Toward Accurate Crowd Counting in Large Surveillance Areas Based on Passive WiFi Sensing

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

***Abstract*— Great efforts have been devoted to solving the crowd counting problem based on vision or other fine-grained measurements. Popular vision and WiFi channel state information based approaches, though are able to achieve relatively high accuracy, suffer from limited scalability. In contrast, passive WiFi sensing-based approaches are capable of supporting large surveillance areas, but often rely on certain global linear or approximately linear regression models, which cannot accurately capture the complex mapping relationship between WiFi sensing data and the corresponding crowd count, especially in a large surveillance area during a long period of time. This paper addresses the issue from the following three aspects. Firstly, in order to combat with these coarse-grained regression models, the large surveillance is partitioned into grids, such that either a local linear model or other implicit local models can be built with respect to each grid. Secondly, sequential WiFi spatial-temporal matrix (SWSTM) is defined in alignment with grids to encode the spatial-temporal information of crowds based on passive WiFi localization and a sliding time window mechanism. Thirdly, the spatial-temporal correlations among crowd features of different grids are mined to better regress such local models by using a recurrent neural network (RNN) with SWSTMs as inputs. Extensive experiments are conducted in a real campus road network with an area of about 4000m^2, and demonstrate that the proposed method significantly reduces the counting error rate from 22.54% to 13.44% compared to several state-of-the-art methods.**

***Index Terms*— WiFi localization, crowd counting, spatial–temporal correlation, multi-objective optimization, deep learning.**

I. Introduction

WITH the rapid development of cities, the surveillance of gathering crowds has become an important research field. But it is still an open problem to accurately estimate the sizes of crowds in large-scale areas, which is vital for many intelligent perception applications [1], [2], [3], including traffic management, business decision, smart city, public safety management, and etc. For example, with the spread of COVID-19, accurately grasping the number of people in a large public space is greatly helpful to the traffic management [4].

Therefore, crowd counting has attracted the attention of the research community. Assisted with the recent advancement in deep learning (DL), computer vision researchers have presented various high-accuracy crowd counting techniques [2], [5], [6]. However, vision-based methods suffer from high installation costs, blind spots, occlusion issues and privacy concerns, and more importantly, cannot completely cover a large-scale area of interest (AOI) with a single camera [7], [8]. As such, algorithms based on analyzing the radio frequency (RF) signals have been introduced [9], [10]. Therein, the WiFi channel state information (CSI)-based approach is a promising one for high-precision crowd counting [11], [12], and can achieve the accuracy of over 85% with the help of DL technologies [13], [14]. But it is still limited to the scale and environmental dynamic due to the fine-grained characteristic of CSI, and can only be applied in indoor scenarios with minor persons in practice.

Finally, only passive WiFi sensing-based methods have the potential for crowd counting in a large AOI [15]. The basic idea is leveraging a special kind of access point (AP), termed WiFi sniffer, to passively sense the nearby pedestrians' existence by capturing the probe (request) frames sent from their mobile devices. However, the challenges from persons with multiple WiFi-enabled mobile devices, uncertain sniffing [16] and MAC address randomization [17] may affect this kind of methods. Fortunately, the strict theoretical deduction in [18] demonstrates both the expected number of mobile devices carried by a pedestrian and the probability that a mobile device can be sensed by a sniffer are constants for a fixed environment. In addition, the rate and pattern of MAC address randomization are proved to have a certain regularity for a long timeslot [17]. As a result, it is expected that the relation between the number of detected devices and the crowd count within a relative large time window is approximate to a linear mapping. Therefore, most of the passive WiFi sensing-based methods attempted to train a global linear [19], [20], [21] or approximately linear regression model [22] for

Manuscript received 15 July 2022; revised 15 June 2023; accepted 4 August 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62262046, Grant 42161070, Grant 62061036, and Grant 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 Grant 2022YFSJ0027 and Grant 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. The Associate Editor for this article was S. Hamdar. *(Corresponding author: Baoqi Huang.)*

Lifei Hao, Baoqi Huang, Bing Jia, and Gang Xu 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: cshbq@imu.edu.cn).

Guoqiang Mao is with the Research Institute of Smart Transportation, Xidian University, Xi'an 710071, China.

Digital Object Identifier 10.1109/TITS.2023.3303700

count counting, and reported a large error rate of more than 20%.

Inspired by [21] and our pilot studies, we find that the mapping between the number of devices detected by WiFi and the realistic crowd count in different sub-regions of the AOI is varying over time and space, and thus we propose to partition the AOI into small grids and then apply local linear or other implicit models with different parameters for each grid, so as to get a finer estimation. In fact, given a gird, the crowd count/density is close to the adjacent girds in space and several previous states in time, showing a strong spatial-temporal correlation [23], [24]. Therefore, we further consider leveraging DL technology to capture this characteristic in order to optimize the mapping parameters.

On these grounds, this paper presents a novel crowd counting approach for large-scale surveillance areas based on passive WiFi sensing. To be specific, all detected devices are located by passive WiFi localization firstly. Then, we partition the whole AOI into grids with equal size, and the localization results within a sliding time window are utilized to construct the WiFi spatial-temporal matrix (WSTM) and several consecutive WSTMs are stacked as a sequential WSTM (SWSTM) to maximally reserve the spatial-temporal information of crowds. At last, three supervised learning methods are proposed to regress the crowd count through the WSTM/SWSTM by gradually adding the merits of area partitioning, utilizing spatial correlation and spatial-temporal correlation, respectively. In addition, a qualitative analysis on the superiorities of our method and several key influence factors of the counting performance are also presented.

For the purpose of performance evaluation, an experimental crowd surveillance system is deployed in a real campus environment with the area of about $4000m^2$, and a sensing dataset is collected during a time period encompassing the peak time after classes. It is shown that the counting accuracy of our method is significantly better than that using the existing global linear or approximately linear regression models [19], [20], [21], [22]. Particularly, the propsed recurrent neural network (RNN) model combining all merits can substantially reduce the counting error rate from 22.54% to 13.44%, which is competitive with vision-based and CSI-based methods in accuracy but for much larger AOIs.

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

- We combine the WiFi localization to construct formatted SWSTMs, so as to reserve the spatial-temporal information of crowds in the sensing data and facilitate the further processing by supervised learning methods.
- Three different optimization/DL technologies are modified as the supervised crowd counter (SCC) of our approach to utilize area partitioning, spatial and spatial-temporal correlation.
- We present the integrated solution in detail in order to make it easily being utilized and modified by practitioners for their applications.
- A labeled WiFi sensing dataset is obtained in a real-world large-scale scenario, and used to validate the effectiveness of three strategies on capturing the fine-grained mapping relationship.

The rest of this paper is organized as follows. Section II surveys the literature in relation to our work. Section III presents our method and a theoretical analysis on its superiorities. In Section IV, three different SCCs are elaborately designed. Section V shows the experimental results, and Section VI concludes the whole paper and sheds light on future works.

II. Related Work

In this section, we shall briefly introduce the literature on crowd counting, including the traditional vision-based approach and the emerging wireless-based approach.

A. Vision-Based Crowd Counting

A recent paper [25] surveyed the existing studies on vision-based crowd scene surveillance, and reported the challenges. First, pixel-level approaches begin with edge detection and use edge features to train a model, and texture-level approaches [26], [27] are coarser-grained than pixel-level ones and aim to analyze image patches. Both of them aim to estimate the crowd count in a scene, rather than identify individuals, and can only achieve coarse-grained results. Second, object-level approaches [28] can obtain more accurate results by identifying individuals but are only suitable for sparse scenes. Third, line counting approaches [29] count the crowd crossing a line of interest rather than the entire AOI, and thus they cannot thoroughly handle the criticality of a situation. Fourth, density mapping approaches [5], [6], [30], [31] estimate the crowd density rather than identifying the number of individuals in the scene, suffering from scale variations [3], [32] and video qualities. Moreover, a common limitation is that only a small area covered by a single camera is considered. Therefore, when dealing with the crowd counting in a large area covered by multiple cameras, it is still an open problem [33], not to mention the possible coverage holes and overlaps.

Based on the above discussion, it can be concluded that, besides the traditional limitations, e.g., illumination conditions and computational complexities, existing vision-based approaches are also restricted by large area surveillance, high pedestrian densities and cross-camera counting.

B. Wireless-Based Crowd Counting

First, with the advantages of low cost, large coverage, scalability, device-free, and convenience for target recognition, passive WiFi sensing has enabled crowd counting [22]. In [15] and [19], the feasibility of the passive WiFi sensing-based crowd counting method was validated through field experiments, and the results showed it suffers higher error rates than 30%. Similarly, [20] adopted WiFi sniffers with directional antennas and video-based crowd counting method, so as to reduce the error rate by close to 20 percent. In [21], a stereoscopic camera was installed at a calibration choke point and helped to reduce the count error. In short, the existing studies though appear to be feasible in practice, are mainly restricted by the relatively low accuracy.

Second, efforts have been devoted to mining a variety of deeper information from sensing data for aiding the crowd

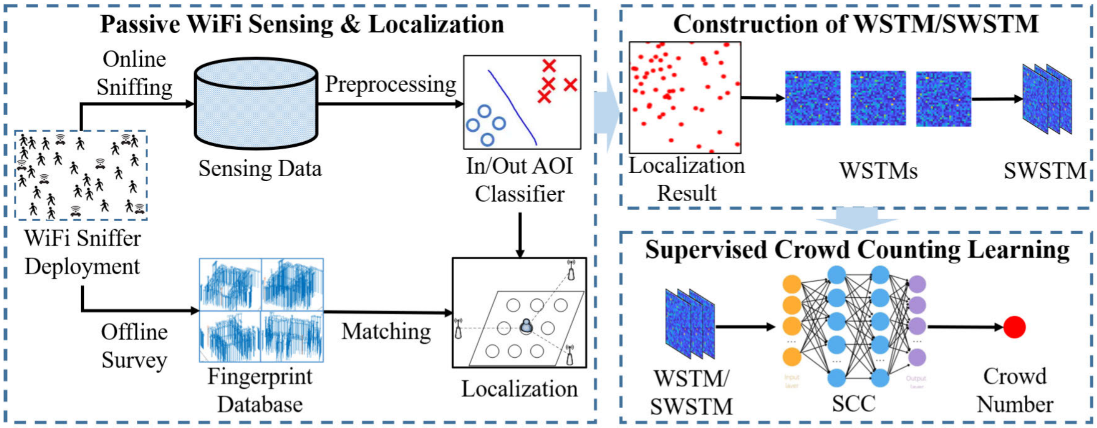


Fig. 1. The overview of the proposed crowd counting approach.

counting. An early study [34] demonstrated the feasibility of using received signal strength (RSS) to estimate crowd counts, and derived a linear formula that relates the crowd count to the RSS average and variance. Nuzzer system [35] further extended this model to work on a large scale. In [12], the authors proposed a probabilistic approach to calculate the RSS probability mass function (PMF) for each case of the crowd count, achieving 25% counting accuracy. However, they need a large number of WiFi sniffers deployed in order to achieve an acceptable counting accuracy. One possible solution to resolve this trade-off is to use CSI instead of RSS [13], [14], [36]. Unfortunately, it is hard to obtain CSI data by commonly used APs/WiFi sniffers, and the fine-grained CSI measurements are highly susceptible to environmental influence.

In summary, neither the traditional vision-based approaches nor the emerging wireless-based approaches can perfectly resolve the issues confronted by crowd counting. However, the wireless-based approaches demonstrate the potential for crowd counting in large areas and/or with high crowd densities, which are the major shortage of vision-based ones. As such, it is imperative to establish a finer mapping relationship between the crowd count and the sensing data, so as to improve the accuracy of passive WiFi sensing-based crowd counting.

III. Passive WiFi Sensing-Based Crowd Counting Approach

In this section, we shall present the overview, the design detail, and a qualitative analysis of the performance of the proposed crowd counting approach.

A. Overview

The overview of the proposed crowd counting approach is shown in Fig. 1, which can be divided into three stages:

1) Passive WiFi Localization: Several WiFi sniffers are uniformly deployed in the AOI according to the coverage requirements of WiFi localization. Then, a offline survey is conducted using several reference devices to construct the WiFi fingerprint database (FD). At last, a localization method such as k-nearest neighbor (KNN) [37], weighted KNN (WKNN) [38] or more advanced DHCLoc [39] is employed to position all available fingerprints in the online sensing.

2) Construction of WSTM/SWSTM: The sensing data within one/several sliding window(s) is collected from WiFi sniffers in the AOI and then preprocessed as fingerprints each of which is a vector consisting of the RSS means from all WiFi sniffers of a detected device, and a simple KNN-based binary classifier trained by both the fingerprints in and out the AOI is adopted to output a logical value denoting whether a detected device is in the AOI. All fingerprints inside the AOI will be located to form a set of positions. Based on the area partitioning and localization results, WSTMs are generated by Algorithm 1 and stacked as the SWSTM.

3) Supervised Crowd Counting Learning: We leverage a traditional multi-objective optimization combined with the local linear model (LLM), or end-to-end/sequence-to-sequence DL method, as the SCC of our approach. The SCC is used to establish a finer and more accurate mapping between WSTMs/SWSTMs and the crowd counts, and exploit spatial-temporal correlations among WSTMs/SWSTMs.

To sum up, all components in our approach will impact the counting performance, and particularly, the classifier and localization influence the accuracy slightly while the alternative SCC directly executes the counting function and affects a lot.

B. Design Detail

The key components of the proposed approach are elaborately formulated in the following.

1) Deployment of WiFi Sniffers: The WiFi sniffer can either be dedicated WiFi sniffers or commercial programmable APs. With the sniffing script, WiFi sniffers can periodically upload the sensing data to the server for further processing. The traditional manual deployment or the state-of-art automatic deployment solutions of APs [40] is adopted for deploying WiFi sniffers. Due to the unapparent effects that the WiFi localization acts on crowd counting, we only ask the deployment of WiFi sniffers satisfying that every point in an AOI should be covered by at least 3 non-collinear WiFi sniffers.

2) Offline Survey: To enable the WiFi localization, a offline survey must be conduct to establish the FD. We first build the cartesian coordinate system for the AOI using a laser range finder measuring the size of AOI. Then, we utilize path-based fingerprint collection [41] which collect WiFi packets while walking along a path with known start and terminal locations. Suppose that there are m_w WiFi sniffers, m_w dimension fingerprint vectors are obtained by averaging the RSS measurements extracted from all sensed packets sent by reference devices. The location label of each fingerprint vector could then be inferred by the timestamps and paths. In addition, to enable the In/Out AOI classifier, both paths inside and outside the AOI are surveyed. At last, all pathes inside the AOI are discretized to n_r reference points (RPs) and the FD with the size of $n_r \times (m_w + 2)$ is obtained, where 2 denotes the length of a RP's 2-D location coordinate.

3) Sensing Data Preprocessing: The original sensing data mainly includes timestamp, RSS, frame type, MAC address of the transmitter, and the following preprocessing is conducted based on these attributes: a) Abnormal Data Cleaning: the frames with missing or over-ranged values will be discarded. b) Device Filtering: fingerprints from non-mobile devices such as wireless APs and staff's devices are filtered out by MAC addresses to ensure detected devices are carried by real pedestrians; fingerprints from devices that are detected by less than 3 WiFi sniffers are also abandoned. c) In/Out AOI classifying: the devices carried by pedestrians before entering or just leaving the AOI would also be detected by sniffers due to the sliding window mechanism, and thus a simple KNN-based classifier trained by the fingerprints of reference devices locating both in and out the AOI is utilized. It is expected to achieve high accuracy for such a simple binary classification problem with m_w features. As last, the online fingerprints outside the AOI are abandoned.

4) WiFi Localization: Before the localization, a fingerprints standardization method [42] is leveraged to alleviate the device heterogeneity. Given a fingerprint vector $\mathbf{F} = [F_1, F_2, \ldots, F_{m_w}]$, the standardized fingerprint vector is calculated by

$$\hat{\mathbf{F}} = [F_1 - \overline{F}, F_2 - \overline{F}, \ldots, F_{m_w} - \overline{F}]/\hat{\sigma}, \tag{1}$$

where $\overline{F} = \frac{1}{m_w}\sum_{i=1}^{m_w} F_i$ and $\hat{\sigma} = \sqrt{\frac{1}{m_w}\sum_{i=1}^{m_w}(F_i - \overline{F})^2}$ denote the mean and standard deviation of the fingerprint itself, respectively. Then, we utilize a sliding window with the size of Δt to filter out the sensing data $\mathcal{D}$, and construct fingerprint vectors of each MAC in $\mathcal{D}^{\Delta t}$, forming a set of fingerprint vectors $\mathbf{FS}^{\Delta t}$. For each standardized fingerprint vector $\hat{\mathbf{F}}$ in $\mathbf{FS}^{\Delta t}$, the KNN-based WiFi localization algorithm [37] is adopted to locate the device. The algorithm selects k RPs with the least Euclidean distances between the standardized FD and $\hat{\mathbf{F}}$, and takes the coordinate means of these RPs as the location estimate. Finally, the set of device positions within Δt, denoted by $\mathbf{L}^{\Delta t}$, is obtained.

5) Construction of WSTM/SWSTM: The WSTM/SWSTM is used to model the spatial and temporal relations with respect to the crowd distribution in spatial domain and the variation of crowd distribution in temporal domain. From the perspective of representation, WSTM/SWSTM acts as the manual feature extraction and can reform the random WiFi modality into a consolidated size of format, facilitating further input into an SCC. To construct the WSTM/SWSTM, we partition the AOI into $M \times N$ grids with equal size according to the cartesian coordinate system, and each of the grids is a rectangle corresponding to a small region. Then, every position in $\mathbf{L}^{\Delta t}$ is assigned to one grid. In temporal domain, WSTMs in multiple consecutive sliding time windows are stacked as sequential data, namely the SWSTM. More details regarding the construction of WSTM/SWSTM is summarized in Algorithm 1.

Algorithm 1 The Construction of WSTM/SWSTM

```
Input: the size of AOI $H \times W$, the partition
       granularity $M \times N$, the sequence length $l$, the
       set of location estimates $\mathbf{L}^{\Delta t}$
Output: the SWSTM $\mathcal{S}_{\mathbf{l}}$
Partition the AOI into $M \times N$ grids with equal size;
Initialize a $M \times N$ matrix $\mathcal{S}^{\Delta \mathbf{t}}$ with 0, $\mathcal{S}_{\mathbf{l}} \leftarrow \emptyset$;
for $L_i \in \mathbf{L}^{\Delta t}$ do
    for $j, k \in M, N$ do
        if $\frac{W}{N}(k-1) \le L_i.x < \frac{W}{N}k$ and
          $\frac{H}{M}(j-1) \le L_i.y < \frac{H}{M}j$ then
            $\mathcal{S}^{\Delta t}_{jk} \leftarrow \mathcal{S}^{\Delta t}_{jk} + 1$;
        end
    end
end
for $t = 1, \ldots, l$ do
    Repeat above processes to generate $\mathcal{S}^{\Delta \mathbf{t}}_{\mathbf{t}}$;
    $\mathcal{S}_{\mathbf{l}} \leftarrow \mathcal{S}_{\mathbf{l}} \bigcup \mathcal{S}^{\Delta \mathbf{t}}_{\mathbf{t}}$;
end
return $\mathcal{S}_{\mathbf{l}}$
```

6) Supervised Crowd Counter: After the original sensing data being processed as WSTMs/SWSTMs, we further transform the crowd counting problem into a supervised non-linear regression with minor annotations of crowd counts. Count labels of different WSTMs/SWSTMs can be obtained by dividing the AOI into sub-regions, counting the crowd counts in each sub-region using cameras, and then summing the counts of all sub-regions. In the regression, three types of supervised method can be adopted to gradually adding the utilizations of the fine-grained mapping between WSTMs and crowd counts, the spatial correlation inside one WSTM and the temporal correlation among sequential WSTMs in an SWSTM.

- For traditional multi-objective optimizations such as the particle swarm optimization (PSO) and genetic algorithm (GA), an LLM is established with respect to the ith row and jth column element in WSTM,

$$c_{ij} = a_{ij} \cdot \mathcal{S}^{\Delta \mathbf{t}}_{\mathbf{ij}} + b_{ij}, \tag{2}$$

where c_{ij} is the crowd count in the corresponding gird, a_{ij} and b_{ij} are the slop and offset of the LLM. Then, we formulate the regression as

$$\min |c^{gt} - \sum_{i,j=1}^{M,N} c_{ij}| = \min |c^{gt} - \sum_{i,j=1}^{M,N} (a_{ij} \cdot \mathcal{S}^{\Delta \mathbf{t}}_{\mathbf{ij}} + b_{ij})|,$$
$$\text{s.t.} \quad a_{ij} \in [a_l, a_u], b_{ij} \in [b_l, b_u], \tag{3}$$

where c^{gt} denotes the real count of the whole AOI in the current moment, and $[a_l, a_u]$, $[b_l, b_u]$ are the lower bound and upper bound of a_{ij}, b_{ij}, respectively. With an applicable optimization algorithm, $2 \cdot M \cdot N$ objectives, i.e., $a_{ij}, b_{ij} (i = 1, \ldots, M; j = 1, \ldots, N)$, are optimized simultaneously to minimize the goal $|c^{gt} - \sum_{i,j=1}^{M,N} c_{ij}|$.

- For non-sequential DL models such as the deep neural network (DNN) and convolutional neural network (CNN), we establish the global non-linear mapping, or implicit local model for each input, by

$$\begin{aligned} \hat{c} &= f^*(\mathcal{S}^{\Delta t}), \\ f^* &\in \mathcal{F} = \{f(\mathcal{S}^{\Delta t}; \theta) | \theta \in \mathbb{R}^D\}, \end{aligned} \tag{4}$$

  where $\mathcal{F}$ is the set of possible mapping functions determined by the models with different hyper-parameters, θ is the hyper-parameters and D is the number of hyper-parameters. Therefore, the regression process is occurring in the build-in mechanism of DL models, and nonlinear activations as well as multiple layers structure of these models even can boost the mapping from WSTMs to counts in a large margin. In addition, the fully connected characteristic links up each element in the WSTM to exploit the spatial correlation among grids, promising a more optimal mapping.
- The sequential variations of DL models such as the RNN and long short term memory (LSTM), can be adopted to further exploit the temporal correlation among WSTMs in an SWSTM for parameter optimizing. The basic idea is that using the weights learned from $(t-1)$th WSTM $\mathcal{S}_{t-1}^{\Delta t}$ to iteratively update the tth one. In analogy with 1-order Markov chain, the temporal correlation is delivering and accumulating step by step, resulting in a finer sequence-to-sequence mapping.

C. Theoretical Analysis on Counting Performance

To clarify the basic idea of our approach for crowd counting, we first conduct a pilot test regarding the linear relation between the ground-truth crowd counts and the number of detected devices located in the designated areas with a $20s$ sliding window. As shown in Fig. 2, the slopes of the linear relations in three sub-regions and the whole AOI (i.e., the sum of all 5 sub-regions) of our testbed with respect to the time variation are given. It can be clearly seen that the slope varies over time and is different for diverse locations in a given AOI (similar conclusion in [21]). On these grounds, we shall give a qualitative analysis on the superiority of our approach.

In the ideal condition, suppose that there are enough people in each of an appropriate partition of the AOI, i.e., $M \times N$ grids, and without considering the impact of localization errors, we follow [18], [19], [20], and [21] to assume the relation between the crowd count and the expected number of detected devices within Δt is a strict proportional mapping $c_{ij}^{gt} = a_{ij} \cdot \mathcal{S}_{\mathbf{ij}}^{\mathbf{\Delta t}}$, or more generally, a linear function $c_{ij}^{gt} = a_{ij} \cdot \mathcal{S}_{\mathbf{ij}}^{\mathbf{\Delta t}} + b_{ij}$. Then, we can get the following conclusions.

First, traditional methods estimate the crowd count in the whole AOI by $\hat{c} = a \cdot \sum \mathcal{S}^{\mathbf{\Delta t}}$ or $\hat{c} = a \cdot \sum \mathcal{S}^{\mathbf{\Delta t}} + b$, which is quite different from $c^{gt} = \sum_{i,j=1}^{M,N} a_{ij} \cdot \mathcal{S}_{\mathbf{ij}}^{\mathbf{\Delta t}}$ or $c^{gt} = \sum_{i,j=1}^{M,N} (a_{ij} \cdot \mathcal{S}_{\mathbf{ij}}^{\mathbf{\Delta t}} + b_{ij})$ and lead to great errors. This is because the compromised a and b would ruin estimates of majority grids (see Fig. 2). Whereas our approach tries to find the approximately optimal $a_{ij}, b_{ij} (i = 1, \ldots, M; j = 1, \ldots, N)$ of the LLM or θs of DL models to minimize estimate errors.

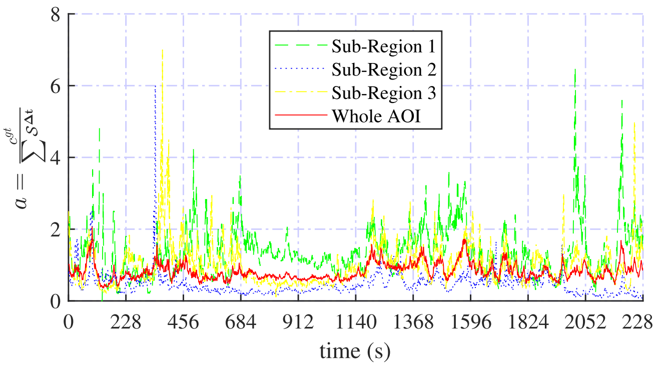


Fig. 2. The slope of the linear relation between the ground-truth crowd count and the number of detected devices varies over time and different sub-regions.

Second, when estimating parameters, localization errors may degrade the counting. However, devices in a given grid are probabilistically located in adjacent grids, and vice versa, the surrounding devices have the probability to be located in this grid, resulting in a compensating effect. Meanwhile, in consideration of the spatial correlation, the expected error of the device number between with and without the consideration of localization error is a trivial value. Therefore, our approach is insensitive to localization errors in some extent, but the counting error would rapidly rise when the localization error is large.

Third, in the real condition, due to the limited number of people in the whole AOI as well as each grid, an additive white Gaussian noise ϵ_{ij} would be introduced into the mapping,

$$\hat{c}_{ij} = a_{ij} \cdot (\mathcal{S}_{\mathbf{ij}}^{\mathbf{\Delta t}} + \epsilon_{ij}) + b_{ij} = a_{ij} \cdot \mathcal{S}_{\mathbf{ij}}^{\mathbf{\Delta t}} + b_{ij} + \epsilon'_{ij}. \tag{5}$$

The noise is heavily influenced by the size of grids, i.e., the partition granularity $M \times N$. Large $M \times N$ both simultaneously boosts the noise and aggravates the impact of localization error. Therefore, a suitable $M \times N$ should be adopted to trade off the count accuracy and errors according to the scales of AOI and the crowd distribution, and a relative large time window would also help to alleviate this contradiction.

Forth, we further introduce the SWSTM for sake of two reasons: the estimated parameters show strong correlations in the temporal domain (see Fig. 2), and thus more consecutive WSTMs are helpful to estimate more accurate parameters; the noise caused by limited number of people is time-varying due to the change of crowd count in grids, and thus different offsets of the LLM and DL model in each WSTM of an SWSTM are updated to compensate them.

IV. Supervised Crowd Counters

As mentioned in Section III-B6, we shall give the detailed description of three elaborately designed SCCs of our approach in the following.

Algorithm 2 The Training of PSO-SCC

```
Input: the number of particles $m_p$, the training set of WSTMs $\mathcal{S} = \{S_1, \dots S_{n_{tr}}\}$ and its label set $\mathbf{C^{gt}} = \{c_1^{gt}, \dots c_{n_{tr}}^{gt}\}$, the thresholds of particles' speeds $v_{th}$, the iterations $I$, the inertia factor $w$, acceleration constants $C_1, C_2$
Output: the optimal solution $\mathbf{X_g}$
Use the least square method to fit $c_k^{gt} = a \cdot \sum S_k + b (k = 1, \dots, n_{tr})$ to get $a$, $b$;
Initialize $m_p$ particles by setting their states $x_i = [\{a\} \times M \cdot N, \{\frac{b}{M \cdot N}\} \times M \cdot N]$ and speeds $v_i = [random(-v_{th}, v_{th}) \times 2 \cdot M \cdot N]$;
Initialize a $m_p$ dimensions vector $\mathbf{P_l} = [\{Inf\} \times m_p]$ recording the minimum counting error of each particle itself and a $m_p$ set $\mathbf{X_l} = \{\mathbf{X_{l1}}, \dots, \mathbf{X_{lm_p}}\} = \{x_1, \dots, x_{m_p}\}$ recording the corresponding state of each particle;
Initialize the global minimum counting error $P_g = Inf$ and the corresponding global optimal particle state $\mathbf{X_g} = x_1$;
for $j = 1, \dots, I$ do
    for $i = 1, \dots, m_p$ do
        Calculate the fitness value of the $i$th particle:
            $FV(x_i) = \sum_{k=1}^{n_{tr}} (x_i^{1:MN} \odot s_k + x_i^{(MN+1):2MN} - c_k^{gt})^2$;   (6)
        if $FV(x_i) < P_l[i]$ then
            $\mathbf{P_l}[i] = FV(x_i)$;$\mathbf{X_l}\{i\} = x_i$;
        end
        if $FV(x_i) < P_g$ then
            $P_g = FV(x_i)$;$\mathbf{X_g} = x_i$;
        end
        Update the speed and state of the $i$th particle:
            $v_i = wv_{i-1} + C_1 \cdot random(0,1)(X_{li} - x_i) + C_2 \cdot random(0,1)(X_g - x_i)$;   (7)
            $x_i = x_{i-1} + v_i$;   (8)
            $x_i^{1:MN} = |x_i^{1:MN}|$   (9)
    end
end
return $\mathbf{X_g}$;
```

A. PSO-Based SCC

We modify the popular multi-objective optimization PSO [43] as the SCC of our approach, termed PSO-SCC, and then give its training process in Algorithm 2. The a_{ij} and b_{ij} of all girds for a WSTM are concatenated as the state of a particle, and the speed of each particle with the same size is used to control the direction and degree when optimizing.

There are three stages in the algorithm: 1) In the initializing stage, states of all particles are initialized by a, b obtained by using the least square method to fit the crowd count and the number of detected devices in the training data. 2) In the fitness calculating stage, we use the sum of squared counting errors (6) as the fitness function, where "$\odot$" is the Hadamard product between slopes and crowd counts in each grids of a WSTM. Then, the fitness values of all particles are calculated, and both the state of each particle when it reaches the minimal fitness values of itself and the state of the particle that reaches the global optimal fitness value as far are recorded. 3) In the updating stage, the speed of each particle is dynamically updated according the speed in the last iteration, the difference between the particle's optimal state and current state, and the difference between the global optimal state and current state. At last, the state is renewed by adding the updated speed, aiming to avoid the local optimum and thus find a more likely optimal solution. In addition, to avoid the a_{ij} becomes a negative value, we use the absolute values of $x_i^{1:MN}$.

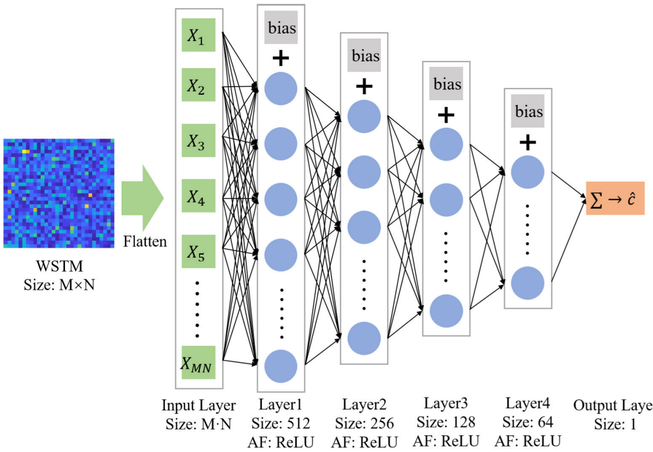


Fig. 3. The architecture of the proposed DNN-SCC. AF is short for the activation function.

B. DNN-Based SCC

To design the DNN-SCC, we utilize a simple but effective triangle-shaped fully connected DNN model with 4 hidden layers, as shown in Fig. 3. The size of hidden layers are set much larger than the possible partition granularity $M \times N$, i.e., 512/256/128/64, in order to extend the model's ability for applying in different scale AOIs. Rectified linear unit (ReLU) [44] is adopted as the non-linear activation function in every node. To satisfy the requirement of one-dimensional input for DNN-SCC, the WSTM will be flattened as a vector with the size of $M \cdot N$, and the outputted crowd count is directly obtained by summing all values in the last hidden layer. The forward propagation process can be indicated as

$$\mathbf{X}^{(i+1)} = ReLU(\mathbf{W}^{(i)}\mathbf{X}^{(i)} + \mathbf{B}^{(i)}), \quad (10)$$

where $\mathbf{X}^{(i+1)}$ denotes the outputs in the $(i+1)$th layer, and $\mathbf{W}^{(i)}$, $\mathbf{X}^{(i)}$, $\mathbf{B}^{(i)}$ denote the weights, inputs, biases of the ith layer, respectively. Particularly, the first inputs $\mathbf{X}^{(0)} = Flatten(\mathcal{S}^{\Delta \mathbf{t}})$ represents feeding WSTM into the model, and the outputted count is a scalar. As for training, the MSE loss is utilized to minimize the counting error, defined as

$$mseloss = \sum_{i=1}^{n_{tr}} (\hat{c}_i - c_i^{gt})^2, \quad (11)$$

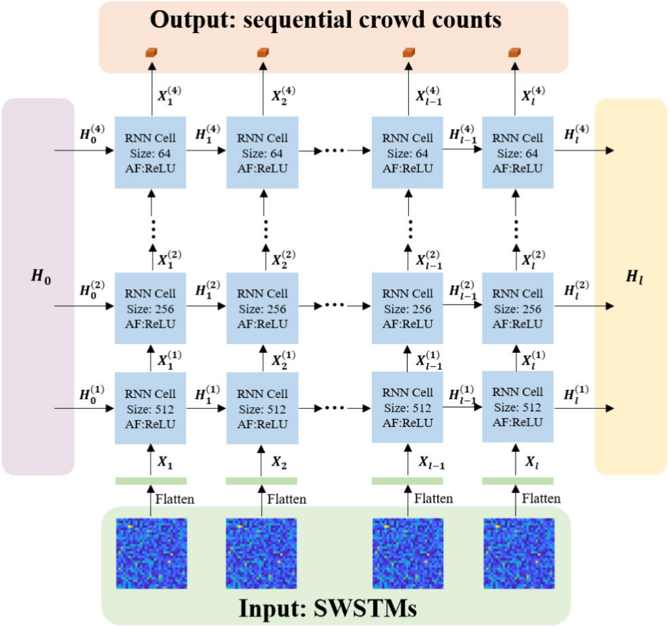


Fig. 4. The architecture of the proposed RNN-SCC. AF is short for the activation function.

where $\hat{c}_i$ and c_i^{gt} are the estimated and ground-truth crowd count of the ith training WSTM, and n_{tr} is the number of training WSTMs.

C. RNN-Based SCC

To further including the temporal correlation, we modify RNN model [45] to RNN-SCC and give the architecture in Fig. 4. The RNN-SCC utilizes RNN cells to replace the nodes in each layer of the traditional DNN model, resulting in absorbing the merit of full connection and hidden state across different time steps, such that it can effectively capture the spatial and temporal correlation simultaneously. As for the architecture, we deliberately leverage the same structure of DNN-SCC for designing RNN-SCC in the vertical direction as well as the activation function and loss function, so as to highlight the difference in temporal domain compared to DNN-SCC. Besides the forward propagation between RNN cells in each layer, a dynamic update mechanism between time steps (WSTMs in the SWSTM) is added into the model, and thus the synthetic forward propagation process is expressed as

$$\mathbf{H}_t^{(i+1)} = ReLU(\mathbf{U}^{(i)}\mathbf{H}_{t-1}^{(i+1)} + \mathbf{W}^{(i)}\mathbf{X}_t^{(i)} + \mathbf{B}_h^{(i)}), \tag{12}$$
$$\mathbf{X}_t^{(i+1)} = ReLU(\mathbf{V}^{(i)}\mathbf{X}_t^{(i)} + \mathbf{B}_x^{(i)}), \tag{13}$$

where $\mathbf{H}_t^{(i+1)}$ and $\mathbf{X}_t^{(i+1)}$ are the hidden states and the outputs of the $(i+1)$th layer and tth time step, $\mathbf{B}_h^{(i)}$ and $\mathbf{B}_x^{(i)}$ are corresponding biases, and $\mathbf{U}^{(i)}$, $\mathbf{V}^{(i)}$ and $\mathbf{W}^{(i)}$ are corresponding weights, respectively. In addition, the sequential outputs only reserve the last one $\mathbf{X}_l^{(4)}$ as the estimated count.

V. Evaluation

To validate the effectiveness and the actual counting performance of the proposed approach, a dataset is collected in a real large-scale scenario, and extensive experiments are conducted.

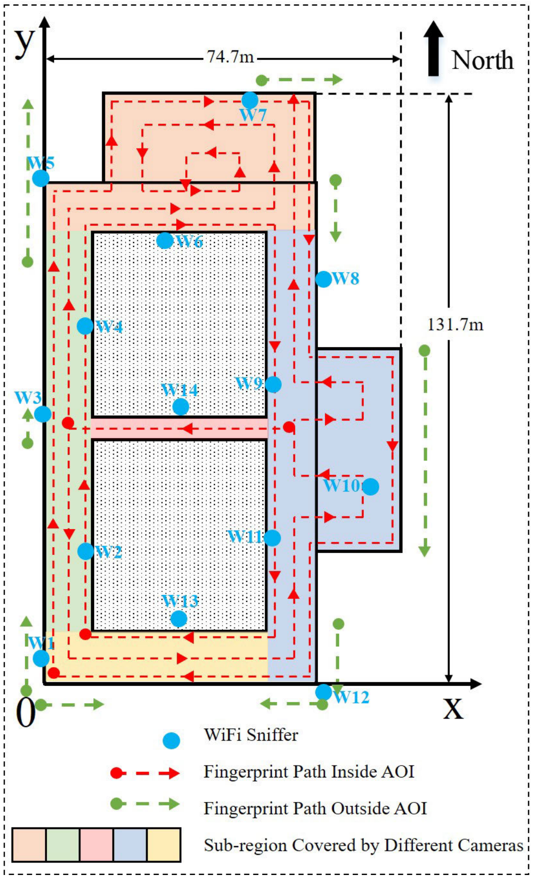

y
74.7m
North
131.7m
0
x
W1
W2
W3
W4
W5
W6
W7
W8
W9
W10
W11
W12
W13
W14
WiFi Sniffer
Fingerprint Path Inside AOI
Fingerprint Path Outside AOI
Sub-region Covered by Different Cameras

Fig. 5. The layout of our testbed.

A. Testbed

We conduct experiments in a road network of our campus with the surveillance area of about $4000m^2$, and the layout is shown in Fig. 5. The AOI is further expended to the minimum rectangle, i.e., $H \times W = 132m \times 75m$, to ensure the whole surveillance area (colored regions) being included. There are 14 customized WiFi sniffers (blue circles) uniformly deployed in the AOI, and numbered as $W1$ to $W14$. Fingerprints are collected from pathes both inside the AOI (red dashed line) and outside the AOI (green dashed line) to constitute the FD and train the In/Out AOI classifier. To obtain the ground truth of crowd counts, 5 smartphones installed on shelves with the height of $2.1m$ are strictly synchronized and deployed in each corner of the testbed to record videos of their corresponding sub-regions (five different colored regions), and then the crowd counts are manually annotated for one frame per second.

B. Dataset

An offline survey using 6 different smartphones with fixed MAC address is conducted, and fingerprints are generated within a sliding window with the size of $\Delta t = 60s$ and step of $1s$. As a result, we collect 8290 fingerprints to train the In/Out AOI classifier and it can achieve the accuracy of over 95% by a simple test. Then, 292 RPs are discretized from the pathes inside the surveillance area, and the KNN algorithm with $k = 3$ is utilized for localization. Finally, after hours of collection during a time period encompassing the peak time after classes, total $2280s$ intact sensing data and corresponding count labels are obtained. The total crowd flow is 145, 617 and the average crowd count in the whole AOI per second is 64.

As for WSTMs/SWSTMs, the partition granularity is set as $M \times N = 18 \times 10$ for all SCCs. To fairly compare baselines, we uniformly divide the dataset with size of $2280s$ into 30 consecutive timeslots, and take 15 of them as training data and the others as testing data to ensure that they are completely isolated. A sliding sequence window with size of $l = 30$ is adopted when constructing SWSTMs.

C. Baselines, Parameters, Metrics and Assumptions

Three traditional global linear or approximately linear regression methods that use the least square method to fit the proportional function $c^{gt} = a \cdot \sum \mathcal{S}^{\Delta \mathbf{t}}$ (denoted by G-PROP) [19], [21], the linear polynomial function $c^{gt} = a \cdot \sum \mathcal{S}^{\Delta \mathbf{t}} + b$ (denoted by G-LPOLY) [20], and the second-degree polynomial function $c^{gt} = a \cdot (\sum \mathcal{S}^{\Delta \mathbf{t}})^2 + b \cdot \sum \mathcal{S}^{\Delta \mathbf{t}} + c$ (denoted by G-SPOLY) [22], between the ground-truth and the number of detected mobile devices located in the AOI, are implemented for comparison. All baselines use our pre-processed data and In/Out AOI classifier, resulting in a slight improvement on counting accuracy compared to their original versions.

The parameters in the PSO-SCC are set as: $m_p = 100$; $v_{th} = 5$; $I = 1000$; $w = 0.95$; $C_1 = C_2 = 2$, according to field experiments of PSO [43] and our attempts on parameters optimizing. Both the DNN-SCC and RNN-SCC use Adam [46] as their optimizers, and are trained by 1000 times. Due to the differences between the structures (non-sequential VS sequential) and the inputs (WSTMs VS SWSTMs) of two models, different learning rates are set to achieve their respective optimums, and thus $lr_{DNN} = 1e-4$ and $lr_{RNN} = 2e-5$.

Three evaluation metrics are adopted to comprehensively evaluate the counting performance, i.e., mean absolute error (MAE), mean square error (MSE) and mean relative error (MRE), which are defined as follows,

$$MAE = \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} |\hat{c}_i - c_i^{gt}|, \tag{14}$$

$$MSE = \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} (\hat{c}_i - c_i^{gt})^2, \tag{15}$$

$$MRE = \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \frac{|\hat{c}_i - c_i^{gt}|}{c_i^{gt}} \times 100\%, \tag{16}$$

where n_{te} is the total number of testing data, and $\hat{c}_i$ and c_i^{gt} are the estimated crowd count and ground-truth crowd count of the whole AOI, respectively. Therein, MAE and MRE reflect the counting accuracy, while MSE shows the counting stability.

In addition, considering that certain strict assumptions need to apply for both our approach and baselines to take place, we summarize a list of assumptions in the following:

- Every point in the AOI must be covered by at least 3 non-collinear WiFi sniffers.
- Every pedestrian pauses or stops short enough to ensure his/her device(s) not being judged as wireless AP(s).
- Every WiFi sniffer upload its sensing data in real-time and the time asynchronism does not exist.
- The pathes outside the AOI are surveyed enough, such that the In/Out AOI classifier can achieve a high accuracy.
- The size of sliding time window is large enough, such that most of mobile devices inside the AOI can be detected.

TABLE I

The Counting Results of Three Baselines and Our Approach With Different SCCs

Method	Metrics		
	MAE	MSE	MRE
G-PROP	14.17	359.16	22.54%
G-LPOLY	13.11	296.96	23.86%
G-SPOLY	13.13	298.38	23.96%
PSO-SCC	10.34	186.75	20.63%
DNN-SCC	8.45	**119.5**	15.79%
RNN-SCC	**7.55**	119.53	**13.44%**

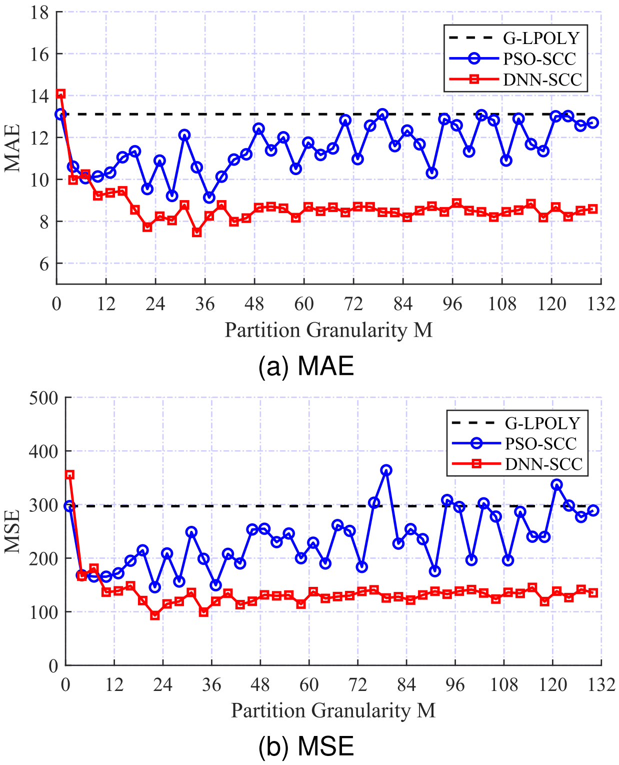
D. Validation of the Crowd Counting Performance

Based on above setups, the counting results of baselines and our approach with different SCCs are summarized in Table I. It can be seen that G-SPOLY and G-LPOLY have a approximate counting accuracy due to the second-degree coefficient (7.88×10^{-6}) of the former is close to 0, which is attribute to the approximately linear characteristic bringed by the huge amount of sensing data within a relative large time window, as we mentioned in the Introduction. Overall, our approach embedded with different SCCs fully exceeds baselines regarding all three metrics by a large margin. Particularly, the PSO-SCC, DNN-SCC and RNN-SCC can reduce the MAE by 21.13%, 35.55% and 42.41% under the same condition, respectively, compared to the best baseline, i.e. G-LPOLY.

For more in-depth discussions, PSO-SCC slightly outperforms baselines by optimizing parameters in LLMs, and it demonstrates the effectiveness of the fine-grained mapping bringed by area partitioning. DNN-SCC has a greater ability than PSO-SCC for mapping WSTMs to counts since the strong fitness ability bringed by the non-linear activation functions and multi-layer structure, and can effectively exploits the spatial correlations among the crowd distribution in each grid due to the full connection. Furthermore, RNN-SCC which has the same vertical structure with DNN-SCC, are further incorporated with horizontal time series to capture the temporal correlations among crowd distributions in adjacent WSTMs of an SWSTM, and transmit them in the whole sequence, resulting in a more accurate counting.

E. Ablation Studies

We conduct the following experiments to further explore how the spatial granularity, the localization and the time sequence length affect the counting accuracy of our approach.



(a) MAE

(b) MSE

Fig. 6. The change curves of the MAE and MSE of our approach with PSO-SCC and DNN-SCC when the partition granularity $M \times N$ increasing.

1) Spatial Granularity: To investigate the variation tendency of counting accuracy with the partition granularity of WSTM varying, we gradually increase M from 1 to $H = 132$ and let $N = round(\frac{W}{H} \cdot M)$. G-LPOLY is adopted as the baseline, and MAEs and MSEs of the PSO-SCC and DNN-SCC which counts mainly relying on the spatial domain, are given in Fig. 6. From the figures, we can conclude that:

First, change curves of both SCCs are fluctuant, which is mainly attributed to the sparsity and randomness of the crowd distribution in the real-world large scenario, rendering the discontinuous tendency. When $M \times N$ is large, the drastic fluctuation of PSO-SCC also indicates its limited ability for optimizing such a large number of parameters, while DNN-SCC shows the superiority of DL.

Second, as we discussed in Section III-C, a suitable partition granularity should be leveraged to balance the limited number of people/devices in grids and the localization errors, and the optimal $M \times N$ is about 18×10 in our testbed. The subsequent curve of PSO-SCC is well accord with our discussion, while DNN-SCC can reach a state of equilibrium, which is not strictly conform to the discussion due to its non-linear fitness ability that aggregates pony-size girds.

Third, MAEs of PSO-SCC are always lower than those of G-LPOLY, because we use the solution of G-LPOLY to initialize PSO-SCC, while the two outliers of PSO-SCC's MSEs are attributed to the difference between the training and testing data. In addition, we also find that DNN-SCC obtains a worse result when $M \times N$ is small, i.e. 2×1 and 1×1, which also verifies the simple utilization of the count obtained by WiFi sniffers is not enough for high-precision counting, even estimating by the DNN model.

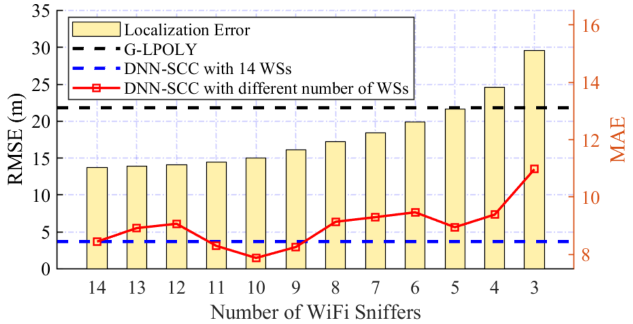


Fig. 7. The variation of the localization error (RMSE) and counting accuracy (MAE) when gradually removing WiFi sniffers. WS is short for WiFi sniffer.

2) Localization Error: As a key component of our approach, WiFi localization plays a vital role in boosting the counting performance. Therefore, we intentionally remove 1 to 11 WiFi sniffers out of all 14 ones, so as to weaken the localization ability and increase the localization error. The root mean square error (RMSE) of localization error, which is obtained by 2 testing devices and the FD constructed by 4 distinct reference devices, is calculated by

$$RMSE = \sqrt{\frac{1}{n_{tf}} \sum_{i=1}^{n_{tf}} (\hat{\mathbf{l}}_i - \mathbf{l}_i^{real})^2}, \tag{17}$$

where n_{tf} is the number of testing fingerprints, and $\hat{\mathbf{l}}_i$ and $\mathbf{l}_i^{real}$ are the estimated and real locations of the ith testing fingerprint's device. The removing order of Wifi sniffers is set as: $W9 \rightarrow W14 \rightarrow W1 \rightarrow W4 \rightarrow W6 \rightarrow W3 \rightarrow W11 \rightarrow W12 \rightarrow W8 \rightarrow W5 \rightarrow W13$, in order to gradually increase the RMSE. At last, we take the G-LPOLY and DNN-SCC with 14 WiFi sniffer as baselines, and plot RMSEs and MAEs of DNN-SCC under the schemes of different numbers of WiFi sniffers in Fig. 7. When the number of WiFi sniffers ≥ 4, MAEs slightly fluctuates around the MAE of DNN-SCC with all WiFi sniffers, which well conforms to our discussions in Section III-C, i.e., the compensation effect of localization and the robustness of our approach. Particularly, in the extreme case, DNN-SCC with only 3 WiFi sniffers still outperforms G-LPOLY on counting accuracy.

3) Time Sequence Length: Another key parameter for constructing SWSTM is the time sequence length l, and thus we test RNN-SCC under the same setup above but varying l from 10 to 120. Since the SWSTM with the length of l contains the sensing data within the timeslot with the length of $\Delta t + l - 1 \approx \Delta t + l$, and thus the cases of DNN-SCC using WSTMs within $\Delta t = (60 + l)s$, are also tested. On these grounds, DNN-SCC using WSTMs within $60s$ is taken as the baseline, and all MAEs are plotted in Fig. 8. According to the results, we conduct the following analyses:

First, we reasonably believe that continually extending l will gradually promote the counting accuracy of RNN-SCC, but we find the MAE will increase when l is larger than 90. We guess this phenomenon is attributed to that the temporal correlation is decreasing when the sensing data is far away from the current moment, and the LSTM or gated recurrent unit (GRU) models may be helpful to relieve it by balancing the long and short term temporal dependencies. In addition,

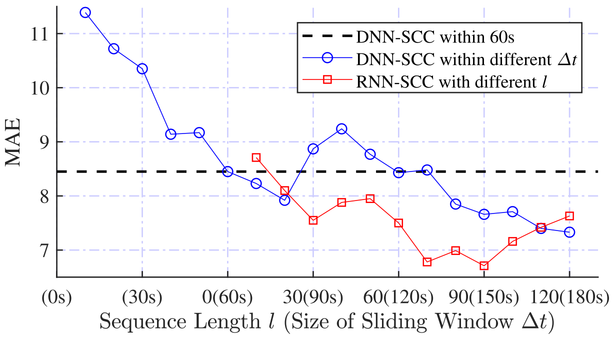


Fig. 8. The MAEs of DNN-SCC within different Δt and RNN-SCC with different l.

RNN-SCC will get a poor result when l is less than about 15, which means that lesser sequence of WSTMs cannot provide enough information for counting, but the added features ruins the counting performance on the contrary.

Second, for cases of DNN-SCC within different Δt, it can be seen that the MAE is continuously decreasing when Δt is less than or equal to $80s$, but it will increase when $\Delta t = 90s$ to $120s$. This is because the pedestrians who just already pass the AOI are included in the WSTM when Δt is large, degrading parameters that map WiFi counts to real counts. Curiously, the MAE can be further reduced by enlarging Δt, and we attribute it to that counting passersby for all WSTMs can reduce noise estimates and the offsets insides the model can then eliminate the redundant count.

Third, to sum up, using the SWSTM by RNN-SCC surpasses handling single WSTM constructed within the same length timeslot of sensing data by DNN-SCC in most cases, since the later one will lose the temporal correlation between crowd distribution of consecutive moments by locating and constructing one WSTM for such a large time window. Another advantage of SWSTMs is that the server only need to deal a small amount of sensing data (within a relative small Δt) including positioning and constructing the WSTM for the current moment, and obtain WSTMs of previous $l-1$ moments in the RAM, which can improve the efficiency.

VI. Conclusion

In this paper, we proposed a passive WiFi sensing-based crowd counting approach for large-scale surveillance areas, which involved a series of advanced technologies including the WiFi localization, optimization theory and DL. The proposed method is a universal and flexible solution for the WiFi-based crowd counting since it has the merits of low-cost and easy-to-deploy, and the alternative components are convenient for relevant researchers or practitioners adapting and modifying the approach in their practices. Extensive experiments in a real scenario demonstrated that our preliminary configurations of the approach is successful, and the ablation studies also given meaningful results about how some key factors influence the counting performance.

In the next step, we plan to extend our approach for providing more indexes on the crowd analyses, such as the crowd density, the flow speed, and the trajectory of a crowd. In addition, more advanced localization method and DL technologies, e.g., the attention mechanism and transformer model, also deserve to be explored.

References

[1] C. C. Loy, K. Chen, S. Gong, and T. Xiang, *Crowd Counting and Profiling: Methodology and Evaluation*. New York, NY, USA: Springer, 2013, pp. 347–382.

[2] V. A. Sindagi and V. M. Patel, "A survey of recent advances in CNN-based single image crowd counting and density estimation," *Pattern Recognit. Lett.*, vol. 107, pp. 3–16, May 2018.

[3] 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.

[4] L. Durán-Polanco and M. Siller, "Crowd management COVID-19," *Annu. Rev. Control*, vol. 52, pp. 465–478, Jan. 2021.

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

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

[7] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 833–841.

[8] 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.

[9] 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. Hum.-Mach. Syst.*, vol. 45, no. 4, pp. 442–452, Aug. 2015.

[10] J. Weppner and P. Lukowicz, "Bluetooth based collaborative crowd density estimation with mobile phones," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. (PerCom)*, Mar. 2013, pp. 193–200.

[11] T. Yoshida and Y. Taniguchi, "Estimating the number of people using existing WiFi access point in indoor environment," in *Proc. 6th Eur. Conf. Comput. Sci. (ECCS)*, 2015, pp. 46–53.

[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, Sep. 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. (ISPA/IUCC)*, Dec. 2017, pp. 967–974.

[15] Y. Fukuzaki, M. Mochizuki, K. Murao, and N. Nishio, "A pedestrian flow analysis system using Wi-Fi packet sensors to a real environment,' in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput., Adjunct Publication*, 2014, pp. 721–730.

[16] Y. Li, J. Barthelemy, S. Sun, P. Perez, and B. Moran, "A case study of WiFi sniffing performance evaluation," *IEEE Access*, vol. 8, pp. 129224–129235, 2020.

[17] 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: Association for Computing Machinery, 2016, pp. 15–20.

[18] 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.

[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.*, 2015, pp. 1519–1526.

[20] 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., Adjunct*, Sep. 2016, pp. 1363–1371.

[21] 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., Services*, 2018, pp. 337–349.

[22] 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.

[23] T. Zang, Y. Zhu, Y. Xu, and J. Yu, "Jointly modeling spatio-temporal dependencies and daily flow correlations for crowd flow prediction," *ACM Trans. Knowl. Discovery Data*, vol. 15, no. 4, pp. 1–20, Mar. 2021.

[24] Y. Miao, J. Han, Y. Gao, and B. Zhang, "ST-CNN: Spatial–temporal convolutional neural network for crowd counting in videos," *Pattern Recognit. Lett.*, vol. 125, pp. 113–118, Jul. 2019.

[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, pp. 1–23, Mar. 2017.

[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*, Dec. 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. G. 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.*, Jun. 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.*, Jun. 2013, pp. 2539–2546.

[30] 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. 15233–15243, Sep. 2022.

[31] 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.

[32] 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.

[33] F. Dittrich, L. E. 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*.

[34] M. Nakatsuka, H. Iwatani, and J. Katto, "A study on passive crowd density estimation using wireless sensors," in *Proc. Int. Conf. Mobile Comput. Ubiquitous Netw.*, Jan. 2008, pp. 1–6.

[35] M. Seifeldin, A. Saeed, A. E. Kosba, A. El-Keyi, and M. Youssef, "Nuzzer: A large-scale device-free passive localization system for wireless environments," *IEEE Trans. Mobile Comput.*, vol. 12, no. 7, pp. 1321–1334, Jul. 2013.

[36] Y.-K. Cheng and R. Y. Chang, "Device-free indoor people counting using Wi-Fi channel state information for Internet of Things," in *Proc. GLOBECOM-IEEE Global Commun. Conf.*, Dec. 2017, pp. 1–6.

[37] Q. Jiang, K. Li, M. Zhou, Z. Tian, and M. Xiang, "Competitive agglomeration based KNN in indoor WLAN localization environment," in *Proc. 10th Int. Conf. Commun. Netw. China (ChinaCom)*, Aug. 2015, pp. 338–342.

[38] L. Yen, C.-H. Yan, S. Renu, A. Belay, H.-P. Lin, and Y.-S. Ye, "A modified WKNN indoor Wi-Fi localization method with differential coordinates," in *Proc. IEEE ICASI*, May 2017, pp. 1822–1824.

[39] 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.

[40] 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, Aug. 2020, Art. no. 102673.

[41] C. Li, Q. Xu, Z. Gong, and R. Zheng, "TuRF: Fast data collection for fingerprint-based indoor localization," in *Proc. Int. Conf. Indoor Positioning Indoor Navigat. (IPIN)*, Sep. 2017, pp. 1–8.

[42] H. Zou, B. Huang, X. Lu, H. Jiang, and L. Xie, "A robust indoor positioning system based on the Procrustes analysis and weighted extreme learning machine," *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1252–1266, Feb. 2016.

[43] I. C. Trelea, "The particle swarm optimization algorithm: Convergence analysis and parameter selection," *Inf. Process. Lett.*, vol. 85, no. 6, pp. 317–325, Mar. 2003.

[44] K. Hara, D. Saito, and H. Shouno, "Analysis of function of rectified linear unit used in deep learning," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2015, pp. 1–8.

[45] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra, "DRAW: A recurrent neural network for image generation," in *Proc. ACM ICML*, 2015, pp. 1462–1471.

[46] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–15.

Lifei Hao received the B.S. degree in applied physics from Chongqing University, Chongqing, China, in 2012, and the M.E. degree in computer technology from Inner Mongolia University, Hohhot, China, in 2019, where he is currently pursuing the Ph.D. degree with the College of Computer Science. His main research interests include the Internet of Things, WiFi localization, and passive WiFi sensing.

Baoqi Huang (Member, IEEE) received the B.E. degree in computer science from Inner Mongolia University (IMU), Hohhot, China, in 2002, the M.S. degree in computer science from Peking University, Beijing, China, in 2005, and the Ph.D. degree in information engineering from The 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 a recipient of the Chinese Government Award for Outstanding Chinese Students Abroad in 2011.

Bing Jia (Member, IEEE) received the Ph.D. degree from Jilin University, 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.

Gang Xu (Member, IEEE) received the M.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 2007, and the Ph.D. degree in computer science from Inner Mongolia University (IMU), Hohhot, China, in 2016. He is currently a Associate Professor with the College of Computer Science, IMU. His research interests include wireless sensor networks, DTN, and mobile computing.

Guoqiang Mao (Fellow, IEEE) received the Ph.D. degree in telecommunications engineering from Edith Cowan University, Australia, in 2002.

He is a Distinguished Professor and the Dean of 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 over 200 papers in international conferences and journals, which have been cited more than 9000 times. His research interests include intelligent transport systems, applied graph theory and its applications in telecommunications, the Internet of Things, wireless sensor networks, wireless localization techniques, and network modeling and performance analysis.

Prof. Mao is a fellow of IET. He received the Top Editor Award for outstanding contributions to the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY in 2011, 2014, and 2015. He is the Co-Chair of the IEEE Intelligent Transport Systems Society Technical Committee on Communication Networks. He has served as the chair, the co-chair, and a TPC member for number of international conferences. He has been an Editor of the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, since 2018, and IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, since 2010. He was an Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, from 2014 to 2019,