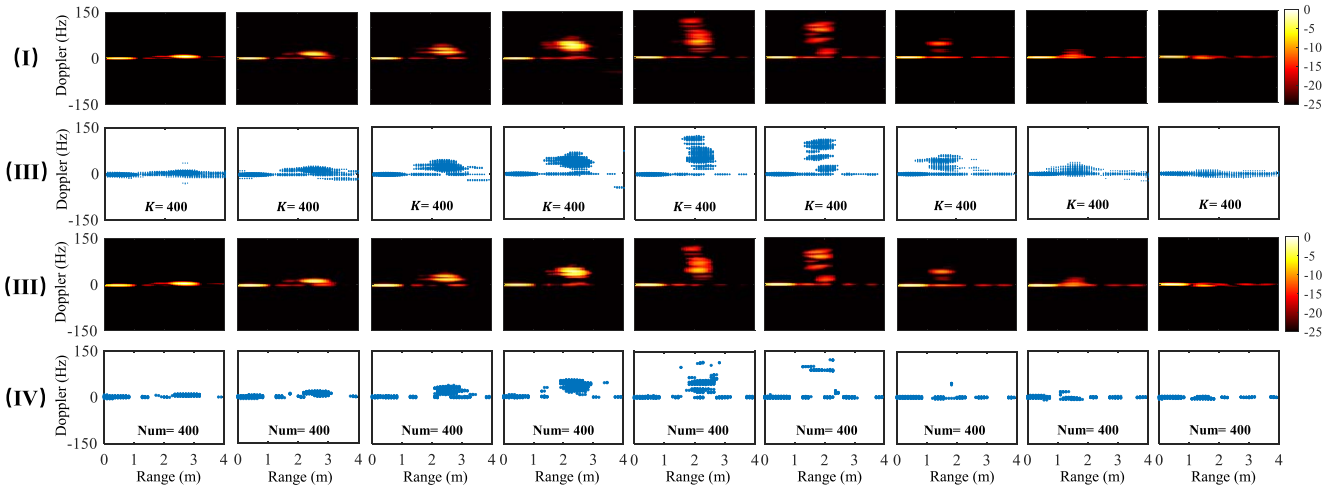

Fig. 4. Feature maps of the example *fall* motion in RD domain. (I) Conventional DRDF with the original signal. (II) DRDSP with the sparsity parameter K equal to 400. (III) Conventional DRDF with the reconstructed signal. (IV) RD point clouds generated by applying CFAR on the conventional DRDF.

dynamic, temporal information is significant in HAR. In our previous study [21], a sliding window with limited duration was adopted to achieve a time sequence of RD maps, i.e., dynamic RD frames. Similarly, RD sparse point clouds can also be transformed into DRDSP to describe time-varying features better. In this work, the sliding window length was set as 0.4 s with an overlap of 50%. Considering real-life situations, nine frames covering 2 s are selected to describe an entire human activity.

Fig. 4 shows feature maps of the example *fall* motion in the RD domain. In detail, row (I) consists of conventional 2-D FFT-based dynamic RD frames. During the entire *fall* motion, the human body got close to the radar with a small Doppler at first. Then, as the *fall* went on, the point target model turned to a body target model, resulting in an extension of the spectrogram in the Doppler axis. This phenomenon peaked at the fifth frame corresponding to the highest radial speed and the widest bandwidth. Finally, after a steep decrease, Doppler returned to zero, and the intensity gathered around the baseband again as the human body lay on the floor.

On the other hand, row (II) illustrated DRDSP obtained by the proposed sparse method. The sparsity K in each frame was set to 400. The blue dots containing range, Doppler, and intensity information can describe the RD distribution of *fall* motion. Particularly, Y_{rec} can be reconstructed with the sparse dictionary $\mathbf{\Phi}$ and the sparse solution $\widehat{\mathbf{X}}$ as follows:

$$\mathbf{Y}_{\text{rec}} = \mathbf{\Phi}\widehat{\mathbf{X}}. \tag{13}$$

Then, the conventional dynamic RD frames based on the reconstructed signal can be achieved, as shown in Fig. 4 row (III). Compared to row (I), a conclusion can be drawn that the proposed method can extract 400 sparse points per frame, instead of the 512×51 pictures, to characterize human activity through the time-varying RD features with a small information loss.

In addition, the 2-D CFAR method (train cell size: 2×2; guard cell size: 1×1; and probability of false alarm: 0.435) was applied on each frame of row (I) to extract CFAR-based point cloud features, as shown in row (IV). The number of CFAR-based feature points was normalized to 400 for each frame. Most of them gathered around clutter components and lost some key information, especially, in the 5th–7th frames. Compared with the sparse point clouds, the extraction of the CFAR-based point clouds highly depends on the empirical parameters of the algorithm, including the size of cells and false alarm probability, and showed a lower efficiency in representing the salient information within the data.

C. PointNet Network

The point clouds have attracted extensive concerns as an essential type of data structure. Usually, they are forced to a regular format, such as images or 3-D voxel grids before fed into conventional neural networks. PointNet network breaks this constraint by supporting point clouds as direct input [40]. Since the sparsity-based feature extraction method succeeds in using sparse point clouds to characterize human activities in time, Doppler, range, and intensity domains, the PointNet network is applied to process TDSP and DRDSP.

The proposed PointNet network is based on the architecture of the one introduced in [36]. As shown in Fig. 5, the input sparse point cloud is rearranged as a $n \times P$ feature matrix, where n is the number of sparse points and P is the number of feature vectors. For TDSP, P is set to 3 to build the 3D-PointNet, corresponding to time, Doppler, and intensity, respectively. For DRDSP, P is 4 to add the range vector as the 4D-PointNet. In particular, unlike conventional space coordinates, the above radar-based features do not satisfy the affine invariance. Therefore, the input transformation from [36] that aligns features from different input by a mini network (T-Net) can be ignored. Then it is mapped to a C_1-dimensional feature by multilayer perceptron (MLP) and aligned by a feature transformation of the T-Net network. Next, the MLP algorithm is adopted again to map the point clouds into C_2 dimensions, and a maximum pooling is used to obtain the global features. Finally, the data are fed to the following classification module for final activity recognition. The first layer is a batch

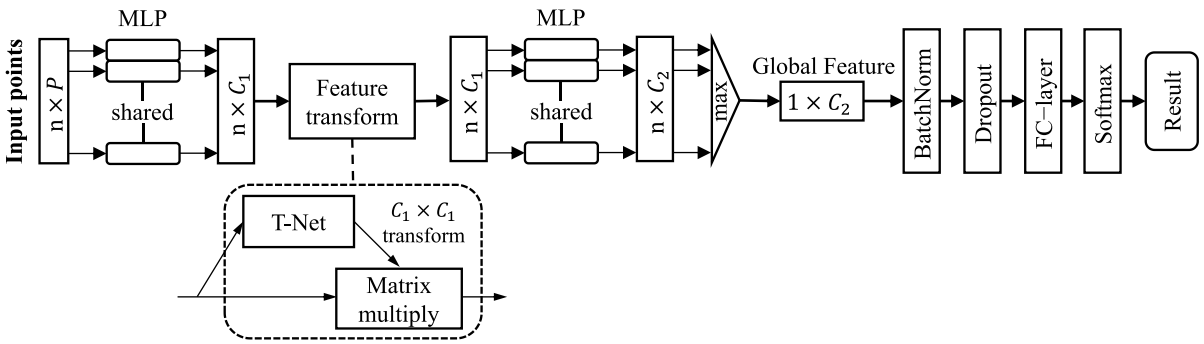

Fig. 5. Architecture of the proposed PointNet network.

TABLE I
KEY PARAMETERS OF THE FMCW RADAR SYSTEM

Center frequency	5.8 GHz
Bandwidth	320 MHz
Sampling frequency	192 KHz
Ramp repetition period	3.3 ms
Range resolution	0.47 m
Maximum detection range	150 m
Unambiguous velocity	3.88 m/s
Transmitted power	8 dBm

TABLE II
TYPICAL HUMAN ACTIVITIES UNDER RESEARCH

Activity	Detailed Description
Fall	Suddenly drop down to the floor by gravity.
Step	Lift the foot and set it down in a new position.
Jump	Spring clear of the ground.
Squat	Sit in a low position with knees bent.
Walk	Proceed through at a moderate pace on foot.
Jog	Run at a slow pace with with fist around chest.

TABLE III
MAIN PHYSICAL PARAMETERS OF VOLUNTEERS

No.	Gender	Age (yr)	Weight (kg)	Height (m)	BMI (kg/m^2)
1	M	23	85	1.75	27.76
2	M	32	78	1.80	24.07
3	M	24	72	1.79	22.47
4	M	25	70	1.77	22.34
5	M	23	68	1.80	20.99
6	F	24	50	1.58	20.03
7	F	23	55	1.62	20.96
8	M	23	72	1.78	22.72
9	F	23	55	1.64	20.45
10	M	23	85	1.91	23.30
11	M	23	54	1.75	17.63
12	F	22	52	1.65	19.10
13	M	23	60	1.72	20.28
14	F	23	52	1.70	17.99
15	M	23	74	1.82	22.34
16	M	24	72	1.70	24.91
Total	M/F(12/6)	23.6±1.4	65.9±11.9	1.74±0.09	21.71±2.61

normalization layer, allowing for much higher learning rates and less dependence on the chosen initialization. The second layer is a drop-out layer and the overfitting problem is suppressed by randomly dropping parameters proportionally. The third layer is fully connected (FC) with the ReLU activation function. The final layer is the softmax layer to achieve the result score. Here, the cross-entropy loss function is used with the Adam optimizer.

III. EXPERIMENTAL SETUP AND DATA

A. System and Data Acquisition

The experimental setup is illustrated in Fig. 6. The system device is a portable FMCW radar whose key parameters are listed in Table I [41]. To verify the versatility and robustness of our method, six typical human daily activities were selected: 1) *fall*; 2) *step*; 3) *jump*; 4) *squat*; 5) *walk*; and 6) *jog*. They are illustrated in Fig. 7 and described in Table II. Sixteen volunteers were enrolled in this study, including 11 males and 5 females. Their main physical parameters are given in Table III. In detail, their ages ranged from 22 to 32 years, and weights ranged from 50 to 85 kg, with the height from 1.58 to 1.91 m. In the data collection, the radar system was set at the height of 0.8 to 1 m. The volunteers were asked to perform the above activities toward the radar at the distance between 2 and 5 m. Each activity was performed 240 times to achieve a balanced data set of 1440 samples. The data set has been shared in IEEE DataPort with the title "Human Activity Data with a 5.8-GHz FMCW Radar"(available at https://dx.doi.org/10.21227/aze5-h339).

B. Implementation Details

In the feature extraction part, STFT-based TD map, and 2-D FFT-based RD frames were utilized as references. Furthermore, a 2D-CFAR detector was applied to the traditional feature maps above to obtain intensity-based point cloud

Fig. 6. Experimental setup with radar system and area where the activities are performed, including a mattress for controlled falls.

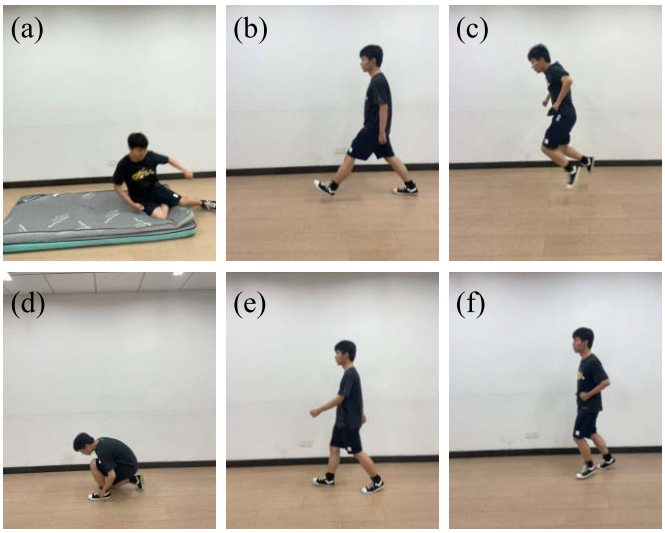

Fig. 7. Illustrations of six typical human activities: (a) *Fall*, (b) *Step*, (c) *Jump*, (d) *Squat*, (e) *Walk*, and (f) *Jog*.

features. In the CFAR-based method, there are very significant differences in the number of points among different maps (i.e., hundreds of points in one map and thousands in another). Therefore, a normalization method, including interpolation and sampling, was applied to align the CFAR-based points to a fixed number, 3000 for a single TD map, and 400 for a single RD frame, respectively.

In the classification part, for the proposed PointNet model, C_1 and C_2 were set to 64 and 1024, respectively. The drop ratio of the drop-out layer was 0.3. Furthermore, we have tried our best to rebuild the CNN [26], 3DCNN [28], and CNN + LSTM [20] networks from the reference studies. A few parameters, such as convolution kernel size, were adjusted to fit our data format. The data set was divided into train and test sets containing 840 and 600 samples.

The feature extraction was implemented with MATLAB R2019b. All the networks were trained based on the Pytorch framework. An NVIDIA GeForce RTX 2060 graphics card (GPU, with a 6-GB memory) and AMD Ryzen 74800H were used on a laptop with 32G of memory.

IV. Results and Discussion

A. Sparsity Versus Information Preservation

The advantage of our proposed sparsity-based feature extraction is to utilize a smaller number of sparse points to represent most key information from received signals. This has been preliminarily confirmed by a qualitative visual comparison between the reconstructed and original feature maps in Section II, Figs. 3 and 4. In this experiment, we analyzed the effect of the sparsity on information preservation first and demonstrated its superiority in a quantitative way.

For TDSP, Fig. 8 shows the results of the example *fall* motion generated by different values of sparsity K. In detail, row (I) shows TDSP, and row (II) shows their corresponding conventional TD maps with the reconstructed signals. Compared with the conventional TD map with STFT in Fig. 3(a), when $K < 100$, noticeable TD information loss can be observed, especially, in the Doppler peak and in the lying down part after the peak. The complete distribution of most TD information appeared to be preserved when $K = 200$. On this basis, when K continued to increase, the added sparse points were all small in size, which indicated low intensity and little contribution to information preservation. As expected, their reconstructed TD maps were almost the same as the one of $K = 200$.

To evaluate the information preservation performance in a quantitative way, a 2-D correlation coefficient between the original TD map from STFT and the reconstructed one from TDSP was calculated as follows:

$$r = \frac{\sum_m \sum_n (P_{mn} - \overline{P})(Q_{mn} - \overline{Q})}{\sqrt{\left(\sum_m \sum_n (P_{mn} - \overline{P})^2\right)\left(\sum_m \sum_n (Q_{mn} - \overline{Q})^2\right)}} \quad (14)$$

where P was the matrix of the TD map from STFT, $\overline{P}$ was its mean value, Q was the matrix of the reconstructed one from TDSP, $\overline{Q}$ was its mean value, m and n represented their row and column indexes. A high value of the correlation coefficient r corresponds to good consistency between the original and reconstructed images, i.e., better information preservation performance. The results of r for the example *fall* motion with different K were illustrated in Fig. 9. The curve of correlation coefficient r grew quickly at first and then slowly after $K = 75$. Accounting also for the results of Fig. 8, after achieving a rate of 0.94 at $K = 200$, the reconstructed noisy parts contributed to the continuous increase of r, hence, the value of K was considered a good compromise.

Furthermore, a similar investigation was conducted in the DRDSP-based method. Taking the 5th frame of the *fall* motion as an example, Fig. 10 shows its DRDSP and reconstructed RD maps with different values of sparsity K. In particular, DRDSP required more sparse points to preserve most key features. This is reasonable, because the TDSP only refers to a single sparse projection, whereas the DRDSP is a set containing multiple Fourier-based sparse projection results among

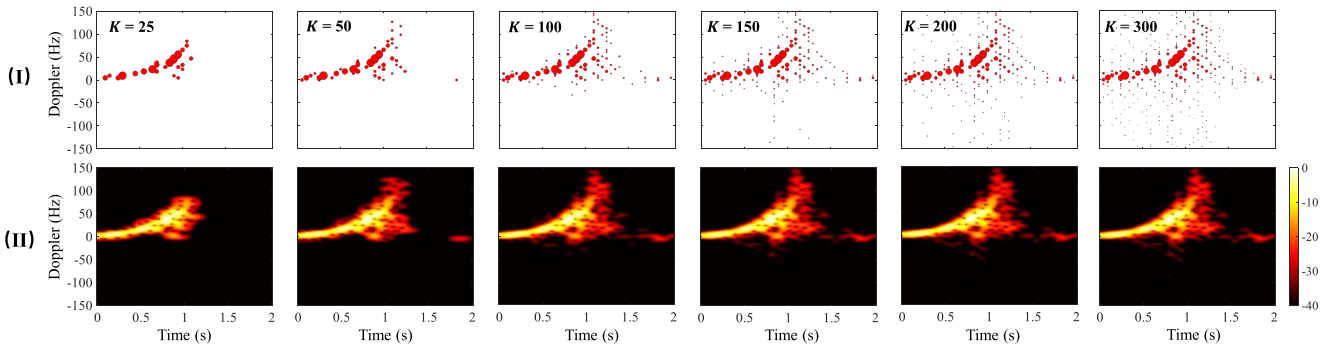

Fig. 8. Results of the example *fall* motion with different values of the sparsity parameter K in TDSP from 25 to 300. (I) TDSP. (II) Corresponding reconstructed TD maps.

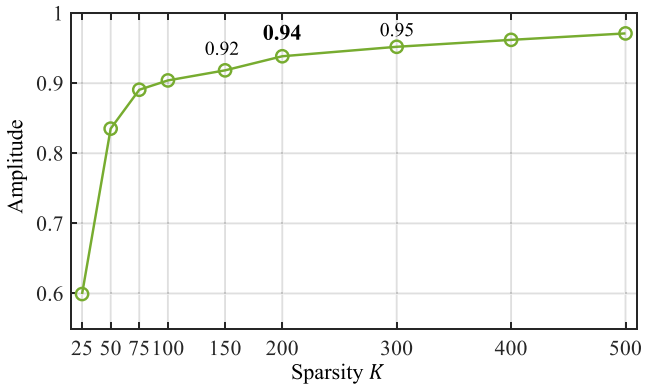

Fig. 9. Correlation coefficient in TDSP versus sparsity parameter K.

all range bins. It can be noticed that when $K \geq 400$, the main parts of the RD distribution have been reconstructed well. The additional points only reflected little information. Fig. 11 illustrates the correlation results with different sparsity K in a line chart. The correlation coefficient r climbed up as the sparsity increased and showed a slowdown in the pace of increase after $K = 400$. Moreover, if we ignored the noisy contributions whose intensity was lower than -25 dB, the correlation coefficients were improved by about 10%, indicated as the dashed green line. Therefore, although noisy contributions reduced the correlation between the reconstructed and the original map, they had little effect on preserving valuable information.

In general, two conclusions about the feature extraction performance of our proposed method can be drawn as follows.

1) A larger sparsity K corresponds to a more significant number of sparse points along with a better information preservation performance.
2) When K exceeds a certain empirical threshold, a continuous increase of sparsity will bring little improvement in the extraction of key features.

B. Sparsity Versus Classification Accuracy and Time Cost

Sparsity K is the key and the only empirical parameter in our proposed feature extraction methods. Its value is directly related to the recognition accuracy and time consumption. In this experiment, the selection of sparsity K was discussed and determined. The accuracy rate was yielded by recognizing six typical human activities based on the proposed PointNet networks, and the time cost indicates the time consumption for the processing to obtain sparse point clouds.

For TDSP, the corresponding results are illustrated in Fig. 12. The red diamonds and blue rectangles indicate accuracy rates and time cost, respectively. The accuracy rate climbed up as the sparsity increased, then remained virtually unchanged around 95% when $K \geq 200$. In addition, an ablation study was carried out to validate the importance of intensity features shown as the red dashed line. Once these features were not taken into account, the time and Doppler coordinates of target points and noise points were treated equally. As the sparsity kept increasing, the number of noise points would also increase and reduce the proportion of target-related ones in model learning. This resulted in a marked drop in the accuracy rate compared with the solid one. On the other hand, the time cost is an exponential curve, which would grow to a nonnegligible degree with the increasing number of sparse points. Therefore, as a tradeoff of accuracy and real-time performance, the sparsity K was set to 200 in the TDSP method corresponding to an average accuracy of 95.0% and an average time cost of 1.14 s.

For DRDSP, a similar investigation was conducted, as shown in Fig. 13. Though DRDSP contains nine frames of sparse points, the time cost in extraction was still short (approximately 0.81 s) and acceptable. Before $K = 400$, the average classification accuracy grew fast along with the increase of sparsity. After that, adding 100 sparse points achieved less than 0.2% improvement. This agrees with the results in Figs. 10 and 11. In particular, when leaving intensity features out, the accuracy drop of DRDSP can be observed but not apparent than the one of TDSP. In general, the sparsity K was set to 400 in the DRDSP method to achieve an average accuracy of 98.0% and an average time cost of 0.81 s.

In general, though the sparsity K is an empirical selection, its accuracy curve is logarithmic. Therefore, one of the most straightforward solutions is to select a value as high as possible while meeting the real-time requirement.

C. Performance Comparison

We evaluated the performance of the proposed methods by comparing them with existing ones. Different comparative

TABLE IV
PERFORMANCES OF THE REFERENCE NETWORKS VERSUS THE PROPOSED APPROACH

Feature	Network/Classifier	Feature size	Extraction time/sample (s)	Training time/epoch (s)	Testing time/sample (ms)	Parameters	Model Size (kB)	Accuracy	Precision	Recall	F1 score
TD map	CNN [26]	128×128	0.10	7.77	3.31	65.7M	256788	92.8%	0.93	0.93	0.93
CFAR-based	3D-PointNet	3000×3	0.58	5.01	0.57	1.61M	6325	92.7%	0.93	0.93	0.93
TDSP	CNN [26]	128×128	1.18	7.77	3.31	65.7M	256788	91.2%	0.92	0.91	0.91
TDSP	**3D-PointNet**	**200×3**	**1.14**	**3.03**	**0.12**	**1.61M**	**6325**	**95.0%**	**0.95**	**0.95**	**0.95**
DRDT	Subspace KNN[21]	9×1×3	0.14	0.28	0.02	NA	1286	91.7%	0.92	0.92	0.92
RD frames	3DCNN [28]	9×64×64	0.08	5.87	2.55	379K	1109	93.5%	0.94	0.94	0.93
RD frames	CNN+LSTM [20]	9×64×64	0.08	8.22	1.03	283K	3768	94.2%	0.94	0.94	0.94
CFAR-based	4D-PointNet	9×400×3	0.75	7.48	2.20	1.61M	6332	96.2%	0.96	0.96	0.96
DRDSP	3DCNN [28]	9×64×64	0.89	5.87	2.55	379K	1109	93.2%	0.93	0.93	0.93
DRDSP	CNN+LSTM [20]	9×64×64	0.89	8.22	1.03	283K	3768	93.6%	0.94	0.94	0.94
DRDSP	**4D-PointNet**	**9×400×3**	**0.81**	**7.48**	**2.20**	**1.61M**	**6332**	**98.0%**	**0.98**	**0.98**	**0.98**

The results of our proposed methods in **bold**.

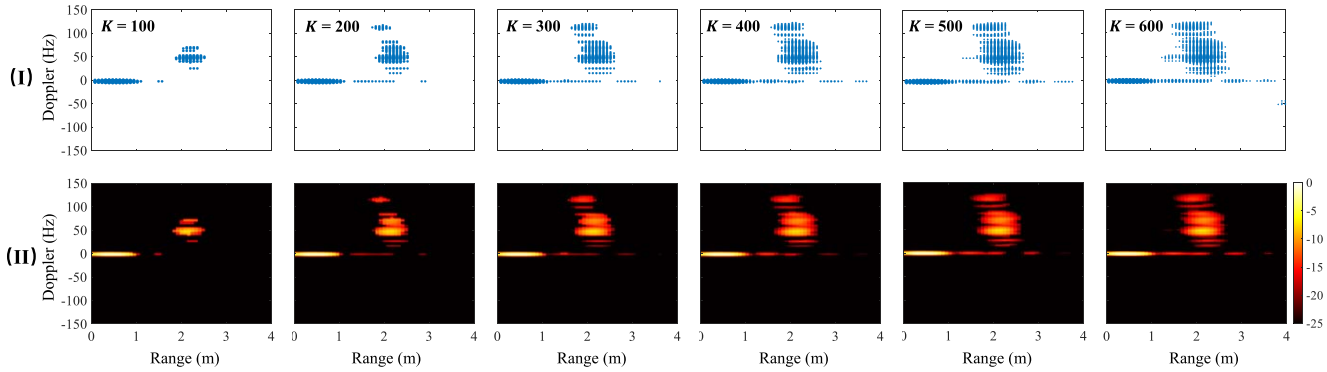

Fig. 10. Results of the example *fall* motion with different values of the sparsity parameter K in DRDSP from 100 to 600. (I) DRDSP. (II) Corresponding reconstructed RD frames.

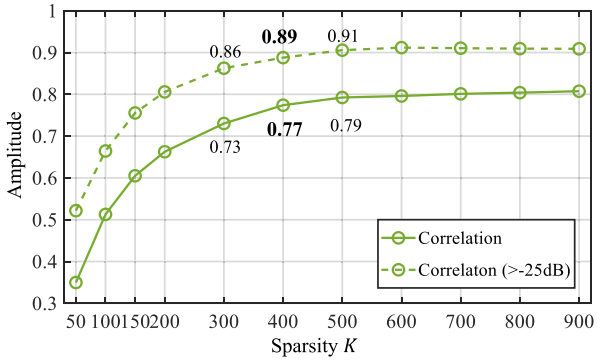

Fig. 11. Correlation coefficient and correlation coefficient (>−25 dB) in DRDSP versus sparsity parameter K.

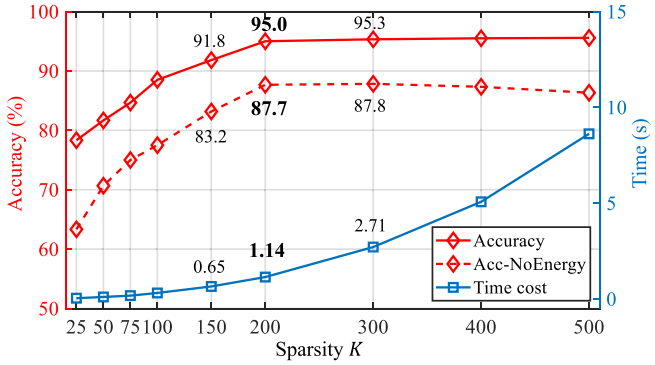

Fig. 12. Average accuracy, average accuracy without energy intensity features (Acc-NoEnergy), and time cost in TDSP versus sparsity parameter K.

performances of each method were computed and listed in Table IV. In detail, "Feature size" is the size of the input used for classification. "Extraction time/sample," "Training time/epoch," and "Testing time/sample" are the average time cost spent for feature extraction, model training, and model testing, respectively. "Parameters" indicates the network computational complexity, and "Model Size" reflects the storage requirement. In addition, "Accuracy," "Precision," "Recall," and "F1 score" are calculated to evaluate the classification performance in more details in term of true positives (TPs), true negatives (TNs), false positive (FP), and false negatives (FNs)

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FN}+\text{FP}+\text{TN}} \tag{15}$$

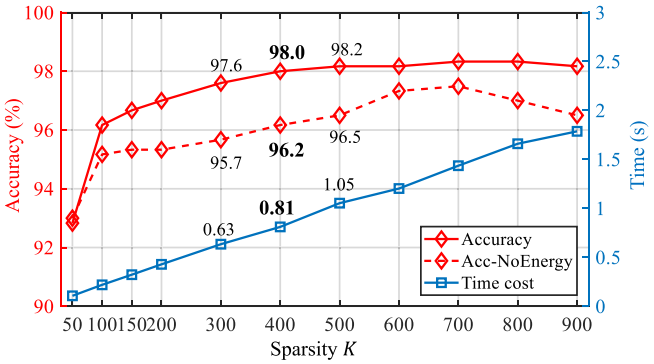

Fig. 13. Average accuracy, average accuracy without energy intensity features (Acc-NoEnergy), and time cost in DRDSP versus sparsity parameter K.

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{16}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{17}$$

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{18}$$

In the TD domain, the TD map with CNN [26] was regarded as the baseline approach. This required the shortest time duration of 0.1 s to extract the TD map by conventional STFT, but this map needed to be scaled into a 128×128 image as input to the CNN network, which resulted in a time cost of 7.77 s in the model training with the recognition accuracy rate of 92.8%. The parameters of the CNN network are 65.7M with a largest model size of 256788 kB. The CFAR-based method utilized 3000 feature points as input to the 3D-PointNet. It required 0.58, 5.01, and 0.57 s in feature extraction, model training, and model testing, respectively. CFAR + 3D-PointNet did not improve the average accuracy for its information loss. In contrast, with the same network, the proposed TDSP only needed 200 sparse points to achieve the highest average recognition accuracy rate of 95.0%, with less time cost of 3.02 and 0.12 s in model training and testing. This result demonstrates the representation efficiency of our proposed method and is consistent with the visual comparison in Fig. 3. On the other hand, we tried the image-based approach by transforming TDSP features into a 128×128 image as input into the CNN network. An increase in required time and a decrease in recognition performance was found. This result validated the effectiveness of PointNet architectures in processing point cloud features. The confusion matrix of the proposed combination of TDSP and 3D-PointNet is shown in Table V. Note that, *walk* motion got the highest recognition accuracy, and there were none samples mislabeled as others. *Fall* motion achieved a 93% recognition accuracy and was easy to be confused with *squat* motion. In addition, *step* motion owned the lowest accuracy rate of 86%, where 2% and 12% of them were misclassified as *fall* and *squat* motion, respectively. This is reasonable that stepping forward with a fast speed and high amplitude is similar to a slow *fall*, while a slow *step* motion acts like a *squat*.

In the RD domain, the DRDT method from our previous study [21] extracted one trajectory point from each RD frame.

TABLE V
CONFUSION MATRIX WITH TDSP

Act. \ Pred.	*Fall*	*Step*	*Jump*	*Squat*	*Walk*	*Jog*
Fall	93%	0	0	7%	0	0
Step	2%	86%	0	12%	0	0
Jump	1%	1%	98%	0	0	0
Squat	4	0	0	96%	0	0
Walk	0	0	0	0	100%	0
Jog	0	0	0	0	3%	97%

TABLE VI
CONFUSION MATRIX WITH DRDSP

Act. \ Pred.	*Fall*	*Step*	*Jump*	*Squat*	*Walk*	*Jog*
Fall	99%	0	0	0	1%	0
Step	0	97%	0	3%	0	0
Jump	0	0	99%	0	1%	0
Squat	3%	3%	0	94%	0	0
Walk	0	0	0	0	99%	1%
Jog	0	0	0	0	0	100%

Therefore, it had the smallest size of input features as $9 \times 1 \times 3$ but only obtained 91.7% recognition accuracy with the subspace K-nearest neighbor (KNN) classifier. Compared with the accuracy rate in our previous work, the performance degradation was caused by the expanded data set, which brought a great challenge to adjust handcraft feature extraction to all records. In addition, each RD frame was scaled to a 64×64 image and fed into 3DCNN [28] and CNN + LSTM [20] networks. They obtained 93.5% and 94.2% classification accuracy, respectively. The parameters and model size of 3DCNN are 379K and 1109 kB, while the ones of CNN + LSTM are 283K and 3768 kB, respectively.

For a better comparison, 400 normalized CFAR-based feature points were selected from each frame as input to the 4D-PointNet network. After fine-tuning, this approach reached 96.2% average accuracy. In particular, the parameters and model size of the 4D-PointNet are 1.61M and 6332 kB. The proposed DRDSP features can obtain the highest recognition accuracy rate of 98.0% with the same network. Similarly, DRDSP features were also processed as images to the CNN-based networks. As expected, a reduction in the accuracy can be observed. The confusion matrix of the proposed combination of DRDSP and 4D-PointNet is shown in Table VI. All six activities had a high recognition accuracy rate of more than 94%. In particular, *fall*, *step*, and *squat* motions are easily confused with each other as they have a similar tendency in time, range, Doppler, and intensity.

In general, compared with traditional image-based features, sparse point clouds can utilize a much smaller feature size but longer time cost for a better human activity representation. On the other hand, PointNet architectures demonstrated their superiority in processing point cloud features.

TABLE VII
ROBUSTNESS PERFORMANCE IN INDIVIDUAL DIVERSITY STUDY

Feature	Network	1	2	3	4	5	6	7	8	9
TD map	CNN [26]	85.0%	90.8%	96.7%	91.7%	97.5%	95.0%	86.7%	99.2%	94.2%
CFAR	3D-PointNet	91.7%	95.0%	93.3%	91.7%	89.2%	90.0%	94.2%	91.7%	84.2%
TDSP	CNN [26]	77.5%	85.8%	90.8%	91.7%	95.0%	96.7%	86.7%	86.7%	82.5%
TDSP	**3D-PointNet**	**96.7%**	**96.7%**	**91.7%**	**91.7%**	**99.2%**	**88.3%**	**92.5%**	**89.2%**	**88.3%**
RD trajectory	DRDT [21]	82.9%	94.9%	89.4%	80.0%	80.0%	81.2%	90.8%	85.8%	80.5%
RD frames	3DCNN [28]	90.8%	95.8%	80.0%	95.8%	97.5%	96.7%	92.5%	90.0%	95.0%
RD frames	CNN+LSTM [20]	91.7%	99.2%	90.0%	91.7%	99.2%	96.7%	90.0%	96.7%	85.8%
CFAR	4D-PointNet	95.8%	100.0%	94.2%	94.2%	92.5%	95.0%	90.0%	93.3%	90.0%
DRDSP	3DCNN [28]	88.3%	91.7%	87.5%	88.3%	97.5%	91.7%	95.0%	93.3%	84.2%
DRDSP	CNN+LSTM [20]	95.0%	90.0%	86.7%	91.7%	89.2%	85.0%	86.7%	94.2%	83.3%
DRDSP	**4D-PointNet**	**98.3%**	**98.3%**	**94.2%**	**98.3%**	**99.2%**	**100%**	**96.7%**	**95.0%**	**95.0%**

Feature	Network	10	11	12	13	14	15	16	Average
TD map	CNN [26]	92.5%	98.3%	84.2%	90.0%	70.0%	73.3%	73.3%	88.65±9.32%
CFAR	3D-PointNet	90.0%	93.3%	95.8%	73.3%	73.3%	63.3%	80.0%	86.87±9.49%
TDSP	CNN [26]	82.5%	95.0%	87.5%	66.7%	76.7%	63.3%	73.3%	83.65±9.93%
TDSP	**3D-PointNet**	**85.8%**	**96.7%**	**95.8%**	**80.0%**	**80.0%**	**86.7%**	**90.0%**	**90.58±5.74%**
RD trajectory	DRDT [21]	82.1%	82.5%	72.5%	60.0%	69.1%	86.7%	89.1%	81.72±8.71%
RD frames	3DCNN [28]	95.0%	95.0%	94.2%	70.0%	83.3%	73.3%	96.7%	90.10±6.76%
RD frames	CNN+LSTM [20]	83.3%	100%	93.3%	70.0%	83.3%	83.3%	96.7%	90.68±8.01%
CFAR	4D-PointNet	89.2%	95.0%	91.7%	83.3%	80.0%	70.0%	100%	90.88±7.62%
DRDSP	3DCNN [28]	91.7%	91.7%	88.3%	70.0%	90.0%	86.7%	100.0%	89.74±6.64%
DRDSP	CNN+LSTM [20]	95.8%	90.0%	92.5%	70.0%	90.0%	80.0%	100.0%	88.75±7.08%
DRDSP	**4D-PointNet**	**90.8%**	**98.3%**	**98.3%**	**80.0%**	**80.0%**	**90.0%**	**100%**	**94.53±6.40%**

The results of our proposed method in **bold**.

D. Individual Effect via Leave-One-Subject-Out Test

Recognizing human activities of unknown persons using well-trained models is essential for practical applications. A leave-one-subject-out test was applied to investigate the robustness of the proposed method when facing individual diversity. Samples from 15 individuals were selected to learn the classification models. Then, the well-trained model was tested by the left-out person. The results of 16 individuals based on different recognition methods are shown in Table VII. Notice that there was a marked drop with a relatively high standard deviation in DRDT and CFAR-based methods. These handcrafted methods were easily influenced by various activity styles and it is challenging to define robust empirical thresholds suitable for all individuals. The TDSP and DRDSP with the PointNet architectures performed well in the robustness test. The TDSP with 3D-PointNet obtained the lowest standard deviation of 5.74%, and the DRDSP with the 4D-PointNet achieved the highest average accuracy rate of 94.53%. In particular, there was no apparent weakness for them among all 16 tests, as only two tests were below 85%. These results demonstrated the robustness of the proposed approach when facing various individuals.

E. Generalization Performance

To investigate the generality of our proposed method, the public data set [42] (https://researchdata.gla.ac.uk/848/) collected at the University of Glasgow (UOG) at six different locations was used. The radar system is an off-the-shelf FMCW radar (by Ancortek) operating at C-band (5.8 GHz) with bandwidth 400 MHz and chirp duration 1 ms, delivering an output power of approximately +18 dBm. The radar is connected to transmitting and receiving Yagi antennas with a gain of about +17 dB and is capable of recording micro-Doppler signatures of the people moving within the area of interest. 106 volunteers of various age groups performed six activities, including: 1) walking; 2) sitting; 3) standing up; 4) drinking water; 5) picking an object from the floor; and 6) falling. A total of 1754 collected samples was used.

Table VIII shows the results of our proposed method and other referenced approaches for the public UOG data set. The TDSP + 3D-PointNet ($K = 200$) and DRDSP + 4D-PointNet ($K = 400$) both achieved state-of-the-art accuracy rates of 92.16% and 95.69% in the term of tenfold cross validation, respectively. This demonstrates the generalization performance of our proposed methods.

TABLE VIII
COMPARISON WITH STATE-OF-THE-ART APPROACHES USING THE PUBLIC UOG DATA SET OF RADAR SIGNATURES [42]

Methods	Samples	Cross Validation	Average Accuracy
WRGAN-GP+DCNN [43]	750	4-fold	92.30%
Customized feature+Hierarchical structure [44]	1080	5-fold	95.40%
Improved PCA+Improved VGG16 [45]	1633	5-fold	96.34%
CWT+RD-CNN [46]	1282	5-fold	95.71%
CA-CFAR+PointNet [47]	1754	10-fold	88.00%
TDSP+3D-PointNet ($K = 200$)	**1754**	**10-fold**	**92.16%**
DRDSP+4D-PointNet ($K = 400$)	**1754**	**10-fold**	**95.69%**

The results of our proposed methods in **bold**.

V. CONCLUSION

This article proposed a novel HAR method, combining sparse theory and PointNet networks. The sparse theory can utilize a limited number of sparse solutions to characterize human activity in the form of TDSP or DRDSP in the TD and RD domains, respectively. Compared with the image-based features, these sparse point clouds had the advantages of clear physical meanings, less redundant noisy contributions, and little information loss. Moreover, the PointNet networks were applied to support multidomain sparse point clouds as direct input for classification. This overcomes the constraint of image-based networks, which require converting point clouds into pictures. Experimental data involving six typical daily human activities and sixteen volunteers were recorded. The feasibility and superiority of the sparsity-based feature extraction method have been demonstrated by comparing the reconstructed and original feature maps in both qualitative visual and quantitative statistical ways. As the key parameter of the sparse representation algorithm, the selection of sparsity was also investigated and discussed to obtain a practical and straightforward solution. In particular, TDSP showed a higher sparse representation efficiency than DRDSP but cost more time. Furthermore, compared with the image-based networks, the proposed PointNet architecture showed a better recognition performance due to its high robustness to small perturbation and corruption of input points caused by noises. Its combination with TDSP and DRDSP achieved the state-of-art recognition accuracy of 95.0% and 98.0% in the TD and RD domains, respectively. Furthermore, a leave-one-subject-out study was applied to validate its robustness when facing individual diversity effects. Finally, a public data set was utilized to demonstrate the generality of our proposed methods by achieving the state-of-the-art recognition performance.

Future research will explore adding range information in TDSP or finding a more effective sparse method in the RD domain. The adaptive selection of sparsity K is also interesting for further investigation. Moreover, a multi-input multi-output (MIMO) system is a worthwhile choice for HAR, as it can supply angle information, which is crucial for a more complex situation.

REFERENCES

[1] M. G. Amin, Y. D. Zhang, F. Ahmad, and K. D. Ho, "Radar signal processing for elderly fall detection: The future for in-home monitoring," *IEEE Signal Process. Mag.*, vol. 33, no. 2, pp. 71–80, Mar. 2016.

[2] J. Le Kernec et al., "Radar signal processing for sensing in assisted living: The challenges associated with real-time implementation of emerging algorithms," *IEEE Signal Process. Mag.*, vol. 36, no. 4, pp. 29–41, Jul. 2019.

[3] R. Zhang, X. Jing, S. Wu, C. Jiang, J. Mu, and F. R. Yu, "Device-free wireless sensing for human detection: The deep learning perspective," *IEEE Internet Things J.*, vol. 8, no. 4, pp. 2517–2539, Feb. 2021.

[4] V. Bianchi, M. Bassoli, G. Lombardo, P. Fornacciari, M. Mordonini, and I. De Munari, "IoT wearable sensor and deep learning: An integrated approach for personalized human activity recognition in a smart home environment," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8553–8562, Oct. 2019.

[5] X. Zhou, W. Liang, K. I.-K. Wang, H. Wang, L. T. Yang, and Q. Jin, "Deep learning enhanced human activity recognition for Internet of Healthcare Things," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6429–6438, Jul. 2020.

[6] A. I. Cuesta-Vargas, A. Galan-Mercant, and J. M. Williams, "The use of inertial sensors system for human motion analysis," *Phys. Ther. Rev.*, vol. 15, no. 6, pp. 462–473, 2010.

[7] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 3, pp. 1192–1209, 3rd Quart., 2012.

[8] L. Ren et al., "Short-time state-space method for micro-Doppler identification of walking subject using UWB impulse Doppler radar," *IEEE Trans. Microw. Theory Techn.*, vol. 66, no. 7, pp. 3521–3534, Jul. 2018.

[9] C. Li et al., "A review on recent progress of portable short-range noncontact microwave radar systems," *IEEE Trans. Microw. Theory Techn.*, vol. 65, no. 5, pp. 1692–1706, May 2017.

[10] E. Cippitelli, F. Fioranelli, E. Gambi, and S. Spinsante, "Radar and RGB-depth sensors for fall detection: A review," *IEEE Sensors J.*, vol. 17, no. 12, pp. 3585–3604, Jun. 2017.

[11] J. A. Ward, P. Lukowicz, G. Troster, and T. E. Starner, "Activity recognition of assembly task using body-worn microphones and accelerometers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1553–1567, Oct. 2006.

[12] M. G. Amin, *Through-the-Wall Radar Imaging*. Boca Raton, FL, USA: CRC Press, 2017.

[13] X. Qiao, G. Li, T. Shan, and R. Tao, "Human activity classification based on moving orientation determining using multistatic micro-Doppler radar signals," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, Aug. 2022.

[14] B. Wang, L. Guo, H. Zhang, and Y.-X. Guo, "A millimetre-waveradar-based fall detection method using line kernel convolutional neural network," *IEEE Sensors J.*, vol. 20, no. 22, pp. 13364–13370, Nov. 2020.

[15] Y. Kim and H. Ling, "Human activity classification based on micro-Doppler signatures using a support vector machine," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 5, pp. 1328–1337, May 2009.

[16] F. Fioranelli, M. Ritchie, and H. Griffiths, "Classification of unarmed/armed personnel using the NetRAD multistatic radar for micro-Doppler and singular value decomposition features," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 9, pp. 1933–1937, Sep. 2015.

[17] B. Erol and M. G. Amin, "Radar data cube processing for human activity recognition using multisubspace learning," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 55, no. 6, pp. 3617–3628, Dec. 2019.

[18] M. Li, T. Chen, and H. Du, "Human behavior recognition using range-velocity-time points," *IEEE Access*, vol. 8, pp. 37914–37925, 2020.

[19] 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.

[20] Y. Kim, I. Alnujaim, and D. Oh, "Human activity classification based on point clouds measured by millimeter wave MIMO radar with deep recurrent neural networks," *IEEE Sensors J.*, vol. 21, no. 12, pp. 13522–13529, Jun. 2021.

[21] C. Ding et al., "Continuous human motion recognition with a dynamic range-Doppler trajectory method based on FMCW radar," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6821–6831, Sep. 2019.

[22] S. Z. Gurbuz and M. G. Amin, "Radar-based human-motion recognition with deep learning: Promising applications for indoor monitoring," *IEEE Signal Process. Mag.*, vol. 36, no. 4, pp. 16–28, Jul. 2019.

[23] A. Li, E. Bodanese, S. Poslad, T. Hou, K. Wu, and F. Luo, "A trajectory-based gesture recognition in smart homes based on the ultra-wideband communication system," *IEEE Internet Things J.*, vol. 9, no. 22, pp. 22861–22873, Nov. 2022.

[24] J. Zhu, X. Lou, and W. Ye, "Lightweight deep learning model in mobile-edge computing for radar-based human activity recognition," *IEEE Internet Things J.*, vol. 8, no. 15, pp. 12350–12359, Aug. 2021.

[25] B. Jokanović and M. Amin, "Fall detection using deep learning in range-Doppler radars," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 54, no. 1, pp. 180–189, Feb. 2018.

[26] H. Sadreazami, B. Miodrag, and R. Sreeraman, "Contactless fall detection using time-frequency analysis and convolutional neural networks," *IEEE Trans. Ind. Informat.*, vol. 17, no. 10, pp. 6842–6851, Oct. 2021.

[27] X. Bai, Y. Hui, L. Wang, and F. Zhou, "Radar-based human gait recognition using dual-channel deep convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9767–9778, Dec. 2019.

[28] W. Li et al., "Real-time fall detection using mmWave radar," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2022, pp. 16–20.

[29] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2013, pp. 6645–6649.

[30] A. Gorji, A. Bourdoux, S. Pollin, and H. Sahli, "Multi-view CNN-LSTM architecture for radar-based human activity recognition," *IEEE Access*, vol. 10, pp. 24509–24519, 2022.

[31] M. Wang, G. Cui, X. Yang, and L. Kong, "Human body and limb motion recognition via stacked gated recurrent units network," *IET Radar Sonar Navig.*, vol. 12, no. 9, pp. 1046–1051, 2018.

[32] A. Shrestha, H. Li, J. Le Kernec, and F. Fioranelli, "Continuous human activity classification from FMCW radar with bi-LSTM networks," *IEEE Sensors J.*, vol. 20, no. 22, pp. 13607–13619, Nov. 2020.

[33] J. Maitre, K. Bouchard, and S. Gaboury, "Fall detection with UWB radars and CNN-LSTM architecture," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 4, pp. 1273–1283, Apr. 2021.

[34] G. Li, R. Zhang, M. Ritchie, and H. Griffiths, "Sparsity-driven micro-Doppler feature extraction for dynamic hand gesture recognition," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 54, no. 2, pp. 655–665, Apr. 2018.

[35] C. Ding, J. Yan, H. Hong, and X. Zhu, "Sparsity-based feature extraction in fall detection with a portable FMCW radar," in *Proc. IEEE Int. Workshop Electromagn. Appl. Stud. Innov. Compet. (iWEM)*, 2021, pp. 1–3.

[36] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 652–660.

[37] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[38] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[39] E. Conte, A. De Maio, and C. Galdi, "CFAR detection of multidimensional signals: An invariant approach," *IEEE Trans. Signal Process.*, vol. 51, no. 1, pp. 142–151, Jan. 2003.

[40] Y. Cheng and Y. Liu, "Person reidentification based on automotive radar point clouds," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, May 2021.

[41] G. Wang, C. Gu, T. Inoue, and C. Li, "A hybrid FMCW-interferometry radar for indoor precise positioning and versatile life activity monitoring," *IEEE Trans. Microw. Theory Techn.*, vol. 62, no. 11, pp. 2812–2822, Nov. 2014.

[42] F. Fioranelli, S. A. Shah, H. Li, A. Shrestha, S. Yang, and J. Le Kernec, "Radar signatures of human activities [data collection]." Accessed: Jul. 5, 2020. [Online]. Available: http://researchdata.gla.ac.uk/848/

[43] L. Qu, Y. Wang, T. Yang, and Y. Sun, "Human activity recognition based on WRGAN-GP-synthesized micro-Doppler spectrograms," *IEEE Sensors J.*, vol. 22, no. 9, pp. 8960–8973, May 2022.

[44] X. Li, Z. Li, F. Fioranelli, S. Yang, O. Romain, and J. Le Kernec, "Hierarchical radar data analysis for activity and personnel recognition," *Remote Sens.*, vol. 12, no. 14, p. 2237, 2020.

[45] Y. Zhao, H. Zhou, S. Lu, Y. Liu, X. An, and Q. Liu, "Human activity recognition based on non-contact radar data and improved PCA method," *Appl. Sci.*, vol. 12, no. 14, p. 7124, 2022.

[46] W. Y. Kim and D. H. Seo, "Radar-based human activity recognition combining range–time–Doppler maps and range-distributed-convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, Mar. 2022.

[47] F. Fioranelli, S. Zhu, and I. Roldan, "Benchmarking classification algorithms for radar-based human activity recognition," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 37, no. 12, pp. 37–40, Dec. 2022.

Chuanwei Ding (Student Member, IEEE) received the Ph.D. degree in electronic and information engineering from Nanjing University of Science and Technology, Nanjing, China, in 2020.

He is currently an Associate Research Fellow with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, where he was a Postdoctoral Fellow from 2020 to 2023. In 2018, he was a Visiting Scholar with the Department of Electrical and Computer Engineering, Texas Tech University, Lubbock, TX, USA. His current research interests include biomedical applications of microwave technology and radar signal processing.

Li Zhang (Student Member, IEEE) received the B.S. degree from Nanjing University of Science and Technology, Nanjing, China, in 2015, where he is currently pursuing the Ph.D. degree.

In 2019, he was a Visiting Scholar with the Department of Electrical and Computer Engineering, Texas Tech University, Lubbock, TX, USA. His current research interests include wireless sensing, signal processing, and deep learning.

Haoyu Chen received the B.S. degree in communication engineering from Nanjing University of Science and Technology, Nanjing, China, in 2020, where he is currently pursuing the master's degree with the School of Electronic and Optical Engineering.

His research interests include radar signal processing and biomedical applications of microwave technology.

Hong Hong (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Nanjing University, Nanjing, China, in 2010.

He is currently a Professor with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing. In 2014, he was a Visiting Scholar with the Institute of Biomedical Engineering and Technology, The Sydney University, Sydney, NSW Australia. In 2019, he was a Visiting Professor with the Department of Electrical and Computer Engineering, University of California at Davis, Davis, CA, USA. His current research interests include biomedical applications of microwave technology, audio signal processing, and radar signal processing.

Xiaohua Zhu (Member, IEEE) received the Ph.D. degree in communication and information system from Nanjing University of Science and Technology, Nanjing, China, in 2002.

He is currently a Professor with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, where he is the Director of Radar and High-Speed Digital Signal Processing Laboratory. He has authored and coauthored four books, and more than 100 papers. He has submitted 18 patent applications. His current research interests include radar system, radar signal theory, and digital signal processing.

Dr. Zhu was a ten-time recipient of the Ministerial and Provincial-Level Science and Technology Award.

Francesco Fioranelli (Senior Member, IEEE) received the Laurea (B.Eng., *cum laude*) and Laurea Specialistica (M.Eng., *cum laude*) degrees in telecommunication engineering from the Università Politecnica delle Marche, Ancona, Italy, in 2007 and 2010, respectively, and the Ph.D. degree from Durham University, Durham, U.K., in 2014.

He is currently a Tenured Assistant Professor with TU Delft, Delft, The Netherlands, and was an Assistant Professor with the University of Glasgow, Glasgow, U.K., from 2016 to 2019, and a Research Associate with the University College London, London, U.K., from 2014 to 2016. He has authored over 135 publications between book chapters, journal and conference papers, edited *Micro-Doppler Radar and Its Applications* and *Radar Countermeasures for Unmanned Aerial Vehicles* (IET-Scitech, 2020). His research interests include the development of radar systems and automatic classification for human signatures analysis in healthcare and security, drones and UAVs detection and classification, automotive radar, wind farm, and sea clutter.

Dr. Fioranelli received three Best Paper Awards.