Cross-domain extendable gesture recognition system using WiFi signals

Yuxi Qin[1], Su Pan[1], and Zibo Li[2]

[1]Nanjing University of Posts and Telecommunications
[2]Nanjing Tech University

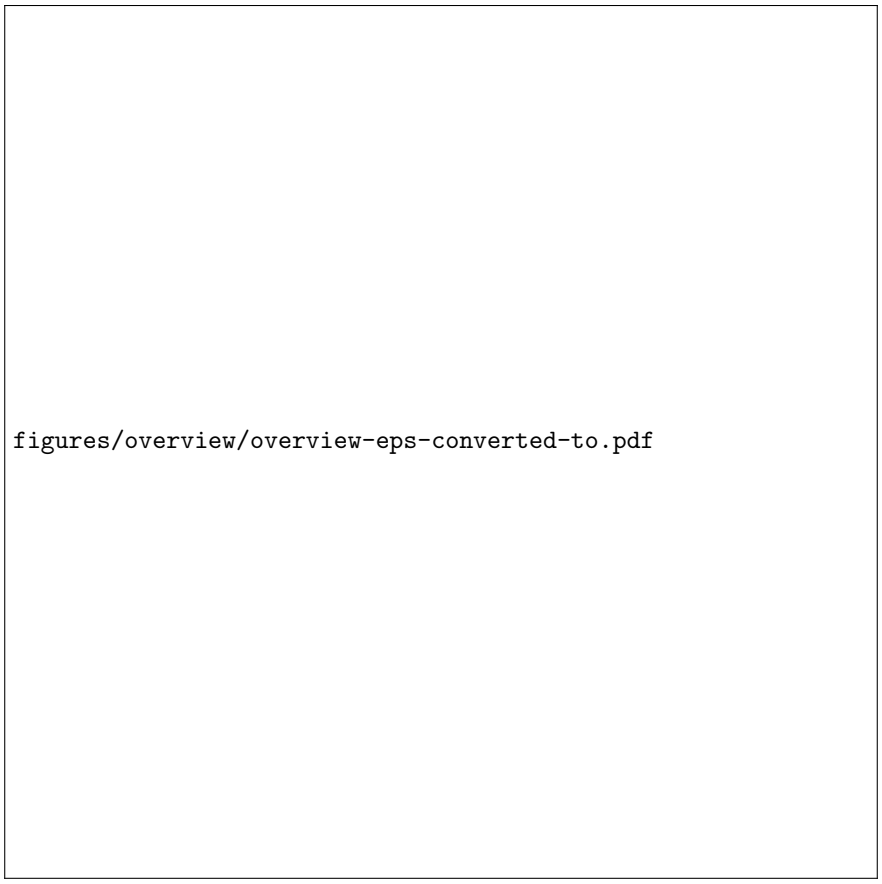
April 11, 2023

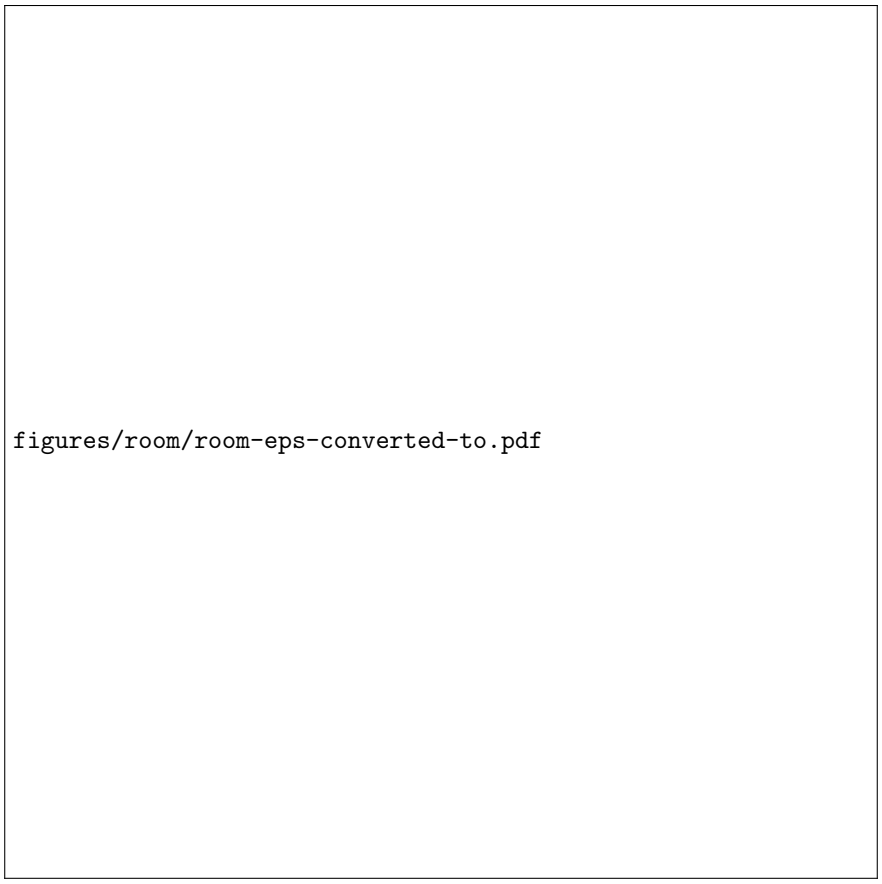
Abstract

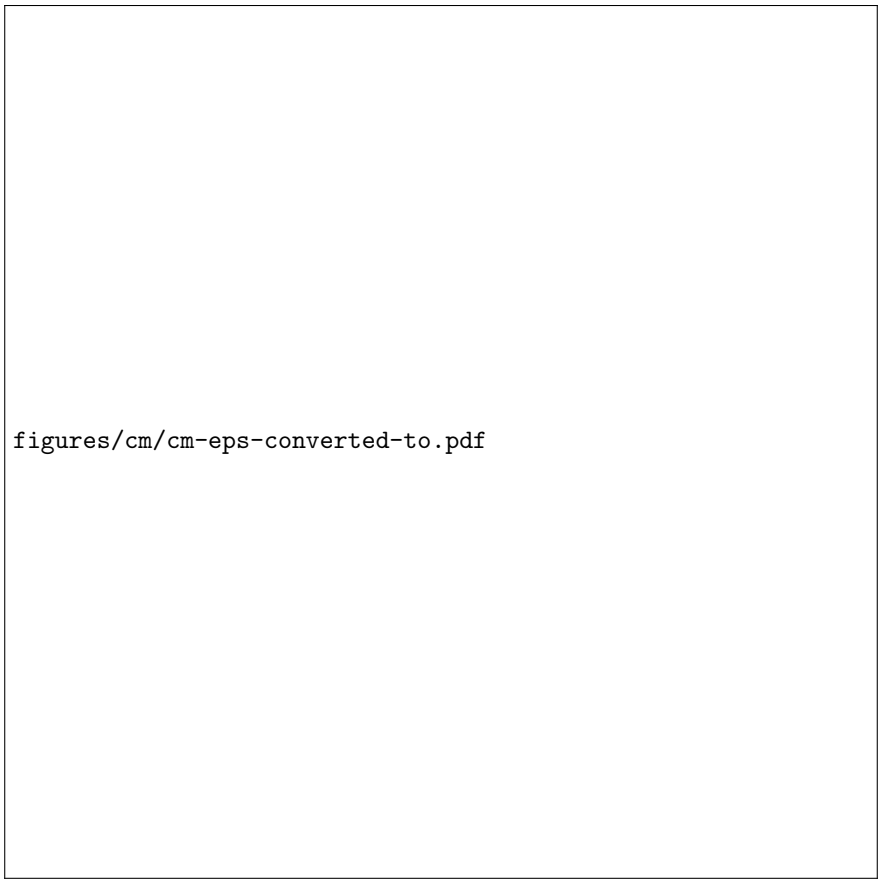
This letter proposes a cross-domain WiFi-based gesture recognition system (WiCross) based on a dynamically weighted multi-label generative adversarial network. Most existing WiFi-based gesture recognition systems are user, orientation, and environment sensitive, which limits the application of WiFi sensing. Compared with the influence of users and environments on WiFi sensing systems, the influence of orientation on WiFi sensing systems is more difficult to remove. To alleviate the confusion caused by the orientation more effectively, we arrange the transmitting and receiving antennas according to the characteristics of the Fresnel region. We propose to dynamically weight different links according to users' orientations and use a multi-label generative adversarial network to obtain domain-independent features. More importantly, WiCross can use domain-independent features to classify some unknown gestures without modifying any code or data set. Lightweight computing resource consumption allows WiCross to respond in real-time. The experimental results show that WiCross can achieve an in-domain recognition accuracy of 93.54% and a cross-domain recognition accuracy of 93.11%

Hosted file

`Cross-domain extendable gesture recognition system using WiFi signals.tex` available at https://authorea.com/users/605341/articles/634841-cross-domain-extendable-gesture-recognition-system-using-wifi-signals







Cross-domain extendable gesture recognition system using WiFi signals

Yuxi Qin, Su Pan, Zibo Li

This letter proposes a cross-domain WiFi-based gesture recognition system (WiCross) based on a dynamically weighted multi-label generative adversarial network. Most existing WiFi-based gesture recognition systems are user, orientation, and environment sensitive, which limits the application of WiFi sensing. Compared with the influence of users and environments on WiFi sensing systems, the influence of orientation on WiFi sensing systems is more difficult to remove. To alleviate the confusion caused by the orientation more effectively, we arrange the transmitting and receiving antennas according to the characteristics of the Fresnel region. We propose to dynamically weight different links according to users' orientations and use a multi-label generative adversarial network to obtain domain-independent features. More importantly, WiCross can use domain-independent features to classify some unknown gestures without modifying any code or data set. Lightweight computing resource consumption allows WiCross to respond in real-time. The experimental results show that WiCross can achieve an in-domain recognition accuracy of 93.54% and a cross-domain recognition accuracy of 93.11%

Introduction: With the popularity of intelligent devices, human-computer interaction is no longer limited to mice, keyboards, and touch screens. In recent years, the gesture recognition based on channel state information (CSI) of WiFi has drawn considerable research attention, such as WiTrace [1], WiHGR [2] and WiPass [3]. Compared with vision-based and wearable sensor-based sensing schemes, the CSI-based scheme has the advantages of low cost, easy deployment, non-light of sight sensing, and device-free. In addition, the CSI-based scheme does not record private information, such as fingerprints, faces, and indoor environments. Therefore, the CSI-based scheme can effectively protect users' privacy while maintaining high recognition accuracy.

Gestures cause changes in the wireless channel when the WiFi signals propagate. Different movement trajectories of the hands lead to unique CSI fluctuations, so we can achieve accurate gesture recognition according to the features of different gestures. Gesture signals are strongly dependent on orientations [4]. In addition, users and environments will also affect recognition accuracy.

To address this problem, many systems have been proposed for cross-domain recognition. Zhang *et al.* [5] presented Widar 3.0 using body-coordinate velocity profiles to obtain domain-independent features. WiDIGR [6] mapped the orthogonal directions to the movement direction and utilized the spectrograms to estimate the walking direction angle. EI [7] adopted adversarial learning to remove the environment and subject specific information. WiGr [8] uses the similarity between the query samples and the class prototypes in the embedding space to perform the gesture classification.

However, these systems have obvious limitations. In Widar 3.0, when gestures move in a non-straight line, the direction of speed changes all the time, which requires Widar 3.0 to process the signals at each time point. Therefore, the complex calculation limits the real-time performance of the system. WiDIGR was designed suitable for the gestures in a single direction. EI and WiGr did not sufficiently consider that a single link cannot perceive objects moving along the elliptic boundaries formed by the Fresnel zones.

In this letter, we propose WiCross, a WiFi-based cross-domain extendable gesture recognition system. WiCross removes domain information, such as users, orientations, and environments. Compared with existing systems, WiCross pays more attention to extracting orientation-independent features. We propose dynamic link weighting to perceive gestures in different orientations better. We evaluate the importance of different links according to the signal characteristics contained in different links and calculate the corresponding weights. We propose to use a multi-label generative adversarial network to extract domain-independent features. We set up four sub-discriminators for gestures and domain information and apply dynamic link weights to all sub-discriminators. When the gesture features extracted by the generator can be recognized by the gesture discriminator, and all the domain discriminators consider the gestures to belong to any domain with the same probability, we obtain fully domain-independent features.

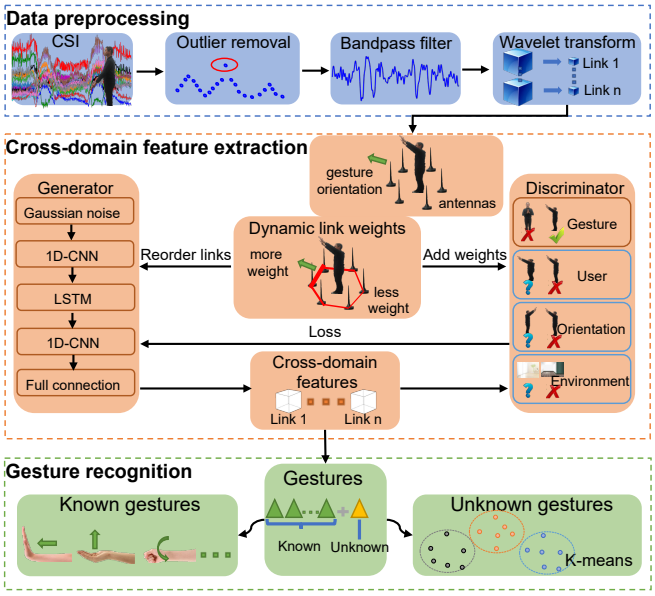


Fig. 1. Overview of WiCross

WiCross uses only one pair of WiFi devices with multiple antennas to reduce system complexity and realize real-time gesture recognition. Most existing systems are not extendable, and adding new gestures requires retraining the models. We constrained the feature distribution of the same class to make the features in the same class more clustered as far as possible. We used the K-Means algorithm to estimate the total number of unknown gesture classes without supervision, and then classify gestures. The main contributions of this work can be summarised as follows:

- We propose a dynamic link weighting algorithm to assign more weights to valuable links and rearrange all links according to the weights to solve the dependency of gestures on orientations.
- We design a novel cross-domain extendable gesture recognition system based on the weighted multi-label generative adversarial network, which can extract domain-independent features from WiFi signals.
- We add the coefficient of variation of the same gestures to the loss function of the gesture discriminator to make the features of the same gestures more clustered, so that WiCross can classify unknown gestures without modifying any code and retrain the model.
- We implement a prototype of WiCross with only one pair of commercial WiFi devices and evaluate its performance in terms of accuracy and efficiency. The experimental results show that the recognition accuracy of cross-domain is 93.11%, and the average response time is 27.3 ms.

System design

Overview: As shown in Figure 1, WiCross contains three parts: data preprocessing, cross-domain feature extraction, and gesture recognition. We first remove noise and non-gesture information from the signals, so the preprocessed data contains cleaner gesture information. However, the data preprocessing cannot remove domain-dependent information because we cannot judge which information is domain-dependent based on a single sample. Therefore, we need to obtain domain-independent features from the preprocessed data. We propose a dynamic link weighting algorithm to remove gestures' dependence on orientations and weight important features. We use a multi-label generative adversarial network to obtain cross-domain features. Finally, we use an extended gesture discriminator to classify known and unknown gestures.

Data preprocessing: The raw CSI contains a large amount of noise and a few outliers, so we need to preprocess the signals first. We use a sliding window to detect outliers by applying the Pauta criterion [9]. Since all the links are obtained from the same receiver device, we remove some noise by conjugate multiplication of the two links [10]. The frequency of gesture signals is usually between 5 and 30 Hz, so a band-pass Butterworth filter is used to remove high-frequency noise and static components. It is worth mentioning that keeping static components in in-domain scenes can effectively improve the recognition accuracy,

but it will hinder cross-domain recognition. Finally, we utilize wavelet transform to reduce the dimension of samples and preserve important features.

Dynamic link weighting: According to many researches on Fresnel zone [6, 11], we know that objects moving along the boundary of the Fresnel zone cannot be perceived, which is determined by the propagation characteristics of wireless signals and cannot be changed by any algorithms. The simplest way to solve this problem is to use multiple links in different orientations. As shown in Figure 2, we evenly arrange the three transmitting antennas and the three receiving antennas alternately into a circle. The human body weakens the links sensing gestures behind the body, and our arrangement allows the links to sense the gestures no matter which direction the user is facing. The importance of each link is different for different gestures and orientations. We believe the higher the proportion of gesture signals in the link, the more valuable it is. We calculate the value of the i^{th} link by equation 1.

$$V_i(X,t) = \frac{\sum \text{peaks}(X,t_g) - \sum \text{troughs}(X,t_g)}{K * \delta^2(X,t_s)}, \tag{1}$$

where X is a CSI segment in the time span t. t_g and t_s denote the time span with and without gesture. $\text{peaks}(X,t_g)$ and $\text{troughs}(X,t_g)$ are functions that calculate the peaks and troughs of the signals in the time span t_g. $\delta^2(X,t_s)$ represents a function that calculates the variance of X in time span t_s, and K is a fixed scaling factor. The link value V is much larger than 1, so we utilize equation 2 to normalize V.

$$W_i = \frac{\exp(V_i)}{\sum_{j=1}^{n} \exp(V_j)}, \tag{2}$$

where n is the total number of links, and W represents the weights of links. Since the orientations and gestures may change, the weights are not fixed and will be recalculated for each sample.

Extractor: We use sub-extractors to extract cross-domain features from each denoised link. We reorder the link according to the W and take the link with the largest W_i as input of the first sub-extractor, and so on. The advantage of this idea is that the features with high value can be fixed in the first few sub-extractors, which is conducive to using prior knowledge to weigh the features and improve the recognition accuracy of the classifiers. Our extractor contains four parts: Gaussian noise, one-dimensional convolutional neural network (1D-CNN), long short-term memory (LSTM), and fully connected part. Adding Gaussian noise to the input effectively prevents over-fitting. LSTM and 1D-CNN show outstanding performance in processing time sequence data. LSTM can find the correlation between data at different time points, and 1D-CNN can efficiently extract the features in each time period. Since $relu$ activation function does not have the vanishing gradient problem, and there is no complex exponential calculation, the training speed is fast, so all activation functions in the extractor are $relu$.

Discriminator: The discriminator consists of two parts: gesture discriminator and domain discriminator. Gesture discriminator is built for gesture recognition. The input features should be usable by most ordinary classifiers, not only by specific network structures. Therefore, the gesture discriminator uses a simple fully connected network consisting of two hidden layers, the first hidden layer and the second hidden layer containing 40 and 20 $relu$ nodes, respectively. The activation function of the output layer uses $softmax$. To recognize the new custom gestures, features of the samples in the same class should be close together. Based on cross-entropy, we add the coefficient of variation of features as the loss function:

$$L_g(X,y,\hat{y}) = -\sum_{i=1}^{n} W_i \left(\sum_{j=1}^{s} y_j \log \hat{y_j} + \frac{\delta(X_i)}{\bar{X}} \right), \tag{3}$$

where s is the number of samples. y_j and $\hat{y_j}$ denote j^{th} true label and prediction label. $\delta(X)$ and $\bar{X}$ represent standard deviation and the mean of sample X.

Domain discriminator aim to extract domain-independent features. We divide domain information into three categories: user, orientation, and environment. We can find differences in user domains if we ask different users to make the same gesture in the same environment. Therefore domain discriminator is designed for finding differences in domains. When the feature from the generator is considered to be equally likely to come from all domains, we can assume that the domain-dependent feature has been removed. It is worth noting that the domain discriminator can not correct to identify domain information that does not represent does not contain domain information. For example, the domain discriminator recognizes all gestures in one room as being in another room, which just incorrectly maps the domain information, but the domain information still exists. Therefore, only when the recognition results are evenly distributed in each domain can we consider domain-dependent information to be removed. We define the loss functions for different users (L_u), as shown in equation 4, and the loss functions for different orientations (L_o) and environments (L_e) can be followed by analogy.

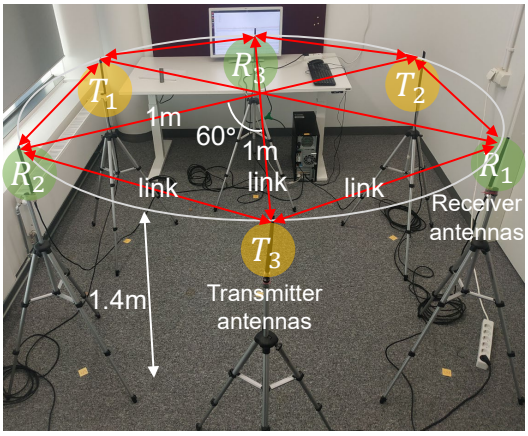


Fig. 2. Experimental scenario

$$L_u = \frac{1}{nd} \sum_{i=1}^{n} W_i \sum_{j=0}^{d} \delta^2 \left(\text{P}(X_i | y_{u,j}; \hat{y}_{u,j}) \right) \in [0,1], \tag{4}$$

where d is the total number of users in the dataset. $\text{P}(X_i|y_{u,j};\hat{y}_{u,j})$ denotes probability distribution of j^{th} user domain in all user domains for link i. $y_{u,j}$ and $\hat{y}_{u,j}$ represent true label and prediction label of j^{th} user domain. Now we obtain the loss function of all the sub-discriminators, so the loss function of the discriminator is:

$$Loss = \alpha L_u + \beta L_o + \gamma L_e + \varepsilon L_g, \tag{5}$$

where α, β, γ, and ε are fixed scaling factors, and we can adjust them according to our preference for certain domains.

Gesture recognition: In order to make the WiCross extendable for unknown gestures, we add an unknown class to the output of the existing gesture discriminator as a new gesture classifier and train this classifier using the features obtained from the generator. The purpose of not including the unknown class in the gesture discriminator is to make the gesture discriminator focus more on optimizing gesture features and reducing the complexity of the discriminator. For known classes, the classifier can indicate what the labels of the gestures are. However, unknown gestures do not have labels, so we use the K-Means clustering algorithm to classify unknown gestures. The K-Means algorithm first needs to traverse the all sample set to determine the number of classes, and then select the central points according to the number of classes for clustering. Therefore, we only need to assign a specific meaning to the clustering results to classify unknown gestures without making any changes to WiCross.

Implementation and evaluation

Experimental setting: We implement WiCross using two computers equipped with Intel 5300 WiFi cards and six antennas. As shown in Figure 2, the six antennas are arranged evenly in a circle with a radius of one meter. All antennas are extended with 3.5 m cables and placed 1.4 m above the ground using tripods. The band and bandwidth of WiFi signals are 5.32 GHz and 20 MHz, respectively. The transmission rate is 200 packets per second. We use the same gestures as Widar 3.0 (push, sweep, clap, slide, circle, zigzag), with each gesture repeated 20 times in every domain. Our dataset contains 6 users, 12 orientations and 6 environments, so the total number of gestures is 8640. In order to train the unknown

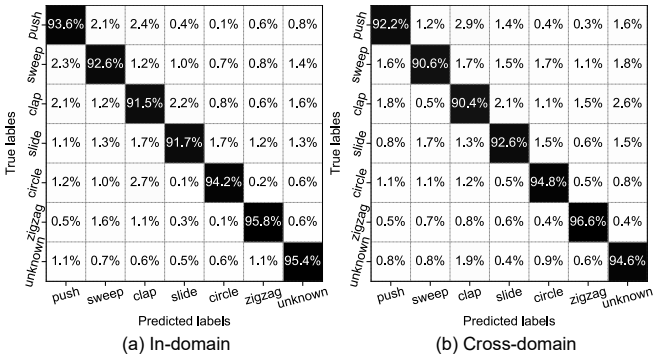


Fig. 3. Overall performance

Table 1: The recognition accuracy systems.

system	cross user	cross orientation	cross environment	in-domain
WiCross	92.4%	92.6%	93.3%	93.5%
Widar 3.0	88.9%	82.6%	92.4%	92.7%
WiGr	$\geq$ 90%	84%	92%	83.5-93.0%

gestures in the classifier, we randomly collect 200 samples of arbitrary gestures, and try to make each gesture as different as possible.

Overall performance: We evaluate the overall performance of WiCross in in-domain and cross-domain scenarios. Four-fold cross validation is used to separate the training set and the test set. Figure 3 displays the confusion matrix for gestures. The average accuracy of WiCross in the in-domain and cross-domain scenarios are 93.54% and 93.11%, and there is a tiny difference between the two scenarios. Therefore, WiCross removes domain information well. Table 1 shows the average accuracy of WiCross, Widar 3.0, and WiGr. We learn from the table that Widar 3.0 and WiGr cannot perform well in cross orientation scenarios, and the accuracy is much lower than that of other scenarios.

Extendibility evaluation: We use 2-6 unknown gestures to test the extendibility of WiCross, and the unknown gestures are inspired by the WriFi [12]. We write letters in the air, and the writing region size used in our experiments is 65×65 cm. We choose the letter 'A', 'B', 'E', 'G', 'H', and 'K' as the unknown gestures, and the experimental results are shown in Figure 4 (a). With the increase of unknown gestures, the detection accuracy does not change significantly, but the clustering accuracy of the K-Means algorithm decreases significantly. When the number of unknown gestures is no more than 3, the recognition accuracy of in-domain and cross-domain exceed 90% and 88%, respectively. In order to prove that the cross-domain features can be used by most classifiers, support vector machine (SVM), K-nearest neighbor (KNN), and decision tree are used to replace our classifier. The recognition accuracy is shown in Figure 4 (b). We learn that SVM has the highest recognition accuracy but little difference with KNN and decision tree. As a result, WiCross has good extendable performance.

Efficiency evaluation: We compare system efficiencies without considering unknown gestures, as shown in Table 2. Our test platform is built on a PC with Ubuntu 18.04 system, Intel i7-9700 processor, and 16G DDR3 memory. We learn from the table that the overall time

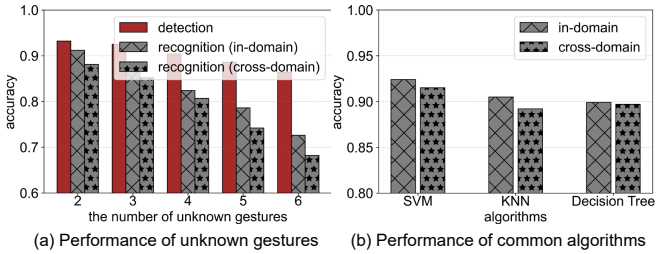


Fig. 4. Performance of extendibility.

Table 2: The time consumption for 1000 samples in different systems.

system	preprocessing time	training time	testing time
WiCross	30.6 s	30.6 s	27.3 s
Widar 3.0	55.7 s	38.2 s	20.5 s
WiGr	21.7 s	54.3 s	50.6 s

consumption of WiCross is less than that of Widar 3.0 and WiGr. Compared with Widar 3.0 and WiGr, WiCross focuses more on using prior knowledge to extract features instead of using complex system structures. WiCross and WiGr utilize one pair of WiFi devices, but Widar 3.0 utilizes at least three pairs of WiFi devices, which allows Widar 3.0 to spend more time preprocessing data.

Conclusion: This letter proposes and implements WiCross, a cross-domain WiFi-based gesture recognition system based on a dynamically weighted multi-label generative adversarial network. WiCross removes domain information of users, orientations, and environments from the CSI and generates domain-independent features that can be used in common classifiers. We propose a dynamic link weights algorithm to solve the cross orientations problem, which achieves better performance than existing systems. In addition, WiCross has good real-time performance while maintaining high accuracy.

Acknowledgment: This work was supported by the Chinese Government Scholarship under Grant 202108320270 and the Graduate Research and Innovation Projects of Jiangsu Province under Grant KYCX20_0717.

References

1 Wang, L., *et al.*: WiTrace: centimeter-level passive gesture tracking using OFDM signals, *IEEE Trans. Mob. Comput.*, **20**(4), pp. 1730-1745 (2021). https://doi.org/10.1109/TMC.2019.2961885.
2 Meng, W., *et al.*: WiHGR: a robust WiFi-based human gesture recognition system via sparse recovery and modified attention-based BGRU," *IEEE Internet Things J.*, **9**(12), pp. 10272-10282(2022). https://doi.org/10.1109/JIOT.2021.3122435.
3 Shen, X., *et al.*: WiPass: 1D-CNN-based smartphone keystroke recognition using WiFi signals, *Pervasive Mob. Comput.*, **73**, pp. 101393 (2021). https://doi.org/10.1016/j.pmcj.2021.101393
4 Zhang, Y., *et al.*: CSI-based location-independent human activity recognition using feature fusion, *IEEE T. Instrum. Meas.*, **71**, pp. 1-12 (2022). https://doi.org/10.1109/TIM.2022.3216419
5 Zhang, Y., *et al.*: Widar3.0: zero-effort cross-domain gesture recognition with Wi-Fi, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44**(11), pp. 8671-8688 (2022). https://doi.org/10.1109/TPAMI.2021.3105387
6 Zhang, L., *et al.*: WiDIGR: direction-independent gait recognition system using commercial Wi-Fi devices, *IEEE Internet Things J.*, **7**(2), pp. 1178-1191 (2020). https://doi.org/10.1109/JIOT.2019.2953488
7 Jiang, W., *et al.*: Towards environment independent device free human activity recognition, *Proc. 24th Annu. Int. Conf. Mobile Comput. Netw.*, pp. 289-304, Oct. (2018). https://doi.org/10.1145/3241539.3241548
8 Zhang, X., *et al.*: WiFi-based cross-domain gesture recognition via modified prototypical networks, *IEEE Internet Things J.*, **9**(11), pp. 8584-8596 (2022). https://doi.org/10.1109/JIOT.2021.3114309
9 Wan, Y., *et al.*: Outlier detection for monitoring data using stacked autoencoder, *IEEE Access*, **7**, pp. 173827-173837 (2019). https://doi.org/10.1109/ACCESS.2019.2956494
10 Yin, Y., *et al.*: Towards fully domain-independent gesture recognition using COTS WiFi device, *Electron. Lett.*, **57**(5), pp. 232-234 (2021). https://doi.org/10.1049/ell2.12097
11 Ren, S., *et al.*: Intelligent contactless gesture recognition using WLAN physical layer information, *IEEE Access*, **7**, pp. 92758-92767 (2019). https://doi.org/10.1109/ACCESS.2019.2927644
12 Fu, Z., *et al.*: Writing in the air with WiFi signals for virtual reality devices, *IEEE Trans. Mob. Comput.*, **18**(2), pp. 473-484 (2019). https://doi.org/10.1109/TMC.2018.2831709