Towards Cross Domain CSI Action Recognition Through One-Shot Bimodal Domain Adaptation

Bao Zhou, Rui Zhou(✉), Yue Luo, and Yu Cheng

University of Electronic Science and Technology of China, Chengdu, China
ruizhou@uestc.edu.cn

Abstract. Human action recognition based on WiFi Channel State Information (CSI) has attracted enormous attention in recent years. Although performing well under supervised learning, the recognition model suffers from significant performance degradation when applied in a new domain (e.g. a new environment, a different location, or a new user). To enable the recognition model robust to domains, researchers have proposed various methods, including semi-supervised domain adaptation, unsupervised domain adaptation, and domain generalization. Semi-supervised and unsupervised solutions still require a large number of partially-labeled or unlabeled samples in the new domain, while domain generalization solutions have difficulties in achieving acceptable accuracy. To mitigate these problems, we propose a one-shot bimodal domain adaptation method to achieve cross domain action recognition with much reduced effort. The method contains two key points. One is that it synthesizes virtual samples to augment the training datatset of the target domain, requiring only one sample per action in the target domain. The other is that it regards the amplitude and the phase as two consistent modals and fuses them to enhance the recognition accuracy. Virtual data synthesis is achieved by linear transformation with dynamic domain weights and the synthesis autoencoder. Bimodal fusion is achieved by the fusion autoencoder and feature concatenation under the criterion of consistency. Evaluations on daily activities achieved the average accuracy of 85.03% and 90.53% at target locations, 87.90% and 82.40% in target rooms. Evaluations on hand gestures achieved the average accuracy of 91.67% and 85.53% on target users, 83.04% and 88.01% in target rooms.

Keywords: Action recognition · Modal fusion · Data synthesis · Domain adaptation

1 Introduction

Action recognition is of great values in many aspects of our daily lives. Existing action recognition approaches are mainly based on visions and sensors. Vision-based solutions can achieve excellent recognition accuracy, but require Line of

S. Longfei and P. Bodhi (Eds.): MobiQuitous 2022, LNICST 492, pp. 290–309, 2023.
https://doi.org/10.1007/978-3-031-34776-4_16

Sight (LOS) and may incur privacy violations. Sensor-based solutions demand the users to carry dedicated devices, thus are inconvenient. As a non-intrusive and pervasive solution, WiFi-based action recognition has attracted increasing attention in recent years, due to the release of Channel State Information (CSI) tools [7,26]. A large number of studies on CSI-based action recognition have been launched since then [1,10], ranging from activity recognition, gesture recognition, to other miscellaneous applications. Most of these solutions exploit supervised learning to achieve high accuracy, requiring the training data and the testing data to follow the same distribution. This indicates that CSI-based action recognition under supervised learning only works well in unchanged domains. If the domain changes (e.g. user diversity, location variation, room change), the recognition accuracy will decline dramatically. To keep high performance in a new domain, a large number of new samples need to be collected in the new domain to retrain or fine-tune the original recognition model. This is impractical for real-world applications. To solve this problem, researchers have proposed various solutions, which may be semi-supervised domain adaptation, unsupervised domain adaptation, or domain generalization. Semi-supervised and unsupervised solutions still need to collect a large number of partially-labeled or unlabeled samples in the new domain, while domain generalization solutions have difficulties in achieving acceptable accuracy because it is difficult to capture the characteristics of the new domain without any real samples.

To minimize the effort of domain adaptation and meanwhile keep high accuracy, we propose a one-shot bimodal domain adaptation method, aiming to achieve cross domain CSI action recognition. Utilizing the source domain data and the only sample per action in the target domain, the method synthesizes a large number of virtual samples for the target domain. Virtual data synthesis is through linear transformation with dynamic domain weights and the virtual samples are made close to the target domain by the synthesis autoencoder. To further enhance the accuracy, the method fuses the amplitude and the phase under the criterion of consistency by the fusion autoencoder. The fused features of all the virtual samples are used to train the action classifier for the target domain. Evaluations proved that the proposed method could achieve one-shot domain adaptation for action recognition across users and locations as well as rooms. The contributions of the paper can be summarized as follows.

- Propose a data synthesis method based on linear transformation and the synthesis autoencoder. With one sample per action in the target domain and the samples in the source domains, the method can synthesize a large number of virtual samples for the target domain. The domain weights in the linear transformation are determined dynamically. The virtual samples are made close to the target domain by the synthesis autoencoder.
- Propose a bimodal fusion method based on Cosine Similarity and the fusion autoencoder, which fuses the amplitude and the phase under the criterion of consistency.
- With virtual data synthesis and bimodal fusion, the method achieves one-shot domain adaptation and hence action recognition across users and locations, with the accuracy of more than 85% in the target domains.

– For the task of cross room action recognition, the method achieves the accuracy of more than 82% in the target rooms with one-shot domain adaptation.

The rest of the paper is organized as follows. Section 2 reviews the related works. Section 3 provides the overview of the method. Section 4 elaborates on the methodology of one-shot bimodal domain adaptation, focusing on virtual data synthesis and bimodal fusion. Section 5 reports the evaluations on cross domain activity recognition and gesture recognition. Section 6 concludes the paper.

2 Related Works

Since the release of WiFi CSI tools [7,26], studies on CSI-based action recognition have been boosting [1,10]. These methods can be model-based or learning-based.

2.1 Model-Based Action Recognition

Model-based methods investigate wireless transmission theories and exploit signal processing techniques to achieve recognition, hence are agnostic to domains. But accurate modeling of WiFi propagation indoors is difficult. These methods usually require a relatively large number of transceivers and have limitations on their placement. Zhang et al. [29] correlated signal propagation with motions in the first Fresnel zone and linked amplitude variations with motions to detect activities. WiDrive [3] recognized in-car activities based on Doppler Frequency Shifts (DFS), employing Hidden Markov Model with Gaussian Mixture emissions Model (HMM-GMM) as the classifier, whose parameters were updated online to adapt to vehicles and drivers. Widar3.0 [32] derived the domain independent feature Body-coordinate Velocity Profile (BVP) from DFS and recognized gestures adaptive to environments, users, locations and orientations. AirDraw [9] achieved learning-free in-air handwriting by gesture tracking using CSI phase. It denoised raw CSI by the ratio between two adjacent antennas, separated reflected signals from noise by Principal Component Analysis (PCA), and corrected tracking by eliminating static components unrelated to hand motions. Niu et al. [16] analyzed DFS to quantify the relationship between signal frequencies and target locations, motion directions and speeds. They proposed movement fragments and relative motion direction changes as two features to recognize gestures across environments, users, locations and orientations.

2.2 Learning-Based Cross Domain Action Recognition

Learning-based methods have less limitations on the number and the placement of transceivers. They try to learn the relationship between CSI measurements and actions. But learning-based methods have dependance on the domains. To achieve cross domain recognition, researchers try to extract domain robust features or transfer knowledge from source to target domains.

Activity Recognition. To achieve location independent activity recognition, FALAR [27] reconstructed the amplitude data by Class Estimated Basis Space Singular Value Decomposition (CSVD) to discard most location information. EI [11] achieved environment and user independent activity recognition using amplitude based on adversarial networks, composed of a Convolutional Neural Network (CNN) feature extractor, an activity recognizer and a domain discriminator. CsiGAN [25] enabled activity recognition adaptive to users based on semi-supervised Generative Adversarial Network (GAN), leveraging limited unlabeled amplitude data to produce diverse fake samples to train a robust discriminator. WiLISensing [6] built a CNN model to recognize activities using amplitude and fine-tuned the model in new locations. Sheng et al. [18] achieved action recognition using amplitude by integrating CNN with Bi-directional Long Short Term Memory (BLSTM) and fine-tuned the model in new scenarios. HAR-MN-EF [20] achieved environment independent activity recognition by leveraging Matching Network with activity enhanced amplitude. Zhang et al. [31] recognized activities using amplitude images by a Dense-LSTM model. They synthesized variant activity data through CSI transformation to mitigate activity inconsistency and subject specific issues. Ma et al. [15] recognized activities across locations and users using amplitude, employing 2DCNN as the activity classifier, 1DCNN as the state machine, and reinforcement learning for neural architecture search.

Gesture Recognition. WiAG [22] achieved gesture recognition independent on locations and orientations as well as environmental dynamics, by generating virtual amplitude samples of gestures in the target domains through a translation function. CrossSense [30] enabled amplitude training samples to be collected once and used across sites, by employing an Artificial Neural Network (ANN) to train a roaming model that generated synthetic training samples of gestures or gaits for each target site. WiADG [35] and JADA [34] exploited unsupervised joint adversarial domain adaptation to realize gesture recognition based on phase difference, mapping the unlabeled target data and the labeled source data to a domain-invariant feature space. Yang et al. [28] enabled one-shot gesture recognition via a Siamese recurrent convolutional architecture based on phase difference, which used transferable pairwise loss to remove structured noise such as individual heterogeneity and various measurement conditions. To alleviate the effort of retraining in a new scenario or for a new user, Wang et al. [23] recognized gestures based on CSI images leveraging a deep similarity evaluation network. Also based on similarity evaluation, Ma et al. [14] achieved gesture recognition based on amplitude image that could recognize new types of gestures, or gestures performed by a new user or in a new scenario. Kang et al. [12] recognized gestures based on DFS through adversarial learning with feature disentanglement and an attention scheme, adaptive to environments, users, locations and orientations.

2.3 Multi-modal Action Recognition

Different information in CSI depicts the action from different dimensions, which can be combined to enhance the performance. WiFit [13] monitored body-weight

exercises robust to environments and users, using amplitude, phase difference and Doppler velocity spectrum. TL-HAR [2] transformed amplitude and phase to images for multiple human activity recognition. MatNet-eCSI [21] employed Matching Network to perform one-shot learning to recognize activities in a new environment. They enhanced activity dependent information and eliminated activity unrelated information and fused amplitude and phase after enhancement. Shi et al. [19] authenticated users by recognizing their activities, using amplitude, relative amplitude and Short Time Fourier Transform (STFT) holograms. They employed an unsupervised domain discriminator to mitigate the impact of locations and environmental changes. FingerDraw [24] tracked finger drawings by exploiting the CSI-quotient model and the Fresnel zone model. It canceled out noise in amplitude and random offset in phase, and quantified the correlation between CSI dynamics and object displacements. DANGR [8] achieved gesture recognition by fusing amplitude and phase. They exploited GAN to augment the dataset and adopted Multi-kernel Maximum Mean Discrepancy (MK-MMD) to shrink the domain discrepancy. WiVi [33] recognized activities by combining vision and WiFi. It employed CNN to extract features from WiFi and a variant of C3D model for vision sensing. An ensemble neural network was constructed for decision. Bakalos et al. [4] detected abnormal activities based on BLSTM, fusing RGB imagery, thermographic imagery and CSI to capture the temporal inter-dependency.

2.4 The Difference

Compared with the prior works, our method is a universal solution for WiFi sensing tasks. Although we only evaluated on activity recognition and gesture recognition, the method can be extended to other CSI sensing tasks. Our method requires only one sample per class in the target domains, hence it achieves one-shot domain adaptation, requiring less effort than semi-supervised and unsupervised methods. Our method fuses amplitude and phase in feature extraction considering the consistency between them, hence enhances the recognition performance, achieving higher accuracy than domain generalization methods. Apart from user diversity and location variation, our method achieves acceptable accuracy for cross room action recognition, which is the most challenging issue in cross domain sensing.

3 Overview

The goal of our work is to recognize actions in new scenarios with the knowledge from the training scenarios. The training scenarios are known as the source domains, whereas the new scenarios, to which the recognition model is adapted, are known as the target domains. The source domains have adequate labeled data to train an accurate recognition model, while the target domains have very few samples, far from training a recognition model. Each scenario is regarded as a domain. Domain changes are caused by user diversity, location variation, room

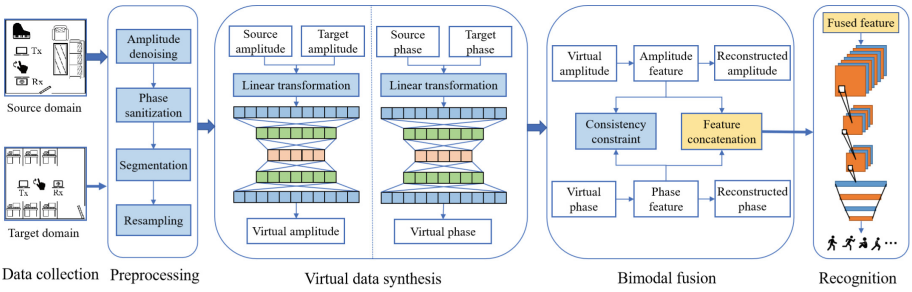


Fig. 1. Framework of the proposed method.

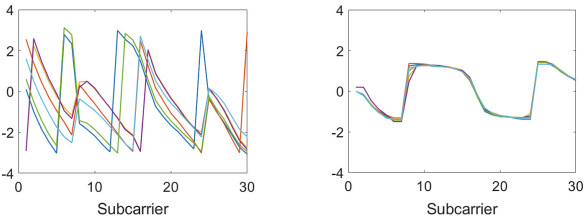
change and etc. Our aim is to achieve cross domain action recognition with little effort by means of one-shot domain adaptation, which requires only one sample per class in each target domain.

To achieve the aim, we propose the method framework as shown in Fig. 1, composed of data collection, preprocessing, virtual data synthesis, bimodal fusion and action recognition. The key components are virtual data synthesis and bimodal fusion. As there is only one sample per action in the target domain, the data synthesis component synthesizes a large number of virtual labeled samples for the target domain, leveraging linear transformation and the synthesis autoencoder. The bimodal fusion component extracts the consistent features from the amplitude and the phase under the constraint of Cosine Similarity via the fusion autoencoder, and concatenates the consistent features for the subsequent action classification. Before virtual data synthesis, the amplitude and the phase are retrieved from CSI, undergoing noise reduction in amplitude and shift removal in phase. After bimodal fusion, the concatenated features of the virtual samples are used to train the action classifier and recognize the actions in the target domain.

4 Methodology

4.1 Data Preprocessing

The amplitude and the phase retrieved from the raw CSI cannot be used directly for action recognition, as the amplitude contains noise and the phase contains shifts caused by Carrier Frequency Offset (CFO) and Sampling Frequency Offset (SFO). CFO is due to the unsynchronization of the central frequencies between the transmitter and the receiver, and SFO is due to the unsynchronized clocks. The data preprocessing goes through amplitude denoising, phase sanitization, data segmentation and data resampling. The raw amplitude sequences are denoised by a median filter, and the raw phase sequences are sanitized by a linear transformation method [17]. The denoised and sanitized data are segmented to retrieve the action part by mean square deviations. Finally the data are resampled to keep a consistent length by interval sampling.



(a) Raw phase (b) Sanitized phase

Fig. 2. Raw and sanitized phase.

As the raw phase values have significant errors due to CFO and SFO, we apply linear transformation to sanitize it [17]. The raw phase of the i-th subcarrier can be expressed as

$$\angle \hat{H}_i = \angle H_i - 2\pi \frac{m_i}{F} \Delta t + \beta + Z \tag{1}$$

where $\angle H_i$ is the true phase, m_i denotes the subcarrier index ranging from -28 to 28, F is the size of Fast Fourier Transform (FFT), Δt is the timing offset due to SFO, β is the unknown phase offset due to CFO, and Z is the measurement noise. Defining two terms k and b as

$$k = \frac{\angle \hat{H}_N - \angle \hat{H}_1}{m_N - m_1}, \quad b = \frac{1}{N} \sum_{i=1}^{N} \angle \hat{H}_i \tag{2}$$

where N is the number of subcarriers, Eq. (1) can be rewritten as

$$\angle \hat{H}_i = \angle H_i - k m_i - b \tag{3}$$

Apply Eq. (3) to the raw phase, we can obtain the sanitized phase. Figure 2 shows the raw and the sanitized phase, each curve representing a packet.

4.2 Virtual Data Synthesis

Suppose $D^s = \{(x_i^s, y_i^s) | i = 1, 2, \cdots, n^s\}$ represents the dataset of the source domains, where x_i^s is an action sample and y_i^s is the action label. Suppose $D^t = \{(x_i^t, y_i^t) | i = 1, 2, \cdots, n^t\}$ represents the datatset of the target domain, where x_i^t is an action sample and y_i^t is the action label. n^s and n^t are the numbers of samples in the source and the target domain, satisfying $n^t \ll n^s$. For one-shot domain adaptation, $n^t = C$, where C is the number of action types, i.e. each action has one sample in the target domain. As the target domain has only C samples, it is too few to train a classifier or fine-tune a pre-trained classifier. To achieve cross domain action recognition, we need to augment the dataset of the target domain. We synthesize virtual samples for the target domain, utilizing the real samples in the source domains and the only sample per action in the target domain. The method of virtual data synthesis is illustrated in Fig. 3.

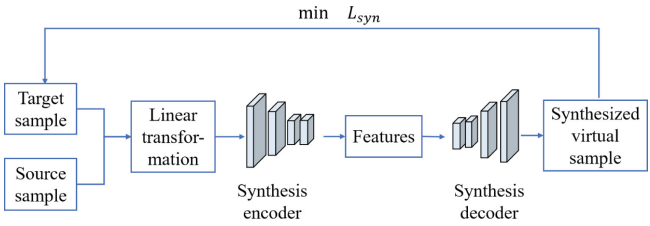


Fig. 3. Virtual data synthesis.

Linear Transformation. The CSI data are the superposition of the signals from all the paths. The static paths are reflected by the static objects, such as floor, ceiling and furniture, while the dynamic paths are reflected by the moving objects, i.e. the sensing targets. Therefore, the dynamic paths are related to the actions, whereas the static paths are environment related. As the synthesized data are for the target domain, they should contain both the action information and the target environment information. The action information mainly comes from the source domains and the target environment information comes from the target domain. Taking a sample from the source domains, denoted as $(x_i^s, y_i^s) \in D^s$, and taking a sample from the target domain, denoted as $(x_j^t, y_j^t) \in D^t$, satisfying $y_i^s = y_j^t$, we apply linear transformation on x_i^s and x_j^t to synthesize the virtual sample as

$$x_i^v = \alpha x_i^s + \beta x_j^t \tag{4}$$

where α and β are the domain weights, to balance the importance of the source domain and the target domain. The label of x_i^v is set as y_j^t.

Dynamic Domain Weights. In the synthesis of a virtual sample, the source domain sample and the target domain sample belong to the same action class, so they share the same action information but have different environment information. Through linear transformation, the common action information and both the environment information are contained in the synthesized sample. Since the synthesized samples are for the target domain, the target environment information should be retained while the source environment information should be removed. This can be achieved by setting the proper domain weights. When α decreases and β increases, x_i^v gets approaching the target domain sample x_j^t. Because different domains have different environmental characteristics, we propose to set the domain weights adaptively. We apply the synthesis autoencoder on the virtual sample x_i^v to determine the values of α and β. The synthesis autoencoder is a CNN, whose structure is shown in Table 1.

Target Domain Enhancement. In addition to the domain weights, the synthesis autoencoder tries to make the virtual data close to the target domain data. Using Mean Square Error (MSE) to calculate the distance between the

Table 1. Structure of the autoencoder

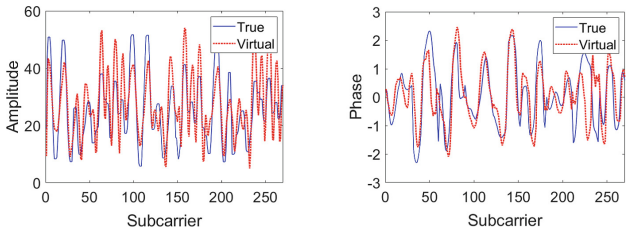
Input layer	(200×270) 2D matrix
Conv2d	Channels=(1,4), kernel size=(5,5), stride=2
Conv2d	Channels=(4,16), kernel size=(3,3), stride=2
Conv2d	Channels=(16,32), kernel size=(3,3), stride=2
Conv2d	Channels=(32,32), kernel size=(2,2), stride=1
ConvTranspose2d	Channels=(32,32), kernel size=(3,2), stride=1
ConvTranspose2d	Channels=(32,16), kernel size=(3,4), stride=2
ConvTranspose2d	Channels=(16,4), kernel size=(3,4), stride=2
ConvTranspose2d	Channels=(4,1), kernel size=(4,4), stride=2
Output layer	(200×270) 2D matrix

synthesized virtual data x_i^v and the target domain data x_i^t, the loss function of the synthesis autoencoder can be defined as

$$\min \quad L_{syn} = \frac{1}{n^s} \sum_{i=1}^{C} \sum_{j=1}^{n^c} (x_{ij}^v - x_i^t)^2 \tag{5}$$

where n^c is the number of samples in the source domains for an action. By training the synthesis autoencoder, the domain weights α and β can be set dynamically and the synthesized virtual samples contain the action information and the target environment information.

Figure 4 shows the virtual amplitude and the virtual phase synthesized from the real sample in the target domain with the action knowledge from the source domains. The virtual samples are similar to the target domain data, thus can be used to train the recognition model for the target domain.



(a) Virtual amplitude (b) Virtual phase

Fig. 4. Synthesized virtual samples

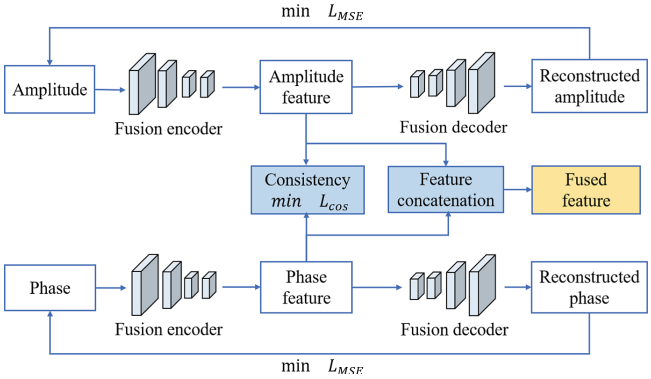


Fig. 5. Fusion of amplitude and phase.

4.3 Fusion of Amplitude and Phase

From CSI we can extract amplitude and phase. The two modals are affected by the actions simultaneously, hence have consistency. To learn the features more accurately, we propose to fuse the two modals in the feature level under the criterion of consistency, as illustrated in Fig. 5. The key in the structure is the fusion autoencoder, which is a CNN as shown in Table 1.

Suppose x^a denotes the amplitude sample and x^p denotes the phase sample. The fusion autoencoder is applied on them to extract the action features in each modal. We denote the features of x^a as $f^a = encoder(x^a)$ and the features of x^p as $f^p = encoder(x^p)$. Considering that x^a and x^p sense the same action simultaneously and have consistency in the feature space, we leverage Cosine Similarity to measure their consistency and minimize their cosine distance. The goal can be expressed as

$$\begin{aligned} \min \quad L_{cos} &= 1 - cos(f^a, f^p) \\ &= 1 - \frac{f^a \cdot f^p}{\parallel f^a \parallel \cdot \parallel f^p \parallel} \\ &= 1 - \frac{\sum_{i=1}^{N} f_i^a f_i^p}{\sqrt{\sum_{i=1}^{N} (f_i^a)^2} \sqrt{\sum_{i=1}^{N} (f_i^p)^2}} \end{aligned} \tag{6}$$

In addition to minimizing the distance between the amplitude and the phase, the extracted features should keep the original feature of each modal. Hence the reconstructions of x^a and x^p, denoted as $\bar{x}^a$ and $\bar{x}^p$, should be as close to the originals as possible, i.e.

$$\begin{aligned} \min \quad L_{MSE}(x^a, \bar{x}^a) &= \frac{1}{N} \sum_{i=1}^{N} (x_i^a - \bar{x}_i^a)^2 \\ \min \quad L_{MSE}(x^p, \bar{x}^p) &= \frac{1}{N} \sum_{i=1}^{N} (x_i^p - \bar{x}_i^p)^2 \end{aligned} \tag{7}$$

The total loss of the fusion autoencoder can be defined as

$$L_{fusion} = L_{cos} + \lambda L_{MSE}(x^a, \bar{x}^a) + \lambda L_{MSE}(x^p, \bar{x}^p) \tag{8}$$

where λ is the balance factor. The extracted features f^a and f^p are concatenated, expressed as $f = [f^a, f^p]$, as the fused feature for action classification.

4.4 Action Classification

We employ CNN as the action classifier, taking the fused feature of amplitude and phase $f = [f^a, f^p]$ as the input. f^a and f^p are the encoded features, having the dimensions of $22 \times 31 \times 32$. After concatenation, the fused feature f has the dimension of $22 \times 31 \times 64$. The action classifier contains 3 convolutional blocks followed by 3 fully connected layers, in which each convolutional block is composed of 1 convolutional layer and 1 max-pooling layer. ReLU is the activation function and cross entropy is the classification loss. The structure of the action classifier is shown in Table 2.

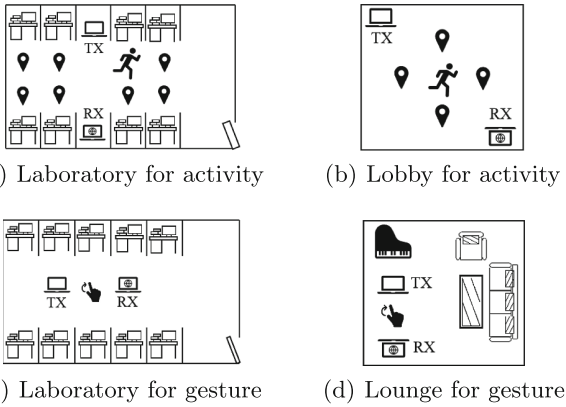
Table 2. Structure of the action classifier

Input layer	$(22 \times 31 \times 64)$ 3D matrix
Convolutional	Channels=(64,64), kernel size=(3,3), stride=1
Convolutional	Channels=(64,32), kernel size=(2,2), stride=1
Convolutional	Channels=(32,16), kernel size=(2,2), stride=1
FullyConnected	Nodes=1024
FullyConnected	Nodes=512
FullyConnected	Nodes=64
Output layer	Nodes=#actions

5 Evaluations

5.1 Experimental Setup and Datasets

To evaluate the proposed method, we conducted extensive experiments on activity recognition and gesture recognition. We deployed two laptops equipped with Intel WiFi Link 5300 as the transmitter and the receiver, each having 3 antennas. The sampling rate was 100 Hz for activity recognition and 1000 Hz for gesture recognition. As there were 9 antenna pairs and each had 30 subcarriers, each CSI packet contained 270 dimensions. After data preprocessing, each action sample had a time sequence of 200 packets with 270 dimensions.



(a) Laboratory for activity (b) Lobby for activity

(c) Laboratory for gesture (d) Lounge for gesture

Fig. 6. Experimental scenarios and setup.

Activity Datasets. We set up two scenarios for activity recognition. One was a laboratory of size 6 m × 8 m and the other was a lobby of size 5 m × 5 m. In the laboratory, as shown in Fig. 6(a), the transmitter and the receiver were placed on both sides of the aisle. The volunteer performed 4 activities (squatting down, standing up, walking and jumping) at 8 locations (as the location pins show in Fig. 6(a), adjacent locations are about one meter apart), and each activity was repeated 20 times per location. In the lobby, as shown in Fig. 6(b), the transmitter and the receiver were placed at two opposite corners of the room. The volunteer performed 6 activities (squatting down, standing up, walking, jumping, falling and climbing) at 5 locations (as the location pins show in Fig. 6(b), adjacent locations are about one meter apart), and each activity was repeated 20 times per location.

Gesture Datasets. We also collected data for gesture recognition in two scenarios: a laboratory of size 6 m × 8 m and a lounge of size 5 m × 5 m. The transmitter and the receiver were placed in the aisle with 2 m apart in the laboratory and 1.6 m apart in the lounge, as shown in Fig. 6(c) and Fig. 6(d). Six volunteers, with different heights and weights, performed 6 gestures in both scenarios. The gestures were common letters {L, O, V, W, Z, S}, and each was repeated 20 times per person.

5.2 Experimental Results of Activity Recognition

Across Locations. In the laboratory, 7 locations were taken as the source domains and the left 1 as the target domain. In the lobby, 4 locations were taken as the source domains and the left 1 as the target domain. We used 20 samples per activity in each source domain and only 1 sample per activity in the target domain to train the adaptation model, and used 19 samples per activity in the target domain to test the adaptation model. The method was evaluated

at the 8 locations in the laboratory and the 5 locations in the lobby, with each location as the target domain in turn. The detailed results are shown in Table 3. The average accuracy of activity recognition at the target locations reached 85.03% in the laboratory and 90.53% in the lobby for one-shot bimodal domain adaptation, outperforming using only the amplitude or the phase.

Across Rooms. For activity recognition across rooms, we took the laboratory as the source room and the lobby as the target room, and vice verse. The experiments were conducted at each location in the target room in turn. We used the data at all the locations in the source room and only 1 sample per activity at the current location in the target room to train the adaptation model, and used 19 samples per activity at the current location in the target room to test the adaptation model. The average accuracy achieved 87.90% from the laboratory to the lobby and 82.40% from the lobby to the laboratory, as shown in Table 5. The accuracy of bimodal fusion outperformed the amplitude or the phase alone.

5.3 Experimental Results of Gesture Recognition

Across Users. For gesture recognition in the laboratory and the lounge, we took 5 users as the source domains and the left 1 as the target domain. We used 20 samples per gesture in each source domain and only 1 sample per gesture in the target domain to train the adaptation model, and used 19 samples per gesture in the target domain to test the adaptation model. The method was evaluated on the 6 users in the laboratory and the lounge, with each user as the target domain in turn. As shown in Table 4, the average accuracy achieved 91.67% in the laboratory and 85.53% in the lounge for the target users using one-shot bimodal domain adaptation, outperforming using only the amplitude or the phase.

Across Rooms. For gesture recognition across rooms, we used the laboratory as the source room and took the lounge as the target room, and vice verse. The experiments were conducted on each user in the target room in turn. We used the data of 5 users in the source room and 1 sample per gesture of the left user in the target room to train the adaptation model, and used 19 samples per gesture of the target user in the target room to test the adaptation model. The average accuracy of cross room gesture recognition achieved 83.04% from the laboratory to the lounge and 88.01% from the lounge to the laboratory, as shown in Table 5. The accuracy of bimodal fusion outperformed the amplitude or the phase alone.

5.4 Comparison with Existing Works

We compared our method with the state of the art. We compared with the data augmentation method in Fido [5], which leveraged Variational Autoencoder (VAE) to synthesize virtual samples from the labeled samples. We compared

Table 3. Activity recognition at target locations (%)

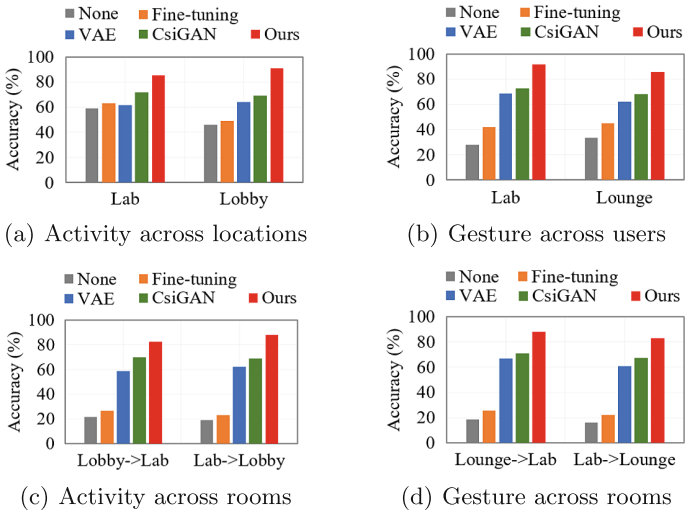
Scenario	Source	Target	Amplitude	Phase	Bimodal
Lab	2-8	1	73.68	75.00	78.95
	1, 3-8	2	77.63	68.42	78.95
	1-2, 4-8	3	80.26	73.68	80.26
	1-3, 5-8	4	89.47	71.05	93.42
	1-4, 6-8	5	72.37	72.37	77.63
	1-5, 7-8	6	86.84	85.53	94.74
	1-6, 8	7	78.95	81.58	90.79
	1-7	8	77.63	75.00	85.53
		Average	**79.60**	**75.32**	**85.03**
Lobby	2-5	1	88.60	85.09	92.11
	1, 3-5	2	84.21	75.44	85.96
	1-2, 4-5	3	85.09	84.21	94.74
	1-3, 5	4	82.46	81.58	92.98
	1-4	5	81.58	78.07	86.84
		Average	**84.30**	**80.88**	**90.53**

Table 4. Gesture recognition on target users (%)

Scenario	Source	Target	Amplitude	Phase	Bimodal
Lab	2-6	1	85.09	79.82	90.35
	1, 3-6	2	88.60	78.95	90.35
	1-2, 4-6	3	92.11	88.60	97.37
	1-3, 5-6	4	83.33	78.07	85.96
	1-4, 6	5	86.84	78.95	92.11
	1-5	6	88.60	71.05	93.86
		Average	**87.43**	**79.24**	**91.67**
Lounge	2-6	1	75.44	69.30	77.19
	1, 3-6	2	81.58	75.44	85.09
	1-2, 4-6	3	91.23	86.84	95.61
	1-3, 5-6	4	79.82	71.93	83.33
	1-4, 6	5	82.46	80.70	85.09
	1-5	6	83.33	68.42	86.84
		Average	**82.31**	**75.44**	**85.53**

Table 5. Action recognition in target rooms (%)

Action	Source	Target	Amplitude	Phase	Bimodal
Activity	Lab	Lobby	83.16	79.21	**87.90**
	Lobby	Lab	76.48	73.03	**82.40**
Gesture	Lab	Lounge	79.53	71.49	**83.04**
	Lounge	Lab	83.77	75.73	**88.01**



(a) Activity across locations (b) Gesture across users

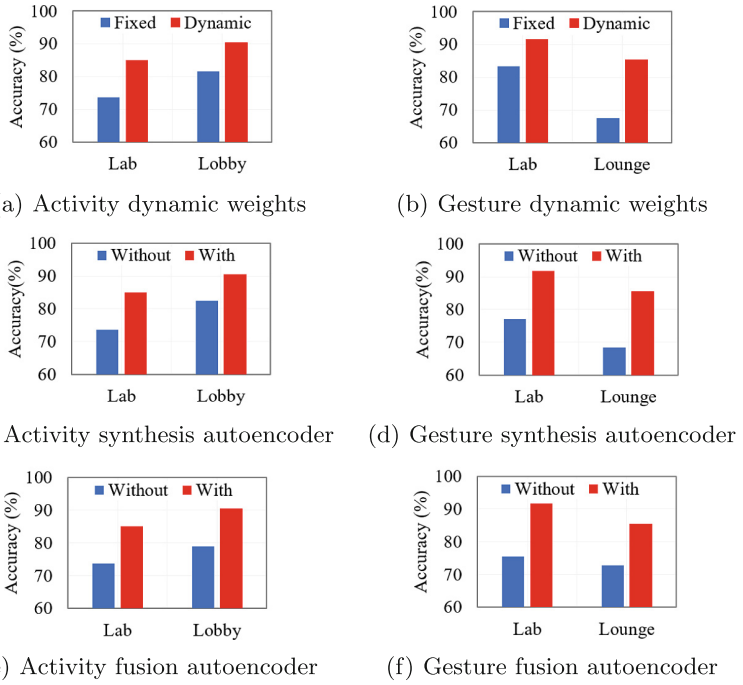
(c) Activity across rooms (d) Gesture across rooms

Fig. 7. Comparison with existing works.

with CsiGAN [25], which leveraged semi-supervised GAN for CSI-based activity recognition, using unlabeled samples to generate virtual samples to train a robust discriminator. We reimplemented the VAE-based data synthesis method in FiDo and reused the original implementation of CsiGAN. These methods were compared with no adaptation and fine-tuning as well. Using the same activity datasets and gesture datasets in Sect. 5.1, the results are shown in Fig. 7. Across locations and across users as well as across rooms in multiple scenarios, our method achieved the best performance in the target domains. The reasons are two-fold. Firstly, our method utilized one real sample per action in the target domain together with the real samples in the source domains to achieve domain adaptation, which could help capture the characteristics of the target domain more accurately. Secondly, our method enhanced the action features by fusing amplitude and phase, while FiDo and CSiGAN only made use of amplitude. We also compared our method with the existing works using only amplitude, and our method still achieved the best performance in the target domains.

5.5 Ablation Study

Effect of Dynamic Domain Weights. We apply linear transformation to synthesize virtual data, which assigns the weights to the source and the target domains, balancing their importance on data synthesis. As different domains have different characteristics, the domain weights should be set differently. Our method sets the domain weights adaptively through the synthesis autoencoder. To verify its effect, we compared it with fixed domain weights ($\alpha = 0.5$, $\beta = 0.5$). The comparison results are shown in Fig. 8(a) for activities and Fig. 8(b) for gestures. Dynamic domain weights outperformed fixed domain weights.



(a) Activity dynamic weights (b) Gesture dynamic weights

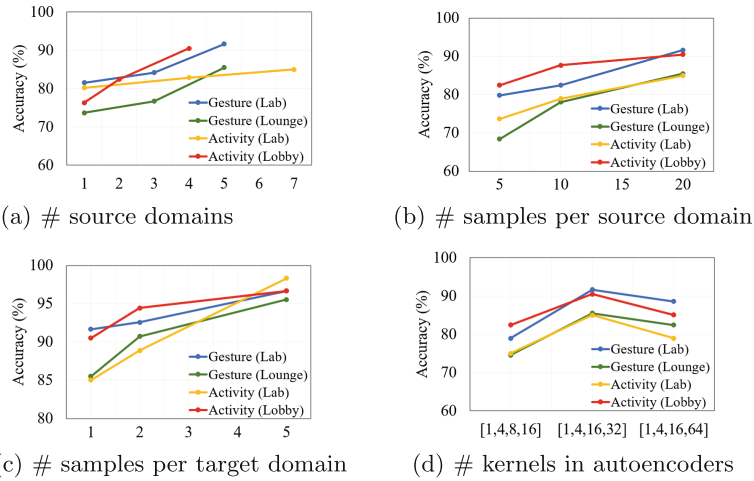
(c) Activity synthesis autoencoder (d) Gesture synthesis autoencoder

(e) Activity fusion autoencoder (f) Gesture fusion autoencoder

Fig. 8. The ablation study.

Effect of Synthesis Autoencoder. In addition to linear transformation, we leverage the synthesis autoencoder to reconstruct the virtual data and make them close to the target domain data, in order to enhance the target environment information. To verify the effect of the synthesis autoencoder, we compared it with not using the synthesis autoencoder. The comparison results are shown in Fig. 8(c) for activities and Fig. 8(d) for gestures. The synthesis autoencoder improved the accuracy by a large margin.

Effect of Fusion Autoencoder. The fusion autoencoder extracted the consistent features from different but consistent modals and meanwhile kept their original action features. To verify its effect, we compared it with not using the fusion autoencoder, which extracted the features from the amplitude and the phase and concatenated them. The comparison results are shown in Fig. 8(e) for activities and Fig. 8(f) for gestures. The fusion autoencoder improved the accuracy by a large margin. Without the fusion autoencoder, the features were extracted from the different modals separately, losing the consistency between them, causing the loss of useful information.



(a) # source domains (b) # samples per source domain

(c) # samples per target domain (d) # kernels in autoencoders

Fig. 9. The parameter study.

5.6 Parameter Study

Number of Source Domains. To evaluate the impact of the number of source domains, we tested on different numbers of source domains with one target domain. For activity recognition across locations in the laboratory, we took 1, 4, 7 locations as the source domains and a different location as the target domain. For activity recognition across locations in the lobby, we took 1, 2, 4 locations as the source domains and a different location as the target domain. For gesture recognition across users in the laboratory and in the lounge, we took 1, 3, 5 users as the source domains and 1 different user as the target domain. The recognition accuracies on the target domains are shown in Fig. 9(a). More source domains led to better performance, but incurred higher cost.

Number of Samples in Source Domains. We evaluated the impact of the number of samples per action in the source domains. For activity recognition across locations in the laboratory and the lobby and for gesture recognition across users in the laboratory and the lounge, we set the number of samples in each source domain as 5, 10, 20 per action and the target domain had 1 sample per action. The recognition accuracies on the target domains are shown in Fig. 9(b), indicating that the accuracy increased with the number of samples in the source domains, but requiring more effort.

Number of Samples in Target Domains. We also evaluated the impact of the number of samples per action in the target domain. For activity recognition across locations in the laboratory and the lobby and for gesture recognition across users in the laboratory and the lounge, we set the number of samples in the target domain as 1, 2, 5 per action, and the number of samples in each source

domain were 20 per action. The recognition accuracies on the target domain are shown in Fig. 9(c). Although more samples achieved higher accuracy, they incurred higher effort. One-shot adaptation achieved acceptable accuracy.

Number of Kernels in Autoencoders. The structure of the synthesis autoencoder and the fusion autoencoder is important to domain adaptation. We evaluated the performance of different numbers of convolution kernels in the autoencoders. We compared 3 groups of kernel numbers, which were [1,4,8,16], [1,4,16,32] and [1,4,16,64] for the encoders and inverse for the decoders. The results on the target domain are shown in Fig. 9(e). [1,4,16,32] kernels achieved the best performance, while [1,4,8,16] kernels could not extract the features effectively and [1,4,16,64] kernels caused overfitting.

6 Conclusions

To achieve cross domain CSI action recognition and sensing, we propose a one-shot domain adaptation method based on virtual data synthesis and bimodal fusion. The method synthesizes the virtual data for the target domains by a linear transformation function and the synthesis autoencoder, which sets the domain weights adaptively for different domains in linear transformation and makes the virtual data close to the target domain data. To improve the recognition accuracy with only one sample per action in the target domain, the method fuses the amplitude and the phase by the fusion autoencoder, which extracts their action features and keeps their consistency to the same actions. Real-world evaluations of activity recognition and gesture recognition in multiple scenarios achieved high accuracy across locations and users as well as rooms. The limitation of the current method is that it still requires one sample per action in the target domain. Zero-shot domain adaptation is our next step work.

References

1. Ahmed, H.F.T., Ahmad, H., Aravind, C.V.: Device free human gesture recognition using Wi-Fi CSI: a survey. Eng. Appl. Artif. Intell. **87**, 103281 (2020)
2. Arshad, S., Feng, C., Yu, R., Liu, Y.: Leveraging transfer learning in multiple human activity recognition using WiFi signal. In: IEEE 20th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM), pp. 1–10 (2019)
3. Bai, Y., Wang, Z., Zheng, K., Wang, X., Wang, J.: Widrive: adaptive wifi-based recognition of driver activity for real-time and safe takeover. In: 39th International Conference on Distributed Computing Systems (ICDCS), pp. 901–911. IEEE (2019)
4. Bakalos, N., Voulodimos, A., Doulamis, N., Doulamis, A., Papasotiriou, K., Bimpas, M.: Fusing RGB and thermal imagery with channel state information for abnormal activity detection using multimodal bidirectional LSTM. In: Abie, H., et al. (eds.) CPS4CIP 2020. LNCS, vol. 12618, pp. 77–86. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-69781-5_6

5. Chen, X., Li, H., Zhou, C., Liu, X., Wu, D., Dudek, G.: Fido: ubiquitous fine-grained wifi-based localization for unlabelled users via domain adaptation. In: Proceedings of the Web Conference (WWW), pp. 23–33. ACM (2020)
6. Ding, X., Jiang, T., Li, Y., Xue, W., Zhong, Y.: Device-free location-independent human activity recognition using transfer learning based on CNN. In: ICC Workshops, pp. 1–6. IEEE (2020)
7. Halperin, D., Hu, W., Sheth, A., Wetherall, D.: Tool release: gathering 802.11n traces with channel state information. ACM SIGCOMM CCR **41**(1) (2011)
8. Han, Z., Guo, L., Lu, Z., Wen, X., Zheng, W.: Deep adaptation networks based gesture recognition using commodity WiFi. In: WCNC, pp. 1–7. IEEE (2020)
9. Han, Z., Lu, Z., Wen, X., Zhao, J., Guo, L., Liu, Y.: In-air handwriting by passive gesture tracking using commodity WiFi. IEEE Commun. Lett. **24**(11), 2652–2656 (2020)
10. He, Y., Chen, Y., Hu, Y., Zeng, B.: WiFi vision: sensing, recognition, and detection with commodity MIMO-OFDM WiFi. IEEE Internet Things J. **7**(9), 8296–8317 (2020)
11. Jiang, W., et al.: Towards environment independent device free human activity recognition. In: MobiCom, pp. 289–304. ACM (2018)
12. Kang, H., Zhang, Q., Huang, Q.: Context-aware wireless based cross domain gesture recognition. IEEE Internet Things J. **8**(17), 13503–13515 (2021)
13. Li, S., Li, X., Lv, Q., Tian, G., Zhang, D.: WiFit: ubiquitous bodyweight exercise monitoring with commodity wi-fi devices. In: 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation, pp. 530–537. IEEE (2018)
14. Ma, X., Zhao, Y., Zhang, L., Gao, Q., Pan, M., Wang, J.: Practical device-free gesture recognition using WiFi signals based on Metalearning. IEEE Trans. Industr. Inf. **16**(1), 228–237 (2020)
15. Ma, Y., et al.: Location- and person-independent activity recognition with WiFi, deep neural networks, and reinforcement learning. ACM Trans. Internet Things **2**(1), 1–25 (2021)
16. Niu, K., Zhang, F., Wang, X., Lv, Q., Luo, H., Zhang, D.: Understanding WiFi signal frequency features for position-independent gesture sensing. IEEE Trans. Mob. Comput. **21**(11), 4156–4171 (2021)
17. Qian, K., Wu, C., Yang, Z., Liu, Y., Zhou, Z.: PADS: passive detection of moving targets with dynamic speed using PHY layer information. In: ICPADS, pp. 1–8 (2014)
18. Sheng, B., Xiao, F., Sha, L., Sun, L.: Deep spatial-temporal model based cross-scene action recognition using commodity WiFi. IEEE Internet Things J. **7**(4), 3592–3601 (2020)
19. Shi, C., Liu, J., Borodinov, N., Leao, B., Chen, Y.: Towards environment-independent behavior-based user authentication using WiFi. In: IEEE 17th International Conference on Mobile Ad Hoc and Sensor Systems (MASS), pp. 666–674 (2020)
20. Shi, Z., Zhang, J.A., Xu, R., Cheng, Q., Pearce, A.: Towards environment-independent human activity recognition using deep learning and enhanced CSI. In: GLOBECOM, pp. 1–6. IEEE (2020)
21. Shi, Z., Zhang, J.A., Xu, R.Y., Cheng, Q.: Environment-robust device-free human activity recognition with channel-state-information enhancement and one-shot learning. IEEE Trans. Mob. Comput. **21**(2), 540–554 (2022)

22. Virmani, A., Shahzad, M.: Position and orientation agnostic gesture recognition using WiFi. In: MobiSys, pp. 252–264. ACM (2017)
23. Wang, J., Gao, Q., Ma, X., Zhao, Y., Fang, Y.: Learning to sense: deep learning for wireless sensing with less training efforts. IEEE Wirel. Commun. **27**(3), 156–162 (2020)
24. Wu, D., et al.: FingerDraw: sub-wavelength level finger motion tracking with WiFi signals. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. (IMWUT) **4**(1), 1–27 (2020)
25. Xiao, C., Han, D., Ma, Y., Qin, Z.: CsiGAN: robust channel state information-based activity recognition with GANs. IEEE Internet Things J. **6**(6), 10191–10204 (2019)
26. Xie, Y., Li, Z., Li, M.: Precise power delay profiling with commodity WiFi. In: MobiCom, pp. 53–64. ACM (2015)
27. Yang, J., Zou, H., Jiang, H., Xie, L.: Fine-grained adaptive location-independent activity recognition using commodity WiFi. In: WCNC, pp. 1–6. IEEE (2018)
28. Yang, J., Zou, H., Zhou, Y., Xie, L.: Learning gestures from WiFi: a siamese recurrent convolutional architecture. IEEE Internet Things J. **6**(6), 10763–10772 (2019)
29. Zhang, F., et al.: Towards a diffraction-based sensing approach on human activity recognition. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. (IMWUT) **3**(1), 1–25 (2019)
30. Zhang, J., Tang, Z., Li, M., Fang, D., Nurmi, P., Wang, Z.: CrossSense: towards cross-site and large-scale WiFi sensing. In: MobiCom, pp. 305–320. ACM (2018)
31. Zhang, J., et al.: Data augmentation and dense-LSTM for human activity recognition using WiFi signal. IEEE Internet Things J. **8**(6), 4628–4641 (2021)
32. Zheng, Y., et al.: Zero-effort cross-domain gesture recognition with Wi-Fi. In: MobiSys, pp. 313–325. ACM (2019)
33. Zou, H., Yang, J., Das, H.P., Liu, H., Spanos, C.J.: WiFi and vision multimodal learning for accurate and robust device-free human activity recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2019)
34. Zou, H., Yang, J., Zhou, Y., Spanos, C.J.: Joint adversarial domain adaptation for resilient WiFi-enabled device-free gesture recognition. In: 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 202–207 (2018)
35. Zou, H., Yang, J., Zhou, Y., Xie, L., Spanos, C.J.: Robust WiFi-enabled device-free gesture recognition via unsupervised adversarial domain adaptation. In: 27th International Conference on Computer Communication and Networks (ICCCN), pp. 1–8. IEEE (2018)