On the Fine-Grained Crowd Analysis via Passive WiFi Sensing

Lifei Hao, Baoqi Huang, *Member, IEEE*, Bing Jia, *Member, IEEE*, and Guoqiang Mao, *Fellow, IEEE*

***Abstract*—Regarding the passive WiFi sensing based crowd analysis, this paper first theoretically investigates its limitations, and then proposes a deep learning based scheme targeted for returning fine-grained crowd states in large surveillance areas. To this end, three key challenges are coped with: to relieve the influences of the randomness and sparsity induced by passive WiFi sensing, an attention-based deep convolutional autoencoder model is designed to recover accurate crowd density maps in a way similar to image reconstruction; to combat the anonymity caused by MAC randomization, following the identification of local high-density crowds (LHDCs) with the density clustering algorithm, i.e., DM-DBSCAN, a bidirectional convolutional LSTM based model is employed to infer LHDC speeds; to overcome the absence of passive WiFi sensing datasets for model training, three semi-synthetic datasets are produced by emulating passive WiFi sensing with practical pedestrian tracking datasets. Extensive experiments confirm that, the proposed scheme significantly outperforms existing WiFi-based methods in terms of crowd density estimation and provides superior crowd speed estimation. More importantly, the scheme can also produce consistent crowd states on a real-world dataset, revealing that it has the ability to support accurate, visualized and real-time crowd monitoring in large surveillance areas.**

***Index Terms*—Crowd analysis, passive WiFi sensing, dataset, crowd density regression, speed estimation.**

I. Introduction

CROWD analysis, the study of how crowds are distributed in space and move over time, is a key focus in research communities [1]. It is crucial for various applications [2] such as crowd management, traffic control, urban planning, and surveillance. Recent stampedes in Itaewon, South Korea [3], and Kanjuruhan Stadium, Indonesia [4], highlight the importance of using crowd analysis to monitor emergency situations and implement control operations to prevent accidents and casualties.

Existing crowd analysis studies can be divided into two categories. First, the vision-based approach extracts crowd state information from images or videos based on detection or regression [5], [6], [7], [8], and suffers from high density scenarios, limited coverage, complex cross-camera processing, high deployment cost, and huge computational complexity [9], [10]. Second, the wireless-based (or specifically WiFi-based) approach infers crowd properties by leveraging the relationship between wireless signals and crowd states [11], [12]. Therein, the WiFi channel state information (CSI)-based approach can achieve relatively high accuracy in crowd counting [13], [14], but incurs high costs in obtaining CSI data, small deployment space, and limited scalability; in contrast, the passive WiFi sensing-based approach, which employs WiFi sniffers to passively sense nearby pedestrians via capturing probe frames sent by their mobile devices, is most promising due to its advantages of low cost, large coverage, and strong scalability [15], [16], [17], and has been validated to be applicable [18], [19] even if only limited accuracy can be obtained.

In specific crowd analysis tasks, estimating accurate and comprehensive static crowd states, including crowd counts and their spatial distribution, is still challenging. Most existing passive WiFi sensing-based methods were dedicated to improving coarse-grained crowd counts [18], [19], [20], whereas some recent researches gradually began to explore scenario-level crowd density maps (CDMs) or heat maps [16], [18], [21] by fusing sensing data and WiFi localization results, but suffered from both the randomness and sparsity of passive WiFi measurements as well as significant localization errors, and additionally, effective verifications and convincing evaluations are still missing. Therefore, it is imperative to understand the limitations of crowd density estimation and further develop effective methods to estimate accurate CDMs via passive WiFi sensing.

Besides, crowd analysis also involves estimating crowd dynamics, e.g., crowd speeds. Even if existing WiFi-based methods are capable of estimating pedestrian speeds in some specific scenarios (e.g., one-way or two-way passages [16], [22]), it is still difficult to deploy them in an arbitrary scenario.

In this paper, a theoretical analysis is conducted to investigate the limitations of crowd analysis based on passive WiFi sensing, motivating us to propose a novel fine-grained crowd analysis scheme for large surveillance areas. To this end, three key challenges must be addressed. First, since current mobile devices often trigger not regular but occasional active scans

Manuscript received 23 March 2023; revised 24 July 2023; accepted 10 October 2023. Date of publication 13 October 2023; date of current version 7 May 2024. This work was supported in part by the National Natural Science Foundation of China under Grants 62262046, 42161070, and U21A20446, in part by the Major Program of Natural Science Foundation of Inner Mongolia A. R. of China under Grant 2021ZD13, in part by the Science and Technology Plan Project of Inner Mongolia A. R. of China under Grants 2022YFSJ0027 and 2021GG0163, and in part by the University Youth Science and Technology Talent Development Project (Innovation Group Development Plan) of Inner Mongolia A. R. of China under Grant NMGIRT2318. Recommended for acceptance by X. Tian. *(Corresponding author: Baoqi Huang.)*

Lifei Hao, Baoqi Huang, and Bing Jia are with the Engineering Research Center of Ecological Big Data, Ministry of Education, the Inner Mongolia A.R. Key Laboratory of Wireless Networking and Mobile Computing, and the College of Computer Science, Inner Mongolia University, Hohhot 010021, China (e-mail: haolifei@mail.imu.edu.cn; cshbq@imu.edu.cn; jiabing@imu.edu.cn).

Guoqiang Mao is with the Research Institute of Smart Transportation, Xidian University, Xi'an 710071, China (e-mail: guoqiangmao@xidian.edu.cn).

Digital Object Identifier 10.1109/TMC.2023.3324334

with randomized MAC addresses [23], [24], random and sparse passive WiFi sensing measurements are produced, and thus only coarse-grained CDMs (termed WDMs) can be obtained using existing methods. Inspired by the similarity between image reconstruction and the recovery of CDMs from WDMs, an attention-based deep convolutional autoencoder (ADCA) model is developed to establish a fine-grained mapping between pixels or patches in WDMs and those in corresponding CDMs by effectively exploiting spatial-temporal information from WDMs via convolutions, sufficiently enriching measurements from random and sparse WDMs via autoencoder, and relieving the influence of non-uniform localization errors cross the whole surveillance area via attention mechanism and structural loss function. Second, the usage of MAC randomization also results in the anonymity of passive WiFi measurements, which makes it difficult to continuously track pedestrians and calculate their speeds. Therefore, we propose a DM-oriented density clustering algorithm, i.e., DM-DBSCAN, to identify local high-density crowds (LHDCs) from an estimated CDM, and a bidirectional convolutional LSTM (BiConvLSTM) model to infer their speed vectors by detecting their changes in space and time from a sequence of CDM patches (SCDMPs). Third, due to the extremely high costs in labeling pedestrians' states in large surveillance areas, labeled passive WiFi sensing datasets for crowd analysis are still absent. As such, by grasping key rules of active scans, MAC randomization and localization error distributions summarized from abundant and complex real-world datasets, we propose to emulate passive WiFi sensing in realistic pedestrian tracking datasets as real as possible, producing three semi-synthetic datasets for enabling the training and evaluating of proposed models.

For the purpose of performance evaluation, we conduct extensive experiments on semi-synthetic datasets and a real-world dataset collected on our campus with the surveillance area of about 4000 m^2, respectively. The former confirms that the proposed scheme significantly outperforms existing WiFi-based methods for crowd density estimation, and can achieve superior speed estimation with an average relative error of velocity estimation less than 23% (37%) and an average absolute error of direction estimation less than 18° (32°) on two semi-synthetic datasets, respectively. The latter demonstrates that the proposed scheme is able to return consistent estimation of crowd states, implying its applicability in accurate, visualized and real-time monitoring of crowd states.

To sum up, our main contributions are four-fold.

- We thoroughly analyse the limitations of passive WiFi sensing-based crowd analysis in terms of crowd count and location estimation from a theoretical perspective.
- We propose the ADCA model to sufficiently extract spatial-temporal features from random and sparse passive WiFi sensing data by adopting an image reconstruction approach.
- We infer crowd speeds according to the changes of crowd densities based on density clustering and BiConvLSTM model.
- We produce semi-synthetic passive WiFi sensing datasets for crowd analysis after systematically investigating the rules of active scan, MAC randomization and localization error distributions in practice.

This paper is organized as follows. Section II surveys related literature. Section III analyzes limitations of WiFi-based crowd analysis. Section IV presents the fine-grained crowd analysis scheme. Section V describes the synthesis process and semi-synthetic datasets. Section VI shows the evaluation results, and Section VII concludes the paper.

II. Related Work

In this section, we shall briefly introduce the literatures of vision-based and WiFi-based crowd analysis.

A. Vision-Based Crowd Analysis

According to the granularity of crowd analysis, the vision-based methods can be divided into two categories. On the one hand, early studies mainly focused on the task of crowd counting [25]. The pixel-level and texture-level methods [26], [27] aim to estimate the crowd count in a scenario, rather than identify each individual, and can only achieve overall results. In contrast, the object-level methods [28] can obtain more accurate results by identifying individual, but are only suitable for sparse scenarios, whereas line counting methods [29] count pedestrians crossing a line of interest rather than the entire area, and thus cannot thoroughly handle critical situations. On the other hand, recent studies [1], [5], [6], [7], [9], [30] estimated comprehensive crowd densities instead of crowd counts, but still suffered from the problems of scale variation and limited video quality [8], [31]. Besides, local perspectives generated by a camera cannot be directly transformed into a global bird's eye view (BEV), imposing difficulties in evaluating the criticality of a crowd, and more importantly it is still an open problem to understand cross-camera scenarios [32].

In summary, it can be concluded that, in addition to the traditional limitations, e.g., illumination conditions and computational complexities, existing vision-based methods are also restricted by large surveillance area, high crowd densities and cross-camera collaboration.

B. WiFi-Based Crowd Analysis

Passive WiFi sensing benefits crowd analysis due to its advantages of low cost, large coverage, scalability, and non-intrusive detection [17]. First, most studies were focused on solving the traditional crowd counting problem. In [19], field experiments validated the feasibility of the passive WiFi sensing-based crowd counting approach, but incurred a large higher error rate of over 30%. Similarly, [18] adopted WiFi sniffers with directional antennas and video-based crowd counting method, which resulted in an error rate of around 20%. Second, some recent studies attempted to estimate crowd densities based on the number and location estimates of sniffed devices. In [21] and [33], a coarse-grained density estimation method was implemented by

taking the number of sniffed devices as the crowd count and positions of sniffers with maximum received signal strength (RSS) measurements from corresponding devices as the crowd location. Similarly, crowd heat maps can be initially generated, and then refined in combination with different techniques, such as linear calibrations [18], the prior knowledge of crowd distributions [34], and the accurate crowd count in each grid [16]. Third, some other studies inferred dynamic crowd states (e.g., crowd speeds) according to the location estimates of pedestrians at different time in specific scenarios [16], [22].

To sum up, existing studies on passive WiFi sensing-based crowd analysis have shown feasibility in practice, but are restricted by relatively low accuracy. Particularly, crowd density estimation relies on simple treatments, and lacks standard datasets and quantitative evaluations, while the crowd speed estimation is limited to simple scenarios, and lacks generality.

III. Limitations of Crowd Analysis Using Passive WiFi Measurements

We shall first introduce preliminaries of passive WiFi sensing, then investigate the limitations of crowd analysis using passive WiFi measurements from a theoretical perspective, and finally, summarize the motivations.

A. Preliminaries on Passive WiFi Sensing

The traditional passive WiFi sensing-based crowd analysis involves two key phases, i.e., the crowd sensing and WiFi localization.

First, to effectively sensing crowds, a certain number of WiFi sniffers are uniformly or automatically deployed [35] in a surveillance area. When a sniffer starts to work, it will continuously sense a probe frame from a surrounding mobile device, extract valuable data including the transmitter's MAC address and RSS measurement, and upload this sensing data item (which is appended with the current timestamp) to a server for further processing. However, due to the existence of WiFi-disabled devices, occasional active scans and MAC randomization, only random and sparse passive WiFi measurements can be obtained, making it difficult to estimate accurate crowd counts.

Second, to obtain the location of each sniffed device, WiFi-based localization approaches [36] such as KNN and WKNN can be employed to localize this device by using the RSS measurements from different sniffers during a time window. Advanced treatments, such as data cleaning, data filtering, and location fingerprint optimization, can also be utilized to improve the localization accuracy. However, the environmental dynamics and pedestrian occlusions lead to severe multi-path effects, making it difficult to estimate accurate crowd location.

In summary, the difficulties arising in the above two phases make WDMs to significantly deviate from actual CDMs, with the result that the passive WiFi sensing-based crowd analysis becomes a challenging task. In what follows, we shall investigate how these difficulties deteriorate WDMs from a theoretical perspective.

B. The Influence of Limited Crowd Count Estimation Accuracy on WDMs

Intuitively, passive WiFi measurements captured during a very short period of time provide little useful information about the corresponding crowd count due to their randomness and sparsity. The common solution in the literature is to employ certain time windows, in the sense that a longer time window provides much more information but incurs larger granularity, and vice versa. As such, given a time window, a probabilistic model is presented to characterize the relationship between the crowd count and passive WiFi measurements.

According to [16], a pedestrian, who carries a (mobile) device and walks at a constant speed along a straight line during the time window of t, can be sniffed with the probability of $1-e^{-ct}$, where $c = lim_{\Delta t\to 0}\frac{q(\Delta t)}{\Delta t}$ and $q(\Delta t)$ is the probability that a device can be sniffed during Δt. Let random variable X denote the number of times by which one device is sniffed during t. Supposing a large natural number n, the observed time period $(0,t]$ is divided into n equal segments, namely

$$s_1 = \left(0, \frac{t}{n}\right], s_2 = \left(\frac{t}{n}, \frac{2t}{n}\right], \ldots, s_n = \left(\frac{(n-1)t}{n}, t\right]. \quad (1)$$

Thus, given s_i, the probability that the device is sniffed is $q(\frac{t}{n}) = lim_{\Delta t\to 0}\{1-[1-q(\Delta t)]^{\frac{t}{n\Delta t}}\} = 1-e^{-\frac{ct}{n}}$; with n increasing, the probability that the device is sniffed twice or more tends to be 0, and the probability that the device is not sniffed is simply $1-q(\frac{t}{n}) = e^{-\frac{ct}{n}}$. Since a sniffer independently sniffs a device every time, thus

$$P(X=i) = \binom{n}{i}\left(1-e^{-\frac{ct}{n}}\right)^i\left(e^{-\frac{ct}{n}}\right)^{n-i}. \quad (2)$$

When $n\to\infty$, we can obtain

$$e^{-\frac{ct}{n}} \to 1-\frac{ct}{n}, \frac{\binom{n}{i}}{n!} \to \frac{1}{i!}, \left(1-\frac{ct}{n}\right)^n \to e^{-ct}. \quad (3)$$

Therefore, the probability that the device is sniffed i times in t is

$$P(X=i) = \frac{e^{-ct}(ct)^i}{i!}. \quad (4)$$

It is evident that X follows a Poisson distribution with the intensity of ct, and particularly, the probability that the device can be sniffed is $P(X\geq 1) = 1-P(X=0) = 1-e^{-ct}$, which conforms to the original derivation in [16].

In practice, however, the residence time of a pedestrian is often not strictly equal to t when he/she is just entering in or leaving out a surveillance area, such that we denote the actual residence time of the ith pedestrian as t_i. Meanwhile, considering the fact that a pedestrian may carry multiple devices and MAC randomization makes one device to be counted multiple times, it is straightforward to obtain that the expected number of sniffed devices can be as large as $\sum_{i=1}^{n_c} bct_i$ (where b denotes the expected number of devices carried by one pedestrian [16]), which is unequal to the crowd count n_c. Since the residence time of each pedestrian differs with the passing area, pedestrian's speed and status, and etc., either the global or any local WDM will be significantly different from its corresponding CDM,

and more importantly, the mapping from any pair of WDM to its corresponding CDM might be different. Therefore, directly assigning a WDM to a CDM [21], [33] or uniformly mapping each WDM to its corresponding CDM [18], [19] will attain poor performance.

C. The Influence of Limited Crowd Location Estimation Accuracy on WDMs

It is evident that large localization errors inevitably result in significantly inaccurate crowd density distributions in WDMs, which shall be investigated from a theoretical perspective in the following.

Define $\mathbf{l} = [x, y] \in R^2$ as the true position coordinate of a sniffed device, $\mathbf{r} = (r_1, r_2, \ldots, r_m)^T$ as the vector of mean RSS measurements from m sniffers in dBm, and $\delta\mathbf{r} = (\delta r_1, \delta r_2, \ldots, \delta r_m)^T$ as the measurement noise vector which is assumed to be independent Gaussian with zero means and covariance matrix $\Sigma = diag(\sigma_{r_1}^2, \sigma_{r_2}^2, \ldots, \sigma_{r_m}^2)$ [37], such that the actual RSS measurement, denoted $\hat{\mathbf{r}}$, is equal to $\mathbf{r} + \delta\mathbf{r}$. Let $\mathbf{g}(\cdot) = (g_x(\cdot), g_y(\cdot))^T : R^m \rightarrow R^2$ be the mapping from RSS measurements to a location estimate by using any localization algorithm, where $g_x(\cdot)$ and $g_y(\cdot)$ are the mappings in x-axis and y-axis, respectively. Without loss of generality, the localization algorithm is supposed to satisfy $\mathbf{l} = \mathbf{g}(\mathbf{r})$, and when only RSS measurements are input, we can have $\hat{\mathbf{l}} = \mathbf{g}(\hat{\mathbf{r}})$, namely $\mathbf{l} + \delta\mathbf{l} = \mathbf{g}(\mathbf{r} + \delta\mathbf{r})$ [38], where $\delta\mathbf{l}$ denotes the localization error.

Supposing that $\mathbf{g}(\cdot)$ is differentiable and δx is the localization error in x-axis, after applying the Taylor expansion on g_x around $\mathbf{r}$ and ignoring higher order items, we can obtain

$$x + \delta x \approx g_x(\mathbf{r}) + \sum_{i=1}^{m} \frac{\partial g_x}{\partial r_i}\delta r_i + \frac{1}{2!}\sum_{i=1}^{m}\sum_{j=1}^{m} \delta r_i \delta r_j \frac{\partial^2 g_x}{\partial r_i \partial r_j}, \tag{5}$$

and thus

$$\delta x = \sum_{i=1}^{m} \frac{\partial g_x}{\partial r_i}\delta r_i + \frac{1}{2!}\sum_{i=1}^{m}\sum_{j=1}^{m} \delta r_i \delta r_j \frac{\partial^2 g_x}{\partial r_i \partial r_j}. \tag{6}$$

Since the noises in RSS measurements are independent Gaussian, the expectation of δx is

$$E(\delta x) = \frac{1}{2!}\sum_{i=1}^{m} \sigma_{r_i}^2 \frac{\partial^2 g_x}{\partial r_i^2}. \tag{7}$$

Similarly, the expectation of δy is

$$E(\delta y) = \frac{1}{2!}\sum_{i=1}^{m} \sigma_{r_i}^2 \frac{\partial^2 g_y}{\partial r_i^2}. \tag{8}$$

It follows from (7) and (8) that, the localization errors are mainly dependent on the magnitudes of the derivatives, which differ cross different locations and sniffers, and in general, since different locations incur different multi-path effects, different derivatives will be generated, such that the errors in location estimates as well as WDMs are often non-uniform cross the whole localization area. Consequently, such non-uniform errors in WDMs will result in severely wrong distributions of crowd density and make it hard to identify critical situations.

D. The Motivations of Our Work

In light of the above analyses, the motivations of our crowd density and speed estimation methods are presented as follows.

1) Crowd Density Estimation: Deficiencies of existing WiFi-based crowd density estimation motivate us to make improvement from the following two aspects. First, both simple crowd counts and coarse heat maps generated by the average count in each local region cannot precisely depict crowd distributions, which motivates us to work on informative density maps. Second, original WDMs or those calibrated by a uniform mapping significantly deviate from the corresponding CDMs due to the randomness and sparsity of passive WiFi sensing and non-uniform localization errors, which motivates us to develop an advanced CDM regression method. To be specific, by making use of the similarities between image reconstruction and the recovery of CDMs from WDMs, such as de-noising, compensation and correction, the state-of-the-art deep learning techniques [39], [40] can be utilized to effectively extract spatial-temporal features from WDMs via convolutions, sufficiently enrich measurements from random and sparse WDMs via autoencoder, and relieve the influence of non-uniform localization errors cross the whole surveillance area via attention mechanism and structural loss function, such that a fine-grained mapping between pixels or patches in WDMs and those in corresponding CDMs can be established.

2) Crowd Speed Estimation: It is crucial to grasp the speeds of all crowds for detecting possible emergencies, but due to MAC randomization, it is hard to track any single pedestrian, such that crowd speed estimation is a challenging task. Alternatively, according to CDMs returned by the above crowd density estimation, one can track some identifiable crowds, so as to estimate their speeds. Specifically, pedestrians often gather in different groups, forming crowds, with the result that the corresponding CDM includes some patches with high densities, so that one can identify the LHDCs in this CDM using any density clustering algorithm and further easily estimate their speeds by detecting their changes in space and time.

IV. Fine-Grained Crowd Analysis Scheme

In this section, we shall present the overview and design details of the proposed fine-grained crowd analysis scheme.

A. Overview

As shown in Fig. 1, the proposed scheme includes three modules, i.e., the WDM generation module, crowd density regression module and speed estimation module. First, a fixed Gaussian kernel function [41] is utilized to generate coarse WDMs based on the traditional sensing and localization results. Second, the ADCA model is designed to reconstruct CDMs from WDMs, an attention mechanism and a fusion loss is applied to improve its feature extraction and local consistency learning capabilities and additionally, to combat with the difficulty in labeling CDMs,

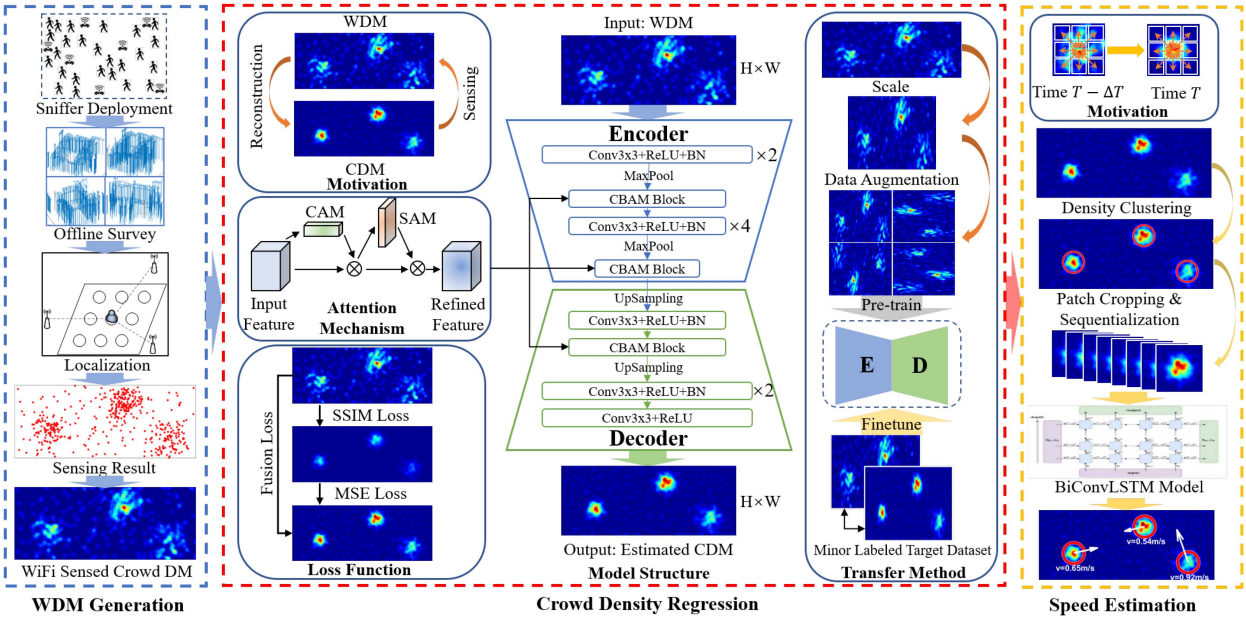


Fig. 1. Overview of the proposed fine-grained crowd analysis scheme.

a transfer learning method is proposed based on fine-tuning the pre-trained model. Third, given a CDM, SCDMPs are constructed based on the LHDCs identified by density clustering, and further utilized to regress the speed vector of each LHDC by on a BiConvLSTM model.

B. The Generation of WDM

Assume that the shape of a relatively large surveillance area is rectangular, because any shape can be included in a bounding box. First of all, we need to transform continuous physical coordinates to discrete image coordinates in pixels, and for ease of processing, a 1 m × 1 m square is transformed to a single pixel in a WDM throughout this paper. By letting n_w be the number of sniffed devices, $\hat{\mathbf{L}} = \{\hat{\mathbf{l}}_1, \hat{\mathbf{l}}_2, \ldots, \hat{\mathbf{l}}_{\mathbf{n_w}}\}$ be the set of their location estimates and $\mathbf{P} = \{p_1, p_2, \ldots, p_{n_w}\}$ be the pixel set transformed from $\hat{\mathbf{L}}$, an image I can be generated as follows

$$I(p) = \sum_{i=1}^{n_w} \delta(p - p_i), \tag{9}$$

where δ is the impulse function. Then, the WDM can be obtained through convolving and smoothing I with a Gaussian kernel function with the size ks, i.e.,

$$WDM(p) = I(p) * G_\sigma(p), \tag{10}$$

$$G_\sigma(p) = \exp\left(-\frac{||p - p_i||^2}{2\sigma^2}\right), \tag{11}$$

where $G_\sigma(p)$ is the Gaussian kernel function and σ is the standard deviation. The parameters ks and σ reflect the impact of a sniffed device on surrounding density values, and are empirically set as $10 < ks < 20$ (odd) and $\sigma = 1$. As a result, a WDM with the size of $H \times W$ is obtained, where H and W are the height and width of the WDM, respectively.

C. ADCA-Based CDM Regression

Considering the 2-D nature of CDM regression, the ADCA model is designed based on the deep convolutional autoencoder [42], as shown in the middle of Fig. 1.

1) Overall Architecture: The model consists of an encoder for extracting valid latent features and a decoder for recovering CDMs. We combine the 3×3 convolutional layer, activation function and batch normalization (BN) as a base block. The convolutional layer can preserve spatial locality by sharing weights, and the latent representation of kth feature map is given by

$$\mathbf{h}^k = ReLU(\mathbf{h^{k-1}} * \mathbf{W}^k + \mathbf{b}^k), \tag{12}$$

where the bias $\mathbf{b}^k$ is broadcasted to the whole map, $ReLU$ activation function is used for avoiding gradient disappearance, $*$ denotes the 2-D convolution, and $\mathbf{W}^k$ is the flip operation over dimensions of weights in the kth layer. The BN is added after each convolutional layer to improve training speed and reduce overfitting [43]. In total, ten base blocks are employed to extract latent features at different scales. Two max-pooling layers are embedded into the encoder to eliminate redundant features and reduce computations, and two bilinear interpolation upsampling layers are accordingly utilized in the decoder to recover the size. As a result, we keep the outputted CDM to have the same size as the original WDM, and ensure that the density value of each pixel is greater than or equal to 0 by the last ReLU. To sum up, the deep architecture enables the model with a strong learning capability, and the full-convolution design allows accepting inputs with an arbitrary size, facilitating the model to be applied in new scenarios without structural adjustment.

2) Attention Mechanism: Three lightweight convolutional block attention modules (CBAMs) [44] are added into the model so as to simultaneously combine the channel and spatial domain

attentions. The former focuses on the meaningful information in input features, identifies and amplifies the importance of a single feature map extracted by different convolution kernels. The latter concerns more on the location information of the crowd density distributions. As a result, each of the three CBAMs can generate attentional maps in both channel and spatial dimensions, which are then sequentially multiplied with the original feature map to adaptively enhance features at different locations and scales.

3) Fusion Loss Function: Most existing image regression methods use the pixel-wise mean square error (MSE) loss to train their models, which relies on the pixel-dependent assumption and ignores the local consistency of DMs [45]. As such, we fuse the MSE loss with the structural similarity index measure (SSIM) which is employed to evaluate local pattern consistency. First, the pixel-wise MSE loss is defined as

$$L_{mse} = \frac{1}{N}||F(\mathbf{X};\mathbf{\Theta}) - \mathbf{Y}||_2^2, \tag{13}$$

where $F(\mathbf{X};\mathbf{\Theta})$ denotes the CDM estimated by the ADCA model with parameters $\mathbf{\Theta}$, $\mathbf{X}$ and $\mathbf{Y}$ represent the WDM and ground-truth (GT) CDM, respectively, and $N = HW$ is the number of pixels in a DM. Second, the SSIM [46] between the estimated and GT CDMs is calculated as follows

$$SSIM(p) = \frac{(2\mu_F(p)\mu_Y(p) + C_1)(2\sigma_{FY}(p) + C_2)}{(\mu_F^2(p) + \mu_Y^2(p) + C_1)(\sigma_F^2(p) + \sigma_Y^2(p) + C_2)}, \tag{14}$$

where p is an arbitrary pixel in DMs, $\mu_F(p)$ and $\sigma_F^2(p)$ are the local mean and variance of F, respectively, $\sigma_{FY}(p)$ is the local covariance, and C_1 and C_2 are small constants used to avoid dividing by 0. Since the SSIM is positively correlated with local consistency, we define the loss as

$$L_{ssim} = 1 - \frac{1}{N}\sum_p SSIM(p). \tag{15}$$

Finally, by weighting the above two losses, the fusion loss function is

$$L_{fusion} = L_{mse} + \alpha L_{ssim}, \tag{16}$$

where α is the weight to balance L_{mse} and L_{ssim}, and is empirically set as 0.1 in experiments. Intuitively, L_{ssim} is used to guarantee the shape similarity between crowd density distributions, while L_{mse} is used to improve the accuracy of the density estimation.

4) Transfer Method: In practical applications, labeling global and accurate CDMs is very hard due to the high cost, so that it would be attractive if the trained ADCA model could be used in a target environment. However, the model learns scenario-specific knowledge to boost the estimation performance in a specific environment, and thus cannot be directly generalized to other environments. As such, based on the idea of reducing the gap between source and target domains, we propose to transfer a model pre-trained in semi-synthetic datasets to any actual dataset with a few labels by means of spatial transformations.

1) Data Scaling: Due to the differences in the geometric size and coordinate setting, crowds in the source and target datasets have different distribution ranges. Define $\mathbf{S}$ as the scaling vector, and thus the location of a pedestrian in the source dataset $\mathbf{l_s} = [x_s, y_s]^T$ can be transformed to $\mathbf{l_t} = \mathbf{l_s}\mathbf{S}$ in the target dataset, i.e., $\mathbf{l_t} = [x_t, y_t]^T$, where $\mathbf{S} = [\frac{x_t^{\max} - x_t^{\min}}{x_s^{\max} - x_s^{\min}}, \frac{y_t^{\max} - y_t^{\min}}{y_s^{\max} - y_s^{\min}}]$.

2) Data Augmentation: To further mitigate the difference between crowd distributions in the source and target datasets, the euclidean transformation is used to augment the scaled dataset. Let $\mathbf{R}$ be a rotation matrix, and thus the location of a pedestrian in the scaled dataset $\mathbf{l_t}$ can be transformed to $\mathbf{l_a} = \mathbf{R}\mathbf{l_t}$ in the augmented dataset. Other transformations, such as reflection, can also be used for augmentation.

3) Fine-tuning: The augmented dataset can be directly used to pre-train a model, but the performance may be poor due to the inevitable complex and scenario-specific factors, such as crowd distributions and RSS measurement noises. Therefore, using a few labeled data, the pre-trained model can be fine-tuned via fixing the encoder and using a reduced learning rate, so as to improve its robustness.

5) Discussions on the Improvement: The advantages of our treatments mainly lie in the following two aspects.

1) Fine-grained crowd counts are estimated by establishing local non-linear mappings. Supposing that the residence time of each pedestrian is t, the expected number of sniffed devices is $E(n_w) = bctn_c = an_c$, where $a = bct$. In [21], [33], $E(n_w)$ is directly used to approximate the crowd count n_c, resulting in large errors, whereas in [16], [18], [19], [34], a constant coefficient a' is estimated to approximate $\frac{1}{a}$, such that the crowd count can be approximated by $a'E(n_w)$. However, the uniform coefficient a' will incur large errors in the resulting crowd counts of most local regions due to the spatial diversity of the pedestrian residence time. In contrast, we transform original WiFi measurements into an image-like WDM to maximally reserve the spatial diversity, and establish local non-linear mapping set $F(\mathbf{WDM};\mathbf{\Theta})$ in pixel level (or patch level) through the training process $\arg\min_{\Theta}[L_{fusion}(F(\mathbf{WDM};\mathbf{\Theta}), \mathbf{CDM})]$, which implicitly considers different t in different local regions.

2) The localization errors are mitigated with the model. Most existing methods in [16], [18], [21], [33] did not handle the coarse crowd location estimation, whereas the method in [34] tried to improve location estimation by using two independent global Gaussian mixture models (GMMs) based on prior crowd distributions in x-axis and y-axis directions, but such simple treatments do not make obvious difference. In contrast, our model connects a pixel in a WDM to its surrounding and even more distant pixels in the corresponding CDM, weighs the importance of each connection that reflects the non-uniform localization error, and finally mitigates the influences of different localization errors through measuring the local inconsistency by the fusion loss involving SSIM in the training process.

D. BiConvLSTM-Based Crowd Speed Estimation

We shall briefly describe the key components of the proposed crowd speed estimation method.

1) The Identification of LHDCs: A density based spatial clustering algorithm, termed DM-DBSCAN, is proposed to identify LHDCs from an estimated CDM based on the well-known DBSCAN [47]. The DM-DBSCAN involves two major

modifications in comparison with the original DBSCAN. First, every pixel in a given CDM functions as a data point and the pixel coordinate defines its spatial coordinate, such that the pixel distance threshold ϵ_p is defined to identify the neighborhood of a point based on the euclidean distance between pixel coordinates. Second, deciding whether a point is a core point by thresholding the sum of the densities of all the points in the neighborhood, such that the density threshold of any point is defined as $MinDes$. Finally, the central pixel coordinate of each LHDC is obtained by rounding the mean coordinate among each cluster.

2) The Construction of SCDMP: Given the central pixel coordinate of a LHDC, a CDM patch with the size of $N_p \times N_p$ (where N_p is the height or width in pixel of the patch) is cropped form the corresponding location of the CDM. Set the sequence length as sl, CDMs at previous $sl-1$ moments are cropped at the same location, and all patches are stacked together to be an $sl \times N_p \times N_p$ SCDMP. In the case of overlapping with the boundary, the corresponding elements are filled with 0 to ensure dimensional consistency. The usage of CDM patches helps to remove redundant information and reduce calculations.

3) The Proposed BiConvLSTM Model: Considering the fact that the forward and reverse density changes of a crowd are closely related to the speed of this crowd, the BiConvLSTM model is adopted to capture the spatial correlation in relation to a 2-D CDM patch via convolution and balance the importance among patches at distant and recent moments. Specifically, a lightweight BiConvLSTM model is designed by including two ConvLSTM layers consisting of 16 3×3 convolutional kernels for extracting sequential features and a fully-connected layer for regressing speed vectors. The input is the SCDMP associated with a LHDC, and the output is the speed vector $\hat{\mathbf{v}} = (\hat{v}_x, \hat{v}_y)^T$ of the LHDC. As for the training, the traditional MSE loss which actually equals to the speed error is adopted and the Adam optimizer is utilized. Finally, the model trained on the source dataset can be directly transferred to a target dataset by using the SCDMPs with the same size due to the similar crowd densities and location changes of different LHDCs in two datasets.

V. Semi-Synthetic Datasets for Passive WiFi Sensing-Based Crowd Analysis

In this section, we shall first investigate the key rules governing passive WiFi sensing from real-world datasets, and then present how to produce semi-synthetic datasets by emulating passive WiFi sensing in existing pedestrian tracking datasets.

A. Learning From Real-World Public Datasets

Prior to emulating passive WiFi sensing, it is a prerequisite to understand how active scans are triggered, how MAC addresses are randomized, and how localization errors are distributed. To this end, we attempt to learn the key rules from several real-word public datasets.

1) Active Scan and MAC Randomization Rules: To find out the rules of determining time intervals between two adjacent active scans and randomizing a MAC address, a recent dataset [48] for MAC de-randomization is adopted. This dataset contains WiFi probe frames sent by 22 mobile devices with 6 major brands and under the standby status in a shielded environment, and a thorough analysis concludes three modes of triggering active scans and three modes of MAC randomization, respectively.

The modes of triggering active scans: 1) Fixed Interval: In this mode, one device triggers any two active scans with a fixed interval, so that given a specific device emulated in the semi-synthetic dataset, its interval is randomly sampled from a uniform distribution, i.e., $I_{fi} \sim U(l_{fi}, u_{fi})$, whose upper bound u_{fi} and lower bound l_{fi} are determined by the union of all fixed intervals. *2) Periodic & Ascending Interval:* In this mode, the interval gradually ascends within a specific range and repeats again after reaching the maximum of the range, which is emulated by designing an algorithm to produce a sequence of increasing time intervals. *3) Random Interval:* In this mode, the interval appears to be random within a range, which is emulated by sampling from a uniform distribution $U(l_{rr}, u_{rr})$.

In addition, a small white Gaussian noise is added to each emulated interval to reproduce sensing delays and performance limitations of sniffers. Some typical and emulated intervals are shown in Fig. 2, which intuitively demonstrates the effectiveness of our emulation for each mode. In semi-synthetic datasets, a discrete probability distribution determined by the proportion of each mode device is used to decide which mode an emulated device belongs to.

The modes of MAC randomization: 1) Non-Randomization or Long Period: In this mode, one device either uses its real MAC address or only uses one randomized MAC address during a relatively long period of time, such that corresponding passive WiFi sensing data can be easily categorized to one common device. *2) Randomization at Every Active Scan:* In this mode, every active scan uses a newly randomized MAC address. *3) Randomization in Each Channel:* In this mode, every active scan uses different randomized MAC addresses across different channels.

Considering the fact that only few old devices adopt the first mode and the third mode is essentially the same as the second mode given sniffers working in one channel, only the second mode is emulated in producing semi-synthetic datasets.

2) Empirical Distributions of Localization Errors: To establish the relationship between error distributions and corresponding scenario configurations, localization errors are investigated by using five public datasets [36], [49], i.e., LAB, OFFICE, CETC, HCXY and SYL, and one self-collected dataset WiCAM (see Section VI-A). The training data in each dataset is used to generate location fingerprint database. Then, the testing data in each dataset is used to perform localization by using the KNN algorithm, in which the value of k is tuned according to different datasets to achieve optimal performance. Finally, the means and standard deviations of localization errors in both x-axis and y-axis, denoted μ_x, σ_x, μ_y and σ_y, are calculated, so as to better respectively investigate the localization error characteristics in two axes. As shown in Fig. 3, it can be found that the histograms of localization errors are in good consistence with the corresponding empirical Gaussian probability density functions (PDFs) in three datasets.

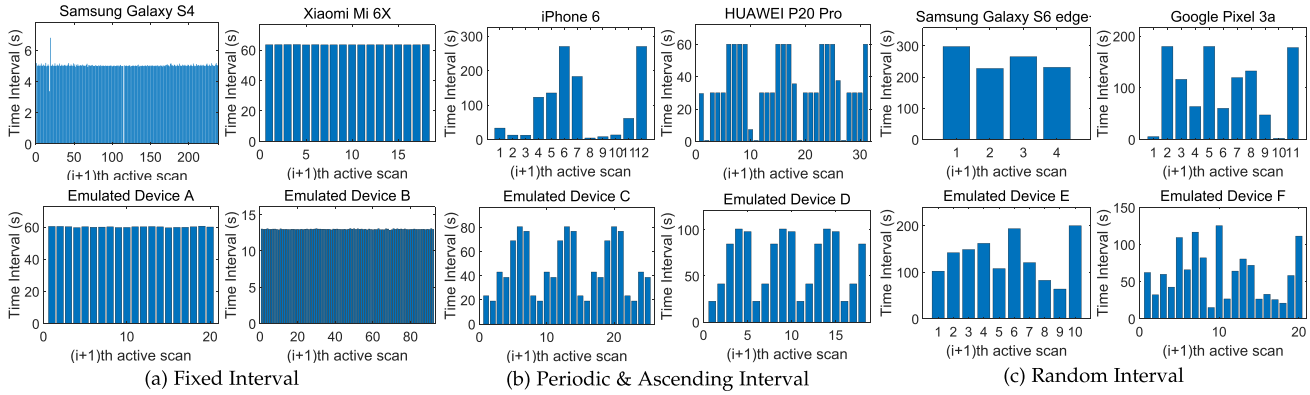


Fig. 2. Three different modes of triggering active scans (real devices in the first row and emulated devices in the second row).

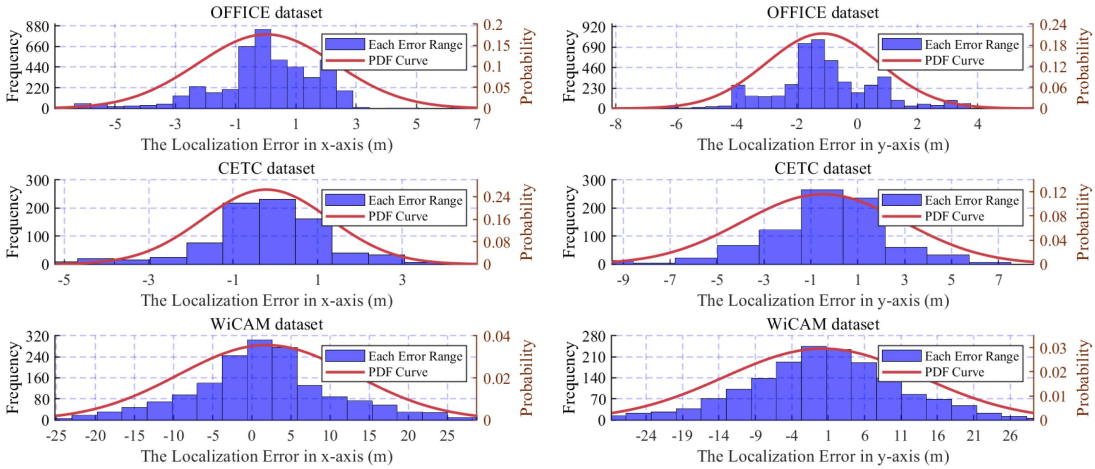


Fig. 3. Histograms of localization errors and the corresponding empirical PDFs of three datasets.

TABLE I
PARAMETERS OF LOCALIZATION ERROR DISTRIBUTIONS AND KEY SETUPS OF THE TESTBED IN EACH DATASET (THE EMULATED VALUES ARE IN BOLD)

Dataset	μ_x	σ_x	μ_y	σ_y	Area	Δx	Δy	AP Number	AP Density
LAB	0.77	1.93	−0.13	1.32	$\approx 81m^2$	$12.00m$	$4.80m$	4	$1/20.25m^2$
OFFICE	0.02	2.28	−1.14	1.87	$\approx 350m^2$	$27.00m$	$16.00m$	6	$1/58.33m^2$
CETC	−0.21	1.50	0.48	3.43	$\approx 1,800m^2$	$16.20m$	$56.80m$	26	$1/69.23m^2$
HCXY	−0.43	3.35	−0.09	2.27	$\approx 3,600m^2$	$117.32m$	$41.00m$	56	$1/64.29m^2$
SYL	−0.12	5.62	−0.59	2.48	$\approx 2,600m^2$	$63.40m$	$28.98m$	23	$1/113.04m^2$
WiCAM	1.86	11.24	0.27	13.48	$\approx 10,000m^2$	$74.20m$	$131.20m$	14	$1/714.29m^2$
WiDIA	**−0.80**	**4.00**	**0.60**	**2.50**	$\approx 970m^2$	$59.55m$	$16.34m$	**15(virtual)**	$1/64.67m^2$
WiATC	**1.20**	**6.00**	**−1.00**	**5.00**	$\approx 4,500m^2$	$88.84m$	$51.48m$	**50(virtual)**	$1/90.00m^2$

More detailed information is summarized in Table I, where the area specifies the coverage area of each testbed; Δx and Δy denote the maximum absolute difference between the coordinates in x-axis and y-axis, respectively; AP number is the number of APs/WiFi sniffers deployed, and AP density is the ratio of the AP number to the area. It can be found that, the localization errors in both axes do not scale with the area, but often decrease with the AP density increasing and Δx or Δy decreasing, which would guide us in determining the emulation parameters given the scenario configurations of existing pedestrian tracking datasets.

B. Emulations in the Semi-Synthetic Datasets

In order to produce semi-synthetic datasets, the following two pedestrian tracking datasets [50], [51] are employed to provide the precisely labelled trajectory of every pedestrian: the DIAMOR dataset, which was collected by using laser range finders in two large straight corridors connecting the Diamor shopping centre in Osaka, Japan, and the ATC dataset, which was collected by using 3-D range sensors in part of the ATC shopping and business center in Osaka, Japan. Both datasets

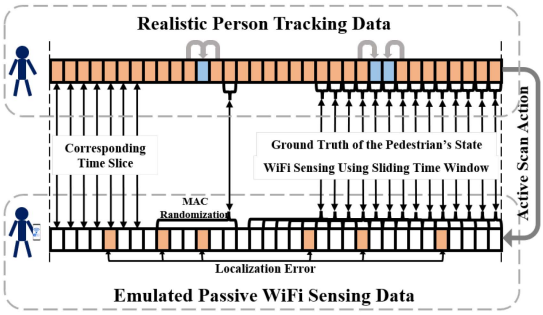


Fig. 4. Emulation of passive WiFi sensing in the semi-synthetic datasets.

include items of timestamp, pedestrian ID, position, speed, and etc.

To synthesize the datasets, a divide-and-conquer approach is employed to emulate corresponding passive WiFi sensing data with the real data of each pedestrian, as shown in Fig. 4. First, the GT positions and speeds of crowds are generated by: 1) divide the whole time into slices with a fixed length, denoted δt; 2) in each slice of time, average the position, velocity and direction of every pedestrian, so as to reduce the influence of noises, or interpolate corresponding data of any pedestrian who was not detected; 3) calculate the mean speed vector of every LHDC detected from a given CDM by using the velocities and directions of all pedestrians residing in the central point of this LHDC. Second, the passive WiFi sensing data is generated by: 1) extract GT positions of each pedestrian by his/her ID in every slice of time; 2) for each pedestrian, generate a list of time slices in which an active scan is triggered according to the emulation method in Section V-A1, and remove the data in the time slices in which the active scan is not triggered; 3) randomize the pedestrian ID in each of the remaining slices to emulate MAC randomization; 4) sample localization errors based on the emulated error distribution, and add them with GT positions in every slice of time.

Based on above two algorithms, the passive WiFi sensing data can be emulated for both DIAMOR and ATC datasets, resulting in the semi-synthetic datasets, termed WiDIA and WiATC, respectively. Besides, a densified dataset, termed WiDATC, is produced based on WiATC.

WiDIA: A total of 37,750 seconds of pedestrian tracking data collected in 2 days [51] are synthesized. By using the sliding time window with the window size $\Delta T = 30$ s and the step $\delta t = 1$ s and the Gaussian kernel with $ks = 11$ and $\sigma = 1$, 37,692 pairs of WDM and CDM with the size of $H \times W = 24 \times 68$ are derived. For the purpose of estimating crowd speeds, DM-DBSCAN is executed with $\epsilon_p = 2.9$ and $MinDes = 5$ to identify a total of 4,937 LHDCs, and corresponding pairs of SCDMP and speed vector.

WiATC: A total of 92,160 seconds of pedestrian tracking data collected in 6 days [50] are synthesized. The same sliding time window and Gaussian kernel as in WiDIA are used to obtain 91,464 pairs of WDM and CDM with the size of $H \times W = 64 \times 96$. Likewise, DM-DBSCAN with a strict $MinDes = 9$ is executed due to the relatively high crowd densities in the ATC dataset, so as to identify a total of 7,861 LHDCs and corresponding pairs of SCDMP and speed vectors.

WiDATC: By stacking the aligned 4 hours data per day of the ATC dataset into one hour, we obtain 22,866 pairs of WDM and CDM with the same size as WiATC.

In summary, the above emulation of passive WiFi sensing based on insightful analyses of real-world datasets conducted in the previous subsection, in combination with the real-world pedestrian tracking datasets, makes it both suitable and feasible to support further crowd analyses. However, since the emulation takes into account noises in localization results, it is unnecessary to directly emulate the multi-path propagation of WiFi signals.

VI. Evaluation

In this section, extensive experiments are conducted to verify the effectiveness and practicality of the proposed crowd analysis scheme based on three semi-synthetic and a real-world datasets.

A. WiCAM: A Realistic Crowd Analysis Dataset

WiCAM was collected on a square with the area of around $10,000\,\mathrm{m}^2$, which is located between several teaching buildings and where several pedestrian walkways and roads connect these buildings. Fourteen Raspberry Pi 3B+ are customized as WiFi sniffers, and uniformly deployed on the walkways, roads and open spaces, as the surveillance area, with the total area of $4,000\,\mathrm{m}^2$. In the offline stage, a path-based fingerprint collection method [52] was employed to quickly collect RSS fingerprints from over 300 locations. In the online stage, after preprocessing the sensing data in the sliding time window with $\Delta T = 30\,\mathrm{s}$ and $\delta t = 1\,\mathrm{s}$, the RSS fingerprint vector associated with one MAC address is produced and further used for localization by the KNN algorithm with $k = 3$. Meanwhile, five time-synchronized smartphones were deployed as cameras to cover the whole surveillance area, such that crowd counts are manually labelled given one video frame per second. At last, totally $2,280\,\mathrm{s}$ sensing data including a peak time, i.e., after classes, and corresponding crowd counts were obtained. Besides, the Gaussian kernel with $ks = 15$ and $\sigma = 2$ is used to derive 2,280 WDMs with the size of $H \times W = 140 \times 80$, and only 60 randomly selected CDMs are labeled due to the difficulty in labeling. DM-DBSCAN with $\epsilon_p = 2.9$ and $MinDes = 3.8$ is employed to identify LHDCs.

B. Baselines and Metrics

Four existing methods are adapted as baselines to estimating crowd densities on all datasets: 1) Early AP Location-based method (APL) [21], [33] simply assigns the position of a nearest sniffer to that of a sniffed device; 2) Calibration Factor-based method (CF) [18] corrects WDMs by multiplying a calibration factor obtained through fitting the number of sniffed devices to the crowd count using training data; 3) Priori knowledge (CF+PK) [34] further corrects the density distribution by two GMMs learned from all pedestrians' real positions; 4) Sequential Filtering based method (SF) [16] partitions the surveillance area into $4\,\mathrm{m} \times 4\,\mathrm{m}$ grids and estimates the crowd density of each grid. Besides, the traditional stacked autoencoder (SAE)

TABLE II
CROWD DENSITY ESTIMATION RESULTS OF ALL METHODS ON THREE SEMI-SYNTHETIC DATASETS

(a) Results on the WiDIA Dataset

Method	PSNR ↑	NSSIM ↑	$\mathbf{RMSE_p}$ ↓	$\mathbf{MAE_c}$ ↓	$\mathbf{MSE_c}$ ↓	PAE ↓	PSE ↓
APL [21], [33]	21.948	0.425	0.089	16.927	409.448	1.135	5.221
CF [18]	28.313	0.309	0.039	4.251	29.064	0.629	1.098
CF+PK [34]	28.290	0.327	0.040	4.334	30.062	0.611	1.113
SF [16]	28.423	0.226	0.039	4.251	29.064	0.658	1.160
SAE Model	29.931	0.470	0.033	7.228	68.494	0.514	0.949
ADCA with L_{mse}	30.010	0.298	**0.032**	3.532	20.183	0.482	0.721
ADCA with L_{ssim}	**30.076**	**0.492**	0.033	3.679	21.000	0.394	**0.535**
ADCA w/o CBAM	30.069	0.461	0.033	4.994	34.648	0.419	0.639
ADCA	30.075	0.475	0.033	**3.500**	**19.122**	**0.384**	0.538

(b) Results on the WiATC Dataset

Method	PSNR ↑	NSSIM ↑	$\mathbf{RMSE_p}$ ↓	$\mathbf{MAE_c}$ ↓	$\mathbf{MSE_c}$ ↓	PAE ↓	PSE ↓
APL [21], [33]	26.303	0.877	0.065	39.319	2945.161	1.876	21.374
CF [18]	30.868	0.832	0.033	5.746	62.338	0.845	3.341
CF+PK [34]	30.424	0.840	0.035	6.966	92.911	1.190	8.324
SF [16]	30.818	0.796	0.034	5.746	62.338	0.943	4.253
SAE Model	33.138	0.887	0.025	9.170	125.439	0.778	3.286
ADCA with L_{mse}	33.058	0.884	0.025	**5.145**	50.030	0.679	2.041
ADCA with L_{ssim}	*Can Not Converge*						
ADCA w/o CBAM	33.898	0.905	0.023	7.132	77.883	0.538	1.513
ADCA	**34.101**	**0.907**	**0.022**	5.472	**50.023**	**0.505**	**1.320**

(c) Results on the WiDATC Dataset

Method	PSNR ↑	NSSIM ↑	$\mathbf{RMSE_p}$ ↓	$\mathbf{MAE_c}$ ↓	$\mathbf{MSE_c}$ ↓	PAE ↓	PSE ↓
APL [21], [33]	18.513	0.874	0.212	158.772	36532.050	6.802	211.114
CF [18]	23.443	0.831	0.084	12.413	265.512	2.071	16.143
CF+PK [34]	22.472	0.817	0.094	15.743	424.682	4.069	75.684
SF [16]	23.054	0.797	0.089	12.413	265.512	2.610	26.345
SAE Model	26.874	0.912	0.051	11.868	225.214	1.591	10.668
ADCA with L_{mse}	27.395	0.920	0.047	11.145	189.634	1.170	4.845
ADCA with L_{ssim}	27.891	**0.922**	0.046	9.818	194.674	1.157	4.770
ADCA w/o CBAM	27.320	0.918	0.048	13.821	302.767	1.214	5.990
ADCA	**27.961**	0.921	**0.044**	**9.501**	**152.607**	**0.985**	**3.873**

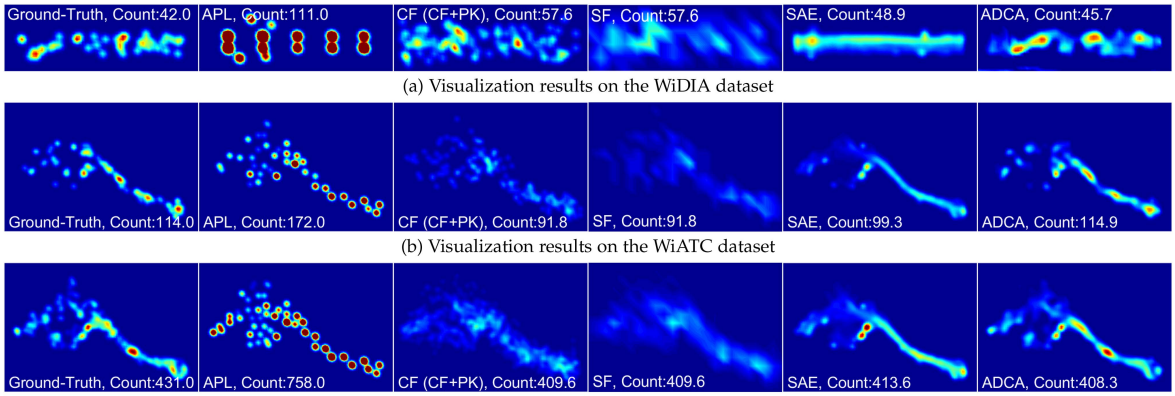
model, ADCA without the attention mechanism (ADCA w/o CBAM), trained with the MSE loss (ADCA with L_{mse}) and SSIM loss (ADCA with L_{ssim}) are also taken for comparison. All semi-synthetic datasets are divided in a Hold-out manner, with 50% as training data and the remainder as test data. All deep learning models are trained by Adam in 300 epoches with the learning rate of 0.001.

Due to the absence of crowd speed estimation methods suitable for passive WiFi sensing, several commonly used sequential models including RNN, LSTM and ConvLSTM with a similar three-layer structure are adopted as baselines to confirm the effectiveness of our method. All models are trained by Adam in 200 epoches with the learning rate of 0.001. SCDMPs are constructed with $N_p = 17$ and $sl = 10$ by default, and a similar partitioning scheme is adopted.

The following seven commonly used metrics are utilized for evaluating crowd density estimation: two image metrics including peak signal-to-noise ratio ($PSNR$) and Normalized SSIM ($NSSIM$) for measuring the overall quality; traditional pixel-wise root MSE ($RMSE_p$); two crowd counting metrics including mean absolute error (MAE_c) and MSE (MSE_c) for measuring the global crowd density; two composite indicators including patch absolute error (PAE) and patch square error (PSE) [5] for simultaneously measuring the density and its spatial distribution by dividing a CDM into 24 patches. As for speed estimation, the mean absolute percentage error of velocity ($MAPE_v$), the MAE of motion angles (MAE_d) and the RMSE of speed vector ($RMSE_s$) are utilized to measure the accuracy in terms of the magnitude, direction and synthesis, respectively.

C. Validating the Accuracy of Crowd Density Estimation

The crowd density estimation results of all methods on all semi-synthetic datasets are listed in Table II. In combination with the analyses in Section III, it can be found that: APL results in the worst performance, probably due to the ignorance of the limitations of passive WiFi sensing; CF and SF attempt to mitigate the influence of the limitation on crowd count estimation, but are still restricted by the inaccurate crowd location estimation; CF+PK further uses the priori knowledge in one scenario to correct location estimates, but is only effective for small and regular scenarios; SAE appears to mitigate the influence of these two limitations to a certain extent, but the limited learning capability leads to difficulties in extracting spatial-temporal features between 2-D DMs; ADCA outperforms all baselines by overcoming the aforementioned challenges, and has the best performance



(a) Visualization results on the WiDIA dataset

(b) Visualization results on the WiATC dataset

(c) Visualization results on the WiDATC dataset

Fig. 5. Visualized crowd density estimation of a randomly selected test sample in each dataset.

TABLE III
ABLATION STUDIES OF CROWD SPEED ESTIMATION ON TWO SEMI-SYNTHETIC DATASETS

(a) Speed Estimation Results by Different Sequential Models

Dataset	WiDIA			WiATC		
Sequential Model	$\mathbf{RMSE_s}\downarrow$	$\mathbf{MAPE_v}\downarrow$	$\mathbf{MAE_d}\downarrow$	$\mathbf{RMSE_s}\downarrow$	$\mathbf{MAPE_v}\downarrow$	$\mathbf{MAE_d}\downarrow$
RNN	0.2585	25.96%	19.45°	0.1667	44.60%	40.49°
LSTM	0.2506	24.55%	18.86°	0.1517	40.45%	35.65°
ConvLSTM	0.2398	23.51%	18.00°	0.1423	38.07%	33.36°
BiConvLSTM	**0.2319**	**22.85%**	**17.83°**	**0.1365**	**36.97%**	**31.89°**

(b) The $RMSE_s$ with Different Sequence Length sl

Sequence Length	2	6	10	14	18
WiDIA	0.2600	0.2445	0.2319	0.2344	0.2369
WiATC	0.1703	0.1476	0.1365	0.1364	0.1391

(c) The $RMSE_s$ with Different Patch Size $N_p \times N_p$

Patch Size	9×9	13×13	17×17	21×21	25×25
WiDIA	0.2751	0.2469	0.2319	0.2238	0.2229
WiATC	0.1642	0.1502	0.1365	0.1310	0.1313

on the fine-grained crowd density estimation. Moreover, ablation studies in the table show that the attention mechanism and fusion loss are effective and necessary. Particularly, it is possible to achieve better performance using L_{ssim} only, which leads to a drastic fluctuation in loss values or even failure to converge on complex WiATC dataset. In contrast, L_{fusion} retains advantages of both L_{mse} and L_{ssim}, and can optimize both densities of local pixels and the global structural consistency of a CDM.

The visualized crowd density estimation results of a randomly selected test sample in each dataset are shown in Fig. 5, and intuitively confirm the above discussions. Moreover, additional conclusions can be drawn: the accuracy of CDMs estimated by APL depends on the number and location of sniffers; the CDMs estimated by CF and CF+PK evidently deviate from the GT CDMs, in that the existing intensive crowds cannot be reflected, which is probably caused by the significant localization errors; SF uses a simple density representation which makes it unable to distinguish dense crowds; SAE can learn crowd distributions and smooth the density estimation, but fails to distinguish different densities; the CDMs estimated by ADCA has the most similar shapes with the GTs and can effectively distinguish different densities.

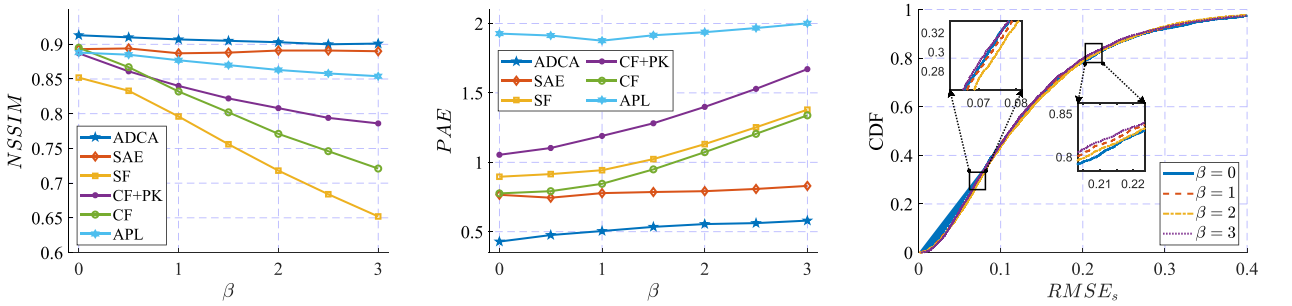
D. Validating the Accuracy of Crowd Speed Estimation

The crowd speed estimation is only tested on WiDIA and WiATC datasets as shown in Table III(a), because the WiDATC dataset virtually puts the pedestrians at different times together and incurs conflicts of crowds moving in different directions. As can be seen, various sequential models can obtain acceptable speed estimation results. Specifically, the performance increases from RNN, LSTM, ConvLSTM to BiConvLSTM, revealing that: LSTM can reasonably balance the importance of features at distant moments with that of recent moments; convolutions facilitate the effective capturing of spatial correlations among 2-D patches; the combination of forward and backward density changes is helpful to improve speed estimation; the proposed model, i.e., BiConvLSTM, achieves a competitive estimation error of about 23% (37%) in velocity and less than 18° (32°) in direction on the WiDIA (WiATC) dataset, confirming the feasibility and potential of our speed estimation idea.

To further investigate the influence of different parameters on speed estimation, another two ablation studies are conducted, as listed in Table III(b) and (c). It turns out that, on both datasets, the performance increases with the sequence length sl increasing, but slightly degrades after sl is beyond 14. Similar phenomenon can be observed in regards to the patch size N_p. However, since the large values of sl and N_p lead to high training costs, it is empirically suggested that $sl = 10$ and $N_p \times N_p = 17 \times 17$ in the experiments.

E. Investigating the Influence of Localization Errors

In practical applications, the diversities of scenarios (e.g., the size, obstacles and multi-path effects), users' own occlusions



(a) $NSSIM$ Curves on CDM Regression (b) PAE Curves on CDM Regression (c) CDFs of $RMSE_s$ on Crowd Speed Estimation

Fig. 6. Influence of localization errors on crowd analysis results, where (a) and (b) are the two comprehensive metrics for crowd density estimation with the scaling factor β increasing, and (c) is the CDFs of $RMSE_s$ of crowd speed estimation with respect to different values of β.

(e.g., pockets, knapsacks and handbags), different deployments of sniffers, and limitations of localization methods together lead to localization errors with different magnitudes. As such, it is necessary to investigate the influence of different magnitudes of localization errors on crowd analysis. To this end, define β to be the scaling factor, such that reducing ($0 \leq \beta < 1$) or increasing ($\beta > 1$) localization errors in a semi-synthetic dataset can be easily realized by letting β take values from (0,1) and $(1, +\infty)$, namely that, the localization error distributions in both axes become $N'_x(\beta \times \mu_x, \beta \times \sigma_x)$ and $N'_y(\beta \times \mu_y, \beta \times \sigma_y)$. The WiATC dataset which involves more data than WiDIA is used for test by re-synthesizing with different values of β. With β increasing, the two comprehensive metrics (i.e., NSSIM and PAE) of all methods for crowd density estimation are shown in Fig. 6(a) and (b), and the cumulative distribution functions (CDFs) of $RMSE_s$ of the proposed speed estimation method is shown in Fig. 6(c).

The figures intuitively indicate that: 1) the crowd density estimation performance of all methods become worse with localization error increasing; 2) different methods incur different degrees of performance reduction, but the ADCA is only slightly affected, indicating that our method is able to learn the localization error distribution through its connectivity characteristic and thus mitigate the impact of localization errors to some extent, as was discussed in Section IV-C5; 3) all the CDFs of $RMSE_s$ under different values of β almost overlap, which is attributable to the fact that our speed estimation method mainly relies on accurate CDMs, such that slight degradation in the quality of estimated CDMs caused by larger localization errors do not significantly impair its accuracy. Note that the slight and inconsistent differences between CDFs are probably caused by the randomness in the processes of re-synthesizing the dataset and initializing the BiConvLSTM model.

F. Validating the Practicality on Real-World Dataset

A random time T is selected for visualization and presentation of crowd density estimation. Four cameras are selected to show the key areas and the GT crowd count in each area, as shown in Fig. 7(a), and accordingly, the manually labeled GT CDM is illustrated in Fig. 7(b). The best baseline, i.e., CF, is implemented (see Fig. 7(c)), in which CDM is relatively scattered and only contains two LHDCs that are seriously inconsistent with the GT. We randomly select 1000 training samples from each semi-synthetic dataset and directly transfer the trained model to WiCAM (see Fig. 7(d)), resulting in the reduction of scattering but keeping the similar shape. Fig. 7(e) and (f) show the CDMs estimated by the ADCA pre-trained with above samples after data scaling and augmentation, and the pre-trained model fine-tuned with 10 labeled CDMs, respectively. Apparently, the former can learn generic knowledge to correct crowd densities and thus to identify more LHDCs, while the latter can learn scenario-specific knowledge from labeled samples to achieve more accurate crowd counts and locations.

In addition, we also test them with 30 labeled test samples and summarize the results in Table IV, where the fine-tuned cases include using different numbers of training samples. It can be found that: except for global crowd counting, the directly transferred and pre-trained models outperform the baseline on most metrics, demonstrating the superiority of our method for fine-grained crowd density estimation; fine-tuned models can significantly boost the accuracy, and particularly, the more training samples are used, the better performance is obtained.

For crowd speed estimation, three LHDCs are identified by DM-DBSCAN, and the speed estimates and corresponding labels are shown in Fig. 7(g) and (a). It can be found that direction estimates are consistent with realities, and velocity estimates are in the normal range and intuitively follow the inverse relationship between crowd densities and velocities. Meanwhile, velocity estimates are slightly smaller than the normal pedestrian walking velocity, i.e., around 1.35 m/s, which is reasonable considering the pedestrians' interaction and compositions of the pedestrian speeds in different directions within a crowd.

To further verify the continuity of crowd analysis, CDM and speed estimates at 11 consecutive moments with the interval of 5 s centered on time T are presented in Fig. 7(h). It can be concluded that: 1) CDM estimates at closer moments have more similar shapes, indicating that CDM estimates have good continuities; 2) the proposed scheme can effectively identify LHDCs, with the result that direction estimates are mostly consistent with road directions, and velocity estimates follow the inverse

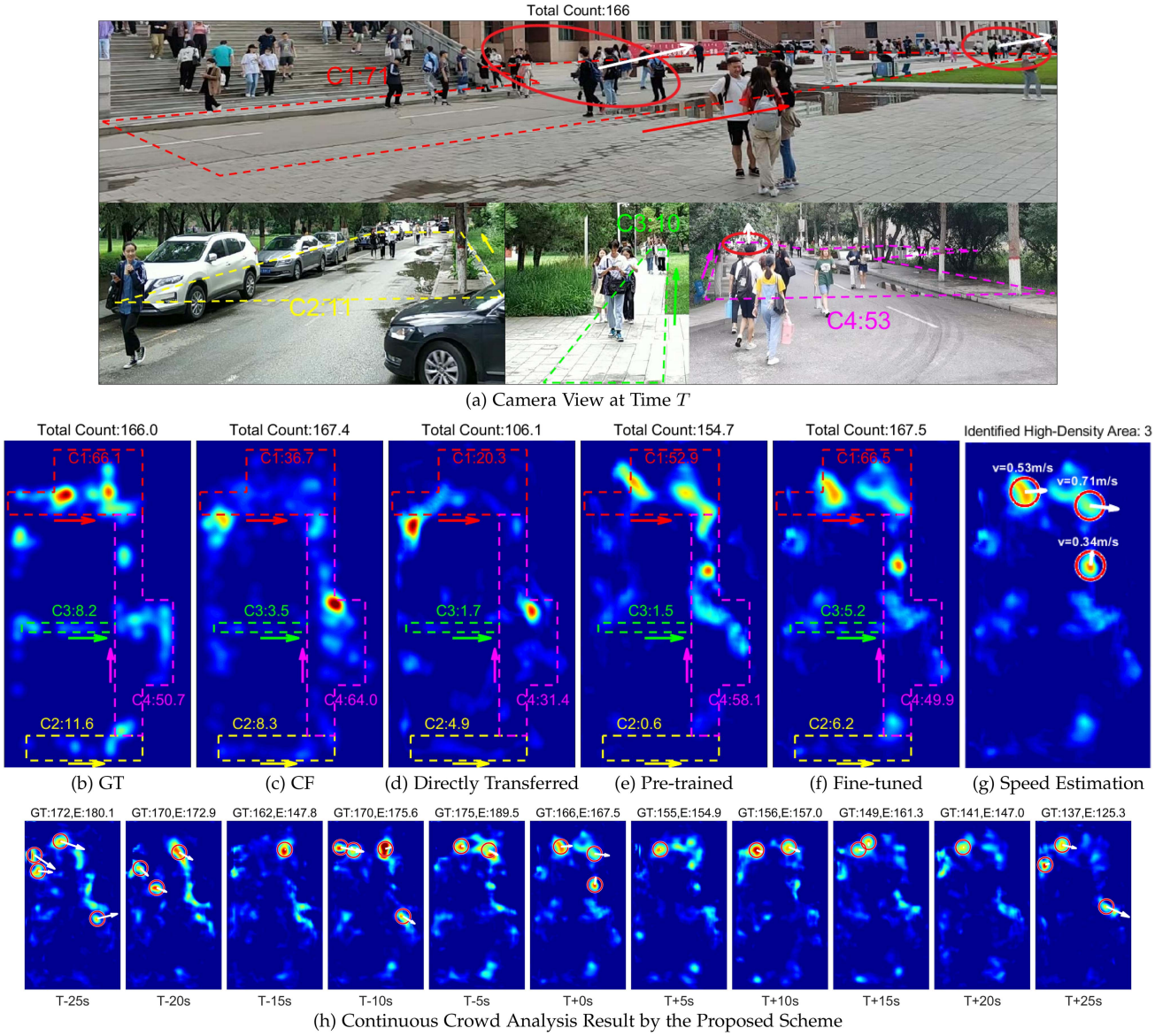


Fig. 7. Visualized crowd analysis results on WiCAM dataset, where (a) camera view and labeled crowd counts in each area at time T; (b) the GT CDM; (c)∼(f) the CDMs estimated by CF, ADCA trained with the mixture of semi-synthetic datasets, pre-trained with the scaled and augmented data, and fine-tuned with 10 labeled training samples, respectively; (g) identified LHDCs and corresponding speed estimates at time T; (h) results of continuous crowd analysis with the interval of 5 s centered at time T (GT is the labeled global crowd count and E is the estimated one).

TABLE IV
CROWD DENSITY ESTIMATION RESULTS OF THE BASELINE (CF) AND DIFFERENT TRANSFER METHODS ON THE WICAM DATASET

Method	PSNR ↑	NSSIM ↑	$\mathrm{RMSE_p}$ ↓	$\mathrm{MAE_c}$ ↓	$\mathrm{MSE_c}$ ↓	PAE ↓	PSE ↓
Baseline (CF [18])	35.676	0.866	0.019	14.173	459.921	1.782	13.979
Directly Transferred	36.697	**0.899**	0.017	33.888	1617.737	1.618	12.223
Pre-trained	**36.804**	0.836	0.016	14.917	701.842	1.534	9.996
Fine-tuned (10 samples)	36.073	0.864	0.017	12.804	595.354	1.578	10.067
Fine-tuned (20 samples)	36.468	0.860	0.017	12.686	517.783	1.549	10.207
Fine-tuned (30 samples)	36.630	0.862	**0.016**	**10.539**	**315.897**	**1.495**	**9.810**

relationship between densities and velocities; 3) looking at the density and speed simultaneously, some interesting phenomena can be found: two LHDCs in the upper left corner at time $T-10\,\mathrm{s}$ converge at time $T-5\,\mathrm{s}$; the LHDC with a small velocity at time $T+10\,\mathrm{s}$ splits into two small crowds at time $T+15\,\mathrm{s}$. In summary, the proposed scheme can be effectively transferred into real scenarios and works well.

VII. Conclusion

This paper presented a passive WiFi sensing based fine-grained crowd analysis scheme for large surveillance areas, including a scenario-level crowd density estimation method and a delicate speed estimation method for LHDCs. An in-depth theoretical analysis was conducted to investigate the limitations of passive WiFi sensing-based crowd analysis, and motivated us to propose the ADCA model to combat these limitations, so as to reconstruct CDMs like image reconstruction. Then, LHDCs were identified from CDM estimates and speed vectors were estimated using the BiConvLSTM model. To implement our scheme and evaluate its performance, three semi-synthetic datasets with different sizes and crowd densities were constructed. In addition, transfer experiments on real-world data confirmed the practicality of our scheme. In summary, our systematic work not only can be referenced by related works in theory, but also paves the way for introducing other innovative applications in practice.

As to future works, we shall focus on how to automatically and efficiently label crowd scenes taken by drones and how to fuse global WiFi measurements with local visual information, so as to achieve accurate crowd analysis and enable more downstream tasks.

References

[1] D. Kang, Z. Ma, and A. B. Chan, "Beyond counting: Comparisons of density maps for crowd analysis tasks—Counting, detection, and tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 5, pp. 1408–1422, May 2019.

[2] M. S. Kaiser et al., "Advances in crowd analysis for urban applications through urban event detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 10, pp. 3092–3112, Oct. 2018.

[3] C. I. Inc., "At least 153 dead, 133 injured in stampede during halloween festivities in seoul," 2022. Accessed: Oct. 29, 2022. [Online]. Available: https://www.cbsnews.com/news/halloween-crowd-surge-seoul-south-korea-dozens-killed-dozens-injured/

[4] B. B. Corporation, "Indonesia: At least 125 dead in football stadium crush," 2022. Accessed: Oct. 02, 2022. [Online]. Available: https://www.bbc.com/news/world-asia-63105945

[5] X. Ding, F. He, Z. Lin, Y. Wang, H. Guo, and Y. Huang, "Crowd density estimation using fusion of multi-layer features," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 8, pp. 4776–4787, Aug. 2021.

[6] Q. Wang and T. P. Breckon, "Crowd counting via segmentation guided attention networks and curriculum loss," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 15 233–15 243, Sep. 2022.

[7] A. Zhu et al., "CACrowdGAN: Cascaded attentional generative adversarial network for crowd counting," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 8090–8102, Jul. 2022.

[8] L. Liu, Z. Cao, H. Lu, H. Xiong, and C. Shen, "NSSNet: Scale-aware object counting with non-scale suppression," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 4, pp. 3103–3114, Apr. 2022.

[9] X. Jiang et al., "Crowd counting and density estimation by trellis encoder-decoder networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6126–6135.

[10] Z. Yan, R. Zhang, H. Zhang, Q. Zhang, and W. Zuo, "Crowd counting via perspective-guided fractional-dilation convolution," *IEEE Trans. Multimedia*, vol. 24, pp. 2633–2647, 2022.

[11] H. Li, E. C. L. Chan, X. Guo, J. Xiao, K. Wu, and L. M. Ni, "Wi-counter: Smartphone-based people counter using crowdsourced Wi-Fi signal data," *IEEE Trans. Human-Mach. Syst.*, vol. 45, no. 4, pp. 442–452, Aug. 2015.

[12] S. Depatla, A. Muralidharan, and Y. Mostofi, "Occupancy estimation using only WiFi power measurements," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 7, pp. 1381–1393, Jul. 2015.

[13] Y. Zhao, S. Liu, F. Xue, B. Chen, and X. Chen, "DeepCount: Crowd counting with Wi-Fi using deep learning," *J. Commun. Inf. Netw.*, vol. 4, no. 3, pp. 38–52, 2019.

[14] S. Liu, Y. Zhao, and B. Chen, "WiCount: A deep learning approach for crowd counting using WiFi signals," in *Proc. IEEE Int. Symp. Parallel Distrib. Process. Appl. IEEE Int. Conf. Ubiquitous Comput. Commun.*, 2017, pp. 967–974.

[15] H. Hong, G. D. De Silva, and M. C. Chan, "CrowdProbe: Non-invasive crowd monitoring with Wi-Fi probe," in *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 2, no. 3, Sep. 2018, Art. no. 115.

[16] B. Huang, G. Mao, Y. Qin, and Y. Wei, "Pedestrian flow estimation through passive WiFi sensing," *IEEE Trans. Mobile Comput.*, vol. 20, no. 4, pp. 1529–1542, Apr. 2021.

[17] A. Lesani and L. Miranda-Moreno, "Development and testing of a real-time WiFi-Bluetooth system for pedestrian network monitoring, classification, and data extrapolation," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 4, pp. 1484–1496, Apr. 2019.

[18] J. Weppner, B. Bischke, and P. Lukowicz, "Monitoring crowd condition in public spaces by tracking mobile consumer devices with WiFi interface," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, New York, NY, USA, 2016, pp. 1363–1371.

[19] Y. Fukuzaki, M. Mochizuki, K. Murao, and N. Nishio, "Statistical analysis of actual number of pedestrians for Wi-Fi packet-based pedestrian flow sensing," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. Proc. ACM Int. Symp. Wearable Comput.*, New York, NY, USA, 2015, pp. 1519–1526.

[20] F.-J. Wu and G. Solmaz, "CrowdEstimator: Approximating crowd sizes with multi-modal data for Internet-of-Things services," in *Proc. 16th Annu. Int. Conf. Mobile Syst. Appl. Serv.*, New York, NY, USA, 2018, pp. 337–349.

[21] L. Schauer, M. Werner, and P. Marcus, "Estimating crowd densities and pedestrian flows using Wi-Fi and bluetooth," in *Proc. 11th Int. Conf. Mobile Ubiquitous Syst.: Comput. Netw. Serv.*, Brussels, 2014, pp. 171–177.

[22] Z. Pu, Z. Cui, J. Tang, S. Wang, and Y. Wang, "Multimodal traffic speed monitoring: A real-time system based on passive Wi-Fi and bluetooth sensing technology," *IEEE Internet Things J.*, vol. 9, no. 14, pp. 12 413–12 424, Jul. 2022.

[23] Y. Li, J. Barthelemy, S. Sun, P. Perez, and B. Moran, "A case study of WiFi sniffing performance evaluation," *IEEE Access*, vol. 8, pp. 129 224–129 235, 2020.

[24] C. Matte, M. Cunche, F. Rousseau, and M. Vanhoef, "Defeating MAC address randomization through timing attacks," in *Proc. 9th ACM Conf. Secur. Privacy Wireless Mobile Netw.*, New York, NY, USA, 2016, pp. 15–20.

[25] J. M. Grant and P. J. Flynn, "Crowd scene understanding from video: A survey," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 13, no. 2, Mar. 2017, Art. no. 19.

[26] X. Wu, G. Liang, K. K. Lee, and Y. Xu, "Crowd density estimation using texture analysis and learning," in *Proc. IEEE Int. Conf. Robot. Biomimetics*, 2006, pp. 214–219.

[27] K. Chen and J.-K. Kämäräinen, "Pedestrian density analysis in public scenes with spatiotemporal tensor features," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 7, pp. 1968–1977, Jul. 2016.

[28] A. Perera, C. Srinivas, A. Hoogs, G. Brooksby, and W. Hu, "Multi-object tracking through simultaneous long occlusions and split-merge conditions," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 666–673.

[29] Z. Ma and A. B. Chan, "Crossing the line: Crowd counting by integer programming with local features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2539–2546.

[30] A. Zhang et al., "Relational attention network for crowd counting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6787–6796.

[31] X. Yu, Y. Liang, X. Lin, J. Wan, T. Wang, and H.-N. Dai, "Frequency feature pyramid network with global-local consistency loss for crowd-and-vehicle counting in congested scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 9654–9664, Jul. 2022.

[32] F. Dittrich, L. E. S. de Oliveira, A. S. Britto Jr., and A. L. Koerich, "People counting in crowded and outdoor scenes using a hybrid multi-camera approach," 2017, *arXiv: 1704.00326*.
[33] M. Uras, R. Cossu, and L. Atzori, "PMA: A solution for people mobility monitoring and analysis based on WiFi probes," in *Proc. 4th Int. Conf. Smart Sustain. Technol.*, 2019, pp. 1–6.
[34] F. Tofigh, G. Mao, J. Lipman, and M. Abolhasan, "Crowd density mapping based on Wi-Fi measurements on train platforms," in *Proc. 12th Int. Conf. Signal Process. Commun. Syst.*, 2018, pp. 1–7.
[35] Y. Tian, B. Huang, B. Jia, and L. Zhao, "Optimizing AP and beacon placement in WiFi and BLE hybrid localization," *J. Netw. Comput. Appl.*, vol. 164, 2020, Art. no. 102673.
[36] L. Hao, B. Huang, B. Jia, and G. Mao, "DHCLoc: A device-heterogeneity-tolerant and channel-adaptive passive WiFi localization method based on DNN," *IEEE Internet Things J.*, vol. 9, no. 7, pp. 4863–4874, Apr. 2022.
[37] S. Jung, C. O. Lee, and D. Han, "Wi-Fi fingerprint-based approaches following log-distance path loss model for indoor positioning," in *Proc. IEEE MTT-S Int. Microw. Workshop Ser. Intell. Radio Future Pers. Terminals*, 2011, pp. 1–2.
[38] Y. Ji, C. Yu, J. Wei, and B. Anderson, "Localization bias reduction in wireless sensor networks," *IEEE Trans. Ind. Electron.*, vol. 62, no. 5, pp. 3004–3016, May 2015.
[39] X.-J. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Red Hook, NY, USA: Curran Associates Inc., 2016, pp. 2810–2818.
[40] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15 979–15 988.
[41] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 589–597.
[42] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Proc. Int. Conf. Artif. Neural Netw. Mach. Learn.*, T. Honkela, W. Duch, M. Girolami, and S. Kaski, Eds., Berlin, Germany: Springer, 2011, pp. 52–59.
[43] G. Chen, P. Chen, Y. Shi, C.-Y. Hsieh, B. Liao, and S. Zhang, "Rethinking the usage of batch normalization and dropout in the training of deep neural networks," 2019, *arXiv: 1905.05928*.
[44] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Cham: Springer, 2018, pp. 3–19.
[45] X. Cao, Z. Wang, Y. Zhao, and F. Su, "Scale aggregation network for accurate and efficient crowd counting," in *Proc. 15th Eur. Conf. Comput. Vis.*, Berlin, Germany: Springer-Verlag, 2018, pp. 757–773.
[46] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[47] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discov. Data Mining*, 1996, pp. 226–231.
[48] L. Pintor and L. Atzori, "A dataset of labelled device Wi-Fi probe requests for MAC address de-randomization," *Comput. Netw.*, vol. 205, 2022, Art. no. 108783.
[49] J. Bi, Y. Wang, B. Yu, H. Cao, T. Shi, and L. Huang, "Supplementary open dataset for WiFi indoor localization based on received signal strength," *Satell. Navigation*, vol. 3, no. 1, Nov. 2022, Art. no. 25.
[50] F. Zanlungo, D. Brščić, and T. Kanda, "Spatial-size scaling of pedestrian groups under growing density conditions," *Phys. Rev. E*, vol. 91, no. 6, Jun. 2015, Art. no. 062810.
[51] F. Zanlungo, T. Ikeda, and T. Kanda, "Potential for the dynamics of pedestrians in a socially interacting group," *Phys. Rev. E Statist. Nonlinear Soft Matter Phys.*, vol. 89, no. 1, Jan. 2014, Art. no. 012811.
[52] B. Huang, Z. Xu, B. Jia, and G. Mao, "An online radio map update scheme for WiFi fingerprint-based localization," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 6909–6918, Aug. 2019.

Lifei Hao received the BS degree in applied physics from Chongqing University, Chongqing, China, in 2012, and the ME degree in computer technology from Inner Mongolia University, Hohhot, China, in 2019, respectively, where he is currently working toward the PhD degree with the College of Computer Science. His main research interests include Internet of Things, WiFi localization, and passive WiFi sensing.

Baoqi Huang (Member, IEEE) received the BE degree in computer science from Inner Mongolia University (IMU), Hohhot, China, in 2002, the MS degree in computer science from Peking University, Beijing, China, in 2005, and the PhD degree in information engineering from Australian National University, Canberra, ACT, Australia, in 2012. He is currently a professor with the College of Computer Science, IMU. His research interests include indoor localization and navigation, wireless sensor networks, and mobile computing. He was the recipient of the Chinese Government Award for Outstanding Chinese Students Abroad in 2011.

Bing Jia (Member, IEEE) received the PhD degree from Jilin Univesity, Changchun, China, in 2013. She is currently an associate professor with the College of Computer Science, Inner Mongolia University, Hohhot, China. Her current research interests include indoor localization, crowdsourcing, wireless sensor networks, and mobile computing.

Guoqiang Mao (Fellow, IEEE) received the PhD degree in telecommunications engineering from Edith Cowan University, Australia, in 2002. He is a distinguished professor and dean with the Research Institute of Smart Transportation, Xidian University. Before that, he was with the University of Technology Sydney and the University of Sydney. He has published more than 200 papers in international conferences and journals, which have been cited more than 9,000 times. His research interests include intelligent transport systems, applied graph theory and its applications in telecommunications, Internet of Things, wireless sensor networks, wireless localization techniques, and network modeling and performance analysis. He received the Top Editor Award for outstanding contributions to the *IEEE Transactions on Vehicular Technology* in 2011, 2014, and 2015. He has been an editor of *IEEE Transactions on Intelligent Transportation Systems* since 2018, *IEEE Transactions on Wireless Communications* since 2014 to 2019, and *IEEE Transactions on Vehicular Technology* since 2010. He is a co-chair of the IEEE Intelligent Transport Systems Society Technical Committee on Communication Networks. He has served as a chair, co-chair, and a TPC member in a number of international conferences. He is a Fellow of the IET.