Human Motion Recognition With Limited Radar Micro-Doppler Signatures

Li, X.; He, Y.; Fioranelli, F.; Jing, X.; Yarovoy, A.; Yang, Y.

DOI
10.1109/TGRS.2020.3028223

Publication date
2020

Document Version
Final published version

Published in
IEEE Transactions on Geoscience and Remote Sensing

Human Motion Recognition With Limited Radar Micro-Doppler Signatures

Xinyu Li, *Graduate Student Member, IEEE*, Yuan He, *Member, IEEE*, Francesco Fioranelli, *Senior Member, IEEE*, Xiaojun Jing, *Member, IEEE*, Alexander Yarovoy, *Fellow, IEEE*, and Yang Yang

***Abstract*— The performance of deep learning (DL) algorithms for radar-based human motion recognition (HMR) is hindered by the diversity and volume of the available training data. In this article, to tackle the issue of insufficient training data for HMR, we propose an instance-based transfer learning (ITL) method with limited radar micro-Doppler (MD) signatures, alleviating the burden of collecting and annotating a large number of radar samples. ITL is a unique algorithm that consists of three interconnected parts, including DL model pretraining, correlated source data selection, and adaptive collaborative fine-tuning (FT). Any of the three components cannot be excluded; otherwise, the performance of the entire algorithm decreases. The experiments with a radar data set of six human motions show that ITL achieves state-of-the-art performance for HMR with limited training samples, outperforming several existing transfer learning approaches. Especially, when there are only 100 samples per person per class, ITL yields an F1 score of 96.7%. Last but not least, ITL is more generalized to human motion differences. Though adapted to recognize the persons' motions in a small-scale target data set, ITL can also classify the persons' motion data used for pretraining, achieving up to 11.0% F1 score enhancement over the conventional FT method.**

***Index Terms*— Deep learning (DL), human motion recognition (HMR), radar micro-Doppler (MD), transfer learning.**

I. Introduction

HUMAN motion recognition (HMR) has emerged as a vital step in many applications, such as assisted living [1], remote health monitoring [2], and human–computer interaction [3], among others. Due to the unique characteristic, e.g., robustness to the environment and the penetrability to opaque objects, radar has been increasingly applied for HMR [4]. The radar spectrogram is the power distribution of target return over time and frequency and is a typical 2-D representation for analyzing radar micro-Doppler (MD) signatures. These spectrograms are individual-unique and motion-unique and have been increasingly used for radar-based HMR [5]–[9].

However, since collecting and annotating radar data manually are time-consuming and expensive, most labeled radar data sets are quite small-scale, and the number of labeled radar data is limited. In this circumstance, training a classification model from scratch with such limited training data, especially deep learning (DL) model, often leads to overfitting. At the same time, due to the difference in data distribution, directly using a trained model to classify the limited data is not effective. As a result, the performance of HMR approaches is often hindered by limited radar data. How to tackle the issue of classification with relatively few training data has become one of the main concerns in the radar community.

Current work that deals with classification tasks with limited training data can be roughly divided into three categories. The first category is to build classifiers robust to limited training data, such as the models in [10]–[12]. The second category is labeled data augmentation with synthetic data [13]–[15]. Transfer learning, which can take advantage of prior knowledge from an existing large-scale data set (source domain) as a supplement for the tasks on a different but related small-scale data set (target domain), is the third category.

Transfer learning can reduce the dependence of models on a vast amount of labeled data by utilizing this prior knowledge from another related domain. Several transfer learning approaches have been presented for radar-based HMR [16]–[19]. For instance, Park *et al.* [18] presented a deep convolutional neural network (CNN) pretrained on ImageNet and fine-tuned the network with measured radar MD spectrograms for human aquatic activity classification. Seyfioğlu *et al.* [19] proposed a residual learning model DivNet trained on the simulated radar MD spectrogram data set and fine-tuned the model with a measured data set to classify seven human activities. The fine-tuning (FT) strategy used in these methods utilizes the target data to fine-tune the pretrained DL models and transfers the source knowledge to compensate for the insufficiency of target domain data. We refer this strategy to the Conventional FT.

Manuscript received March 18, 2020; revised August 21, 2020; accepted September 26, 2020. Date of publication October 13, 2020; date of current version July 22, 2021. This work was supported in part by the National Nature Science Foundation of China under Grant 61901049, in part by the BUPT Basic Research Fund under Grant 500419757, and in part by the BUPT Excellent Ph.D. Students Foundation under Grant CX2020208. *(Corresponding author: Yuan He.)*

Xinyu Li, Yuan He, and Xiaojun Jing are with the Key Laboratory of Trustworthy Distributed Computing and Service, Beijing University of Posts and Telecommunications (BUPT), Beijing 100876, China (e-mail: lixinyu@bupt.edu.cn; yuanhe@bupt.edu.cn; jxiaojun@bupt.edu.cn).

Francesco Fioranelli and Alexander Yarovoy are with the Department of Microelectronics, Delft University of Technology, 2628CD Delft, The Netherlands (e-mail: f.fioranelli@tudelft.nl; a.yarovoy@tudelft.nl).

Yang Yang is with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: yang_yang@tju.edu.cn).

Digital Object Identifier 10.1109/TGRS.2020.3028223

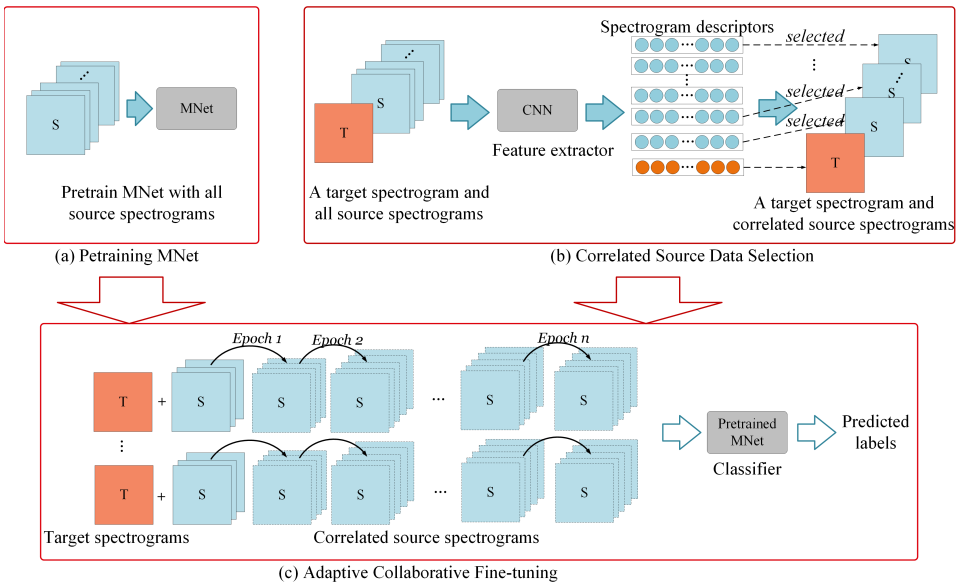

Fig. 1. Pipeline of the proposed *ITL* for HMR. (a) Pretraining the proposed MNet with the source spectrogram data set. (b) CSDS: the process of constructing spectrogram descriptors (represented as vectors) and selecting the correlated source instances for every spectrogram in the target domain. (c) ACFT: the process of adaptively FT the pretrained MNet with target spectrograms and the selected source instances for several epochs. In the pipeline, the MNet is first pretrained with the source data. At the same time, the CSDS process is performed to select correlated source samples for every target sample. Then, the ACFT process is performed for FT the pretrained model.

However, the performance of Conventional FT approaches often degrades when the amount of labeled data drops. Furthermore, the catastrophic forgetting effect [20] (the tendency of DL models to abruptly forget previously learned tasks after being trained for a new task) usually occurs in the Conventional FT. In other words, when the model fine-tuned on the target data set is applied to classify the persons' motions in the source data set, the performance usually decreases. As a result, the Conventional FT method often lacks generalization and cannot scale well to the persons in different domains simultaneously.

In this article, aiming at effectively training an HMR model with limited training samples and improving the performance on the target task, we propose a novel instance-based transfer learning (ITL) approach with radar MD signatures. To recognize the motions in the target data set (target domain) with limited training data, a larger data set (source domain) whose data distribution is different but relevant to that of the target data set is used as supplementary. Furthermore, ITL utilizes some source instances that are similar to the target data to jointly fine-tune a DL model to recognize human motions in the target domain.

The overall flow of the proposed ITL is shown in Fig. 1. First, we design a deep CNN MNet for radar-based HMR as the backbone of ITL and pretrain it with all available source data [see Fig. 1(a)]. At the same time, a correlated source data selection (CSDS) algorithm is designed to pick up partial instances from the source domain as supplements for the target data [see Fig. 1(b)]. Then, an adaptive collaborative fine-tuning (ACFT) algorithm [see Fig. 1(c)] is presented to fine-tune the pretrained MNet with the whole target data set and the selected source data. With ACFT, ITL can perform the target task while retaining partial source knowledge. This property allows the fine-tuned MNet to be used for classifying the target motion data, as well as accurate identification of the motions in the source data set. In other words, ITL is more generalized to cope with the data distribution discrepancy between the two domains, which is often caused by human motion differences.

The major contributions of this article can be summarized as follows.

1) We propose an ITL approach for radar-based HMR with limited training data, alleviating the need of labeling a large amount of radar data.
2) ITL is a unique algorithm that consists of three interconnected parts, including DL model pretraining, CSDS, and ACFT. Any of the three components cannot be excluded; otherwise, the performance of the entire algorithm for HMR decreases.
3) The experimental results demonstrate that ITL has good performance for recognizing human motions with limited radar data, outperforming several state-of-the-art HMR methods. When there are only 100 samples per person per class, an F1 score of 96.7% is yield. Furthermore, though trained for the classification task in the target data set, ITL can generalize to human motion differences and recognize different persons' motions in both the source domain and the target domain.

The rest of this article is organized as follows. Section II presents some related work on ITL. Section III describes the measured data collection and preprocessing process and presents some data analysis. Section IV introduces the structure of ITL. Section V presents the analysis and discussion of the experimental results. Furthermore, some ablation studies on ITL are performed in Section VI. Finally, Section VII concludes this article.

II. Related Work

Transfer learning can improve the performance of predictive models to a new domain (target domain) by utilizing the knowledge in a preexisting domain (source domain). In this way, compared with training a DL model from scratch, transfer learning approaches do not require a large amount of labeled data, so they can be applied for the tasks that have only limited data. Among transfer learning methods, the ITL is a typical approach and has been employed in many applications [21]–[25]. The instance-based methods assume that some source data can be reused for learning in the target domain by reweighting. In general, the reweighted source data are used in two ways: pretraining a DL model and FT a pretrained model together with the target data [26].

For instance, Ngiam *et al.* [22] proposed an instance-based approach where the source data are given diverse weights of importance, and the weighted source data are employed for pretraining a DL model. Then, the proposed method is applied to fine-grained image classification tasks. Arnold and Cohen [23] proposed a cross-language ITL approach for multilingual deep retrieval. Wang *et al.* [24] proposed a Minimax Game-based transfer learning method and utilize the adversarial scheme to select source data for training the DL model with the target data. The proposed method is applied for item recommendation and text retrieval tasks. In addition, Ge and Yu [21] proposed an ITL approach for natural image classification tasks. The Kullback–Leibler (KL) divergence is applied as the similarity metric between the source samples and the target samples. Partial source data that are highly similar to the target data are used to fine-tune the pretrained DL model.

In this article, we innovatively apply ITL to radar-based HMR and present the ITL method to tackle the issue of insufficient labeled radar data. Especially, ITL takes the advantage of partial prior knowledge that is learned from the sufficient motion data of the existing persons (source) to classify the motion data of new persons (target). Based on the general ITL idea, we select a part of the source data, instead of using merely target data, and adopt the target data and the selected source data to fine-tune a pretrained DL model.

Unlike the existing ITL approaches, we utilize a different similarity metric to compare the similarity between the source data and the target data. With that, a series of source samples are specially selected for every piece of target data. Furthermore, during the FT process, the selected source samples are assigned diverse importance by reweighting their training losses. In this way, the source samples with less domain discrepancy can contribute more to HMR in the target domain. Last but not least, we design a deep CNN model especially for radar-based HMR and use it as the backbone of ITL. The proposed CNN model has a simple structure and good performance for classifying radar spectrograms. With these improvements, the proposed ITL can recognize the human motions accurately when there is only limited data in the target domain. Furthermore, ITL is more generalized to the data distribution discrepancy and can scale well to recognize different persons' motions.

TABLE I
Basic Physical Information of the Six Subjects

	Sub ♯1	Sub ♯2	Sub ♯3	Sub ♯4	Sub ♯5	Sub ♯6
Age	23	25	23	23	23	24
Height (cm)	173	178	172	166	188	169
Weight (kg)	73	71	75	66	92	52

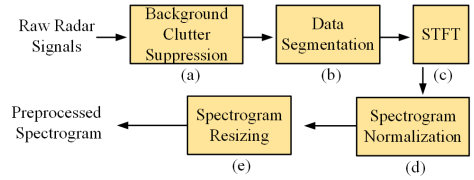

Fig. 2. Pipeline of radar raw signal preprocessing. (a) MTI is used on the raw radar signals for background clutter suppression. (b) Processed radar data is divided into several segments of 1 s with an overlap of 0.36 s between adjacent segments. (c) STFT is applied to transform the radar signals into 2-D complex data. (d) 2-D complex data are normalized, and the modules are obtained to form a spectrogram. (e) After resizing, the preprocessed spectrograms are obtained and used for further experiments.

III. Data Collection, Preprocessing, and Analysis

A. Data Collection

We utilize an ultrawideband (UWB) radar PulsON 440 for the experiments. PulsON 440 is composed of two antennas for transmitting and receiving C-band radio signals. The waveform generator generates chirp signals with a bandwidth of 1.8 GHz and a center frequency of 4.0 GHz. The UWB radar can distinguish the main scattering points of the human target due to its high range resolution. By accumulating the echo signals of multiple strong scattering points, the SNR of the received signals can be improved. Thereby, the target recognition ability of the UWB radar is enhanced.

The experiments are conducted in an indoor environment. The radar is set at the height of 1 m, and six motions (M1: running forward, M2: running in a circle, M3: jumping forward, M4: sitting on a chair, M5: walking forward, and M6: boxing in place) are performed by six persons in the line-of-sight of the radar with an aspect angle of 0°. All the subjects are limited to move within the range from 1.5 to 7.5 m. Each of the six motions is continuously performed by an individual for approximately 1.5 min. In each scenario, the process is repeated one to three times. The basic physical information of the six subjects is listed in Table I.

B. Data Preprocessing

In this study, we employ MD spectrograms as input to the network, treating HMR as a spectrogram classification problem. The radar data preprocessing process is illustrated in Fig. 2. First, the moving target indicator (MTI) is adopted on the raw radar echo signals to remove the static background clutter. Next, the processed radar data are divided into several segments of 1 s so that there is an approximately complete cycle of each of the six motions. The overlap between adjacent segments is 0.36 s.

Based on this, a 1024-point short-time Fourier transform (STFT) is used to process these data segments. Since the

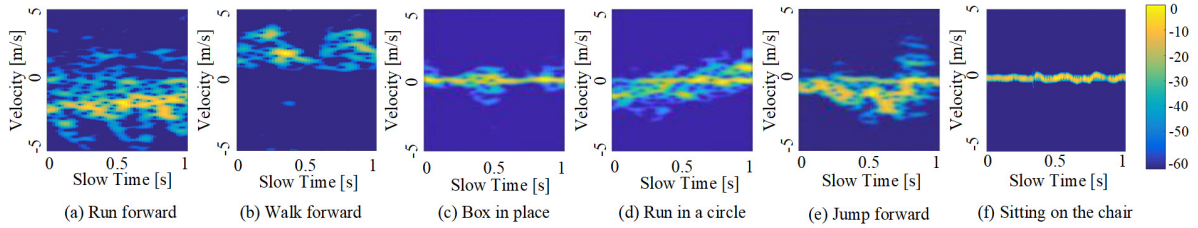

Fig. 3. Several typical MD spectrograms of human motions. (a) Run forward. (b) Walk forward. (c) Box in place. (d) Run in a circle. (e) Jump forward. (f) Sitting on the chair.

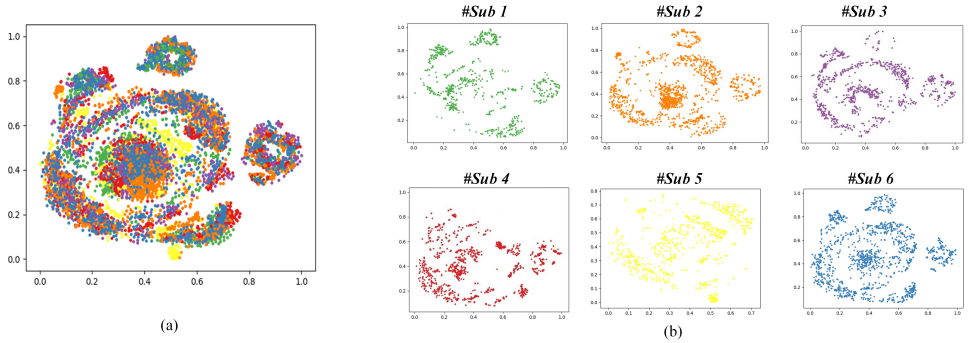

Fig. 4. Visualization results of the whole spectrogram data set with t-SNE. (a) Distribution of all motion data of the six persons. (b) Distributions of the six persons' motion data, separated for each individual.

human bodies are distributed targets, the scattered data from the bodies are spread over a few range cells. Thus, the STFT is performed on the radar data that are summed over several resolution cells. The obtained 2-D radar data after the STFT is still complex-valued and the modules of the 2-D complex data are utilized to form the spectrograms.

Then, we normalize all the spectrograms to make the values in the spectrograms fall into [0, 1]. Data normalization can prevent the value of a particular dimension from being too large. In this way, the convergence of DL models can be facilitated [27]. Finally, the spectrograms are resized into 150×150 pixels for further processing. In the radar MD data set, there are 300 spectrograms per person per motion. Several typical preprocessed spectrograms are shown in Fig. 3.

C. Data Analysis

Due to the human individual motion differences, different persons' motion data often have some discrepancy and are varied in distribution. In this circumstance, when a DL model that is trained with the motion data of several persons is directly applied to recognize the motions of new persons, the performance of the model often decreases.

To show the differences in the distributions of the six persons' motion data, we reduce these motion data to a series of 2-D vectors and visualize the dimensionality-reduced data with t-distributed stochastic neighbor embedding (t-SNE) [28]. The visualization results are shown in Fig. 4. It can be seen that, though related, the distributions of the six persons' motion data are different, indicating that there are individual motion differences between these persons. Furthermore, a quantitative similarity comparison between the six persons' motion data is also performed. In detail, we assume that the six persons' motion data follow independent multivariate Gaussian

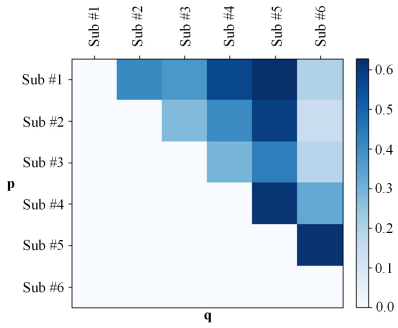

Fig. 5. KL divergence $\mathrm{KL}(p||q)$ between the motion data of one person to the others. p and q are the probability distributions of the motion data of any two of the six people. It can be seen that when the KL divergence between the two distributions is smaller, the data distributions are more similar.

distributions. Then, the KL divergence from the motion data distribution of one person to the others can be calculated. The KL divergence $\mathrm{KL}(p||q)$ is shown in Fig. 5, where p and q are the probability distributions of the motion data of any two of the six people. It can be seen that when the KL divergence between the two distributions is small, the similarity between the data is relatively high.

IV. Description of ITL

In this section, we introduce the algorithmic components of our proposed HMR approach, ITL, in detail.

A. Problem Formalization

Mathematically, the problem is described as follows. Let the source domain training data set $D_s = \{x_i^{(s)}, y_i^{(s)}\}_{i=1}^{N_s}$, where there are N_s data in the source domain. $x^{(s)} \in \mathbf{R}^{m \times n}$ denotes an $m \times n$ matrix corresponding to the radar MD signature

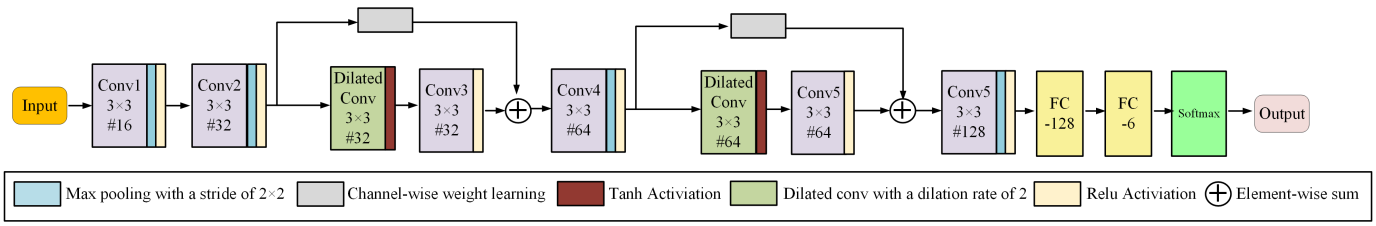

Fig. 6. Architecture of the proposed backbone (MNet) for HMR. The proposed MNet is composed of six convolutional layers, two dilated convolutional layers, two channelwise attention layers, and two fully connected layers.

of the human motions in the source domain. $y^{(s)}$ denotes the corresponding label of C_s categories. A source classification network $f_s(\cdot)$ is trained with D_s from scratch. Let the target data set $D_t = \{x_i^{(t)}, y_i^{(t)}\}_{i=1}^{N_t}$, where $x^{(t)} \in \mathbf{R}^{m\times n}$ denotes an $m \times n$ matrix corresponding to the radar MD signature in the target domain. The data in the target domain belong to the same C_t categories as those in the source domain. However, there is a distribution discrepancy between D_s and D_t, which makes $f_s(\cdot)$ not suitable to classify the target data. Furthermore, there are only a limited number of instances in D_t, which are insufficient to train a sufficiently generalized classification model.

Our goal is to train a target classification network $f_t(\cdot)$ to recognize the motions accurately in D_t when there is limited target training data. To this end, an instance-based deep transfer learning approach ITL is presented. The proposed ITL transfers the relevant knowledge from the sufficient motion data (source domain) as a supplement to classify the motion data in a new data set (target domain). The details of our proposed algorithm are summarized in Algorithm 1.

B. Structure of the Pretrained Deep Model

In this article, we design a deep neural network MNet for radar-based HMR and use it as the backbone of ITL. In radar spectrograms, each pixel of the spectrogram has both an intensity and a sample of time and frequency values, distinguishing it from optical images. Due to the unique properties of MD signatures, the proposed DL approach is designed to be more tailored to the radar data. The architecture of MNet is illustrated in Fig. 6.

As shown in this figure, convolutional layers, together with max-pooling, are the basic components of the network. Furthermore, to extract more discriminative features from the MD signatures, we apply the dilated convolution mechanism and the channelwise attention mechanism within MNet. Then, two fully connected layers are connected with the last convolutional layer sequentially. The softmax function is employed at the end of MNet to predict the labels of the input spectrograms.

1) Channelwise Attention: The channelwise attention mechanism enhances the network performance by accounting for the different importance that each feature channel has in the classification process. The more useful feature channels are weighted accordingly to emphasize their contribution and the other way round for less important feature channels [29]. By explicitly modeling the channel-interdependencies and recalibrating the features, the proposed network is more

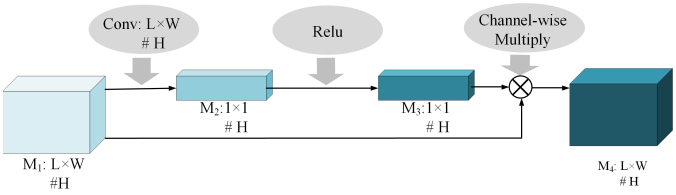

Fig. 7. Illustration of the channelwise attention mechanism. M_1 represents the input feature maps of $L \times W$ from H channels, and so, M_2, M_3, and M_4. *Conv* represents the convolution with H kernels of $L \times W$.

focused and oriented to pay more attention to the more informative data.

The process of channelwise attention mechanism is illustrated in Fig. 7. First, the feature maps M_1 from H channels are fed into a convolutional layer. In order to obtain the importance of every channel, the convolutional layer is designed with H kernels with the size of $L \times W$, which has the same size as the input feature maps. Hence, the output feature maps M_2 are H real numbers and have a global receptive field. Next, the 1×1 feature maps are excited with an activation function, and the output values (feature maps M_4) are treated as the weights of importance corresponding to these channels. Finally, the channel recalibration is completed by multiplying the weights with the original feature maps M_1 channel-by-channel. In this way, the original M_1 is transformed into the weighted feature maps. The channels with larger weights are paid more attention.

2) Dilated Convolution: In CNNs, pooling is utilized to decrease the redundancy of the feature maps and enlarge the receptive fields. The receptive field is the size of the activation area on the feature map during a convolution operation. However, pooling has many drawbacks, such as missing spatial information and small-object information. For example, when there are three pooling layers with a kernel of 2×2, the information of the objects smaller than 8×8 is lost.

To tackle this problem, dilated convolution [30] is adopted in this article. Instead of downsampling, dilated convolution is achieved by zero-padding on the convolution kernels, as shown in Fig. 8. This mechanism makes the dilated convolution able to increase the receptive fields without losing the structured information of data. The size of the receptive field is proportional to a parameter called dilation rate due to the number of zero-padding increasing as the dilation rate increases. When the dilation rate is set to 1, dilated convolution is equivalent to the conventional convolution. The dilated convolution

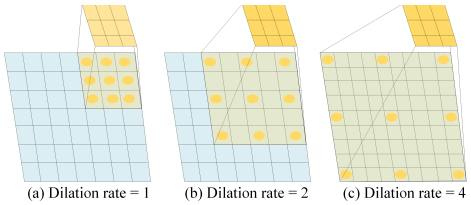

Fig. 8. Dilated Convolution with different dilation rates. The blue area is the input feature map, and the yellow area is the convolution kernel. The pale yellow area is the receptive field. The yellow dots are the pixels that are convolved with the convolution kernel.

operation is able to retain more useful information of the input without increasing the parameters of the network and is helpful to obtain more globally representative information about the original data.

C. CSDS

In this section, we propose a novel CSDS algorithm to select the most appropriate data for the collaborative FT. Instead of only using the target data to fine-tune the pretrained network, we make a partial selection of source data with high similarity to the target data and utilize them to fine-tune MNet, along with all target data.

1) MD Signature Descriptor: As efficient feature extractors, deep CNNs can learn the high-level semantic representation of the input data. The representation can be used to describe input data. In this article, we utilize AlexNet [32], a typical CNN for image classification, to further obtain the descriptors of the input MD signatures. It is noted that instead of AlexNet, many other CNNs, such as VGG-Net, residual network (ResNet), and Inception-Net, can also implement this function. We select AlexNet because it can extract semantic information effectively, and its structure is relatively simple. In detail, we treat the convolution kernels of the last convolutional layer in AlexNet as filters $\{F_0, F_1, F_2, \ldots, F_{255}\}$. All data in D_s and D_t are input to a pretrained AlexNet. Then, the feature maps output from the last convolutional layer is represented into histograms corresponding to the input MD signature. Let $M_i(x, y)$ denote the output feature map of the ith filter F_i and $\mathbf{h}_i$ its histogram, where $i = \{0, 1, 2, \ldots, 255\}$. In the beginning, the pixel value range of all histograms is set from 0 to 255, and the width of every histogram bin is set to 0.5.

To obtain more discriminative descriptors, we refine the histograms and avoid a large percentage of pixels falling into the same bin. Especially, we first obtain the maximum pixel value $p^u_{\max}$ and the minimum pixel value $p^{\min}_i$ of $M_i(x, y)$ by scanning the whole ith feature maps in the source data set D_s. Then, the pixel value range of $\mathbf{h}_i$ is set from $p^{\min}_i$ to $p^u_{\max}$. Furthermore, we iterate through the original $\mathbf{h}_i$'s of D_s and D_t and adaptively set the width of the histogram bins so that there is a roughly equal percentage of pixels in each bin. The percentage is set to 2% so that there are no more than 50 bins in every type of histograms $\{\mathbf{h}_i\}^{255}_{i=0}$. This setting makes a compromise between computing complexity and efficiency of the representation, which allows the further designed descriptor to have a proper dimension and to be discriminative at the same time.

Algorithm 1 ITL: An Instance-Based Transfer Learning Method for HMR With Limited Radar Data

Input:

Motion network *MNet*, a source data set $D_s = \{x^{(s)}_i, y^{(s)}_i\}^{N_s}_{i=1}$, a small-scale labeled target data set $D_t = \{x^{(t)}_i, y^{(t)}_i\}^{N_t}_{i=1}$, and number of epochs I_t.

Output:

The fine-tuned *MNet* for classifying unlabeled target data.

pretraining *MNet* with the source data set D_s, with the parameters initialized by *ImageNet* [31].
Obtaining the histogram descriptor H^k for each sample x in D_t and D_s.
For each spectrogram $x^{(t)}$ in D_t and all spectrograms $\{x^{(s)}_i\}^{N_s}_{i=1}$ in D_s, calculating the EMDs between $x^{(t)}$ and $\{x^{(s)}_i\}^{N_s}_{i=1}$ as Eq. 1.
$i \Leftarrow 0$.
For each spectrogram $x^{(t)}$ in D_t, selecting the top 100 correlated spectrograms in EMD similarity from D_s.
While not converged **or** $i < I_t$ **do** s
    Fine-tuning the parameters of *MNet* with the selected source spectrograms and all spectrograms in D_t.
    Calculating the loss weights of the selected source spectrograms.
    if $i \leq 5$ **then**
        Calculating the information entropy E for each spectrogram $x^{(t)}$ with Eq. 5.
        Increasing the number of the selected source spectrograms for each $x^{(t)}$ as Eq. 6.
        Re-weighting the selected source spectrograms.
    $i = i + 1$
return The fine-tuned parameters $w^{(m)}$, $b^{(m)}$ of *MNet* (m $= 1, 2, \ldots, M$ and M is the number of layers to be optimized in *MNet*).

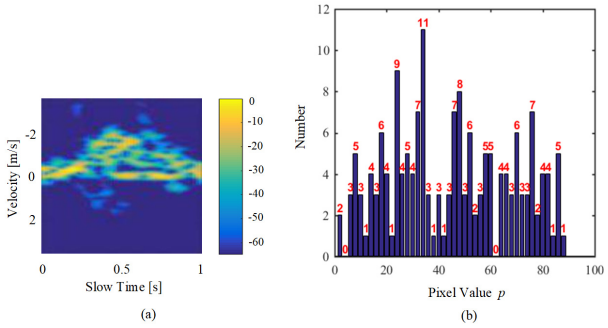

Fig. 9. Typical spectrogram and its histogram. (a) Radar spectrogram. (b) Histogram corresponding a specific convolution kernel in the last convolutional layer of AlexNet.

In this way, the inhomogeneous intervals are acquired, and the new ith histogram $\mathbf{h}'_i$ of a spectrogram is obtained. Fig. 9 illustrates the refined histograms of a radar spectrogram corresponding to a filter F_i. Finally, for the spectrogram x_k, the corresponding histograms $\{\mathbf{h}'_i\}^{255}_{i=0}$ are concatenated to form an MD signature descriptor, namely, $\mathbf{H}^k = \{\mathbf{h}_0, \mathbf{h}_1, \ldots, \mathbf{h}_{255}\}$.

2) Similarity Metrics of MD Signatures: The Earth mover's distance (EMD) [33] is the minimal cost that must be paid to transform one distribution into another distribution. It is

proposed based on the solution to a typical transportation problem but can be used to measure the distance between two generalized distributions irrespective of the underlying application. As a similarity metric of two histograms, EMD is more efficient than other possible histogram matching techniques due to its feasibility of operating on variable-length representations of the distributions.

A histogram can be formulated as a set $S = \{s_j = (w_j, m_j)\}_{j=1}^{N}$, where the histogram values are denoted as the weights w_j, and the indices of bins are denoted as positions m_j. N denotes the number of bins in the histogram. Given two histograms $P = \{(p_i, u_i)\}_{i=1}^{m}$ and $Q = \{(q_i, v_i)\}_{i=1}^{n}$, with size m, n, respectively, the EMD of P and Q is defined as the minimum work required to resolve the supply–demand transports, namely

$$\mathrm{EMD}(P, Q) = \min_{F=\{f_{ij}\}} \frac{\sum_{i,j} f_{ij} d_{ij}}{\sum_{i,j} f_{ij}} \tag{1}$$

with the constrains

$$\sum_j f_{ij} \le p_i, \quad \sum_i f_{ij} \le q_j$$
$$\sum_{i,j} f_{ij} = \min\left\{\sum_i p_i, \sum_j q_j\right\}, \quad f_{ij} \ge 0 \tag{2}$$

where p_i represents the histogram values of P and q_i represents the histogram values of Q. u_i and v_i represent the indices of bins of P and Q, respectively. Furthermore, $F = \{f_{ij}\}$ denotes a flow set. Each flow f_{ij} represents the amount transported from the ith supply to the jth demand. d_{ij} denotes the distance between the position u_i and v_j.

In this article, EMD is employed to measure the similarity of the histogram descriptors $\mathbf{H}$'s corresponding to the MD signatures in D_t and D_s. Given the histogram descriptors $\mathbf{H}$'s of a specific target spectrogram $x^{(t)}$ and all source spectrograms $\{x_i^{(s)}\}_{i=1}^{N_s}$, the EMDs $\{\mathrm{EMD}_i\}_{i=1}^{N_s}$ of $x^{(t)}$ and $\{x_i^{(s)}\}_{i=1}^{N_s}$ are calculated with $\mathbf{H}^{(t)}$ and $\{\mathbf{H}_i^{(s)}\}_{i=1}^{N_s}$ according to (1) and (2). A small EMD value between $x^{(t)}$ and $x^{(i)}$ means that these two spectrograms are highly correlated. Furthermore, the source radar spectrograms with smaller EMD values are preferred as the correlated instance of $x^{(t)}$. In detail, the source spectrograms are ranked in ascending order based on their EMD values. For each target spectrogram, K source spectrograms (corresponding to the top 2.0% of the whole set of source spectrograms) are chosen as the most correlated set at the outset, based on the EMD metric. This ensures that the initially selected spectrograms are the most similar to the target spectrogram. Then, these instances are utilized to fine-tune the pretrained MNet along with all the target data.

D. ACFT

1) Source Instances Reweighting: During FT the pretrained MNet, there are a series of source spectrograms selected as the correlated instances by more than one target spectrogram. Compared with treating them equally, attaching more importance on the source instances that are selected more than once can make the FT process more efficient. Thus, the importance of the selected source instances differs. Especially, we put the importance to the loss function and reweight the FT loss of the selected source instances. Suppose that, in an epoch, the ith source instance $x_i^{(s)}$ is selected as the correlated instance by w target spectrograms. Then, we design the loss function L of FT as follows:

$$L = \sum_i L_{\mathrm{cls}}\left(y_i^{(t)}, \hat{y}_i^{(t)}\right) + \sum_j \sin\left(\frac{\pi}{2} * \frac{w_j}{w_{\max}}\right) * L_{\mathrm{cls}}\left(y_j^{(s)}, \hat{y}_j^{(s)}\right) \tag{3}$$

where w_j denotes the number of target instances that select $x_j^{(s)}$ as its correlated source instance. $w_{\max}$ denotes the maximum among w's corresponding to all the selected source instances in an epoch. $y_i^{(t)}$ and $\hat{y}_i^{(t)}$ are the true label and the predicted label of the ith target instance $x_i^{(t)}$, respectively. Similarly, $y_j^{(s)}$ and $\hat{y}_j^{(s)}$ are the true label and the predicted label of the jth source instance $x_j^{(s)}$. The classification loss L_{cls} adopts the cross-entropy loss, whose definition is given as follows:

$$L_{\mathrm{cls}} = -[p \log(\hat{p}) + (1 - p) \log(1 - \hat{p})] \tag{4}$$

where p and $\hat{p}$ are the ground-truth one-hot label and the predicted probability, respectively.

2) Adaptive Source Data Search: Subsequently, we present the adaptive searching scheme to employ more nearest source spectrograms in the following FT epochs, which is able to facilitate the target spectrograms classification. We calculate the information entropy E_i^m to measure the classification uncertainty of the target training sample $x_i^{(t)}$ after the mth epoch

$$E_i^m = -\sum_{c=1}^{C_t} p_{i,c}^m \log\left(p_{i,c}^m\right) \tag{5}$$

where C_t is the number of motion categories in D_t, and $p_{i,c}^m$ is the probability that $x_i^{(t)}$ is classified as the cth class by the softmax layer of MNet in the mth epoch. The larger E_i^m, the higher the classification uncertainty of $x_i^{(t)}$. We set the threshold θ for the classification uncertainty E. When E_i^m is larger than θ, we increase the number of correlated source samples for $x_i^{(t)}$ in the next epoch.

Furthermore, we stop the adaptive searching scheme after five FT epochs because too many epochs can lead to more source instances that are not highly correlated with the target data employed in the FT process. The overall adaptive source data search is then given as

$$N_i^{m+1} = \begin{cases} N_i^m + \alpha, & m \le 5 \text{ and } \hat{y}_i^{(t)} \ne y_i^{(t)} \\ N_i^m + \beta, & m \le 5 \text{ and } \hat{y}_i^{(t)} = y_i^{(t)} \\ & \quad \text{and } E_i^m \ge \theta \\ N_i^m, & \text{others} \end{cases} \tag{6}$$

where N_i^m and N_i^{m+1} are the numbers of the selected nearest source samples for $x_i^{(t)}$ in the mth and $m + 1$th epochs, respectively. $N_i^1 = K$. $\hat{y}_i^{(t)}$ and $y_i^{(t)}$ are the predicted label and the true label of $x_i^{(t)}$, respectively. α and β are set to

$K/2$ and $K/4$. θ is set to 0.25 empirically so that every target spectrogram tends to be classified into a particular category with high probability.

V. Experimental Implementation and Results

A. Evaluation Methodology

In the experiments, a leave-two-individual-out cross-validation method is adopted to split the data set into two parts: the source data set and the target data set. Especially, we randomly select the motion data of $(n-2)$ persons as the source data set D_s, where n equals 6 in the experiments. The data of the other two persons are utilized as the target data set D_t. Hence, the process is repeated $\binom{n}{2} = 15$ times to obtain the average performance. Since it is infeasible to perform thousands of trials to get a statistical characterization of the experimental results, we assume that the leave-two-individual-out cross-validation can approximate the statistical results. Furthermore, with the leave-two-individual-out cross-validation, the generalization to human motion differences of ITL can be demonstrated well.

To evaluate the efficiency of ITL with limited training data, the motion data per person per class in the source domain are divided for training and validation according to the ratio of 8:2. The target data set is also divided according to the ratio of 2:1 in the same way. Furthermore, we randomly select N instances per person per class from the target training set for FT and evaluate the classification performance of ITL on the target validation data set.

B. Implementation Details

We employ Tensorflow [34] that is a widely used DL framework developed by Google Brain to train our model. The proposed MNet is pretrained from scratch with D_s. The batch size is set to 32, and the learning rate λ_1 is set to 10^{-3}. The model is pretrained for 400 epochs, and L2 normalization is employed during the training process. For each human individual, 70% of the motion spectrograms in D_t are selected for FT and the others for validation. During FT, the basic learning rate λ_2 is set to 10^{-5}, and an exponentially learning rate decay γ is set to 0.9. The model is fine-tuned for 50 epochs. The batch size is also set to 32. All experiments are performed on a CPU and Ti 1080 GPUs with CUDA for acceleration.

C. Comparison Methods

To further investigate the performance of ITL, we compare the model with several state-of-the-art transfer learning methods, including two radar-based transfer learning approaches and three typical instance-based approaches designed for optical image classification. These comparison approaches are also implemented with the data set that is described in Section III.

DivNet [19] is specially designed for radar-based HMR with radar MD spectrograms. The network is pretrained with diverse Kinect-based simulated motion data and fine-tuned with a limited number of measured radar data.

DuNet [17] is presented for radar-based HMR. The ResNet is adopted as the backbone of the method. The prior knowledge from simulated MOCAP radar data is transferred by FT the pretrained backbone with the limited target samples.

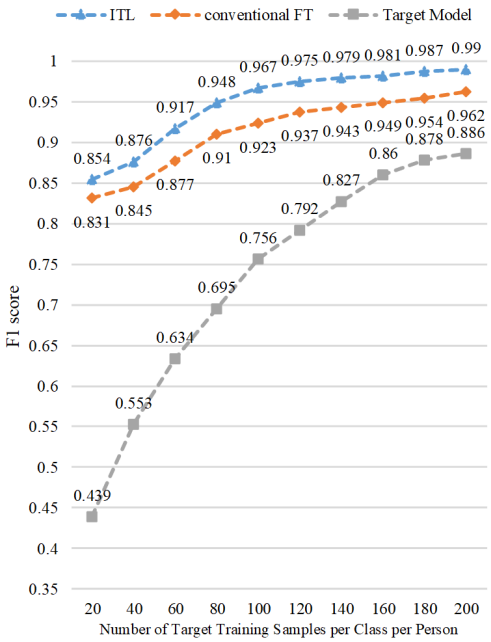

Fig. 10. F1 score performance of ITL, the Conventional FT, and the Target Model for classifying the target validation data when diverse amounts of target samples are used for training.

NgiamNet [22] is an ITL approach. In this method, the source data are first reweighted based on their similarity to the target data. A DL backbone is first pretrained with the reweighted source data.

GeNet [21] is an ITL approach. The source data are reweighted based on a similarity metric between the source data and the target data. Then, a pretrained backbone model is fine-tuned with the reweighted source data and all target data.

AsgarianNet [35] is another ITL approach that reweights the source data and uses them for FT. Distinctively, this approach proposes the hybrid weight for source data, which measures the similarity of a source sample to the target domain and the importance of the sample in the target task.

D. Experimental Results

1) Performance With Limited Numbers of Target Samples: To evaluate the performance of the proposed ITL for HMR with limited training data, different amounts of target samples per class per person are provided for training ITL. The experimental results are shown in Fig. 10. Furthermore, to demonstrate the efficiency of ITL, we select two baseline methods for comparison. In detail, we train the proposed backbone model MNet from scratch with the limited target samples, and the test F1 scores are shown with gray marks in Fig. 10. Then, a Conventional FT method that utilizes target samples to fine-tune a pretrained model is adopted for comparison. Especially, the MNet is pretrained on the source training data set and fine-tuned with the available target samples. The results are shown with orange marks in Fig. 10.

From Fig. 10, we can see that the proposed ITL yields the best performance among the three methods. Especially,

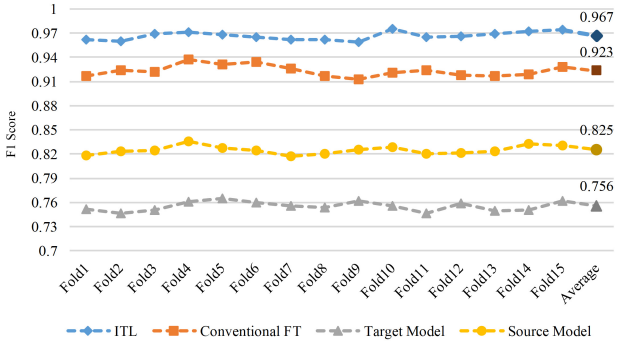

Fig. 11. Results of the leave two-individual-out cross-validation when there are 100 samples per class per person.

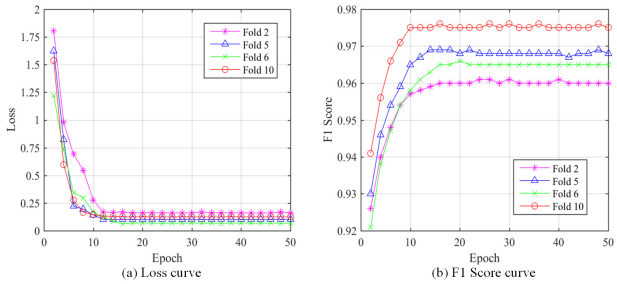

Fig. 12. Loss curves and F1 Score curves of Fold 2, Fold 5, Fold 6, and Fold 10.

when there are 100 target samples per class available, ITL outperforms the Conventional FT by the largest margin of 4.4% F1 score. In detail, since there are insufficient samples for training MNet from scratch, the Target Model is susceptible to overfitting, and the performance is the worst. Furthermore, the performance of the Conventional FT and ITL improves as the number of the target samples increases. ITL outperforms the Conventional FT all the time, demonstrating its better performance for the classification task with limited training data. Since ITL has an obvious advantage over the Conventional FT in the F1 score when there are 100 target samples per person per class, we select 100 target training samples as the typical setting in the following experiments.

2) Cross-Validation Performance With 100 Target Samples per Person per Class: The results of the leave-two-individual-out cross-validation when there are 100 target samples per class per person are further shown in Fig. 11.

It can be seen that the performance of ITL during the two-individual-out cross-validation is steady. Regardless of the difference between the source domain data and the target domain data, the method can achieve an F1 score of about 96.7%. The standard deviation of the 15 folds in the F1 score is merely 4.88e-3. In detail, the average F1 scores of 96.7%, 92.3%, and 75.6% are achieved by ITL, the Conventional FT, and the Target Model, respectively.

Besides, to demonstrate the infeasibility of directly using the pretrained MNet to classify the target samples, we introduce another baseline model Source Model. The Source Model is obtained by training the backbone MNet with the whole source data set, and no FT is involved. It can be found that Source Model achieves an average F1 score of merely 82.5% for classifying the target samples, indicating that there are some differences between the source data and the target data.

Furthermore, we select fourfolds (Fold 2, Fold 5, Fold 6, and Fold 10) from the 15-fold two-individual-out validation experiments without any adjective, and their convergence properties are shown in Fig. 12 in detail. From the loss curves and the F1 score curves, we can see that ITL often begins to converge after 10 epochs and yields a stable performance after 30 epochs.

E. Analysis on Generalization of ITL

Generally speaking, the Conventional FT method often forgets how to perform the source task as training the new target task progresses. As a result, the DL model fine-tuned with new persons' motion data often cannot achieve good performance to recognize the persons' motions in the previous (source) domain. In contrast, the proposed ITL is more generalized to the human motion differences, and the fine-tuned model can also scale well to the persons' motions in the source domain.

1) Generalization of ITL With Diverse Numbers of Target Samples: First, as shown in the leave two-individual-out cross-validation results of Fig. 11, ITL has good performance to recognize the persons' motions in the target domain.

Especially, ITL is first pretrained with the source data and fine-tuned with varying numbers of target samples for the task on the target domain. Then, the source validation samples are employed for classification to test the generalization ability of ITL. The experimental results are shown in Fig. 13. For comparison, the Conventional FT model that is fine-tuned with different amounts of target data is also employed to classify the source samples. The Source Model is trained with the source training samples, and no FT is involved.

As shown in Fig. 13, since the Source Model is trained with the source data, the performance of classifying the source validation samples is the best. In contrast, the performance of the Conventional FT is poor. In addition, the performance of the Conventional FT decreases with the increase in the number of target samples. It is because with the amount of the target samples increasing, the distribution of the available target samples tends to be closer and closer to the real distribution of the target domain data, which is different from that of the source data. In this circumstance, when the fine-tuned model performs well on the target domain, its performance for the source domain usually drops.

In contrast, no matter how many target samples are available for FT, the performance of ITL is better than that of the Conventional FT. Especially, when there are more than 100 samples per person per class, the performance of classifying the source samples is about 90.0% F1 score, exceeding that of the Conventional FT by over 11.0%. Furthermore, when there are more than 160 target samples per class for FT, the F1 score of ITL is only about 7.0% lower than that of the Source Model. Good performance is achieved because, in ITL, some source samples that are highly correlated with the target samples are selected for collaborative FT. As a result, ITL can be adapted to the new target task while retaining partial source knowledge. This property makes ITL generalized to the motion differences between different domains and scales well to the persons' motions in both the source and target domains.

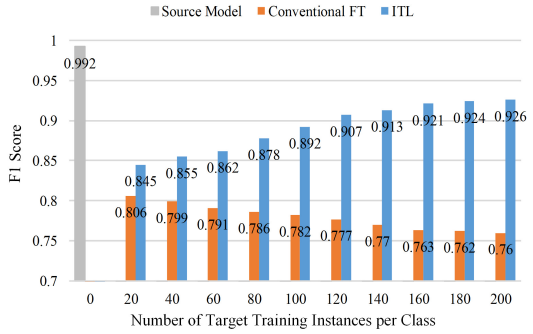

Fig. 13. Performance of Source Model, Conventional FT, and the proposed ITL for classifying the validation source samples. Among the three methods, the Source Model is trained with the source data set. The Conventional FT and ITL are pretrained with the source data and fine-tuned with different numbers of target samples. It can be seen that, though worse than the performance of Source Model, the performance of ITL trained with varying numbers of target samples is better than that of Conventional FT. It is demonstrated that compared with Conventional FT, the proposed ITL is more generalized to the human motion differences between the source domain and the target domain.

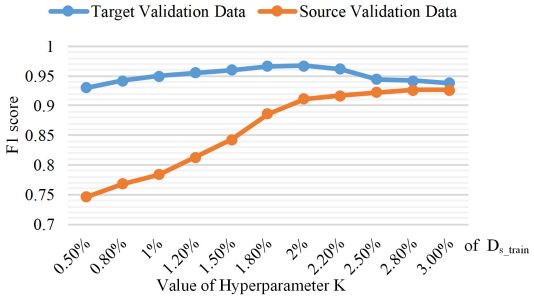

Fig. 14. Performance variation of ITL to diverse values of K when there are 100 target samples available for FT. The blue marks represent the results of using ITL to classify the target data. Orange marks: results of using ITL to classify the source data.

2) Impact of the Value of K on the Performance of ITL: According to (6), a certain amount of source samples is selected for every target training sample to perform the collaborative FT. Based on this setting, we change the number of the selected source data by adjusting the value of hyperparameter K and explore the impact of K on the generalization ability of *ITL*. The experimental results are shown in Fig. 14.

From this figure, we can find that, with the value of K increasing, the performance of ITL for classifying the source validation samples improves. It is because with more source training samples that are similar to the target data are selected and involved in the FT process, ITL can preserve more knowledge of the source domain while performing well on the target domain. At the same time, the performance for classifying the target samples improves with K increasing when K is less than 2.0%. However, when K is more than 2.0%, a decreasing trend is shown. It is because with the value of K increasing, more source samples that are not highly correlated are selected, which is of little help to the task of the target domain.

Furthermore, when K is between 2.0% and 2.5%, the F1 score of ITL for classifying the target data is over 94.0%, and the performance of classifying the source samples is over 90.0%. Thus, we can conclude that, when K is set between 2.0% and 2.5%, ITL is generalized to the differences of the source and the target domains and can scale well to recognize the motions of diverse persons. In further experiments, we set K to 2.0%.

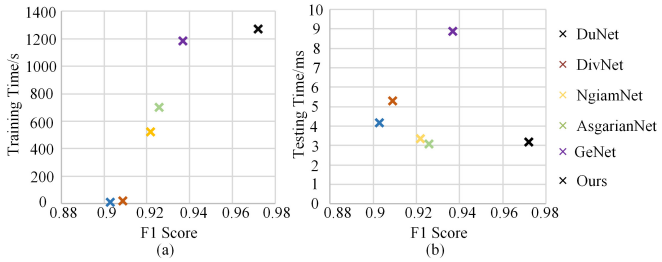

Fig. 15. Comparison in terms of computational time and F1 score for different methods. In detail, (a) depicts the training time and the F1 scores of the six approaches and (b) depicts the testing time per sample and the F1 scores of the six approaches.

F. Comparison With the State of the Art

To verify the efficacy of ITL, we compare it with several state-of-the-art transfer learning approaches. Especially, we vary the number of the target training samples and perform the leave-two-individual cross-validation on these methods. Then, the average F1 scores are obtained for comparison. The results are depicted in Table II.

Comparison in F1 Score: From Table II, we can find that our proposed ITL obtains the best performance when there are more than 20 target samples per class for FT, indicating the feasibility of ITL for radar-based HMR with limited data. Though the performance of ITL is not the best when there are 20 target samples, the F1 score of ITL is merely 0.4% lower than that of GeNet, which yields the best performance.

Comparison in the Number of Parameters: The number of parameters in these DL models is listed in the last column of Table II, reflecting the spatial complexity of these methods. It can be seen that GeNet has the most parameters due to the complicated backbone. As for our method, with the effective but relatively simple structure of MNet, there are only 22.88M parameters in ITL.

Comparison in Training/Testing Time: Fig. 15 illustrates the training time and the testing time of the six methods when there are 100 target training samples per person per class. In detail, in Fig. 15(a), the model training time and the F1 scores of the six approaches are shown. It can be seen that the training time of DuNet and DivNet is much shorter than the other methods. The reason is mainly due to that the two methods only use only the target samples to fine-tune their backbone models. As a result, the similarity between the source domain and the target domain is not required to be calculated. In this way, their training time is greatly shortened.

As for the other four instance-based methods, NgiamNet, AsgarianNet, GeNet, and our approach, the training time includes two parts: the time of calculating the similarity of the source data and the target data and the time of FT the DL backbone model. Due to the operation of the similarity calculation, the training time of the four methods is much longer than that of DuNet and DivNet. Among the four methods, the training time of ITL is the longest since this method requires to select the correlated source samples for

TABLE II
COMPARISON WITH THE STATE-OF-THE-ART METHODS IN F1 SCORE

	♯Target Training Instances										♯Param
	20	**40**	**60**	**80**	**100**	**120**	**140**	**160**	**180**	**200**	
DuNet	0.866	0.876	0.881	0.894	0.903	0.908	0.914	0.919	0.926	0.931	33.16M
DivNet	0.863	0.867	0.882	0.898	0.909	0.911	0.915	0.924	0.930	0.941	54.52M
NgiamNet	0.858	0.864	0.883	0.911	0.922	0.934	0.941	0.940	0.946	0.952	24.73M
GeNet	**0.865**	0.881	0.904	0.926	0.937	0.940	0.945	0.952	0.954	0.959	119.69M
AsgarianNet	0.854	0.873	0.905	0.919	0.926	0.938	0.941	0.947	0.950	0.953	22.55M
Ours	0.861	**0.882**	**0.921**	**0.951**	**0.967**	**0.972**	**0.974**	**0.980**	**0.984**	**0.988**	22.88M

[1] '♯Target Training Instances' denotes the number of target instances per motion that are used for fine-tuning.
[2] '♯Param' denotes the number of parameters in the DL models.

every target sample. However, though it takes more training time, ITL has gained a performance boost and yields the highest F1 score.

Furthermore, Fig. 15(b) shows the comparison results in terms of testing time per sample and the F1 score for the six methods. In general, the training process of a model is often performed offline. Compared with the training time, the testing time per sample has a greater impact on whether the model can be applied in practice. During testing, the similarity calculation operation is not required, and the running time is greatly shortened. As shown in this subfigure, though the training time is long, the proposed ITL takes a short time to classify a sample.

VI. ABLATION STUDY ON *ITL*

To better prove the effectiveness of ITL for the HMR task with limited training data, some ablation studies on ITL are performed. During the ablation study, we performed the experiments under the typical setting where there are 100 target samples per person per class available. The 15-fold cross-validation is employed to obtain the average F1 score.

A. Ablation Study on MNet

To demonstrate the good performance of MNet for recognizing human motions with radar MD spectrograms, we change the structure of MNet slightly. Three variants of MNet are designed, which are referred to as MNet-v1 and MNet-v2, respectively. Especially, to obtain MNet-v1, the two channelwise attention modules in MNet are removed. In MNet-v2, the dilation rate in the two dilated convolutional layers is set to 1, and the dilated convolution operations are converted into the general convolutions.

Then, we compare the performance of the two transfer learning methods (the Conventional FT and ITL) when using MNet, MNet-v1, and MNet-v2 as the backbone, respectively. The comparison results in the F1 score are listed in Table III. As shown in this table, when the channelwise attention modules are removed, the performance of both the Conventional FT and ITL decreases, indicating that the two channelwise attention modules are vital to the performance of MNet. At the same time, the performance of MNet-v2 is not as good as that of MNet regardless of whether the source data are fine-tuned or not, demonstrating the efficiency of dilated convolution operations.

TABLE III
PERFORMANCE COMPARISON WITH OTHER DEEP MODELS AS BACKBONE

	Conventional FT	***ITL***	**Difference**
MNet-v1	0.914 ±0.03	0.949 ± 0.03	+0.035
MNet-v2	0.913 ± 0.01	0.950 ± 0.01	+0.037
MNet	**0.923** ± 0.01	**0.967 ± 0.02**	+0.038
VGG16	0.918 ± 0.02	0.962 ± 0.01	**+0.044**
ResNet10	0.921 ± 0.03	0.959 ± 0.02	+0.036
Inception-v3	0.905 ± 0.02	0.944 ± 0.01	+0.039

Furthermore, to demonstrate the superiority of MNet for radar-based HMR, we replace MNet with several typical CNN models, including VGG16, ResNet10, and Inception-v3. We select these three DL models among the existing state-of-the-art models for comparison because they have a similar number of convolution layers to MNet. Then, their performance as the backbone of Conventional FT method and ITL is compared. The comparison results are listed in Table III. We can see that regardless of whether the source data are used for FT or not, our model achieves the best results and is more suitable for the HMR tasks with radar MD spectrograms than the other DL models. Furthermore, when using VGG16 as the backbone, ITL outperforms the Conventional FT the most, with a difference of 0.44% F1 score. In addition, VGG16 yields similar performance to MNet when used as the backbone of ITL, with an F1 score of 96.2%. It is indicated that compared with ResNet10 and Inception-v3, VGG16 is more suitable to transfer the motion characteristic in radar spectrograms.

B. Ablation Study on CSDS

1) Analysis on the MD Signature Descriptor: First, we replace AlexNet with three typical CNNs VGG16, ResNet18, and Inception-v3 as the feature extractor and utilize the last convolutional layers of these models as filters to obtain the MD signature descriptors. Furthermore, the optical image data set ImageNet, instead of a radar image data set, is utilized to train the feature extractor. Though radar spectrograms have different characteristics from optical images, several

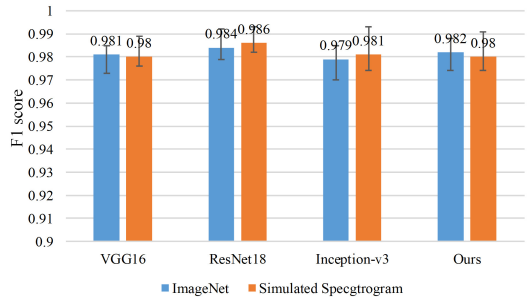

Fig. 16. Performance of ITL in average F1 score when using different deep models that are pretrained on ImageNet and a simulated radar data set, respectively.

radar-based HMR literature demonstrated the feasibility of extracting features from radar spectrograms with a model that is trained on a large-scale optical image data set [17], [18], [36], [37]. Furthermore, to make our work more complete and comprehensive, we utilize a simulated MOCAP radar data set [38] instead of ImageNet to train the feature extractor. The results are illustrated in Fig. 16.

As we can see in Fig. 16, the performance of using the four deep models to obtain the MD signature descriptors is similar, demonstrating the feasibility of applying these typical deep models as the feature extractor of ITL. However, despite the good performance, the complexity of these models is to obtain MD signature descriptors due to different numbers of kernels in the last convolutional layers. In this circumstance, by balancing complexity and F1 score, we can conclude that AlexNet is a better choice. In addition, the performance of using simulated radar spectrograms and ImageNet is broadly similar. Though the simulated radar data are more similar to our measured data, there is no noticeable performance advantage. However, as the error lines in the figure show, the maximum F1 score when using simulated data to train the feature extractor is often more significant than the maximum F1 score when using ImageNet. Furthermore, the performance of ResNet18 that is trained with simulated radar data is the best, with an average F1 score of 97.1%. Based on these results, we have reasons to believe that using simulated radar data to train the feature extractor has more potential to achieve good performance for HMR [19], [39].

2) Analysis on Source Instance Selection: To demonstrate the efficiency of the EMD-based source instance selection algorithm, we compare this solution with the other three source instance selection solutions.

*S*1: FT the pretrained MNet with only target data set D_t (Conventional FT).
*S*2: FT the pretrained MNet with the whole source data set D_s and the target data set D_t.
*S*3: FT the pretrained MNet with randomly selected source data and the whole D_t.

The comparison results are listed in Table IV. As shown in the table, the proposed EMD-based source instance selection algorithm yields the best performance. Especially, the F1 score of the Conventional FT (**S**1) is 92.3%. The F1 score of FT with both D_s and D_t (**S**2) is 91.8%. The F1 score of FT with the randomly selected source data and D_t (**S**3) is 91.5%. It can be seen that the performance of **S**2 and **S**3 is not improved and even worse than **S**1. It is because using the whole source data set or using randomly selected source data for FT can bring some negative knowledge transfer to the network. In contrast, the EMD-based source instance selection algorithm achieves the best performance, outperforming the solution **S**1 by 4.4% F1 score.

TABLE IV
CLASSIFICATION RESULTS ON THE TARGET VALIDATION DATA SET WITH DIFFERENT FT ALGORITHMS

Method	Transfer Performance
S1: Fine-tuning with D_t only	92.3%
S2: Fine-tuning with D_t and the whole D_s	91.8%
S3: Fine-tuning with D_t and the randomly selected source data	91.5%
S4: Fine-tuning with D_t and the source data selected with EMD-based algorithm *(Ours)*	**96.7%**

TABLE V
COMPARISON STUDY FOR THE PROPOSED ACFT ALGORITHM

Method	Transfer Performance
C1: Assigning equal instance importance	93.8%
C2: Neglecting the limitation of E	95.7%
C3: Reconstruction loss function	96.0%
C4: The proposed ACFT algorithm (*Ours*)	**96.7%**

C. Ablation Study on ACFT

1) Analysis on the Comparison Experiments: To investigate the effect of the two elements (adaptive source data search and source instance reweighting) in ACFT on the performance of ITL, we perform the three following comparison experiments.

*C*1: Assigning equal importance to the selected source instances, and setting the same loss weight to all of the instances in (3).
*C*2: In each of the first five epochs, if $\hat{y}_i^{(t)} = y_i^{(t)}$, β correlated source instances are selected for each target instance, without the limitation of information entropy E.
*C*3: Replace the loss weights $\sin((\pi/2)*(w_i/w_{\max}))$ of source samples in (3) with $(\pi/2)*(w_i/w_{\max})$.

Table V shows the comparison results. It can be seen that our method yields the best performance. When assigning equal importance on the selected source instances ($C1$), an F1 score of 93.8% is yielded, which is 2.9% lower than that of using the proposed ACFT algorithm. When selecting the same number β of correlated source instances for each target instance if $\hat{y}_i^{(t)} = y_i^{(t)}$ ($C2$), the performance drops to 95.7%. Furthermore, when replacing $\sin((\pi/2)*(w_i/w_{\max}))$ with $(\pi/2)*(w_i/w_{\max})$ in the loss function L ($C3$), the performance of ITL decreases to 96.0%. Though using sine function is the result of heuristic attempts, the comparison results demonstrate the efficiency of using sine instead of linear loss weights.

2) Visualization of Diverse Importance of the Source Samples: To reveal the diverse importance of the selected source

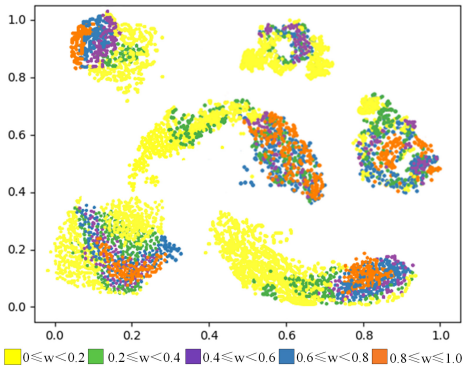

Fig. 17. Visualization of the loss weights w assigned to the source instances. A larger loss weight represents that the corresponding source sample is attached with more importance, which means that it is more important to the motion recognition task in the target domain.

instances, we visualize the loss weights assigned to the source data with t-SNE. In particular, the loss weights of the selected source samples in the fifth FT epoch are recorded when there are 100 target samples per person per class available for training. For those source training samples that are not used for FT, the loss weights are 0. Then, all of the source training samples are input to the Source Model, and the feature vectors output by the last convolutional layer is visualized with t-SNE. The visualization results are shown in Fig. 17. A larger loss weight means that the source instance is attached more importance for the HMR task, while a smaller weight means that the instance is less correlated with the target data and is less important to the collaborative FT process. From Fig. 17, we can see that in ITL, only partial source domain data are helpful for the classification task of the target domain, and selected for the ACFT process.

VII. Conclusion

To address the performance limitation of the DL algorithms caused by insufficient training data, this article proposed an ITL approach for radar-based HMR. The approach is composed of three interconnected and necessary parts (MNet pretraining, CSDS, and ACFT) rather than a collection of three distinct pieces. Measured human motion data were collected using a pulsed UWB radar. Six human subjects, each performing six motions, were involved in this study.

Experimental results showed that the proposed ITL was able to accurately recognize human motions with limited radar data, with an F1 score of 96.7% when there are only 100 samples per person per class. Moreover, though ITL is adopted to the recognition task in the target domain, it is generalized to human motion differences and can scale well to recognize the persons' motions in the source domain simultaneously. Furthermore, this proposed approach outperformed several state-of-the-art transfer learning methods when there is a limited number of training data. In addition, some ablation studies were conducted to demonstrate the uniqueness of the components in ITL. Any exclusion of these components resulted in performance degradation. Finally, despite the effectiveness of ITL, how to reduce the computational cost of the model needs to be further researched.

References

[1] Q. Wu, Y. D. Zhang, W. Tao, and M. G. Amin, "Radar-based fall detection based on Doppler time–frequency signatures for assisted living," *IET Radar, Sonar Navigat.*, vol. 9, no. 2, pp. 164–172, 2015.
[2] G. Gennarelli, F. Soldovieri, and M. Amin, "Radar for indoor surveillance: State of art and perspectives," *Proc. SPIE*, vol. 11059, Jun. 2019, Art. no. 1105903.
[3] A. Jalal, Y.-H. Kim, Y.-J. Kim, S. Kamal, and D. Kim, "Robust human activity recognition from depth video using spatiotemporal multi-fused features," *Pattern Recognit.*, vol. 61, pp. 295–308, Jan. 2017.
[4] X. Li, Y. He, and X. Jing, "A survey of deep learning-based human activity recognition in radar," *Remote Sens.*, vol. 11, no. 9, p. 1068, May 2019.
[5] Y. He, X. Li, and X. Jing, "A mutiscale residual attention network for multitask learning of human activity using radar micro-Doppler signatures," *Remote Sens.*, vol. 11, no. 21, p. 2584, Nov. 2019.
[6] Y. Yang, C. Hou, Y. Lang, D. Guan, D. Huang, and J. Xu, "Open-set human activity recognition based on micro-Doppler signatures," *Pattern Recognit.*, vol. 85, pp. 60–69, Jan. 2019.
[7] H. Du, T. Jin, Y. He, Y. Song, and Y. Dai, "Segmented convolutional gated recurrent neural networks for human activity recognition in ultra-wideband radar," *Neurocomputing*, vol. 396, pp. 451–464, Jul. 2020.
[8] 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.
[9] Y. Yang, C. Hou, Y. Lang, T. Sakamoto, Y. He, and W. Xiang, "Omnidirectional motion classification with monostatic radar system using micro-Doppler signatures," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3574–3587, May 2020.
[10] Z. Lin, K. Ji, M. Kang, X. Leng, and H. Zou, "Deep convolutional highway unit network for SAR target classification with limited labeled training data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 7, pp. 1091–1095, Jul. 2017.
[11] F. Deng, S. Pu, X. Chen, Y. Shi, T. Yuan, and S. Pu, "Hyperspectral image classification with capsule network using limited training samples," *Sensors*, vol. 18, no. 9, p. 3153, Sep. 2018.
[12] S. Tian, C. Wang, H. Zhang, and B. Bhanu, "SAR object classification using the DAE with a modified triplet restriction," *IET Radar, Sonar Navigat.*, vol. 13, no. 7, pp. 1081–1091, Jul. 2019.
[13] X. Zhang, Z. Wang, D. Liu, and Q. Ling, "DADA: Deep adversarial data augmentation for extremely low data regime classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 2807–2811.
[14] A. Davari, H. C. Ozkan, A. Maier, and C. Riess, "Fast and efficient limited data hyperspectral remote sensing image classification via GMM-based synthetic samples," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2107–2120, Jul. 2019.
[15] B. Erol, S. Z. Gurbuz, and M. G. Amin, "Motion classification using kinematically sifted ACGAN-synthesized radar micro-Doppler signatures," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 56, no. 4, pp. 3197–3213, Aug. 2020.
[16] H. Du, T. Jin, Y. Song, Y. Dai, and M. Li, "Efficient human activity classification via sparsity-driven transfer learning," *IET Radar, Sonar Navigat.*, vol. 13, no. 10, pp. 1741–1746, Oct. 2019.
[17] H. Du, Y. He, and T. Jin, "Transfer learning for human activities classification using micro-Doppler spectrograms," in *Proc. IEEE Int. Conf. Comput. Electromagn. (ICCEM)*, Chengdu, China, Mar. 2018, pp. 1–3.
[18] J. Park, R. Javier, T. Moon, and Y. Kim, "Micro-Doppler based classification of human aquatic activities via transfer learning of convolutional neural networks," *Sensors*, vol. 16, no. 12, p. 1990, Nov. 2016.
[19] M. S. Seyfioglu, B. Erol, S. Z. Gurbuz, and M. G. Amin, "DNN transfer learning from diversified micro-Doppler for motion classification," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 55, no. 5, pp. 2164–2180, Oct. 2019.
[20] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2935–2947, Dec. 2018.
[21] W. Ge and Y. Yu, "Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 10–19.
[22] J. Ngiam, D. Peng, V. Vasudevan, S. Kornblith, Q. V. Le, and R. Pang, "Domain adaptive transfer learning with specialist models," 2018, *arXiv:1811.07056*. [Online]. Available: http://arxiv.org/abs/1811.07056
[23] A. O. Arnold and W. W. Cohen, "Instance-based transfer learning for multilingual deep retrieval," 2019, *arXiv:1911.06111*. [Online]. Available: http://arxiv.org/abs/1911.06111

[24] B. Wang *et al.*, "A minimax game for instance based selective transfer learning," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Alaska, AK, USA, Jul. 2019, pp. 34–43.
[25] T. Wang, J. Huan, and M. Zhu, "Instance-based deep transfer learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Honolulu, HI, USA, Jan. 2019, pp. 367–375.
[26] S. Jialin Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[27] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Barcelona, Spain, Dec. 2016, pp. 901–909.
[28] L. Van Der Maaten, "Accelerating t-SNE using tree-based algorithms," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3221–3245, 2014.
[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7132–7141.
[30] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: http://arxiv.org/abs/1511.07122
[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami Beach, FL, USA, Jun. 2009, pp. 248–255.
[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, CA, USA, Dec. 2012, pp. 1097–1105.
[33] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, Nov. 2000.
[34] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Operat. Syst. Design Implement.*, Nov. 2016, pp. 265–283.
[35] A. Asgarian *et al.*, "A hybrid instance-based transfer learning method," 2018, *arXiv:1812.01063*. [Online]. Available: http://arxiv.org/abs/1812.01063
[36] Y. Kim, J. Park, and T. Moon, "Classification of micro-Doppler signatures of human aquatic activity through simulation and measurement using transferred learning," *Proc. SPIE*, vol. 10188, May 2017, Art. no. 101880V.
[37] X. Shi, Y. Li, F. Zhou, and L. Liu, "Human activity recognition based on deep learning method," in *Proc. Int. Conf. Radar (RADAR)*, Brisbane, QLD, Australia, Aug. 2018, pp. 1–5.
[38] Y. He, Y. Yang, Y. Lang, D. Huang, X. Jing, and C. Hou, "Deep learning based human activity classification in radar micro-Doppler image," in *Proc. 15th Eur. Radar Conf. (EuRAD)*, Madrid, Spain, Sep. 2018, pp. 230–233.
[39] M. S. Seyfioglu, B. Erol, S. Z. Gurbuz, and M. G. Amin, "Diversified radar micro-Doppler simulations as training data for deep residual neural networks," in *Proc. IEEE Radar Conf. (RadarConf)*, Oklahoma City, OK, USA, Apr. 2018, pp. 0612–0617.

Xinyu Li (Graduate Student Member, IEEE) received the B.Eng. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2017, where she is pursuing the Ph.D. degree with the School of Information and Communication Engineering.

Her main research interests are human detection and activity recognition, machine learning, and small-sample learning.

Yuan He (Member, IEEE) received the B.Sc. and M.Sc. degrees from the National University of Defense Technology, Changsha, China, in 2007 and 2010, respectively, and the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands, in 2014.

He is an Assistant Professor at the Beijing University of Posts and Telecommunications, Beijing, China. His main research interests are machine learning, signal processing, and electromagnetic computation.

Francesco Fioranelli (Senior Member, IEEE) received the Ph.D. degree from Durham University, Durham, U.K., in 2014.

He is an Assistant Professor at TU Delft, Delft, The Netherlands, and was an Assistant Professor at the University of Glasgow, Glasgow, U.K., from 2016 to 2019 and a Research Associate at the University College London, London, U.K., from 2014 to 2016. 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.

Xiaojun Jing (Member, IEEE) received the M.S. and Ph.D. degrees from the National University of Defense Technology, Changsha, China, in 1995 and 1999, respectively.

He is a Professor at the Beijing University of Posts and Telecommunications, Beijing, China. His research interests include information security and image processing.

Alexander Yarovoy (Fellow, IEEE) received the diploma degree (Hons.) in radiophysics and electronics, and the Candidate Phys. & Math. Sci. and Doctor Phys. & Math. Sci. degrees in radiophysics from Kharkov State University, Kharkov, Ukraine, in 1984, 1987, and 1994, respectively.

In 1987, he joined the Department of Radiophysics, Kharkov State University, as a Researcher and became a Full Professor there in 1997. From September 1994 to 1996, he was with the Technical University of Ilmenau, Ilmenau, Germany, as a Visiting Researcher. Since 1999, he has been with the Delft University of Technology, Delft, The Netherlands. Since 2009, he has been the Chair of microwave sensing, systems, and signals. He holds six patents. He has authored or coauthored more than 450 scientific or technical articles and 14 book chapters. His main research interests are in high-resolution radar, microwave imaging, and applied electromagnetics (in particular, UWB antennas).

Dr. Yarovoy was a recipient of the European Microwave Week Radar Award for the paper that best advances the state-of-the-art in radar technology in 2001 (together with L. P. Ligthart and P. van Genderen) and in 2012 (together with T. Savelyev). In 2010 together with D. Caratelli, he got the Best Paper Award of the Applied Computational Electromagnetic Society (ACES). He served as the Chair and TPC Chair of the 5th European Radar Conference (EuRAD08), Amsterdam, The Netherlands, as well as the Secretary of the First European Radar Conference (EuRAD04), Amsterdam. He also served as the Co-Chair and TPC Chair of the Xth International Conference on GPR (GPR2004) in Delft. He served as an Associate Editor of the *International Journal of Microwave and Wireless Technologies* from 2011 to 2018 and a Guest Editor of five special issues of the IEEE TRANSACTIONS and other journals. In the period 2008–2017, he served as the Director of the European Microwave Association (EuMA).

Yang Yang received the Ph.D. degree from Tianjin University, Tianjin, China, in 2019.

He is an Assistant Professor at Tianjin University. His research interests include deep learning, pattern recognition, object detection, human activity recognition, and micro-Doppler radar.