Privacy-Preserving Cross-Environment Human Activity Recognition

Le Zhang, Wei Cui, Bing Li*, Zhenghua Chen, Min Wu and Teo Sin Gee

Abstract—Recent studies have demonstrated the success of using the channel state information (CSI) from WiFi signal to analyze human activities in a fixed and well-controlled environment. Those systems usually degrade when being deployed in new environments. A straightforward solution to solve this limitation is to collect and annotate data samples from different environments with advanced learning strategies. Although workable as reported, those methods are often privacy sensitive because the training algorithms need to access the data from different environments, which, may be owned by different organizations. We present a practical method for the WiFi-based privacy-preserving cross-environment human activity recognition. It collects and shares information from different environments, while maintaining the privacy of individual person being involved. At the core of our approach is the utilization of the Johnson-Lindenstrauss transform, which is theoretically shown to be differentially private. Based on that, we further design an adversarial learning strategy to generate environment-invariant representations for human activity recognition. We demonstrate the effectiveness of the proposed method with different data modalities from two real-life environments. More specifically, on the raw CSI dataset, it shows 2.18% and 1.24% improvements over challenging baselines for two environments respectively. Moreover, with the DWT features, it further yields 5.71% and 1.55% improvements respectively.

Index Terms—Human Activity Recognition, Deep Learning, Adversarial Learning, Differential-privacy, Federated Learning, Multi-Modal Learning

I. INTRODUCTION

Human activity recognition (HAR) are becoming popular due to its far-reaching applications in human-centric problems such as intruder detection, monitoring patients in hospitals and children and elderly monitoring in smart homes [1]. Over the past ten years, HAR systems using various techniques have been proposed and studied [2] and the most popular prototypes could be categorized into two families: wearable sensor based and vision based techniques. In the former scenario [3], people are required to wear specially-designed devices such as accelerometers and gyroscopes. This could be user-unfriendly, especially for the children and elderly. Instead, vision based HAR systems can effectively recognize certain activities with high-resolution cameras [4], [5], [6] and a recent survey can be found in [7]. Nevertheless, the system may fail to work if the people of interest are occluded. These systems also infringe users’ privacy and raise security concerns significantly.

Recently, due to the rapid development of ubiquitous WiFi technologies, non-invasive HAR based on WiFi-signals has been extensively studied. They are said to be “non-invasive” (or “device-free” or “passive”) because that subjects are not assumed to carry any device. The rationale behind such systems is that a certain human activity could bring unique multi-path reflections in wireless signals from a transmitter (TX) to a receiver (RX) [8], as illustrated in Fig 1 WiFi based systems are shown to be advantageous in the following aspects. First, they are easy to deploy in a plug and play manner as the WiFi infrastructures are widely available in indoor environments. Second, they are user-friendly and do not require any devices on the target. Third, they work effectively in a wide range of environment conditions including non-line-of-sight (NLOS) and absence of lighting.

![Fig. 1: Multi-path propagation with human activity. A human with different activities brings unique multi-path reflections in wireless signals from a transmitter (TX) to a receiver (RX).](image)

Amongst all the WiFi-based HAR systems, the received signal strength (RSS) and detailed channel state information (CSI) [9] are widely used. However, recent studies show that the RSS from the WiFi MAC layer is single-valued and suffers from multi-path fading in complex indoor environments [10]. In contrast, the CSI from the physical layer could provide more fine-grained information on multiple channels and is more robust to environmental variations.

Although being widely deployed, existing WiFi based HAR systems mainly work in a fixed and well-controlled environment [11] and their performance may drop significantly when being applied in new environments without system refinement/re-training. A straightforward solution for this limitation is to collect and annotate data from different environments and train the system with the combined dataset [11]. Although this procedure could successfully improve the gen-
eralization of the WiFi HAR system, it is still far from optimal from the privacy point of view. The reason is that the training algorithms have to access the data from different environments. This could often be impractical in privacy-concerned scenarios. For example, a recent measurement study of \[12\] reveals that there exists a high level of information leakage from Wireless networks. This could be used to further infer the location \[13\], \[14\], identity \[15\] and other biometric information \[16\] of the users. This further allows malicious users abuse services like illegally accessing restricted resources and creating bogus alibis. Motivated by this, we study the problem of “Privacy-Preserving Cross-Environment Human Activity Recognition” in wireless networks. In our setting, the HAR system is trained on data samples from different environments while should not expose private information in these datasets. We use the well-established Johnson-Lindenstrauss transform \[17\], \[18\], \[19\], which is theoretically shown to be differentially private. An adversarial learning strategy is further used to generate environment-invariant representations for human activity recognition. Real-world experiments are conducted to demonstrate the effectiveness of the proposed method. In summary, the core contributions and novelty comes from the following three perspective.

- The new task of “Privacy-Preserving Cross-Environment Human Activity Recognition”. In our setting, the learning processing should not expose private information in these datasets. We believe this research could motivate more advanced privacy-concerned wireless sensing applications.
- A novel learning strategy to learn more generalizable activity representations. Through an advanced adversarial learning procedure, we are able to factorize the learnt discriminative features into environment and human activity oriented attributes. Thus, the final representation for activities are more environment-invariant.
- Versatility illustration. We show the feasibility of the proposed methods with different input modalities. By aggregating their results, improved performances are further achieved.

II. RELATED WORK

In the past years, there has been a plethora of work for wireless technologies based HAR, which, from our view, could be divided into two categories: customized hardware based HAR, commercial device based HAR.

A. Customized Hardware based HAR

Several customized hardware based approaches have been explored to sense the human activity by using various hardware \[20\], \[21\], \[22\], \[23\], \[24\], \[25\]. For example, WiVi \[20\] leverages Inverse Synthetic Aperture Radar (ISAR) to identify simple gestures through walls and behind closed doors. WiTrack \[21\] tracks the 3D motion of a user by utilizing a daughterboard for the USRP software radio. Based on WiTrack, WiTrack2.0 is further developed to localize up to five users simultaneously using time of flight (TOF) measurements \[22\]. WiSee \[23\] uses USRP to extract doppler shifts from received signals and achieves promising performance for hand gesture analysis. Chen et al. exploits an SDR based custom receiver design which enables fine-grained keystroke detection and identification using wireless signals \[24\]. Besides, there are a numerous works are proposed based on wearable sensors and cameras for human action recognition. For example, Mahmud et al. \[26\] utilized data from various wearable sensors to automatically recognize actions by an efficient deep CNN architecture which could combine varieties of features extracted from diverse perspectives. Afza et al. \[27\] implemented an action recognition technique based on camera with features fusion and feature selection. While workable in certain scenarios as reported, those approaches are not user-friendly and difficult to deploy.

B. Commercial device based HAR

By contrast, commercial device based HAR leverage the ubiquitous commercial wireless devices for activity recognition. According to the signal characteristic, the commercial device based HAR can be mainly classified into RSS based and CSI based.

1) RSS based HAR: RSS based HAR systems mainly work by utilizing RSS to measure the variations of human activities on wireless signals from WiFi devices \[28\], \[29\], \[30\], \[31\], \[32\]. A statistical model which models the relationship of the RSS variance and the position of moving person was proposed in \[28\]. In \[29\], the authors employ a dedicated transmitter to classify the radio-based activity recognition. Wilson et al. show that their proposed radio tomographic imaging (RTI) based model is able to capture the changes in attenuation caused by human beings \[30\], \[31\] accurately. However, RSS values only provide coarse-grained information about wireless channel and sometimes may suffer from multi-path fading \[33\]. Therefore, RSS-based systems may not perform well due to the unreliable value of RSS signals. Although some effort is devoted to address the errors from the multi-path effect in complex indoor environments \[34\], it usually involve a dense deployment of cooperative wireless devices and thus is less efficient in practical scenarios.

2) CSI Based HAR: In CSI Based systems, the fine-grained physical layer CSI can be exploited from commercial-off-the-shelf (COTS) WiFi network interface cards (NICs) (such as Intel 5300 and Atheros 9390). E-Eyes \[35\] developed by Wang et al. utilizes histograms of CSI amplitudes as fingerprints to identify and distinguish in-place activities and walking movements inside a home. WiHear \[36\] is introduced to detect and identify the fine-grained CSI dynamics from mouth movements. CARM \[37\] quantifies the correlation between CSI dynamics and specific human activities using commercial WiFi devices. Ali et al. adopted unique changes in the CSI values to recognize keystrokes \[38\]. Those methods work under a well-designed three-stage pipeline: signal pre-processing, feature extraction, and Classifier training. Band-pass filtering \[39\] and principal component analysis (PCA) \[37\], \[38\] are applied as preprocessing mechanisms to filter out the interference noise. After de-noising of CSI data, discrete wavelet transform (DWT) is employed to extract
features [36], [37], [38]. Then, handcrafted features are used as input to different machine learning techniques for multi-class classification. Popular choices for classifier include but are not limited to hidden markov model (HMM) [37] and support vector machines (SVMs) [39]. Motivated by the recent success of deep neural networks in various domains such as image classification [40], nonlinear-regression [41] edge detection [42], segmentation [43], crowd counting [44], video analytics [45], [46] and Internet of Things [10], deep learning methods for CSI based activity recognition have been proposed as well. Yousefi et al. first comprehensively reviewed the state-of-art of HAR. Then they proposed a deep learning approach, i.e., long short-term memory (LSTM), for activity recognition using WiFi CSI signals [2]. In order to process CSI measurements in both forward and backward directions, Chen et al. presented an advance BLSTM network for human activity recognition using CSI measurements [42]. Wang et al. presented a sparse autoencoder based deep learning approach for realizing device-free wireless localization and activity recognition [49].

Although WiFi based HAR has been extensively studied in the literature, most of them could only work in a fixed and well-controlled environment. When being deployed in a new environment, they usually fail to recognize the activities accurately. The reason is that different environments usually have various spatial layout and thus the objects in the environment could introduce different multi-path reflections in wireless signals from a TX to a RX. Existing machine learning based HAR systems typically work as a black-box without considering the changes of environment and thus do not generalize well to new environments. Manually collecting and annotating new data in the new environment and re-training the system could solve the problem [11]. However, this procedure could be privacy-sensitive because the training algorithm have to access the data samples from different environments, which could be owned by different organizations. As shown in [12], this could brings a high level risk of information leakage from Wireless networks and enables other abused services [49], [13], [14] and finally allows malicious users abuse services like illegally accessing restricted resources and creating bogus alibis [50]. To this end, we fill in this gap and study the problem of “Privacy-Preserving Cross-Environment Human Activity Recognition” in wireless networks. In our setting, the HAR system is trained on data samples from different environments while should not expose private information in these datasets. We embark on the well-established Johnson-Lindenstrauss transform [17], [18], [19], which is theoretically shown to be differentially private and thus make it impossible to reverse-engineer the sensitive data from what has been released. Based on that, we further designed a deep network which encodes the human activity and the environment attributes in a disentangled manner. This essentially allows us to learn environment-invariant feature representations for different activities in an adversarial manner. Moreover, learning disentangled representations for both environment and activity, as done in the proposed system, allows us to further seamlessly transfer environment-free activity knowledge from other environments. We test the feasibility of the proposed method with different input modalities.

### III. Methodology

#### A. Privacy-Preserving Module

We employ the well-established differential privacy strategy [51], [52], [53] to preserve the privacy of the data samples from different environments. Originally introduced by Dwork et al. [51], differential privacy constitutes nowadays a standard for privacy guarantees. Differential privacy makes it possible to share information about a dataset by describing the patterns of groups within the dataset while withholding information about individuals in that dataset.

**Definition 1.** A randomized mechanism \( M : \mathcal{X}^n \rightarrow \mathbb{R}^d \) is defined to be \((\xi, \delta)\)-differential privacy if for any two datasets \( X, X' \in \mathcal{X}^n \) differing by a single element and for any set of possible output \( O \subseteq \text{Range}(M) \):

\[
\mathbb{P}(M(X) \in O) \leq e^\xi (\mathbb{P}(M(X') \in O) + \delta)
\]

The rationale behind differential privacy, as defined Eqn. [1] is that if the effect of making an arbitrary single substitution in the database is small enough, the query result cannot be used to infer much about any single individual, and therefore provides privacy. Differential privacy has been extensively studied in the machine learning community to protect privacy and examples include but are not limited to logistic regression [54], principal component analysis [55], boosting [56], support vector machine [57] and so on. However, for the WiFi based HAR task, differential privacy has not been explored before.

With the definition of differential privacy, we are now able to show how this can be guaranteed by the Johnson-Lindenstrauss transform [18]. More formally, Let \( N(0, \sigma I) \in \mathbb{R}^{k \times l} \) be a \( k \times l \) matrix where each entry is drawn i.i.d. from \( N(0, \sigma) \). The Johnson-Lindenstrauss Lemma [19] shows that a linear transformation of a set of \( n \) data points to a much smaller subspace by such a random matrix has a high probability to retain the pairwise distances within \((1 \pm \eta)\) for some \( \eta \in [0, 0.5] \). Mathematically, we have

**Theorem 1** Let \( x_1, x_2, \ldots, x_n \) be a set of \( n \) points in \( \mathbb{R}^k \). For any \( \eta, l > 0 \) and an \( N(0, \frac{1}{2}I) \in \mathbb{R}^{k \times l} \) matrix \( M \), with probability at least \( 1 - \frac{2}{e^{\ln(2\pi)}} \), the following holds for every \( i, j \):

\[
1 - \eta \leq \frac{\|x_i^TM - x_j^TM\|^2}{\|x_i - x_j\|^2} \leq 1 + \eta
\]

The Johnson-Lindenstrauss transform is widely used across in different communities including computational speedups, machine learning, information retrieval and so on [58]. For differential privacy, the Johnson-Lindenstrauss transform was studied as well. Blocki et al. [18] proved that the Johnson-Lindenstrauss transform is able to preserve the differential privacy. Furthermore, the Theorem 1 in [17] clearly shows that the mechanism of publishing \( M \) and the noisy data \( \bar{X} = XM + \Delta \) satisfies the \((\eta, \delta)\)-differential privacy.

**Definition 1.** For any \( \eta, \delta > 0 \) and \( \sigma = \frac{w \sqrt{2ln(\frac{1}{\epsilon})}}{\xi} \) where \( w = \max_{1 \leq i \leq k} \left( \sum_{j=1}^{l} M_{ij}^2 \right)^{\frac{1}{2}} \), that is the so-called \( l2 \)-norm sensitivity of \( M \), which is tightly concentrated around 1 [17].

As the Johnson-Lindenstrauss transform is able to preserve the privacy without losing the geometric structure of the
dataset, we suggest to each environment send the transformed data to the central server, as illustrated in Algorithm 1. However, we empirically find that directly sending out the original dataset to the server may lead to less satisfied results. The reason could be that, the signals directly from the sensors turn out to be highly noisy, generating large intra-class variance. Directly applying the Johnson-Lindenstrauss transform on the original dataset further inserts noises in it and makes the following optimization procedure difficult. To this end, we introduce the feature encoding networks for each environment to pre-process the original data samples to get highly discriminative features, which will be further processed by the Johnson-Lindenstrauss transform and finally sent out to the central server. More specifically, we have:

$$F^i(X^i) = f_j^i(X^i, W^{f_j^i}),$$

(3)

where $f_j^i$ is the feature encoding network parameterized by $W^{f_j^i}$ for the $i^{th}$ environment. In our study, $f_j^i$ is implemented by a multi-layer perception and the details could be found in Section III-C1.
Algorithm 1: Privacy-Preserving Module

**Inputs:** features and labels \((F^i, Y^i)\) \(i = 1, \ldots, K\) in each of the \(K\) environments and \(\delta, l > 0\)

**while** Environment iterator \(i\) is less than \(K\) **do**

1. Environment \(i\) generates an \(N'(0, \frac{1}{l})^{k \times l}\) matrix \(M^i\) and an \(N'(0, \sigma)^{k \times l}\) matrix \(\delta^i\);
2. Environment \(i\) sends \(\{F^i, M^i + \Delta^i, Y^i\}\) to the central server;
3. \(i = i + 1\);

**end**

B. Adversarial Learning for Environment-Oriented Attributes

Human bodies and the other objects in the environments are reflectors of WiFi signal [59]. A person with different activity may affect the wireless signal from a fixed environment during a specific time when the activity are being carried out. In this way, the wireless signals received by the devices usually encode informative information which could be specific to both the environment and the activities. Motivated by this, we propose to encode the features of the input signal in a factorized manner. More specifically, we aim to learn the decoupled environment-oriented attributes \((E)\) and activity-oriented attributes \((A)\) in the latent space. As \(E\) and \(A\) are fully optimized by their corresponding loss functions which are elaborated in the following section, the partition for \(E\) and \(A\) could be arbitrary and in this paper we simply set the first \(e\) features as \(E\) and keep the remaining as \(A\). A similar concept has been successfully applied in computer vision tasks [60].

More specifically, for a specific input \(F_i\) on the central server, we have:

\[
\begin{align*}
\hat{F}^i & = F^i M^i + \Delta^i, \\
F^i & = [E_1^i, E_2^i, \ldots, E_e^i, A_1^i, A_2^i, \ldots, A_a^i]
\end{align*}
\]

in which \(e\) and \(a\) stands for the dimensionality of \(E\) and \(A\), respectively.

Based on the received representation \(\hat{F}^i = F^i M^i + \Delta^i\), an activity classifier, denoted as \(f_a\) with parameters \(W^a\), is appended to the system to classify the activities. The classifiers is optimized by the commonly used softmax cross-entropy loss in each environment:

\[
\mathcal{L}^a(f_a(W^a, F^i)) = - \sum_{i=1}^{N^i} \sum_{c=1}^{C} \mathbb{I}[y_k = c] \log q_{ik}^c, \quad i = 1, 2, \ldots, K
\]

where \(\mathbb{I}\) is an indicator function with \(\mathbb{I}(\text{True}) = 1\) otherwise 0. The parameter \(\lambda_1\) controls the weight of this loss function. \(q_{ik}^c\) corresponding to the softmax of the environment classifier \(q_{ik}^e\), that is:

\[
q_{ik}^e = \text{softmax}(f_e(W^e, E_i)),
\]

and the probability of the \(e^{th}\) class is:

\[
q_{ik}^e = \frac{e_{ik}^e}{\sum_{c=1}^{C} e_{ik}^c}
\]  

The features from different environment usually have different structure due to the different environment conditions. Learning representations from each environment individually without considering the coherence amongst them may not generalize well in practice. We argue that a better representation should have good transferability and thus should be environment-invariant. This means that the learned representations are coherent for different data samples from different environment [61], [62]. Motivated by this assumption, we design an auxiliary environment classifier to distinguish which environment the input datum comes from based on the environment-oriented attributes \(E\). After convergence of the system, it cannot distinguish them because the final representation is environment-invariant. This is also beneficial via reducing over-fitting by learning a generalizable representation, which is applicable not only to the environment in question, but also to other environments with significant commonalities.

In our setting, since a shared network backbone is employed by two environments, additional data from the other environment act as a regularization which requires the system to perform well on the current environment.

To this end, we train an environment classifier, denoted as \(f_e\) with parameters \(W^e\), to classify which environment a particular datum comes from. For each feature representation from one specific feature extraction network \(f\), we learn the environment classifier with the following softmax loss. In each mini-batch, the loss is as follows:

\[
\mathcal{L}^e(f_e(W^e, E)) = - \lambda_1 \sum_{i=1}^{N^i} \sum_{k=1}^{K} \mathbb{I}[y_i = k] \log q_{ik}^e
\]

where \(\mathbb{I}\) is an indicator function with \(\mathbb{I}(\text{True}) = 1\) otherwise 0. The parameter \(\lambda_1\) controls the weight of this loss function. \(q_{ik}^e\) corresponding to the softmax of the environment classifier \(q_{ik}^e\), that is:

\[
q_{ik}^e = \text{softmax}(f_e(W^e, E_i)),
\]

and the probability of the \(k^{th}\) environment is:

\[
q_{ik}^e = \frac{e_{ik}^e}{\sum_{k=1}^{K} e_{ik}^e}
\]

As in [62], an adversarial-like learning objective is further introduced which aims at “maximally confusing” the two environments by computing the cross entropy between the output predicted environment labels and a uniform distribution over environment labels:

\[
\mathcal{L}^{adv}(f_e(W^e, E)) = - \lambda_2 \sum_{i=1}^{N^i} \sum_{k=1}^{K} \frac{1}{K} \log q_{ik}^e
\]

This confusion loss aims at learning an environment-invariant representation in a sense that the best environment classifier performs poorly. The parameter \(\lambda_2\) controls the weight of this loss function.

The two losses \(\mathcal{L}^e(f_e(W^e, E))\) and \(\mathcal{L}^{adv}(f_e(W^e, E))\) stand in direct opposition to one another. More specifically, learning a fully environment invariant representation requires the environment classifier give poor classification accuracy, and
learning an effective environment classifier means that the representation is environment distinguishable. Motivated by the recent success in adversarial learning [63], we optimize \( L^a(f_c(W^e, E)) \) and \( L^{adv}(f_c(W^e, E)) \) in an iterative manner. More specifically, we first optimize the \( W^e \) in \( L^a(W^e, E) \) with \( E \) being fixed to improve the environment classifier \( W^e \). Then we optimize \( E \) in \( L^{adv}(W^e, E) \) with \( W^e \) being frozen, aiming at learning an environment-invariant representation which could fuse the optimized environment classifier.

While training the network with the proposed adversarial learning works to align the final representation, the activity classifier has not been fully adopted to the new representations yet. To tackle this, for each environment, we propose to transfer the activity-oriented attributes from other environments. More specifically, for each environment \( i \), we manipulate the activity-oriented attributes \( A^i \) by transferring the the activity-oriented attributes \( A^i(j \neq i) \) from the other environments. That is, for each environment \( i \), we refine the system with the following loss function

\[
L^a(W^e, f_a(E^i, A^i)) = -\sum_{k=1}^{N} \sum_{c=1}^{C} \mathbb{I}[y_i = c] \log q_{k,c}^i, \tag{12}
\]

where \( q_{k,c}^i \) is defined in the same way as in Eqn. 7. In this way, the activity-oriented attributes \( A^i \), which is assumed to be disentangled from the environment-oriented attributes \( E^i \) in the source environment, are interacted with the environment-oriented attributes \( E^i \) from all other environments. By this way, the activity-oriented attributes from the other environment are well-transferred into the current environments.

C. Overall system

An illustration of the proposed system is presented in Fig. 2. Below we provide more details on the network structure and training procedures.

1) Network Structure: As mentioned above, our system consist of feature encoding network \( f_f \) in each environment, an activity classifier \( f_a \) and an environment classifier. We also include an auxiliary environment classifier \( f_e \). For the feature extraction network \( f_f \), we use a three layer fully convolutional network. The output channels for those layers are all set to be 10 and the kernel sizes are all set to be 5. We use ReLU activation function and insert a batchnorm layer after the ReLU operator. A maxpooling layer with a kernel size of 3 is also used after each batchnorm layer. As for the Working in this way, we get the factorized representation of \( F^i \) which has 720 dimensions. \( F^i \) is further splitted into two non-overlapped subset for \( E^i \) and \( A^i \) which has 500 and 960 dimensions, respectively. As for the Johnson-Lindenstrauss transform, we did not perform any dimensionally reduction and thus use a 710 × 720 square matrix. The noise matrix is sampled from \( N(0,0.01) \). For the environment classifier, we use two fully-connected network with 512, 256 and 2 outputs in each layer, respectively. We also insert a ReLU and a batchnorm layer between them. For the activity classifier, we use a fully-connected layers network with 256 and 7 outputs in each layers, respectively.

Algorithm 2: Privacy-Preserving Cross-Environment Human Activity Recognition

**Inputs:** data and labels \((X^i,Y^i)\) in each environment where \( i = 1, \cdots, K; \)

**initialization:** initialize the parameter of \( W_f^i, W^a \) and \( W^e \), which corresponds to the parameter of feature encoding network in the \( i^{th} \) environment \((i = 1, 2, \cdots, K)\), the activity classifier, and the environment classifier respectively.

**Outputs:** Fully optimized \( W_f^i \) \((i = 1, 2, \cdots, K)\), \( W^a \) and \( W^e \);

**Main Optimization Procedure:**

1. The local server in each environment sends its encoded features to the central server by Algorithm 1.
2. The central sever optimize \( W^a \) by Eqn. 5 and sent the gradient to each environment;
3. The local server in each environment updates the \( W_f^i \) based on the received gradient and Eqn. 4 and Eqn. 3 .
4. The central server optimize the \( W^e \) by Eqn. 8 .
5. The central server send the gradient to each environment by Eqn. 11 .
6. The local server in each environment updates the \( W_f^i \) based on the received gradient and Eqn. 4 and Eqn. 3 .
7. The central server optimize \( W^a \) by Eqn. 12 and sent the gradient to each environment;
8. The local server in each environment updates the \( W_f^i \) based on the received gradient and Eqn. 4 and Eqn. 3 .
9. epoch=epoch+1.

end

IV. EXPERIMENTAL VALIDATION

In this section, we elaborate the details of our experiment. We first describes the dataset we used in this study followed by our experimental setup. Then, the experimental results are summarized and discussed. Moreover, we provide some ablative study to further understand the merits of the proposed system.

A. Dataset and Experimental Setup

We collect the dataset from two typical real-life indoor environments: (a) Environment 1 is a conference room covering a 7.2 × 12 m² area and consisting of standard furniture: tables, chairs, boards, etc. (b) Environment 2 is computer laboratory at a campus which contains many tables and desktops and covers an area of 16 × 32 m². Fig. 3 shows the layouts of the meeting room and the laboratory. In all environments, we use a commercial WiFi router (TP-LINK) with three antenna as a transmitter. For a receiver, we use a Lenovo laptop equipped with an Intel 5300 NICs, running with Ubuntu operating system,
and installed with the tool provided in [64]. We use three antennas of receiver to collect data packets at a transmission rate of 500 Hz. At each time instant, $1 \times 3 \times 30$ CSI streams are recorded during the measurements. A sliding window with a window size of $4s$ is used to construct CSI vector. Thus, the shape of each sample is $2000 \times 90$.

Fig. 3: The layouts of the two environments for experiments. Environment 1 is a conference room and environment 2 is a laboratory room. The CSI measurements for each activity (except run and walk) are collected at the location marked with triangles. For the activities of “Run” and “Walk”, the volunteers follow the path marked with dashed lines near Tx and Rx. In each testing environments, we ask all volunteers to perform each activity at each location 22 times. To ensure the generality of the collected CSI data, the volunteers are requested to change their orientation randomly during performing their activities.

Totally, seven volunteers participate in our experiment, and perform seven common daily activities of “Empty”, “Jump”, “Pick up”, “Run”, “Sit down”, “Wave hand” and “Walk”. The volunteers include 5 male and 2 female graduate/undergraduate students with ages in the range of $18 - 27$, and heights in the range of $1.6m$ to $1.8m$. The experiments are conducted in different days in a month, with 4312 samples for each environment and 8624 samples in total. The entire data collection process is recorded by the camera to obtain ground truth.

The proposed system is implemented in Pytorch [65] in Ubuntu 16.04 environment with a Nvidia Titan-X graphic card. The batchsize is set to be 32. The epoch number $N$ is set to be 200 and the learning rate is $5e-3$. For all the learning module, we decrease the learning rate by a factor of 0.1 for every 50 epochs. For the weight parameter $\lambda_1$ and $\lambda_2$ in Eqn. [8] and [11] we set them as $1e-5$. We conduct a stratified sampling strategy to select 20% samples from each class and each environment as the testing set and the remaining as the training set. The parameters of all the approaches are carefully tuned using a small validation set from the training data.

B. Main Results

We firstly present our main results on the raw data in Table [1] We compare the proposed method with different baseline methods. More specifically, in “Baseline 1”, we train a network consisted of only feature encoding module and activity classifier module with data samples from environment 1. In the same way, in “Baseline 2” we train the system only with data samples in Environment 2. For the method of “Combined Training”, we train the system with all the data samples from all the environments. We also compare the proposed method with the federated learning strategy introduced in [60]. It protects the data privacy from an engineering perspective by using training data distributed on the different devices, and learning a shared model by aggregating locally-computed updates. From the results we can see that i) Both Baseline 1 and Baseline 2 perform significant worse than the other three methods. The reason is that, different environment may have different spatial layout and thus the objects in the environment could generate different multi-path reflections in wireless signals from a TX to a RX. Hence, directly deploying the trained system in the new environment may lead to poor performances. ii) The proposed method performs on par with the “Combined Training” strategy. This is in line with the existing literature by showing that the Johnson-Lindenstrauss transform performs well in maintaining the geometric structure of the input features. Most importantly, it also satisfies the differential privacy in Definition 1 in a sense that the privacy of individual person being involved are preserved. In Fig. 4, we visualize the confusion matrix of the proposed method with one baseline method. More specifically, in Fig. 4a, we train the network with the training data in Environment 1 and test it under both Environment 1 and Environment 2. Accordingly, in Fig. 4b, we train the network with the proposed method. As shown in Fig. 4, we can observe that: 1) For the baseline method, the “wave” would be easily misclassified, as “wave” is of slighter body movement (i.e., only has hand movements) than others, which make it vulnerable to environment changes. 2) The class of “walk” could be easily fooled by “run” and vice versa. This is straightforward to understand because these two activities behave similarly and thus could yield similar motion patterns. 3) The proposed method show consistent improvement over all the categories due to the leverage of the information from both environments.

After showing the advantage of the proposed method on the raw input data, we further investigate its performance on other input modalities. More specifically, we use 5 levels DWT [59] to extract the handcrafted features from the raw data, and the extracted features of dimension 189 are obtained. Then we apply the same proposed method on those features and the

1 We use the public available implementation from https://github.com/AshwinRJ/Federated-Learning-PyTorch
TABLE II: Performance Comparison of the proposed method with different baseline methods on the raw data.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy 1 (%)</th>
<th>Accuracy 2 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline 1</td>
<td>83.59</td>
<td>55.91</td>
</tr>
<tr>
<td>Baseline 2</td>
<td>59.83</td>
<td>87.45</td>
</tr>
<tr>
<td>Combined Training</td>
<td>84.02</td>
<td>92.18</td>
</tr>
<tr>
<td>Federated Learning</td>
<td>83.39</td>
<td>91.07</td>
</tr>
<tr>
<td>Ours</td>
<td>85.57</td>
<td>92.31</td>
</tr>
</tbody>
</table>

By comparing the results in both Table II and Table III, we could easily find that by using the raw data, a better performance could be achieved. The reason could be that, during the extracting of the handcrafted features, some information could be inevitably lost. However, we argue that the handcrafted feature based models may still contain complementary knowledge and could be used to further enhance the final recognition performances [67]. To demonstrate this, we perform a late fusion strategy, as done in other computer vision tasks [68], [6], to fuse the predicted probabilities of each classes from those two models. More specifically, we fused the softmax scores of the two models, which are trained with the raw data and the handcrafted features respectively, with a weighted averaging strategy. The final performances are summarized in Table IV. From the results we can see that the performances are further enhanced with this strategy.

TABLE IV: Performance Comparison of the proposed method on multi-modality inputs.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy 1 (%)</th>
<th>Accuracy 2 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Handcrafted Features</td>
<td>71.57</td>
<td>62.57</td>
</tr>
<tr>
<td>Raw Data</td>
<td>85.57</td>
<td>92.31</td>
</tr>
<tr>
<td>Fusion</td>
<td>86.67</td>
<td>92.88</td>
</tr>
</tbody>
</table>

C. Ablation Studies

In this section, we provide more ablative studies on several key parameters of the proposed method to further understand the its merits. The parameter we consider here are : 1) the dimensionality of the environment and activity-oriented attributes in Eqn. 4 and 2) The weights parameter \( \lambda_1 \) and \( \lambda_2 \) in Eqn. 8 and 11 respectively. 3) The effect of the J-L transform in terms of recognition accuracy. For simplicity, we use the system trained on the raw data as a study case.

Dimensionality of the environment and activity-oriented attributes: In order to investigate this parameter, we introduce a parameter \( \rho \) here to measure the dimensionality of both attributes. Here in this setting we consider the environment-oriented attributes have \( \rho \) percentage of whole features in the latent space and the activity-oriented attributes have the remaining \( 1-\rho \) percentage of features. A larger value of \( \rho \) means that the network focuses more on modelling the environment and a less value indicates that more efforts are spent to modelling the activities. In the experiments, as explained in Section III-C1, we use the first 240 features as the environment-oriented attributes which essentially makes \( \rho = 33.33\% \). More detailed studies are summarized in Table V. From the results, we can see that setting the dimensionality of the environment-oriented attributes to be 240, which means \( \rho = 33.33\% \), achieves best performances.

TABLE V: The effect of the dimensionality of environment and activity-oriented attributes.

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>Accuracy 1 (%)</th>
<th>Accuracy 2 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>73.88</td>
<td>87.62</td>
</tr>
<tr>
<td>33.33%</td>
<td>85.57</td>
<td>92.31</td>
</tr>
<tr>
<td>80%</td>
<td>81.21</td>
<td>89.78</td>
</tr>
</tbody>
</table>

Weight Parameter: the weights parameter \( \lambda_1 \) and \( \lambda_2 \) in Eqn. 8 and 11 control the importance of the adversarial learning part in the system. A larger value for these parameters indicates that the system pays more attention to learn environment-invariant representation, which in principle could be beneficial but also bring the risk of over-fitting. For simplicity in this study we keep \( \lambda_1 = \lambda_2 \) and we empirically observe that setting \( \lambda_1 \) and \( \lambda_2 \) to be \( 1e - 5 \) could lead to satisfactory results and detailed results are summarized in Table VI.

Effect of the J-L Transform: The J-L transform is used to preserve the privacy, as illustrated in Section III-A. However, it
with different input modalities on real-world datasets. We show an average of accuracy improvement of 1.47% over the challenging federated learning framework of \cite{66}. This work could open a path towards designing more advanced privacy preserving methods for wireless sensing applications and serve as a new baseline therein. One limitation of this work is that it only focuses on single-person activity recognition. In the future we plan to design more advanced network structures and learning strategies for group behavior analysis.

**REFERENCES**


Le Zhang is a professor with the School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC). He received his Ph.D. degree from Nanyang Technological University (NTU) in 2016. From 2016 to 2018, he was a postdoc fellow at Advanced Digital Sciences Center (ADSC), Singapore. From 2018 to 2021, he was a research scientist at Institute for Infocomm, Research (I2R), Agency for Science, Technology, and Research (A*STAR), Singapore. His research interests include computer vision and machine learning. He has published over 50 papers in leading journals and conferences, such as IEEE TPAMI, IJCV, IEEE TIP, IEEE CVPR, etc. He is an associate editor of Neurocomputing and IET Biometrics.

Wei Cui received the M.E. and Ph.D. degrees in pattern recognition and intelligent system from Northeastern University, Shenyang, China, in 2013 and 2017, respectively. Currently, she is a scientist at Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore. Her current research interests include wireless sensor networks, wireless sensing, and machine learning.

Bing Li received his Ph.D. degree from Northeastern University, Shenyang, China, in 2018. From 2018 to 2021, he was a research fellow at University of New South Wales, Australia. Currently, he is a research scientist at IHPC, A*STAR, Singapore. His research interests include Text Mining, DB+AI, NLP, and AI-based IoT applications. He published more than 20 papers on prestigious conferences or journals, i.e., AAAI, IJCAI, TKDE, and served as PC members and invited reviewers for many conferences and journals, such as ICLR, ICML, AAAI, TKDE, etc. He was awarded ACM China SIGMOD DOCTORAL DISSERTATION AWARD in 2019.

Zhenghua Chen received the B.Eng. degree in mechatronics engineering from University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2011, and Ph.D. degree in electrical and electronic engineering from Nanyang Technological University (NTU), Singapore, in 2017. He has been working at NTU as a research fellow. Currently, he is a Scientist and Lab Head at Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore. He has won several competitive awards, such as First Place Winner for CVPR 2021 UG2+ Challenge, A*STAR Career Development Award, First Runner-Up Award for Grand Challenge at IEEE VCIP 2020, Finalist Academic Paper Award at IEEE ICPHM 2020, etc. He serves as Associate Editor for Elsevier Neurocomputing and Guest Editors for five journals. He is currently the Vice Chair of IEEE Sensors Council Singapore Chapter and IEEE Senior Member. His research interests include smart sensing, data analytics, machine learning, transfer learning and related applications.

Min Wu is currently a Senior Scientist in Machine Intelligence Department, Institute for Infocom Research, Agency for Science, Technology and Research (A*STAR), Singapore. He received his Ph.D degree in Computer Science from Nanyang Technological University, Singapore in 2011 and B.S. degree in Computer Science from the University of Science and Technology of China, Hefei, China in 2006. He received best paper awards in InCoB 2016 and DASFAA 2015. He also won the IJCAI competition on repeated buyers prediction in 2015. His current research interests include machine learning, graph mining and bioinformatics.

Teo Sin Gee is a research scientist at the Institute for Infocomm Research (I2) in Singapore. He obtained the PhD degree from Monash University, Australia in 2016. His research interests include applied cryptography, data privacy and security, malware and network anomaly classification, deep learning, federating learning and adversarial machine learning.