The following publication D. Khan and I. W. -H. Ho, "CrossCount: Efficient Device-Free Crowd Counting by Leveraging Transfer Learning," in IEEE Internet of Things Journal, vol. 10, no. 5, pp. 4049-4058, 1 March1, 2023 is available at https://dx.doi.org/10.1109/JIOT.2022.3171449.

CrossCount: Efficient Device-free Crowd Counting by Leveraging Transfer Learning

Danista Khan, *Graduate Student Member, IEEE* and Ivan Wang-Hei Ho, *Senior Member, IEEE*

***Abstract*—Recently, wireless sensing is gaining immense attention in the Internet of things (IoT) for crowd counting and occupancy detection. As wireless signals propagate, they tend to scatter and reflect in various directions depending on the number of people in the indoor environment. The combined effect of these variations on wireless signals is characterized by the channel state information (CSI), which can be further exploited to identify the presence of people. State-of-the-art CSI-based supervised crowd counting systems are vulnerable to temporal and environmental dynamics in practical scenarios as their performance degrades with fluctuations in the indoor environments due to multipath fading. Inspired by the breakthroughs of transfer learning and advancement in edge computing, we have leveraged in this work the concept of transfer learning to minimize this problem via exploiting the trained model from source environment for other indoor environments to perform device-free crowd counting (CrossCount) at the target rooms. Our results show that this technique can combat the dynamics of the environment and achieves 4.7% better accuracy with 40% reduction in training time as compared to conventional convolutional neural networks. In essence, our results imply the future possibility of harnessing crowdsourced CSI data collected at different indoor environments to boost the accuracy and efficiency of local crowd counting systems.**

***Index Terms*—Crowd counting systems, channel state information (CSI), convolutional neural networks (CNN), transfer learning, Internet of things, cloud computing.**

I. Introduction

In the past decade, wireless sensing has gained enormous attention owing to its wide range of applications in the Internet of things (IoT) for indoor localization [1], user-identification [2], crowd counting [3], and human activity recognition (HAR) [4]. Crowd counting holds great importance in many applications such as allocating free spaces at restaurants, bus, and train terminals to optimize customer satisfaction. Moreover, real-time crowd counting can reduce energy consumption in indoor buildings by automatically controlling the air-conditioner and other electrical appliances based on occupancy information. Besides this, an emerging interest is being developed by different governments amid the Covid-19 situation for monitoring quarantine rules and regulations by the deployment of crowd counting systems.

Existing crowd counting approaches can be further classified into two sub-categories: sensor-based and vision-based counting. In the case of the sensor-based approaches [5], the user is required to carry the sensor everywhere for accurate detection, thus making it is intrusive. For vision-based techniques [6], the detections are performed using video cameras which have several constraints. First, people have potential privacy concerns for having cameras in certain locations; therefore such systems cannot be deployed everywhere. Second, the problem of blind spots will hinder continuous monitoring, which occurs due to low light conditions and non-line of sight (NLOS) between the person and camera. Finally, image processing of such systems requires high computation power, which makes it cost-ineffective.

Recently, device-free monitoring has acquired popularity due to the widespread of wireless local area networks in indoor environments. Received signal strength indicator (RSSI) is a MAC layer quantitative parameter that has been proposed to address the issue of wireless crowd counting [7]. However, the decline in monitoring may occur due to multipath fading and environmental noises in RSSI measurements. Channel state information (CSI) is a physical-layer parameter that can overcome the challenges in RSSI-based monitoring. CSI is a fine-grained signature that is sensitive to variations in surroundings due to multipath effect [8].

Previous studies on CSI-based crowd counting systems exploit the traditional supervised machine and deep learning models such as support vector machine (SVM), deep neural networks (DNN), and long short-term memory (LSTM) for the classification of crowd. However, to the best of our knowledge, these techniques do not take the dynamics of the environment into account, which is significant for successful real-time execution of CSI-based crowd counting systems in practical scenarios. Transfer learning is a machine learning technique adapted to learn new tasks for target domain ($\mathcal{D}_{\mathcal{T}}$) by transferring knowledge from already learned tasks of source domain ($\mathcal{D}_{\mathcal{S}}$) [9], [10]. This helps to minimize the overall cost via reducing the dependence on a large amount of target domain data by transferring knowledge across domains [11]. The need for this technique arises when the data are dynamic in nature, and will become outdated as time passes by [12]. Hence, with the recent advancement in deep transfer learning, for the first time in literature, we present a device-free crowd counting system "CrossCount" (XCount) that provides a potential solution to the aforementioned problem by accurately classifying humans in different indoor environments. Our proposed system is computationally light while giving better accuracy as compared to conventional convolutional neural network (CNN) based crowd counting systems.

In this paper, we aim to improve the accuracy and efficiency of crowd counting via transfer learning in CNN. We try to

Danista Khan and Ivan Wang-Hei Ho are with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong e-mail: danista.khan@connect.polyu.hk and ivanwh.ho@polyu.edu.hk

adapt the trained model for the source indoor environments to the temporal and environmental dynamics of the target environment, which makes XCount an ideal candidate to implement crowd counting by leveraging indoor CSI crowdsourced data from intelligent devices [13], [14]. Federated learning [15] trains a centralized cloud model by harnessing decentralized data from distributed edge devices. Therefore, the concept of utilizing crowdsourced data can be further extended to establish a federated transfer learning framework for CSI-based crowd counting.

The major contributions of this paper are as follows:

- We propose a transfer learning-based deep learning framework to categorize the number of people in a room by exploiting the collected CSI data and convolutional neural networks (CNN). Moreover, we deploy the proposed framework on different training sets to empirically verify the ability of transfer learning for accurately characterizing the number of people under different environmental dynamics.
- We demonstrate the environmental adaptability of our model under different conditions, which motivates the possibility of using crowdsourced CSI data in the future to develop a centralized federated transfer learning framework for generating device-free local crowd counting models.
- We experimentally study the limitations of this kind of CSI-based system to provide insights into the implementation of real-time crowd counting systems.

The rest of the paper is organized as follows. In Section II we give an overview of existing research of device-free crowd counting systems. Section III describes our system's design and the proposed transfer learning technique, while in Section IV we show the implementation of our crowd counting system. Section V discusses the results and limitations of our XCrowd system. Finally, in section VI we conclude the paper and provide insights into our future work.

II. Existing Device-free Crowd Counting Systems

Crowd counting and occupancy detection play a vital role for various real-time purposes such as energy-saving, multi-functional space management (e.g., restaurants, airports, and train stations) by updating the customers regarding the available capacity. Existing studies on crowd counting and occupancy detection [16], [17] mainly focus on surveillance using cameras, which results in high deployment cost and privacy concerns. Moreover, studies [18], [19], based on intrusive detection requires the user to carry or wear a sensor device. Device-free crowd counting provides a potential solution to the drawbacks of the aforementioned techniques as these device-free approaches only require existing WiFi infrastructure for the classification of the crowd in an indoor environment.

Device-free crowd counting techniques can be further divided into RSSI and CSI-based approaches.

A. RSSI-based crowd counting systems

The fluctuations in RSSI occur when the person moves in the LOS of the transmitter-receiver wireless link [20]. It has been verified that with the addition of more people in a room, there is more impact on wireless signals. This change helps to recognize the number of people in a room. In studies [21], [22], [23], researchers concluded that:

- RSSI remains at stable values when no subject is present in the sensing area.
- RSSI reading decreases drastically when subjects enter the area of interest.
- Increase in the number of people may block the LOS which will result in a significant drop in RSSI readings.

Overall, RSSI-based approaches require complicated maintenance and high cost as they require dense RF links to detect the exact number of people.

TABLE I
CSI based crowd counting systems

Reference	Input signal	Signal processing	Algorithms
[26]	20 features	DWT denoising	SVM
[3]	Amplitude estimation	PEM matrix	DNN regression
[24]	Phase difference	FE model	SVM
[25]	Phase	BF	SVM
[27]	Amplitude and phase	BF	DNN
[28]	CSI	BF	LSTM network

B. CSI-based crowd counting systems

CSI provides fine-grained information with multiple subcarriers. The changes in CSI measurements due to the presence of people in a room can be extracted to infer the number of people in an indoor environment. Existing works on CSI-based crowd counting systems perform different pre-processing techniques like discrete wavelet transform (DWT), Butterworth filters (BF) such as band-pass filter (BPF), and feature expansion (FE) model for removing the noise from raw CSI and feeding it into classification models.

Based on the variations in signals, authors in [3] formulated a monotonic function for the relationship between CSI fluctuations and the number of moving people in a room. The function calculates the percentage of nonzero elements (PEM) in the dilated CSI matrix as an input to DNN. In [24], authors proposed a crowd counting system based on the phase difference extraction and FE space model. The proposed system is able to achieve an accuracy of 97% for eight classes based on the SVM model. WiCount [25] uses the amplitude and phase to form 180 subcarriers for the classification of people based on three types of datasets: 1) stationary; 2) semi-stationary; and 3) non-stationary. The classification learning algorithm is based on a two-layered DNN. Table I summarizes the major CSI-based crowd counting techniques.

To the best of our knowledge, most of the existing WiFi-based crowd counting works do not consider the impact of temporal and environmental dynamics, which play a crucial role when implementing these systems in real-time practical scenarios for accurate classification.

III. Preliminaries and Rationales

A. Preliminaries of CSI

The infrastructure of CSI-based crowd counting system is comprised of an access point (AP) for packet transmission and a receiver (Rx) for CSI extraction. The AP can be any market-available wireless router that supports 802.11n whereas the Rx is a wireless interface that supports 802.11n with modified firmware to extract CSI [29]. CSI is a fine-grained physical layer measurement that represents the channel frequency response (CFR) between the access point and the receiver by combining the effects of signal scattering and multipath fading. For high data transmission, most WiFi devices are equipped with more than one transmitting and receiving antennas which creates multiple-input multiple-output (MIMO) wireless links. Based on orthogonal frequency domain multiplexing (OFDM), each physical link is divided into multiple subcarriers at the same instant, which helps to transmit more data at the same time. The collected CSI comprises a CSI channel matrix **H**, which is dependent on the total transmitting and receiving antennas. The CSI can be expressed as:

$$\mathbf{H} = [H_1, H_2, H_3.., H_i, ..H_N]^T, i \in [1, N] \quad (1)$$

Where N is the total number of subcarriers and T represents the total number of wireless links between transmitter and receiver. The received CSI contains the combined effect on the wireless channel by the surrounding environment. This combined effect will produce a unique pattern in the time series of the CSI due to the multipath generated by human motion. This CSI measurement can be represented as a 3D matrix for a wireless system, where the complex values provide the details of the attenuation in amplitude and phase shift due to multipath in wireless channels. CSI of a single subcarrier can be represented as:

$$H_i = |H_i| e^{jsin(\angle H_i)} \quad (2)$$

where i represents the i-th subcarrier, $|H_i|$ is the CSI amplitude and $\angle H_i$ is the phase angle of the CSI [30].

B. System Design and Experimental Setup

We propose to count the number of people in a room by leveraging transfer learning in order to deal with the impact of temporal and environmental dynamics on CSI. The framework of the proposed system is shown in Fig. 1.

Fig. 2a and Fig. 2b show the actual setup for crowd counting data collection in rooms BC621 and CD634 of the Hong Kong Polytechnic University. The access point was configured at 2.4 GHz with a channel width of 40 MHz. A minicomputer (hummingboard pro) and a laptop were connected to this access point to develop a wireless link. The laptop was used as a client of AP to send packets at the rate of 20 Hz to the receiver device equipped with the network interface card (NIC) 5300. Sixteen people of different body sizes, height, weight, gender, and age participated in this data collection for crowd counting in the two rooms. The participants were allowed to do anything while being seated on their chairs which means semi-stationary movements were captured. Several combinations were made by positioning participants in different positions to capture the environmental dynamics so that the deep learning models could be trained optimally for practical scenarios.

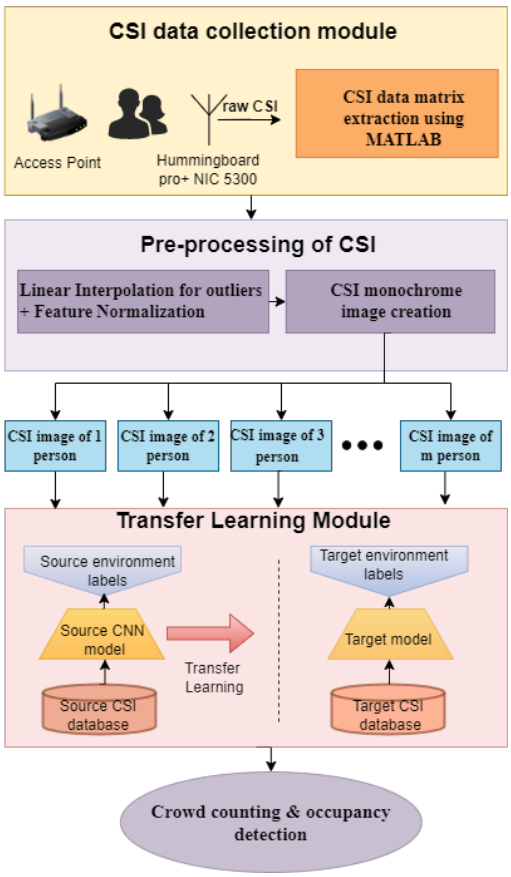
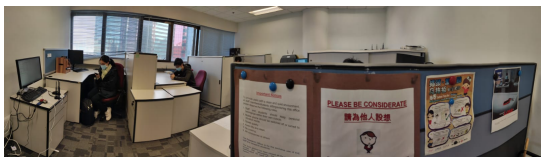


Fig. 1. Framework of XCount.



(a) Research office BC621

(b) Conference room CD634

Fig. 2. Data collection setups for crowd counting.

In the research office BC621, the access point is used under the non-light of sight (NLOS) condition by having a bookshelf between the transmitter and receiver. The data collection was conducted in January 2021 and March 2021 to capture the environmental dynamics. A total of 31 seating combinations were made on the five research desks to capture the CSI data for five classes: one, two, three, four, and five people. CSI monochrome frames were made of the mentioned combinations by taking 50 consecutive samples and 90 subcarriers (30 $\times$ 3 wireless links) for each combination.

TABLE II
CONFIGURATION FOR BC621 AND CD634

Configuration	BC621	CD634
Rx	NIC 5300	NIC 5300
Dimensions of room	5 m by 5.75 m	6.25 m by 8.5 m
No. of Tx	2	2
No. of Rx	3	3
Time difference	2 months	1 week
Frequency Band	2.4 GHz	2.4 GHz
Packet Rate	20 Hz	20 Hz
Total combinations	32	15
Training frames	64	30
Testing frames	34	15
Classes	5	5

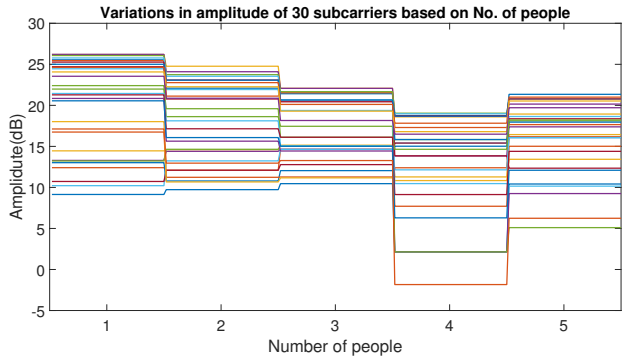


Fig. 3. Variations in amplitude of 30 subcarriers.

The total size of the training data matrix is thus 64 × 90 × 50 × 1, where 64 are the total number of frames, 90 are the subcarriers, 50 are the samples in each frame, and 1 represents the monochrome image. To evaluate the performance of deep learning models, 34 CSI frames were made with different files as a testing dataset to introduce variance in the source and target distributions.

For room CD634, the access point is placed in the line of sight (LOS) of the receiver, and the data was collected in April 2021 with a difference of one week. A total of fifteen seating combinations were made randomly to capture the CSI data for five classes: one, two, three, four, and five people. For the training dataset, a total of 30 CSI frames were formed to train the models. Thus, the dimensions of the training data matrix are 30 × 90 × 50, where 30 are the total CSI frames, 90 and 50 are the subcarriers and samples, receptively. Table II summarizes the experimental configuration for both rooms. Fig. 3 shows the changes in the CSI amplitude of 30 subcarriers for the five classes. It can be observed from Fig. 3 that with an addition of a person in the room, there is a noticeable difference in the CSI amplitude due to the multipath formed by the reflections of WiFi signals. These variations in the amplitude of subcarriers help the learning model to differentiate the number of people in a room even if the people are performing semi-stationary activities.

C. Data pre-processing and synthesis

The received raw CSI from the CSI tool contains many outliers which directly impact the training of the deep learning models. Hence, these outliers are removed by applying linear interpolation to all samples across each subcarrier for all classes, and feature normalization is performed for 90 subcarriers. The input of a CNN model is a time series. As we are dealing with supervised learning, so data has to be labeled according to the classes $Y_{dataset} \in \mathbb{R}^{m\times 5}$ based on the Euclidean distance, where five represents the total number of classes, and m represents the total number of samples in the dataset.

D. Splitting and Shuffling of data

In order to make the training of the proposed model more effective, shuffling of the dataset is performed by collecting different files for every combination. The shuffled dataset is then further split into training, validation, and testing datasets by following the ratio 5:2.5:2.5 for data collected in BC621 and CD634.

E. Performance metrics

To provide a comprehensive performance evaluation for our XCount system, we evaluate our proposed model based on the following performance metrics of machine learning: accuracy, recall, precision, error rate, training time, and F_1 score [31].

F. Software

CSI data was collected using a CSI tool, and then MATLAB was exploited for the extraction of CSI by using the codes provided on [32]. Pre-processing and training of the models were also performed on MATLAB on a Lenovo ThinkPad i5-6200U CPU @ 2.30GHz, 2400 MHz.

IV. IMPLEMENTATION OF XCOUNT

In this section, we present the generic architecture of CNN model, which is used for transfer learning. It is followed by the discussion of how it fits our work.

A. Generic CNN architecture

CNN is a type of DNN that has at least one layer which involves a convolutional operation. They are useful for applications that require image recognition by learning to find objects in the input image. The CNN comprises three types of layers. These are convolutional layers, pooling layers, and fully connected layers. The basic functionality of the CNN model can be divided into four main modules [33].

- The input layer of a CNN holds the pixel values of the input image.
- The convolutional layer takes in the input (or activation from the output of previous layers) and performs convolution operations based on the size of the filters. The purpose of rectified linear unit (ReLU) activation function is to return zero for the negative input so that only the positive part can be returned.

- Pooling layer is used to perform down-sampling to reduce the dimensionality of the input; hence it results in increasing the computational efficiency.
- The purpose of fully connected layers is similar to that in DNNs as they produce class scores from the activation for classification. ReLU function is usually used between these layers to improve the performance.

B. Why do we need CNN models for crowd counting?

Each image of a MNIST dataset [34] has the dimensions of 24 $\times$ 24 $\times$ 1, which means the total number of neurons required at the input layer is equal to 24 $\times$ 24 = 784. As the size of the image increases, the neuron size increases which makes the model computationally ineffective. As a result, there is a requirement for deep learning models to extract features of the input data without losing the main characteristics of the image. Hence, CNN models serve the exact purpose as they reduce the dimensions of the input data and extract the required features automatically. This is helpful in the case of using raw CSI as CNN models can automatically extract the required features from the CSI input images.

C. CNN model for crowd counting

We design and develop a CNN model for classifying the number of people in a room using CSI. Our proposed model comprises 21 layers. Fig. 4 shows the layout of the number of layers in our model. The input of the network is a monochrome image with the dimensions of 50 $\times$ 90 $\times$ 1. Then a two-dimensional convolutional layer is stacked with a filter size of 3 $\times$ 3 and padding of 1. Batch normalization (BN) and ReLU activation function are used after each convolution operation and the number of filters is increased with the order of two with each convolutional layer.

To train the CNN network, the learning rate is set to 0.01, weight initialization is set to "He" [35], the training batch size is 64 and the number of epochs is 30. The training model is optimized and updated by the stochastic gradient descent with the momentum algorithm (SGDM) optimizer. Each frame consists of 50 sample points which represent a mini-batch. The overlapping between the samples of the input frames is zero. The 30 CSI frames from room CD634 with balanced classes are used for training. Fifteen frames are used as the validation dataset and fifteen frames are used to evaluate the performance of the trained model as the testing set.

D. Transfer learning model for crowd counting

We first identify the CNN model that gives the best accuracy for the testing data of CD634 (source domain), and this twenty-one-layered trained network is saved and used to perform transfer learning on different target environments. The weights of the lower layers (closer to input) are kept constant by freezing the layers. For our case, freezing the weights of the first sixteen layers gives the best results. Hence, the pre-trained model with frozen layers is used to apply transfer learning on the dataset of BC621. The layout of the transfer learning model is shown in Fig. 4.

To fine-tune the last five layers of the network with the training data of BC621 (target domain which was collected in a different room on different day, the training parameters are kept constant as that of the pre-trained model. The overlapping between the samples of the input frames is zero. The 64 CSI frames from room BC621 with unbalanced classes are used for retraining to cover all possible seating arrangements. 34 frames are used as the validation dataset and 34 frames are used to evaluate the performance of the model as the testing set. The proposed transfer learning scheme is summarized in Algorithm 1.

Algorithm 1 Algorithm for transfer learning

Input:

- CSI source dataset ($\mathcal{D}_{\mathcal{S}}$)
- CSI target data ($\mathcal{D}_{\mathcal{T}}$)
- Learning rate γ
- Batch size B
- Epoches E

Output: Crowd classification labels (l)

1: **Initialize** the weights with He initialization.
2: Train CNN model with $\mathcal{D}_{\mathcal{S}}$.
3: **Repeat** the training to get the base model with maximum accuracy.
4: Freeze the layers of the pre-trained base model.
5: Fine-tune the output layers with $\mathcal{D}_{\mathcal{T}}$.

Additionally, we verify the adaptability of XCrowd for different target environments by leveraging the same pre-trained model for the CSI data collected in BC602 classroom after nine months. In this scenario, we use 60 CSI frames with balanced classes to fine-tune the pre-trained model, while 30 different frames with the same distribution are used for validation and testing datasets.

V. Experimental results and performance evaluation

To visualize the test dataset attained after feature normalization from rooms BC621 and CD634, we plot the t-SNE visualizations [36] of the test data in the two-dimensional space to observe the overlapping and scattering of different classes and combinations. It can be seen from Fig. 5a that the five classes are scattered over the 2-D plane due to thirty-one different combinations. The t-SNE plot of room CD634 shown in Fig. 5b also shows the scattering of testing data based on the five classes for fifteen combinations. The two plots show the massive difference in the two testing datasets, which verifies the temporal and environmental dynamics in CSI data due to the changes in the room and time.

We compare the performance of the two networks based on the following parameters: average accuracy, maximum accuracy, F_1 score, recall, error rate, training time, and precision. Fig. 6a and Fig. 6b show the results of the two networks based on ten iterations. As can be observed in the two heatmaps, the average accuracy achieved for the same dataset with transfer learning is 81.47% whereas for the simple CNN model the mean accuracy attained is 76.764% which is 4.7%

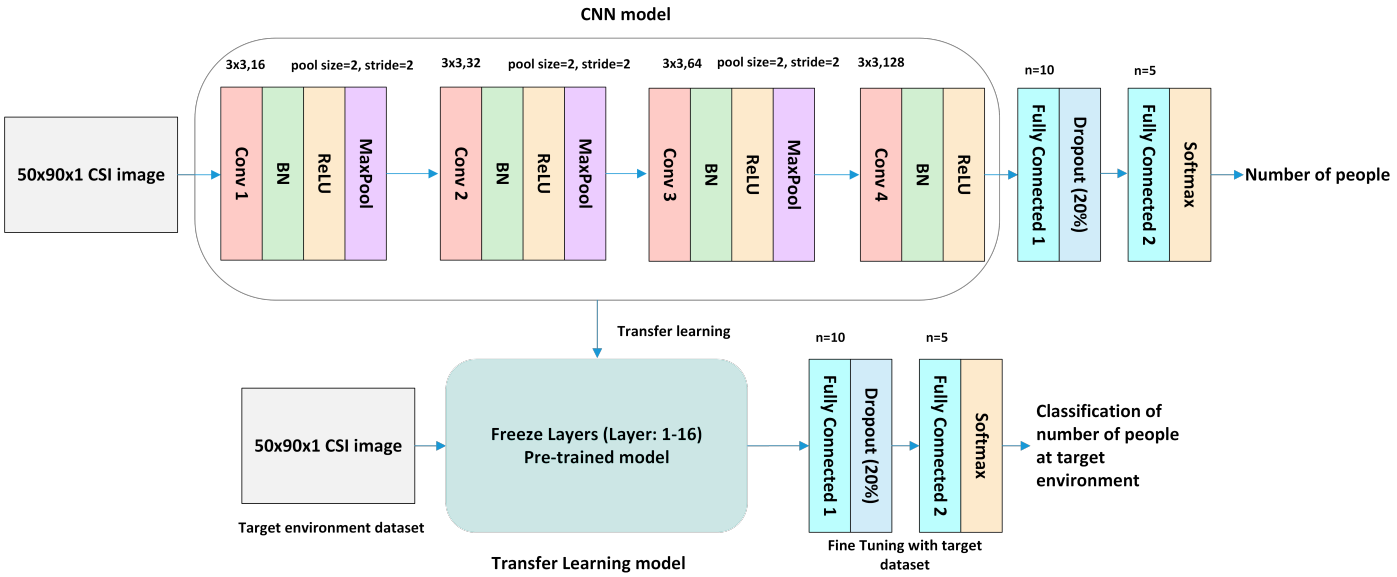
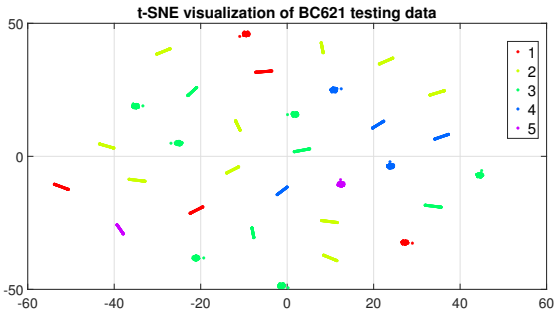
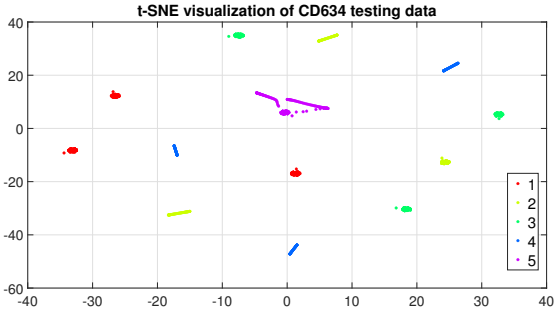


Fig. 4. CNN and transfer learning models for crowd counting. The upper layout shows the 21-layered CNN model, where Conv1-Conv4 layers perform the convolution operations. After every convolutional layer, batch normalization is performed to speed up the learning process before applying the ReLU activation function that helps to make the deep learning model more robust. The MaxPool layers after the ReLU function perform pooling operations on the feature maps. The lower layout shows the transfer learning model for XCrowd where the weights of the first sixteen layers are pre-trained for transfer learning and the last five layers are fine-tuned with the target domain data.



(a) t-SNE visualization of testing data for BC621



(b) t-SNE visualization of testing data for CD634

Fig. 5. t-SNE visualization to observe testing data in 2D.

less accurate. Moreover, the average training time for our proposed model and the generic CNN model is 9.3 and 15.6 seconds, respectively. This huge difference in time shows that our proposed model computationally trains 40% faster than the CNN model and with 4.7% performance boost. We further evaluate the performance of our dataset by exploiting a LSTM network which is suitable for time-series data. To train the LSTM network, the learning rate is set to 0.01 and the number of hidden units is 100. The results are summarized in Fig.6c, which shows that the transfer learning model outperforms both the CNN and LSTM models in every performance metric for the application of CSI-based crowd counting.

A. Impact of the number of layers on the transfer learning model for crowd counting

To analyze the impact on detection accuracy by changing the number of frozen layers for transfer learning, we change the frozen layers from two to eighteen. The graph in Fig. 7 shows the variations in the maximum accuracy and training time due to the changing of the number of frozen layers. It can be observed from the graph that the maximum accuracy and reduced training time is achieved by freezing the first sixteen layers of the model.

B. Impact of the distance between the transmitter and receiver and LOS/NLOS

To analyze the impact of the receiver's distance from the transmitter, we collected the CSI data from two receivers placed at a distance of r_1 = 5.8 m and r_2 = 3.3 m from the transmitter, respectively. The floor layout is shown in Fig. 8 for room CD634 in May 2021. To create a NLOS scenario, a whiteboard is placed in between the transmitter and receivers. Fig. 9a and Fig. 9b show the results for the LOS scenario with two receivers r_1 and r_2. The results for the NLOS scenario are shown in Fig. 9c and Fig. 9d.

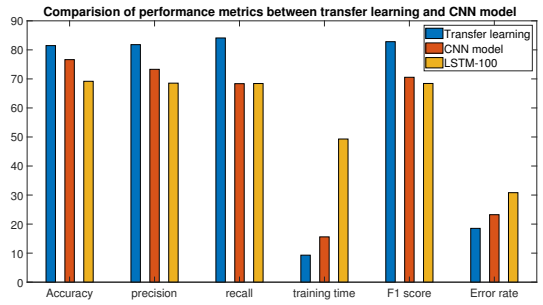
The summary of the average accuracy for all of the four cases is plotted in Fig. 10. It can be observed from the

Iteration	1	2	3	4	5	6	7	8	9	10	Average
Accuracy	88.24	82.35	73.53	70.59	94.12	85.29	82.35	82.35	79.41	76.47	81.47
Precision	92	82	70	70	96	82	82	88	80	76	81.8
Recall	88.29	88.01	58.5	80.77	96	91.38	86.2	84.56	78.89	88.39	84.099
Time	7	6	8	6	11	11	11	11	11	11	9.3
F1 score	90.1	84.9	64.32	75	96	86.44	84.05	86.25	79.44	81.73	82.823
Error rate	11.76	17.65	26.47	29.41	5.88	14.71	17.65	17.65	20.59	23.53	18.53

(a) Performance metrics for transfer learning

Iteration	1	2	3	4	5	6	7	8	9	10	Average
Accuracy	76.47	79.41	85.29	67.65	73.53	67.65	82.35	67.65	82.35	85.29	76.764
Precision	72	74	84	66	68	64	79	66	76	84	73.3
Recall	61.33	63.57	89.44	56.77	62.46	55.32	84.94	55.4	66.06	88.29	68.358
Time	15	15	16	16	18	16	15	15	15	15	15.6
F1 score	66.24	68.39	86.64	61.04	65.11	59.34	81.86	60.24	70.68	86.09	70.563
Error rate	23.53	20.59	14.71	32.35	26.47	32.35	17.65	32.35	17.65	14.71	23.236

(b) Performance metrics for CNN model



(c) Comparison of performance metrics between transfer learning model and CNN model

Fig. 6. Performance metrics for deep learning models are represented in the two heatmaps. The values encircled in blue shows the average values of the two models whereas values encircled in orange and black shows the best and worst accuracy values of the two techniques, respectively.

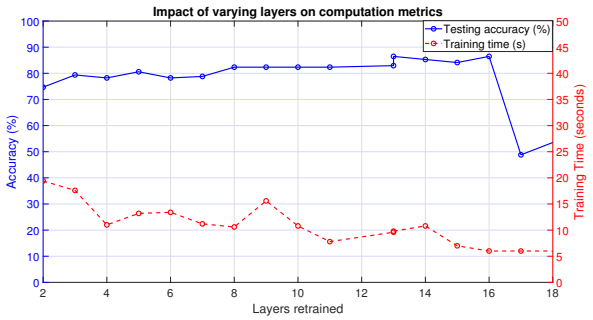


Fig. 7. Impact of changing layers (frozen) on accuracy and training time.

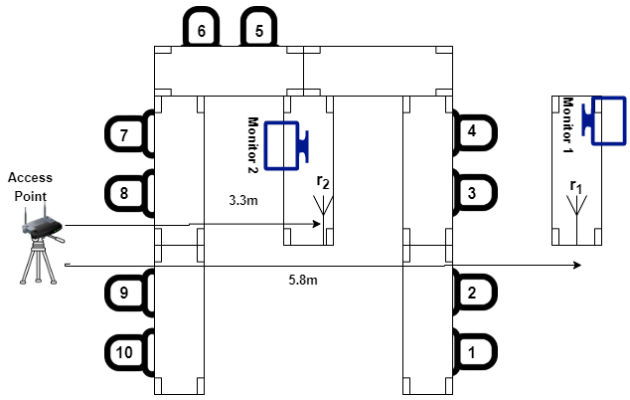


Fig. 8. Floor layout of CD634 to monitor the impact of transmission distance and LOS/NLOS.

Iteration	1	2	3	4	5	6	7	8	9	10	Average
Accuracy	73.33	60	53.33	73.33	66.67	73.33	60	66.67	80	60	66.666
Precision	84	61.67	70.83	85	76.67	82	76	82	85.33	73.33	77.683
Recall	73.33	60	53.33	73.33	66.67	73.33	60	66.67	80	60	66.666
Time	6	6	6	6	6	6	6	6	6	6	6
F1 score	78.31	60.82	60.85	78.74	71.32	77.42	67.06	73.54	82.58	66	71.664
Error rate	26.67	40	46.67	26.67	33.33	26.67	40	33.33	20	40	33.334

(a) Performance evaluation for LOS with r_1

Iteration	1	2	3	4	5	6	7	8	9	10	Average
Accuracy	73.33	80	86.67	66.67	80	86.67	86.67	80	80	80	80.001
Precision	85	87	86.67	82	83.33	90	90	87	87	87	86.5
Recall	73.33	80	86.67	66.67	80	86.67	86.67	80	80	80	80.001
Time	6	6	6	6	6	6	6	6	6	6	6
F1 score	78.74	83.35	86.67	73.54	81.63	88.3	88.3	83.35	83.35	83.35	83.058
Error rate	26.67	20	13.33	33.33	20	13.33	13.33	20	20	20	19.999

(b) Performance evaluation for LOS with r_2

Iteration	1	2	3	4	5	6	7	8	9	10	Average
Accuracy	80	80	80	93.33	80	73.33	73.33	66.67	66.67	86.67	78
Precision	81.67	81.67	87	95	81.67	78.33	78.33	68.67	81.9	90	82.424
Recall	80	80	80	93.33	80	73.33	73.33	66.67	66.67	86.67	78
Time	6	6	6	6	6	6	6	6	6	6	6
F1 score	80.82	80.82	83.35	94.16	80.82	75.75	75.75	67.65	73.5	88.3	80.092
Error rate	20	20	20	6.67	20	26.67	26.67	33.33	33.33	13.33	22

(c) Performance evaluation for NLOS with r_1

Iteration	1	2	3	4	5	6	7	8	9	10	Average
Accuracy	86.67	80	86.67	86.67	86.67	86.67	73.33	100	93.33	93.33	87.334
Precision	88.33	83.33	88.33	88.33	90	88.33	85	100	95	95	90.165
Recall	86.67	80	86.67	86.67	86.67	86.67	73.33	100	93.33	93.33	87.334
Time	6	6	6	6	6	6	6	6	6	6	6
F1 score	87.49	81.63	87.49	87.49	88.3	87.49	78.74	100	94.16	94.16	88.695
Error rate	13.33	20	13.33	13.33	13.33	13.33	26.67	0	6.67	6.67	12.666

(d) Performance evaluation for NLOS with r_2

Fig. 9. Performance evaluation of XCount by changing transmission distance in LOS/NLOS scenarios.

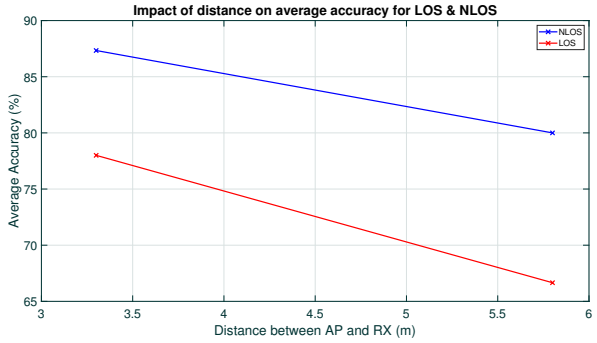


Fig. 10. Summary of average accuracy.

graph that transmission distance and positioning of transmitter have a significant impact on the overall average accuracy of XCount. The results clearly indicate that the average accuracy for NLOS is better as compared to LOS. Additionally, the positioning of the receiver also plays a vital role in classifying the correct number of people in the room. Therefore, by decreasing the transmission distance, the accuracy of classification can be improved in practical scenarios.

C. Impact of antenna selection

To determine the impact of the antenna on the average accuracy, we conducted different experiments with the CNN model by choosing 30 subcarriers. Fig. 11 shows the accuracy of different antennas of receiver r_1 for the LOS and NLOS scenarios. It can be observed that the accuracy of antenna 2 is the highest under LOS condition, while the accuracy of antenna 3 is the highest under NLOS condition. This is because each antenna behaves differently with respect to the motion of humans, and a slight change in the scenario could

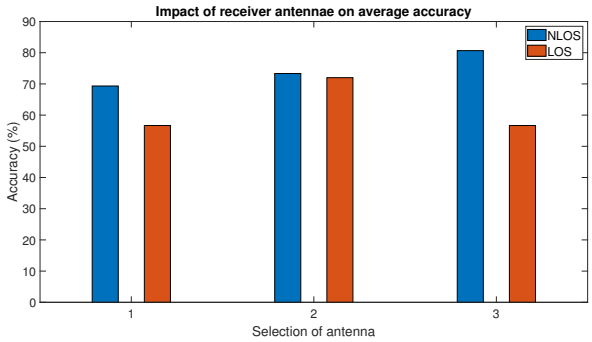


Fig. 11. Performance of three antennas for LOS and NLOS.

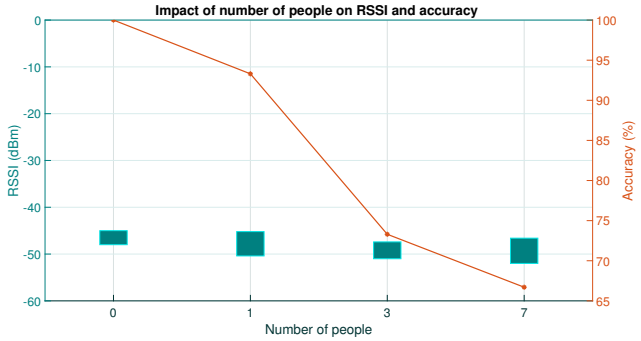


Fig. 12. Impact of RSSI on average accuracy.

deeply impact the accuracy of the antenna. This is the reason that we chose 90 subcarriers that have the combined effect of antennas for our XCount system.

D. Impact of RSSI

To evaluate the threshold of RSSI which gives the optimum result, we carried out a number of iterations for different classes to find the optimal threshold under the NLOS scenario for r_2. Fig. 12 shows the accuracy and range of RSSI for class zero, one, three, and seven people in CD634. It can be observed from the graph that as the number of people increases in a room, the received RSSI decreases which directly impacts the accuracy of the classification. As a result, the performance of the model decreases with the degradation in received RSSI. There is a significant decrease in accuracy when the RSSI of received samples is less than -50 dBm. Hence, in practical scenarios, the edge device can be programmed to automatically discard samples that are lesser than the threshold to ensure the detection accuracy of the model.

E. Impact of the number of people

To identify the limitations of our proposed system, we increased the number of classes from five to eight and examined if the transfer learning model is adaptable to the increase in the number of people. For five, six, and seven classes, the validation and testing data contains CSI frames for one to five, one to six, and one to seven people in a room whereas for eight classes, the validation and testing data contains CSI frames for zero to seven people in the CD634 room. The average accuracy

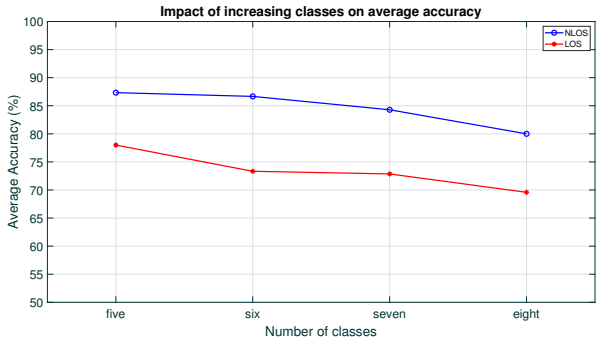


Fig. 13. Average accuracy for different classes.

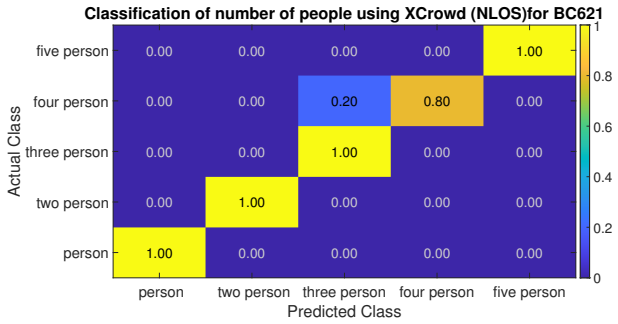
results for both NLOS and LOS are shown in Fig. 13 for r_2. It can be observed from Fig. 13 as the number of classes increases from five to eight, the average accuracy decreases for both LOS and NLOS conditions, which shows that the average accuracy of XCrowd decreases with the increasing number of people to detect.

F. Crowd classification based on transfer learning for different target indoor environments

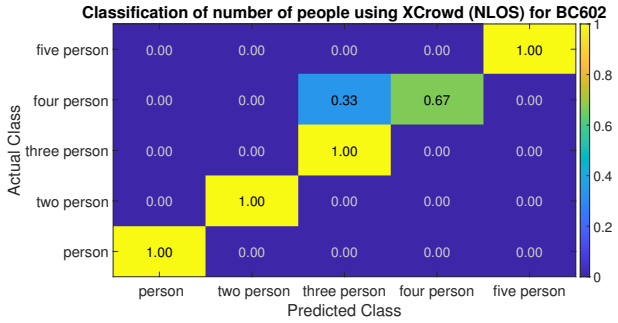
We evaluated the performance of our XCrowd system for classifying the number of people in BC621 with unbalanced-NLOS classes. The confusion matrix for this scenario is shown in Fig. 14a. The result shows that our XCrowd model is able to classify "one-person", "two-person", "three-person" and "five-person" cases with 100% accuracy whereas for the "four-person" class, the model may mis-regard it as the three-person class with a 20% chance, leading to an overall average accuracy of 96%. To further validate the performance of XCrowd, the model is evaluated for the balanced-NLOS CSI data at a classroom BC602 with the dimensions of 5.4 $\times$ 4.2 m^2. The layout of this target indoor environment is totally different from the source CD634 meeting room. The result is shown in the confusion matrix in Fig. 14b, and the overall average accuracy is 93.4%. These results demonstrate that our proposed model can accurately count the number of people in different target indoor environments, which can be further extended for energy-saving and quarantine applications to monitor the presence of human activity in an indoor environment".

VI. Conclusion and Future Work

In this paper, we have leveraged transfer learning to adapt the temporal and environmental dynamics of CSI for a crowd counting system (XCrowd). To evaluate the performance of the proposed model, we have collected CSI data that corresponds to different number of people using the CSI tool installed on readily available NIC 5300 and hummingboard pro devices. We have analyzed the results using various performance metrics, and our experimental results indicate that our proposed transfer learning model is 40% computationally lighter and 4.7% more accurate as compared to conventional CNN models. We have further identified the limitations of the system by exhaustively evaluating XCrowd under many



(a) Crowd classification NLOS for BC621



(b) Crowd classification NLOS for BC602

Fig. 14. Crowd classification by XCount for different indoor environments

practical scenarios such as increasing the distance between the access point and receiver, LOS or NLOS scenarios, increasing the number of output classes from five to eight, studying the relationship between RSSI and average accuracy, and the impact of antenna selection on the accuracy of XCrowd.

Since our XCount framework is computationally efficient and can adapt to the dynamics of the environment with high accuracy, local models can be further enhanced via harnessing crowdsourced CSI data collected from different intelligent devices at different locations. This will facilitate the development of a federated transfer learning framework for device-free crowd counting in practical scenarios.

ACKNOWLEDGMENT

This work was supported in part by the Key-Area Research and Development Program of Guangdong Province (2020B 090928001).

REFERENCES

[1] J. M. Rocamora, I. W.-H. Ho, W.-M. Mak, and A. P.-T. Lau, "Survey of CSI fingerprinting-based indoor positioning and mobility tracking systems," *IET Signal Processing*, vol. 14, no. 7, pp. 407–419, 2020.

[2] T. Xin, B. Guo, Z. Wang, M. Li, Z. Yu, and X. Zhou, "Freesense: Indoor human identification with Wi-Fi signals," in *2016 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2016, pp. 1–7.

[3] R. Zhou, X. Lu, Y. Fu, and M. Tang, "Device-free crowd counting with WiFi channel state information and deep neural networks," *Wireless Networks*, pp. 1–12, 2020.

[4] D. Khan and I. W.-H. Ho, "Deep Learning of CSI for Efficient Device-free Human Activity Recognition," in *2021 IEEE 7th World Forum on Internet of Things (WF-IoT)*. IEEE, 2021, pp. 19–24.

[5] J. Weppner and P. Lukowicz, "Bluetooth based collaborative crowd density estimation with mobile phones," in *2013 IEEE international conference on pervasive computing and communications (PerCom)*. IEEE, 2013, pp. 193–200.

[6] M. Kim, W. Kim, and C. Kim, "Estimating the number of people in crowded scenes," in *Visual Information Processing and Communication II*, vol. 7882. International Society for Optics and Photonics, 2011, pp. 171 – 178.

[7] E. Cianca, M. De Sanctis, and S. Di Domenico, "Radios as sensors," *IEEE Internet of Things Journal*, vol. 4, no. 2, pp. 363–373, 2016.

[8] Y. Ma, G. Zhou, and S. Wang, "WiFi sensing with channel state information: A survey," *ACM Computing Surveys (CSUR)*, vol. 52, no. 3, pp. 1–36, 2019.

[9] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global, 2010, pp. 242–264.

[10] Y. Wang, G. Gui, H. Gacanin, T. Ohtsuki, H. Sari, and F. Adachi, "Transfer learning for semi-supervised automatic modulation classification in ZF-MIMO systems," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 10, no. 2, pp. 231–239, 2020.

[11] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[12] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[13] X. Hu, Z. Ning, K. Zhang, E. Ngai, K. Bai, and F. Wang, "Crowdsourcing for Mobile Networks and IoT," 2018.

[14] S. Yu, X. Chen, S. Wang, L. Pu, and D. Wu, "An edge computing-based photo crowdsourcing framework for real-time 3D reconstruction," *IEEE Transactions on Mobile Computing*, 2020.

[15] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[16] S.-Y. Cho, T. W. Chow, and C.-T. Leung, "A neural-based crowd estimation by hybrid global learning algorithm," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 29, no. 4, pp. 535–541, 1999.

[17] R. Ma, L. Li, W. Huang, and Q. Tian, "On pixel count based crowd density estimation for visual surveillance," in *IEEE Conference on Cybernetics and Intelligent Systems, 2004.*, vol. 1. IEEE, 2004, pp. 170–173.

[18] J. Weppner and P. Lukowicz, "Collaborative crowd density estimation with mobile phones," *Proc. of ACM PhoneSense*, 2011.

[19] P. G. Kannan, S. P. Venkatagiri, M. C. Chan, A. L. Ananda, and L.-S. Peh, "Low cost crowd counting using audio tones," in *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems*, 2012, pp. 155–168.

[20] M. Youssef, M. Mah, and A. Agrawala, "Challenges: device-free passive localization for wireless environments," in *Proceedings of the 13th annual ACM international conference on Mobile computing and networking*, 2007, pp. 222–229.

[21] M. Nakatsuka, H. Iwatani, and J. Katto, "A study on passive crowd density estimation using wireless sensors," in *The 4th Intl. Conf. on Mobile Computing and Ubiquitous Networking (ICMU 2008)*. Citeseer, 2008.

[22] Y. Yuan, J. Zhao, C. Qiu, and W. Xi, "Estimating crowd density in an RF-based dynamic environment," *IEEE Sensors Journal*, vol. 13, no. 10, pp. 3837–3845, 2013.

[23] C. Xu, B. Firner, R. S. Moore, Y. Zhang, W. Trappe, R. Howard, F. Zhang, and N. An, "SCPL: indoor device-free multi-subject counting and localization using radio signal strength," in *Proceedings of the 12th international conference on Information Processing in Sensor Networks*, 2013, pp. 79–90.

[24] J. Zong, B. Huang, L. He, B. Yang, and X. Cheng, "Device-free crowd counting based on the phase difference of channel state information," in *2020 IEEE International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA)*, vol. 1. IEEE, 2020, pp. 1343–1347.

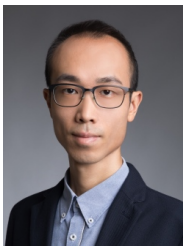
[25] L. Zhang, Y. Zhang, B. Wang, X. Zheng, and L. Yang, "WiCrowd: Counting the Directional Crowd With a Single Wireless Link," *IEEE Internet of Things Journal*, vol. 8, no. 10, pp. 8644–8656, 2020.

[26] H. Zou, Y. Zhou, J. Yang, W. Gu, L. Xie, and C. Spanos, "Freecount: Device-free crowd counting with commodity WiFi," in *GLOBECOM 2017-2017 IEEE Global Communications Conference*. IEEE, 2017, pp. 1–6.

[27] S. Liu, Y. Zhao, and B. Chen, "WiCount: a deep learning approach for crowd counting using WiFi signals," in *2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC)*. IEEE, 2017, pp. 967–974.

[28] D. Konings and F. Alam, "LifeCount: A Device-free CSI-based Human Counting Solution for Emergency Building Evacuations," in *2020 IEEE Sensors Applications Symposium (SAS)*. IEEE, 2020, pp. 1–5.
[29] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Predictable 802.11 packet delivery from wireless channel measurements," *ACM SIGCOMM Computer Communication Review*, vol. 40, no. 4, pp. 159–170, 2010.
[30] J. Liu, H. Liu, Y. Chen, Y. Wang, and C. Wang, "Wireless sensing for human activity: A survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1629–1645, 2019.
[31] D. Sánchez-Rodríguez, M. A. Quintana-Suárez, I. Alonso-González, C. Ley-Bosch, and J. J. Sánchez-Medina, "Fusion of channel state information and received signal strength for indoor localization using a single access point," *Remote Sensing*, vol. 12, no. 12, p. 1995, 2020.
[32] "dhalperi/linux-80211n-CSI tool: 802.11n CSI tool based on iwlwifi and linux." [Online]. Available: https://github.com/dhalperi/linux-80211n-csitool/
[33] P. Kim, "Convolutional neural network," in *MATLAB deep learning*. Springer, 2017, pp. 121–147.
[34] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.
[35] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.
[36] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE." *Journal of machine learning research*, vol. 9, no. 11, 2008.
[37] Y. Xu, *Autonomous Indoor Localization Using Unsupervised Wi-Fi Fingerprinting*. Kassel university press GmbH, 2016.
[38] S. Depatla, A. Muralidharan, and Y. Mostofi, "Occupancy estimation using only WiFi power measurements," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 7, pp. 1381–1393, 2015.
[39] M. Muaaz, A. Chelli, M. W. Gerdes, and M. Pätzold, "Wi-Sense: A passive human activity recognition system using Wi-Fi and convolutional neural network and its integration in health information systems," *Annals of Telecommunications*, pp. 1–13, 2021.
[40] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

Danista Khan (Graduate Student Member, IEEE) received her MSc degree in Electrical Engineering from University of Engineering and Technology, Lahore in 2017. She completed her graduation in Electrical Engineering in 2013 from The University of Lahore by securing a gold medal. She is currently pursuing the Ph.D. degree with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong. From 2013 to 2019, she was a Lecturer in the Department of Electrical Engineering at The University of Lahore, Lahore. She has been awarded with HKAUW Postgraduate Scholarship, in 2020. Her research is focused on device free human activity recognition and crowd counting.



Ivan Wang-Hei Ho (M'10–SM'18) received the B.Eng. and M.Phil. degrees in information engineering from The Chinese University of Hong Kong, Hong Kong, in 2004 and 2006, respectively, and the Ph.D. degree in electrical and electronic engineering from the Imperial College London, London, U.K., in 2010. He was a Research Intern with the IBM Thomas J. Watson Research Center, Hawthorne, NY, USA, and a Postdoctoral Research Associate with the System Engineering Initiative, Imperial College London. In 2010, he cofounded P2 Mobile Technologies Ltd., where he was the Chief Research and Development Engineer. He is currently an Associate Professor with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong. His research interests include wireless communications and networking, specifically in vehicular networks, intelligent transportation systems (ITS), and Internet of things (IoT). He primarily invented the MeshRanger series wireless mesh embedded system, which received the Silver Award in Best Ubiquitous Networking at the Hong Kong ICT Awards 2012. His work on indoor positioning and IoT also received a number of awards, including the Gold Medal at iENA 2019, the Gold Medal with the Organizer's Choice Award at iCAN 2020, and the Gold Medal at the International Exhibition of Inventions Geneva in 2021. He is currently an Associate Editor for the IEEE Access and IEEE Transactions on Circuit and Systems II, and was the TPC CoChair for the PERSIST-IoT Workshop in conjunction with ACM MobiHoc 2019 and IEEE INFOCOM 2020.