WiCAR: WiFi-based in-Car Activity Recognition with Multi-Adversarial Domain Adaptation

Fangxin Wang
Simon Fraser University
Burnaby, British Columbia, Canada
fangxinw@sfu.ca

Jiangchuan Liu
Simon Fraser University University
Burnaby, British Columbia, Canada
jcliu@sfu.ca

Wei Gong
University of Science and Technology of China
Hefei, Anhui, China
weigong@ustc.edu.cn

ABSTRACT

In-car human activity recognition is playing a critical role in detecting distracted driving and improving human-car interaction. Among multiple sensing technologies, WiFi-based in-car activity recognition exhibits unique advantages since it does not rely on visible light, avoids privacy leaks and is cost-efficient with integrated WiFi signals in cars. Existing WiFi-based recognition systems mostly focus on the relatively stable indoor space, which only yield reasonably good performance in limited situations. Based on our field studies, the in-car activity recognition, however, is much more complicated suffering from more impact factors. First, the external moving objects and the surrounding WiFi signals can cause various disturbances to the in-car activity sensing. Second, considering the compact in-car space, different car models can also lead to different multipath distortions. Moreover, different people can also perform activities in different shapes. Such extraneous information related to specific driving conditions, car models and human subjects is implicitly contained for training and prediction, inevitably leading to poor recognition performance for new environment and people.

In this paper, we consider the impact of different *domains* including driving conditions, car models and human subjects on the in-car activity recognition with field measurements and experiments. We present *WiCAR*, a WiFi-based in-car activity recognition framework that is able to remove domain-specific information in the received signals while retaining the activity related information to the maximum extent. A deep learning architecture integrated with domain adversarial training is applied to domain independent activity recognition. Specifically, we leverage multi-adversarial domain adaptation to avoid the discriminative structures mixing up for different domains. We have implemented WiCAR with commercial-off-the-shelf WiFi devices. Our extensive evaluations show that WiCAR can achieve in-car recognition accuracy of around 95% in untrained domains, where it is only 53% for solutions without domain adversarial network and 83% for the state-of-the-art domain adversarial solution.

IWQoS '19, June 24–25, 2019, Phoenix, USA

ACM ISBN 978-1-4503-6778-3/19/06…$15.00
https://doi.org/10.1145/3326285.3329054

CCS CONCEPTS

• **Networks** → Wireless access points, base stations and infrastructure; Network performance modeling; • **Human-centered computing** → Ubiquitous and mobile computing systems and tools; • **Hardware** → Robustness.

KEYWORDS

In-Car Human Activity Recognition, Domain Adversarial Network, WiFi Signal Processing, Deep Learning

ACM Reference Format:
Fangxin Wang, Jiangchuan Liu, and Wei Gong. 2019. WiCAR: WiFi-based in-Car Activity Recognition with Multi-Adversarial Domain Adaptation. In *IEEE/ACM International Symposium on Quality of Service (IWQoS '19), June 24–25, 2019, Phoenix, AZ, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3326285.3329054

1 INTRODUCTION

Human activity recognition in cars [18] is playing a significant role in safe driving and human-car interaction. On one hand, it is helpful to effectively detect the driver's distraction behaviors [13, 14], such as watching cellphone during highway driving and forgetting shoulder check before changing lanes. Such detection together with a realtime warning can help avoid the potential driving accidents so as to greatly improve the driving safety. On the other hand, activity-based human-car interaction [19, 36] brings more possibilities to in-car entertainment. Particularly, in the emerging AR/VR and autonomous driving scenarios, people can enjoy the immersive experience with simple gesture-based control.

Many existing sensing technologies use cameras [31, 33], wearable sensors [11, 15] and radio-frequency identification (RFID) [5, 26] for activity recognition. However, cameras are limited to visible light and have privacy leakage concerns. Wearable sensor-based approaches usually lead to poor experience due to wearing extra devices. RFID-based solutions will introduce extra cost since current cars are not equipped with RFID devices. The recent WiFi-based sensing technologies have seen great success in indoor activity recognition [12, 17, 23–25, 27, 32, 34] and are promising to be applied in cars as WiFi is replacing Bluetooth for in-car entertainment [30]. The basic idea of WiFi-based recognition is that in-car activities will affect the surrounding WiFi signals, and the reflected signals by different activities exhibit distinct characteristics, which can be further classified by proper learning tools.

While sharing some commonalities with indoor activity recognition, the in-car situations still have unique challenges, as illustrated in Fig. 1. First, different from an indoor space that is relatively stable, a car usually experiences various and fast-changing driving

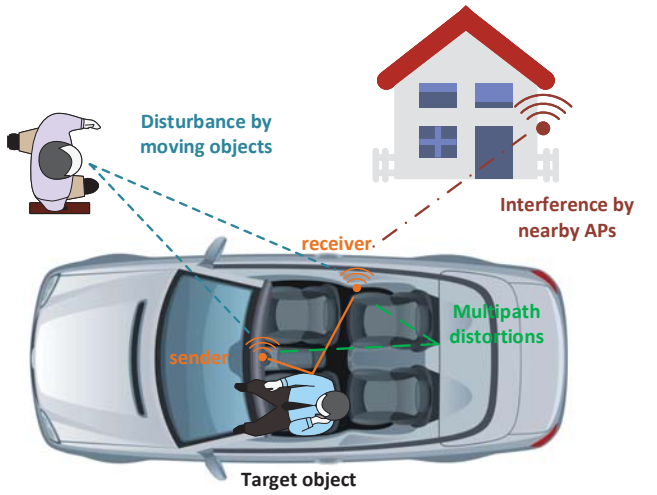


Figure 1: An illustration of multiple impacts on in-car activity recognition.

conditions, which will largely affect the recognition result. On one hand, the different wireless signals outside the car will interfere with the sensing signal in the car. For instance, the downtown area is full of APs and the nearby WiFi signals in the same channel can conflict with the target signal [20], while in open place the received signals can be much cleaner with less interference. On the other hand, external moving objects such as pedestrian and other cars may also cause signal fluctuations, affecting the in-car activity recognition accuracy. Second, the interior space, facilities and materials of different car models are highly heterogeneous and lead to distinct multipath distortions [22] in WiFi signals, which further degrade the recognition accuracy differently. At last, people varying in ages, genders, shapes and habits may affect the WiFi signals in different ways even when performing the same activity. Therefore, a well-trained recognition model in one domain-specific situation (e.g., specific driving condition, car model or human subject) may not work effectively when the target domain changes. This renders a practical in-car recognition system deployment infeasible since there are innumerable domain-specific situations.

To address this problem, we present *WiCAR*, a WiFi-based in-Car Activity Recognition framework that is able to remove the domain-specific information in the received signals while retaining the activity related information to the maximum extent. In this way, our recognition model that is trained over several particular domains can be well applied to other untrained driving conditions, car models and human subjects. WiCAR mainly consists of three components: a *feature encoder*, an *activity predictor* and a set of *domain discriminators*. Given temporally continuous in-car activities, the feature encoder employs stacked convolutional neural network (CNN) architectures to extract the characteristics in both time dimensions and frequency dimensions from WiFi spectrograms. The feature encoder cooperates with the activity predictor to achieve high activity recognition accuracy and simultaneously prevents the domain discriminators from distinguishing different domains. The state-of-the-art solutions [10, 35] only consider single-adversarial domain adaptation in the indoor scenario. The in-car scenarios, however, are much more complicated, and the discriminative structures of different domains can be easily mixed up under such solutions [16], leading to false domain discrimination. To this end, WiCAR leverages multi-adversarial domain adaptation to play against the feature encoder.

We have implemented WiCAR using Commercial Off-The-Shelf (COTS) Intel 5300 WiFi cards. With channel monitoring tools [9], we characterize the activity features by observing corresponding channel state information (CSI) changes. We have also performed extensive evaluations on 8 common in-car activities, involving 6 different driving conditions, 6 different car models, 8 volunteers and over 20,000 activity samples. The results show that WiCAR can achieve in-car activities accuracy of around 95% for new driving conditions, car models and human subjects, where it is only 53% for solutions without domain adversarial network and 83% for the state-of-the-art single-adversarial domain adaptation solution.

The rest of this paper is organized as follows. Section 2 provides a basic overview of our WiCAR framework. Section 3 describes our data preprocessing scheme to convert the raw signals into discriminative input spectrograms. We introduce our domain adversarial learning model in Section 4 in detail. Section 5 presents extensive experiments to evaluate the performance of our approach compared to the state-of-the-art solutions. Section 6 introduces the existing researches that are related to our work. We conclude this paper in Section 7.

2 SYSTEM OVERVIEW

In this paper, we present a deep learning-based environment independent in-car activity recognition framework named WiCAR. Being able to remove the domain-specific information in the collected CSI metrics while retaining as much activity related information as possible, WiCAR can be easily deployed in different car models and adapted to different driving conditions and human subjects after one pre-training process. WiCAR mainly consists of three components, i.e., CSI measurement, data preprocessing and deep domain adaptation, as illustrated in Fig. 2.

CSI measurement. We use a pair of WiFi transceiver devices deployed in cars to collect CSI metrics. For each specific domain combination (e.g., a person in a car in one particular driving condition), the background CSI is first collected as baseline. We then collect the CSI metrics when people are performing corresponding activities, which are required for further model training.

Data preprocessing. The collected raw CSI metrics need to be processed as feature representations before training. Low-pass filter and principal component analysis (PCA) are first applied to remove the high-frequency signal noise and extract the main wave characteristics from multiple subcarriers, respectively. We then design an activity detection method to automatically segment the data belonging to an activity. Short time Fourier transform (STFT) is further leveraged to generate feature spectrograms on both time dimension and frequency dimension, which are fed to the activity recognition component.

Deep domain adaptation network. We design an advanced deep learning model to effectively identify the discriminative features while preventing the impact of underlying domain information. In particular, a CNN-based feature encoder and a deep neural

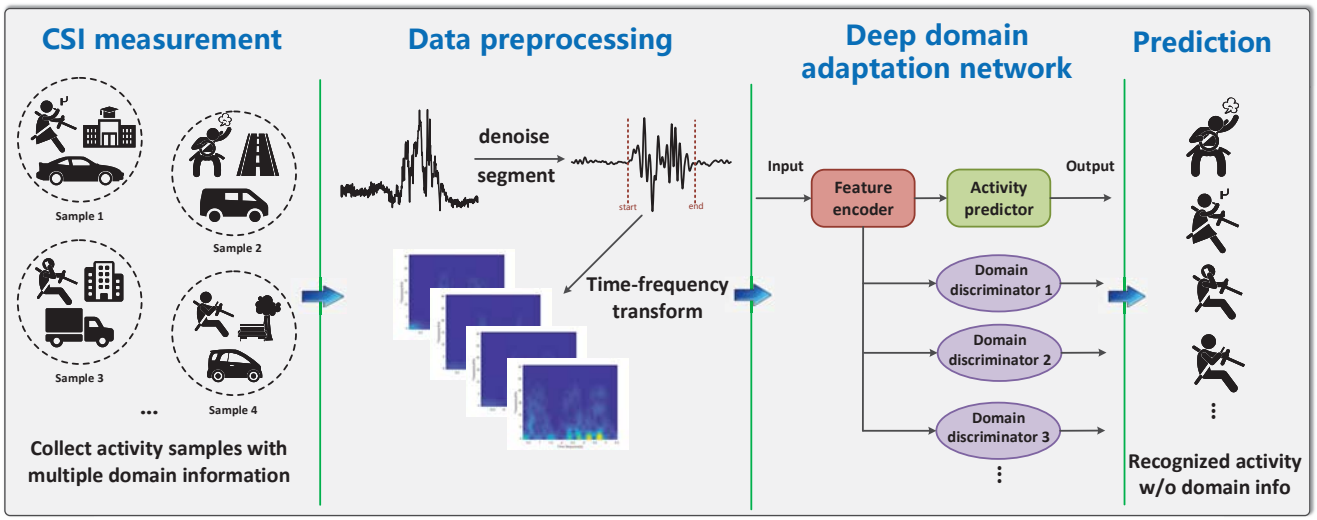


Figure 2: The framework of WiCAR.

network-based activity predictor cooperate to maximize the activity recognition accuracy. Besides, a set of domain discriminators are incorporated in our model to prevent the feature encoder from extracting domain related information in car. Note that our domain adaptation approach is able to learn the transferred features from the source domains to the target domains. In this way, the trained model can be directly applied to other untrained domains such as new cars and new drivers for activity recognition.

3 DATA PREPROCESSING

We first introduce the preliminary data preprocessing steps before the learning model, including CSI denoising, activity segmentation, and feature representation.

3.1 CSI Extraction and Denoising

When a person is performing activities inside a car, the received signals actually come through multiple paths, including *static paths* and *dynamic paths*, known as the multipath effect. The static paths include the line of sight path and those reflected paths by car seats, mirrors, etc., whose paths keep static during activities. The dynamic paths consist of those reflected by moving bodies and the objects outside the car since there exists relative movement when the car is running. As a result, we represent the CFR as:

$$H(f,t) = e^{-j2\pi\Delta f t}\left(H_s(f) + \sum_{k=1}^{P_d} a_k(f,t)e^{-j2\pi f \tau_k(t)}\right) \quad (1)$$

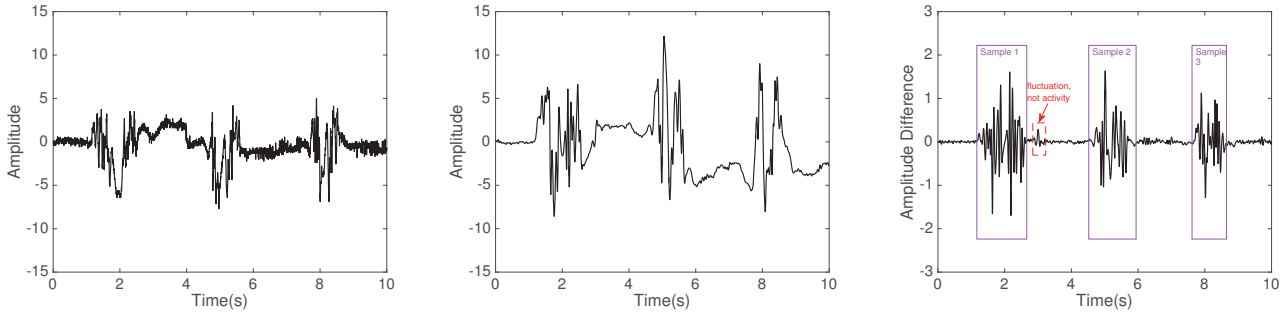
where $H_s(f)$ denotes the combined CFR of static paths, P_d is the number of dynamic paths, and $a_k(f,t)$ and $\tau_k(t)$ indicate the complex channel attenuation and time of flight for path k, respectively. Note that the COTS WiFi devices can have carrier frequency offset (CFO) [7] due to the lack of synchronization, which can induce unknown phase shift. Like prior works [3, 28], we use the CFR power (e.g., the multiplication of $H(f,t)$) to eliminate the phase noise so that the CFR power frequency and dynamic wave length change can be correlated.

We use a pair of transceivers deployed in a car to collect raw CSI metrics when people are performing activities. Modern WiFi devices that support 802.11n/ac standards have multiple antennas and can work in multi-input multi-output (MIMO) mode. The channel between each antenna pairs consists of N_s OFDM subcarriers. The Intel 5300 NICs used in our system can report CSI metrics of 30 selected OFDM subcarriers according to [9]. Thus, given N_T transmitting antennas and N_R receiving antennas, we can obtain a total of $30 * N_T * N_R$ streams, where we call the time-series CFR value of an OFDM subcarrier as a stream.

The collected raw streams are too noisy and cannot be directly applied as the input features. As the CFR frequency variance caused by human activities are mostly low-frequency component, we first use a low-pass filter (e.g., Butterworth filter) to remove high-frequency components such as white noise. Given the redundant similar CSI streams and the correlations therein [28], we then apply principal component analysis (PCA) to reduce the data dimensionality as well as extract the common characteristics from multiple subcarriers. Since the first component contains too much noise [28], we use the average of the second and third components (denoted as p-stream) for further processing. Fig. 3(a) and Fig. 3(b) compare a randomly selected raw CSI stream and the denoised p-stream. We can easily find that after denoising the p-stream becomes more smooth with little high-frequency noise. Besides, the fluctuation features in p-stream are more obvious than the raw stream, which also indicates that PCA can effectively extract the key features related to human activities.

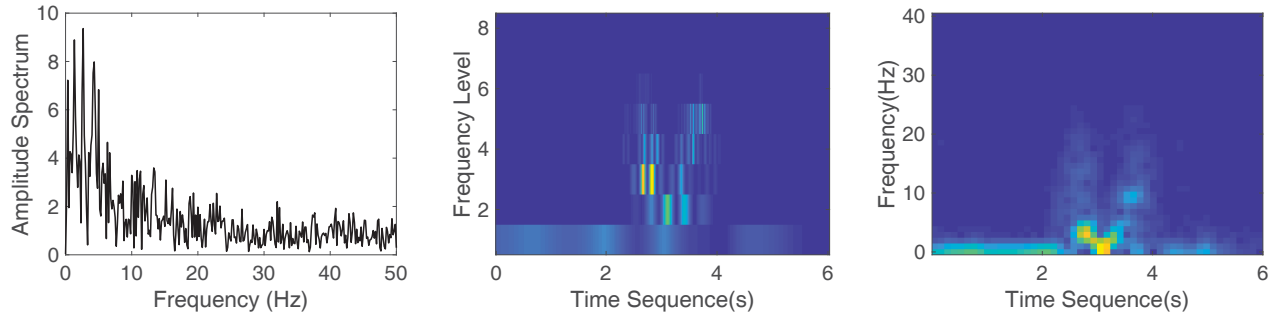
3.2 Activity Detection and Segmentation

With the denoised representative feature stream, we need to detect whether there exists an activity and segment the effective part from the stream. Since in real in-car recognition scenario different activities can have various time durations and gaps, a dynamic detection method is required to enable the adapted and real-time detection and segmentation. Fig. 3(b) shows the p-stream of three



(a) The CSI pattern of a randomly selected raw stream. (b) The CSI pattern of p-stream after low-pass filtering and PCA process. (c) First order difference of p-stream for activity sample segmentation.

Figure 3: The selected raw signal, PCA processed signal and first order difference signal for data preprocessing.



(a) The processed amplitude spectrum using FFT. (b) The processed spectrogram using DWT. (c) The processed spectrogram using STFT.

Figure 4: The feature representations of FFT, DWT and STFT.

"pushing right" activities. We can find that the signal is more volatile during activity and is more stable in absence of an activity. Such wave characteristics however are not clean enough for detection and segmentation with a lot of wave shifts. We therefore calculate the first order difference of the p-stream (denoted as dp-stream) for representation.

We denote an original p-stream signal as $\mathbf{h} = \{h_1, h_2, ..., h_n\}$, where each h_i is the value of a sampling point. Then the first order difference of the p-stream $\mathbf{g} = \{g_1, g_2, ..., g_n\}$ can be calculated as $g_i = h_i - h_{i-1}$. Fig. 3(c) shows the amplitude value of the corresponding dp-stream and we have two key observations for activity detection. First, when there is an activity, the dp-stream has visually obvious fluctuation and the wave variance is several orders of magnitude larger than when there is no activity. Second, the wave fluctuation lasts a time period rather than a moment, which is consistent with the duration of the activity. Based on the two observations, we develop an effective method to detect and segment an activity automatically. Given a time point t we consider a small time window $[t - T/2, t + T/2]$ around it. The standard deviation of t can be calculated as $\sigma_t = \sqrt{\frac{1}{K-1}\sum_{\tau_i=t-T/2}^{t+T/2}(g_{\tau_i} - \bar{g})^2}$, where K is the number of sample points within this time window and $\bar{g}$ is the mean value of them. We first calculate the average standard deviation σ_s when there is no activity and the average standard deviation σ_a of when there is an activity. A variance threshold δ_V is defined as $\delta_V = \alpha\sigma_s + (1-\alpha)\sigma_a$, where α is defined as the variance ratio. We segment an activity sample from the dp-stream based on the following rules.

$$Max : t_q - t_p \tag{2}$$

$$s.t.\ \sigma_{t_i} > \delta_V \text{ and } t_q - t_p \geq \delta_T, \forall\ t_i \in [t_p, t_q] \tag{3}$$

where δ_T is the time duration threshold, indicating the shortest possible time for an activity. Then the p-stream samples within $[t_p, t_q]$ are detected as an activity. As illustrated in Fig. 3(c), our detection method can effectively segment three real activity samples and the short pulse will not be detected as an activity. In our practical experiments, we set T as 0.2 s, α as 0.7 and δ_T as 1.6 s.

3.3 Feature Representation

The segmented p-stream clearly demonstrates the CSI wave patterns from the correlated multiple CSI streams. Yet this is not a good feature representation because it only shows the amplitude over time without exhibiting the frequency characteristics explicitly. Meanwhile, the frequency component is a good indicator since different activities usually have different frequency characteristics. We therefore seek time-frequency transform tools to reveal the

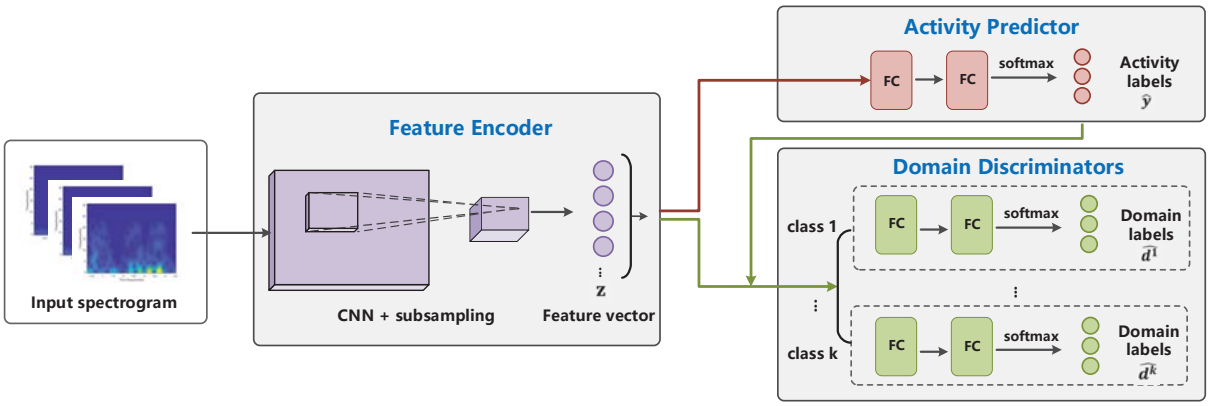


Figure 5: The main components of the learning model.

frequency features to feed the learning model. Fourier transform is the most common used method to obtain the frequency component. Fig. 4(a) shows the amplitude spectrum of the dp-stream of a sample signal using fast Fourier transform (FFT). This transform however loses the information of time dimension, which is not good for feature representation.

To obtain features on both time dimension and frequency dimension, our WiCAR system uses short-time Fourier transform (STFT) on every segmented p-stream samples to extract the frequency characteristics. Different from most existing systems using discrete wavelet transform (DWT) [24], STFT converts the time domain wavelet patterns to the time-frequency domain spectrogram, enabling more fine-grained and even resolution on frequency dimensions. Fig. 4(b) and Fig. 4(c) compare the spectrograms extracted using DWT and STFT, respectively. We can find that STFT is able to achieve similar resolution granularity in both time and frequency dimensions, which also benefits the future learning model with CNN architecture. In our system, we set the sampling rate of each antenna pair at 500 Hz so that we can detect a maximal frequency of 250 Hz, which is sufficient for human activities.

4 MULTI-ADVERSARIAL DOMAIN ADAPTATION MODEL

In this section, we present the methodology of the learning model of our WiCAR system. Fig. 5 illustrates the main components of our learning model, including a *feature encoder*, an *activity predictor* and a set of *domain discriminators*. These components are stacked together as a deep learning model to achieve the domain independent activity recognition. The feature encoder tries to only extract the effective features related to the human activity from the collected information while filtering out the domain related information. The activity predictor then maximizes the prediction accuracy based on the features. Besides, a set of domain discriminators are introduced to discriminate between the source and the target domains. Recall that our goal is to remove the domain-specific features while retaining the activity related features. Thus, the final objective is to minimize the loss of the activity predictor and maximize the loss of the domain discriminators. Through such a min-max game, the feature encoder will finally extract the domain-independent features for in-car activity recognition.

4.1 Feature Encoder

Let $x_i \in X$ be an input data sample and X is the whole input sample space. Each input data sample has an activity label $y_i \in Y$, where Y is the label space. Besides, each x_i also has its corresponding domain labels. The domain space D consists of the source domain space D_s with labels and the target domain space D_t without labels.

The input data are first fed into a feature encoder G_f to map the original complex high dimensional features to low dimensional feature representations $\mathbf{z}$. As a popular feature extraction network model, convolutional neural network (CNN) has exhibited powerful ability in extracting the spatial relationships from figures. Our input data samples have similar figure-like structures so that can be well applied with CNN-based feature extraction network model. Note that the CNN model requires a uniform input feature representation. However, the spectrogram converted from each segmented p-stream samples can have diverse time duration due to different activity durations. We observe from the collected data samples that the durations of all the activities are less than 5 seconds. Thus, we set a fixed time length at 5 seconds for every spectrogram and fill those short spectrograms with padding zeros. In this way, all the activity samples are transformed as uniform spectrogram feature matrices as the input data.

We use two stacked CNN layers to extract features from the input spectrograms, each followed with a rectified linear unit (ReLU) layer as the activation function. Max pooling layers are also applied to reduce the feature dimensions. As illustrated in Fig. 5, the output feature from the feature encoder can be represented as

$$\mathbf{z} = G_f(\mathbf{x}; \theta_f) = CNN(\mathbf{x}; \theta_f) \tag{4}$$

where θ_f denotes all the parameters of the feature encoding layers.

4.2 Activity Predictor

The activity predictor G_y takes the feature representation $\mathbf{z}$ from the feature encoder as input. In our model, two fully connect layers with corresponding activation layers are employed to learn the discriminative features. At last, a softmax layer is used to map the

features to a latent space with the same size as the activity label space. In this way, we can represent the predicted activity label distribution probabilities for input $\mathbf{x}$ as

$$\hat{\mathbf{y}} = G_y(G_f(\mathbf{x};\theta_f);\theta_y) \tag{5}$$

where θ_y denotes all the parameters in the activity predictor.

We therefore have the integrated loss function of G_f and G_y by calculating the cross-entropy function between the actual labels and the predicted label predictions as

$$\begin{aligned}\mathcal{L}_y(G_f, G_y) &= \mathbb{E}_{\mathbf{x},y}[-log\ G_y(G_f(\mathbf{x};\theta_f);\theta_y)] \\ &= -\frac{1}{|D_s|}\sum_{x_i \in D_s}\sum_{j=1}^{M} y_{ij}\ log(G_y(G_f(x_i;\theta_f);\theta_y))\end{aligned} \tag{6}$$

where $|D_s|$ is the number of samples belonging to the source domains and M is the number of activities labels. As mentioned, the target of the feature encoder and the activity predictor is to achieve a maximized recognition accuracy. Thus, during the training process, the feature encoder G_f cooperates with the activity predictor G_y to minimize the label prediction loss $\mathcal{L}_y$.

4.3 Multiple Domain Discriminators

Domain adaptation has seen success in transfer learning [21] due to its ability to learn transferable features between source domains and target domains. For the in-car activity recognition scenario, it is impossible to collect data samples from every domain for training since there always exist new driving conditions and new drivers. Therefore, we consider leveraging unsupervised domain adaptation [6] to train a generic in-car recognition model through filtering out those domain-specific characteristics even when the target domains are fully unlabeled.

In particular, a recent state-of-the-art indoor activity recognition model [10] uses single domain discriminator for domain adaptation, which, however, is not sufficient for in-car activity recognition. The activity recognition in cars can be affected by multiple impact factors (classes) such as driving conditions, human subjects and car models. The underlying features within each particular class usually exhibit specific structures, indicating the boundaries of different classes. Besides, the collected data for different classes can be not evenly distributed. Yet the existing single-adversarial domain adaptation methods require mixing up the discriminative structures, which easily leads to false alignment of discriminative structures and further degrades the domain independent activity recognition. To address this issue, we for the first time propose to apply multi-adversarial domain adaptation for in-car activity recognition.

In our model, we incorporate a set of domain discriminators $G_d = \{G_d^1, G_d^2, ..., G_d^K\}$, where K is the number of classes. Each domain discriminator consists of two fully connected layers together with ReLU activation function. And a softmax layer is also applied at last to generate the domain distributions for each class. Similarly, the domain space can be divided into K classes as $D = \{D^1, D^2, ..., D^K\}$. Each input sample should have a domain label d_i^k for every class D^k. Each domain discriminator takes as input the concatenation of the feature representations $\mathbf{z}$ from the feature encoder and the label distributions $\hat{\mathbf{y}}$ from the activity predictor, and predicts the domain labels distributions $\hat{\mathbf{d^k}}$ of the corresponding class k as:

$$\hat{\mathbf{d^k}} = G_d^k(\mathbf{z}, \hat{\mathbf{y}};\theta_d^k) = G_d^k(G_f(\mathbf{x};\theta_f), G_y(G_f(\mathbf{x};\theta_f);\theta_y);\theta_d^k) \tag{7}$$

where θ_d^k is the total parameters of the k-th domain discriminator. Note that the feature encoder G_f and the domain discriminators G_d play a minimax game to remove the domain-specific characteristics from the input data to achieve domain-independent activity recognition. To do this, we first compute the integrated loss functions of the two components as:

$$\begin{aligned}\mathcal{L}_d^k(G_f, G_d) &= \mathbb{E}_{\mathbf{x},d}[-log\ G_d^k(\mathbf{z}, \hat{\mathbf{y}};\theta_d^k)] \\ &= -\frac{1}{|D|}\sum_{x_i \in D}\sum_{j=1}^{|D^k|} d_{ij}^k\ log(G_f(x_i;\theta_f), G_y(G_f(x_i;\theta_f);\theta_y);\theta_d^k)\end{aligned} \tag{8}$$

where $|D|$ is the number of samples belonging to the whole sample space, $|D^k|$ is the number of labels for class k and d_{ij}^k is the corresponding domain label. Integrating the loss of all the K discriminators together, we get the total loss for discriminators G_d as

$$\mathcal{L}_d(G_f, G_d) = \sum_{k=1}^{K}\mathcal{L}_d^k(G_f, G_d) \tag{9}$$

The final goal of our model is to minimize the label prediction loss $\mathcal{L}_y(G_f, G_y)$ and maximize the domain discrimination loss $\mathcal{L}_d(G_f, G_d)$, while these two objectives contradict with each other. To make these objectives consistent, we introduce the gradient reversal layer proposed in [6]. Based on Eq. 6, Eq. 8 and Eq. 9, we have the final joint loss function as follows:

$$\mathcal{L}(G_f, G_y, G_d) = \mathcal{L}_y(G_f, G_y) - \lambda\mathcal{L}_d(G_f, G_d) \tag{10}$$

where λ is a hyper-parameter to trade-off the two objectives in the final optimization. Our learning model tries to minimize the loss function so as to achieve domain independent in-car activity recognition.

5 EVALUATION AND DISCUSSION

In this section, we conduct real trace-driven experiments comparing our WiCAR system with state-of-the-art WiFi-based activity recognition systems with a discussion.

5.1 Implementation and Evaluation Setup

Prototype. We fully implement the WiCAR prototype using the commercial-off-the-shelf (COTS) devices without hardware or software change. Two Dell Latitude D820 laptops both equipped with an Intel 5300 WiFi card are used as the wireless sender and receiver. In our experiment, we use 2 antennas in the sender ($N_T = 2$) and 3 antennas in the receiver ($N_R = 3$) because we find that when using 3×3 MIMO mode the signal amplitude for each antenna pair obviously decreases and will further undermine the recognition accuracy. We select channels in the 5G frequency band rather than 2.4G since the shorter wavelength in 5G leads to higher resolutions for activity movement.

Data collection. In our experiment, we select a total of 8 common activities in cars for recognition, including pushing forward (PF), pushing right (PR), raise hand twice (RT), right shoulder check (RSC), left shoulder check (LSC), texting message (TM), answering cellphone (AC) and picking up things (PK). The different activities

(a) Pushing Forward (b) Pushing Right (c) Raise Hand Twice (d) Right Shoulder Check

(e) Left Shoulder Check (f) Texting Message (g) Answering Cellphone (h) Picking Up

Figure 6: Collecting activity samples in different driving conditions and different cars involving multiple volunteers.

and measurement settings are illustrated in Fig. 6. Each activity contains about 2500 samples collected in multiple classes, including 4 different driving conditions, 4 types of cars, and involving 4 volunteers varying in genders, heights and habits, which corresponds to 64 different domains in total. For each collection, the WiFi transceivers of our prototype are located at the same position, i.e., the co-pilot seat, to guarantee the measurement consistency. For in-car context, people are restricted by the safety belts at their own seats and will mostly perform activities at specific locations toward one same direction. Thus, the impact of activity location and orientation is quite marginal. As to the practical in-car deployment in the future, the WiFi transceivers can be integrated with the central control system and will not occupy extra space.

Learning setup. We implement the WiCAR learning model using tensorflow [1] and train the learning model based on the collected dataset on a desktop equipped with GTX 1080 Ti GPU cards, dual Intel I7 3.6 GHz CPU cards and 32GB memory. The filters of convolutional layers are 5×5 and are applied at stride 1. And the filters of max pooling layers are 2×2 with the stride of 2. We set the default neuron numbers in the two fully connected layers in both activity predictor and domain discriminators as 150 and 80, respectively.

To verify the generality of our learning model, we test WiCAR using collected data in new target domains that are different from the source domains. In particular, we ask another 4 volunteers to perform activities in different driving conditions and different cars, and collect about 2000 activity samples for testing.

Baseline methods. We compare our approach with several existing learning models, including Random Forest (RF), WiBot [18] and EI [10]. WiBot is a state-of-the-art activity recognition system that uses the traditional learning method without domain adaptation to recognize activities in cars. In its original design, WiBot did not consider the scenario of multiple driving conditions and car models. We train this model only considering the activity labels rather than the domain labels in our experiment. EI is a state-of-the-art learning framework for WiFi-based indoor activity recognition that considers single-adversarial domain adaptation. We incorporate its learning model and apply it in the in-car recognition. RF is one of the most used classification methods due to its simplicity even without hyper-parameter tuning. By constructing a multitude of decision trees and introducing the bagging method at the training time, RF is able to inhibit the overfitting effect compared with decision tree. We extract 13 features of both time dimension and frequency dimension for training RF. Besides the ten features as in EI, we also extract the maximum, minimum, and the variance of frequency from the spectrogram.

5.2 General Activity Recognition Performance

We first evaluate the general performance of our WiCAR system. Fig. 7 and Fig. 8 compares the detailed recognition situations on 8 target activities between WiCAR (with multi-adversarial domain adaptation) and the state-of-the-art in-car activity recognition system WiBot (without domain adaptation). Each cell represents the probability of recognizing an actual label as a predicted label. We can find that our WiCAR system achieves an average of 94.3% accuracy for all activities even under new domains, while the state-of-the-art solution only has an average accuracy of 58.2%. Specifically, the prediction result without domain adaptation seems to be disordered that one activity can be recognized as many other activities. For example, the texting message (TM) activity was recognized as all other 7 activities, some even with relatively high probabilities such as RT and AC. This result is because traditional learn model will inevitably contain much domain-specific information in the collected samples. Such extraneous information can blur the boundaries of different activities so that the recognition accuracy will drop dramatically. In contrast, the domain adaptation architecture

Actual Label	Predicted Label							
	PF	PR	RT	RSC	LSC	TM	AC	PK
PF	0.57	0.15	0.02	0	0.06	0.17	0.03	0
PR	0.13	0.66	0	0	0	0.12	0.09	0
RT	0.02	0.03	0.71	0	0	0.09	0.12	0.03
RSC	0.05	0.09	0	0.54	0.13	0.08	0	0.11
LSC	0.04	0.1	0	0.15	0.59	0.07	0	0.05
TM	0.02	0.05	0.14	0.08	0.05	0.51	0.12	0.03
AC	0.04	0	0.13	0.08	0.04	0.08	0.54	0.09
PK	0.03	0.03	0.07	0.19	0.07	0	0.1	0.51

Figure 7: The confusion matrix for recognition accuracy using the state-of-the-art approach without domain adaptation.

Actual Label	Predicted Label							
	PF	PR	RT	RSC	LSC	TM	AC	PK
PF	0.92	0.02	0	0	0	0.04	0.02	0
PR	0	0.95	0	0.02	0	0.01	0.02	0
RT	0	0	0.98	0.02	0	0	0	0
RSC	0	0	0	0.96	0.02	0	0	0.02
LSC	0.04	0	0	0.03	0.91	0	0.02	0
TM	0.01	0	0.02	0	0	0.92	0.05	0
AC	0	0.03	0	0	0	0.03	0.94	0
PK	0	0	0	0.02	0	0.01	0	0.97

Figure 8: The confusion matrix for recognition accuracy using our WiCAR approach.

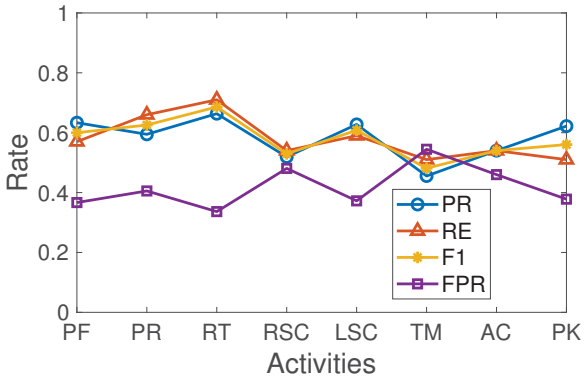


Figure 9: The statistic metrics of activity recognition when using the state-of-the-art approach without domain adaptation.

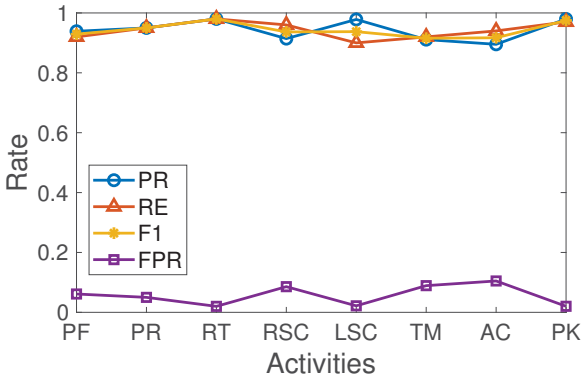


Figure 10: The statistic metrics of activity recognition when using WiCAR approach.

in our learning model helps the feature encoder to remove such domain-specific features and only retains the activity related features, which finally boost the recognition accuracy even for new domains.

To comprehensively evaluate the classification result, we also consider the following metrics used in statistics: 1) *False Positive Rate* (*FPR*) denotes the ratio of falsely labeled activities as another activity. 2) *Precision* (*PR*) is defined as $\frac{TP}{TP+FP}$, where TP is the ratio of correctly labeled activities, FP is the ratio of falsely labeled activities as another activity. 3) *Recall* (*RE*) is $\frac{TP}{TP+FN}$, where FN is the ratio of mislabeled true activities. 4) *F1-score* (*F1*) is a combined metric for precision and recall, defined as $\frac{2*PR*RE}{PR+RE}$. Fig. 9 and Fig. 10 illustrates these metrics for WiBot and WiCAR. We can find that WiBot can only achieve about 50% for precision and recall. This result indicates that with traditional learning methods, a large portion of in-car activity will be misclassified under new domains, which can hardly satisfy the requirement of safety-oriented activity detection or human-car interaction. In contrast, both the precision and recall of WiCAR achieve more than 90%, which is a 40% improvement compared with the traditional method.

Fig. 11 demonstrates the general recognition accuracy of our WiCAR approach and the baseline approaches when we use different training domain settings. The x-axis of $k_1 * k_2 * k_3$ means how many domains we use for the training, where k_i indicates the number of domains in the i-th class. We can observe that our WiCAR approach outperforms all other baseline approaches with multiple classes. In particular, RF and WiBot keep a relatively low accuracy under 50% since they never eliminate the impact of the extraneous domain-specific information in the collected data. The general accuracy of both the single-adversarial domain adaptation approach EI and the multi-adversarial domain adaptation approach WiCAR keep increasing as the number of training domains grows. When we use enough domains (e.g., $4 * 4 * 4$) for training, EI can only achieve an accuracy of about 83%, while WiCAR can achieve about 95% recognition accuracy, leading to a 12% improvement. This comparative experiment shows that our WiCAR approach is more capable of distinguishing the inherent structures of different domains and can remove such domain-specific impacts more effectively.

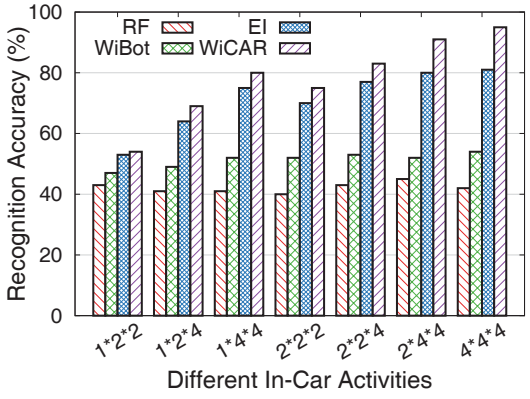


Figure 11: Comparison of different approaches for activity recognition accuracy when training model with different domain settings.

6 RELATED WORK

6.1 WiFi CSI-based Activity Recognition

WiFi CSI-based sensing technology has seen great success in recent years. The basic rationale is that human movements can affect the signals between WiFi antennas, and through profiling the CSI pattern changes we can retrieve the corresponding human movement. WiFi CSI-based sensing enables activity recognition in numerous application scenarios. Wang et al. [29] propose WiFall that uses the fine-grained CSI changes to detect only one falling activity. Wang et al. [28] establish the relationship between human movement velocity and CSI dynamics, and uses such quantitative relationship for indoor activity recognition. Wang et al. [25] develop WiHear, which is able to identify the subtle impact of different mouth shapes on the WiFi signals so as to recognize simple human pronunciation. Ali et al. [3] present WiKey, which is able to detect the different CSI changes when fingers are tapping on different keys. Virmani et al. [24] build up the correlations between CSI features and the location as well as orientation of target subject, and translates CSI measurements to the corresponding virtual samples for recognition.

In-car WiFi-based activity recognition emerges as a hot topic given its important role in autonomous driving and human-car interaction. The pioneer research WiBot [18] uses WiFi CSI to recognize simple gestures in dedicated conditions and car models. Different from previous researches, our WiCAR approach for the first time considers removing the extraneous information related to different driving conditions, human subjects and car models so as to achieves the environment/subject independent in-car activity recognition.

6.2 Domain Adversarial Learning

The learning model in WiCAR is related to the domain adversarial network [2]. Adversarial network has been widely used in the machine learning community toward many applications. The most representative model is generative adversarial network (GAN) [8], which utilizes a minimax two-player game to fool the discriminative model so that the generative model is able to generate high-quality data. Similar to GAN, the domain adversarial network also employs a domain discriminative model for adversarial learning. Yet the difference is that we encourage the model to learn a feature representation of the original input, which is discriminative for the main task in the source domain and invariant with respect to the shift between domains [6]. Ganin et al. [6] first propose the basic domain adversarial network and the corresponding backpropagation training methods. And many approaches [4] have also been proposed to improve the performance of the adversarial training.

Based on the theory of domain adversarial learning, pioneer researches have been proposed toward the WiFi-based activity recognition. Zhao et al. [35] develop a modified domain adversarial network-based model to remove the individual and condition-specific information during sleeping and utilize the extract features for accurate sleep stage prediction. Similarly, Jiang et al. [10] propose an adversarial network-based learning model to remove the environment and subject-specific information for indoor activity recognition. They both use single-domain adversarial architecture, that is, combining the multiple potential feature classes into one class. Different from the indoor recognition scenario, the in-car recognition scenario is much more complicated and subject to many impact factors, where the discriminative structures can be easily mixed up. In this paper, WiCAR incorporates multi-adversarial domain adaptation network and integrates them for in-car activity recognition, achieving a much higher recognition accuracy.

7 CONCLUSION

In this paper, we for the first time present a WiFi-based environment/subject independent in-car activity recognition framework named WiCAR. WiCAR employs a data preprocessing scheme to convert raw collected WiFi signals into effective feature representations for each activity. Leveraging a multi-adversarial domain adaptation network model, WiCAR is able to remove the domain-specific information in the received raw signals while retaining the activity related information so as to achieve environment/subject independent in-car activity recognition. Extensive evaluations further demonstrate the superiority of WiCAR compared to the state-of-the-art solutions.

ACKNOWLEDGMENTS

This work is supported by a Huawei-SFU Visual Computing Joint Lab Grant. The work of Wei Gong is supported by the Fundamental Research Funds for the Central Universities No. WK2150110013. The corresponding author is Jiangchuan Liu.

REFERENCES

[1] Martín Abadi et al. 2016. Tensorflow: a system for large-scale machine learning.. In *Proceeding of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, Vol. 16. 265–283.

[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 (MobiCom)*. ACM, 90–102.

[4] Martin Arjovsky and Léon Bottou. 2017. Towards principled methods for training generative adversarial networks. *arXiv preprint*

arXiv:1701.04862 (2017).

[5] Xiaoyi Fan, Wei Gong, and Jiangchuan Liu. 2018. TagFree Activity Identification with RFIDs. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 2, 1 (2018), 7.

[6] Yaroslav Ganin and Victor Lempitsky. 2014. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495* (2014).

[7] Jon Gjengset, Jie Xiong, Graeme McPhillips, and Kyle Jamieson. 2014. Phaser: Enabling phased array signal processing on commodity WiFi access points. In *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking (MobiCom)*. ACM, 153–164.

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

[9] Daniel Halperin, Wenjun Hu, Anmol Sheth, and David Wetherall. 2011. Predictable 802.11 packet delivery from wireless channel measurements. *ACM SIGCOMM Computer Communication Review* 41, 4 (2011), 159–170.

[10] Wenjun Jiang, Chenglin Miao, Feilong Ma, Shuochao Yao, et al. 2018. Towards Environment Independent Device Free Human Activity Recognition. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking (MobiCom)*. ACM.

[11] Oscar D Lara, Miguel A Labrador, et al. 2013. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys and Tutorials* 15, 3 (2013), 1192–1209.

[12] 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 (Ubicomp)*. ACM, 250–261.

[13] Zhaojian Li, Shan Bao, Ilya V Kolmanovsky, and Xiang Yin. 2018. Visual-manual distraction detection using driving performance indicators with naturalistic driving data. *IEEE Transactions on Intelligent Transportation Systems* 19, 8 (2018), 2528–2535.

[14] Tianchi Liu, Yan Yang, Guang-Bin Huang, Yong Kiang Yeo, and Zhiping Lin. 2016. Driver distraction detection using semi-supervised machine learning. *IEEE Transactions on Intelligent Transportation Systems* 17, 4 (2016), 1108–1120.

[15] Francisco Javier Ordóñez and Daniel Roggen. 2016. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors* 16, 1 (2016), 115.

[16] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. 2018. Multi-adversarial domain adaptation. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*.

[17] Kun Qian, Chenshu Wu, Zimu Zhou, Yue Zheng, Zheng Yang, and Yunhao Liu. 2017. Inferring motion direction using commodity wi-fi for interactive exergames. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 1961–1972.

[18] Muneeba Raja, Viviane Ghaderi, and Stephan Sigg. 2018. WiBot! In-Vehicle Behaviour and Gesture Recognition Using Wireless Network Edge. In *Proceeding of the 38th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 376–387.

[19] Jorge Santos, Natasha Merat, Sandra Mouta, Karel Brookhuis, and Dick De Waard. 2005. The interaction between driving and in-vehicle information systems: Comparison of results from laboratory, simulator and real-world studies. *Transportation Research Part F: Traffic Psychology and Behaviour* 8, 2 (2005), 135–146.

[20] Gordon L Stüber. 1996. *Principles of mobile communication*. Vol. 2. Springer.

[21] 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 (ICCV)*. 4068–4076.

[22] Deepak Vasisht, Swarun Kumar, and Dina Katabi. 2016. Decimeter-Level Localization with a Single WiFi Access Point.. In *Proceedings of the 13th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, Vol. 16. 165–178.

[23] 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 (MobiSys)*. ACM, 401–413.

[24] Aditya Virmani and Muhammad Shahzad. 2017. Position and Orientation Agnostic Gesture Recognition Using WiFi. In *Proceedings of the 15th International Conference on Mobile Systems, Applications, and Services (MobiSys)*. ACM, 252–264.

[25] 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 (TMC)* 15, 11 (2016), 2907–2920.

[26] Jue Wang, Deepak Vasisht, and Dina Katabi. 2014. RF-IDraw: virtual touch screen in the air using RF signals. In *ACM SIGCOMM Computer Communication Review*, Vol. 44. ACM, 235–246.

[27] 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 (TWC)* 17, 4 (2018), 2432–2442.

[28] 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 (MobiCom)*. ACM, 65–76.

[29] Yuxi Wang, Kaishun Wu, and Lionel M Ni. 2017. Wifall: Device-free fall detection by wireless networks. *IEEE Transactions on Mobile Computing (TMC)* 16, 2 (2017), 581–594.

[30] Jefferey L Wilson. 2016. Automotive WiFi availability in dynamic urban canyon environments. *Navigation: Journal of The Institute of Navigation* 63, 2 (2016), 161–172.

[31] Lu Xia and JK Aggarwal. 2013. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2834–2841.

[32] Tong Xin, Bin Guo, Zhu Wang, Pei Wang, Jacqueline Chi Kei Lam, Victor Li, and Zhiwen Yu. 2018. Freesense: a robust approach for indoor human detection using wi-fi signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 2, 3 (2018), 143.

[33] Chenyang Zhang and Yingli Tian. 2012. RGB-D camera-based daily living activity recognition. *Journal of Computer Vision and Image Processing* 2, 4 (2012), 12.

[34] 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 (CoNEXT)*. ACM, 83–96.

[35] 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 *Proceedings of International Conference on Machine Learning (ICML)*. 4100–4109.

[36] Rencheng Zheng, Kimihiko Nakano, Hiromitsu Ishiko, Kenji Hagita, Makoto Kihira, and Toshiya Yokozeki. 2016. Eye-Gaze Tracking Analysis of Driver Behavior While Interacting With Navigation Systems in an Urban Area. *IEEE Transactions on Human-Machine Systems* 46, 4 (2016), 546–556.