Towards Environment Independent Device Free Human Activity Recognition

Wenjun Jiang [1], Chenglin Miao [1], Fenglong Ma [1], Shuochao Yao [2], Yaqing Wang [1], Ye Yuan [3],
Hongfei Xue [1], Chen Song [1], Xin Ma [1], Dimitrios Koutsonikolas [1], Wenyao Xu [1], and Lu Su [1*]
[1] State University of New York at Buffalo, Buffalo, NY USA
[2] University of Illinois at Urbana-Champaign, Urbana, IL USA
[3] Beijing University of Technology, Beijing, China
Email: [1] {wenjunji, cmiao, fenglong, yaqingwa, hongfeix, csong5, xma24, dimitrio, wenyaoxu, lusu}@buffalo.edu, [2] syao9@illinois.edu, [3] yuanye91@emails.bjut.edu.cn

ABSTRACT

Driven by a wide range of real-world applications, significant efforts have recently been made to explore device-free human activity recognition techniques that utilize the information collected by various wireless infrastructures to infer human activities without the need for the monitored subject to carry a dedicated device. Existing device free human activity recognition approaches and systems, though yielding reasonably good performance in certain cases, are faced with a major challenge. The wireless signals arriving at the receiving devices usually carry substantial information that is specific to the environment where the activities are recorded and the human subject who conducts the activities. Due to this reason, an activity recognition model that is trained on a specific subject in a specific environment typically does not work well when being applied to predict another subject's activities that are recorded in a different environment. To address this challenge, in this paper, we propose EI, a deep-learning based device free activity recognition framework that can remove the environment and subject specific information contained in the activity data and extract environment/subject-independent features shared by the data collected on different subjects under different environments. We conduct extensive experiments on four different device free activity recognition testbeds: WiFi, ultrasound, 60 GHz mmWave, and visible light. The experimental results demonstrate the superior effectiveness and generalizability of the proposed EI framework.

CCS CONCEPTS

• **Networks** → *Wireless access points, base stations and infrastructure*; • **Human-centered computing** → *Interaction techniques*;

KEYWORDS

Human Activity Recognition; Device Free; Environment Independent

ACM Reference Format:
Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, Wenyao Xu, and Lu Su. 2018. Towards Environment Independent Device Free Human Activity Recognition. In *The 24th Annual International Conference on Mobile Computing and Networking (MobiCom '18), October 29-November 2, 2018, New Delhi, India.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3241539.3241548

1 INTRODUCTION

Human Activity Recognition (HAR) plays an important role in a wide range of real-world applications, such as smart home, health care and fitness tracking. Traditionally, smart mobile devices, including phones, watches, and other wearables, are widely used to recognize human activities. However, device-based approaches have many limitations due to the extra burden and discomfort brought to those who wear devices. To address this challenge, significant efforts are recently made to explore device-free human activity recognition techniques that utilize the information collected by various wireless infrastructures without the need for the monitored subject to carry a dedicated device.

These approaches, though different in various aspects, share the same idea: by extracting and analyzing information carried by the wireless signal transmitted between a pair

[*] Lu Su is the corresponding author.

MobiCom '18, October 29-November 2, 2018, New Delhi, India
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5903-0/18/10…$15.00
https://doi.org/10.1145/3241539.3241548

of wireless devices (e.g., smartphone, laptop, WiFi access point), we can infer the activities of a person located between the sender and receiver, since his/her activities would incur changes to the transmission pattern of the wireless signals.

Thus far, various device free human activity recognition approaches and systems have been developed. However, a major challenge has not been addressed. That is, the wireless signals arriving at the receiving devices usually *carry substantial information that is specific to the environment where the activities are recorded and the human subject who conducts the activities*. On one hand, the signals, when being transmitted, may be penetrating, reflected, and diffracted by the media (e.g., air, glass) and objects (e.g., wall, furniture) in the ambient environment. On the other hand, different human subjects with different ages, genders, heights, weights, and body shapes affect the signals in different ways, even if they are taking the same activity. As a result, an activity recognition model that is trained on a specific subject in a specific environment will typically not work well when being applied to predict another subject's activities that are recorded in a different environment.

To address this challenge, in this paper, we propose **EI**, a deep-learning based device free activity recognition framework that can *remove the environment and subject specific information* contained in the activity data and *extract environment/subject-independent features* shared by the data collected on different subjects under different environments.

The core of EI is an *adversarial network*, which consists of three main components: *feature extractor*, *activity recognizer*, and *domain discriminator*. The feature extractor, which is a *Convolutional Neural Network* (CNN), cooperates with the activity recognizer to carry out the major task of recognizing human activities, and simultaneously, tries to fool the domain discriminator to learn the environment/subject-independent representations.

To deal with the practical yet challenging scenarios where for most of the environments/subjects, the collected activity data are *unlabeled*, the proposed model not only makes use of labeled data, but also takes advantage of the information contained in the unlabeled data. In addition, to tackle various practical issues, in the proposed model, we also design three constraints that can significantly improve the prediction performance.

We conduct extensive experiments on **FOUR** different device free activity recognition testbeds, based on different wireless technologies: **WiFi**, **ultrasound**, **60 GHz mmWave**, and **visible light**. The experimental results demonstrate the superior effectiveness and generalizability of the proposed EI framework.

The rest of this paper is organized as follows. We first provide an overview of the proposed EI framework in Section 2. Then we elaborate on each component of the proposed deep learning model in Section 3. In Section 4, we conduct a series of experiments on four different device free activity recognition testbeds to evaluate the performance of the proposed framework. We discuss the related work in Section 5 and conclude the paper in Section 6.

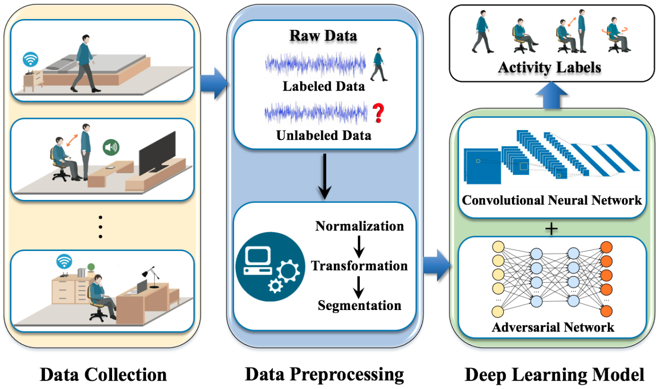


Figure 1: System framework

2 SYSTEM OVERVIEW

In this section, we provide an overview of the proposed EI framework. As shown in Fig. 1, EI consists of three components: data collection, data preprocessing and deep learning model.

- **Data Collection.** In this paper, we consider a scenario where the human activities are monitored in different environments (e.g., different rooms), and in each environment there are some ambient devices whose generated signals (e.g., WiFi and acoustic) can be affected by human activities. Our system first collects the activity data (i.e., the affected signals) in each environment during the monitoring process.
- **Data Preprocessing.** For some environments, part of the collected data are manually labeled, and for the others, the label information is not provided. Our goal is thus to train a prediction model based on all the collected data including both labeled and unlabeled data to predict the label of each unlabeled activity. In order to achieve the goal, we first normalize the acquired signal and then transform the signal to a form suitable for analysis. Finally we split the transformed signal into short segments to train the activity recognition model. The detailed descriptions of the data preprocessing for different signals are provided in Section 4.
- **Deep Learning Model.** The collected activity data, after being preprocessed, may still be very complex. This makes it difficult for traditional machine learning algorithms to characterize the underlying patterns of such data. To address this challenge, we make use

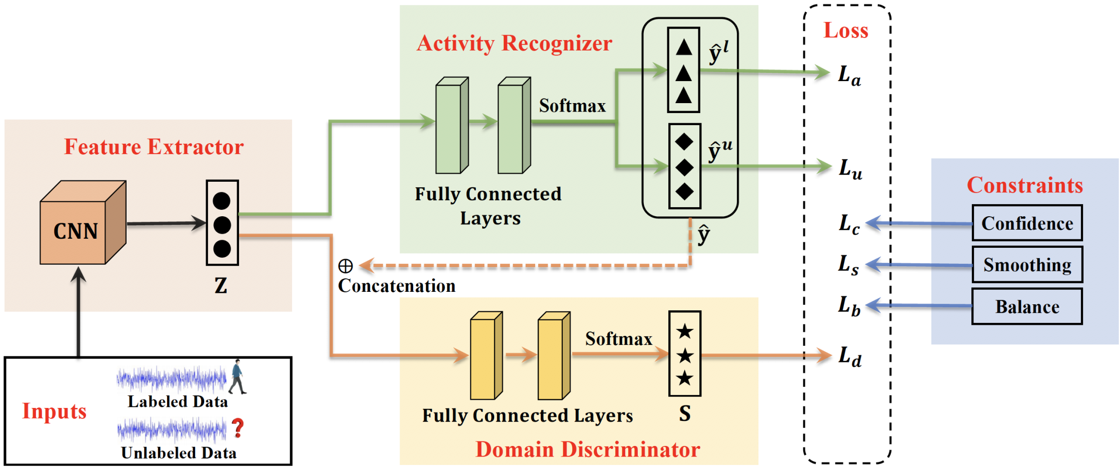


Figure 2: Model Overview.

of deep learning techniques which have been proved effective in deriving discriminative representations from complex data. In particular, we propose a deep learning model, which incorporates an adversarial network, to predict the label of unlabeled activities. The proposed deep learning model can not only make use of labeled data, but also take advantage of the information contained in the unlabeled data that can help improve the predictive performance. Additionally, the proposed model is able to remove the uniqueness of each ***domain*** (defined as *a pair of environment and human subject*), and extract commonness shared across different domains. Therefore, it can be used to predict the labels of the activities recorded under unseen environments.

3 METHODOLOGY

An overview of the proposed deep learning model is shown in Fig. 2. The input data of our model includes both labeled and unlabeled human activities. In this paper, we consider a general and practical problem setting: the environments for collecting labeled data are different from the ones where unlabeled data are collected. This problem setting requires that the proposed approach must be able to learn transferable features for different environments, i.e., *environment-independent representations.*

Towards this goal, the input data are first transformed into low-dimensional representations $\mathbf{Z}$ by the component of **feature extractor**, which consists of three-layer convolutional neural networks (CNNs). Using the learned feature representations, the **activity recognizer**, whose goal is to maximize the prediction accuracy, can obtain the predictions $\hat{\mathbf{y}}$ on all the input data. To remove domain-specific features, a **domain discriminator** is designed to label each domain (i.e., to identify which activities are conducted by *which subject* under *which environment*). The input of the domain discriminator is the concatenation of $\mathbf{Z}$ and $\hat{\mathbf{y}}$. After two fully connected layers with softmax, we can obtain the domain label distributions $\mathbf{S}$. The goal of domain discriminator is to maximize the performance of domain label prediction, which seemingly contradicts with our ultimate goal of learning domain-independent features of activities. To address this contradiction, in our design, the feature extractor tries its best to cheat the domain discriminator (i.e., minimize its predictive accuracy), and at the same time, boost the performance of the activity recognizer. Through this minimax game, the proposed model can finally learn the common environment-independent features for all the activities.

Besides, we design three **constraints** that can significantly improve the prediction performance. The details of our model will be elaborated in the rest of this section.

3.1 Model Inputs

The proposed model can recognize human activities with different types of signals, including WiFi, ultrasound, 60 GHz millimeter wave, and visible light. Below we provide a general description of the model inputs. The details on how these signals are transformed into the input to the model can be found in Section 4.

First, we refer to the domains with and without label information as source and target domain, respectively. In this paper, we consider the scenario of multiple source and target domains. Let $\mathbf{X}$ be the input activity data of the proposed model, which includes two parts: labeled human activities $\mathbf{X}^l$ and unlabeled ones $\mathbf{X}^u$. Each data $\mathbf{X}_i$ has a corresponding domain label $d_i \in \mathcal{D}$, where $\mathcal{D}$ denotes the set of all the source and target domains. Each labeled data $\mathbf{X}_i^l \in \mathbf{X}^l$ also has a true activity label $y_i^l \in \mathcal{Y}$, where $\mathcal{Y}$ is the set of all the activities. Let $\mathbf{d}$ denote the domain label vector of $\mathbf{X}$, and $\mathbf{y}^l$ be the ground truth vector of $\mathbf{X}^l$. Thus, the inputs of our model are the activity data $\mathbf{X}$, the domain label vector $\mathbf{d}$ and the ground truth data $\mathbf{y}^l$. The output is the estimated label y_i^u of each unlabeled activity $\mathbf{X}_i^u \in \mathbf{X}^u$.

3.2 Feature Extractor

We employ CNNs to extract activity features, which are widely used in the human activity recognition task [58]. In the proposed approach, we use three-layer stacked CNNs to extract features. In each layer of CNNs, 2D kernels are used as the filters, followed by a batch norm layer to normalize the mean and variance of the data at each layer. At last, we add a rectified linear unit (ReLU) to introduce nonlinearity and a max-pooling layer to reduce the size of representation. Let $\mathbf{\Theta}$ be the set of CNN parameters. Given the input data $\mathbf{X}$, we can obtain their feature representations as follows:

$$\mathbf{Z} = \mathrm{CNN}(\mathbf{X}; \mathbf{\Theta}). \tag{1}$$

3.3 Activity Recognizer

Based on the outputs of feature extractor (i.e., $\mathbf{Z}$), a fully-connected layer followed by an activation function is used to learn the representation $\mathbf{V}_i$ of $\mathbf{X}_i$ as follows:

$$\mathbf{V}_i = \mathrm{Softplus}(\mathbf{W}_z \mathbf{Z}_i + \mathbf{b}_z), \tag{2}$$

where $\mathbf{W}_z$ and $\mathbf{b}_z$ are the parameters to be learned and the softplus function is an activation function to introduce nonlinearity. In order to predict the labels of human activities, we need to map the feature representation $\mathbf{V}_i$ into a new latent space $\mathbf{H}_i \in \mathbb{R}^C$, where C is the number of human activities. Moreover, a softmax layer is used to obtain the probability vector of activities as follows:

$$\hat{\mathbf{y}}_i = \mathrm{Softmax}(\mathbf{H}_i) \text{ and } \mathbf{H}_i = \mathbf{W}_v \mathbf{V}_i + \mathbf{b}_v, \tag{3}$$

where $\mathbf{W}_v$ and $\mathbf{b}_v$ are parameters. The input data of the proposed model include labeled and unlabeled activities, and thus $\hat{\mathbf{y}} = [\hat{\mathbf{y}}^l, \hat{\mathbf{y}}^u]$, where $\hat{\mathbf{y}}^l$ denotes the predicted probabilities of labeled data, and $\hat{\mathbf{y}}^u$ represents the predicted probabilities of unlabeled data.

For the labeled data, cross entropy function can be used to calculate the loss between the predictions and the ground truths as follows:

$$L_a = -\frac{1}{|\mathbf{X}^l|} \sum_{i=1}^{|\mathbf{X}^l|} \sum_{c=1}^{C} \mathbf{y}_{ic}^l \log(\hat{\mathbf{y}}_{ic}^l), \tag{4}$$

where $|\mathbf{X}^l|$ is the number of data with labels. Actually, directly optimizing Eq. (4) suffices to learn model parameters and make predictions on unlabeled data. However, when label information is limited, incorporating unlabeled data can help the proposed model improve the predictive performance. Actually, for unlabeled data, we also can calculate their losses using entropy as follows:

$$L_u = -\frac{1}{|\mathbf{X}^u|} \sum_{i=1}^{|\mathbf{X}^u|} \sum_{c=1}^{C} \hat{\mathbf{y}}_{ic}^u \log(\hat{\mathbf{y}}_{ic}^u), \tag{5}$$

where $|\mathbf{X}^u|$ is the number of unlabeled data. By minimizing the entropy in Eq. (5), we can increase the confidence of the predictions on unlabeled data, and thus drive the classifier's decision boundary away from unlabeled data [13].

In this paper, we consider a practical yet challenging scenario of human activity recognition, that is, for a significant portion of the domains (i.e., environment-subject pairs), no activity data are labeled. This requires the classifier to be able to learn the common activity features shared by all the domains, i.e., transferable activity representations for new or unseen domains. Such features should be environment-independent and do not contain any domain-specific information. To achieve this goal, we need to remove the uniqueness of activities in each domain. Specifically, we use domain adaption technique to capture the environment-independent activity features.

3.4 Domain Discriminator

Domain adaptation is a technique that aims to learn a mapping among domains. When the target domains are fully unlabeled, the technique is called unsupervised domain adaptation [10]. In this paper, we employ the technique of unsupervised domain adversarial training [10, 11] to fully make use of unlabeled data to remove the domain-specific uniqueness of activities. In particular, we aim to design a domain discriminator, whose goal is to recognize the environment where the activities are recorded, to force the feature extractor (whose goal is to cheat the domain discriminator) to produce environment-independent activity features.

To achieve this goal, similar to [60], we first concatenate the output matrix of feature extractor (i.e., $\mathbf{Z}$) and the prediction matrix $\hat{\mathbf{y}}$ as follows:

$$\mathbf{F} = \mathbf{Z} \oplus \hat{\mathbf{y}}, \tag{6}$$

where $\oplus$ is the concatenation operation. Since $\mathbf{Z}$ contains both domain-independent and domain-specific features, to identify the commonness shared across different domains, we need to take $\mathbf{Z}$ into consideration. Moreover, some features, though being domain-specific, are helpful to the activity recognition task. Thus, we still need to keep such features. This can be achieved by concatenating $\mathbf{Z}$ and $\hat{\mathbf{y}}$ as the input of domain discriminator.

Then, two fully connected layers with corresponding activation functions are used to project $\mathbf{F}$ into domain distributions $\mathbf{S}$, as follows:

$$\mathbf{U}_i = \mathrm{Softplus}(\mathbf{W}_f \mathbf{F}_i + \mathbf{b}_f), \tag{7}$$

$$\mathbf{S}_i = \text{Softmax}(\mathbf{W}_u \mathbf{U}_i + \mathbf{b}_u), \tag{8}$$

where $\mathbf{W}_f$, $\mathbf{b}_f$, $\mathbf{W}_u$ and $\mathbf{b}_u$ are parameters. $\mathbf{U}_i$ is the representation in the latent space. In order for the domain discriminator to identify the domain labels of the input activities, we define the loss between the domain distributions and true domain labels as follows:

$$L_d = -\frac{1}{|\mathbf{X}|} \sum_{i=1}^{|\mathbf{X}|} \sum_{j=1}^{|\mathcal{D}|} \mathbf{d}_{ij} \log(\mathbf{S}_{ij}), \tag{9}$$

where $|\mathcal{D}|$ denotes the number of domains, and $\mathbf{d}_i$ is the one-hot vector of true domain labels. The goal of the domain discriminator is to minimize the loss function L_d so as to maximize the performance of domain label prediction, which contradicts with our ultimate goal of learning domain-independent features of activities. To address this contradiction, we propose to maximize the domain discriminator loss L_d in our final objective function. Based on Eq. (4), Eq. (5) and Eq. (9), we can obtain the loss function as follows:

$$L = L_a + \alpha L_u - \beta L_d, \tag{10}$$

where α and β are the weighting parameters. From Eq. (10), we can observe that the feature extractor tries its best to cheat the domain discriminator by maximizing L_d, and at the same time, boost the performance of the activity recognizer by minimizing both L_a and L_u. Through this minimax game, we can learn the common environment-independent features for all the activities and finally obtain the predicted labels for unlabeled data.

3.5 Constraints

It is known that without sufficient data, deep neural networks are prone to overfitting, which often leads to unsatisfactory performance. In practical device-free activity recognition scenarios, it is usually difficult to collect sufficient activity data. Therefore, how to prevent overfitting with limited data is vital for the design of our unsupervised domain adaptation model. In order to tackle the overfitting problem, we propose two effective constraints: *confidence control constraint* and *smoothing constraint*. They are designed to handle the overconfidence and the unsmooth latent space of deep neural networks, two typical symptoms of overfitting [41].

To further improve the model's performance, we also propose a *balance constraint* that can incorporate the prior knowledge of the labels' distribution in the training data to improve the stability of training process.

3.5.1 Confidence Control Constraint. One symptom of overfitting is the overconfidence of the model when it places all probability on a single class in the training set [41]. If the model is overconfident on the estimation of the unlabeled data, it may converge prematurely and get stuck in an inferior local optimum, which may degrade the performance of the model on testing.

To address this issue, we propose a confidence control constraint, which penalizes $\hat{\mathbf{y}}_{ic}$ when it is too confident. The loss of the confidence control constraint is defined as follows:

$$L_c = -\frac{1}{|\mathbf{X}|} \sum_{i=1}^{|\mathbf{X}|} \sum_{c=1}^{C} (\log(\hat{\mathbf{y}}_{ic}) + \log(1 - \hat{\mathbf{y}}_{ic})), \tag{11}$$

In this way, if $\hat{\mathbf{y}}_{ic}$ approaches 0 or 1, the penalty will go to infinity.

3.5.2 Smoothing Constraint. Unsmooth latent space is another common symptom of overfitting. It happens when the prediction on a data point $\mathbf{X}_j$ is significantly different from those of its neighbors in the feature space $\mathbf{Z}$ (i.e., the classifier abruptly changes its predictions across neighboring data samples). In such a situation, the proposed model will learn an unreliable estimation [39]. Under the unsupervised domain adaption setting, there is no labeled information to penalize the wrong predictions for the unlabeled data in target domains through the loss function (i.e., Eq. (10)), which will aggravate the unsmoothing problem. To avoid this problem, we add a smoothing constraint to the loss function Eq. (10).

In supervised domain adversarial training models [54], it is easy to add a smoothing constraint. If a pair of data has the same label, then the distance between them in the feature space is short. However, in the unsupervised domain adaptation setting, some data samples do not have labels. Thus, such approaches cannot be directly applied. To solve this problem, we propose to add M ϵ-neighbors to each input sample $\mathbf{X}_i$ in its latent feature space $\mathbf{V}_i$. This is equivalent to adding Gaussian noise $\mathbf{r_m}$ to $\mathbf{V}_i$, denoted as $\mathbf{V}_i^m = \mathbf{V}_i + \mathbf{r}_m$. Then the Jensen-Shannon divergence between the predictions of $\mathbf{V}_i$ and $\mathbf{V}_i^m$ is calculated as the loss value of the smoothing constraint.

Mathematically, we add M small centered isotropic Gaussian noise $\mathbf{r}_m \sim \mathcal{N}(\mathbf{0}, \epsilon\mathbf{I})$ ($m \in \{1, \cdots, M\}$) to the latent representation $\mathbf{V}_i$. We also enforce that after passing the label predictor (i.e., Eq. (3)), the label distribution predicted from the noisy representation denoted as $\hat{\mathbf{y}}_i^m$ should be close to that predicted from the original latent representation (i.e., $\hat{\mathbf{y}}_i$). We achieve this through minimizing the Jensen-Shannon divergence between them. Jensen-Shannon divergence is a method of measuring the similarity between two probability distributions. It is based on Kullback-Leibler divergence, but is symmetric and always returns a finite value. Assuming that the Kullback-Leibler divergence between distributions $\hat{\mathbf{y}}_i$ and $\hat{\mathbf{y}}_i^m$ can be expressed as $\text{KL}(\hat{\mathbf{y}}_i || \hat{\mathbf{y}}_i^m)$, then the Jensen-Shannon divergence between them is defined as follows:

$$\mathrm{JS}(\hat{\mathbf{y}}_i||\hat{\mathbf{y}}_i^m) = \frac{1}{2}\mathrm{KL}(\hat{\mathbf{y}}_i||\frac{\hat{\mathbf{y}}_i + \hat{\mathbf{y}}_i^m}{2}) + \frac{1}{2}\mathrm{KL}(\hat{\mathbf{y}}_i^m||\frac{\hat{\mathbf{y}}_i + \hat{\mathbf{y}}_i^m}{2}). \quad (12)$$

Thus, the average loss of the smoothing constraint can then be formulated as follows:

$$L_s = \frac{1}{|\mathbf{X}|}\sum_{i=1}^{|\mathbf{X}|}\frac{1}{M}\sum_{m=1}^{M}\mathrm{JS}(\hat{\mathbf{y}}_i||\hat{\mathbf{y}}_i^m). \quad (13)$$

3.5.3 Balance Constraint. We observe that, in some cases, the model tends to assign the same label to the data samples corresponding to multiple similar but different activities. To deal with this issue, we propose to add a balance constraint to the loss function, which first estimates the percentage of each activity according to our prior knowledge or labeled data, and then enforces the estimated percentage in the final prediction of the activities. In particular, let P_c be the estimated or known overall percentage of activity c. After predicting the labels of $|\mathbf{X}|$ samples, we can obtain a prediction matrix with size $|\mathbf{X}| \times C$. $\hat{\mathbf{y}}_{ic}$ is the probability of $\mathbf{X}_i$ being labeled as the c-th activity, and d_i is its domain label. We introduce an auxiliary distribution $\mathbf{q}_i$ to be the balanced label prediction probability. We calculate $\mathbf{q}_{ic}$ by normalizing the total number of predictions on activity c with the same domain label d_i:

$$\mathbf{q}_{ic} = \frac{P_c \cdot \hat{\mathbf{y}}_{ic}/\sum_{i'}\hat{\mathbf{y}}_{i'c} \cdot \mathbb{1}_{d_{i'}=d_i}}{\sum_c P_c \cdot \hat{\mathbf{y}}_{ic}/\sum_{i'}\hat{\mathbf{y}}_{i'c} \cdot \mathbb{1}_{d_{i'}=d_i}} \quad (14)$$

After obtaining the auxiliary distribution $\mathbf{q}_i$, we define the balance constraint as the Jensen-Shannon divergence between $\hat{\mathbf{y}}_i$ and $\mathbf{q}_i$ as follows:

$$L_b = \frac{1}{|\mathbf{X}|}\sum_{i=1}^{|\mathbf{X}|}\mathrm{JS}(\hat{\mathbf{y}}_i||\mathbf{q}_i). \quad (15)$$

3.6 Objective and Training

With all the above constraints, we can finally give the overall loss function as follows:

$$J = L + \gamma L_s + \eta L_b + \pi L_c, \quad (16)$$

where γ, η and π are predefined hyper-parameters.

In the training process, we iteratively update the parameters. Let $\Omega = \{\Delta, \Gamma\}$ be the set of all the parameters, where $\Delta = \{\mathbf{W}_f, \mathbf{b}_f, \mathbf{W}_u, \mathbf{b}_u\}$ denotes the parameters in the domain discriminator, and $\Gamma = \Omega - \Delta$. We first fix Δ and update the remaining parameters (i.e., Γ) according to Adam [22], and then fix Γ to update Δ.

4 EXPERIMENTS

In this section, we conduct experiments on four different device free activity recognition testbeds, i.e., WiFi, ultrasound, 60 GHz mmWave and visible light, to evaluate the performance of the proposed system.

4.1 Baseline Methods

We compare our approach with two state-of-the-art domain adaptation deep learning models CAT [60] and VADA [39] as well as random forest, one of the most widely used traditional classification models. In its original design, CAT model uses only labeled data on the source domain. For a fair comparison with our model, we let CAT incorporate unlabeled data on the target domains. We also slightly change the loss function of the domain discriminator in VADA so that it can fit our multi-source, multi-target domain adaptation scenario. In addition, both of the deep learning models adopt the same CNN architecture as our approach for a fair comparison. For random forest, we extract 10 statistic features from both time and frequency domain. The time-domain features include: mean, standard deviation, relative standard deviation, mean absolute deviation, max, min, energy, and interquartile range. The frequency-domain features include dominant frequency and mean frequency. Especially, for acoustic signals, we utilize MFCC features.

4.2 Experiment with WiFi Signals

4.2.1 Channel State Information. In this section, we make use of the Channel State Information (CSI) to analyze the effect of the human activities on the WiFi signal. CSI refers to known channel properties of a communication link in wireless communications. This information describes how a signal propagates from the transmitter to the receiver and represents the combined effect of, for example, scattering, fading, and power decay with distance [1]. Modern WiFi devices supporting IEEE 802.1n/ac standards have multiple transmitting and receiving antennas, and thus can transmit data in MIMO (Multiple-Input Multiple-Output) mode. In an Orthogonal Frequency Division Multiplexing (OFDM) system, the channel between each pair of transmitting and receiving antennas consists of multiple subcarriers. The narrow band flat-fading channel with N_t transmitters and N_r receivers on the s-th subcarrier ($s \in \{1, 2, \cdots, N_s\}$) can be modeled as:

$$y = {H_s}^T \times x + n, \quad (17)$$

where $y \in \mathbb{C}^{N_r \times 1}$ denotes the received vector, $H_s \in \mathbb{C}^{N_t \times N_r}$ is the channel matrix over the s-th subcarrier, $x \in \mathbb{C}^{N_t \times 1}$ is the transmitted vector, and $n \in \mathbb{C}^{N_r \times 1}$ represents the noise vector. Noise is often modeled as circularly-symmetric complex normal with $n \sim \mathcal{CN}(0, S)$ where the mean value is zero and the noise covariance matrix S is known. The CSI value for each subcarrier is an estimate of H_s. Since there are N_s subcarriers, the final CSI can be represented by a multi-dimensional matrix $H \in \mathbb{C}^{N_s \times N_t \times N_r}$. We use the tool in [16] to report CSI values of 30 OFDM subcarriers. Thus, the dimensionality of H is $30 \times N_t \times N_r$. The reason why

[1] https://en.wikipedia.org/wiki/Channel_state_information

CSI can be used for recognizing human activities is mainly because it is easily affected by the presence of humans and their activities. Specifically, the human body may block the Line-of-Sight (LOS) path and attenuate the signal power. Additionally, the human body can introduce more signal reflections and change the number of propagation paths. Thus, the variance of CSI can reflect the human movements in the WiFi environments.

4.2.2 Experimental Settings. In this experiment, we employ 11 volunteers (including both men and women) as the subjects and collect CSI data from 6 different rooms in two different buildings. Figure 3 shows the Experimental setting in one of the rooms. In particular, we build a WiFi infrastructure, which includes a transmitter (a wireless router) and two receivers. We choose to use the Intel Wireless Link 5300 NIC to collect the CSI data, and the transmission rate is set as 200 packets per second. The human activities (shown in Fig. 4) conducted by the subjects include wiping the whiteboard, walking, moving a suitcase, rotating the chair, sitting, and standing up and sitting down. We let the subjects repeat these six activities in each room for 5 rounds and in each round, the subjects are asked to take each type of activity for 51 seconds. Totally, we collect the activity data of 40 subject-room pairs, corresponding to 40 different domains.

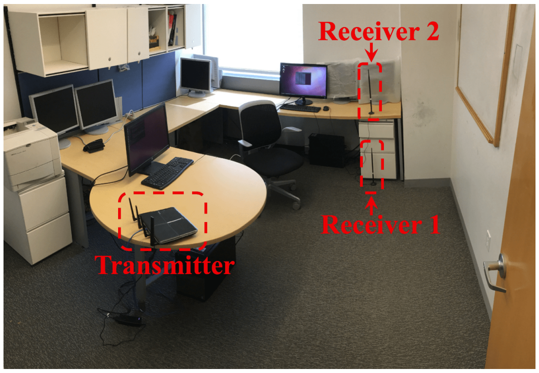


Figure 3: Experimental setting for human activity recognition with WiFi signals.

4.2.3 Data Preprocessing. In this experiment, the CSI measurements we use are the amplitude information of the subcarriers. Due to the packet loss during the data collection process, we first interpolate the CSI measurements to obtain uniform sampling periods and then normalize the CSI measurements to have a mean of zero and standard deviation of one. After that we use the Hampel filter [7] to remove outliers and downsample the CSI measurements into 25 Hz. We segment the CSI measurements every 128 samples with 32 samples overlap, which corresponds to the human activity of about 5.12 seconds. For each segment from the two receivers, we calculate the correlation between the segment and the segments lagged by no more than τ time units. We set τ to be 128 in our experiments. Then we combine them with the FFT of each segment as the input to the deep learning model.

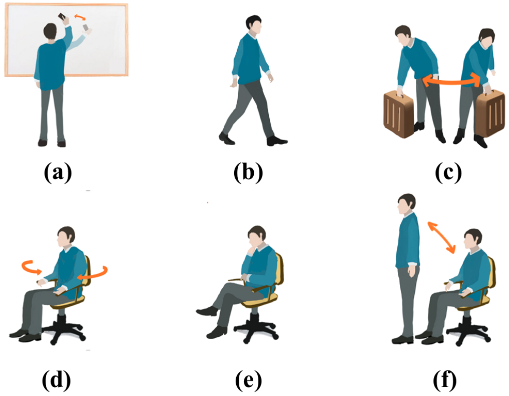


Figure 4: Human activities used to evaluate the performance of EI. (a) Wiping the whiteboad; (b) Walking; (c) Moving a suitcase; (d) Rotating the chair; (e) Sitting; (f) Standing up and sitting down.

4.2.4 Performance Evaluation. We first quantitatively analyze the performance of the proposed EI framework on the CSI dataset and compare it with the baselines. We randomly divide the CSI dataset into source domains (i.e., the subject-room pairs with labeled activities) and target domains (where no activities are labeled), and at the same time, ensure that the rooms in source and target domains are different. In this experiment, there are 22 source domains (11 volunteers in 3 rooms) and 18 target domains (10 volunteers in 3 rooms), and 10 volunteers are involved in both source and target domains. We gradually increase the number of source domains from 2 to 22, and use accuracy as the measure of evaluation. Figure 5 shows the results on the CSI dataset.

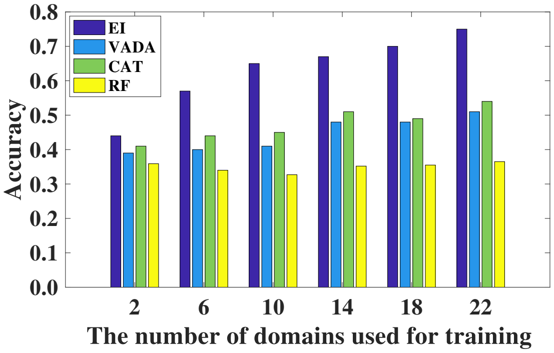


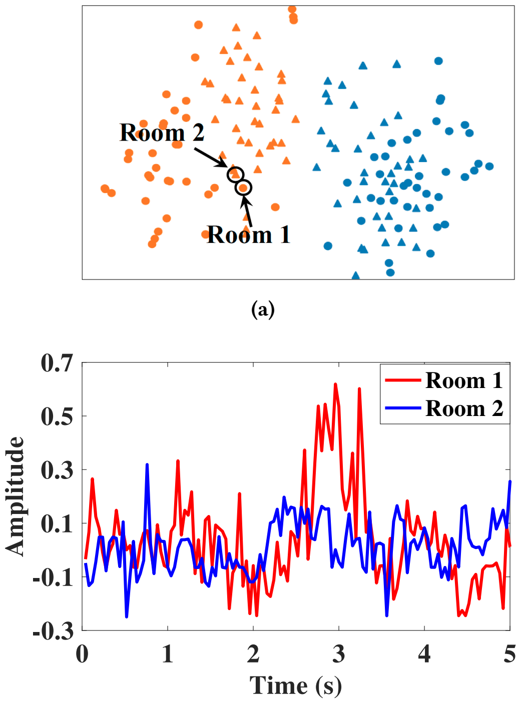
Figure 5: Accuracy of the proposed model on CSI data.

From Fig. 5, we can observe that all the approaches have low accuracy when there are only 2 source domains. This is because the labeled samples are too few to learn a good classifier for each approach. However, the approaches that utilize unlabeled data on the target domains (i.e., EI, VADA, CAT) are able to learn better classifiers than random forest which takes as input only labeled data. The WiFi signals are sensitive to the surrounding environments, and thus the signals collected in the source domains and target domains are quite different, which makes random forest unable to achieve a good performance on the target domains. For this reason, even when the number of source domains increases, the performance of random forest does not have significant improvement. On the contrary, the other three deep learning based approaches are able to extract the common features shared by both source and target domains, which enables them to utilize label information more effectively. Therefore, their performance is better than that of random forest. Among them, the proposed EI framework can achieve the best performance. By adding the balance constraint and confidence control constraint, the proposed approach can significantly increase the exploration ability, and is suitable for the task of activity recognition with WiFi signals even when the boundaries among different activities are ambiguous.

The ultimate goal of the proposed EI framework is to learn environment-independent representations of activities. To qualitatively evaluate the learned representations, we conduct the following experiment on the WiFi CSI dataset. From the unlabeled data in target domains, we first select one subject who collected data of two different activities in two different rooms, i.e., four activity and room pairs. Then we randomly select 40 data samples for each activity and room pair, and finally plot the learned representations of these samples according to Eq. (1) on a 2-D space with t-SNE [28] shown in Fig. 6a.

In Fig. 6a, we use orange and blue colors to represent different activities, and circle and triangle markers to represent different rooms. Note that the activity labels of those samples are unknown. It can be observed that the samples in the latent feature space $\mathbf{Z}$ can form two clearly separate clusters, where each cluster corresponds to an activity. Moreover, we can observe that within each activity cluster, samples from different rooms are mixed with each other. This demonstrates the effectiveness of the proposed EI framework, i.e., *learning environment-independent features.*

To further illustrate the aforementioned observation, we first pick two samples with the same activity label. As seen in Fig. 6a, they are close to each other in the latent feature space, though being collected from two different rooms. We then plot their original one-channel CSI waveforms in Fig. 6b. As one can see, their waveforms are quite different. With such different input data, the proposed EI framework can still learn similar representations. This again validates that the proposed EI framework is able to remove domain-unique features and extract environment-independent information from unlabeled data.



(a)

(b)

Figure 6: Learned representation (a) and raw signal (b).

4.3 Experiment with Ultrasound Signals

4.3.1 Experimental Settings. In this experiment, we aim to study the effect of human activities on ultrasound signals and evaluate the performance of the proposed system. To achieve the goal, we employ 12 volunteers (including both men and women) as the subjects to conduct the 6 different activities (wiping the whiteboard, walking, moving a suitcase, rotating the chair, sitting, as well as standing up and sitting down) that are shown in Fig. 4. The activity data are collected from 6 different rooms in two different buildings. Figure 7 shows the experiment setting in one of the rooms. The transmitter is an iPad on which an ultrasound generator app is installed, and it can emit an ultrasound signal of nearly 19 KHz. The receiver is a smartphone and we use the installed recorder app to collect the sound waves. The sound signal received by the receiver is a mixture of the sound waves traveling through the Line-of-Sight (LOS) and those reflected by the surrounding objects, including the human bodies in the room. We let the subjects repeat these six activities in each room for 5 rounds and in each round, the subjects are asked to take each type of activity for 51 seconds. Totally, we collect the activity data of 40 subject-room pairs (i.e., 40 domains).

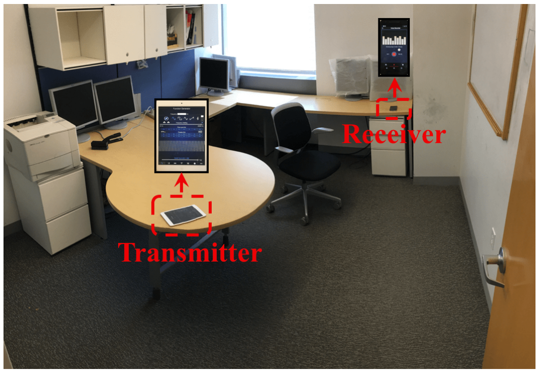


Figure 7: Experimental setting for human activity recognition with ultrasound signal.

4.3.2 Data Preprocessing. While the ultrasound signal is being transmitted, it may be reflected by the ambient objects, such as the human body. When the human subject moves, the phase of the received signal will get increased/decreased with the change of its propagation distance.

Thus, we can view the received ultrasound wave at the receiver as a phase-modulated signal whose phase changes with the movement of subject. As suggested in [51], we can extract the phase information through demodulating the received signal. Assume that the transmitted signal can be represented by $T(t) = A\cos(2\pi ft)$, then we can represent the received signal as $R(t) = A'\cos(2\pi ft - 2\pi fd/c)$, where A and A' are the amplitude of the transmitted and received signal respectively, f is the frequency, c is the speed of sound, and d is the length of the propagation which will be influenced by the movement of subject. Then d/c is the propagation delay and $2\pi fd/c$ is the phase lag caused by the propagation delay. The demodulation algorithm is to multiply the received signal with $\cos(2\pi ft)$ to extract the signal around frequency f:

$$\begin{aligned} &A'\cos(2\pi ft - 2\pi fd/c) \times \cos(2\pi ft) \\ &= \frac{A'}{2}(\cos(-2\pi fd/c) + \cos(4\pi ft - 2\pi fd/c)). \end{aligned} \tag{18}$$

After passing the output signal through a low pass filter of frequency f', we only keep the signal whose original frequency was between $[f - f', f + f']$, which represents the influence of the human movement on the ultrasound signal. Using similar method, we multiply the received signal with $-\sin(2\pi ft)$ to get $\frac{A'}{2}(\sin(-2\pi fd/c))$.

Then, we downsample signal to 345 Hz and segment the signal for every 2048 points with 512 overlapping points. Finally, we use $\frac{A'}{2}(\cos(-2\pi fd/c))$ and $\frac{A'}{2}(\sin(-2\pi fd/c))$ as well as their FFTs as the input to the deep learning model.

4.3.3 Performance Evaluation. In this experiment, we divide the rooms into two disjoint sets as source and target domains. There are 21 subject-room pairs (11 volunteers and 3 rooms) used as the source domains, and 19 pairs (10 volunteers and 3 rooms) as the target domains. Nine volunteers are involved in both source and target domains. Figure 8 shows the accuracy of all the approaches on the ultrasound dataset with different number of source domains.

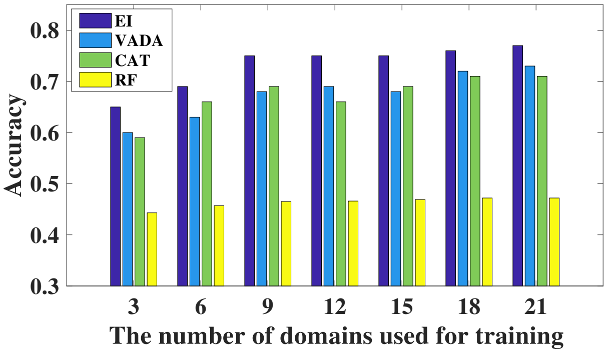


Figure 8: Accuracy of the proposed model on ultrasound data.

From Fig. 8, we can observe that the proposed EI framework can achieve better performance compared with the baselines in all cases. We also notice that the performance of random forest is the worst. For random forest, though we use Mel-frequency Cepstral Coefficients (MFCCs) [2], a feature commonly used for audible sound based recognition tasks, as its input data, its accuracy is still not as good as that of the deep learning models. Moreover, it can be observed that as the number of source domains increases, all the methods have a general trend of increasing-and-stabilizing. This means that with a few labeled data, all the approaches are able to learn good classifier boundaries on the ultrasound dataset.

In Fig. 9a, we first show the learned representations of acoustic signals that correspond to a single subject performing two different activities in two different rooms. From Fig. 9a, we can observe similar patterns as those in the experiment with WiFi signals, but the boundary between these two activities is more clear. Figure 9b lists two acoustic signals of the same activity collected from different rooms on the same volunteer. Though they are different, the representations of them are quite close in the learned latent space.

4.4 Experiment with 60 GHz mmWave

In recent years, the 60 GHz millimeter-wave (mmWave) technology has been introduced to further increase the throughput of wireless networks. In addition to improving the communication performance, 60 GHz millimeter-wave signals

[2] https://en.wikipedia.org/wiki/Mel-frequency_cepstrum

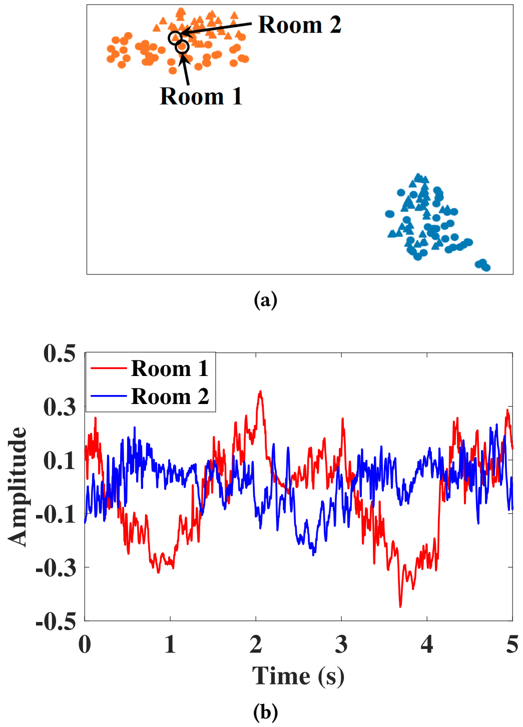


Figure 9: Learned representation (a) and raw signal (b).

can also be leveraged for sensing tasks such as human activity recognition. In this section, we study the effect of human activities on the mmWave signals.

4.4.1 Experimental Settings. In this experiment, 10 volunteers (including both men and women) are employed as the subjects and the activity data are collected from 4 different rooms in two different buildings. Figure 10 shows the experiment setting in one of the rooms. The platform we use to collect the mmWave activity data is X60 [35]. Each X60 node is based on National Instruments' mmWave Transceiver System [18] and equipped with a user-configurable 24-element (12 for TX and 12 for RX) phased antenna array from SiBeam. Previous gesture tracking systems [53] used receivers equipped with narrow-beam horn antennas (e.g., 3.4 degrees in [53]), essentially eliminating multipath, which enabled them to perform the passive tracking using physics laws. In contrast, commercial mmWave systems using phased array antennas generate imperfect beams with wide main lobes and often strong side lobes due to the discretization of the individual antenna element phase shift and the relatively small number of antenna elements. For example, the main lobe in the beams generated by our hardware is 30-35 degrees. In Fig. 11, we illustrate the pattern of the beam we used (beam 12) in polar coordinates. Such imperfect beams often result in non-negligible multipath propagation (although still weaker than in WiFi) [29, 34, 35]. Thus, using only the physics laws it is very difficult to precisely model the complex ambient environments as well as the unique characteristics of different human subjects. Deep learning technique is an ideal solution for this problem due to its superior feature extraction ability.

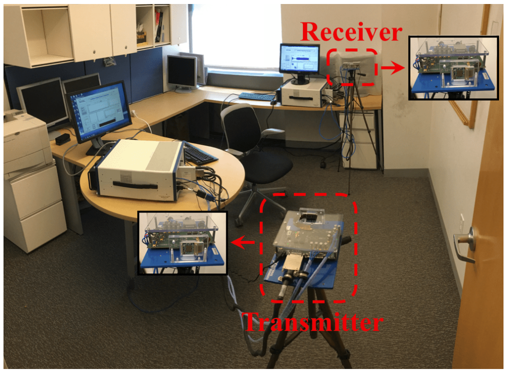


Figure 10: Experimental setting for human activity recognition with mmWave signals.

In our experimental setting, we ask the subjects to conduct 5 types of activities (walking, moving a suitcase, rotating the chair, sitting, as well as standing up and sitting down) that are shown in Fig. 4. The subjects are also asked to repeat these five activities in each room for 4 rounds and in each round, we collect 10 segments of mmWave signal (5 seconds for each segment) for each activity of one subject. Totally, we collect the activity data of 19 different domains.

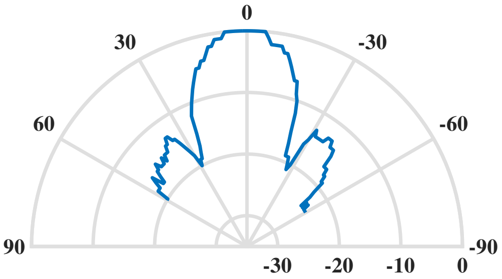


Figure 11: The pattern of the 12th beam of the mmWave signal in polar coordinates.

4.4.2 Data Preprocessing. With the accompanying software API on this platform, we are able to obtain a channel impulse response (CIR) sample (each has 1024 points) every 40 ms. For each data segment, we collect samples for 5 seconds, hence there are 125 CIR samples in each segment. Also, in order to characterize the frequency response of the wireless channel, we transform each CIR sample to a frequency response sample through simply calculating the Fourier transform of each CIR sample. After that, we downsample each frequency response sample to 32 points to compose a 32 × 125 feature matrix as the input to our model.

4.4.3 Performance Evaluation. In the experiment on the mmWave dataset, there are 11 source domains (9 volunteers in 2 rooms) and 8 target domains (6 volunteers in 2 rooms), and 5 volunteers are involved in both source and target domains. Figure 12 shows the accuracy of all the approaches on the mmWave dataset. We can observe that the proposed EI performs better than all the baselines, but the improvement is not significant compared with the results on both WiFi and ultrasound datasets. This phenomenon is caused by the unique properties of the collected mmWave data. The 60 GHz mmWave is usually made directional [3], and this directionality makes the collected data not as sensitive to the surrounding environments as WiFi and acoustic signals.

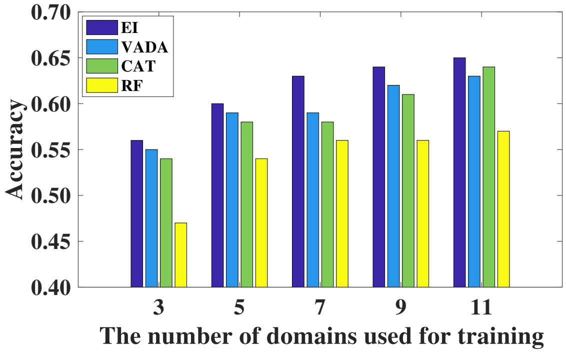


Figure 12: Accuracy of the proposed model on mmWave data.

We also conduct experiments to analyze the representations learned by the proposed EI framework, which is shown in Fig. 13. As seen, though the environment-specific information contained in the mmWave signals is not as much as in WiFi and acoustic signals, the proposed EI framework can still remove it and improve the prediction performance.

4.5 Experiment with Visible Light

4.5.1 Experimental Settings. To evaluate the performance of the proposed system in the visible light environments, we build an optical system using photoresistors to capture the in-air body gesture. Given the light source, the system is able to precisely detect the illuminance change (lux) caused by the body interaction. Specifically, we employ the cadmium-sulfide (CdS) cells, which are basically resistors that change their resistive value in ohms depending on the amount of light which is shining onto the squiggly face. To measure the resistor, we employ Arduino Uno and connect one end of the cell to the power (5V) and the other to a pull-down resistor to ground. With each board equipped with 6 analog input pins (A0 A5), we developed 3 boards with 18 CdS cells in total (as shown in Fig. 14). The resistor value of each cell is monitored and recorded through the serial port at the sampling rate of 15 Hz. To simultaneously record the data from three boards, we implemented the reading program using processing sdk so that the logged system clock on each board is synchronized. For the ambient light source, we chose Qooltek Portable USB lamp because it provides three lighting options: natural mode, warm mode and cool mode, which covers most of the lighting conditions in daily life.

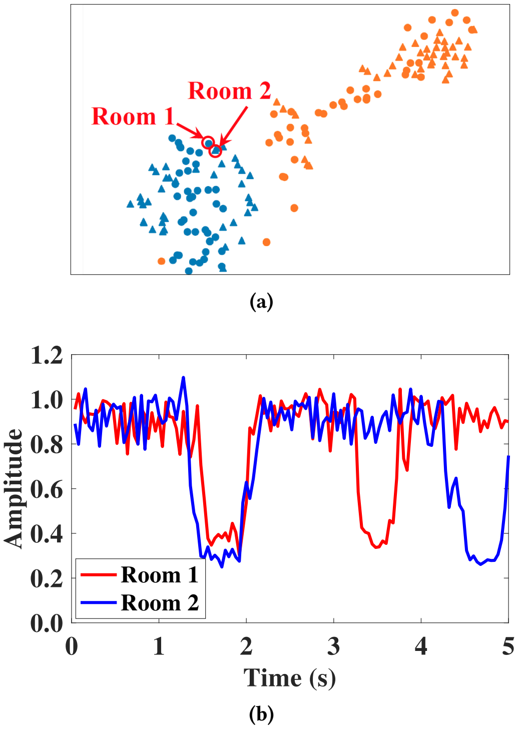


Figure 13: Learned representation (a) and raw signal (b).

In this experiment, we treat the above three lighting options (i.e., natural mode, warm mode, and cool mode) as three different environments, and then design four hand gestures (i.e., drawing an anticlockwise circle, drawing a clockwise circle, drawing a cross, and shaking hand side to side). Specifically, we employ 6 volunteers (including both men and women) as the subjects and each of them performs 20 trials of every gesture under a given lighting condition. In total, we collect the activity data of 18 different domains.

[3] 60 GHz mmWave is highly attenuated due to its high frequency. To mitigate its high attenuation characteristics, directionality is usually employed.

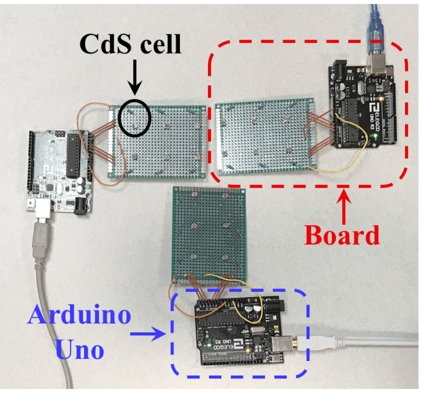


Figure 14: The optical system used for collecting visible light signals.

4.5.2 Data Preprocessing. Due to the unavoidable small variation in the sampling length between trials, we need to segment the data into the uniformed length. Since all 18 photoresistors are synchronized in one trial, we randomly pick one as the pivot (e.g., the first one) and segment all the data based on the same timestamp. The hand gesture generates the peak (or valley) when it covers (or leaves) the surface of the photoresistor. To detect peak values, we adopt a peak detection algorithm with adaptive threshold [63].

Given the recorded signal $x(n)$, this algorithm obtains the relationship between the counted peaks and the threshold value. Specifically, it goes through all the threshold values from 0 to the maximal magnitude difference, and the corresponding number of peaks (or valleys) is detected. Then, we search for the stage where the number of the peaks stays unchanged when the thresholds increase, which implies that most of the random noise is ignored and only the true peaks are counted. In this way, we can accurately find the gesture-relevant peaks in $x(n)$. Based on the empirical knowledge, we select the entire gesture window as 2100 ms to make sure it covers all the peaks and segment the data from all photoresistors according to the timestamp. Eventually, each gesture is represented by a data sequence of 480 samples.

4.5.3 Performance Evaluation. Different from the previous three experiments, the environment in this case is the lighting option, not the room. In practice, the collected visible light data are not sensitive to the lighting options, but the quality of the data mainly depends on the gestures of subjects. Therefore, the domain-specific information in this experiment comes more from the uniqueness of subjects than environments.

In this experiment, the lighting options are fixed (three options), and we have 6 source domains (2 volunteers) and 12 target domains (4 volunteers). Note that there is no common volunteer in both source and target domains.

Figure 15 shows the experimental results on the Visible Light dataset. We can observe that the proposed EI framework still outperforms all the baselines in terms of accuracy. Since there is no common subjects between source and target domains, for each approach, higher accuracy means better ability of learning transferable feature representations. Random forest cannot extract such features, and thus performs the worst. Figure 16 presents the case study on the learned representations and raw visible light signals. Here we select two different subjects who collect data of two different activities in one environment, i.e., four activity and subject pairs. Figure 16a shows the learned representations. We use orange and blue colors to represent different activities, and circle and triangle markers to represent different subjects. Figure 16b lists two light signals of the same activity collected by different subjects in the same environment. Both Fig. 15 and 16 show that the proposed EI framework has the ability of removing unique characteristics of different subjects and is effective for the device free human activity recognition task.

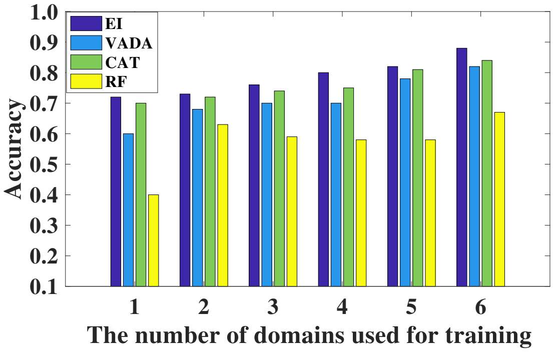
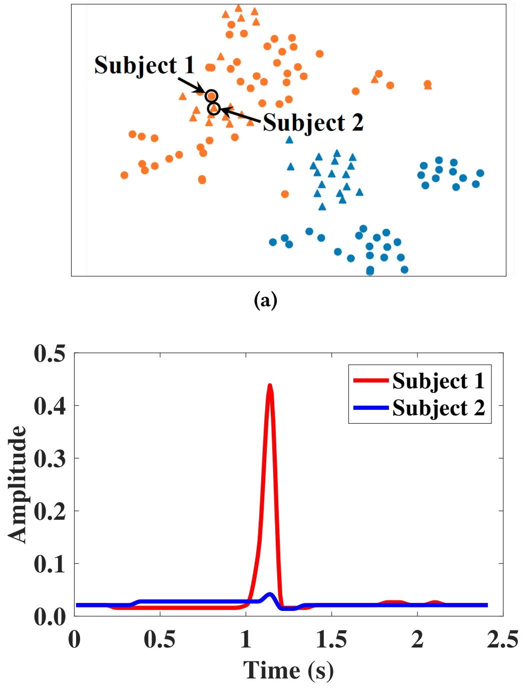


Figure 15: Accuracy of the proposed model on visible light data.

4.6 The Effect of the Balance Constraint

As described in Section 3.5, in the proposed EI framework, we add a balance constraint to control the percentage of the data labeled as each activity by the model. The percentage of each activity, in our design, can be estimated according to either prior knowledge or labeled data in source domains. In practice, however, the real percentage of each activity in unlabeled data from target domains may not exactly equal to the estimated percentage. In this section, we evaluate how sensitive our model is to the percentage of each activity.

Here we take the CSI dataset as an example. We set the number of source domains to be 22 and the number of target domains to be 18. Then, for each target domain, we randomly select some activities and discard a proportion of the data of these activities to make the percentage of activities in the



(a)

(b)

Figure 16: Learned representation (a) and raw signal (b).

Table 1: Accuracy of EI framework when the activity percentage in the target domains does not match that in the source domains.

Number of activities	1	2	3	4	5
Discard ratio = 0.25	0.73	0.73	0.72	0.72	0.72
Discard ratio = 0.50	0.72	0.71	0.71	0.71	0.69

target domains different from that in the source domains. In this experiment, we consider two cases where the ratios of the discarded data are set as 0.25 and 0.50, respectively. For each case, we vary the number of the selected activities from 1 to 5. Table 1 reports the accuracy of the EI framework for the two cases. From this table, we can observe that compared with the ideal scenario when the percentage of each activity in the target domains equals to that in the source domains (the accuracy is 0.75 according to Fig. 5), the performance of the EI framework drops slightly. Additionally, the results in Table 1 also show that the accuracy of the EI framework decreases slightly when the ratio of the discarded data increases from 0.25 to 0.50. The results of this experiment verify that the proposed EI framework can still achieve good performance even when the percentage of each activity in unlabeled data does not match that in labeled data.

To further evaluate the effect of the balance constraint, we also implement the EI framework on the CSI dataset without taking balance constraint into account, and then compare it with the EI framework with balance constraint as well as the baseline methods. The adopted experimental setting here is the same as that in Section 4.2.4. The comparison results are shown in Fig. 17, from which we can see that even when we remove the balance constraint, the EI framework can still achieve better performance than the baselines. However, the performance of the EI framework without balance constraint is not as good as that when the balance constraint is enabled. For example, when the number of domains used for training is 22, the accuracy of the EI framework with balance constraint is 0.75 while that of the EI framework without balance constraint is only 0.61, which is also much lower than the accuracy of any of the unbalanced settings shown in Table 1.

In summary, the above experimental results show that the designed balance constraint plays an important role in the human activity recognition tasks, even if the percentage of the activities in the target domains does not exactly match that in the source domains.

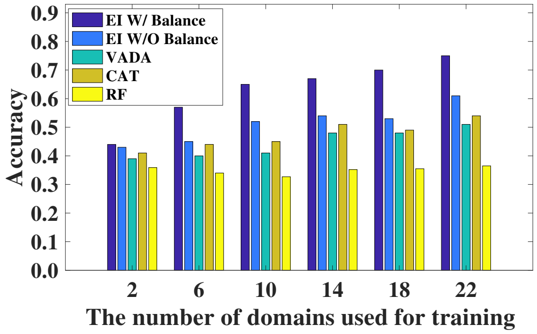


Figure 17: Accuracy of the proposed model without balance constraint.

5 RELATED WORK

Device-free Human Activity Recognition: Human activity recognition (HAR) has been widely studied in recent years. However, traditional methods such as vision based [5, 30, 56] and wearable device based [9, 19–21] methods either have privacy and complexity problems or require subjects to wear special devices. To address these challenges, researchers start to leverage wireless signals (e.g., ultrasound, WiFi, mmWave, visible light, etc.) to achieve device-free human activity recognition. Based on the type of adopted wireless signal as well as the feature extracted from the signal, those methods can be clustered into the following categories:

- **Acoustic-based methods:** Acoustic signals emitted and recorded by Commercial-Off-The-Shelf (COTS) mobile devices can achieve frequency higher than 17 KHz, which is inaudible to most people [32]. When the acoustic signals reflect off moving objects, such as human body, they get frequency shift due to the Doppler effect. In some recent work [6, 15, 31, 33], the authors propose to recognize human gestures and activities through analyzing frequency shift over a period of time.

- **RSSI-based methods:** As an indicator of the power level of the signal received at the receiver, the received signal strength indicator (RSSI) can be used to measure the distance as well as the channel conditions between the transmitter and receiver. Some researchers [1, 36, 40, 49] propose to recognize human activities through analyzing the RSSI values. For example, by analyzing the changes in WiFi signal strength, it is possible to recognize in-air hand gestures around the user's mobile device [1].
- **CSI-based methods:** As a known channel property of a communication link, CSI can reflect the combined effects of scattering, fading and even the power decay with distance. Thus, compared with RSSI, CSI can capture the fine-grained changes of wireless channels. Because of the release of Linux 802.11n CSI Tool [16], recently a lot of research work have been conducted to utilize CSI for the task of human activity recognition [3, 8, 14, 38, 47, 48, 50, 52, 55] or gesture recognition [17, 24, 37, 42, 45, 46, 59].
- **mmWave-based methods:** Compared with WiFi, which uses 2.4/5 GHz frequency bands, 60 GHz mmWave has a much shorter carrier wavelength. The shorter wavelength of 60 GHz mmWave can create stronger reflection from small objects since wireless signals cannot easily bypass objects larger than wavelength [53]. Moreover, 60 GHz mmWave is usually made directional and the signal strength of 60 GHz mmWave is highly correlated with the object material [23]. Therefore, researchers have begun to use this technology to recognize/tracking different gestures [27, 53], monitor vital signs [57], and image the objects [61, 62]. To the best of our knowledge, our work is the first that uses 60 GHz mmWave to recognize whole-body activities.
- **Light-based methods:** Since each human activity can produce unique continuous shadow map under visible light, some recent work [4, 25, 26] propose to recognize human activities or gestures by analyzing those shadow maps.

The above device free activity recognition approaches and systems, though having good performance in certain cases, are all challenged by the environment/subject-specific information contained in the wireless signals.

Domain Adversarial Training: Technically, our work is related to domain adversarial training approaches [2, 10, 11, 39, 43, 44, 60]. Domain adversarial training shares with the generative adversarial network [12] the use of adversarial objective, and its goal is to encourage a neural network to learn a representation that is predictive to learning task on the source domain, but uninformative to the domain of the input. [2, 10, 11] are the first domain adversarial training approaches that are proposed to tackle the unsupervised domain adaptation problem. To further improve the domain adaptation performance, Zhao *et al.* [60] propose a conditional adversarial architecture, which can retain the information relevant to the predictive task when removing the domain-specific information. Although this architecture is effective, it is mainly designed for supervised tasks without taking the unlabeled data into account. To take advantage of the unlabeled data, the authors of [39] propose to force the classifier to be confident on the unlabeled data to improve the adversarial training. Different from previous work, our proposed model incorporates the unlabeled data into conditional adversarial architecture. Moreover, we find out that merely increasing the confidence on the unlabeled data may lead to premature convergence and even extreme cases where most samples are incorrectly assigned to the same activity category. In order to tackle these problems, we further add a confidence control constraint and make use of the prior knowledge, i.e., the percentage of activities on the labeled data, to design a balance regularization.

6 CONCLUSIONS

In this paper, we propose an effective and general framework to recognize device free human activities. Especially, the proposed framework can remove environment and subject specific information and learn transferable features of activities. The proposed framework is composed of a feature extractor, an activity recognizer, a domain discriminator, and several constraints. The feature extractor tries to its best to cheat the domain discriminator by minimizing domain label accuracy, and at the same time, maximizes the performance of the activity recognizer. Through this minimax game, the proposed framework can finally derive environment-independent features. Extensive experiments on four different testbeds, including WiFi, ultrasound, 60 GHZ mmWave and visible light, demonstrate the effectiveness of the proposed framework.

ACKNOWLEDGMENTS

We thank our anonymous shepherd and reviewers for their insightful comments and suggestions on this paper. This work was supported in part by the US National Science Foundation under grants CNS-1652503 and CNS-1629929.

REFERENCES

[1] Heba Abdelnasser, Moustafa Youssef, and Khaled A Harras. 2015. Wigest: A ubiquitous wifi-based gesture recognition system. In *Computer Communications (INFOCOM), 2015 IEEE Conference on*. IEEE, 1472–1480.

[2] Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. 2014. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446* (2014).

[3] Kamran Ali, Alex X Liu, Wei Wang, and Muhammad Shahzad. 2015. Keystroke recognition using wifi signals. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. ACM, 90–102.

[4] Chuankai An, Tianxing Li, Zhao Tian, Andrew T Campbell, and Xia Zhou. 2015. Visible light knows who you are. In *Proceedings of the 2nd International Workshop on Visible Light Communications Systems*. ACM, 39–44.

[5] Robert Bodor, Bennett Jackson, and Nikolaos Papanikolopoulos. 2003. Vision-based human tracking and activity recognition. In *Proc. of the 11th Mediterranean Conf. on Control and Automation*, Vol. 1.

[6] Ke-Yu Chen, Daniel Ashbrook, Mayank Goel, Sung-Hyuck Lee, and Shwetak Patel. 2014. AirLink: sharing files between multiple devices using in-air gestures. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 565–569.

[7] Laurie Davies and Ursula Gather. 1993. The identification of multiple outliers. *J. Amer. Statist. Assoc.* 88, 423 (1993), 782–792.

[8] Shihong Duan, Tianqing Yu, and Jie He. 2018. WiDriver: Driver Activity Recognition System Based on WiFi CSI. *International Journal of Wireless Information Networks* (2018), 1–11.

[9] Matthew Field, David Stirling, Zengxi Pan, Montserrat Ros, and Fazel Naghdy. 2015. Recognizing human motions through mixture modeling of inertial data. *Pattern Recognition* 48, 8 (2015), 2394–2406.

[10] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*. 1180–1189.

[11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* 17, 1 (2016), 2096–2030.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[13] Yves Grandvalet and Yoshua Bengio. 2005. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*. 529–536.

[14] Xiaonan Guo, Bo Liu, Cong Shi, Hongbo Liu, Yingying Chen, and Mooi Choo Chuah. 2017. WiFi-Enabled Smart Human Dynamics Monitoring. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*. ACM, 16.

[15] Sidhant Gupta, Daniel Morris, Shwetak Patel, and Desney Tan. 2012. Soundwave: using the doppler effect to sense gestures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1911–1914.

[16] Daniel Halperin, Wenjun Hu, Anmol Sheth, and David Wetherall. 2011. Tool release: Gathering 802.11 n traces with channel state information. *ACM SIGCOMM Computer Communication Review* 41, 1 (2011), 53–53.

[17] Wenfeng He, Kaishun Wu, Yongpan Zou, and Zhong Ming. 2015. Wig: Wifi-based gesture recognition system. In *Computer Communication and Networks (ICCCN), 2015 24th International Conference on*. IEEE, 1–7.

[18] National Instruments. 2017. Introduction to the NI mmWave Transceiver System Hardware - National Instruments. http://www.ni.com/white-paper/53095/en/. Accessed on 06/25/2017.

[19] Wenjun Jiang, Qi Li, Lu Su, Chenglin Miao, Quanquan Gu, and Wenyao Xu. 2018. Towards Personalized Learning in Mobile Sensing Systems. In *Distributed Computing Systems (ICDCS), 2018 IEEE 38th International Conference on*. IEEE.

[20] Matthew Keally, Gang Zhou, Guoliang Xing, Jianxin Wu, and Andrew Pyles. 2011. Pbn: towards practical activity recognition using smartphone-based body sensor networks. In *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*. ACM, 246–259.

[21] Adil Mehmood Khan, Ali Tufail, Asad Masood Khattak, and Teemu H Laine. 2014. Activity recognition on smartphones via sensor-fusion and kda-based svms. *International Journal of Distributed Sensor Networks* 10, 5 (2014), 503291.

[22] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[23] B. Langen, G. Lober, and W. Herzig. 1994. Reflection and transmission behaviour of building materials at 60 GHz. In *Personal, Indoor and Mobile Radio Communications, 1994. Wireless Networks-Catching the Mobile Future., 5th IEEE International Symposium on*, Vol. 2. IEEE, 505–509.

[24] Hong Li, Wei Yang, Jianxin Wang, Yang Xu, and Liusheng Huang. 2016. WiFinger: talk to your smart devices with finger-grained gesture. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 250–261.

[25] Tianxing Li, Qiang Liu, and Xia Zhou. 2016. Practical human sensing in the light. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 71–84.

[26] Tianxing Li, Xi Xiong, Yifei Xie, George Hito, Xing-Dong Yang, and Xia Zhou. 2017. Reconstructing hand poses using visible light. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 71.

[27] Jaime Lien, Nicholas Gillian, M Emre Karagozler, Patrick Amihood, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev. 2016. Soli: Ubiquitous gesture sensing with millimeter wave radar. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 142.

[28] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, Nov (2008), 2579–2605.

[29] Thomas Nitsche, Guillermo Bielsa, Irene Tejado, Adrian Loch, and Joerg Widmer. 2015. Boon and bane of 60 GHz networks: practical insights into beamforming, interference, and frame level operation. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies*. ACM, 17.

[30] Ronald Poppe. 2010. A survey on vision-based human action recognition. *Image and vision computing* 28, 6 (2010), 976–990.

[31] Yang Qifan, Tang Hao, Zhao Xuebing, Li Yin, and Zhang Sanfeng. 2014. Dolphin: Ultrasonic-based gesture recognition on smartphone platform. In *Computational Science and Engineering (CSE), 2014 IEEE 17th International Conference on*. IEEE, 1461–1468.

[32] A Rodríguez Valiente, A Trinidad, JR García Berrocal, C Górriz, and R Ramírez Camacho. 2014. Extended high-frequency (9–20 kHz) audiometry reference thresholds in 645 healthy subjects. *International journal of audiology* 53, 8 (2014), 531–545.

[33] Wenjie Ruan, Quan Z Sheng, Lei Yang, Tao Gu, Peipei Xu, and Longfei Shangguan. 2016. AudioGest: enabling fine-grained hand gesture detection by decoding echo signal. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 474–485.

[34] Swetank Kumar Saha, Hany Assasa, Adrian Loch, Naveen Muralidhar Prakash, Roshan Shyamsunder, Shivang Aggarwal, Daniel Steinmetzer, Dimitrios Koutsonikolas, Joerg Widmer, and Matthias Hollick. 2018. Fast and infuriating: Performance and pitfalls of 60 ghz wlans based on consumer-grade hardware. In *2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE.

[35] Swetank Kumar Saha, Yasaman Ghasempour, Muhammad Kumail Haider, Tariq Siddiqui, Paulo De Melo, Neerad Somanchi, Luke Zakrajsek, Arjun Singh, Owen Torres, Daniel Uvaydov, Josep Miquel Jornet, Edward Knightly, Dimitrios Koutsonikolas, Dimitris Pados, and Zhi Sun. 2017. X60: A programmable testbed for wideband 60 ghz wlans with phased arrays. In *Proceedings of the 11th Workshop on Wireless Network Testbeds, Experimental evaluation & CHaracterization*. ACM, 75–82.

[36] Markus Scholz, Till Riedel, Mario Hock, and Michael Beigl. 2013. Device-free and device-bound activity recognition using radio signal strength. In *Proceedings of the 4th Augmented Human International Conference*. ACM, 100–107.

[37] Jiacheng Shang and Jie Wu. 2017. A robust sign language recognition system with sparsely labeled instances using Wi-Fi signals. In *Mobile Ad Hoc and Sensor Systems (MASS), 2017 IEEE 14th International Conference on*. IEEE, 99–107.

[38] Cong Shi, Jian Liu, Hongbo Liu, and Yingying Chen. 2017. Smart user authentication through actuation of daily activities leveraging WiFi-enabled IoT. In *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing*. ACM, 5.

[39] Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. 2018. A DIRT-T Approach to Unsupervised Domain Adaptation. In *International Conference on Learning Representations*.

[40] Stephan Sigg, Shuyu Shi, Felix Buesching, Yusheng Ji, and Lars Wolf. 2013. Leveraging RF-channel fluctuation for activity recognition: Active and passive systems, continuous and RSSI-based signal features. In *Proceedings of International Conference on Advances in Mobile Computing & Multimedia*. ACM, 43.

[41] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.

[42] Sheng Tan and Jie Yang. 2016. WiFinger: leveraging commodity WiFi for fine-grained finger gesture recognition. In *Proceedings of the 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing*. ACM, 201–210.

[43] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. 2015. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*. 4068–4076.

[44] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, Vol. 1. 4.

[45] Raghav H Venkatnarayan, Griffin Page, and Muhammad Shahzad. 2018. Multi-User Gesture Recognition Using WiFi. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 401–413.

[46] Aditya Virmani and Muhammad Shahzad. 2017. Position and orientation agnostic gesture recognition using wifi. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 252–264.

[47] Guanhua Wang, Yongpan Zou, Zimu Zhou, Kaishun Wu, and Lionel M Ni. 2016. We can hear you with wi-fi! *IEEE Transactions on Mobile Computing* 15, 11 (2016), 2907–2920.

[48] Jie Wang, Liming Zhang, Qinghua Gao, Miao Pan, and Hongyu Wang. 2018. Device-Free Wireless Sensing in Complex Scenarios Using Spatial Structural Information. *IEEE Transactions on Wireless Communications* 17, 4 (2018), 2432–2442.

[49] Jie Wang, Xiao Zhang, Qinhua Gao, Hao Yue, and Hongyu Wang. 2017. Device-free wireless localization and activity recognition: A deep learning approach. *IEEE Transactions on Vehicular Technology* 66, 7 (2017), 6258–6267.

[50] Wei Wang, Alex X Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. 2015. Understanding and modeling of wifi signal based human activity recognition. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. ACM, 65–76.

[51] Wei Wang, Alex X Liu, and Ke Sun. 2016. Device-free gesture tracking using acoustic signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. ACM, 82–94.

[52] Yan Wang, Jian Liu, Yingying Chen, Marco Gruteser, Jie Yang, and Hongbo Liu. 2014. E-eyes: device-free location-oriented activity identification using fine-grained wifi signatures. In *Proceedings of the 20th annual international conference on Mobile computing and networking*. ACM, 617–628.

[53] Teng Wei and Xinyu Zhang. 2015. mtrack: High-precision passive tracking using millimeter wave radios. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. ACM, 117–129.

[54] Jason Weston, Frédéric Ratle, Hossein Mobahi, and Ronan Collobert. 2012. Deep learning via semi-supervised embedding. In *Neural Networks: Tricks of the Trade*. Springer, 639–655.

[55] Dan Wu, Daqing Zhang, Chenren Xu, Yasha Wang, and Hao Wang. 2016. WiDir: walking direction estimation using wireless signals. In *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing*. ACM, 351–362.

[56] Lu Xia, Chia-Chih Chen, and JK Aggarwal. 2012. View invariant human action recognition using histograms of 3d joints. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE, 20–27.

[57] Zhicheng Yang, Parth H Pathak, Yunze Zeng, Xixi Liran, and Prasant Mohapatra. 2016. Monitoring vital signs using millimeter wave. In *Proceedings of the 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing*. ACM, 211–220.

[58] Ming Zeng, Le T Nguyen, Bo Yu, Ole J Mengshoel, Jiang Zhu, Pang Wu, and Joy Zhang. 2014. Convolutional neural networks for human activity recognition using mobile sensors. In *Mobile Computing, Applications and Services (MobiCASE), 2014 6th International Conference on*. IEEE, 197–205.

[59] Ouyang Zhang and Kannan Srinivasan. 2016. Mudra: User-friendly Fine-grained Gesture Recognition using WiFi Signals. In *Proceedings of the 12th International on Conference on emerging Networking EXperiments and Technologies*. ACM, 83–96.

[60] Mingmin Zhao, Shichao Yue, Dina Katabi, Tommi S Jaakkola, and Matt T Bianchi. 2017. Learning sleep stages from radio signals: A conditional adversarial architecture. In *International Conference on Machine Learning*. 4100–4109.

[61] Yibo Zhu, Yanzi Zhu, Zengbin Zhang, Ben Y Zhao, and Haitao Zheng. 2015. 60GHz mobile imaging radar. In *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*. ACM, 75–80.

[62] Yanzi Zhu, Yibo Zhu, Ben Y Zhao, and Haitao Zheng. 2015. Reusing 60ghz radios for mobile radar imaging. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. ACM, 103–116.

[63] Yan Zhuang, Chen Song, Aosen Wang, Feng Lin, Yiran Li, Changzhan Gu, Changzhi Li, and Wenyao Xu. 2015. SleepSense: Non-invasive sleep event recognition using an electromagnetic probe. In *Wearable and Implantable Body Sensor Networks (BSN), 2015 IEEE 12th International Conference on*. IEEE, 1–6.