Human Behavior Recognition Using Range-Velocity-Time Points

MENG LI, TAO CHEN, AND HAO DU

ABSTRACT Radar-based sensors do not require optimal lighting and atmospheric conditions and nonocclusion, making them a promising solution for human behavior analysis in complex environments. Existing radar-based models generally retrieve features from either the time-velocity domain or the time-range domain. Such two-dimensional representations cannot fully depict dynamic human motion features. In this paper, we propose a temporal range-Doppler PointNet-based method to analyze human behavior. We transform human echoes to 3D point sets and then feed them into the hierarchical PointNet model for classification. The proposed point network can learn structural features from the micromotion trajectory more effectively than directly processing the raw point cloud. To further improve our model’s robustness in practical applications, we design an outlier detection module for detecting anomalies such as in multitarget scenarios. The results of experiments on motion capture databases and range-Doppler radar measurements demonstrate that our method realizes outstanding performance in terms of the classification accuracy, noise robustness, and anomaly detection accuracy.

INDEX TERMS Human activity recognition, micro-Doppler effect, deep learning, point set, range-Doppler processing, ultra-wideband radar.

I. INTRODUCTION

The recognition of human activities plays an important role in people’s daily life, for instance, in medical, security, and law enforcement applications. Since deep learning techniques enable human behavior recognition systems to automatically extract and learn hierarchical features, many of these systems have been designed and achieved promising results. These end-to-end frameworks have been successfully applied in home behavior classification [1], gait analysis [2], and violence detection [3]. Despite their success, existing activity recognition systems, which mainly rely on visible light or structured light, often suffer from occlusion by opaque obstacles (such as doors and walls). Although several methods aim to capture human motion in low light [4] or partial occlusion conditions [5], these camera-based models fail completely when a person is fully occluded. However, radar-based models can address such issues since an electromagnetic wave with a frequency range of a few gigahertz can easily penetrate occlusions [6], [7]. In particular, ultrawideband radar can obtain detailed information instead of abstracting the human body as a single point. Moreover, Chen, et al. [8] introduced the micro-Doppler effect to the radar community, which has enabled researchers to analyze human behavior depending on micro-motions induced by different body segments. This effect is widely used in cost-effective radar systems whose antenna size is too limited to spatially resolve a human skeleton [9].

However, micro-Doppler-based methods only extract the information from the time-velocity domain, which will cause the loss of a large amount of information contained in the range profile. In fact, the micro-Doppler signature for certain types of activities (for instance, falling and sitting) may be insignificant while the corresponding range extent varies significantly [10]. Therefore, Jokanović et al. [11] chose to further combine range and Doppler features to improve activity recognition accuracy. However, this combination method still analyzed the human behavior in separable 2D (two-dimensional) domains and then combined disjoint features in the classification procedure. Abdulatif et al. [12] exploited micromotion in the joint range-Doppler map, but the lack
of time scale made it difficult to utilize the temporal information. Recently, there have been some attempts [13], [14] to sense human motions by extracting point clouds in the range-velocity-time space, but the inherent redundancy of these point sets hindered its performance. Du et al. [15], [16] developed a 3D deep learning framework for human activity classification. However, their work cannot capture local structures of the micromotion signature in a hierarchical way and has not been conducted in an outlier analysis.

In this work, we present a temporal range-Doppler PointNet-based approach to track and recognize human behaviors. Our work builds upon the existing body of knowledge in the radar domain, aiming to comprehensively screen the range, velocity, and time signatures of radar echoes. As illustrated in Fig. 1, human backscattering echoes are first transformed into a set of range-velocity-time points that represent the visible surface of the micromotion trajectory. Then, the hierarchical PointNet module takes the coordinates of the point sets as input and predicts their corresponding labels. In fact, the designed hierarchical module not only learns the dynamic human motion profiles in an efficient way but also improves its performance in detecting out-of-distribution echoes. The results of experiments on motion capture databases and range-Doppler radar measurements demonstrate that our model achieves significant improvement compared to all the baseline methods that extract information in either image form or raw point clouds.

The key contributions are summarized as follows:

- We propose a 3D deep learning approach, the temporal range-Doppler PointNet, for tracking and estimating the micromotion of body segments over time. This method yields a higher classification accuracy, generalization capability, and noise robustness than existing methods.
- To increase the robustness of our method for practical applications, we add 3D adversarial perturbations to points to enlarge the margin between in-distribution and out-of-distribution echoes. Thus, the proposed model can detect anomalies (such as overlapped echoes) among test samples.
- The proposed pipeline is useful for fully occluded conditions, providing a complementary input when visible light does not work. This network can also be extended to other tasks using range-Doppler radar systems.

II. RELATED WORK

There are primarily two types of human activity recognition research: video-based models and sensor-based models [17]. Video-based systems extract high-dimensional features from videos or images [18], while sensor-based systems rely on the motion data recorded by smart sensors [19]. Considering that microwaves in a frequency range of a few GHz can penetrate walls and opaque objects, a radar-based sensor is an appropriate alternative to analyzing human behavior when visible light is blocked.

With the development of sensing technology, researchers have focused on analyzing human behavior via radar backscattering echoes. The primary idea behind these works is to adapt the inputs to form a virtual image [20]–[22]. With especially designed 2D or even 3D antenna arrays, these researchers project the human target onto the corresponding range plane and process the figure similar to the way of a visual system would. However, these antenna arrays, whose size should be comparable to the wavelength (according to diffraction theory, the longer the wavelength is, the better the penetration [23]), would make the whole radar system inevitably nonportable. Hence, some researchers analyze human behavior with very limited antennae size (generally one transmitting and one receiving antenna [24]). Although such a configuration can only detect the relative radial position of the target, a much shorter transmission interval can be obtained due to the simplified transmitting/receiving module. Thus, rather than detecting the keypoints in the captured figure from a single radar pulse, researchers tend to collect multiple reflected pulses and exploit the temporal variations [25]. Our work continues along this line of work, analyzing human behavior without recovering the skeletal structure.

It is known that when reflected from the moving target, the carrier frequency of the radar signal will be shifted. This phenomenon is called the Doppler effect [8]. The micro-Doppler effect refers to an additional frequency modulation induced by the nonuniform motions of the target. This effect enables us to depict motion profiles in more detail. For a human target, the micro-Doppler signatures, which are produced by the vibration and rotation of limbs, provide complementary information to identify various human activities [9]. Ram and Ling [26] measured the Doppler signature of a human gait and demonstrated the difference between the human and the quadruped motion of animals, but
they did not provide methods to quantify such differences. Kim and Ling [27] extracted six features from the micro-Doppler spectrogram and employed a support vector machine (SVM) to classify different human activities. However, the features extracted from the micro-Doppler spectrogram are solely statistical information, for example, the mean, period, variance and amplitude. These hand-crafted features might be limited to specific tasks and are hard to apply to other research topics. Conventional methods such as independent component analysis [28], empirical mode decomposition [29], and mutual cross entropy [30] are also hindered by the same problem.

Recently, inspired by the successful application of deep learning in various fields, different types of deep neural networks (convolutional neural networks [31], [32] and recurrent neural networks [33]) have been employed to improve micro-Doppler spectrogram based human activity recognition. In fact, most existing methods focus on the joint time-velocity maps and lose information contained in the high-range-resolution profiles. Hence, Jokanović and Amin [11] proposed the sparse autoencoder to fuse information from both the time-velocity and time-range domains to make classification. However, their method is in fact a two-dimensional architecture: the whole micromotion signatures are projected onto two parallel 2D domains and are then processed by separate autoencoders. He et al. [34] presented the range-velocity-time information in a comprehensive way, which is actually a visualization tool that does not perform quantitative analyses.

Different from these methods, our model employs a generalized point cloud model to simultaneously represent the range-velocity-time signature. Motivated by the idea of MoSculp [35], we trace the motion signatures in each joint range-velocity map along the time axis and construct 3D trace sculptures and sample points from its surface. In this way, both the micro-Doppler spectrogram and range profile of human backscattering echoes can be seen as rendered views of our model.

Our work is also inspired by geometry deep learning. Qi et al [36], [37] proposed PointNet to directly treat point clouds as network inputs. In addition, deep kd-tree [38], PointCNN [39], and So-net [40] are recently proposed network architectures that directly process point clouds. These deep networks have shown promising performances on object classification and scene understanding, but none have been applied to radar signal processing. Although there are studies on extracting points from RGB-D [41] or LiDAR scanners [42], constructing and analyzing points from a range-Doppler radar system are different concepts. Instead of directly processing the input points, we add an intensity-based reconstruction module before the hierarchical PointNet model, which incorporates the radar domain knowledge into the 3D network.

Moreover, deep neural networks typically perform poorly when the training and testing samples are from different distributions. Thus, it is necessary for deep neural networks to recognize whether the input samples are out-of-distribution samples. This research is related to open-set recognition [43]. Researchers observed that the prediction scores for out-of-distribution data tend to differ from those for in-distribution data [44]. Probability models [45], linear models [46], and reconstruction models [47] have been proposed for enhancing networks’ anomaly detection performance. Deep anomaly detection methods for 2D images are the most popular [48], [49]. However, methods for 3D objects lag behind their 2D counterparts [50]. Furthermore, scant research has been conducted on the analysis of outliers of radar-based micromotion in non-Euclidean space.

III. METHODOLOGY

In this section, we will describe the temporal range-Doppler PointNet-based pipeline in detail. As illustrated in Fig. 1, our approach mainly contains three parts. (1) In the range-velocity-time point acquisition part, the radar echoes are transformed via range-Doppler processing along the time axis. Then, the target information is gathered by CFAR (constant false alarm rate) detection. (2) In the intensity-based point cloud reconstruction part, based on the intensity values of the detection points, the point features are aggregated by motion sculpture construction and iterative farthest point sampling. (3) In the point cloud classification part, the point form features are finally processed by the hierarchical PointNet module to recognize the corresponding motion labels. In this part, we also consider the anomaly detection issue, making the network detect out-of-distribution echoes.

A. RANGE-VELOCITY-TIME POINTS ACQUISITION

The range-velocity-time points are transformed from the radar signal, which is actually a superposition of the response from the whole body. Assuming that different body segments are treated as discrete scattering centers, the backscattering echoes of the human, denoted by \( y(r, t_s) \), can be calculated as follows:

\[
y(r, t_s) = \sum_{i=1}^{N_t} \rho_i (t_s) \delta (r - r_i(t_s))
\]

where \( N_t \) is the total number of scattering segments of the target, \( \rho \) is the reflectivity parameter determined by the transmitting frequency, surface texture, local geometry, and distance from the target, \( \delta \) is the Dirac delta function, and \( r_i(t_s) \) denotes the radial distance between the radar and the \( i_{th} \) body part within a pulse denoted by \( t_s \).

As illustrated in Fig. 2, all the radar echoes can be operated at two distinct time scales [23]: the fast time and the slow time. The fast time dimension refers to a single pulse while the slow time is related to multiple radar pulses. By measuring the interval between transmitting and receiving a single pulse, the instantaneous distance between the target and the radar can be calculated. Moreover, since the target is not an ideal point, more details of the extended target in the range dimension can also be captured by the signal along the fast
time. When more than one radar pulse is transmitted, more dynamic motions of the target can be obtained by analyzing the reflected pulses along the slow time.

Different from applying the Fourier transform to radar pulses within fixed range bins, we apply the operation on the whole range dimension at a certain slow time interval and then repeat the process iteratively. Therefore, the whole motion profile representation can be obtained by the following function:

$$S_{\text{motion}}(r, v, t_s) = \text{Spec} \left( t_f, f, t_s \right) \mid t_f = \frac{2r}{c}, f = \frac{2v}{\lambda}$$

where $S_{\text{motion}}(r, v, t_s)$ is the whole motion trajectory about the range $r$, velocity $v$ along the slow time $t_s$, $f$ is the Doppler frequency shift, $\lambda$ is the wavelength of the radar signal, and $c$ is the speed of light. It should be noted that the whole range bins, rather than the range bins containing the target (which are used in the micro-Doppler spectrogram), are transformed in this step. Therefore, it is necessary to remove the target-absent regions in 3D space.

As illustrated in Fig. 3, we slide a 2D CFAR detector [51] across each range-velocity frame pixel by pixel to obtain the pixels whose intensity exceeds the detection threshold. The CFAR detector comprises three cells: test cells (red) cover the region to be detected, reference cells (green) estimate the intensity of the covered region to offer the detection threshold, and guard cells (yellow) are barriers that separate the test cells and reference cells. After CFAR detection, the detected scatters of each range-velocity frame can be seen as point clouds in range-velocity-time space.

**B. INTENSITY-BASED POINT CLOUD RECONSTRUCTION**

Based on the above analysis, it is known that the range-velocity-time domain, compared with the 2D representation, retains all the information of humans carried in the raw echoes (see Fig. 4). However, the inherent redundancy of these point sets makes it difficult, even though the recently proposed 3D point network cannot address this issue (more details about the experimental results are in Section IV-E), because raw point clouds in the range-velocity-time space suffer from redundancy and disorder issues. For example, we may encounter the following: (1) the number of point sets during an observation period is enormous (thousands of points in one frame and hundreds of frames in one second); (2) points existing in one range-velocity frame have information solely about a subset of the limbs and often miss other body parts (the human target acts as a reflector rather than a scatterer; thus, not all body parts reflect the signal back to the radar [20]); and (3) the interval between range-velocity frames is nontrivial and lacks a correlation procedure for the points in consecutive frames.

To address the above issues, we introduce the intensity-based reconstruction method. As shown in Fig. 5, separate body parts, transformed in the range-velocity map, have high intensity values. It has been found that different body parts are separated by the intensity signature, and the same body part does not exhibit abrupt intensity changes within a limited period. Therefore, guided by the intensity signature, the point cloud can be reconstructed in the following steps: (1) For the redundancy problem, we retain the contour of each body part and remove the inside region for simplification. (2) For the lost information in a single frame and the lack of correlating consecutive frames, we interleave the point between consecutive frames and align the point belonging to the same body part among different frames based on the intensity value.

In practice, these reconstruction steps can be achieved by sampling the farthest points [52] from the motion sculpture. Specifically, we first construct the motion sculpture by connecting the CFAR detection points of equal intensity much in the same way contour lines join points of equal elevation [32]. Then, we re-sample points from the sculpture surface via iterative farthest point sampling. This point cloud sampling process not only simplifies the point cloud but also aggregates useful information about human motion profiles.
C. POINT CLOUD CLASSIFICATION

The reconstructed point cloud is finally classified by a hierarchical PointNet models. The detailed network architecture of the hierarchical network is presented in Fig. 6, which is based on the architecture of PointNet proposed in [36] and [37].

As seen from the illustration, the network comprises three point set abstraction levels. The first level groups the point cloud into 512 ($N_1 = 512$) local regions while the second abstraction level groups into 128 ($N_2 = 128$) regions. The centroid of each region is obtained by iterative farthest point sampling. Then, points inside each local region are collected by a k-nearest neighbor search ($k = 64$) and remain within a radius in Euclidean space. After sampling and grouping operations, each local region is fed into a shared basic PointNet (the dashed box shown in Fig. 6) to extract a $C_l$-dimensional feature ($l = 1, 2$). Then, the $C_l$-dimensional features are concatenated with the $d$-dimensional coordinates of the corresponding points. At the third level, all the input points are fed into a single PointNet model to extract the global feature. Finally, the global feature is input into a multilayer perceptron to make the classification.

The basic PointNet module [36], which is the dashed box in Fig. 6, maps each input into a $C_l$-dimensional feature vector by the MLP (multilayer perceptron). The weights of the MLP are shared among all the input points. Then, all the point features are aggregated by the max pooling operation to yield a single $C_l$-dimensional feature vector. For the hierarchical PointNet model, the input features are grouped into local regions and mapped into the basic PointNet model separately. The extracted feature of the lower levels is also processed region by region. This process enables the network to capture features from the original point set at increasingly larger scales along a multiresolution hierarchy. In particular, by processing the point sets region by region, in much the way convolutional filters slide across the 2D image, can analyze the whole motion profiles on a fine-grained scale.

D. ANOMALY DETECTION

Deep learning classifiers tend to fail if their test set distribution differs from the training set distribution. Moreover, if confronted with anomalies, these classifiers fail silently by yielding high-confidence predictions. In this section, we focus on our model’s performance in detecting out-of-distribution radar echoes, which is critical for practical applications. In radar research, training samples are often obtained in single-mover scenarios. When more than one target is present, the classification performance will deteriorate substantially. Here, the overlapping echoes are out-of-distribution samples. To detect these outliers, we preprocess the range-velocity-time points prior to feeding them into the hierarchical network.

Specifically, the preprocessing operation is inspired by ODIN [48], a recent work that introduces pixel perturbations into the input image to increase the margin between the maximum softmax scores of in-distribution and out-of-distribution samples. Here, we modify the input point cloud by perturbing its position values instead of the pixels in the images. For each input point cloud $\mathcal{P}$, the modified point cloud is expressed as

$$\mathbf{p}_i' = \mathbf{p}_i - \epsilon \text{sign} (\nabla \mathbf{p}_i \log S(\mathcal{P}; T))$$

where $\mathbf{p}_i$ is the $i^{th}$ point coordinate in the original point set $\mathcal{P}$, $\mathbf{p}_i'$ is its corresponding point of the modified point set $\mathcal{P}'$, $\epsilon$ is the perturbation magnitude, $S$ is the softmax score, and $T$ is the temperature scaling factor. In temperature scaling [48], the logits are scaled by the constant factor $T$ prior to being fed into the softmax layer. This operation can also increase the softmax score margin between three-dimensional in-distribution and out-of-distribution point samples.

Via this input preprocessing step, outliers can be detected automatically by the designed network in the test phase without retraining the network with anomalies. We find that the hierarchical architecture is also essential for anomaly detection (more details will be clarified in Section IV-D).

E. IMPLEMENTATION DETAILS

All the radar echoes are trimmed to 2-second clips. The input point sets consist of two-part information: the normalized coordinate value from the range, velocity, and time domains, and the corresponding surface normal according to [53]. The number of points in one sample is 1024. Their orientation is normalized by performing principal component analysis on the corresponding 3D coordinate indexes. The radius of the $k$-nearest neighborhood is set to 0.1 at the first abstraction level and 0.3 at the second level.

The model is trained in a supervised way, calculating the similarity with two-part objective functions consisting of the cross-entropy loss and an L2 norm regularization term. We use the SGD optimizer with an initial learning rate of 0.01 which is divided by 10 after 20 epochs, and a regularization strength of 0.001. The batch size is 32. The number of training epochs is 100. Jittering and rotating operations are applied during data augmentation.

For anomaly detection, the temperature scaling factor $T$ is set to 1000, and the perturbation magnitude $\epsilon$ in Eq. 3 is 0.007. The above two factors are tuned on 100 in-distribution single-target samples that are rotated 90 degrees and added to the original copy.
FIGURE 6. Overview of the proposed framework. Input point sets are processed by three point set abstraction levels. The point sets are sampled and grouped as local regions, and then fed into a shared basic PointNet level by level. Finally, the global feature is put into a multi-layer perceptron to make the classification.

IV. EXPERIMENTS

In this section, we evaluate the proposed method on MoCap (motion capture) datasets and ultrawideband radar measurements.

A. DATASETS

1) CMU MoCap DATASET

We first examine the performance of the proposed approach on the CMU MoCap dataset. This dataset, generated by Graphics Laboratory at the Carnegie Mellon University [54], facilitates research on human movements. It contains over 2600 different clips of full body motions performed by 144 subjects. Each motion clip records the temporal three-dimensional coordinate information of 31 body segments. In our work, the reflected signals from the recorded human body are synthesized by the ellipsoid-based human backscattering model. As shown in Fig. 7, we model every two adjacent key points as a body segment and approximate the segment as a prolate ellipsoid. The value size of each part is described in [55]. With the known volume size and measured distance between adjacent key points, the radar cross section of body segments can be calculated. By summing the echoes from all the ellipsoids, the backscattering echoes of the whole target can be obtained. The bandwidth of the ultrawideband radar is 1.5 GHz, and the center frequency is 4.0 GHz. The MoCap data are interpolated at 500 Hz. To diversify the simulated data, the radar position is changed in every simulation situation. There are 400 samples per activity in the training set and 100 samples per activity in the testing set.

FIGURE 7. Simulated model of a boxing motion based on the CMU database.

2) KINECT-BASED MoCaP

Kinect sensor, which is able to capture the time-varying information of human skeletons, can also be used to simulate micromotion signatures [56]. The difference between the CMU- and Kinect-based simulations is the number of recorded human joints. The former records 31 joints, and the latter records 20 joints (see Fig 8). Here, we test the proposed approach on the UTD-MHAD dataset [57], which is a Kinect-based database containing 27 actions performed by 8 subjects (4 females and 4 males). In our setting, the radar bandwidth is also set to 1.5 GHz. The center frequency is 4.3 GHz. We interpolate the MoCap data file to obtain a pulse repetition interval of 2 milliseconds. To evaluate our method’s robustness to noise, we add Gaussian noise to the

VOLUME 8, 2020
3) THROUGH-OBSTACLE EXPERIMENT

The employed ultrawideband radar system operates between 3.1 GHz and 4.8 GHz and transmits microwaves at 300 Hz. The antenna module comprises one transmission and one reception antenna port. The sensor is set at 1.2 meters high to match the human’s center of gravity. We put the radar in a wooden box, where the forward, left, and right sides of the sensor are all blocked by opaque wooden board (see Fig. 9). In the experiment, five subjects (4 males and 1 female) are measured for 10 seconds as they perform different kinds of activities each time. Each activity is measured twenty times per subject, and the measurement range is between 1 m and 6 m. The measured data are trimmed to a 2-second duration to perform the point-based processing step. Among the processed point sets, we choose 400 samples per class as the training set and 100 point clouds per label as the testing set.

4) ANOMALY DETECTION

This dataset contains 300 radar echoes from two-person and three-person moving scenarios that were synthesized from the CMU MoCap dataset. To address the multi-mover echoes, the designed model should be able to detect and separate the echoes. Here, we test whether our trained method can detect overlapping echoes during inference after being trained only on single-target cases.

B. CLASSIFICATION PERFORMANCE COMPARISON

For comparison, we select several existing models for human activity recognition as baselines: (1) MD-CNN [31], a convolutional neural network that processes micro-Doppler spectrograms; (2) R-CNN, a similar architecture that processes the time-range domain instead of the time-velocity domain; (3) MD-SVM [27], a machine learning method that extracts six statistical features from the micro-Doppler spectrogram and feed them into a support vector machine; (4) R-SVM, a support vector machine that takes the SIFT features [58] of the range-time domain as input instead of the micro-Doppler spectrogram; (5) MDR-SA [11], a recently proposed sparse autoencoder that takes both micro-Doppler spectrograms and range profiles as input, and the hidden features of the autoencoder are fed into an MLP for classification; and (6) MD-SA and (7) R-SA which use the same architecture proposed in [11] and process on the micro-Doppler and range profiles separately. We also compare our model with its naive version, P-Net [15], which consumes the whole range-Doppler-time point sets with PointNet [36]. All these methods are tested on the CMU dataset and the UTD-MHAD datasets.

We select eight activities (walking, running, jumping upward, jumping forward, standing with slight movement, boxing, kicking and climbing) from the CMU dataset. The intermediate results (reconstructed range-velocity-time point sets) of the eight activities are shown in Fig. 10. The results of the different methods are shown in Table 1.

As can be observed from the table, (1) both point-based models, namely, the temporal range-Doppler PointNet and P-Net models, achieve better results on most of the behaviors as compared to other baselines. This finding verifies our claim that incorporating the rich information of both the time-velocity and time-range domains could help improve the model performance. (2) Our hierarchical model outperforms its basic model (P-Net). This finding indicates that capturing features from the point set at increasingly larger scales enables the model to better distinguish various kinds of range-Doppler-time points. (3) The autoencoder using the fusion of the range and velocity temporal information achieves third
place, outperforming other methods that solely rely on either micro-Doppler spectrogram or range information. The reason is that the autoencoder captures more information than the single source-based models. (4) The micro-Doppler based model performs better than the same architecture using range profile features, demonstrating that the micro-Doppler spectrogram is more distinctive than range information for this configuration of radar signals.

For the UTD-MHAD dataset, we gathered micromotion signatures from seven activities (jogging, walking, sitting to standing, lunging, throwing, boxing and clapping). The results of the different methods are shown in Table 2. This table reveals that our method outperforms the others in terms of the mean classification accuracy. Compared with the other methods, the average accuracy improves from 87.7% to 89.0%. For several categories, such as clapping, the classification accuracy achieves second place by 7 percentage points. Although improvement is not realized on all classes, the overall performance is increased by 5.9 percentage points when compared with the basic version (P-Net).

### C. THROUGH-OBSTACLE EXPERIMENT

Here, we test the radar-based sensor’s performance in fully occlusion conditions. The experimental setting is clarified in Section IV-A. We employ the leave-one-user-out (LUOU) cross-validation scheme to conduct the experiment. In this setting, the reflected signal from each subject is used for testing once, while the remaining four subjects are used for training. Six motion classes (boxing, jumping, kicking, running, standing, walking) for 5 subjects (4 males and 1 female) are used for retraining our model.

### D. OVERLAPPING ECHO DETECTION

We adopt five metrics that are used in [48] to evaluate our method in terms of overlapping samples: (1) the FPR at 95% TPR is the probability that an outlier sample is mistaken as an in-distribution sample when the true-positive rate is 95%; (2) the detection error is the misclassification probability over all the detection thresholds when the true-positive rate is above 95%. (3) the AUROC is the area under the receiver operating characteristic curve; (4) AUPR-In is the area under the precision-recall curve when the in-distribution is specified as positive; (5) AUPR-Out is the area under the precision-recall curve when the outlier is specified as positive.

The per-subject and mean activity classification results are shown in Table 3. Our model outperforms the second-place model by over 7 percentage points accuracy, which indicates that our model has a competitive generalization capability even in fully occluded conditions. Almost all the models achieve a low accuracy for Subject 4 because Subject 4 is a female whose velocity and range extent of each motion are different from those of the other subjects. It should also be noted that although the MDR-SA achieve a competitive classification accuracy in the previous test, its generalization capability is the worst.

### TABLE 1. Human activity classification accuracy on the CMU dataset (%).

| Method   | Mean | Box | Climb | Walk | Jump | Run | Front | Kick | Stand |
|----------|------|-----|-------|------|------|-----|-------|------|-------|
| MD-CNN   | 91.1 | 93  | 89    | 97   | 92   | 96  | 97    | 59   | 98    |
| R-CNN    | 83.5 | 83  | 70    | 80   | 98   | 83  | 96    | 79   | 79    |
| MD-SVM   | 90.3 | 89  | 91    | 99   | 94   | 96  | 79    | 74   | 100   |
| R-SVM    | 75.6 | 92  | 35    | 87   | 100  | 77  | 38    | 79   | 97    |
| MDR-SA   | 91.4 | 100 | 96    | 99   | 98   | 93  | 98    | 64   | 83    |
| MD-SA    | 86.3 | 96  | 75    | 99   | 70   | 93  | 94    | 64   | 100   |
| R-SA     | 80.6 | 95  | 86    | 73   | 98   | 72  | 89    | 55   | 77    |
| P-Net    | 93.6 | 91  | 96    | 97   | 100  | 96  | 100   | 69   | 100   |
| Ours     | 95.2 | 94  | 96    | 96   | 97   | 96  | 97    | 87   | 98    |

### TABLE 2. Human activity classification accuracy on the UTD-MHAD dataset (%).

| Method   | Mean | Jog | Walk | Sit | Stand | Lung | Thow | Box   | Clap |
|----------|------|-----|------|-----|-------|------|------|-------|------|
| MD-CNN   | 81.9 | 88  | 84   | 87  | 73    | 79   | 81   | 81    |
| R-CNN    | 78.3 | 81  | 78   | 78  | 70    | 62   | 76   | 75    |
| MD-SVM   | 80.6 | 83  | 82   | 85  | 78    | 74   | 77   | 85    |
| R-SVM    | 67.6 | 63  | 70   | 69  | 65    | 64   | 69   | 73    |
| MDR-SA   | 83.1 | 88  | 87   | 80  | 85    | 76   | 84   | 81    |
| MD-SA    | 76.1 | 78  | 83   | 82  | 75    | 64   | 80   | 71    |
| R-SA     | 70.4 | 72  | 79   | 77  | 73    | 56   | 76   | 69    |
| P-Net    | 87.7 | 91  | 93   | 88  | 82    | 84   | 90   | 86    |
| Ours     | 89.0 | 92  | 93   | 87  | 85    | 82   | 91   | 93    |

### TABLE 3. Classification results in leave one user out (LOUO) cross-validation scheme (%).

| Methods   | Mean | S1 | S2 | S3 | S4 | S5 |
|-----------|------|----|----|----|----|----|
| MD-CNN    | 81.16| 78.13| 87.04| 84.25| 75.58| 80.78|
| R-CNN     | 65.94| 69.57| 68.19| 73.48| 71.32| 47.14|
| MD-SVM    | 77.76| 80.10| 78.61| 82.19| 68.63| 79.34|
| R-SVM     | 72.45| 78.03| 75.41| 76.61| 73.11| 59.08|
| MDR-SA    | 69.83| 71.12| 73.19| 75.41| 61.31| 68.13|
| Ours      | 90.35| 87.32| 91.76| 95.05| 86.06| 91.57|

The candidate methods are as follows: (1) **PointNet-based methods**, namely, P-Net (a three-dimensional deep learning framework that proposed in [15]), P-Net+OSVM (an architecture proposed in [46] that combines PointNet model and a one-class support vector machine), P-Net+Scale (a PointNet model that uses temperature scaling [48]), and P-Net+Adv (a PointNet model that uses temperature scaling and perturbation, which was introduced in Eq. 3); and (2) **hierarchical methods**.
### TABLE 4. Anomaly detection performance for overlapping echoes (%).

| Method               | FPR (95 TPR) | Detection Error | AUROC In | AUROC Out | AUPR  |
|----------------------|--------------|-----------------|----------|-----------|-------|
| P-Net                | 91.3         | 36.9            | 64.6     | 70.2      | 51.9  |
| P-Net+OSVM           | 89.0         | 46.1            | 63.7     | 61.1      | 64.4  |
| P-Net+Scale          | 89.5         | 44.7            | 52.7     | 58.9      | 44.2  |
| P-Net+Adv            | 88.5         | 44.3            | 53.5     | 59.4      | 44.9  |
| HP-Net               | 34.3         | 15.1            | 92.1     | 94.4      | 89.1  |
| HP-Net+OSVM          | 87.3         | 45.6            | 52.5     | 46.6      | 58.7  |
| HP-Net+Scale         | 31.5         | 14.5            | 91.7     | 93.6      | 89.4  |
| Ours                 | 29.8         | 15.1            | 93.3     | 95.2      | 91.8  |

PointNet-based methods, namely, HP-Net (the hierarchical PointNet model introduced in Fig. 6), HP-Net+OSVM (an architecture that combines the hierarchical PointNet model and a one-class support vector machine that was proposed in [46]), and HP-Net+Scale (a hierarchical PointNet model that uses temperature scaling). Both types of modules are only trained on the in-distribution samples, that is, the CMU MoCap dataset. In the test phase, the test samples are from both single target echoes and overlapping echoes. We analyze each method’s output to determine whether the overlapping echoes can be detected as anomalies.

Table 4 presents a performance comparison on multimover overlapping echoes. This table reveals that the hierarchical architecture significantly contributes to anomaly detection. According to Table 4, HP-Net and its extensions outperform the P-Net-based models (P-Net, P-Net+OSVM, P-Net+Scale, and P-Net+Adv). Our model (HP-Net with 3D adversarial perturbations) outperforms the other hierarchical models in terms of the 4 metrics. Compared with its baseline (HP-Net), the FPR at 95% TPR decreases from 34.3% to 29.8%, and the AUROC and AUPR values also improve.

Fig. 11 compares the histograms of the confidence distributions on the in-distribution and out-of-distribution (overlapping echo) samples for the eight methods. We assume that the classifier outputs a higher confidence score for the in-distribution samples than for the out-of-distribution samples. In a deep network, the confidence is measured by the maximum softmax value. Our method (Fig. 11(h)) has less overlap between the in-distribution and out-of-distribution samples compared with the other methods and, thus, separates the distributions more effectively.

**E. IMPACT OF POINT CLOUD RECONSTRUCTION**

Considering that the CFAR detection result is also presented in point form, we investigate the impact of our proposed point reconstruction modules. Specifically, we sample the same number of points directly from the CFAR detection points and feed them into the same temporal range-Doppler PointNet model to test its performance.

Fig. 12 shows the average and per-class motion classification accuracy in the CMU MoCap dataset. The average accuracy of the CFAR points-based methods is over 10% lower than that of the range-velocity-time point method. When confronted with sudden motions such as kicking and boxing, the CFAR-points-based model deteriorates rapidly. Fig. 13 shows the classification accuracy on the practical experiment.
measured by the ultrawideband radar. The average accuracy of the CFAR points-based methods is more than 9% lower than that of the range-velocity-time point method. Fig. 14 shows the average classification accuracy versus different SNRs (signal-to-noise ratio). To mimic a real application, we introduce different levels of Gaussian white noise to the test samples, while the training samples have an SNR of 25 dB. The figure shows that our model is more robust to different levels of noise. However, the CFAR-points-based method is more sensitive to noise and fails completely when the SNR is less than 10 dB.

As shown in the above figures, the network’s performance deteriorates significantly when using CFAR points, suggesting that points directly sampled from the original CFAR results cannot effectively depict various human motions. The finding verifies that extracting from the surface and interleaving between each transmission interval can represent the motion profiles more effectively.

F. SELF-COMPARISONS

Here, we evaluate the effects of the number of abstraction levels and the size of the final max layer output. The hyper-parameters are chosen based on 5-fold cross-validation using the CMU MoCap training set. In Fig. 15, we show how our model’s performance changes with regard to both the numbers of levels (1, 2, 3, and 4) and the final max layer output (256, 512, 1024, and 2048). It can be seen that performance improves as we increase the numbers of levels and output channels; however, the network is saturated at 3 levels and approximately 1024 output channels. Considering that a larger model size may lead to overfitting, the number of abstraction levels is set to 3 and the number of max layer output is set to 1024.

V. CONCLUSION AND FUTURE WORK

In this paper, we propose a temporal range-Doppler PointNet-based method to analyze human behavior using an ultrawideband radar. Our approach can make full use of all the informative signatures gathered from the range-velocity-time points, providing a new comprehensive way to process the reflected signal wave. In the experiments, we show that our method significantly outperforms existing methods. In particular, the reconstruction module aggregates the range-velocity-time point features to 3D deep neural networks, which effectively depicts human behavior dynamics and addresses disorder and redundancy issues. Moreover, our work in anomaly detection suggests that hierarchical features are critical in three-dimensional point cloud outlier analysis. In the future, we will explore cross-modality learning by leveraging the information from both radar and optical sensors.

REFERENCES

[1] C. Debes, A. Merentitis, S. Sukhanov, M. Niessen, N. Frangiadakis, and A. Bauer, “Monitoring activities of daily living in smart homes: Understanding human behavior,” IEEE Signal Process. Mag., vol. 33, no. 2, pp. 81–94, Mar. 2016.
[2] X. Ben, P. Zhang, Z. Lai, R. Yan, X. Zhai, and W. Meng, “A general tensor representation framework for cross-view gait recognition,” Pattern Recognit., vol. 90, pp. 87–98, Jun. 2019.
[3] A. Singh, D. Patil, and S. N. Omkar, “Eye in the sky: Real-time drone surveillance system (DSS) for violent individuals identification using ScatterNet hybrid deep learning network,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2018, pp. 1629–1637.
[4] C. Chen, Q. Chen, J. Xu, and V. Koltun, “Learning to see in the dark,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 3291–3300.
[5] S. Zhang, J. Yang, and B. Schiele, “Occluded pedestrian detection through guided attention in CNNs,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6995–7003.
[6] M. G. Amin, Through-the-Wall Radar Imaging. Boca Raton, FL, USA: CRC Press, 2017.
[7] T. Jin, B. Chen, and Z. Zhou, “Image-domain estimation of wall parameters for autofocus of Through-the-Wall SAR imagery,” IEEE Trans. Geosci. Remote Sens., vol. 51, no. 3, pp. 1836–1843, Mar. 2013.
[8] V. C. Chen, F. Li, S.-S. Ho, and H. Wechsler, “Micro-Doppler effect in radar: Phenomenon, model, and simulation study,” IEEE Trans. Aerosp. Electron. Syst., vol. 42, no. 1, pp. 2–21, Jan. 2006.
[9] D. Tahmoush, “Review of micro-Doppler signatures,” IET Radar, Sonar Navigat., vol. 9, no. 9, pp. 1140–1146, Dec. 2015.
MENG LI was born in Zhengzhou, China, in 1991. She received the B.Sc. degree from the China University of Geosciences, Wuhan, China, in 2013, and the M.Sc. degree from the University of Science and Technology of China, Hefei, China, in 2016. She is currently pursuing the Ph.D. degree with Tsinghua University, Beijing, China. Her main research interests are motion recognition, machine learning, and geometric deep learning.

TAO CHEN received the Ph.D. degree from the University of Science and Technology of China, in 2004. He is currently a Tenure-Track Professor and a Doctoral Supervisor of the Institute of Public Safety Research, Tsinghua University, China. He is also the Deputy Director of the Security Department, the Assistant Dean of the THHI (Hefei Institute for Public Safety Research, Tsinghua University). He is also the Level 1 Certified Fire Engineer in China, a Council Member of the Institute of Building Fire Protection-ASC, a member of the Society for Risk Analysis—China, and a member of the China Association for Public Safety. He is mainly engaged in research on national emergency response platform, public safety decision-making methods, human behavior and evacuation dynamics, and resettlement technologies during emergency. He has successively engaged in more than 40 research projects of China, and was responsible for four projects of the National Natural Science Foundation of China, two projects of the National Key Technology Research and Development Program, and one project of the Science Foundation for Post Doctorate Research. He has published over 80 articles and two books in recent years. He won the State Scientific and Technological Progress Award (First Class) for one time and the Provincial or Ministerial Science and Technology Award (First Class) for four times.

HAO DU received the B.Sc. and M.Sc. degrees from the Electronic Engineering Institute, Hefei, China, in 2013 and 2016, respectively. He is currently pursuing the Ph.D. degree with Tsinghua University, Beijing, China. His main research interests are radar signal processing, machine learning, and geometric deep learning.