Data Augmentation and Dense-LSTM for Human Activity Recognition using WiFi Signal

Jin Zhang, Fuxiang Wu, Bo Wei, Qieshi Zhang, Hui Huang, Syed W. Shah, Jun Cheng, *Member, IEEE*

***Abstract*—Recent research have devoted significant efforts on the utilization of WiFi signals to recognize various human activities. An individual's limb motions in the WiFi coverage area could interfere wireless signal propagation, that manifested as unique patterns for activities recognition. Existing approaches though yielding reasonable performance in certain cases, are ignorant of two major challenges. The performed activities of the individual normally have inconsistent speed in different situations and time. Besides that the wireless signal reflected by human bodies normally carry substantial information that is specific to that subject. The activity recognition model trained on a certain individual may not work well when being applied to predict another individual's activities. Since only recording activities of limited subjects in certain speed and scale, recent works commonly have moderate amount of activity data for training the recognition model. The small-size data could often incur the overfitting issue that negative affect the traditional classification model. To address these challenges, we propose a WiFi based human activity recognition system that synthesize variant activities data through 8 CSI transformation methods to mitigate the impact of activity inconsistency and subject-specific issues, and also design a novel deep learning model that cater to the small-size WiFi activity data. We conduct extensive experiments and show synthetic data improve performance by up to 34.6% and our system achieves around 90% of accuracy with well robustness in adapting to small-size CSI data.**

***Keywords*-WiFi, channel state information, human activity recognition, data augmentation, neural network**

I. Introduction

Human activity recognition (HAR) that aims to identify the actions of subjects is of great importance for a wide range of real-world applications, such as smart home, health-care services, and fitness tracking, etc. With recent intelligent products like Google Nest and Xbox Kinect Sensor, the video cameras [1] [2] [3] are used to analyze human activities so that the smart home can capture the households' actions and provide the personalized connections to physical objects for controlling purposes. In assisted living places, the wearable sensors [4] [5] [6] [7] are used to continuously track activities of elderlies in case of any emergencies happening. Camera, phone, and wearable are widely used in these applications to recognize activities. However, the device-based approaches have many limitations due to security concerns raised by cameras and extra burdens from wearables that cumbersome to carry for a long time. To address these challenges, the recent proposed radio-based approaches demonstrate its ability for human sensing and intrinsic merits, which exploit existing WiFi infrastructure and are less intrusive.

WiFi-integrated devices are ubiquitous in urban indoor areas of modern societies. By closely examining the perturbations of WiFi radio spectrum, it is feasible to recognize various human activities. Prior works [8] [9] [10] [11] are designed to recognize basic human movement such as sitting, walking, and running, etc. Recent works rely on the fine-tuned signal processing approaches in achieving subtle motion recognition, e.g. finger gestures [12], hand gestures [13] [14], typing on a keyboard [15] and even breath monitoring [16] [17]. Prior works [18] [19] [20] has proven that individuals' gait (i.e. walking style) could be recognized by calculating multiple hand-crafted features.

Various WiFi based HAR approaches and systems have been developed. However, a major challenge has not been addressed. That is, the perturbations of WiFi signals are subject to change upon the diversified motion speed and body types of individuals who conduct activities. Human motions are composed of multiple limb (i.e. arms and legs) movements. Their moving speed and scale of one same activity naturally change when an individual in different situations and time. Moreover, the characteristics of human body shape, height, and frame for each individual are unique. As such the human activity patterns are prone to be varied for different individuals. As a result, an activity recognition model that is trained on a specific subject in certain situation could not work well on another subject at different time and situations. One simple solution could be the traversal method that recording activities in numerous of moving speed for each individuals to be monitored. However, it is tedious and painful for smart home settings and healthcare services.

Under such circumstance, recent works mostly record activities of limited amount of subjects and only consider the certain speed and scale of the activities. The scale of activities dataset, therefore, is commonly small in contrast with the traditional computer vision based activities dataset. For instance, prior works [21] [7] involve 8 groups of daily human activities such as walking, sitting etc. and contain less than few hundreds of

Manuscript received May 09, 2020; revised June 22, 2020, August 21, 2020; accepted September 22, 2020.

Jin Zhang, Fuxiang Wu, Qieshi Zhang, Hui Huang, Jun Cheng are with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China. Email: {jin.zhang, fx.wu1, qs.zhang, jun.cheng}@siat.ac.cn Bo Wei is with Computer and Information Sciences, Northumbria University, UK. Email: bo.wei@northumbria.ac.uk Hui Huang is with Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg. Email:hui.huang@uni.lu Syed W. Shah is with Computer Science and Engineering, The University of New South Wales, Australia. Email: z5038389@zmail.unsw.edu.au Corresponding author: Jun Cheng Email: jun.cheng@siat.ac.cn

This work was supported in part by National Natural Science Foundation of China (61772508, U1713213, U1913202, U1813205), Shenzhen Technology Project (JCYJ20170413152535587, JCYJ20180507182610734, JCYJ20180302145648171), CAS Key Technology Talent Program

Fig. 1: Operational Scenario

data samples, whereas, ImageNet [22] often used as benchmarks contains millions of images and thousands of synsets. The dataset scarcity could lead to the overfitting problem [23] [24] that the activity recognition model corresponds too closely to a particular limited set of data and fail to fit the future unseen data. Thus, it is challenging for traditional machine learning algorithms and deep learning techniques to effectively characterize the inherent patterns of activities while preventing the occurrence of the overfitting problem.

To address these problems, we are the first to propose a WiFi based device free activity recognition system that synthesizes diversified activity data to mitigate the impact of varied motion speed and subject specific information, and incorporates a novel deep learning model that optimized for the small-size dataset. Fig. 1 demonstrates our system's operation scenario. We employ three commercial off-the-shelf WiFi devices. One transmitter node (mini-PC with built-in WiFi card) continuously broadcasts packets to two receiver nodes which simultaneously record Channel State Information (CSI) from the received WiFi packets as shown in Fig. 1. The CSI is able to capture the aggregate impact of multi-path, shadowing, constructive and destructive interference brought from the human body upon the WiFi signals. Two receivers obtain CSI of WiFi signals from two independent radio paths and form different sensing angles. When a person performs certain activities, their limb motions influence the radio transmission paths of each receiver. It is expected that these impacts are in turn manifested as similar but unique perturbations which could be used for recognizing activities.

There are several challenges in realizing the system. The CSI data are usually interfered due to the imperfection of hardware and sensitiveness with surroundings, such as external moving objects and radio environment changes. We adopt Principal Component Analysis (PCA) to filter the noisy CSI time series in each subcarrier. The spectral and tempo patterns carried in CSI streams mostly represent the characteristics of varied activities. Thus, we employ Short-Time Fourier Transform (STFT) to convert each CSI streams into spectrogram as the activities' CSI dataset. The generated spectrograms of two receivers are combined and used for extracting the unique patterns of each activity without hand-crafted feature calculation.

The next challenges is, the diversified activities in speed and scale and subject-specific variants cause significant inconsistency in activities' CSI dataset which negatively affect the recognition accuracy when applied in realistic environment. To alleviate these problems, we artificially generate additional synthetic CSI samples by transforming the training activities data through multiple forms of transformations. The idea behind data synthesis is to increase and extend the activity pattern variations through transformations and encourage the activity recognition model to become invariant to these transformations, and eventually alleviate the influence of diversified motion speed and scale and subject-specific issue. Specifically, we design 3 groups of CSI data synthesis approaches, i.e. data-independent, deformation, and task-specific methods, 8 types of transformation methods in total. The data-independent methods contains dropout, Gaussian noise; the deformation methods includes time stretching, spectrum shifting, spectrum scaling, frequency filtering; the task-specific methods are the mixture of multiple activities samples, and changing principle components coefficients of activities samples. The synthetic activities dataset combined with the originals referred to as the augmented training dataset are used for the activity recognition model.

The third challenge is to exploit the previous augmented CSI dataset and differentiate activities' types. The small-size CSI activities dataset could often cause overfitting issue for the traditional recognition model. Herein, we propose a novel Deep Neural Network (DNN) model, Dense-LSTM that optimized for the small-size WiFi CSI data. The proposed Dense-LSTM model is composed of the composite dense neural network (Dense) and the bi-directional Long Short-Term Memory (LSTM) neural network. The dense network that characterized by concatenating and reusing features is able to keep the model in a compact manner and alleviate the overfitting issue in small-size data scenarios. The compact dense network is to extract spatial features and the LSTM a variant of Recurrent Neural Network (RNN) is to extract the temporal features of activities in CSI data.

The fourth challenge is to identify the model's appropriate hyperparameters for the synthetic dataset. The architecture and hyperparameters of the DNN model account for the model's representation capacities which could be adjusted for the extended synthetic CSI dataset to further improve performance and prevent the occurrence of the overfitting issue. Prior works heavily rely on the trial-and-error method to handcraft the DNN model structure for WiFi CSI-based recognition tasks. Instead, we employ the Bayesian optimization and Hyperband (BOHB) method to automatically search the optimized hyperparameters of the model when applying on the synthetic CSI dataset.

In summary, we made the following contributions:

- Design and implement the data augmentation approach that systematically transform and synthesize CSI data to alleviate the influence of motion inconsistency and subject-specific issue. To the best of our knowledge, our system is the first in the literature that systematically bring data augmentation in WiFi CSI sensing area.
- Design a novel deep learning model Dense-LSTM that optimized for the small-size CSI dataset and avoid the overfitting issue. To the best of our knowledge, the Dense-LSTM model is the first in the literature that optimized for the small-size WiFi CSI dataset. We adopt a hyperparameter optimization approach to automatically search the optimized configurations of the model for the synthetic CSI data.

- Implement the system using off-the-shelf WiFi devices and collect 10 types of activities from 5 volunteers. The comprehensive experiments show the data augmentation approach significantly improve recognition accuracy by up to 34.6% and the Dense-LSTM model increase the accuracy by up to 21.2% compared with the traditional DNN model. Our system achieves a stable recognition accuracy around 90.0% in small-size data scenarios.

The rest of the article is organized as follows. Section II discusses related works. Section III explains the whole architecture of our proposed system. Section III-A introduces WiFi CSI data and our designed data pre-processing and augmentation methods. Section III-B shows the details of the proposed deep learning model. Section III-C presents the hyperparameter optimization approach. Section IV explains the implementation and comprehensively evaluate the system. Section V concludes the paper.

II. Related Work

A. Data Augmentation

Data augmentation using label-preserving transformation has been widely used in computer vision and image recognition tasks and speech analysis. The purposes of applying data augmentation is mainly to address the data sparsity issue which often limit the performance of neural network classifier. The image translation, deformation and reflection [25], [26], [27] have led to significant improvements in image recognition performance. Artificially adding noisy data with various types to original speech data [28], [29] is a commonly used strategy for noise robust speech recognition. [30], [31] demonstrate the audio transformations including pitch shifting and time stretching are beneficial for speed recognition. As for WiFi CSI analysis, one work [32] reports the resampling (equiv. time stretching) of original data and adding noise improve the person detection accuracy. Our proposed method in this paper bring the data augmentation concepts originated from the image and speech recognition areas into WiFi CSI sensing area, and is a natural but significant evolution of previous research works.

B. Deep Learning

The WiFi CSI activities dataset normally is characterized with the limited size due to the cumbersome data collection process. Prior research [33] reveals that the deep learning model lead towards a steady decline in accuracy as the decrease of dataset size. It can be explained by the improper configurations of model as a result of overfitting on a small dataset size. Therefore, the dedicated deep learning models have been carefully designed for activities recognition to prevent the occurrence of overfitting issue when applying on the small scale of CSI dataset. [34] exploits 3 stacked convolutional neural network (CNN) layers to recognize 6 kinds of daily activities. [21] leverages attention-based bi-directional long short-term memory (Bi-LSTM) network which contains 2-layers of LSTM to recognize 6 kinds of basic human movements, e.g. walk, run, sit, etc. [35] incorporates 1-layer CNN and 2-layers LSTM and [36] chooses 4-layers LSTM to recognize 8 types of activities. [37] uses 3-layers CNN to classify 8 basic activities. [38] uses 6-layers CNN to discriminate 4 activities. [39] uses 9-layers CNN for sign language recognition and the dataset contains 276 sign words. It is obvious that researchers use additional stacked neural networks to fit the increased complexity of dataset, since the coordinated scaling of data set and model size predictably improve the model's performance [40]. The relationship between dataset size and model size exists across a range of model configurations such as optimizers, regularizers (i.e. weight decay inherently reduce model capacity), etc. However, these specialized model could not be easily reused in other applications with different amount of dataset. Our proposed model Dense-LSTM in this paper is optimized for the small scale of CSI data and prevent from the existence of the overfitting issue. We further employ a hyperparameter optimization method to alleviate the effort of parameter tuning of the model. We believe our proposed system is an important progress to WiFi based activity recognition.

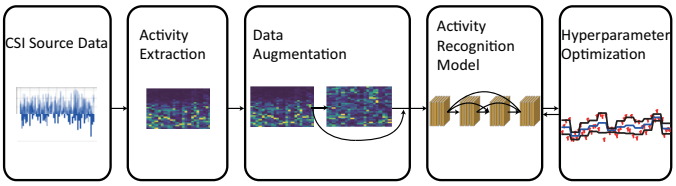


Fig. 2: The system overview.

III. System Design

We consider the typical operational scenario for the proposed system is depicted in Fig. 1. Our system consists of one transmitter which periodically broadcast packets which are simultaneously captured by two receivers which are capable of extracting the CSI as discussed in Section III-A1. Our system has two operational phases - training and testing. In each phase we ask the same set of individuals to conduct a set of activities listed in Table III for a short period of time. The CSI data collected during the training phase would be systematically transformed in a range of methods. The augmented dataset extend the training set and encourage the proposed deep-learning activity model to become invariant to these transformation brought from diversified activities speed and subjects. At the testing phase, the collected CSI data are kept in the original form which are then matched with those in the augmented training data to uniquely recognize the activity types.

Fig. 2 depicts the system's basic building blocks explained as the following.

- Section III-A1 introduces the WiFi CSI data. Additional effort is needed since the original CSI data contain numerous system noise and radio environment interference in surroundings.
- The PCA along with a band-pass filter is used to remove the dominant noise and extract the pronounced human movements. Then we apply STFT to transform CSI into spectrogram as the training data (see Section III-A2).

- We employ the combination of 3 groups and 8 types of data augmentation methods to synthesize samples together with the original as the augmented training dataset (see Section III-A3).
- We train a deep-learning based activity recognition model Dense-LSTM on the augmented dataset to determine the activities' types (see Section III-B).
- A hyperparameter optimization approach is to find the appropriate configurations of the model corresponding to the synthetic dataset to further improve performance (see Section III-C).

A. Data Pre-processing and Augmentation

1) Channel State Information (CSI): Most modern off-the-shelf WiFi NICs continuously monitor the frequency response of WiFi subcarriers as CSI [41]. Suppose X_i^p and Y_i^p denote the transmitted and received symbols in subcarrier i and antenna pair p, and H_i^p represent the CSI measured in frequency domain at the moment. Then,

$$Y_i^p = H_i^p \times X_i^p \;\; i \in [1, C] \;\; p \in [1, N_t \times N_r], \qquad (1)$$

N_t and N_r denote the number of transmit and receive antennas, respectively. Our hardware has 2 transmit and 3 receive antennas, i.e. $N_t = 2$, $N_r = 3$. We make use of the CSI amplitude $\|H_i^p\|$ simplified as H_i^p for signal stability and reliability. The CSI depicts power fading of channels thus can effectively reflect the human body movements. Our experiments use the Qualcomm Atheros AR9590 WiFi NICs and the customized driver [42] provides 114 (i.e. $C = 114$) OFDM subcarriers of 802.11n between each antenna pair; as a result, the dimension of the CSI time series is $114 \times N_t \times N_r$.

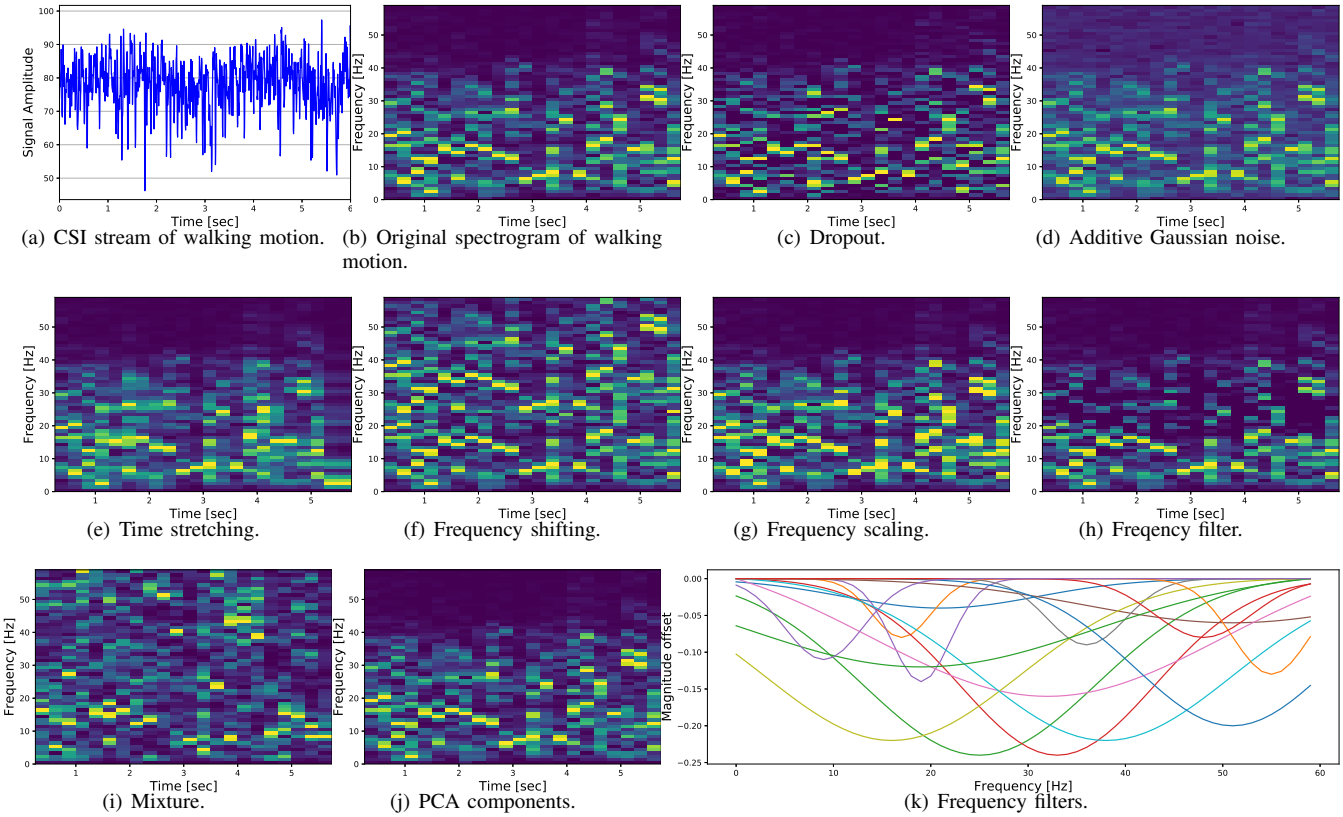
2) Data Pre-processing: WiFi NICs originally designed for communications have a set of internal state transitions including transmission rate adaptions and power variations which often incur significant noise on the collected CSI information. Since the CSI streams of neighboring subcarriers are highly correlated due to each subcarrier only differs slightly in their frequencies [43], we employ PCA [44] to emphasize variation and extract principal patterns in dataset and remove the ambient and internal noise. Recall that each transmit-receive antenna pair consists of 114 subcarriers mentioned in Section III-A1. We apply PCA on the CSI streams of every 114 subcarriers and select p subset of principal components. p is set arbitrarily as 3. The activities listed in Table III involve large body motions, therefore we use a Butterworth band-pass filter to extract signals between 0.2 Hz to 40 Hz from the first principal component. Then, we inverse transform the cleaned p components back to its original space to retrieve sanitized CSI data. An augmentation method explained in Section III-A3 is designed to change the importance weights of the components to synthesize transformed CSI data. Since the spectral-tempo patterns normally are representative of different activity types, we employ STFT on each CSI streams and the window size w of Discrete Fourier Transform (DFT) is set as 0.25 seconds. Let f denote the sampling rate and t represent the duration of CSI streams. The dimension of CSI spectrogram is to be $t/w \times f/2$. We stack the CSI spectrogram of overall subcarriers similar with a multi-layer lasagna resulting in the three dimension space $(114 \cdot N_t \cdot N_r) \times t/w \times f/2$. Next, the data augmentation approach tries to exploit this feature space for data synthesis.

TABLE I: Summary of Notations.

Notations	Meaning
η	dropout rate of additive random noise
σ	standard deviation of Gaussian noise
α	stretched duration in Time Stretching
β	shifted amount in Spectrum Shifting
γ	scaled amount in Spectrum Scaling
δ	scaling factor of filters in Frequency Filtering
μ	mean value of filters in Frequency Filtering
σ	deviation of filters in Frequency Filtering
ϵ_i	mixture ratio of samples
ζ_i	principle coefficients weights
c	number of stacked composite dense network layers
k	growth rate of dense network
d	dropout rate of dense network

3) Data Augmentation: We devise a range of augmentation methods that can be efficiently implemented to transform the CSI spectrograms. Of which two methods are data-independent that adding random noise, four methods are used to deform spectrograms, and two methods are specific to the activity recognition task. All of them can be easily applied in the training phase and generalized the training dataset to accommodate the negative impact of activity inconsistency and subject-specific problems. Each of augmentation method has one single parameter that defines the effect strength of the specific method. Fig. 3 demonstrates the influence of the data augmentation methods on CSI spectrogram. We take one CSI time series sample of walk activities of an individual as the example (see Fig. 3(a)), and show the corresponding spectrogram (see Fig. 3(b)), and illustrate the resulting modified CSI spectrogram with different amount of the effect strength of each augmentation method (see Figs. 3(c) to 3(j)). We employ the label-preserving transformation method, that the augmented training samples are to preserve their class labels (i.e. activity types). The transformation approaches are categorized in the following three groups, i.e. data-independent method, deformation method and task-specific method. Before explaining the details, for the sake of convenience, we summarize the notations in Table I, which will be used in the later data augmentation and activity recognition model descriptions.

Data-independent Methods: WiFi CSI dataset would still contain significant amount of noise even after the previous activity extraction approach mentioned in Section III-A2. An obvious and straightforward way to increase the robustness of activity recognition model is to proactively corrupt the training dataset with random noise. We consider the two simple methods, dropout - setting a part of the CSI spectrogram to zero with a given probability η and additive Gaussian noise - adding noise with the mean of zero and a given standard deviation σ. Such methods are independent of the kind of the input data and directly applied on the CSI spectrogram. Fig. 3(c) shows an example of spectrograms with $\eta = 3\%$



(a) CSI stream of walking motion. (b) Original spectrogram of walking motion. (c) Dropout. (d) Additive Gaussian noise.

(e) Time stretching. (f) Frequency shifting. (g) Frequency scaling. (h) Freqency filter.

(i) Mixture. (j) PCA components. (k) Frequency filters.

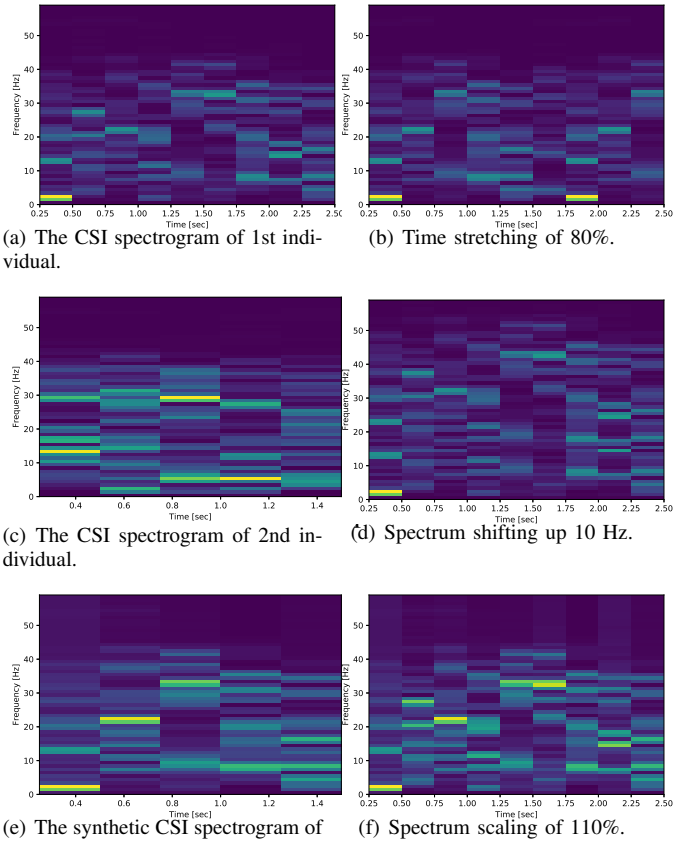
Fig. 3: Illustration of data augmentation methods on CSI spectrograms.

dropout and Fig. 3(d) shows an example of spectrogram that corrupted with the Gaussian noise of $\sigma = 0.2$.

Deformation Methods: The activities of monitored individuals are varied in speed and scale so as for the corresponding CSI spectrograms which could be misleading to the activity recognition model. We transform the CSI spectrograms by moderate amounts to mitigate the impact of the activities inconsistency but do not significantly change the overall patterns. We implement a set of deformation methods to modify the spectrograms as follows:

- Time Stretching: slow down or speed up the CSI streams (while spectrum unchanged). We exploit the linear interpolation to horizontally stretch spectrograms. The stretching rate is α percentage of the duration of the original CSI streams. The synthetic spectrogram are to be extended or cropped to keep their original duration. Fig. 3(e) shows the resulting spectrogram with $\alpha = 6\%$ of stretching.
- Spectrum Shifting: raise the spectrum of the CSI streams (while duration unchanged). Specifically, we vertically shift up the frequency band between 5 Hz and 40 Hz with a given amount and retain the bottom part unchanged. The shifting amount is β Hz. Fig. 3(f) shows a modified spectrogram with shifting up of 10 Hz.
- Spectrum Scaling: scale spectrograms (equal to add random offsets) in the previous frequency band by multiplying a factor, defined as (γ + 1). Fig. 3(g) demonstrates an amplified spectrogram by a certain amount ($\gamma + 1 = 102\%$).
- Frequency Filtering: the random frequency filters are applied to the CSI spectrograms as the fourth deformation method. The idea behind is to help the recognition model not focus on the specific frequency band but grasp the overall patterns of activities. Specifically, we design a filter response as a Gaussian function $f(x) = -\delta \cdot exp(-0.5 \cdot (x-\mu)^2/\sigma^2)$, with μ randomly chosen in the range between 5Hz and 60Hz, and σ randomly chosen between 1 and 20, δ randomly chosen in a given range. The width of the range of δ defines the filters' effect strength. Fig. 3(k) shows 15 of such designed filters as examples. Fig. 3(h) displays a filtered spectrogram.

Task-specific Methods: For the radio based HAR task, the monitored indoor environment mostly contain the rich ambient noise such as background radio noise, moving objects disturbance, etc. Such external interference could cause the inconsistency and performance degradation especially when an individual being monitored at different time period. The 1st task-specific data augmentation is we create additional training samples by mixing several samples together to alleviate the external interference's negative influence. Specifically, we consider the case of blending a given training sample A with two randomly chosen samples B and C, and the resulting sample D inherits the sample A's label. The mixture involve the linear addition of spectrograms as $D = (1-\epsilon_1) \cdot A + \epsilon_2 \cdot B + \epsilon_3 \cdot C$, with ϵ_i chosen uniformly at random in a given range after any other augmentation methods. We control the effect strength



(a) The CSI spectrogram of 1st individual.
(b) Time stretching of 80%.
(c) The CSI spectrogram of 2nd individual.
(d) Spectrum shifting up 10 Hz.
(e) The synthetic CSI spectrogram of 1st individual.
(f) Spectrum scaling of 110%.

Fig. 4: Illustration of data augmentation methods on CSI spectrograms.

via the width of range of ϵ_i to any given samples. Fig. 3(i) demonstrates a resulting modified spectrogram.

The 2nd task-specific data augmentation consists of altering the intensities of the principal components in each CSI dataset. Recall that in Section III-A2 we perform PCA on the CSI streams of every 114 subcarriers, then select and inverse transform 3 principal components for data sanitation. The selected principal components may contain different amount of noise and activities due to the changing environment. Thus to each training samples, we multiply the principal component coefficients by a random variable in a given range. An certain cleaned CSI samples H_x are constructed as following:

$$H_x = [\boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3][(\zeta_1 + 1)\lambda_1, (\zeta_2 + 1)\lambda_2, (\zeta_3 + 1)\lambda_3]^T, \quad (2)$$

where $\boldsymbol{P}_i$ and λ_i are ith principal components (i.e. eigenvector) and its coefficients (i.e. eigenvalues), respectively, and ζ_i is the aforementioned random variable. We control its effect strength via changing the width of range of ζ_i. Both of the previous methods contribute to identify the activity recognition model that invariant with the varied surroundings and individuals being monitored. Fig. 3(j) shows a resulting spectrogram.

During training, the above 8 factors for each CSI sample are chosen uniformly at random in a given range, such as changing in a range of 10, equiv. from 0% to 10% for η, σ, α, γ, δ, ϵ_i and ζ_i, 0 to 10 Hz for β. The width of the range defines the effect strength of particular transformation methods. We evaluate the impact of varied range width (from 5 to 26) of the 8 factors in Section IV. After a combination of the whole eight transformation methods, the modulated spectrograms combined with the originals become the augmented training dataset. Fig. 4 shows an simple example of data augmentation process. Two individuals, a girl and a boy, were asked to conduct one same activity (i.e. checking their wristwatch) for once, shown in Figs. 4(a) and 4(c), respectively. The boy performed faster than the girl at that moment. Thus, the spectrogram in Fig. 4(c) has short duration and strong spectrum across a wide frequency range, compared with the one in Fig. 4(a). The distinct difference demonstrate the activities inconsistency and subject-specific problem that influence the applications of activity recognition performance in realistic environment. The moderate amount of transformations are able to alleviate the negative impact. Figs. 4(b), 4(d) and 4(f) shows the resulting CSI spectrogram of 1st individual with different amounts of stretching, shifting and scaling. With the appropriate combinations, the synthetic spectrogram of 1st individual (i.e. Fig. 4(e)) contains much similarities with the spectrogram of 2nd individual. Therefore, as the above example illustrated, we believe through a set of CSI data synthesis the activities inconsistency and subject-specific variants can be mitigated to improve CSI based activities recognition performance.

B. Dense-LSTM Based Activity Recognition

In this section, we design a deep neural network model Dense-LSTM for WiFi-based activity recognition. In WiFi based sensing area, due to the cumbersome data collection process the size of the CSI dataset is normally quite small, especially compared with the traditional image analysis tasks. For instance, the well-known ImageNet [22] dataset often used as benchmarks contains over 21 thousand synsets and 14 million images in total, whereas the WiFi CSI dataset recently published normally involves less than 100 classes to be differentiated. The typical state-of-the-art image analysis approaches [45], [46] would leverage over hundreds or thousands of neural layers. A complex neural network with a large number of layers have better capabilities for classification, but are prone to suffer from the overfitting problems and performance depredation when applying on the limited amount of CSI dataset. Traditional methods such as weight decay, small batch size and learning rate might be insufficient to alleviate such issue. Therefore the prior WiFi CSI works mentioned in Section II have to employ a dedicated number of certain types of neural layers to ensure the performance. In this paper we propose a deep learning model Dense-LSTM that optimized for the WiFi CSI dataset of the small size scenarios.

The Dense-LSTM's architecture is presented in Fig. 5. Systematically, we take advantage of the design concept of the Dense neural Network (DenseNet) [47], which connects each layer to every other layers, and employ a dense connected convolutional network structure to extract the abstract features initially. In this structure every layer concatenate and reuse the features maps of all preceding layers. It could strengthen feature propagation and reduce the number of parameters and

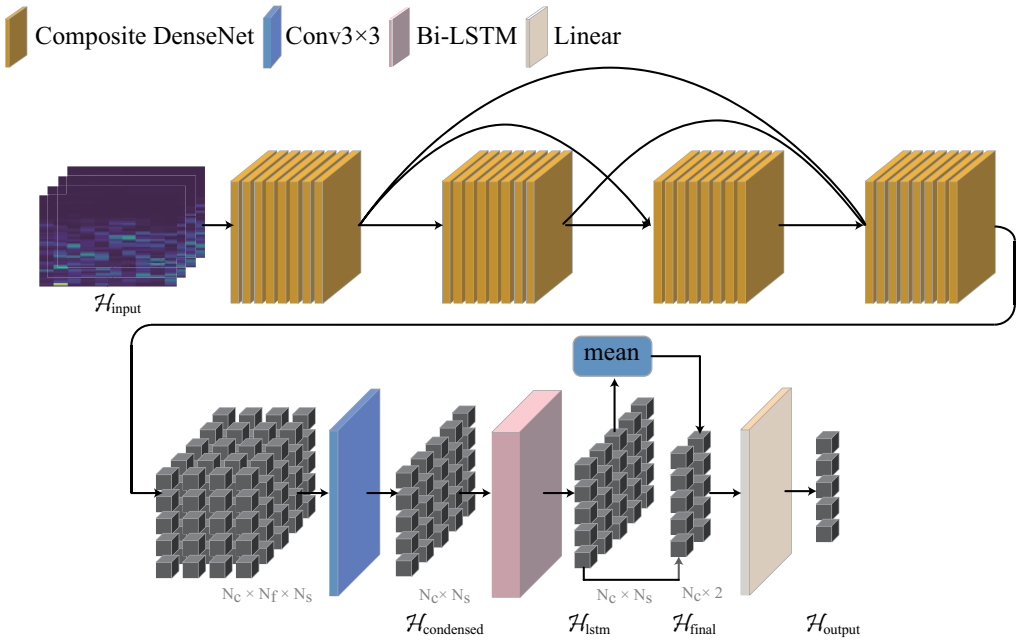


Fig. 5: The Dense-LSTM architecture is presented from left to right. The stacked CSI spectrogram denoted as $\mathcal{H}_{input}$ are fed into the network. Then, 4 composite DenseNet layers (i.e. yellow cuboids) are used to extract spatial patterns of activities. Next, the resulting intermediate features (i.e. the black cubs) is condensed by a convolutional layer (Conv 3×3) (see blue panel). An Bi-LSTM RNN (see red panel) is used to extract tempo patterns and a linear layer (see light yellow panel) predicts the activity types of $\mathcal{H}_{input}$.

alleviate the small data size problem. Thereafter, we use the bi-directional Long Short-Term Memory (Bi-LSTM) [48] to further learn the temporal features in the time series streams of WiFi CSI dataset. Thus, the Dense-LSTM could capture the spatial-tempo patterns in CSI data to uniquely identify the human activities. In details, the CSI spectrogram dataset indicated as $\mathcal{H}_{input}$ is processed by a series of neural models, i.e. varied dense network layers, a convolutional (Conv) network layer, a Bi-LSTM layer, and a linear layer in the end to predict the activities' types.

First, we explain the CSI data preparation and arrangement for the proposed DNN model. As shown in Fig. 1 the dual receivers form two independent sensing angles. The limb and torso of individuals may distinctively interfere the two propagation paths. We overlap and stack the CSI spectrogram of the dual receivers to make use of our deployment settings. Recall that the calculated CSI spectrogram space results in three dimensions discussed in Section III-A2. The stacked feature space along the 1st dimension is $2 \cdot (114 \cdot N_t \cdot N_r) \times t/w \times f/2$, simplified as $\mathcal{H}_{input} \in R^{N_c \times N_s \times N_f}$. The CSI spectrogram of overall subcarriers are combined which could be considered as images with multiple color channels. This interleaving feature arrangement is beneficial for the DNN model to capture the miniature differences between the dual receivers.

Then, we employ DenseNet to extract distinguishing spatial features from $\mathcal{H}_{input}$. Traditional methods such as CNN [49] tend to draw representational power from deep network architectures (i.e. multiple stacked layers) for better performance but are prone to be negatively affected by the overfitting problem. In contrast, DenseNets concatenate all forwarding layers' features and exploit network potential through reusing features, resulting in an efficient and light model. Specifically, the dense network comprises L layers, each of which implements a non-linear transformation $H_l(\cdot)$, where l indexes the network layer. $H_l(\cdot)$ denotes a composite function of operations which consists of Batch Normalization (BN), rectified linear units (ReLU), Convolution (Conv), and Dropout. The dropout layer at the end of the consecutive operations has been proved to be able to alleviate the overfitting problem [50]. The dropout rate denoted as d as a hyperparameter determines its effect strength. It has been noted that a normalized 1×1 convolution combined with a 3×3 convolution which referred to as the Bottleneck layer structure contribute to better computational efficiency. Thus we define $H_l(\cdot)$ as the composite function of several consecutive operations: BN-ReLU-Conv(1×1)-Dropout followed by BN-ReLU-Conv(3×3)-Dropout. The composite DenseNet is illustrated as the yellow cuboid comprised of 8 panels in Fig. 5. Similar with DenseNet we connect every layer $H_l(\cdot)$ to all subsequent layers. Consequently, one certain l^{th} layer receives the combined feature-maps of overall preceding layers as inputs,

$$x_l = H_l([x_0, x_1, \cdots, x_{l-1}]), \tag{3}$$

where x_l denotes the output of the l^{th} layer, and $[x_0, x_1, \cdots, x_{l-1}]$ refer to the concatenation of the feature-maps produced in overall preceding l^{th} layers.

The concatenation of each composite layers expand the size of produced feature maps. Assume each layer $H_l(\cdot)$ produces k feature-maps, then the l^{th} layer will have $k_0 + k \cdot (l-1)$ feature maps, k_0 stands for the number of channel of the input dataset $\mathcal{H}_{input}$. We empirically select the 3rd dimension N_f as the expanded channel resulting as N'_f. The hyperparameter

TABLE II: The Dense-LSTM feature size and architecture.

Layers	Feature Size	Dense-LSTM
DenseNet	$N_c \times N_s \times N_f$	$\begin{bmatrix} 1 \times 1\ H_l(\cdot) \\ 3 \times 3\ H_l(\cdot) \end{bmatrix} \times c$
Conv	$N_c \times N_s$	$3 \times 3\ Conv$
Bi-LSTM	256	128 hidden $\times 2$
Linear	10	

k referred to as Growth Rate that regulates the amount of information each layer contributes to the model. Additionally, we consider the number of the composite layers $H_l(\cdot)$ as one hyperparameter, denoted as c. Fig. 5 shows an example of 4 dense layers (i.e. $c = 4$). The depth and width and robustness of model depend upon c, k and d, respectively, that needed to be chosen properly to ensure performance which will be discussed in Section III-C. We randomly choose $c = 4$, $k = 11$, and $d = 0.3$ as default initial settings.

Next, a single convolutional layer (Conv 3×3) is to condense and further extract features on the N'_f dimension (the yellow panel in Fig. 5). The generated feature space of $\mathcal{H}_{condensed}$ is $N_c \times N_s$. Bi-LSTM [48] RNN is effective to learn long-term dependencies of feature space so as for the temporal characteristics of CSI time series. Thus, one Bi-LSTM RNN layer is applied on the $\mathcal{H}_{condensed}$ and the dropout ratio of LSTM is also set as the hyperparameter d. To summarize the generated $\mathcal{H}_{lstm}$, the concatenation of its first column $\mathcal{H}^0_{lstm}$ and the mean of $\mathcal{H}_{lstm}$ along its first dimension $\overline{\mathcal{H}_{lstm}}$ become the final representation:

$$\mathcal{H}_{final} = [\overline{\mathcal{H}_{lstm}}; \mathcal{H}^0_{lstm}]. \tag{4}$$

Finally, a linear layer projects features to the unnormalized probabilities $\mathcal{H}_{output}$ and a Softmax function calculates the probability vector to predict activity types. The Dense-LSTM configuration details are illustrated in Table II. The LSTM network is chosen to use 128 of hidden states, thus the resultant feature size of Bi-LSTM layer is 256. The linear layer has 10 hidden states corresponding to 10 types of activities to be classified. For training the model, the negative log-likelihood loss (equiv. cross entropy function) is adopted to optimize the loss between predictions and ground truths. We choose the Adam optimizer and set the weight decay rate as 0.0001. As the WiFi CSI dataset size is small, the learning rate and batch size is set relatively low, as 0.0008 and 10, respectively.

C. Hyperparameter Optimization

The proposed data augmentation approaches in Section III-A3 expand the existing training dataset by twofold. As the increase of training data size the representation capacity of the deep learning model need grow to fully take advantage of the augmented data and improve the identification performance. The architecture (i.e. the number of stacked dense network layers) and hyperparameter (i.e. growth rate and dropout rate) mostly determine the representation capacity and behavior of the deep learning model. As per this the proper settings need to be found to increase the Dense-LSTM's representation abilities, meanwhile, avoid the overfitting issue when applying on the expanded synthetic dataset.

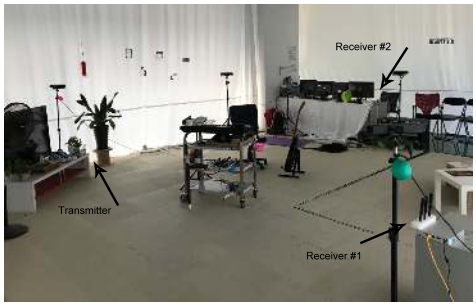


Fig. 6: A living room where experiments conducted.

We exploit the Bayesian optimization and Hyperband (BOHB) method [51] for the joint neural architecture and hyperparameter optimization. BOHB uses multivariate Kernel Density Estimators (KDEs) to select promising configurations and dynamically allocates more resources (i.e. the number of epochs during training model) for the selected configurations. Specifically, we set the optimization method to initially try the default configurations of model and early stop at 133th epoch. The top best-performing configurations continue to be evaluated with the full budget with the epochs of 400. The configuration to be optimized consists of the three hyperparameters mentioned in Section III-B, c decides the number of stacked DenseNet layers, k represents the growth rate and d determines the dropout rate, which are the key factors in determining the Dense-LSTM's architecture and behaviors. Every hyperparameters is presumably spread in the predefined range, i.e., $c \in [3, 10]$, $k \in [3, 20]$, and $d \in [1, 9] * 0.1$. The promising integers within the range would be evaluated and the configurations with best-performing accuracy is to be used as the optimized settings of the model. We apply the BOHB approach on the original and the augmented dataset to replace the tedious manual tuning of architectures and hyperparameters. The optimized configurations contribute to alleviate the overfitting problem and further improve the recognition performance upon the synthetic dataset. Section IV-C demonstrates the detailed process and the corresponding benefits of the hyperparameter optimization approach.

IV. Evaluation

In this section we present a comprehensive evaluation of the system. Section IV-A outlines the experimental setup. Section IV-B evaluates the effect of the augmentation methods in different settings. Section IV-C analyzes the hyperparameter optimization approach and demonstrates the optimized configurations of model and corresponding performance. Section IV-D shows the training and testing time costs. Section IV-E discusses the effect of the dropout rate. Section IV-F compares our system with traditional methods in various settings.

A. Experiment Setup

We implemented a prototype of the system using off-the-shelf PCs with built-in WiFi NICs. We specifically use three

TABLE III: The performed activities.

1	Make phone calls	6	Jumps
2	Check wristwatch	7	Lie down
3	Walk	8	Play guitar
4	Walk (fast)	9	Play piano
5	Run	10	Play basketball

identical mini-PCs equipped with Intel i7 CPU, 2G RAM, and Qualcomm Atheros AR9590 WiFi NIC. We installed Ubuntu 14.10 with modified Atheros NIC drivers in the mini-PCs. Three (i.e. $N_r = 3$) WiFi antennas are mounted on top of the mini-PC at receivers and two (i.e. $N_t = 2$) antennas at transmitter resulting in 114 (subcarriers) $\times 2 \times 3 = 684$ per packet time-series CSI data streaming. An enclosed furnished space is chosen to imitate a standard living room (30 m^2) wherein residents conduct various activities. Fig. 6 shows the furniture such as TV, plants, tables, etc. The transmitter was on the TV stand and receivers were 3.5 meters away on a cabinet and a table respectively. The recruited volunteers conduct activities within this area and may cut across or close to the direct Line-of-Sight (LoS) path between transceivers. We set the center frequency of WiFi channel as 5.23 GHz, bandwidth as 40 MHz. The transmitter broadcasts packets every 5ms (i.e. sampling rate is 200Hz). The CSI data extracted in MATLAB at receivers are processed by the proposed model written in python. The model is trained and tested in a NVIDIA server equipped with Tesla P40 GPU. The activities data collection was on the 7th floor of our building on campus where the coexisted WiFi network operated the entire time.

We recruited 5 college students who were both male and female and aged between 25 and 28 years. Each student was asked to perform 10 types of activities listed in Table III. To simulate the realistic scenarios, every subject conducts each of activities only once lasting for 10 seconds. The activities with inherently of finite duration would be repeated for a number of times. Such short duration of data collections are performed for once and feasible for smart home and healthcare applications. The recorded WiFi CSI from the dual WiFi receivers that simultaneously monitoring students are combined as one activity observation. The gathered CSI activities are separated into 5 segments, each with 2 seconds resulting in 25 observations of each activity types. We choose 12, 8 and 5 out of 25 observations of each activities that correspond to 120, 80 and 50 data samples as training, validation and test dataset, respectively. The augmented CSI dataset has twofold 240 training data samples. We employ true detection rate (accuracy) to evaluate our system performance.

B. Effect of Augmentation Methods

We train our deep-learning model Dense-LSTM with each of the eight different augmentation methods on the activities data and evaluate the effect of each augmentation method on the unmodified test dataset. Fig. 7 depicts the comparisons of augmentation methods with eight different changing range width of 5, 8, up to 26 in a step of 3. The parameters of the eight augmentation methods i.e. η, σ, α, β, γ, δ, ϵ_i, and ζ_i are to be randomly chosen in the selected changing range. We repeat experiments for evaluating every parameters in each range width for 200 times to enable a fair comparison of the augmentation methods. Fig. 7 displays the recognition accuracy of each augmentation method and shows the result of the base system with and without augmented dataset. Note we use the default settings of the hyperparameters (i.e. c, k, and d) of the deep-learning model as mentioned in Section III-B in this experiment.

In Fig. 7, we observe that dropping out a small amount of CSI signals in a range width of 11% increases the recognition accuracy by 6.3% on average compared to the base system without augmentation. Corrupting the CSI spectrogram with Gaussian noise does not have a strong effect. The CSI spectrogram contain much noise due to the imperfect hardware and disturbing radio environment. Thus the dropout possibly remove the influence of inconsistent external noise and help the recognition model focus on the innate patterns of human activities. Time stretching has strong effect on the recognition performance indicating the mitigation of tempo inconsistency in the training data, specially at the range width of 26% with the average 8.6% increase. It seems appropriately fill in the gaps of varied speed of motions which not covered in our small training dataset. Spectrum shifting has moderate improvement on performance at the range width of 5 Hz but diminish the accuracy at larger range width. Spectrum scaling have a modest improvement with the best setting at the range width of 26%. Random frequency filters have a similar effect with the best changing range width of 8%. The prior three methods modulate CSI spectrograms and effectively mitigates the spectral inconsistency in the training sets. Mixing with random samples improve the accuracy by 5.1% approximately with the best setting at the range width of 5%. Random coefficient weights of PCA also has a modest improvement at a range width of 17%. We believe that both methods help to form realistic samples and drive the network model to recognize activities irrespective of varied radio environment and individuals being monitored.

We further compare the recognition performance of the original data and the augmented data that combining dropout of 11%, time stretching of 26%, spectrum shifting of 5%, spectrum scaling of 26%, frequency filters of 8%, mixture of 5% and PCA of 17%, except from the Gaussian noise method. In Fig. 7, the bars from the right indicated as 'no augm.' denotes the recognition accuracy of 79.0% of the base system without data augmentation. In contrast, the combination of the selected augmentation methods (denoted as 'combined' in Fig. 7) achieve the recognition accuracy of 89.0% on average. While performance improvement do not add up linearly, we do observe the combined augmentation methods bring an additional 10.0% relative accuracy improvement comparing with original dataset without augmentation, and also outperform any single method on the mitigation to activities inconsistency. We keep using the previous selected best settings of the combined augmentation methods in the following experiments.

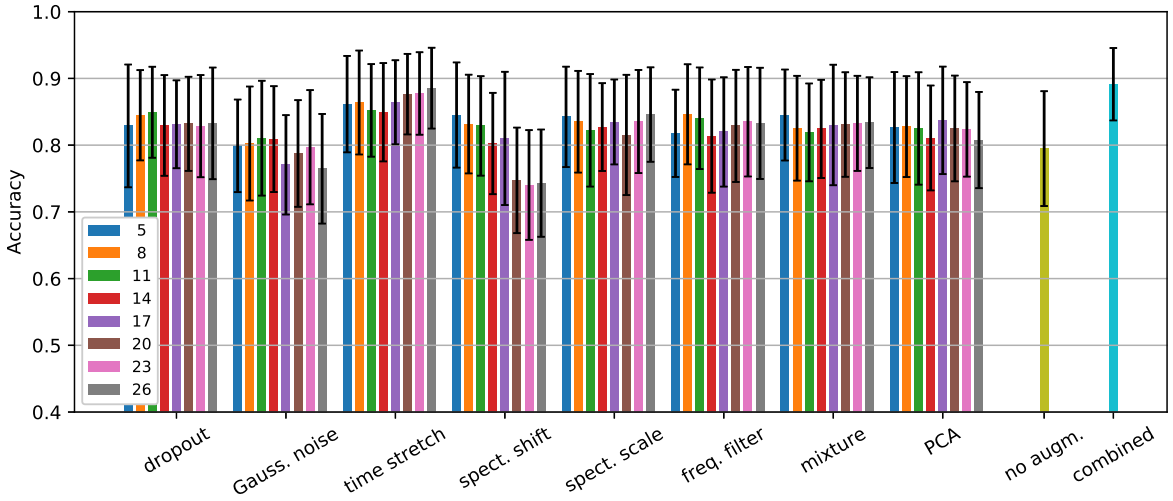
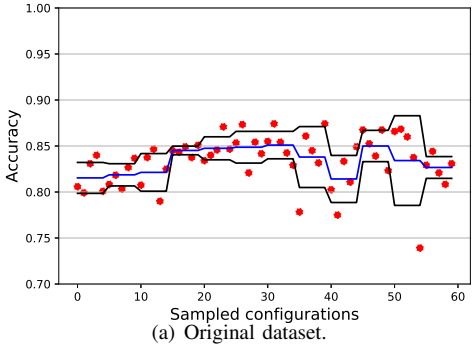
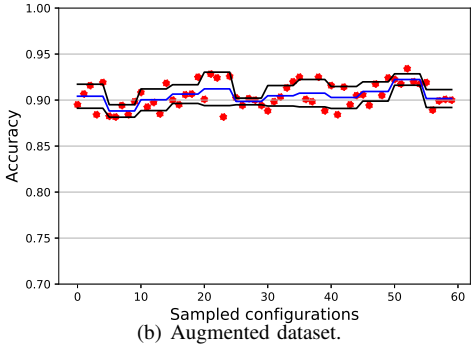


Fig. 7: Effect of augmentation methods in different changing range width from 5 to 26.



(a) Original dataset.



(b) Augmented dataset.

Fig. 8: Recognition accuracy of the sampled configurations evaluated during the whole optimization procedure. The red dots denote the recognition accuracy of specific configurations. The blue and black line represents the mean and standard deviation of accuracy, respectively.

C. Hyperparameter Optimization Analysis

We experiment the hyperparameter optimization approach to determine the network model with well-suited architecture and behavior for the original and the augmented training dataset. The appropriate network model should have better recognition accuracy and avoid the overfitting problem at the same time. We evaluate the proposed network model on the generated dataset and search for around 60 groups of the optimized hyperparamters. We repeat experiments of every configuration for 25 times to ensure the whole optimization process end in finite time period. In Fig. 8, the red dots represent the recognition accuracy of the network model with specific configurations, the blue and black lines represent the mean and standard deviation of accuracy of the optimized network models. We observe the performance of the sampled configurations tend to increase over time for both the original and augmented dataset. We also observe that most of configurations applied on the augmented dataset have relatively comparable results. It exhibits that the augmentation methods help network models with different structures to reliably identify activities. The configurations with best-performing accuracy are chosen as the optimized hyperparameters of the model which shown in Table IV. In details, the number of densenet layers as $c = 6$ and the growth rate as $k = 15$ fit the augmented dataset, whereas, both factors are set as 4 for the original dataset. The dropout rate is set as $d = 0.5$ for both dataset. It demonstrates the extra representation capabilities are preferred to fit the augmented dataset. In order to evaluate the benefits of the increased growth rate and densenet layers, we experiment the settings fitted with the original dataset on the augmented dataset, as shown in Fig. 9. It shows the data augmentation increase the average accuracy from 84.0% to 87.0%, and the optimized network model ($c = 6$ and $k = 15$) further improve the recognition accuracy by 3.4% from 87.0% to 90.4%. When comparing with the default settings depicted in Fig. 7, the optimized network model increase the accuracy by 5.0% for the original dataset and 1.4% for the augmented dataset. Therefore, the recognition performance improvements demonstrate the benefits of optimizing the hyperparameters of the proposed deep learning model. We keep using the optimized hyperparameter settings for the following experiments.

TABLE IV: The optimized network configurations.

	Densenet layers c	Growth rate k	Dropout rate d
Original	4	4	0.5
Augmented	6	15	0.5

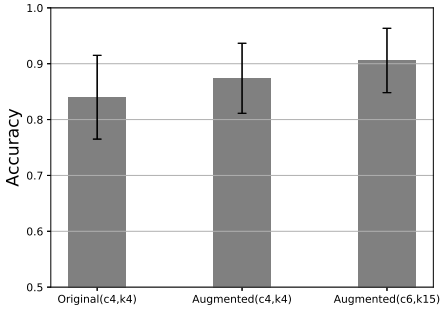


Fig. 9: Comparing the performance of the original and the augmented CSI data.

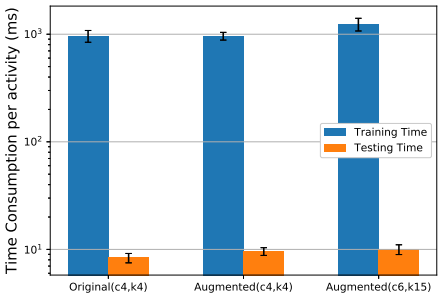


Fig. 10: Comparing the training and testing time of two configurations applied on the original and augmented CSI data.

D. Time Consumption of Training and Testing

Fig. 10 shows the time consumption of training and testing of different configurations of the activity model applied on the original and augmented dataset. We observe that the training and testing time of the configuration ($c = 4, k = 4$) for the original dataset is 962.5ms and 8.3ms for each activity samples, respectively. The augmented dataset with the same configuration takes similar amount of training time but cost a bit more testing time, 9.6ms on average. In contrast, the configuration ($c = 6, k = 15$) fitted with augmented data takes 1238.75ms and 10ms to finish training and testing, respectively. The extra composite densenet layers and growth rate increase the training time, however training usually is performed offline once. Nonetheless, the raised testing time is 2ms and 0.4ms compared to the prior settings which is negligible. Therefore, it is beneficial to employ the optimized model and practical to be implemented in real-time scenarios.

E. Effect of Dropout Rate

As discussed in Section III-B, the composite dense network and Bi-LSTM network incorporate the dropout method to alleviate the overfitting issue. Herein, we evaluate the effect of different dropout rate d from 0.1 to 0.9 as shown in Fig. 11. We find the dropout rates of 0.5 is the best-performing setting for both original and augmented data. The extreme values may bring unnecessary noise and diminish the deep-learning model's accuracy.

F. Comparing with Traditional Methods

In this subsection, we firstly experiment the effectiveness of the data augmentation approaches on the traditional

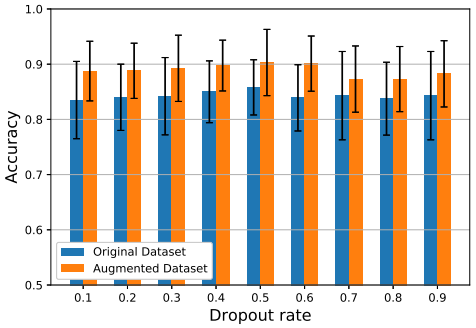


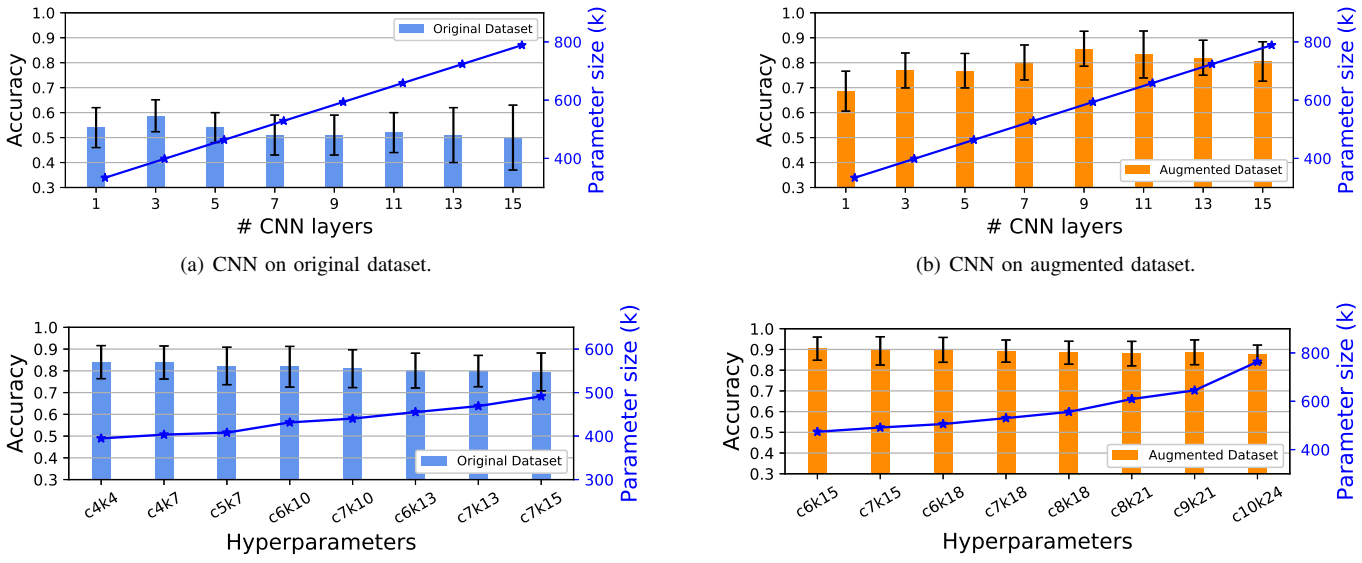
Fig. 11: Effect of dropout rate.

TABLE V: The performance of SAC, SVM, CNN, LSTM, CNN-LSTM on the original and augmented CSI dataset.

	Original		Augmented	
	mean	deviation	mean	deviation
SAC [18]	0.150	0.047	0.300	0.085
SVM	0.360	0.070	0.502	0.085
CNN-3 [37]	0.588	0.064	0.769	0.070
LSTM-4 [36]	0.686	0.065	0.726	0.072
CNN-1+LSTM-2 [35]	0.658	0.043	0.778	0.064

recognition methods. We evaluate the performance of two standard classifiers, i.e. Sparse Approximation based Classification (SAC), Support Vector Machine (SVM), and three deep learning methods, i.e. CNN, LSTM, CNN-LSTM on the original and augmented dataset shown in Table V. Prior works [9][18] use SAC for its merits that resilient with noise. In Table V, we find the augmented dataset increase the average accuracy by 15% and 14.2% for SAC and SVM, respectively. The deep learning based approaches, 3-layers CNN [37], 4-layers LSTM [36], 1-layer CNN 2-layers LSTM [35] denoted as CNN-3, LSTM-4, CNN-1+LSTM-2 in Table V have better accuracy and the augmented dataset consistently help improve performance. However, these traditional methods still have unsatisfactory performance since the standard classifiers lack ability to extract representative features from raw CSI data and the traditional DNN models have ill-suited representation capacity for small-size CSI dataset.

Next, we further compare the proposed deep learning model Dense-LSTM with the traditional convolutional neural network. As mentioned in Section II, SignFi [39] makes use of 9-layers CNN for sign language recognition and involves 276 types of sign words to be classified. The architecture of the designed CNN model is a typical deep learning model to extract representative feature in CSI data and has been widely adopted in other WiFi based activity recognition works [37], [35], [52], [53], [54]. Thus, we implement the CNN model in [39] as benchmark for comparison. We keep the existing settings such as kernel size and dropout rate, and only vary the number of CNN layers to evaluate the performance of CNN on the original and augmented CSI dataset, as shown in Fig. 12(a) and Fig. 12(b). It is obvious the 3-layers CNN performs best on the original dataset but only achieves the average accuracy of 58.8% (see Fig. 12(a)). The reason behind is that the stacked CNN model increase the usage of parameters and representa-



(a) CNN on original dataset.

(b) CNN on augmented dataset.

(c) Dense-LSTM on original dataset.

(d) Dense-LSTM on augmented dataset.

Fig. 12: The performance of the CNN network [39] and our system. The synthetic data increases recognition accuracy of the 9-layers CNN by 34.6%. The Dense-LSTM model increases the accuracy by 21.2% compared with the 3-layers CNN on the original dataset and achieves around 90% of accuracy with well robustness on the augmented dataset.

tion capacity whereas the consequent overfitting issue make performance even worse. In contrast, the 9-layers CNN on the augmented dataset achieves an average accuracy of 85.6% (see Fig. 12(b)). Herein, the synthetic data significantly improve the recognition accuracy of the 9-layers CNN model by 34.6%, increasing from 51.0% to 85.6%. However, the performance of CNN is mostly unstable, for instance the accuracy of 1-layer and 15-layers CNN decrease by 15.6% and 5.1% respectively. The deficiency and overflow of representation capacity both negatively affect the performance, therefore, the CNN model would require dedicated configurations for specific small-size dataset to ensure accurate result. As such it demands much effort from practitioners to optimize the CNN model for every specific recognition tasks with varied small size of data.

As discussed above the traditional methods that in need of dedicated configurations is ill-suited for the WiFi based recognition tasks with small-size data. Next, we demonstrate the Dense-LSTM model is optimized for the applications that characterized by the small-size data, through following experiments. We change the values of densenet layers c and growth rate k to evaluate the performance of Dense-LSTM on the original and augmented data, as shown in Fig. 12(c) and Fig. 12(d). It is evident that the Dense-LSTM model with varied c and k on the original dataset achieves over 80% of accuracy (see Fig. 12(c)) which increase the performance by 21.2% compared to the 3-layers CNN model (see Fig. 12(a)). The stable recognition result shows the Dense-LSTM model performs well on the small-size dataset. Moreover, the Dense-LSTM model on the augmented dataset achieves around 90% of accuracy in all settings shown in Fig. 12(d). The model performs stable with low deviation error, and the synthetic data help increase the average accuracy by around 10%. As per the original and augmented data scenarios in Fig. 12(c) and Fig. 12(d), the Dense-LSTM model in varied configurations and parameter size has consistent performance and shows well robustness against the overfitting issue. The Dense-LSTM model concatenates and reuses features which keep the model in a compact manner to prevent from the overfitting issue and consistently and accurately recognize activities' types. Combined with the data augmentation and hyperparameter optimization, we believe the proposed system is a generalized solution for WiFi based recognition tasks in small-size data scenarios. This is an important step towards realizing robust WiFi-based human sensing in natural environment.

V. Conclusion and Future Work

In this paper we present a WiFi based activity recognition system that synthesize diversified activities data to eliminate the influence of activity inconsistency and subject-specific issues, and propose a novel deep learning model Dense-LSTM that optimized for the small-size WiFi CSI dataset. Human limb motions naturally change in speed and scale in different situations and time. The characteristics of human body are also unique for different persons. The system exploits 3 groups of data synthesis approaches and 8 types of transformation methods, including dropout, Gaussian noise, time stretching, spectrum shifting, spectrum scaling, frequency filtering, sample mixture and principle component coefficient changes to synthesize activities data. The proposed Dense-LSTM model that concatenate and reuse the feature maps which keep the model in a compact manner to avoid the overfitting issue. The comprehensive experiments show the synthetic data improve performance by up to 34.6%. The Dense-LSTM model achieves consistent 90% of accuracy in varied configurations without performance degradation compared to traditional methods. Combined with the data augmentation and

hyperparameter optimization, we envision the proposed system is to be a generalized solution for WiFi based recognition applications, especially in small-size data scenarios.

While our work is effective in addressing the activity inconsistency and subject-specific influence, we also acknowledge several challenges that needed to be solved towards real world implementations. In the future we plan to consider some more configurations of sensing environment to be monitored including variant locations with different ambient radio environment, transceiver and individuals' relative coordinates, etc. Prior works rely on training the transfer learning and generative adversarial learning models on dataset collected from multiple environments. On the contrary, we plan to estimate the variability of environment and synthesize dataset corresponding specific system location and geometry to mitigate the impact of different or even unseen new environment.

REFERENCES

[1] R. Bodor, B. Jackson, and N. Papanikolopoulos, "Vision-based human tracking and activity recognition," in *Proc. of the 11th Mediterranean Conf. on Control and Automation*, vol. 1, 2003.

[2] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010.

[3] C. Wang *et al.*, "Human identification using temporal information preserving gait template," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2164–2176, 2012.

[4] A. M. Khan, A. Tufail, A. M. Khattak, and T. H. Laine, "Activity recognition on smartphones via sensor-fusion and kda-based svms," *International Journal of Distributed Sensor Networks*, vol. 10, no. 5, p. 503291, 2014.

[5] M. Field, D. Stirling, Z. Pan, M. Ros, and F. Naghdy, "Recognizing human motions through mixture modeling of inertial data," *Pattern Recognition*, vol. 48, no. 8, pp. 2394–2406, 2015.

[6] F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, and Y. Amirat, "Physical human activity recognition using wearable sensors," *Sensors*, vol. 15, no. 12, pp. 31 314–31 338, 2015.

[7] W. Jiang, Q. Li, L. Su, C. Miao, Q. Gu, and W. Xu, "Towards personalized learning in mobile sensing systems," in *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2018, pp. 321–333.

[8] Y. Wang, J. Liu, Y. Chen, M. Gruteser, J. Yang, and H. Liu, "E-eyes: device-free location-oriented activity identification using fine-grained wifi signatures," in *Proceedings of the 20th International Conference on Mobile computing and networking*. ACM, 2014, pp. 617–628.

[9] B. Wei *et al.*, "Radio-based device-free activity recognition with radio frequency interference," in *International Conference on Information Processing in Sensor Networks*. ACM, 2015, pp. 154–165.

[10] Y. Zeng *et al.*, "Analyzing shopper's behavior through wifi signals," in *Proceedings of the 2nd workshop on Workshop on Physical Analytics*. ACM, 2015, pp. 13–18.

[11] B. Wei, W. Hu, M. Yang, and C. T. Chou, "From real to complex: enhancing radio-based activity recognition using complex-valued csi," *ACM Transactions on Sensor Networks (TOSN)*, vol. 15, no. 3, p. 35, 2019.

[12] S. Tan *et al.*, "Wifinger: leveraging commodity wifi for fine-grained finger gesture recognition," in *Proceedings of the 17th International Symposium on Mobile Ad Hoc Networking and Computing*. ACM, 2016, pp. 201–210.

[13] H. Abdelnasser, M. Youssef, and K. A. Harras, "Wigest: A ubiquitous wifi-based gesture recognition system," in *International Conference on Computer Communications (INFOCOM)*. IEEE, 2015, pp. 1472–1480.

[14] J. Yang, H. Zou, Y. Zhou, and L. Xie, "Learning gestures from wifi: A siamese recurrent convolutional architecture," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10 763–10 772, 2019.

[15] K. Ali *et al.*, "Keystroke recognition using wifi signals," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. ACM, 2015, pp. 90–102.

[16] J. Zhang, W. Xu, W. Hu, and S. S. Kanhere, "Wicare: Towards in-situ breath monitoring," in *Proceedings of the 14th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. ACM, 2017, pp. 126–135.

[17] D. Zhang, Y. Hu, Y. Chen, and B. Zeng, "Breathtrack: Tracking indoor human breath status via commodity wifi," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 3899–3911, 2019.

[18] J. Zhang, B. Wei, W. Hu, and S. S. Kanhere, "Wifi-id: Human identification using wifi signal," in *International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 2016, pp. 75–82.

[19] J. Zhang *et al.*, "Human identification using wifi signal," in *International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*. IEEE, 2016, pp. 1–2.

[20] C. Shi, J. Liu, H. Liu, and Y. Chen, "Smart user authentication through actuation of daily activities leveraging wifi-enabled iot," in *Proceedings of the 18th International Symposium on Mobile Ad Hoc Networking and Computing*. ACM, 2017, p. 5.

[21] Z. Chen, L. Zhang, C. Jiang, Z. Cao, and W. Cui, "Wifi csi based passive human activity recognition using attention based blstm," *IEEE Transactions on Mobile Computing*, vol. 18, no. 11, pp. 2714–2724, 2018.

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 248–255.

[23] Y. Liu, J. A. Starzyk, and Z. Zhu, "Optimized approximation algorithm in neural networks without overfitting," *IEEE transactions on neural networks*, vol. 19, no. 6, pp. 983–995, 2008.

[24] D. M. Hawkins, "The problem of overfitting," *Journal of chemical information and computer sciences*, vol. 44, no. 1, pp. 1–12, 2004.

[25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[26] P. Y. Simard, D. Steinkraus, J. C. Platt *et al.*, "Best practices for convolutional neural networks applied to visual document analysis." in *International Conference on Document Analysis and Recognition (ICDAR)*, vol. 3, no. 2003, 2003.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[28] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," in *Sixth International Conference on Spoken Language Processing*, 2000.

[29] F.-H. Liu, Y. Gao, L. Gu, and M. Picheny, "Noise robustness in speech to speech translation," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[30] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *Proceedings of ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, 2013.

[31] J. Schlüter and T. Grill, "Exploring data augmentation for improved singing voice detection with neural networks." in *International Society for Music Information Retrieval (ISMIR)*, 2015, pp. 121–126.

[32] H. Huang and S. Lin, "Widet: Wi-fi based device-free passive person detection with deep convolutional neural networks," in *Proceedings of the 21st International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*. ACM, 2018, pp. 53–60.

[33] D. Soekhoe, P. Van Der Putten, and A. Plaat, "On the impact of data set size in transfer learning using deep neural networks," in *International Symposium on Intelligent Data Analysis*. Springer, 2016, pp. 50–60.

[34] H. Zou, J. Yang, Y. Zhou, L. Xie, and C. J. Spanos, "Robust wifi-enabled device-free gesture recognition via unsupervised adversarial domain adaptation," in *2018 27th International Conference on Computer Communication and Networks (ICCCN)*. IEEE, 2018, pp. 1–8.

[35] F. Wang, W. Gong, and J. Liu, "On spatial diversity in wifi-based human activity recognition: A deep learning-based approach," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2035–2047, 2018.

[36] F. Wang, W. Gong, J. Liu, and K. Wu, "Channel selective activity recognition with wifi: A deep learning approach exploring wideband information," *IEEE Transactions on Network Science and Engineering*, 2018.

[37] S. Arshad, C. Feng, R. Yu, and Y. Liu, "Leveraging transfer learning in multiple human activity recognition using wifi signal," in *Proceedings of 20th International Symposium on" A World of Wireless, Mobile and Multimedia Networks"(WoWMoM)*. IEEE, 2019, pp. 1–10.

[38] B. Sheng, F. Xiao, L. Sha, and L. Sun, "Deep spatial–temporal model based cross-scene action recognition using commodity wifi," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 3592–3601, 2020.

[39] Y. Ma, G. Zhou, S. Wang, H. Zhao, and W. Jung, "Signfi: Sign language recognition using wifi," *Proceedings of Interactive, Mobile, Wearable and Ubiquitous Technologies (UbiComp)*, vol. 2, no. 1, pp. 1–21, 2018.

[40] J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. Patwary, M. Ali, Y. Yang, and Y. Zhou, "Deep learning scaling is predictable, empirically," *arXiv preprint arXiv:1712.00409*, 2017.
[41] Z. Yang *et al.*, "From rssi to csi: Indoor localization via channel response," *ACM Computing Surveys (CSUR)*, vol. 46, no. 2, p. 25, 2013.
[42] Y. Xie, Z. Li, and M. Li, "Precise power delay profiling with commodity wifi," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking (MobiCom)*. New York, NY, USA: ACM, 2015, p. 53–64.
[43] W. Wang *et al.*, "Understanding and modeling of wifi signal based human activity recognition," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. ACM, 2015, pp. 65–76.
[44] I. Jolliffe, *Principal component analysis*. Springer, 2011.
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.
[46] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," in *European conference on computer vision*. Springer, 2016, pp. 646–661.
[47] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 4700–4708.
[48] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.
[49] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves *et al.*, "Conditional image generation with pixelcnn decoders," in *Advances in neural information processing systems (NIPS)*, 2016, pp. 4790–4798.
[50] M. Cogswell, F. Ahmed, R. Girshick, L. Zitnick, and D. Batra, "Reducing overfitting in deep networks by decorrelating representations," *arXiv preprint arXiv:1511.06068*, 2015.
[51] S. Falkner, A. Klein, and F. Hutter, "BOHB: Robust and efficient hyperparameter optimization at scale," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018, pp. 1437–1446.
[52] H. Zou, Y. Zhou, J. Yang, H. Jiang, L. Xie, and C. J. Spanos, "Wifi-enabled device-free gesture recognition for smart home automation," in *Proceedings of 14th International Conference on Control and Automation (ICCA)*. IEEE, 2018, pp. 476–481.
[53] D. A. Khan, S. Razak, B. Raj, and R. Singh, "Human behaviour recognition using wifi channel state information," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7625–7629.
[54] H. Zou, Y. Zhou, J. Yang, and C. J. Spanos, "Towards occupant activity driven smart buildings via wifi-enabled iot devices and deep learning," *Energy and Buildings*, vol. 177, pp. 12–22, 2018.

Jin Zhang received his PhD degree in Computer Science and Engineering from University of New South Wales, Australia in 2017. He is a postdoctoral researcher in Lab for human machine control at Shenzhen institutes of advanced technology, Chinese Academy of Sciences. His research interests include WiFi human sensing, Internet of Things and computer network.

Fuxiang Wu received the Ph.D. degree in pattern recognition and intelligent system from Beihang University, Beijing, China, 2017. He works in Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China. His research interests include text-to-image synthesis, multimodal deep learning, and natural language processing. His current research focuses on text analysis and image editing in multimodal deep learning.

Bo Wei has been a senior lecturer in the Department of Computer and Information Sciences at Northumbria University. He was a Postdoctoral research assistant in University of Oxford. He obtained his PhD degree in Computer Science and Engineering in 2015 from the University of New South Wales, Australia. His research interests are Mobile Computing, Internet of Things, and Wireless Sensor Networks.

Qieshi Zhang (S'07-M'14) received the Ph.D. degree from Waseda University, Japan, in 2014. From 2010 to 2012, he was a Research Fellow with the Japan Society for the Promotion of Science (JSPS), Japan. From 2012 to 2019, he was a Research Assistant, Research Associate and Adjunct Researcher with the Information, Production and Systems Research Center, Waseda University, Japan. He is currently an Associate Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China. His current research interests include computer vision, behavior recognition, autonomous driving, and intelligent robots.

Hui Huang received the M.Sc. degree in computing science from the University of Glasgow, UK, in 2013, and the Ph.D. degree the University of New South Wales, Australia in 2018. He is currently a research associate with Interdisciplinary Centre on Security Reliability and Trust, University of Luxembourg. His research interests include V2X communications, autonomous driving and intelligent transportation systems.

Syed W. Shah received his Ph.D. degree in Computer Science and Engineering from the University of New South Wales (UNSW), Sydney, Australia, and an M.S. degree in Electrical and Electronics Engineering from the University of Bradford, U.K. He is currently working as a Postdoctoral Research Fellow at Deakin University, Australia. His research interests include pervasive/ubiquitous computing, user authentication/identification, Internet of Things, signal processing, data analytics, privacy and security.

Jun Cheng received the B.Eng. and M.Eng. degrees from the University of Science and Technology of China, Hefei, China, in 1999 and 2002, respectively, and the Ph.D. degree from the Chinese University of Hong Kong, Hong Kong, in 2006. He is currently a Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, and the Director of the Laboratory for Human Machine Control. His current research interests include computer vision, robotics, machine intelligence, and control.