Environment-Robust WiFi-based Human Activity Recognition using Enhanced CSI and Deep Learning

Zhenguo Shi, *Member, IEEE,* Qingqing Cheng, *Member, IEEE,* J. Andrew Zhang, *Senior Member, IEEE,* Richard Yida Xu, *Member, IEEE,*

***Abstract*—Deep learning has demonstrated its great potential in Channel State Information (CSI)-based Human Activity Recognition (HAR), and hence has attracted increasing attention in both the industry and academic communities. While promising, most existing high-accuracy methodologies require to re-train their models when applying the previous-trained ones to a new/unseen environment. This issue has limited their practical usabilities. In order to overcome this challenge, this paper proposes an innovative scheme, which combines an activity-related feature extraction and enhancement (AFEE) method and Matching Network (AFEE-MatNet). The proposed scheme is "one-fits-all", meaning that the trained model can be directly applied in new/unseen environments without any retraining. We introduce the AFEE method to enhance CSI quality by eliminating noise. Specifically, the approach mitigates environmental noises unrelated to activity while better compressing and preserving the behaviour-related information. Moreover, the size of feature signals generated by AFEE are reduced, which in turn significantly shortens the training time. For effective feature extraction, we propose to use the MatNet architecture to learn transferable features shared among source environments. To further improve the recognition performance, we introduce a prediction checking and correction scheme to rectify some classification errors that do not abide by the state transition of human behaviours. Extensive experimental results demonstrate that our proposed AFEE-MatNet significantly outperforms existing state-of-the-art HAR methods, in terms of both recognition accuracy and training time.**

***Index Terms*—WiFi, Device free sensing, Deep learning, Channel state information, Human activity recognition.**

I. Introduction

WIFI signals, one of the most pervasive wireless signals, have received paramount interest as a promising source for device-free human activity recognition (HAR). For WiFi-based HAR, the sensing task can be accomplished by analyzing the different characteristics of WiFi channels induced by various human behaviors [1], [2]. In such a case, the targets do not need to be equipped with any devices such as cameras or watches, which are necessary for traditional device-based HAR. WiFi-based HAR has the advantages of better user privacy protection and low deployment cost, thereby gaining its potentially popular adoption [3], [4], [5]. Since the channel state information (CSI) is able to provide fine-grained information about communication links such as amplitude and phase diversity, it has been widely explored for WiFi-based HAR [6], [7].

Z. Shi was with University of Technology Sydney, Australia and now is with the School of Computing, Macquarie University. Q. Cheng is with School of Electrical Engineering and Telecommunications, University of New South Wales, Australia. J. A. Zhang is with the School of Electrical and Data Engineering, University of Technology Sydney, Australia. R. Y. D. Xu is with Department of Mathematics, Hong Kong Baptist University (HKBU). E-mail: zhenguo.shi@mq.uts.edu.au, qingqing.cheng@unsw.edu.au andrew.zhang@uts.edu.au, xuyida@hkbu.edu.hk

A. Related Work and Motivation

Recently, significant progress has been made in CSI-HAR, by leveraging signal processing techniques and deep learning networks (DLNs) [8], [9], [10]. To name a few, a recent solution in [11] first transformed CSI measurements from multiple channels into radio images and then extracted the color and texture information from those radio images. On this basis, a successful HAR is accomplished, by using the deep features extracted by a sparse autoencoder (SAE). To further improve sensing performance, the authors in [12] extracted discriminated features from CSI streams, and then employed the long-short term memory recurrent neural networking (LSTM-RNN) to achieve a reliable sensing result using the extracted information. For the same purpose, the bi-directional long short-term memory (BDLSTM) was adopted in [13], through which the representative features in two directions from raw sequential CSI measures can be effectively extracted. Similarly, the authors in [14] proposed a recognition model drawing support from the improved linear discriminant analysis and softmax regression algorithm. In [15], the authors also developed a deep learning based HAR solution which uses a channel selection and combination mechanism to improve C-SI quality. In [16], we developed a feature extraction method to enhance the CSI quality and learn distinguishable information. Then, this extracted data was fed into LSTM-RNN for HAR, obtaining reliable recognition results. Although the above CSI-based HAR works, including our own in [16], can achieve desirable sensing results, these solutions are highly specific to environments where the HAR model is trained [8], [14], [15], [17]. As a result, the recognition accuracy usually drops dramatically if the classifier trained with primitive features in source/seen environments is used to recognize activities in new/unseen environments. In other words, well-trained schemes in the above works cannot be directly used for HAR in unseen environments.

Given the above challenge, significant research effort has been made to improve the generalization ability of HAR techniques by leveraging various DL networks. For instance,

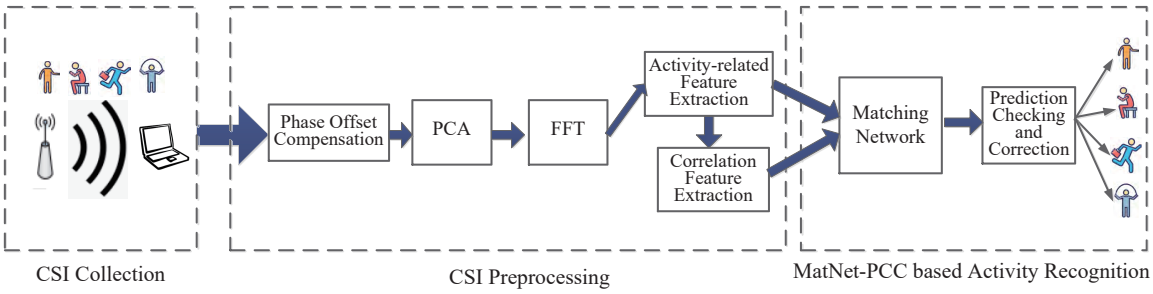


Fig. 1. Main modules for the AFEE-MatNet Scheme.

a transfer learning network was employed in [18] to facilitate environment-robust HAR, by extracting the features commonly shared across source environments and the target/testing environment. Another work [19] exploited a transfer neural network for environment-robust HAR, by capturing the common knowledge in time and spatial domains shared by the testing and source environments. A solution proposed in [20] attempted to remove the environment-specific information contained in the activity and learn the environment-independent features by exploiting the property of the adversarial network. Similarly, the authors in [21] and [22] explored Dense-LSTM and attention-based bidirectional LSTM respectively, to extract behavior-related information and reduce the number of training samples from the testing environment. For the same purpose, the graph few-shot learning network and meta-learning were used in [23] and [24], respectively, for behavior-related feature extraction. Consequently, the recognition models can work well in the testing environment after they are fine-tuned with a few training samples from the testing environment. Another recent HAR solution was proposed in [25] where Doppler frequency shift was utilized to distinguish different activities, especially for some intensive behaviors, e.g., running and jogging. Additionally, the authors of [26] proposed a method to learn environment-robust features, which are then input into a convolutional neural network (CNN) and RNN for cross-environment HAR.

Although environment-independent recognition has been achieved to some extent, the above methods have some limits. To be specific, their recognition accuracies heavily rely on the number of different source environments, and the performance will degrade dramatically when the diversity of source environments is insufficient (e.g., [19], [20]). It is also challenging for them to extract high-quality and discriminative features across different environments due to the limitation of feature extraction processes and deep learning architectures. Moreover, some solutions are designed for line-of-sight (LOS) scenarios only (e.g., [21], [23], [24]), and their sensing accuracies would degrade significantly when LOS conditions are not held. Additionally, some works (e.g., [25], [26]) concentrate on recognizing intensive (i.e., highly dynamic) behaviors only, failing to identify small activities such as standing and laying.

Apart from the aforementioned works, our earlier attempts also targeted the generalized HAR. Specifically, our work in [27] accomplished sensing tasks using only one sample of each activity from the testing environment. To achieve that, we first extracted behavior-dependent features, and then designed an innovative training strategy to bridge the gap between source environments and the testing environment. Although this method can achieve environment-robust HAR, the recognition model still needs to be re-trained when being applied to a new/unseen environment, and training samples from that environment are required. To further improve the generalization ability, we proposed an environment-independent HAR in [28], drawing support from principal component analysis techniques and one-shot learning. As a result, the trained model can be directly applied to new environments, without requiring retraining. However, the solution in [28] is designed for a relatively simple setting, and its recognition performance degrades notably when the number of source environments is insufficient.

B. Main Contributions

To address the above challenges, in this work, we propose a scheme using an activity-related feature extraction and enhancement (AFEE) method and matching network (AFEE-MatNet) to accomplish a cross-environment HAR. Our proposed novelties of AFEE-MatNet scheme has three major novelties. First, it can realize "one-fits-all", meaning that once the model is trained using source/seen environments, it can be directly used in new/unseen environments, without requiring re-training. Second, the proposed AFEE-MatNet is capable of achieving much better recognition performance than other related HAR methods. Third, our proposed AFEE-MatNet only requires a limited number of source environments to train the HAR model well, which is difficult for state-of-the-art HAR techniques to realize. To achieve the above objectives, we develop and apply novel CSI preprocessing techniques (i.e., AFEE) and the matching network with prediction checking and correction (MatNet-PCC), which are significantly different from our prior works and state-of-the-art HAR techniques.

The major contributions of this work are summarized as follows.

- To improve the CSI quality, we propose AFEE to remove activity-unrelated but environment-specific data and enhance behavior-related information. The proposed AFEE is composed of two steps: CSI cleaning and enhancement, and frequency domain feature extraction and signal compression. The step of CSI cleaning and enhancement can remove the phase offset completely, reduce the noise

and eliminate activity-unrelated data in the CSI. The step of frequency-domain feature extraction extracts activity-dependent features in the frequency domain, enhancing activity-related information and reducing feature dependence on the environment. On top of that, the proposed AFEE is capable of decreasing the volume of feature signals, considerably reducing the training time.

- We propose a MatNet with a prediction checking and correction (MatNet-PCC) algorithm to realize environment-independent HAR. We first employ the MatNet architecture to extract transferable features across source environments. To achieve better sensing results, we then design a prediction checking and correction (PCC) algorithm to further rectify some recognition errors that do not follow the state transition of human activities. The designed PCC algorithm can also be applied to other HAR schemes for performance improvement (see Table II in Section V-B).
- We design and perform numerous experiments under various conditions and scenarios. The results demonstrate that our proposed AFEE-MatNet holds many advantages over state-of-the-art HAR methods, in ameliorating recognition accuracies and reducing the training time.

We organize the rest of the paper as follows. In Section II, a brief overview of the proposed AFEE-MatNet is provided. Section III details the information on our proposed AFEE. The details of the developed MatNet-PCC scheme are described in Section IV. We discuss the experimental results in Section V. Section VI concludes the main contributions of this work.

II. Overview of Proposed AFEE-MatNet Scheme

To design a "one-fits-all" recognition model, we propose the AFEE-MatNet scheme by leveraging the discriminative features extracted from CSI and the developed MatNet-PCC scheme. As Fig. 1 depicts, the proposed AFEE-MatNet scheme consists of three main modules: CSI Collection, CSI Preprocessing and MatNet-PCC based Activity Recognition. The first module is used to collect and store the CSI that reflects the changes in wireless signal propagations caused by human behaviors. The second module aims to clean and enhance the acquired CSI matrix through processing in both time and frequency domains. The last module identifies different human activities, drawing support from the enhanced CSI and MatNet.

CSI Collection: In an indoor WiFi network, a person performs different activities, causing various influences on wireless channels. This may absorb, reflect or diffract WiFi signals, changing the characteristics of CSI, such as amplitude and number of multiple paths. To acquire variations on CSI, we employ the Intel 5300 network interface card (NIC), a widely adopted commercial off-the-shelf (COTS) WiFi device. As regulated by IEEE 802.11n, the CSI can be effectively collected by the CSI tools, with 30 subcarriers for each pair of transmitter-receiver antennas [29]. More detailed information about the experimental setup can be found in Section V-A.

CSI Preprocessing: In this module, we intend to improve the quality of the CSI matrix by mitigating the noise, removing activity-unrelated data and condensing activity-related information. Another goal of this module is to reduce the size of the CSI matrix, so as to decrease the complexity. To this end, our proposed AFEE method consists of two key steps: *CSI cleaning and enhancement* and *frequency domain feature extraction*.

In the step of *CSI cleaning and enhancement*, we first perform the conjugate multiplication (CM) method to address the effect of phase offset in the CSI. On top of that, we use the PCA algorithm to further improve CSI quality by eliminating the noise and removing activity-unrelated data. The output can be significantly cleaned and enhanced, while it still contains the residual noise and residual activity-unrelated information. To solve that issue, in the step of *frequency domain feature extraction*, we transfer the CSI from the time domain to the frequency domain before extracting activity-related features. The output of AFEE contains condensed activity-related information, significantly reducing the dimension compared to the original CSI matrix.

MatNet-PCC based Activity Recognition: The purpose of this module is to distinguish different behaviors by leveraging the enhanced CSI from the former module and MatNet-PCC method. In particular, we first employ MatNet to automatically learn hidden features from the enhanced CSI, so as to extract features commonly shared among different environments. As a result, the information, which is activity-related and environment-independent, can be effectively extracted for HAR. After that, we propose a prediction checking and correction (PCC) scheme to further fix recognition errors and improve sensing accuracy. It is noteworthy that MatNet architecture is trained using samples from source environments in an offline manner. The well-trained model is then applied to identify various activities in an online manner.

III. AFEE based CSI Preprocessing

In this section, we will describe the design of AFEE to improve the quality of the CSI matrix. We first present the CSI cleaning and enhancement method, followed by the discussion of frequency-domain feature extraction method.

A. CSI Cleaning and Enhancement

Let N_t and N_r denote the number of antennas at the transmitter and receiver, respectively. The CSI vector $\mathbf{h}(m)$ acquired from the m-th received packet can be represented as

$$\mathbf{h}(m)=[H_{1,1}(m),H_{1,2}(m),\ldots,H_{l,k}(m),\ldots,H_{L,K}(m)]^T, \quad (1)$$

where $H_{l,k}(m)$ indicates the CSI data collected in the lth wireless link at the kth subcarrier; $L = N_t \times N_r$ represents the total number of wireless links; K denotes the total number of subcarriers in the wireless link; and T indicates the transpose operation. The CSI matrix $\mathbf{H}$, which is composed of CSI vectors collected from M packets, can be given by

$$\mathbf{H} = [\mathbf{h}(1), \ldots, \mathbf{h}(m) \ldots, \mathbf{h}(M)]. \quad (2)$$

Although putting the original CSI matrix $\mathbf{H}$ into a deep learning network directly can realize behavior recognition, it is not a good option. The reason for this is that $\mathbf{H}$ contains much activity-unrelated information that could severely influence the recognition result. Moreover, the noise and the phase offset in

$\mathbf{H}$ also affect the recognition performance. To address these problems, we first propose to apply a conjugate multiplication (CM) method [30] to improve the quality of the CSI. The key insight of CM is to take the acquired CSI with the best quality as a reference $\mathbf{h}_{\text{ref}}$, and then calculate a conjugate multiplication of $\mathbf{h}_{\text{ref}}$ and $\mathbf{h}$. The criterion for selecting $\mathbf{h}_{\text{ref}}$ is to choose a CSI vector collected from the antenna with a maximum ratio of amplitudes and standard deviations (MRASD). To obtain $\mathbf{h}_{\text{ref}}$, we first calculate the wireless link with MRASD, by

$$L_{\text{ref}} = \arg\max_{l \in L} \frac{1}{K} \sum_{k=1}^{K} \frac{\text{mean}(|\mathbf{h}_{l,k}|)}{\text{std}(|\mathbf{h}_{l,k}|)}, \tag{3}$$

where L_{ref} represents the index of wireless link with MRASD; mean(.) and std(.) stand for the mean operation and the standard deviation operation, respectively. Then $\mathbf{h}_{\text{ref}}$, acquired from the m-th packet, can be formed in equation (4) as shown on the top of next page.

Upon obtaining $\mathbf{h}_{\text{ref}}$, the reference CSI matrix $\mathbf{H}_{\text{ref}}$ for M packets can be expressed as

$$\mathbf{H}_{\text{ref}} = [\mathbf{h}_{\text{ref}}(1), \ldots, \mathbf{h}_{\text{ref}}(m) \ldots, \mathbf{h}_{\text{ref}}(M)]. \tag{5}$$

Based on equation (5), the conjugate multiplications between all the wireless links and reference links can be obtained by

$$\mathbf{C} = \mathbf{H}_{\text{ref}} \odot \mathbf{H}^{*}, \tag{6}$$

where $\odot$ stands for dot product, and $*$ denotes the Hermitian.

Through the above operations, the output $\mathbf{C}$, with size $LK \times M$, is expected to overcome the effect of phase offset [30]. However, it still contains some activity-unrelated information and random noise, negatively affecting recognition results. To this end, we propose to perform PCA to retain the activity-related information and eliminate noise contained in $\mathbf{C}$. Specifically, we divide $\mathbf{C}$ into L sub-matrices, written as

$$\mathbf{C} = [\overline{\mathbf{C}}_1, \ldots, \overline{\mathbf{C}}_l, \ldots, \overline{\mathbf{C}}_L]^T, \tag{7}$$

$$\overline{\mathbf{C}}_l = \begin{bmatrix} \mathbf{C}_{l,1}(1) & \ldots & \mathbf{C}_{l,1}(m) & \ldots & \mathbf{C}_{l,1}(M) \\ \mathbf{C}_{l,2}(1) & \ldots & \mathbf{C}_{l,2}(m) & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \mathbf{C}_{l,K}(1) & \ldots & \mathbf{C}_{l,K}(m) & \ldots & \mathbf{C}_{l,K}(M) \end{bmatrix}. \tag{8}$$

Next, we apply a PCA operation to each $\overline{\mathbf{C}}_l$ to obtain the top $p+1$ principal components. Note that most of the noise is involved in the first principal component. To mitigate it, we construct the principal component matrix $\widehat{\mathbf{C}}_l$ by discarding the first principal component and keeping the rest of components, represented as

$$\widehat{\mathbf{C}}_l = \overline{\mathbf{C}}_l \times \mathbf{\Phi}_l^{\{2:p\}}, \tag{9}$$

where $\mathbf{\Phi}_l^{\{2:p\}}$ stands for the matrix constructed by elements from the 2nd to the pth eigenvectors. The new feature matrix based on the principal components of all wireless links can be obtained by

$$\widehat{\mathbf{C}} = [\widehat{\mathbf{C}}_1, \widehat{\mathbf{C}}_2, \ldots, \widehat{\mathbf{C}}_L]^T. \tag{10}$$

Note that the size of $\widehat{\mathbf{C}}$ is $P \times M$, $P = pK$, and P is used to achieve a tradeoff between computational complexity and recognition accuracy. The impact of P on the recognition performance will be illustrated in Fig. 8 later. Although $\widehat{\mathbf{C}}$ can be used as the input signal to train the DLN, feeding $\widehat{\mathbf{C}}$ into the DLN directly would cause a significant increase in the training complexity due to its large dimension. For instance, in this paper, the time window for each activity is approximately 1s, the rate of samples f_s is 1KHz, and we set $P = 60$ empirically in the experiments, so $\widehat{\mathbf{C}}$ is a matrix of size 60×1000. It would cause extremely high training complexity if we directly took $\widehat{\mathbf{C}}$ as the input signal for the DLN. Consequently, it is important to decrease the size of the $\widehat{\mathbf{C}}$, thereby lowering the training complexity.

B. Frequency Domain Feature Extraction

Through the above operations, $\widehat{\mathbf{C}}$ has extracted unique characters for different human activities. However, it has a large dimension, resulting significant training overhead. Moreover, $\widehat{\mathbf{C}}$ still contains residual activity-unrelated information and residual noise, leading to performance degradation. To deal with these problems, we propose the frequency domain feature extraction method to extract reliable features in the frequency domain.

We perform Fast Fourier Transform (FFT) on each row of $\widehat{\mathbf{C}}$ to get the frequency domain feature matrix, by

$$\mathbf{C}_F = \text{FFT}(\widehat{\mathbf{C}}), \tag{11}$$

where FFT(.) stands for the Fast Fourier Transform operation. $\mathbf{C}_F$ indicates the extracted frequency domain feature matrix.

It is notable that, most of CSI variations caused by human activity in daily life are in a relatively low frequency range (less than 100Hz), due to the limited speed and space of movements. On this basis, we discard the data in the high frequency range contained in $\mathbf{C}_F$ to remove activity-unrelated information whilst retraining the activity-related features. To further enhance the activity-related CSI, we remove the zero frequency component (i.e., the first column of $\mathbf{C}_F$) which is mainly environment-specific but behavior-unrelated. Let $\widehat{\mathbf{C}_F}$ be the compressed feature matrix, and q be the cutoff frequency used to filter out activity-unrelated features. The value of q is determined based on the types of activity to be recognized. In this paper, we intend to recognize six activities $\{laying, standing, walk, fall, standup, empty\}$. The maximum frequency for these behaviors is about 80Hz [31], so we set $q = 80$Hz. Through the aforementioned operations, the size of $\widehat{\mathbf{C}_F}$ is $P \times qM/f_s$, which is much smaller than that of $\mathbf{C}_F$. Thus, using $\widehat{\mathbf{C}_F}$ as the input signal to train the DLN can result in a notable decrease in the training complexity.

It is noteworthy that the correlation features between different wireless links of transmitter-receiver pairs can provide distinguished information for identifying different behaviors [32]. To provide more information for input signals to DLN, we also compute the correlation feature of $\widehat{\mathbf{C}_F}$, by

$$\mathbf{U}_C = \widehat{\mathbf{C}_F} \times (\widehat{\mathbf{C}_F})^T. \tag{12}$$

Note that, $\mathbf{U}_C$ and $\widehat{\mathbf{C}_F}$ can provide correlated and complementary features for behavior recognition. On the one hand, $\widehat{\mathbf{C}_F}$ contains features in different wireless links, while

$$\mathbf{h}_{\text{ref}}(m)=[\underbrace{H_{L_{\text{ref}},1}(m),\ldots,H_{L_{\text{ref}},K}(m)}_{1},\underbrace{H_{L_{\text{ref}},1}(m),\ldots,H_{L_{\text{ref}},K}(m)}_{2},\ldots,\underbrace{H_{L_{\text{ref}},1}(m),\ldots,H_{L_{\text{ref}},K}(m)}_{L}]^T. \tag{4}$$

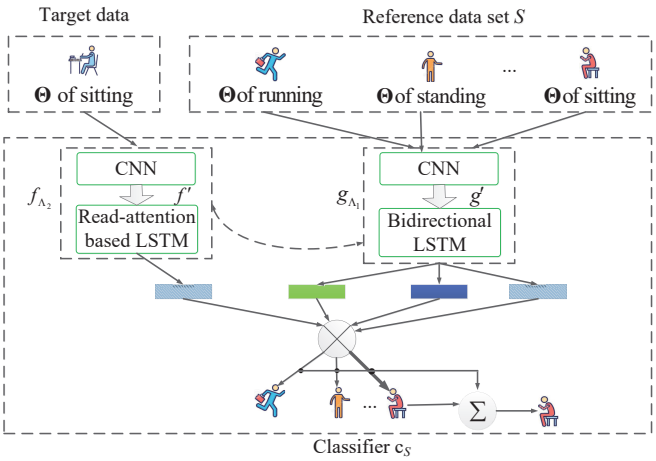


Fig. 2. Architecture of MatNet based activity recognition.

it fails to present other types of information such as the correlation features between different links. On the other hand, $\mathbf{U}_C$ highlights the correlation information between different wireless links, but it loses some information during correlation operations. Therefore, we take $\mathbf{\Theta} = \{\widehat{\mathbf{C}_F}, \mathbf{U}_C\}$ as the input signal of MatNet for extracting reliable and distinguished features.

IV. MatNet-PCC based Human Activity Recognition

As presented in Section III, the output of our AFEE method is expected to contain significantly enhanced CSI, by retaining the activity-related information whilst mitigating activity-unrelated data. However, it is difficult to fully eliminate the impact of the environment. To overcome this problem, we propose to use MatNet to learn and extract transferable features shared among different environments, meaning that these features are robust to environments. Moreover, we propose a prediction checking and correction method to further improve recognition accuracy.

A. Architecture of MatNet

To realize activity classification, we propose to employ MatNet to automatically learn and extract hidden features from the enhanced CSI, as depicted in Fig. 2. Given a source data set S, MatNet is able to build a classier c_s for each S, mapping S to c_S, $S \rightarrow c_S(.)$. The source data set S with N samples can be obtained by

$$S = \{(x_i, y_i)\}_{i=1}^{N}, \tag{13}$$

where (x, y) indicates the input-label pairs; $x = \mathbf{\Theta}$ stands for the input feature signal; y denotes the output label for the corresponding human activity.

After that, the estimated output label $\hat{y}$ can be obtained based on S and the target sample $\hat{x}$, by

$$\hat{y} = \arg\max_{y} P(y|\hat{x}, S), \tag{14}$$

where $P(.)$ stands for the probability distribution, defined as

$$P(\hat{y}|\hat{x}, S) \triangleq S \rightarrow c_S(\hat{x}). \tag{15}$$

To estimate $\hat{y}$, we calculate the linear combination of y in the source data set S. In this paper, an attention mechanism in the form of softmax over the *cosine similarity* is employed to combine y. Let x_i, y_i stand for the input signal and the corresponding label from the source data set $S = \{(x_i, y_i)\}_{i=1}^{N}$, so $\hat{y}$ can be rewritten as

$$\hat{y} = \sum_{i=1}^{N} \frac{e^{\cos(f(\hat{x}), g(x_i))}}{\sum_{j=1}^{N} e^{\cos(f(\hat{x}), g(x_j))}} y_i \tag{16}$$

where $\cos(\alpha, \beta)$ is the cosine similarity function [33], given by

$$\cos(\alpha, \beta) = \frac{\alpha \cdot \beta}{\| \alpha \| \| \beta \|}. \tag{17}$$

In (16), f and g are defined as the embedding functions of $\hat{x}$ and x_i, respectively. Note that $\hat{x}$ and x_i are embedded fully conditioned on the whole source data set S, expressed as $f(\hat{x}, S)$ and $g(x_i, S)$, respectively. This enables MatNet to extract generalised features from different source data (environments).

As shown in Fig. 2, both f and g are composed of CNN with LSTM, which can extract discriminated and generalized features from the input data to reliably detect behaviors. In particular, f consists of a CNN with read-attention based LSTM [34]. Let $f'(\hat{x})$ denote the output features of CNN given a target sample $\hat{x}$. $g(S)$ represents the embed data set via g over the whole source data set S. In such a case, the embedded feature $f(\hat{x}, S)$ can be obtained by

$$f(\hat{x}, S) = \text{attLSTM}(f'(\hat{x}), g(S), N_p), \tag{18}$$

where attLSTM(.) stands for the read-attention based LSTM; N_p denotes the number of unrolling steps in LSTM. Thus, h_{n_p}, the state of the read-attention based LSTM after n_p processing steps, is expressed as

$$h_{n_p} = \hat{h}_{n_p} + f'(\hat{x}), \tag{19}$$

$$\hat{h}_{n_p}, c_{n_p} = \text{LSTM}(f'(\hat{x}), [h_{n_p-1}, r_{n_p-1}], c_{n_p-1}), \tag{20}$$

$$r_{n_p-1} = \sum_{i=1}^{N_s} \text{softmax}\big(h_{n_p-1}^T g(x_i)\big) g(x_i), \tag{21}$$

where $\text{LSTM}(f'(\hat{x}), [h_{n_p-1}, r_{n_p-1}], c_{n_p-1})$ follows the same structure defined in [35]; c_{n_p} represents the cell; r_{n_p-1} is

read-out from $g(S)$; N_s indicates the length of $g(S)$. After conducting N_p steps of "reads", we can obtain

$$\text{attLSTM}(f'(\hat{x}), g(S), N_p) = h_{N_p}. \tag{22}$$

For the embedding function g, it includes a CNN with a bidirectional LSTM [36]. For a given x_i, discriminative features $g'(x_i)$ are extracted by the CNN network, then $g(x_i, S)$ can be obtained with the help of bidirectional LSTM, given as

$$g(x_i, S) = \vec{h}_i + \overleftarrow{h}_i + g'(x_i), \tag{23}$$

$$\vec{h}_i, \vec{c}_i = \text{LSTM}(g'(x_i), \vec{h}_{i-1}, \vec{c}_{i-1}), \tag{24}$$

$$\overleftarrow{h}_i, \overleftarrow{c}_i = \text{LSTM}(g'(x_i), \overleftarrow{h}_{i+1}, \overleftarrow{c}_{i+1}), \tag{25}$$

where $\vec{h}_i$ and $\overleftarrow{h}_i$ stand for the outputs of forward LSTM and backward LSTM, respectively; $\vec{c}_i$ and $\overleftarrow{c}_i$ denote the cells of the forward LSTM and the backward LSTM, respectively; $\text{LSTM}(g', h, c)$ follows the same structure as described in [35].

According to the above discussion, g is a function of the whole source set S. Note that g is critical for conducting embedding operations on x_i, especially when the value of x_j is close to that of x_i. For instance, let x_i and x_j be input signals for two similar activities (e.g., stand up and standing), respectively, then we can train g to transfer x_i and x_j to two recognizable domains conditional on the whole source data set.

B. Training Strategy

We define a task, denoted as $\mathcal{T}$, as the distribution for potential label sets of human behaviors. In each episode, a set of human activities $\mathcal{L}$ are sampled from $\mathcal{T}$, $\mathcal{L} \sim \mathcal{T}$, consisting of six different behaviors: $\{standing, laying, walk, standup, empty, fall\}$. Next, $\mathcal{L}$ is used for sampling both the source data set S and the batch of target set B, achieving $\mathcal{S} = S \sim \mathcal{L}$ and $\mathcal{B} = B \sim \mathcal{L}$. The purpose of training MatNet is minimizing the error between the estimated and the actual labels in $\mathcal{B}$ under the condition of $\mathcal{S}$. In this case, we can obtain the loss function of MatNet based HAR, by

$$Loss = -\text{E}_{\mathcal{L}\sim\mathcal{T}}\left[\text{E}_{\mathcal{S},\mathcal{B}}\left[\sum_{(x,y)\in B} \log P_\Lambda\left(y|x,\mathcal{S}\right)\right]\right], \tag{26}$$

where $\Lambda = \{\Lambda_1, \Lambda_2\}$, Λ_1 and Λ_2 stand for parameter sets of embedding functions g and f, respectively. The core objective of training MatNet is to minimize $Loss$ over $\mathcal{B}$ conditioned on the source data set $\mathcal{S}$, which is

$$\Lambda = \arg\min_{\Lambda} Loss(\Lambda). \tag{27}$$

Note that the training process is conducted fully conditional on the whole data set S that is collected form multiple different source environments. Under this situation, the relationship between different source environments can be built by drawing support from g and f. As a result, the generalized features among different environments can be learned and extracted for HAR. In other words, the impact of a specific environment on recognition performance is significantly reduced. Therefore, the proposed AFEE-MetNet scheme is robust to environments, contributing to environment-independent recognition.

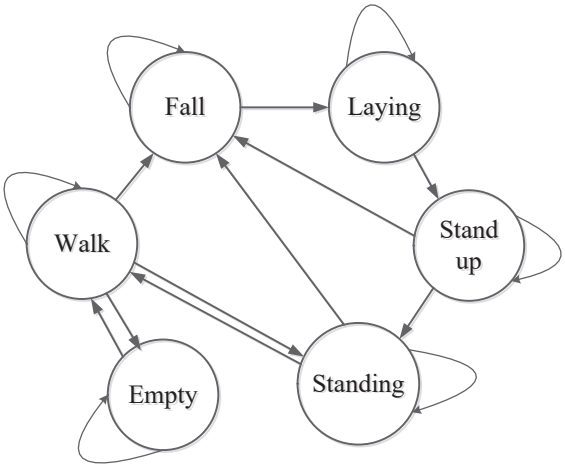


Fig. 3. Simplified state transition diagram for six different activities.

C. Prediction Checking and Correction

The core of the proposed prediction checking and correction (PCC) method is to rectify certain recognition errors that do not match the state transition of human behaviors, so as to further improve the recognition accuracy. When a person performs a set of different activities continuously, these behaviors are not independent but belong to a circle of continuous states. Fig. 3 shows the simplified state transition diagram for the case when people perform six different activities $\{laying, standing, walk, fall, standup, empty\}$. To be specific, when a person conducts "stand up" at the current time slot, he/she may perform "fall", "standing" or "stands up" at next time slot. In such a case, if the output of MatNet at the next time slot is "fall", "standing" or "stands up", the result follows the state transition diagram. Under this situation, we treat the output of MatNet as a logical result and keep it as the final output. Otherwise, the output of MatNet at the next time will be regarded as an incorrect result, and we will correct it using the detailed scheme as follows.

We let N_a stand for the total number of activities to be recognized, and N_a is set to 6, including $\{laying, standing, walk, fall, standup, empty\}$; $Y(t_n - 1)$ denotes the final output of our proposed method at time slot $t_n - 1$, and $\hat{y}(t_n)$ represents the output of MatNet at time slot t_n; $Y(t_n - 1), \hat{y}(t_n) \in [1, N_a]$; Υ is the state transition diagram for different activities. If $Y(t_n - 1)$ and $\hat{y}(t_n)$ abide by the state transition diagram, i.e., $[Y(t_n - 1), \hat{y}(t_n)] \sim \Upsilon$, we set $\eta = 1$, otherwise, $\eta = 0$.

If $\eta = 0$, it means that the output of the MatNet is incorrect and should be rectified. We let $\mathbf{P}_m$ with size $N_a \times N_a$ be the confusion matrix; $\mathbf{P}_m(i, j)$ stands for the probability of the activity i being recognized as the activity j; $i, j \in [1, N_a]$. We will rectify the incorrect sensing result based on Υ and $\mathbf{P}_m$. Specifically, we first find the possible activity set for the time

Algorithm 1: Prediction Checking and Correction.

begin
  Initialize: the final output $Y(t_n-1)$,
  the outputs of MatNet $\hat{y}(t_n), \hat{y}(t_n-1), \ldots, \hat{y}(t_n-\tau)$,
  the state transition diagram Υ;
  the confusion matrix $\mathbf{P}_m$;
  if $[\hat{y}(t_n-1), \ldots, \hat{y}(t_n-\tau)] \nsim \Upsilon$
    PCC method is inactivated;
    $Y(t_n) = \hat{y}(t_n)$;
  else
    if $[\hat{y}(t_n), Y(t_n-1)] \sim \Upsilon$
      $\eta = 1$;
    else
      $\eta = 0$;
      Compute A_{t_n} according to equation (28);
      Compute j^* according to equation (29);
    end
    Compute $Y(t_n)$ according to equation (30);
  end
  Update $\mathbf{P}_m$;
end

slot t_n conditioned on $Y(t_n-1)$, denoted as A_{t_n}, which is given by

$$A_{t_n} \triangleq \{j | [Y(t_n-1), j] \sim \Upsilon, j \in [1, N_a]\}. \tag{28}$$

After that, the probability of misjudging A_{t_n} as $\hat{y}(t_n)$ can be obtained with the help of $\mathbf{P}_m$. Moreover, we can get the activity j^* that holds the highest probability of incorrect classification in $\mathbf{P}_m$, expressed as

$$j^* = \arg \max_{j \in A_{t_n}} \mathbf{P}_m(j, \hat{y}(t_n)), \tag{29}$$

where j^* can be treated as the final output of our proposed scheme at time slot t_n. Thus, the final output of our proposed method at time slot t_n can be expressed as

$$Y(t_n) = \begin{cases} \hat{y}(t_n), & \eta = 1, \\ j^*, & \eta = 0. \end{cases} \tag{30}$$

To this end, we can see that the value of $Y(t_n)$ heavily relies on $Y(t_n-1)$. In other words, the accuracy of $Y(t_n-1)$ significantly affects $Y(t_n)$, restricting the performance of PCC. To address this issue, we propose an activation mechanism for PCC to guarantee its reliability, i.e., preventing over-correction cases. To be specific, the proposed PCC method can be activated to correct $\hat{y}(t_n)$, only if the outputs of MatNet $[\hat{y}(t_n-1), \ldots, \hat{y}(t_n-\tau)]$ abide by Υ but $\hat{y}(t_n)$ does not. In this paper, the value of τ is empirically set as 3. The details of proposed PCC are summarized in Algorithm 1.

V. Implementation and Evaluation

In this section, numerous simulations are designed and conducted to verify the recognition performance of the proposed AFEE-MatNet.

A. Experimental Setup

To validate the recognition performance, we implement the proposed AFEE-MatNet in seven indoor environments (such as laboratory, meeting room, office, two types of bedrooms, dining room, living room) with various wireless environmental

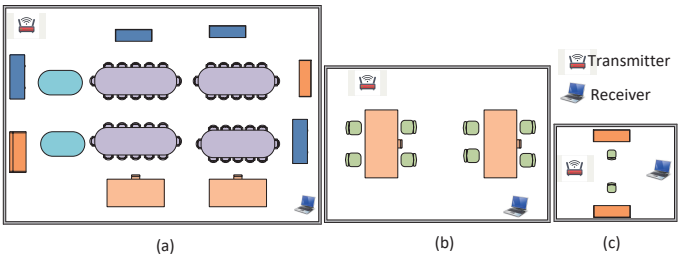


Fig. 4. Layout of three indoor experimental areas: (a) $6m \times 7m$ laboratory. (b)$4m \times 6m$ meeting room. (c)$3m \times 4m$ office.

complexities. Notably, the complexity of the wireless environment is dependent on many factors, e.g., distance and obstacles between transmitter and receiver. In each configuration/environment, each of five people performs six types of behaviors, including falling, laying, stand up, standing, walking, and empty room. Each behavior is conducted 200 times by each person. Among these seven environments, the first four environments are taken as source environments, and the collected data in them are regarded as the training dataset, which is used for training the proposed AFEE-MatNet scheme. The remaining three configurations are treated as testing environments, and the acquired data in these environments is the testing dataset being used for evaluating the recognition performance of the proposed AFEE-MatNet scheme. Due to limited space, we illustrate the layouts of three testing environments in Fig. 4. In particular, the first configuration/environment is a $6m \times 7m$ laboratory room, the second one is a $4m \times 6m$ meeting room, and the third one is a $3m \times 4m$ square area.

In the experiments, two computers are employed as transmitter and receiver, respectively, and each equips with an Intel WiFi NIC5300 network card operating under the 802.11n standard. The WiFi exposure level of Intel NIC5300 is around 0.05W/m^2, which is much lower than the guideline (i.e., 10W/m^2) [37], significantly mitigating health risk [38], [39]. Both the transmitter and receiver conduct data transmission at the operating frequency 5.32 GHz. The transmitter, having one antenna ($N_t = 1$), keeps emitting signals, and the receivers continuously collect data via three antennas ($N_r = 3$). The CSI tool in [29] is adopted to acquire and store signals (i.e., CSI). There are 30 subcarriers ($S = 30$) available for each pair of transmitter-receiver antennas. We propose to use a sliding window to collect data/samples for each behavior from raw CSI streams, and the time length for this window is set as 1s. In the training process, if multiple behaviors are involved in one time window, its label will be the activity with the largest ratio. We leave the work of designing a sliding window with an adaptive time length based on different activities to be a future task. The sample rate is 1 kHz, thus the dimension of CSI matrix ($\mathbf{H}$) is 90×1000.

In the training stage, each embedding function of the proposed MatNet-eCSI includes a CNN with 6 convolutional layers. In each layer, there are 3×3 convolution, a ReLU non-linearity operation, and a 2×2 max-pooling. We employ

TABLE I
AVERAGE RECOGNITION ACCURACY OF DIFFERENT METHODS IN THREE INDOOR CONFIGURATIONS

Method	1st Exp.	2nd Exp.	3rd Exp.
Proposed AFEE-MatNet	0.734	0.763	0.803
EI	0.531	0.577	0.605
TNNAR	0.485	0.511	0.544
BVP	0.687	0.699	0.715

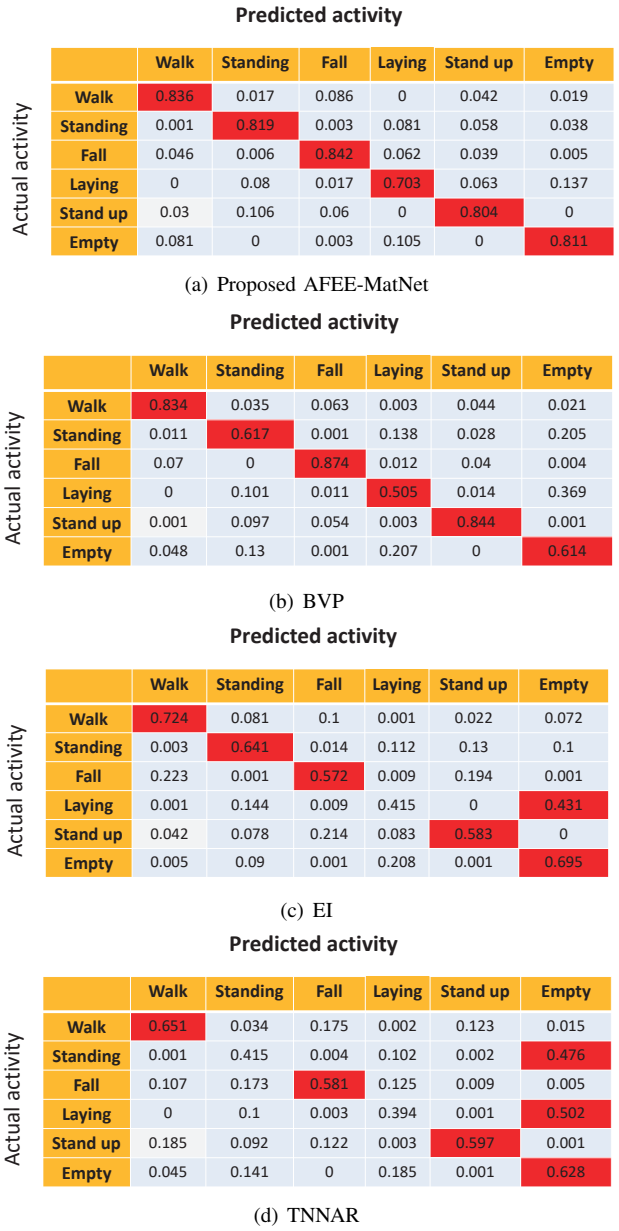
a 2.3 GHz PC with Nvidia GeForce GTX 1070Ti graphic card (8GB memory) to train the proposed AFEE-MatNet. We select 64 and 0.001 as the batch size and learning rate, respectively. Note that we obtain robust scaling for the developed AFEE-MatNet in experiments. Specifically, we perform a normalization operation on the input signal before feeding it into the MatNet architecture. Next, in the training process, we utilize the Batch Normalization method [40] to normalize the input signal of each layer.

B. Performance Evaluation

In this section, we first present extensive simulation results under various conditions and parameters, to analyze the performance of the proposed AFEE-MatNet and other three recognition schemes (i.e., EI [20], TNNAR [19], and BVP [26]). Then, we comprehensively validate the sensing capability of our developed AFEE-MatNet from a wide range of aspects.

1) Recognition Performance for Different Methods: Table I compares the average recognition accuracy of six activities for different sensing methods under different testing environments. The training dataset from four source environments are collected to train each scheme, and each trained model is used to identify activities in testing environments. As can be observed from this table, the proposed AFEE-MatNet notably outperforms the other three sensing methods in each testing environment, which gives credit to the property of the proposed AFEE scheme and MatNet. Specifically, the AFEE is proposed to mitigate most activity-unrelated data and condense the behavior-related information. As a result, the impact of activity-unrelated components (e.g., caused by noise or environment) on feature signals is significantly reduced. Moreover, we propose to employ MatNet architecture to automatically build a relationship among different source environments and extract generalized features for HAR. In other words, the features commonly shared among different source environments are extracted for HAR, and the information subject to a certain environment would be discarded. This enables our proposed AFEE-MatNet to achieve robustness to environments. For EI [20] and TNNAR [19], the number of source environments limits their recognition accuracies. When the number is not sufficient, it is hard for these methods to achieve accurate sensing results. BVP [26] fails to reliably classify some light activities (such as laying), degrading its detection accuracy.

To provide more details, we present the confusion matrix of four methods in Fig. 5. In this figure, we select the third testing environment to examine the performance of each method. We can observe that the sensing accuracy of the proposed AFEE-MatNet is far better than those of the other three methods.



(a) Proposed AFEE-MatNet

(b) BVP

(c) EI

(d) TNNAR

Fig. 5. Confusion matrix for different human activity recognition methods.

To be specific, each estimated behavior is in accordance with the corresponding actual one with high probability, implying that the proposed work is capable of accomplishing reliable detections. By contrast, for EI [20] and TNNAR [19], their prediction results cannot match the corresponding actual ones. Although the prediction activity of BVP [26] is consistent with the actual behaviors, the accuracy for light activities is low (e.g., laying and standing).

In Fig. 6, we demonstrate the number of source environments on the average recognition accuracy for four sensing schemes. The third testing environment is selected for performance evaluation in this figure. As can be seen, the proposed AFEE-MatNet, EI [20] and TNNAR [19] can achieve improved detection results with the number of source environments increasing. The reason is that, with more source environments, these three schemes are able to extract more

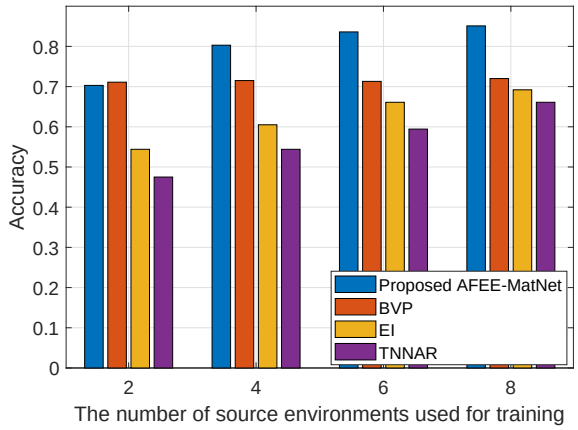


Fig. 6. Recognition accuracy with increased number of source environments

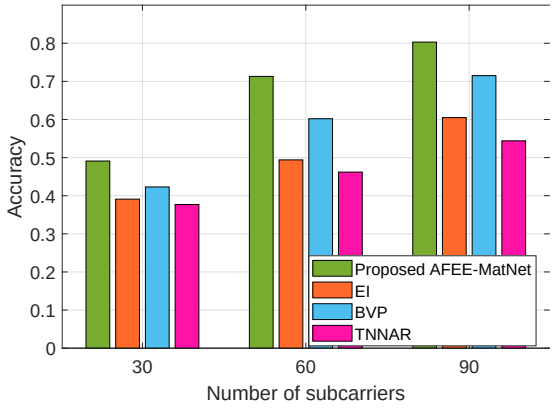


Fig. 7. Impact of the number of used subcarriers on the recognition accuracy.

transferable features shared among these environments, resulting in better recognition performance. In contrast, increasing the number of source environments does not necessarily lead to improvement for BVP [26]. This is because the authors of BVP proposed velocity-related data for HAR, which has little relationship with the number of source environments. Moreover, the RNN architecture adopted in that paper cannot learn generalized features shared among source environments.

In Fig. 7, we show how the number of receiving antennas, each containing 30 subcarriers, affects the average recognition accuracy. We examine the sensing performance of different sensing schemes in the third testing environment. We can observe from this figure that all methods can obtain improved recognition results when the number of subcarriers increases. Moreover, the proposed work is able to achieve larger improvement with more subcarriers, compared to the other three methods.

In Table. II, we also present how well the proposed PCC inproves the classification results of different methods. The recognition stage is conducted in the third testing environment. From this table, it is clear that PCC can also be applied to other methods, enabling these methods to achieve higher sensing accuracies. The reason is that PCC can correct some detection errors that are not in accordance with the state

TABLE II
IMPACT OF PCC ON RECOGNITION ACCURACY FOR DIFFERENT METHODS

Method	Without PCC	With PCC
Proposed AFEE-MatNet	0.752.	0.803
BVP	0.715	0.758
EI	0.605	0.621
TNNAR	0.544	0.559

TABLE III
IMPACT OF AFEE ON PROPOSED METHOD

Method	Accuracy	Training Time
With AFEE	0.767	131.7 mins
Without AFEE	0.464	531.2 mins

transition of human behaviors. Moreover, we can see that our proposed work can achieve a more obvious improvement than the other methods. This is because our proposed work achieves higher sensing accuracies, thereby having more opportunities to activate PCC to further improve detection performance, as can be seen from the activation mechanism of PCC in Section IV-C).

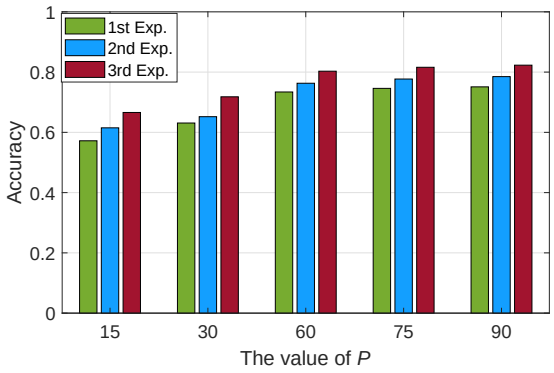
2) Effect of AFEE on AFEE-MatNet: In this subsection, we investigate the significance of AFEE on the proposed scheme.

In Table III, we demonstrate how AFEE affects the performance of the proposed scheme in the sense of recognition accuracy and training time. In this table, we show the average sensing accuracy of three testing environments. As can be seen, the setup with AFEE achieves much higher training accuracy and less training time, compared to the setup without AFEE. The reason is that the proposed AFEE has the capability of significantly suppressing activity-unrelated data and enhancing activity-related information. As a result, the impact subject to a specific environment but not helpful for HAR can be effectively reduced, making the proposed scheme more robust to variations of environments. Moreover, we can find that the training time in the case with AFEE is much less than that without AFEE. This is because AFEE can greatly decrease the size of the CSI matrix that are the input signals for the training stage, shortening the training time.

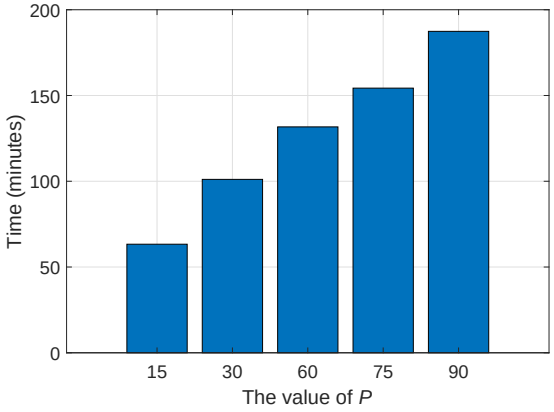
Fig. 8 illustrates how the number of principal components for all wireless links (P) in AFEE influences the sensing accuracy and training time. As Fig. 8 depicts, a larger P can result in improved recognition accuracy for each testing environment, while the speed of improvement slows down when P is sufficiently large. This is because a larger P, i.e., more principal components, can provide more useful information for HAR, contributing to better recognition results. However, the proportion of distinctive features for HAR contained in the principal component with a larger index becomes less. Additionally, the training time becomes longer with a larger P. Therefore, there exists a tradeoff between recognition accuracy and training time when choosing the value of P.

3) Impact of Input Signals and Human Diversity on AFEE-MatNet: To further examine the performance of our AFEE-MatNet, we discuss how its performance changes with various input signals and human subjects.

In Table IV, we investigate changes in recognition perfor-



(a) Recognition Accuracy



(b) Training time

Fig. 8. Impact of P on the recognition accuracy and training time.

TABLE IV
RECOGNITION ACCURACY USING DIFFERENT INPUT SIGNALS.

Method	1st Exp.	2nd Exp.	3rd Exp.
AFEE-MatNet	0.734	0.763	0.803
AFEE-MatNet-C	0.698	0.724	0.766
AFEE-MatNet-U	0.669	0.715	0.749

mance with different input signals in three testing environments. "AFEE-MatNet" refers to the case of using the output of AFEE (i.e., $\mathbf{\Theta}$) to train MatNet for HAR. "AFEE-MatNet-C" and "AFEE-MatNet-U" stand for cases of putting $\widehat{\mathbf{C}_F}$ and $\mathbf{U}_C$ (refer to Section III-B) into the MatNet architecture for training, respectively. From this table, it is obvious that AFEE-MatNet performs much better than both "AFEE-MatNet-C" and "AFEE-MatNet-U", because it can extract more distinguishable and generalized features for recognition.

In Fig. 9, we demonstrate how detection accuracies vary with different human beings in each testing environment. Two persons performed different activities in the training stage to build the training dataset. In the testing process, the trained model is used to recognize activities performed by the other three volunteers. In this figure, we provide the average recognition accuracy for three persons. As can be seen, the recognition accuracy changes differently with various testing persons in each testing environment, implying that different persons have diverse impacts on sensing performance. Another observation is that the average recognition results

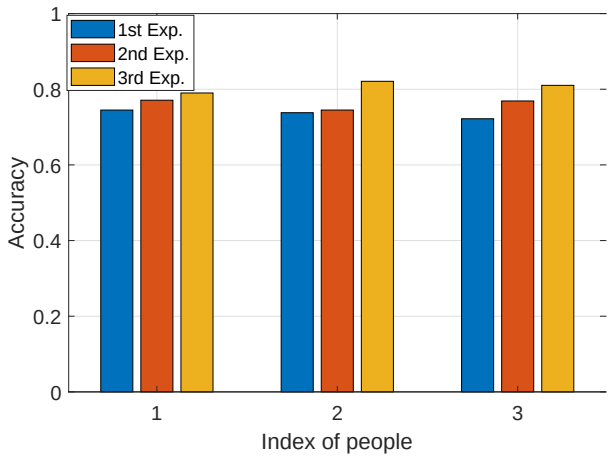


Fig. 9. Average recognition accuracy for different people

across various persons are still reliable (e.g., above 72%) in all testing environments. This demonstrates that the proposed AFEE-MatNet is robust to human diversity.

C. Application potential of AFEE-MatNet

The above simulations verify that the proposed AFEE-MatNet is able to provide a reliable "one-for-all" HAR solution, which brings great application potential. Specifically, AFEE-MatNet can be applied to a wide range of applications, including health, safety, security and entertainment in aged care centers, hospitals, and smart homes. For instance, aged care centers may apply our proposed AFEE-MatNet to detect aged people's abnormal activities (such as falling down) and report an alert to the monitor center, without exposing people's private information (e.g., facial data). Moreover, our AFEE-MatNet can be applied by hospitals to monitor patients' behaviors, ensuring their safety. In addition, AFEE-MatNet can be used by smart home systems, so that people can control appliances via their activities or gestures without using conventional remote-controls.

VI. CONCLUSION

In this paper, we propose an innovative CSI-based HAR scheme, denoted as AFEE-MatNet, to accomplish the "one-fits-all" human activity recognition. The proposed scheme is shown to have the following major novelties. First, the proposed scheme only requires an initial training, and then it can be directly applied to new environments without an extra re-training process. Second, the proposed scheme is able to achieve much better recognition performance, compared to other HAR techniques. Third, our scheme requires fewer seen environments for training the recognition model, compared to other HAR methods. These were achieved by the combined AFEE and MatNet-PCC methods. The AFEE method is able to mitigate noise, eliminate the impact of behavior-unrelated elements subject to the specific environment, and retain the activity-related information. It can also significantly decrease the size of input signals, lowering the computational complexity and shortening the training time. We propose to employ

the MatNet architecture to extract generalized features shared among source environments, facilitating cross-environment recognitions. To further improve sensing performance, we propose a prediction checking and correction method to rectify detection errors that do not abide by the state transition of behaviors. We design and conduct numerous experiments to validate the performance of our proposed AFEE-MatNet from a wide range of aspects. The extensive results verify that our proposed scheme significantly outperforms existing state-of-the-art techniques, in terms of improving the recognition accuracy and lowering the training time.

In this work, LSTM was adopted in MatNet to process activity signals and extract effective features. Its memorized structure demands more resources in training and implementation, compared to techniques without requiring memory. How to apply techniques without memory for deep learning-based HAR is an interesting open research problem.

ACKNOWLEDGMENT

This research was supported partially by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (project DP210101411).

REFERENCES

[1] H. Fei, F. Xiao, J. Han, H. Huang, and L. Sun, "Multi-variations activity based gaits recognition using commodity wifi," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 2, pp. 2263–2273, 2020.

[2] Y. Wang, K. Wu, and L. M. Ni, "Wifall: Device-free fall detection by wireless networks," *IEEE Transactions on Mobile Computing*, vol. 16, no. 2, pp. 581–594, Feb 2017.

[3] W. Zhang, S. Zhou, L. Yang, L. Ou, and Z. Xiao, "Wifimap+: High-level indoor semantic inference with wifi human activity and environment," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 7890–7903, 2019.

[4] J. Huang, B. Liu, C. Chen, H. Jin, Z. Liu, C. Zhang, and N. Yu, "Towards anti-interference human activity recognition based on wifi subcarrier correlation selection," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 6, pp. 6739–6754, 2020.

[5] J. Huang, B. Liu, P. Liu, C. Chen, N. Xiao, Y. Wu, C. Zhang, and N. Yu, "Towards anti-interference wifi-based activity recognition system using interference-independent phase component," in *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, 2020, pp. 576–585.

[6] Z. Wang, B. Guo, Z. Yu, and X. Zhou, "Wi-fi csi-based behavior recognition: From signals and actions to activities," *IEEE Communications Magazine*, vol. 56, no. 5, pp. 109–115, 2018.

[7] B. Sheng, Y. Fang, F. Xiao, and L. Sun, "An accurate device-free action recognition system using two-stream network," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 7, pp. 7930–7939, 2020.

[8] C. Xiao, D. Han, Y. Ma, and Z. Qin, "Csigan: Robust channel state information-based activity recognition with gans," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10 191–10 204, 2019.

[9] S. Fang, C. Li, W. Lu, Z. Xu, and Y. Chien, "Enhanced device-free human detection: Efficient learning from phase and amplitude of channel state information," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 3048–3051, 2019.

[10] K. Chen, L. Yao, D. Zhang, X. Wang, X. Chang, and F. Nie, "A semisupervised recurrent convolutional attention model for human activity recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 5, pp. 1747–1756, 2020.

[11] Q. Gao, J. Wang, X. Ma, X. Feng, and H. Wang, "Csi-based device-free wireless localization and activity recognition using radio image features," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 11, pp. 10 346–10 356, Nov 2017.

[12] S. Yousefi, H. Narui, S. Dayal, S. Ermon, and S. Valaee, "A survey on behavior recognition using wifi channel state information," *IEEE Communications Magazine*, vol. 55, no. 10, pp. 98–104, Oct 2017.

[13] Z. Chen, L. Zhang, C. Jiang, Z. Cao, and W. Cui, "Wifi csi based passive human activity recognition using attention based blstm," *IEEE Transactions on Mobile Computing*, vol. 18, no. 11, pp. 2714–2724, 2019.

[14] J. Zuo, X. Zhu, Y. Peng, Z. Zhao, X. Wei, and X. Wang, "A new method of posture recognition based on wifi signal," *IEEE Communications Letters*, vol. 25, no. 8, pp. 2564–2568, 2021.

[15] F. Wang, W. Gong, J. Liu, and K. Wu, "Channel selective activity recognition with wifi: A deep learning approach exploring wideband information," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 1, pp. 181–192, 2020.

[16] Z. Shi, J. A. Zhang, R. Xu, and Q. Cheng, "Deep learning networks for human activity recognition with csi correlation feature extraction," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, May 2019, pp. 1–6.

[17] F. Wang, J. Liu, and W. Gong, "Wicar: Wifi-based in-car activity recognition with multi-adversarial domain adaptation," in *2019 IEEE/ACM 27th International Symposium on Quality of Service (IWQoS)*, 2019, pp. 1–10.

[18] J. Zhang, Z. Tang, M. Li, D. Fang, P. Nurmi, and Z. Wang, "Crosssense: Towards cross-site and large-scale wifi sensing," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '18. New York, NY, USA: ACM, 2018, pp. 305–320. [Online]. Available: http://doi.acm.org/10.1145/3241539.3241570

[19] J. Wang, V. W. Zheng, Y. Chen, and M. Huang, "Deep transfer learning for cross-domain activity recognition," in *Proceedings of the 3rd International Conference on Crowd Science and Engineering*, ser. ICCSE?8. New York, NY, USA: Association for Computing Machinery, 2018. [Online]. Available: https://doi-org.ezproxy.lib.uts.edu.au/10.1145/3265689.3265705

[20] W. Jiang, C. Miao, F. Ma, S. Yao, Y. Wang, Y. Yuan, H. Xue, C. Song, X. Ma, D. Koutsonikolas, W. Xu, and L. Su, "Towards environment independent device free human activity recognition," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '18. New York, NY, USA: ACM, 2018, pp. 289–304. [Online]. Available: http://doi.acm.org/10.1145/3241539.3241548

[21] J. Zhang, F. Wu, B. Wei, Q. Zhang, H. Huang, S. W. Shah, and J. Cheng, "Data augmentation and dense-lstm for human activity recognition using wifi signal," *IEEE Internet of Things Journal*, vol. 8, no. 6, pp. 4628–4641, 2021.

[22] C. C. Q. Z. C. Z. X. D. Yong Tian, Sirou Li, "Small csi samples-based activity recognition: A deep learning approach using multidimensional features," in *Security and Communication Networks*, 2021.

[23] Y. Zhang, Y. Chen, Y. Wang, Q. Liu, and A. Cheng, "Csi-based human activity recognition with graph few-shot learning," *IEEE Internet of Things Journal*, pp. 1–1, 2021.

[24] Y. Zhang, X. Wang, Y. Wang, and H. Chen, "Human activity recognition across scenes and categories based on csi," *IEEE Transactions on Mobile Computing*, pp. 1–1, 2020.

[25] Y. Ge, S. Li, M. Shentu, A. Taha, S. Zhu, J. Cooper, M. Imran, and Q. Abbasi, "A doppler-based human activity recognition system using wifi signals," in *2021 IEEE Sensors*, 2021, pp. 1–4.

[26] Y. Zheng, Y. Zhang, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Zero-effort cross-domain gesture recognition with wi-fi," in *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '19. New York, NY, USA: ACM, 2019, pp. 313–325. [Online]. Available: http://doi.acm.org/10.1145/3307334.3326081

[27] Z. Shi, J. A. Zhang, Y. D. R. Xu, and Q. Cheng, "Environment-robust device-free human activity recognition with channel-state-information enhancement and one-shot learning," *IEEE Transactions on Mobile Computing*, pp. 1–1, 2020.

[28] Z. Shi, J. A. Zhang, R. Xu, Q. Cheng, and A. Pearce, "Towards environment-independent human activity recognition using deep learning and enhanced csi," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, 2020, pp. 1–6.

[29] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Tool release: Gathering 802.11n traces with channel state information," *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 1, pp. 53–53, Jan. 2011. [Online]. Available: http://doi.acm.org/10.1145/1925861.1925870

[30] X. Li, D. Zhang, Q. Lv, J. Xiong, S. Li, Y. Zhang, and H. Mei, "Indotrack: Device-free indoor human tracking with commodity wi-fi," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 3, Sep. 2017. [Online]. Available: https://doi.org/10.1145/3130940

[31] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Device-free human activity recognition using commercial wifi devices," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 5, pp. 1118–1131, May 2017.

[32] Z. Shi, J. A. Zhang, R. Xu, and G. Fang, "Human activity recognition using deep learning networks with enhanced channel state information," in *2018 IEEE Globecom Workshops (GC Wkshps)*, 2018, pp. 1–6.

[33] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Computer Vision – ACCV 2010*, R. Kimmel, R. Klette, and A. Sugimoto, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 709–720.

[34] O. Vinyals, S. Bengio, and M. Kudlur, "Order matters: Sequence to sequence for sets," *arXiv preprint arXiv:1511.06391*, 2015.

[35] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[36] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

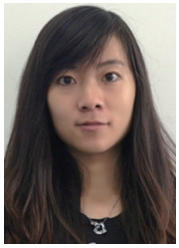
[37] ICNIRP, "Guidelines for Limiting Exposure to Electromagnetic Fields (100 kHz to 300 GHz)," *Health Phys*, 2020.

[38] FCC, "CFR 2.1091 - Radiofrequency radiation exposure evaluation: mobile devices." *Federal Communications Commission*, 2010.

[39] FCC.report, "Maximum Permissible Exposure Report," *SHENZHEN LCS COMPLIANCE TESTING LABORATORY LTD.*, 2018.

[40] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

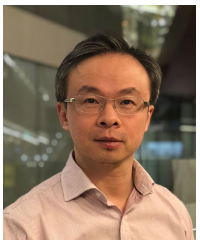
Zhenguo Shi received his M.S.(in 2011) degree and PhD (in 2016) from Harbin Institute of Technology (HIT), China. He received his second PhD degree (in 2022) from University of Technology Sydney, Australia. He was a visiting student at the University of Leeds (2012-2013) and a visiting scholar at the University of Technology Sydney, Australia (2016-2017). He is currently working as a Postdoctoral Research Associate in the School of Computing at the Macquarie University, Australia. His research interests include Wireless sensing, Human activity recognition, IoT, Deep learning, Cognitive radio, Interference alignment, Massive MIMO and UWB wireless communication.



Qingqing Cheng received her M.E. degree from the Harbin Institute of Technology, China in 2014 and her Master of Research (MRes) degree from the Macquarie University, Australia, in 2016. She is currently working as a PhD candidate in the School of Electrical and Data Engineering at the University of Technology Sydney, Australia. Her research interests include 5G security, privacy preservation, cognitive radio, deep learning, and massive MIMO.

J. Andrew Zhang (M'04-SM'11) received B.Sc. degree from Xi'an JiaoTong University, China, in 1996, M.Sc. degree from Nanjing University of Posts and Telecommunications, China, in 1999, and Ph.D. degree from the Australian National University, in 2004. Currently, he is an Associate Professor in the School of Electrical and Data Engineering and the director of the Radio Sensing and Pattern Analysis (RaSPA) laboratory, University of Technology Sydney, Australia. He was a researcher with Data61, CSIRO, Australia from 2010 to 2016, the Networked Systems, NICTA, Australia from 2004 to 2010, and ZTE Corp., Nanjing, China from 1999 to 2001.

Dr. Zhang's research interests are in the area of signal processing for wireless communications and sensing, with current focuses on joint communications and radio/radar sensing, and autonomous vehicular networks. He has published over 210 journal and conference papers, and has won 5 best paper awards including in IEEE ICC 2013. He is a recipient of CSIRO Chairman's Medal and the Australian Engineering Innovation Award in 2012 for exceptional research achievements in multi-gigabit wireless communications.



Richard Yi Da Xu is currently a Professor with the Department of Mathematics, Hong Kong Baptist University (HKBU).He has authored about 50 papers, including the IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, Pattern Recognition, ACM Transactions on Knowledge Discovery from Data, the Association for the Advancement of Artificial Intelligence, and the International Conference on Image Processing. His current research interests include machine learning, computer vision, and statistical data mining.