ORIGINAL ARTICLE

Round trip time meets transformers: high-fidelity human counting in cluttered environments

Haruki Yonekura[1] · Hamada Rizk[1,2,3] · Hirozumi Yamaguchi[1,3]

Received: 18 December 2024 / Accepted: 23 July 2025

Abstract

Accurate human counting in indoor environments is essential for optimizing people-centric applications, such as crowd management, disaster response, and monitoring in settings like shopping malls and healthcare facilities. Traditional vision approaches face challenges with poor lighting conditions and raise privacy concerns. WiFi-based solutions enable device-free human counting by detecting disruptions in wireless signals caused by human presence. However, methods using received signal strength indicator are unreliable due to physical obstructions, multipath fading, radio interference, and fluctuating access point power. While WiFi channel state information-based systems are more sensitive to environmental changes, they lack standardization, limiting their practicality. To overcome these limitations, this paper presents *Time4Count*, an innovative device-free indoor human counting system that leverages round trip time measurements to achieve high accuracy and scalability. *Time4Count* capitalizes on human-induced fluctuations in signal propagation time to accurately estimate the number of individuals in a space. By employing a multivariate transformer-based feature extraction method, the system effectively mitigates non-line-of-sight errors and signal distortions, ensuring robust performance even in cluttered indoor environments. Additionally, *Time4Count* integrates spatial discretization and multi-label classification techniques, enabling it to count an unlimited number of individuals in real-time. The system was rigorously evaluated in two realistic, cluttered environments using commodity hardware, involving up to 15 participants. Experimental results reveal that *Time4Count* achieves an high counting accuracy of 92.7%. To our knowledge, *Time4Count* is the first RTT-based indoor counting system, providing a precise solution for indoor monitoring. Implementation is available at: https://github.com/mclab-osaka/time4count.

Keywords People counting · Round trip time · Multivariate transformer · Multi-label classification · Cluttered environments

1 Introduction

Tracking the number of people within a specified area is essential for various applications, including intelligent guidance in museums, energy efficiency in intelligent buildings, indoor analysis, and emergency evacuations. In retail environments, for instance, lighting and climate control can be adjusted automatically based on customer concentration, thereby enhancing both energy efficiency and customer experience. Analyzing visitor traffic also aids strategic planning for retail spaces.

The significance of people counting has garnered research interest, particularly in computer vision, where deep learning advancements have led to accurate counting solutions. These systems, often based on convolutional neural networks (CNNs), process images, or videos to estimate crowd density [1–4]. However, these methods

https://doi.org/10.1007/s00521-025-11540-8

Published online: 29 August 2025

often rely on specialized hardware, suffer from environmental conditions like lighting, and face significant challenges in privacy-sensitive contexts, especially in spaces where cameras or other visual-based systems are not practical. The need for cost-effective, privacy-preserving, and robust solutions in cluttered environments underlines the importance of exploring alternative technologies such as WiFi sensing.

Recent advancements in WiFi link blockage analysis [5–8] have introduced innovative approaches to people counting. These methods utilize the interruption of WiFi signals by human bodies to estimate presence and count individuals. However, their accuracy can vary under real-world conditions due to simplified assumptions. The received signal strength indicator (RSSI) of WiFi signals, significantly influenced by human body blockages, presents a potential avenue for device-free sensing [9, 10]. However, RSSI-based systems face challenges such as signal degradation from physical obstructions, multipath fading, radio interference, and variable transmission power of access points, all of which can compromise system performance [11–15]. For example, the work done by Januszkiewicz et. al. [11] verified that placing a single adult in the line-of-sight lifts the median path loss in the 2.4 GHz band by 9.5 dB, and that three adults push the penalty to about 11 dB, implying that designers must reserve on the order of ten decibels of fade margin whenever people may occlude the link. In addition, the RSSI received by receiver strongly depend on the device itself [16]. Conversely, WiFi channel state information (CSI) has been explored by several systems [17–20] for its sensitivity to changes in radio waves, which can indicate human presence. However, the lack of standardization in CSI necessitates the use of specialized hardware or software to acquire it, rendering it impractical for numerous applications and limiting its utility in counting in diverse environments, and because every WiFi packet captures a full matrix of complex numbers for up to 256 subcarriers on each antenna, sampled hundreds of times per second, the raw logs can swell from megabytes to gigabytes in minutes.

Recently, time-based techniques have shown promising solutions, particularly in device-based settings. These techniques estimate the distance between a mobile device (e.g., smartphone) and access points by measuring the signal's propagation time and utilizing the known propagation velocity of the signal. Various approaches have been proposed for measuring propagation time, including time of arrival (ToA) [21], time difference of arrival (TDoA) [22], and RTT [23]. ToA and TDoA methods necessitate precise time synchronization among all devices, posing a challenge. In contrast, RTT utilizes the difference in recorded times to measure the time required for the signal to travel to a destination node and return, thereby mitigating the synchronization problem. Unlike RSSI-based methods, RTT demonstrates enhanced resilience against the challenges common in cluttered indoor environments, including multipath interference, signal attenuation, variations in transmission power, and radio interference. The fine time measurement (FTM) protocol, introduced in the IEEE 802.11mc-2016 standard, enables RTT measurements between mobile devices and access points. Its growing support from commercial access points and consumer devices has made it a practical and viable choice for indoor applications. Building on this foundation, the IEEE 802.11az-2023 standard further enhances these capabilities, delivering greater accuracy and robustness in time-based localization. It provides improved handling of challenges such as multipath effects and propagation latency, making RTT-based methods more reliable for precise indoor positioning and human sensing applications. However, RTT still faces limitations, including the potential for distance overestimation caused by indirect signal paths [24]. Approaches such as map matching and advanced filtering techniques have been explored to mitigate these issues [25]. Despite these challenges, time-based techniques, particularly those leveraging advancements in IEEE standards, continue to hold significant promise for practical indoor human sensing and localization [23, 26–28].

In this paper, we introduce *Time4Count*, an innovative human counting system that leverages the precision of RTT measurements to overcome the limitations of conventional counting methods. By utilizing RTT variations caused by human presence, *Time4Count* offers a fully device-free solution, relying solely on RTT data as input. This approach eliminates the need for additional hardware or invasive technologies, making it a practical and efficient choice for real-world applications. At the core of *Time4Count* is an efficient multivariate transformer-based feature extraction mechanism, meticulously designed to capture and interpret the intricate signal dynamics associated with human movement and environmental interference. This method significantly enhances the system's ability to decode complex patterns in RTT signals, ensuring robust performance even in cluttered or dynamic indoor settings. A defining feature of *Time4Count* is its ability to count an unlimited number of individuals without being restricted to predefined count classes. This flexibility is achieved through the integration of spatial discretization and multi-label classification mechanisms, which together enable precise detection and counting in diverse and unpredictable scenarios.

The system demonstrates a counting accuracy of 92.7%, showcasing its effectiveness and potential applicability in various indoor environments. To the best of our knowledge, this is the first counting system based on RTT, setting a new standard for accuracy and adaptability in the field of crowd management technologies.

Our contribution is described as follows: We leverage the 1D CNN and multivariate transformer encoder to extract features to predict the number of individuals present. To the best of our knowledge, this is the first work to apply transformer models to WiFi RTT data. Through this approach, we underscore the necessity of adapting models capable of handling time-series data for accurate prediction using WiFi signals. Secondly, we integrate a multi-label classification methodology, thereby transcending the confines of traditional people counting models and evolving towards an unlimited number counting framework. This adaptation allows for predicting varying numbers of individuals with enhanced flexibility and accuracy. Additionally, we substantiate our proposed system's efficacy by conducting extensive evaluations using data collected from readily available devices on the market in two realistic, cluttered environments. This empirical validation underscores the practical applicability and reliability of our approach in real-world settings.

This paper is organized as follows: Sect. 2 explains the concept of RTT. Section 3 reviews related work in RTT-based localization system and people counting technology and method. In Sect. 4, we explain the basic concept, and Sect. 5 presents an overview of our methodology. Section 6 describes the components our proposed system has and their role. Section 7 discusses the experimental setup and analyzes the results. Finally, Sect. 9 concludes the paper.

2 Background

RTT is defined as the time interval between the initiation of a network request at a source transmitter, its reception at a receiver, and the subsequent return of the response to the source transmitter. In this study, RTT is employed to measure the distance between two WiFi stations—specifically, commodity devices such as a smartphone and an access point.

The primary advantage of the RTT technique lies in its ability to estimate the distance between two stations without requiring synchronization. Synchronization is a critical challenge in time-based systems, and RTT's independence from synchronization simplifies its application. This capability is facilitated by the fine timing measurement (FTM) protocol, introduced in the IEEE 802.11mc standard, which provides native support for RTT measurements. The RTT-based distance estimation process begins with the smartphone, acting as the initiator, sending a WiFi signal to an access point, the receiver, to ascertain its availability. Upon receipt, the access point responds with an acknowledgment signal (ACK), initiating a two-way communication exchange. This exchange enables the smartphone to measure the RTT and compute the distance. Repeated measurements can be performed to enhance accuracy. Furthermore, RTT estimation is conducted for all access points within the smartphone's range, providing a comprehensive spatial analysis.

One benefit of RTT is its ability to compute distances locally on the edge device, thereby preserving user privacy. Unlike cloud-based processing methods, the localized computation reduces exposure to potential data breaches or misuse.

Figure 1 illustrates the FTM protocol workflow. The process starts with the smartphone transmitting an FTM request to the access point to confirm its availability. Upon confirmation, the access point transmits an ACK signal, enabling the smartphone to compute RTT through the exchange of multiple FTM packets.

The RTT value is calculated using the following equation:

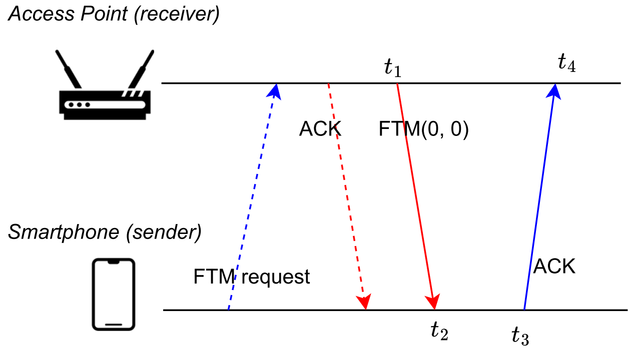


Fig. 1 Workflow of the IEEE 802.11mc fine timing Measurement(FTM) protocol for unsynchronized ranging. FTM protocol exchange between a smartphone (initiator) and an access point (responder). At t_1, the smartphone sends an FTM request (blue dashed arrow); the AP acknowledges receipt (red dashed arrow) and, at t_2, replies with an FTM frame (solid red arrow). The smartphone receives this frame at t_3 and returns an ACK (solid blue arrow) at t_4. The round trip time is then computed as $RTT = (t_4 - t_1) - (t_3 - t_2)$ from which the one-way distance estimate follows as $\frac{1}{2} RTT \times c$

$$RTT = t_4 - t_1 - (t_3 - t_2) \tag{1}$$

Here, $(t_3 - t_2)$ represents the processing time on the access point. The distance between the two devices is then computed as:

$$Distance = \frac{1}{2} RTT \times c \tag{2}$$

where c is light speed, whose value is about $3 \times 10^8 m/s$.

Because IEEE 802.11 mc/az Fine Timing-Measurement (FTM) exchanges discard a station's internal processing delay when computing the round trip time (RTT), transient queuing or back-off experienced by other devices does not bias the distance estimate itself.

Notably, the smartphone performs RTT measurements for all RTT-capable access points within its vicinity. Unlike traditional multi-lateration approaches [24, 29, 30], *Time4Count* leverages the collected RTT values via the FTM protocol as unique fingerprints to estimate the number of individuals present in indoor environments. This process is described in detail in the Sect. 6.

3 Related work

In this section, we present a survey of the relevant literature related to our *Time4Count* system.

3.1 RTT-based indoor localization systems

To the best of our knowledge, there are no existing RTT-based counting approaches. However, due to its resilience and robustness in indoor localization, the RTT technique gained more traction in recent years. WiNar [23] employs RTT fingerprints to develop a probabilistic model using Bayesian inference. This model estimates the likelihood of the user's presence at predetermined reference points, offering valuable insights for localization purposes. DeepNar [26] utilizes RTT data collected during an offline phase to train a multi-layer deep learning model functioning as a multi-class classifier. In the online phase, the user's device captures RTT measurements from nearby access points. It inputs them into the trained model, which then calculates the probability of the user's presence at the reference points. Conversely, RRLoc [31] demonstrates improved performance compared to previous systems based on received signal strength indication (RSSI) and RTT. It employs a hybrid approach that combines RSSI and RTT measurements, integrating them through a DeepCCA network to extract high-level features. These features are subsequently used to train a deep classification model for precise localization. MagTT [32] integrates magnetic field measurements with RTT to achieve submeter-level accuracy. Utilizing a CNN-LSTM architecture, MagTT illustrates the potential of fusing various sensor data for enhanced indoor localization accuracy. The authors of [33] proposed the robot localization system that fuses WiFi RTT information and improves the positioning error. This system employs adaptive data filtering and other positioning error smoothing methods to enhance the performance. LocFree [34] is the single-person localization method using only RTT in a static environment and has achieved a median localization error of 1.56m in an area of $5.8m \times 8.3m$ that consists of many kinds of furniture. Additionally, the RTT-based localization system developed by WhereArtThou [35] employs the extended Kalman filter with a random walk motion model (EKF-RW) and the step-and-heading-based filter (EKF-SH), which integrate distance measurements with inertial sensor readings to enhance accuracy. The final two references concentrate on the tracking of a single object or individual and do not address the tracking of multiple objects or people.

While RSSI-based systems benefit from not requiring specific hardware and utilizing signal strength for localization, they are vulnerable to obstacles, interference, and multipath effects. In contrast, RTT provides more precise distance measurements by directly capturing signal propagation time delay. The standardization of RTT by IEEE 802.11mc has made it widely available in commercial off-the-shelf (COTS) devices such as smartphones and access points.

The standardization of the RTT-based technique in the WiFi technology and the robustness of the time-based techniques gave the localization systems the ability to present an enhanced performance with fine-grained accuracy. Motivated by this, our work focuses on leveraging RTT measurements as features to recognize the presence of multiple people within an indoor environment and thus counting them in a device-free fashion.

3.2 Counting systems

There are many techniques for counting the number of people in both indoor and outdoor environments, leveraging a diverse array of sensors and their combinations. These techniques can be, in general, categorized into two main categories of methods: detection-based methods and regression-based methods.

STEERER [36] represents one of the detection-based and image-based methods designed to count and localize people, addressing the challenge of scale variation by cumulatively selecting and inheriting discriminative features from the optimal scale. Crowd++ [37] introduces an unsupervised method for detecting the number of speakers using microphones installed in smartphones. In [38], authors have tried to identify the number of people in the proximity scenario from Point cloud data with a combination of convolutional neural networks (CNN) and k-means clustering.

Meanwhile, MSCNN [39] and COUNT Forest [40] focus on crowd density estimation from image data utilizing regression mechanisms. These methods extract density maps from images and estimate density in crowded situations. An environmental sensor-based method, as described in [41], partially utilizes carbon dioxide concentration to understand the presence of individuals and optimize energy consumption by heating, ventilation, and air-conditioning (HVAC) systems. With their design to understand the number of people, they try to save energy consumed by heating, ventilation, and air-conditioning (HVAC) systems. Additionally, WiFree [42] utilizes channel state information (CSI) to achieve crowd counting with a classification model.

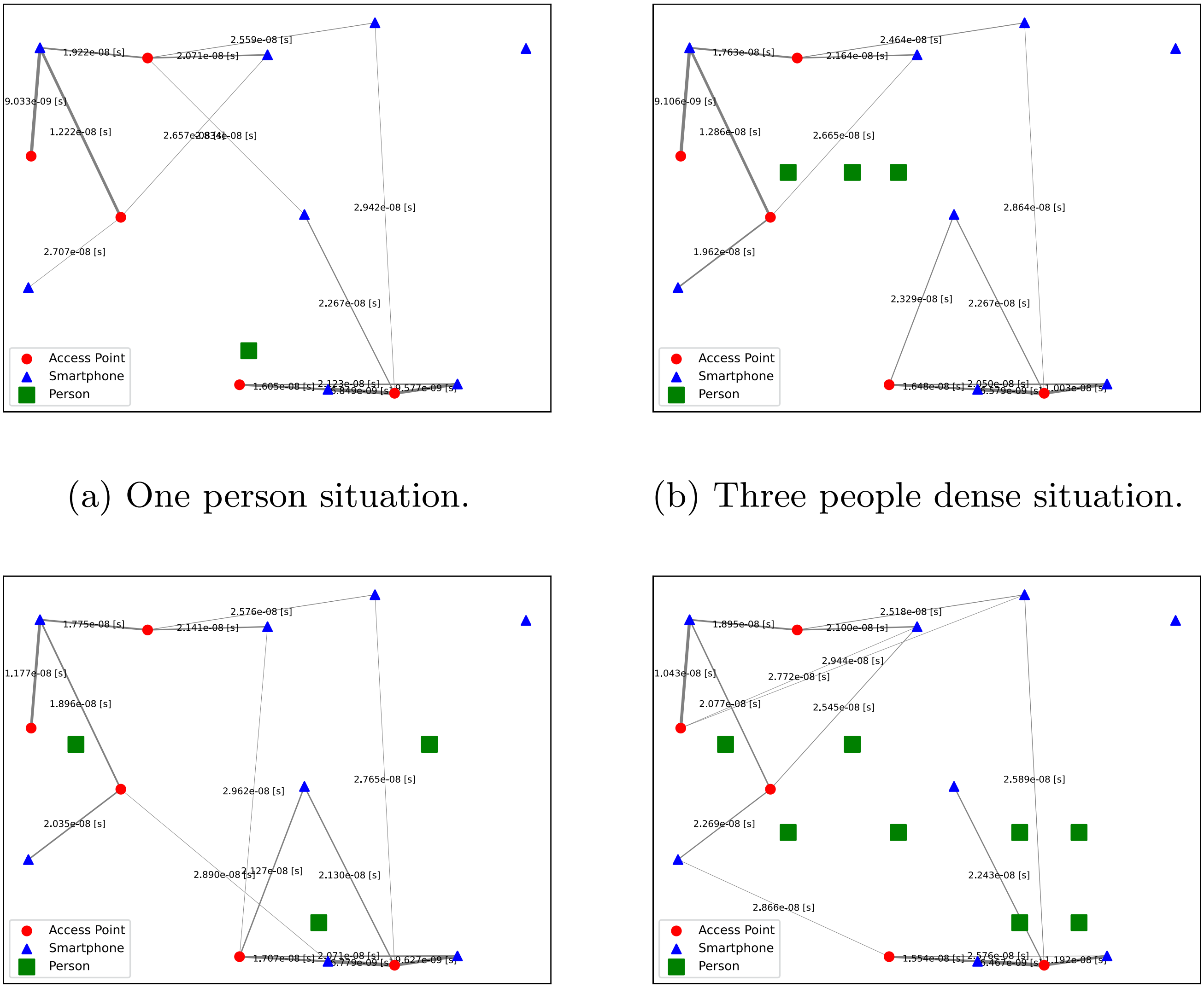
Furthermore, there are works that utilize WiFi CSI for people counting [39, 43–46]. Our method stands out by leveraging transformer-based classification, allowing for more flexible and accurate predictions across varying numbers of individuals. This novel approach enhances the reliability and adaptability of our people counting system, distinguishing it from previous methodologies.

https://doi.org/10.1007/s00521-025-11540-8

4 The basic idea

The core concept of *Time4Count* is to utilize the propagation time measurements between transmitters (e.g., access points) and receivers (e.g., smartphones) to accurately count users in various environments. *Time4Count* leverages the disruption caused by a user to the direct line-of-sight (LOS) path between a transmitter and a receiver. When a person obstructs this path, the signal is forced to take an indirect, non-line-of-sight (NLoS) route, resulting in increased travel time and decreased signal strength due to blockage by the human body. By analyzing the combined data from all NLoS and LoS paths for each transmitter–receiver pair, *Time4Count* can accurately map out the spatial presence of users.

The challenge arises from the diverse distribution and varying number of users, which complicates the development of a robust model capable of precise user counting. In environments cluttered with furniture or



(a) One person situation.

(b) Three people dense situation.

(c) Three people situation.

(d) Eight people situation.

Fig. 2 RTT data visualization of links with values less than 3.0×10^{-8} seconds (The bolder the line, the shorter time (LoS), which is the smaller RTT value)

densely populated with individuals in close proximity, Fig. 2 demonstrates that variations in RTT, resulting from signals navigating through these impediments, introduce additional layers of complexity.

To address these challenges, *Time4Count* employs a sophisticated feature extraction framework designed to analyze RTT measurements over time, leveraging both spatial and temporal characteristics of the data. This approach ensures a detailed understanding of signal variations and their underlying patterns, which is crucial for achieving robust performance in dynamic environments.

The methodology combines advanced techniques for capturing localized fluctuations in signal data and identifying sequential dependencies. Spatial analysis focuses on detecting subtle changes in the environment, such as obstructions or user movements, while temporal analysis emphasizes the order and evolution of measurements over time, enabling the differentiation of overlapping signals and mitigation of noise. By integrating these capabilities, *Time4Count* creates a comprehensive and high-dimensional representation of the data, effectively isolating noise and providing reliable insights into user interactions, as described in Sect. 6.

5 System overview

The architecture of the *Time4Count* system is depicted in Fig. 3, consisting of two main phases: the offline training phase and the online counting phase. In the offline phase, RTT measurements (affected by the density of humans) of different access points are captured by smartphones distributed throughout the target area while users occupy arbitrary reference points. This data collection leverages a **Data Recorder** mobile application, running on each WiFi-enabled device, utilizing the RTT API [47] to gather RTT readings. This data is then uploaded to a server for further processing. The **Pre-processor** module normalizes measurements and constructs pairs of fixed-size vectors from the RTT data, which represent the signals captured from access points within the vicinity. Subsequently, the **Feature Extractor** module processes these vector pairs to derive high-level, counting-discriminative features through a complex nonlinear transformation of the original signals into a new embedding space. This is facilitated by a transformer network, enhancing the distinction between user counts even their locations are arbitrary. In order to enable counting any number of users without pre-definied bound, the **Spatial Discretizer** module superimpose a virtual grid to the target environment to facilitate the recognition of users' presence in each discrete cell and thus counting them. The features are then utilized by the **Counting Model Builder** module, which trains a classification model to estimate the presence of different users and count them. The output of this phase includes the trained feature extraction and counting models, which are stored for subsequent retrieval during the online phase.

During the online phase, the system provides real-time estimations of the number of people present. It continuously collects and preprocesses RTT data, ensuring proper data normalization and shaping. This preprocessed data are then input into the pre-trained models to extract embeddings and to accurately count the number of users. This real-time estimation enables effective counting-based service in realistic, cluttered environments.

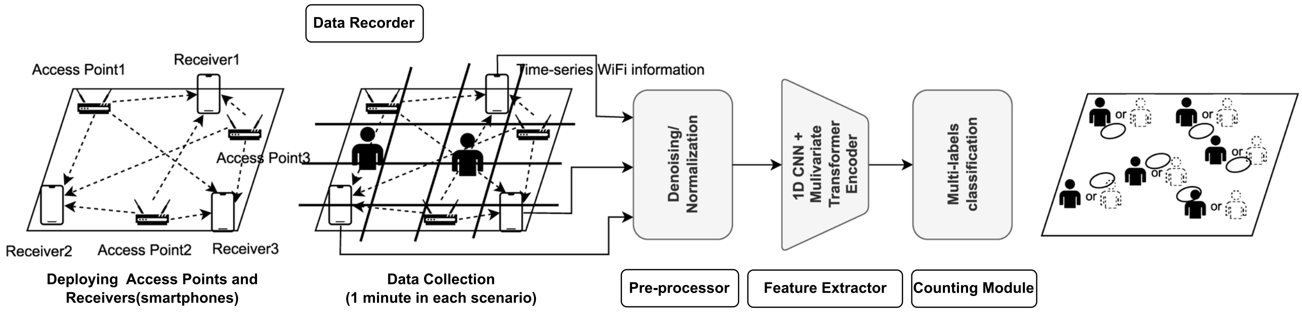

Data Recorder
Access Point1
Receiver1
Access Point3
Receiver2
Access Point2
Receiver3
Deploying Access Points and Receivers(smartphones)
ime-series WiFi information
Data Collection (1 minute in each scenario)
Denoising/ Normalization
1D CNN + Mulivariate Transformer Encoder
Multi-labels classification
Pre-processor
Feature Extractor
Counting Module
or
or
or
or
or
or

Fig. 3 System overview

6 System details

6.1 Deployment and data collection by Data Recorder

In our experimental setup, we distribute multiple access points and smartphones as signal receivers throughout the environment. Given practical constraints, such as limited resources, there will inevitably be a finite number of access points and smartphones available for deployment. To ensure comprehensive coverage, we locate these devices in a manner that maintains uniform spacing between each device type throughout the room. Subsequently, once installed, the locations of these devices remain fixed throughout the duration of the experiments.

6.2 The pre-processing module

The Pre-processing module is pivotal in both the offline training and online localization phases of *Time4Count*, focusing on data preparation and anomaly correction for deep learning applications. This module processes RTT data from smartphones to construct input vectors for machine learning models, with each vector element representing a signal measurement from a smartphone to an access point. Given the dynamic nature of real-world environments, not all access points are consistently detected in each scan, which leads to input vectors of variable lengths. To address this variability, undetected access points in a scan are assigned a placeholder RTT value, set to represent an improbably high distance—specifically, 0.2×10^{-3} ms in RTT, which exceeds typical values for access points within detection range. Additionally, the module removes anomalies such as the Android API reporting incorrect negative RTT values, likely due to internal calibration of WiFi cards or multipath effects. Such anomalies, including latency variations at reference points, can degrade traditional trilateration techniques. The module also standardizes the feature set by normalizing the input vector elements to a range of [0, 1]. This normalization is crucial for improving the effectiveness of the optimization algorithms during the training of the models, thereby enhancing the overall accuracy and robustness of *Time4Count* in user counting in cluttered environments.

6.3 Spatial discretization methodology

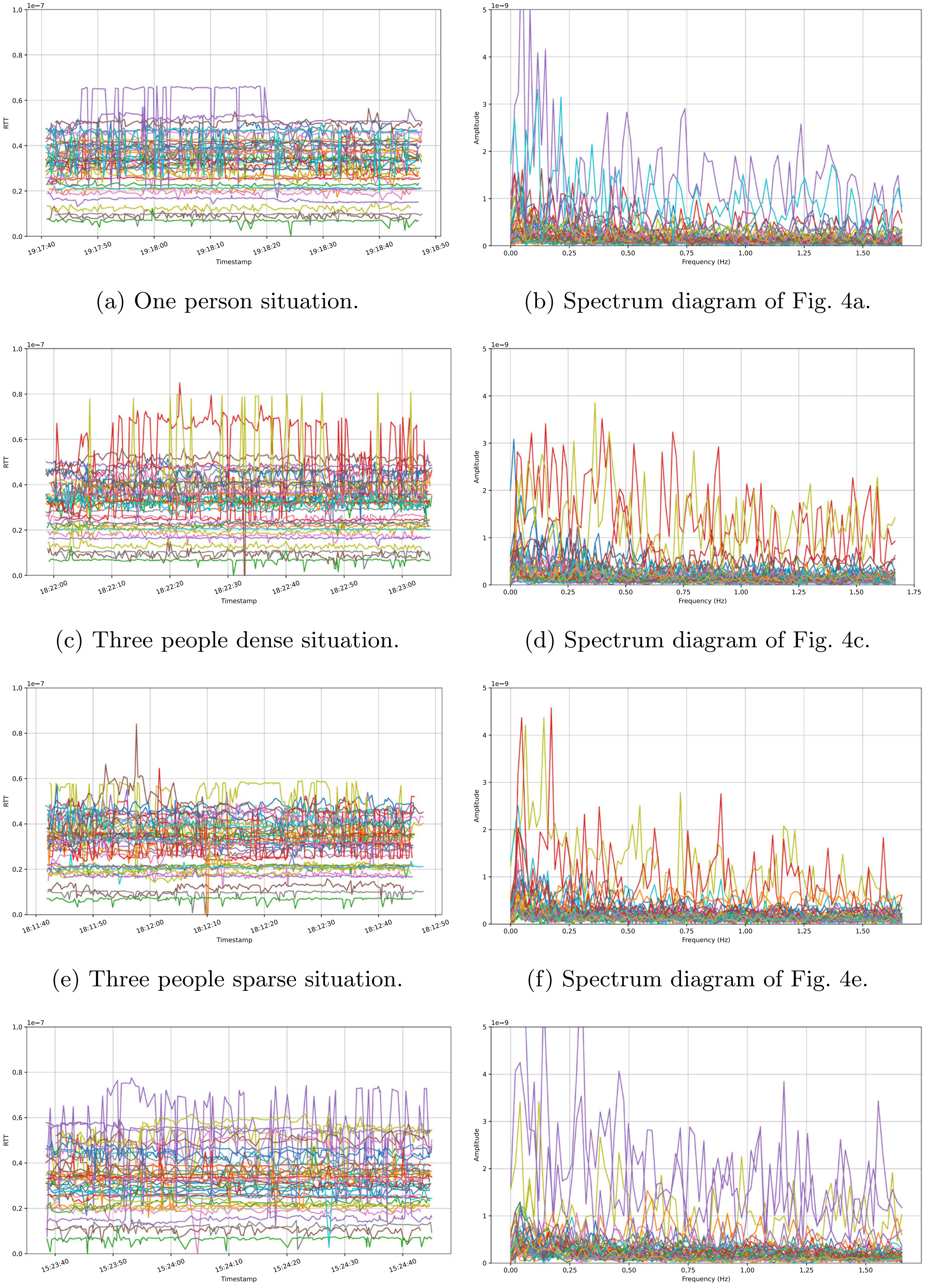
To facilitate the analysis of human counting within an environment using wifi data, we implement a spatial discretization strategy. This approach involves segmenting the space into a virtual grid, where each cell within the grid represents an area of 1 m × 1 m dimensions. Each individual is presumed to occupy a single cell at any given time. This assumption allows for a structured analysis of the interactions between the users and the wireless signal, as each cell can be independently evaluated for its signal characteristics and associated human presence. The granularity of this grid, with its 1-meter square cells, strikes a balance between spatial resolution and computational manageability. It ensures that the spatial distribution of wifi data is adequately captured, providing a detailed yet manageable framework for analyzing how human presence and movements affect wireless signals.

6.4 Feature extractor

Our feature extractor employs a 1D CNN and transformer encoder, adapted from natural language processing to time series analysis, to predict the total number of individuals. This model, demonstrated by Vaswani et al. [48] and further validated in recent studies [49], utilizes a pre-training and fine-tuning approach, which enhances performance by extracting nuanced features relevant to our task.

Figure 4 presents exemplar WiFi RTT traces collected while subjects were seated. Even in this ostensibly static setting, the RTT series displays noticeable fluctuations, underscoring the need to model RTT as a genuine time-varying process rather than as isolated snapshots. To diagnose the temporal structure of these traces, we computed

https://doi.org/10.1007/s00521-025-11540-8



(a) One person situation.

(b) Spectrum diagram of Fig. 4a.

(c) Three people dense situation.

(d) Spectrum diagram of Fig. 4c.

(e) Three people sparse situation.

(f) Spectrum diagram of Fig. 4e.

(g) Eight people situation.

(h) Spectrum diagram of Fig. 4g.

Fig. 4 RTT data and spectrum diagram visualization(The same color in different figures represents the same signal value)

the power spectral density via the fast Fourier transform (FFT). The coexistence of pronounced low-frequency (long-term) and high-frequency (short-term) components confirms that the RTT signal possesses multi-scale temporal dependencies. This spectral evidence cautions against relying solely on conventional convolutional neural networks, which are optimised for recognizing local patterns in fixed-size receptive fields and therefore risk neglecting extended dependencies [50]. Effective modelling of RTT dynamics instead demands architectures capable of integrating information across a wide temporal horizon, such as dilated temporal convolutions or transformer-style attention mechanisms, so that both enduring trends and fleeting variations are captured in a unified representation.

To address this limitation, we incorporate a transformer with positional encoding. One of the key advantages of the transformer is its ability to capture long-range dependencies in the data. This capability is crucial for identifying patterns that develop over extended periods—patterns that traditional models like RNNs and LSTMs often struggle to capture due to vanishing gradient issues. The transformer's self-attention mechanism resolves this limitation by directly modeling the relationships between all elements of the sequence, making it particularly effective for our application. This approach is particularly advantageous when dealing with noisy time series data. The transformer's ability to recognize contextual relationships allows it to assign appropriate weights to different parts of the sequence, even in the presence of noise, improving the detection of human presence.

The feature extraction process begins with a series of 1D CNN layers, which are used to detect local features in the RTT data. CNNs apply a set of filters that scan through the time series, identifying short-term patterns or fluctuations that are indicative of immediate changes in RTT values. Mathematically, the output $h^{(l)}_{b,c,t}$ of the l-th convolutional layer for batch index b, output channel c, and time step t is given by

$$h^{(l)}_{b,c,t} = \sigma\left(\sum_{f=1}^{Cin}\sum_{\tau=0}^{k-1} W^{(l)}_{c,f,\tau}\, h^{(l-1)}_{b,f,t+\tau} + b^{(l)}_{c}\right) \tag{3}$$

where $\sigma(\cdot)$ denotes the ReLU activation function, k is the kernel size, C_{in} is the number of input channels, and $W^{(l)}$ and $b^{(l)}$ are the learnable weights and biases, respectively.

However, these local features alone cannot account for the broader temporal relationships that unfold over longer periods. To address this, the outputs from the CNN layers are passed to the transformer encoder, which serves to capture long-range dependencies. Given a sequence of feature vectors $H \in \mathbb{R}^{K\times d}$, where K is the sequence length and d is the dimensionality of each vector, the transformer encoder operates through multi-head self-attention:

$$Q = HW_Q, \quad K = HW_K, \quad V = HW_V, \tag{4}$$

$$A = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right), \tag{5}$$

$$Z = AV, \tag{6}$$

$$\tilde{H} = \text{LayerNorm}(H + ZW_O), \tag{7}$$

where W_Q, W_K, W_V, and W_O are learnable projection matrices, and $\tilde{H}$ represents the output of the attention block. This output is further passed through a position-wise feedforward network with residual connections and layer normalization.

Positional encoding is added to H prior to attention computation to encode the temporal order of the input:

$$\mathrm{PE}_{(t,2i)} = \sin\left(\frac{t}{seq\_length^{2i/d}}\right), \quad \mathrm{PE}_{(t,2i+1)} = \cos\left(\frac{t}{seq\_length^{2i/d}}\right), \tag{8}$$

where t is the time step and i indexes the feature dimensions. This enables the model to reason about sequence order—a property not inherently captured by attention alone.

Despite the fact that the outputs of the convolutional layers are continuous-valued vectors (unlike the discrete token embeddings in classical NLP), prior works [51–53] demonstrate that Transformers effectively process such continuous inputs in time-series domains. This hybrid design thus enables the model to jointly learn localized and global temporal features.

The integration of the transformer encoder is essential for modeling the long-term dependencies present in the RTT data. The CNN layers extract localized patterns, but the transformer, with its self-attention mechanism, enables the model to capture relationships between distant time steps, allowing it to consider the global context. The self-attention mechanism directly models the interactions between all parts of the sequence, effectively bypassing the vanishing gradient problem that hampers traditional sequence models such as RNNs and LSTMs. This is particularly advantageous for handling noisy or irregular time series data, as it allows the model to discern important patterns over extended periods.

Positional encoding is introduced at this stage to ensure that the transformer can interpret the temporal order of the RTT sequences. Unlike CNNs, which naturally preserve spatial information, Transformers do not inherently recognize sequence order, making positional encoding critical. This encoding provides the necessary temporal structure, allowing the model to understand the chronological progression of RTT data and capture both local and long-term dependencies in a unified manner.

The architecture of our feature extraction module, as shown in Fig. 5, processes input sequences from N receiver devices, specifically smartphones. The model is designed to handle a multivariate time series input where each of the N entries corresponds to a distinct receiver device. For each device, the input comprises a sequence of M time-series features of length K, where M represents the number of access points. This design allows our system to analyze interactions between each receiver and multiple transmitters, enhancing the model's ability to discern patterns indicative of user presence.

Our approach is validated by the t-SNE visualization in Fig. 7b, where the transformer encoder's output shows clear and distinguishable user counting levels, in contrast to the ambiguous clusters observed in the raw data visualization in Fig. 6a. This clear grouping demonstrates the efficacy of our feature extraction strategy in identifying and counting users accurately in complex environments.

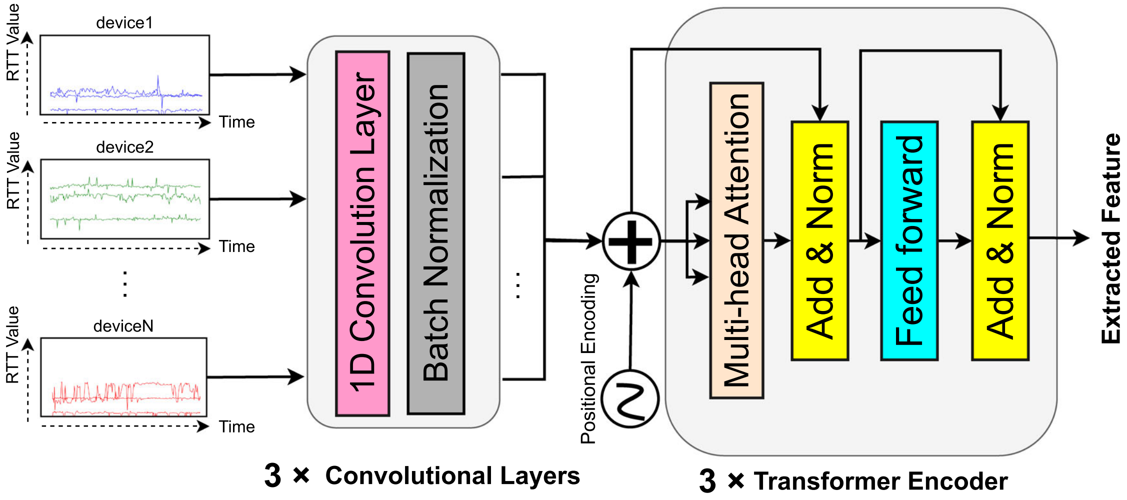


Fig. 5 The overall architecture of the feature extractor

6.5 Counting model

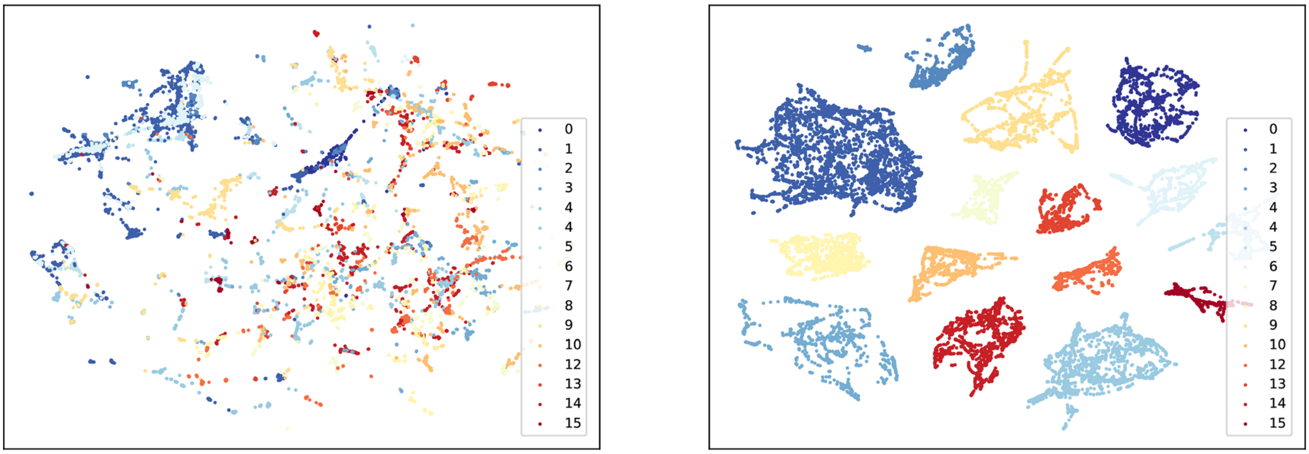
In this section, we discuss how to train a counting model leveraging the embeddings derived from the feature extraction module.

The model adopts a multi-label classification approach [55], where each cell within a physical environment is independently assessed to determine the probability of user presence, effectively treating the detection problem as a series of binary classifications—one for each cell. This multi-label classification approach enables the model to concurrently evaluate multiple cells, making it adept at handling scenarios where multiple individuals are present in various spatial divisions of the environment. The binary classification for each cell predicts whether it is occupied or not, providing a granular level of detail that enhances the overall accuracy of the counting process. The effectiveness of this approach is underpinned by the use of a specialized loss function, described mathematically as follows:

$$\begin{aligned} loss(x,y) = -\frac{1}{C}\sum_{i}\Bigg(y[i] \times log\Big(\frac{1}{1+\exp(-x[i])}\Big) \\ + (1-y[i]) \times log\Big(\frac{\exp(-x[i])}{(1+\exp(-x[i]))}\Big)\Bigg) \end{aligned} \tag{9}$$

where $C \in \mathbb{Z}_{>0}$ represents the total number of cells (classes), $x[i] \in \mathbb{R}$ denotes the model's output logits for each cell, and $y[i] \in 0, 1$ indicates the actual presence (1) or absence (0) of a user in cell i. This MultiLabel Soft-Margin Loss function, designed for differentiability, facilitates the optimization of the model during training by penalizing the prediction error across all cells simultaneously. This structured approach not only improves learning efficiency but also enhances the model's ability to generalize across different environmental configurations and user distributions.

By implementing this multi-label classification technique, we effectively bypass the limitations of traditional counting methods that rely on direct classification of user counts, thereby achieving a more scalable and flexible solution for dynamic and densely populated environments.



(a) From raw data.

(b) From embedded features.

Fig. 6 t-SNE [54] visualization(perplexity = 5, num of iteration = 250)

6.6 Convergence properties of the hybrid feature extractor

Let $f_\theta : \mathbb{R}^{M\times K} \to \mathbb{R}^C$ denote the hybrid CNN–Transformer described in Sect. 6.4, with parameters $\theta \in \mathbb{R}^d$. Given an RTT sequence $\boldsymbol{x}$, it outputs logits $z = f_\theta(\boldsymbol{x})$. With the MultiLabel Soft-Margin loss $\ell(z, y)$ in Equation (9), the empirical risk reads

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\big(f_\theta(\boldsymbol{x}_i), \boldsymbol{y}_i\big), \qquad L^\star = L(\theta^\star).$$

Choose

$$r = \frac{\sigma_0}{2L_J},$$

so every $\theta \in B(\theta^\star, r)$ satisfies $\sigma_{\min}\big(J_{agg}(\theta)\big) \geq \sigma_0/2$ [56–59]. Within this neighbourhood, the following Polyak–Łojasiewicz (PL) inequality holds [60]:

$$\|\nabla_\theta L(\theta)\|_2^2 \;\geq\; \mu\,\big(L(\theta) - L^\star\big), \qquad \mu = \frac{\sigma_0^2}{C}. \tag{10}$$

The result rests on two standard assumptions.

- Jacobian regularity: The aggregated Jacobian $J_{agg}(\theta) = \frac{1}{\sqrt{n}}[J_1(\theta)^\top; \ldots; J_n(\theta)^\top]$ obeys $\sigma_{\min}(J_{agg}(\theta)) \geq \sigma_0/2$ in $B(\theta^\star, r)$ for sufficiently wide networks [56–59].
- Quadratic growth of the loss: For the sigmoid cross-entropy loss, the stacked error $E(\theta) = [e_1^\top, \ldots, e_n^\top]^\top$ with $e_i = \sigma(z_i(\theta)) - y_i$ satisfies the two-sided bound

$$\frac{\beta_{\min}}{C}\,\|E(\theta)\|_2^2 \;\leq\; n\big(L(\theta) - L^\star\big) \;\leq\; \frac{1}{4C}\,\|E(\theta)\|_2^2, \qquad \beta_{\min} = \min_{|t|\leq\delta} \sigma(t)\big(1 - \sigma(t)\big) > 0, \tag{11}$$

whenever $\|z_i(\theta) - z_i(\theta^\star)\|_\infty \leq \delta$. The right-hand inequality exploits $\sigma'(t) = \sigma(t)\,[1 - \sigma(t)] \leq 1/4$; the left arises from the local strong convexity obtained by the second-order Taylor expansion of the sigmoid-BCE loss. Rearranging (11) gives $\|E(\theta)\|_2^2 \geq 4nC\big(L(\theta) - L^\star\big)$, the form required to derive the Polyak–Łojasiewicz constant.

Since $\nabla_\theta L(\theta) = \frac{1}{\sqrt{n}\,C}\, J_{agg}(\theta)^\top E(\theta)$, combining the two bounds yields (10).

The model is trained with AdamW and a `ReduceLROnPlateau` scheduler,[1] which multiplies the learning rate by a factor $\gamma \in (0, 1)$ after p stagnant epochs. Because the scheduler only decreases η_t, the condition $0 < \eta_t < 1/L$ persists, leading to the piecewise-linear decay

$$L(\theta_{t+1}) - L^\star \leq \big(1 - \eta_t\mu\big)\big(L(\theta_t) - L^\star\big).$$

Hence, the loss declines at least linearly while η_t is fixed, and each reduction produces a tighter geometric envelope.

The theoretical analysis presumes that both the CNN blocks and the Transformer encoder possess sufficient width to keep the aggregated Jacobian well conditioned. To confirm that the proposed hyperparameters indeed

[1] https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLROnPlateau.html.

place the model in this over-parameterised regime, we trained the feature extractor 32 times with independent random seeds while keeping all other settings fixed. Figure 7a plots the resulting training losses, and Fig. 7 shows the relative weight movement. Across all seeds, the loss curves collapse and converge monotonically, while the normalised weight movement plateaus below 10^{-1}. These observations indicate (i) initialization does not materially affect optimisation behaviour, and (ii) the model consistently converges during training, as experimentally observed under the proposed hyperparameter settings.

7 Experiments and results

7.1 Data collection

In this study, data collection took place within a realistic cluttered environment, as illustrated in Figs. 8a and Fig. 9a, c. The configurations of these testbeds are summarized in Table 1. For the pilot testbed, they are limited to multi-label classification for location estimation in scenarios where a single individual is present at a designated location. During the others, groups of up to 15 individuals each sat for one minute at predetermined locations in testbed1 (Fig. 10), and groups of up to six individuals each stood for one minute at predetermined locations in testbed2, following the setup experienced in several research [61–63].

Data collection was streamlined using an Android application installed on multiple smartphones, which was designed to continuously scan for nearby access points. To ensure uniformity and synchronization in data collection, the application was configured to operate concurrently across all devices. One device was designated as the master to initiate the scanning process, ensuring that all participating smartphones collected data simultaneously and uniformly. This setup was crucial for maintaining the integrity and consistency of the data collected during the experiment. To achieve high-precision distance measurements, the IEEE 802.11az[2] WiFi standard was utilized for RTT measurements, improving location accuracy and ensuring reliable data collection. The application was deployed on eight Pixel 3 devices (receivers), while five Google Nest WiFi access points (transmitters) were utilized for signal transmission. Access points and smartphones were positioned to maximize line-of-sight coverage across each room. One side of the smartphone faces a spacious area like the ceiling, and the access points are positioned in open and elevated places, such as the tops of shelves, so that no objects block them from above.

7.2 System parameters

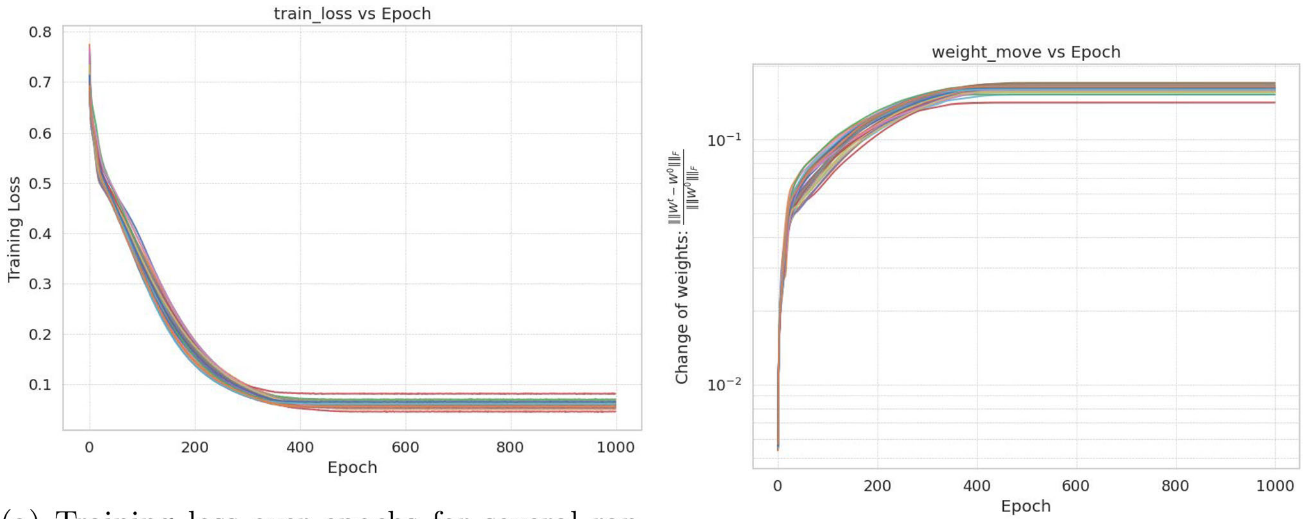
The dataset is partitioned into a training dataset and a test dataset at an 80% to 20% ratio, respectively.

Additionally, we performed data augmentation on the normalized training data by adding noise with a mean of 0 and a standard deviation of 0.001 to the training data. Using different seed values, we conducted data augmentation twice for each training data point—that is, including the synthetic data, we tripled the size of the training dataset. Moreover, we applied a dropout rate of 0.2 during training to suppress overfitting.

7.3 Results

The system parameters and the model configuration are described in Table 2. These parameters are determined by grid search.

[2] https://standards.ieee.org/beyond-standards/newly-released-ieee-802-11az-standard-improving-wi-fi-location-accuracy-is-set-to-unleash-a-new-wave-of-innovation/.



(a) Training loss over epochs for several random initialisations.

(b) From embedded features.

Fig. 7 Learning analysis

7.3.1 Pilot study: preliminary evaluation of model performance

To assess the performance of our multi-label classification model, we adopt *example-based* evaluation metrics, which evaluate predictions at the instance level by comparing predicted and ground-truth label sets per sample. Specifically, we report **Example-based Precision**, **Recall**, and **F1-score**, which are defined as follows for an input instance $i \in \{1, \ldots, N\}$:

$$\text{Precision}_i = \frac{|\hat{\mathbf{y}}_i \cap \mathbf{y}_i|}{|\hat{\mathbf{y}}_i|} \quad \text{Recall}_i = \frac{|\hat{\mathbf{y}}_i \cap \mathbf{y}_i|}{|\mathbf{y}_i|}, \quad \text{F1}_i = \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{12}$$

Here, $\hat{\mathbf{y}}_i$ and $\mathbf{y}_i$ denote the predicted and true binary label vectors, respectively. The overall performance is then computed by averaging over all N instances:

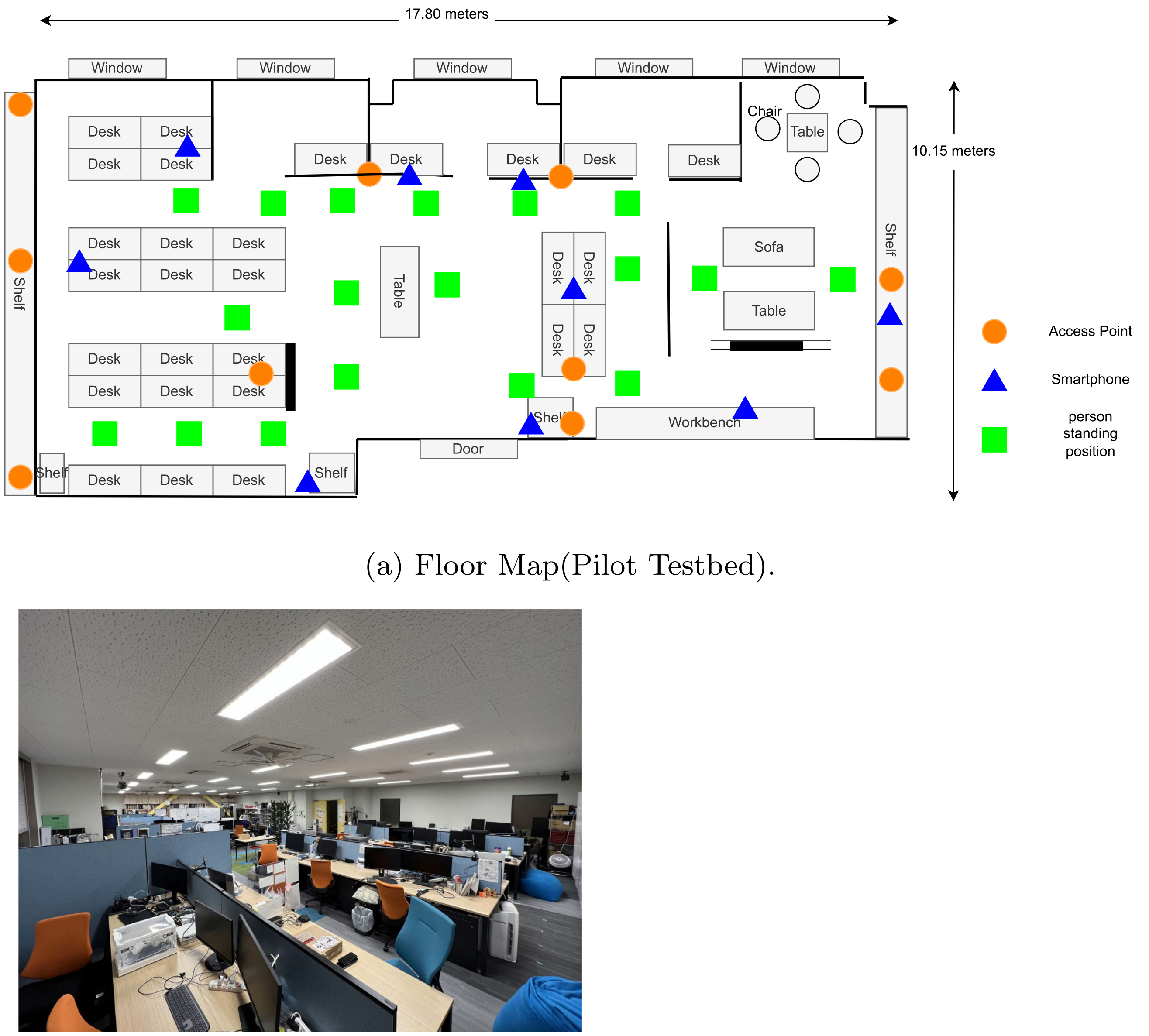
$$\text{Example-based F1} = \frac{1}{N} \sum_{i=1}^{N} \text{F1}_i \tag{13}$$

In addition, we report the subset accuracy, also known as the Exact Match Ratio, which strictly measures the proportion of instances where the predicted label set exactly matches the ground-truth set:

$$\text{Subset Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[\hat{\mathbf{y}}_i = \mathbf{y}_i] \tag{14}$$

These metrics provide a comprehensive evaluation: the example-based metrics allow for partial correctness, while subset accuracy serves as a stringent criterion of correctness in multi-label settings.

As shown in Table 3, multi-label classification is able to be used for this task, detecting an individual is at the candidate position.



(a) Floor Map(Pilot Testbed).

(b) Photograph of the interior space.

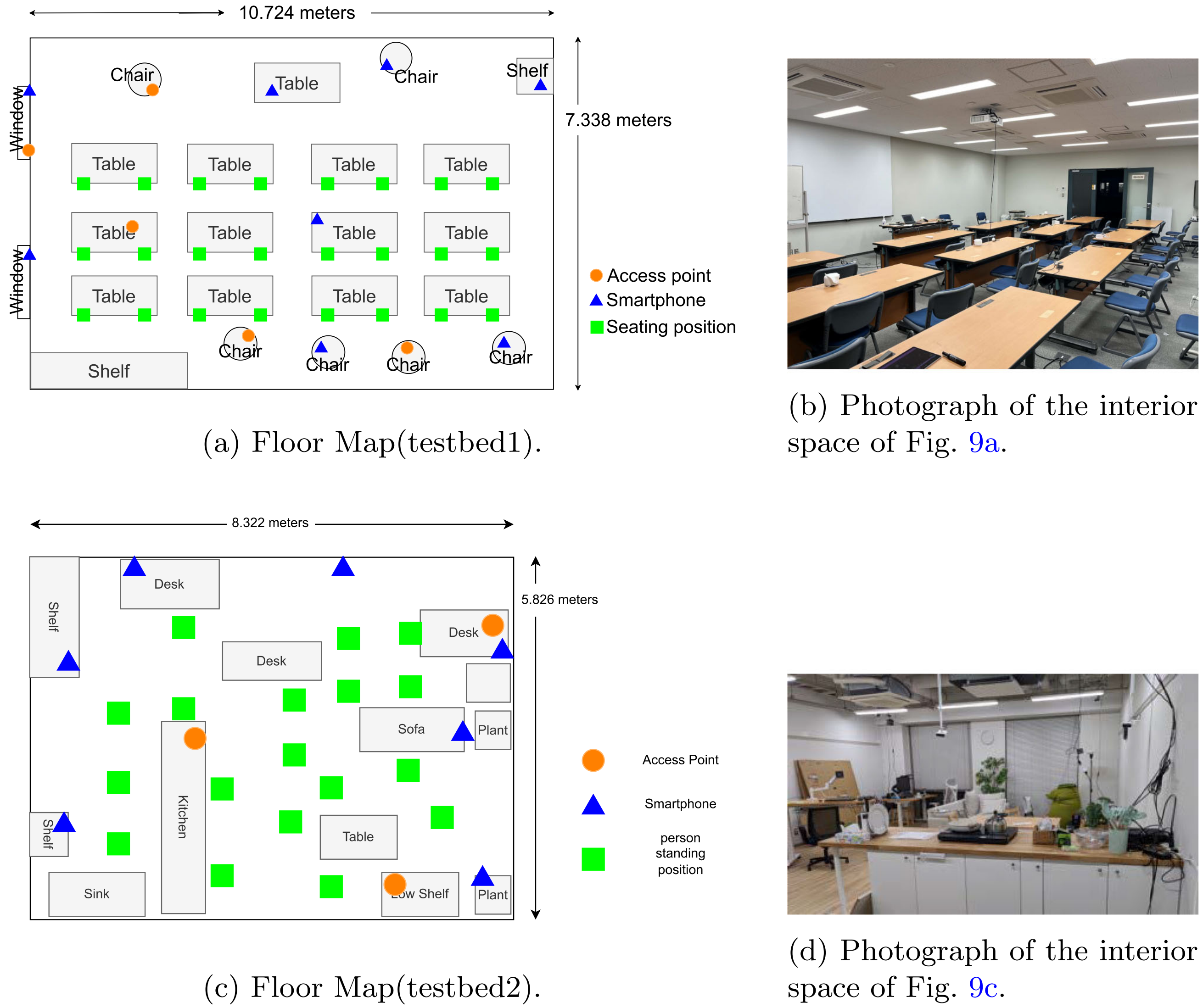
Fig. 8 Preliminary validation environment

7.3.2 Counting performance evaluation

In this section, we assess the performance of our proposed counting system. We specifically evaluate our transformer-based method and compare it to other existing methods for counting using WiFi signals, such as Wi-cal [45], RPCNet [65], LocFree [34] and ImgFi [66].[3]

It is important to note that *Time4Count* is the first RTT-based counting technique. However, to rigorously evaluate the system's model design, we applied state-of-the-art counting models to the RTT data to ensure a fair comparison. Figure 11 shows the proposed models surpass all the compared schemes. This improvement is attributed to the transformer's superior ability to handle sequential data and its enhanced capacity to capture long-

[3] It is noteworthy that RPCNet and ImgFi are implemented as multi-class classification, and we modified the last layer of their model to fit multi-label classification(The task in Wi-cal is regression).



(a) Floor Map(testbed1).

(b) Photograph of the interior space of Fig. 9a.

(c) Floor Map(testbed2).

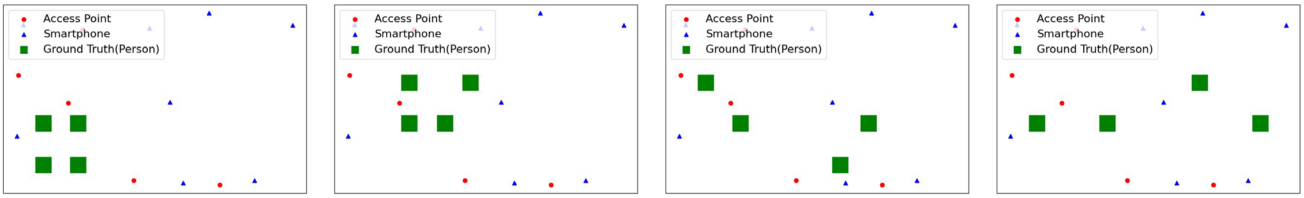
(d) Photograph of the interior space of Fig. 9c.

Fig. 9 Experiment environment for counting

Table 1 Summary of the testbeds

Item	Pilot Testbed	Testbed 1	Testbed 2
Participant Arrangement	1 person standing	15 people sitting	6 people standing
Number of Candidate Points	18	24	18
Number of Position Configurations	18	138	50
Total Number of Records	13,512	27,123	11,083
Average Records per Pattern	751	197	221
Number of Receivers	9	8	7
Number of Access Points	10	5	3

range dependencies within the signal data. Wi-cal [45] uses limited signal features (maximum, minimum, average, etc.) and does not explicitly consider the correlation between different signals. From the results, we can conclude that the summarized features are insufficient for analyzing RTT time-series data. LocFree [34] originally uses the multiple layers perceptron with utilizing dropout for a single person localization. The result shows that the simple architecture is not able to handle the time-series data in our dataset. RPCNet [65] employs 1D CNN



(a) Dense Situation1. (b) Dense Situation2. (c) Sparse Situation1. (d) Sparse Situation2.

Fig. 10 Dense and sparse situation settings

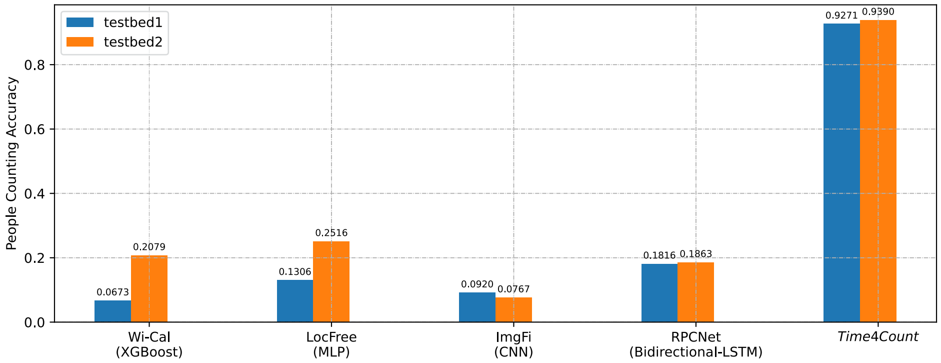
Table 2 System hyperparameters

Parameter	Explore range (**bold** means default)
Batch size	{256, 512, **1024**, 2048}
Learning rate	{0.001, 0.005, **0.0001**, 0.00005}
Optimizer	AdamW [64]
Epochs	500
Input sequence length	8
1D CNN output channels	{32, **64**, 128, 256}
1D CNN hidden channels	{32, 64, **128**, 256}
1D CNN kernel size	{3, 5, **7**, 9}
The number of Attention heads in Transformer	{1, **2**, 4, 8}
Dimension of the feedforward network in Transformer	{32, 64, **128**, 256}
Data Augmentation Rate	Double(2x)
Dropout rate	{0.1, **0.2**, 0.3, 0.4}

Table 3 Example-based and Subset Accuracy Evaluation Metrics

Metric	Score
Subset Accuracy	0.9071
Example-based Precision	0.9073
Example-based Recall	0.9075
Example-based F1-score	0.9074

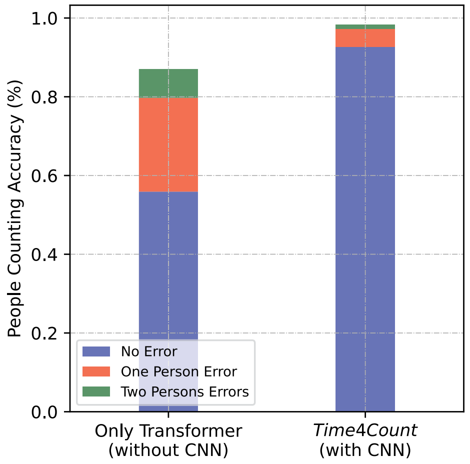
Fig. 11 Counting performance of the proposed *Time4Count* system and the baseline approaches

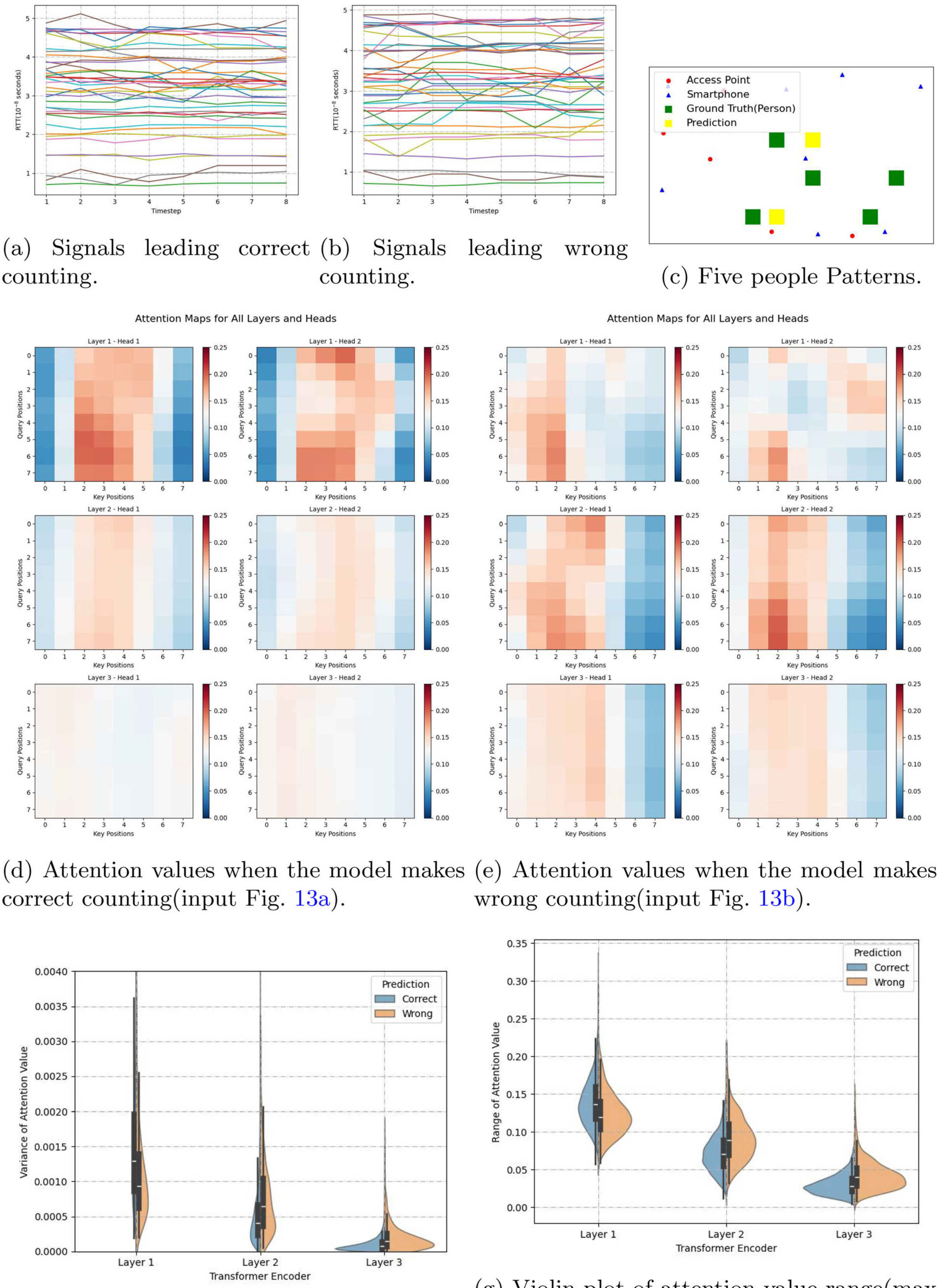


and bidirectional-LSTMs to grasp spatial and temporal features. Although the architecture is well-known for processing time series data, it does not succeed in accurately predicting the number of people in our dataset. ImgFi [66] utilizes a method that converts time-series signal data into images through transformations such as the Gramian angular field, Markov transition field, or recurrence plot. This image representation is then processed using a CNN model to identify human activities. However, this approach predominantly emphasizes derivative values over the actual data values, which could result in lower accuracy compared to the method we propose. Additionally, the method's reliance on image-based transformations means that it requires a sufficiently long sequence of input data, making it less effective for shorter time-series sequences. The superior performance of an architecture combining 1D CNN with a Transformer over using a transformer alone for time series data analysis can be attributed to the ability of the convolutional layer to effectively extract significant local features and emphasize critical information shown in Fig. 12, thereby reducing noise and highlighting essential characteristics. In this scenario, a LocFree [34] that has 75,636 trainable parameters required 54 MB of memory for inference, whereas our proposed method that has 280,820 trainable parameters required 394 MB. Although this increase in memory usage still allows operation on typical edge devices, it could be reduced by decreasing the number of CNN layers or simplifying the Transformer hyperparameters. By first feeding the per AP RTT streams into a shallow 1D CNN, *Time4Count* distils short-range convolution motifs that capture the sub-second co-fluctuations shared by neighbouring access points and the handset. Prior WiFi counting schemes based on RSSI or CSI compress each window to simple statistics or rely on handcrafted features, which blur the cross-channel relationships. The subsequent Transformer encoder then fuses those local embeddings through self-attention, learning long-range inter AP dependencies that RNN-based models or earlier RTT RSSI CSI pipelines cannot express. This enables accurate counts from only a brief slice of RTT data and establishes *Time4Count* as the first people counting system to combine IEEE 802.11mc az RTT measurements with a modern attention backbone.

Figure 13 illustrates a scenario in which the trained model produces both correct and incorrect predictions on test data. Figure 13a, b show datasets with identical spatial distribution patterns, as depicted in Fig. 13c. From the raw data, incorrect predictions are notably associated with signal waveforms exhibiting significant amplitude disturbances. The attention maps highlight the model's focus during processing. For correct predictions, the attention map at Layer 1, shown in Fig. 13d, effectively identifies essential regions. As the layers deepen, attention values diminish, indicating that the model condenses input information and concentrates on the most salient features. This reduction in attention values also suggests that the model filters out unnecessary information, focusing its resources efficiently on relevant features. In contrast, incorrect predictions are characterized by persistently high attention values even in deeper layers, such as Layer 2 and Layer 3, as shown in Fig. 13e, which was led by the perturbation in the raw data in Fig. 13b. This behavior suggests that the model struggles to isolate critical features, leading to information diffusion and confusion. Additionally, the sustained attention to non-essential information indicates that the model inefficiently allocates resources, contributing to prediction errors. The input to the transformer consists of features extracted by a preceding CNN. The results imply that the model performs effectively when the CNN reduces noise and captures relevant local patterns. When feature extraction is accurate, the transformer can efficiently focus attention in the early layers, minimizing the need for extensive reprocessing in deeper layers. Finally, the difference in attention dynamics between correct and incorrect predictions is further validated through statistical analysis of the variance and the range (i.e., the

Fig. 12 Effect of CNN layers(testbed1)





(a) Signals leading correct counting.

(b) Signals leading wrong counting.

(c) Five people Patterns.

(d) Attention values when the model makes correct counting(input Fig. 13a).

(e) Attention values when the model makes wrong counting(input Fig. 13b).

(f) Violin plot of attention value variances.

(g) Violin plot of attention value range(max - min).

◀**Fig. 13** Case study about five people counting

difference between maximum and minimum values) of the attention map values, as demonstrated in Fig. 13f, g. The greater variance observed in deeper layers during incorrect predictions suggests that the model experiences confusion, resulting in unstable focus across layers.

7.3.3 The impact of input sequence length

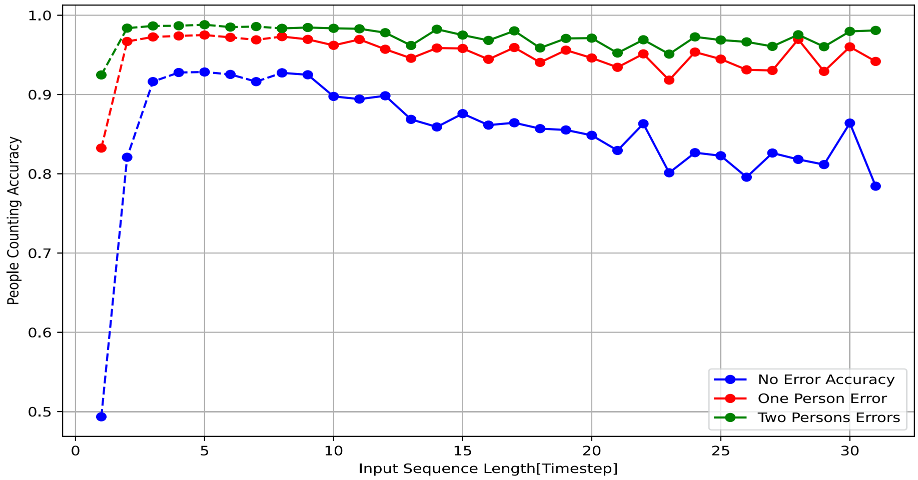
Figure 14 illustrates the effect of input sequence length on the accuracy of user counting. The results indicate that very short sequences lead to a drop in system accuracy, as they do not provide sufficient data for the model to make reliable predictions. On the other hand, too long sequences introduce variability in the data due to changes in the number of individuals or their positions over time, which can confuse the model and degrade performance in our dataset. An optimal balance is achieved with a sequence length of 8(about 2.2 s), which provides enough contextual information without introducing significant variability, resulting in an accuracy of 92.7%. The system parameters and model configuration values were fixed as shown in Table 2.

7.3.4 Robustness of the number of people

Figure 15 showcases the system’s performance as tested with an increasing maximum number of people observed during the training dataset. Notably, the model demonstrates slightly superior performance in scenarios involving three and five to eight people, surpassing the results of the one, two, and four-people classifications. This enhancement is attributed to the key factor that our approach utilizes multi-label classification, significantly improving the accuracy and robustness of people counting. This method allows the model to simultaneously predict multiple labels, making it more effective in complex scenarios where multiple people may be present. The inclusion of diverse environmental data during training equips the model to better understand and adapt to different situational variables. By analyzing detailed features from each candidate position, the model can more accurately assess and predict the presence of individuals, thus enhancing both its accuracy and performance in real-world applications.

Additionally, Fig. 15 provides detailed insights into the accuracy of our proposed system for predicting user counts. In the context of people counting, allowing a margin of error of plus or minus two people, the results demonstrate that our model achieves over 98% accuracy in estimating the number of individuals present. This high level of precision underscores the effectiveness of our approach in accurately counting users in various settings.

Fig. 14 The effect of input sequence length on *Time4-Count* performance(Lengths smaller than the CNN kernel size is indicated by dot lines)



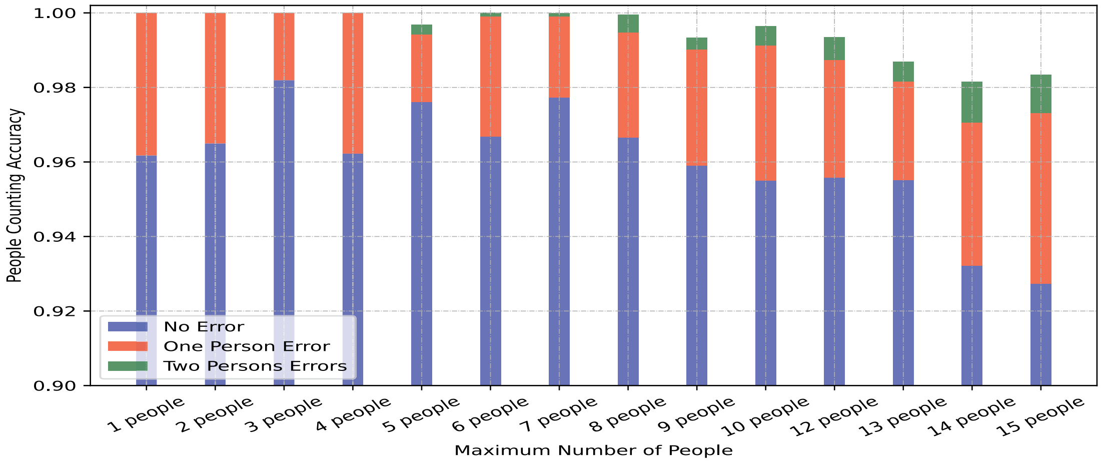


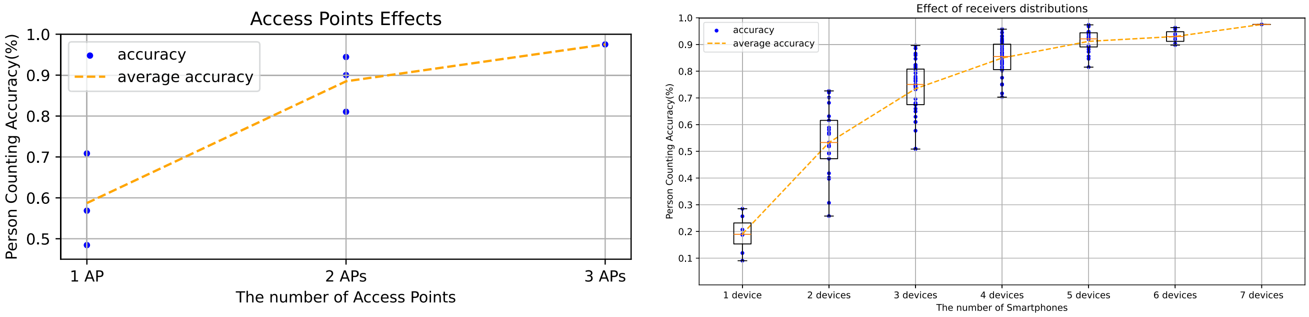
Fig. 15 Robustness against the number of people.

7.3.5 Robustness of the number of devices

We explore the requisite number of devices, including access points and smartphones, necessary for achieving accurate predictions. Fig. 16a, b present the results showcasing the robustness of our approach in terms of the number of devices. Each blue point shows the counting accuracy.

The result of Fig. 16a highlights the impact of varying the number of access points. When we track the average score, we can do an analysis that employing three access points enables achieving over 90% accuracy in counting. This observation aligns with the intuitive understanding derived from localization systems based on time of arrival (ToA), which mathematically necessitates a minimum of three base stations for accurate positioning. However, intriguingly, simulation results demonstrate that employing only two access points can also yield 90% accuracy in two specific configurations, underscoring the potential effectiveness of a strategically positioned pair of access points in capturing the room's dynamics comprehensively. Conversely, employing a single access point results in a significant decline in prediction accuracy, emphasizing the vulnerability of such setups to signal interference in multi-person scenarios, even when not in direct line-of-sight.

The result of Fig. 16b elucidates the influence of the number of receivers on prediction accuracy. As shown in Fig. 9a pointed out in blue triangle icons, the smartphones are installed along the walls of the room, effectively determining the room's visibility from the perspective of the access points. Unsurprisingly, increasing the number of smartphones enhances model performance, as it directly correlates with an increase in the number of features available for analysis. On average, employing more than five smartphones enables achieving over 90% accuracy



(a) The effect of access point density.

(b) Effect of the number of receivers.

Fig. 16 Investigations of the number of devices(testbed2)

in person counting. However, certain distribution patterns may fail to consistently achieve this threshold, necessitating the deployment of at least seven devices to ensure reliable performance within our experimental room and environment.

7.4 Data augmentation study

Finally, we evaluate the impact of different levels of data augmentation on the model's performance. Data augmentation is used to increase the diversity of the training data and enhance the model's generalization capability. We systematically investigate how augmenting the dataset by various factors (e.g., doubling and tripling) influences the model's predictive accuracy.

Figure 17 illustrates how varying the amount of augmented data affects the model's performance. The improvements observed with data augmentation can be attributed to the increased variety in the training set, which allows the model to generalize better to unseen data. However, we observed that the model's performance degrades when the augmentation factor exceeds 4x. This decline is believed to result from a loss of data diversity due to excessive augmentation with noise, making the augmented data less representative of the true data distribution.

8 Limitations

The present study demonstrates that *Time4Count* reaches state-of-the-art accuracy in a single 80 m^2 office, yet three open challenges remain. First, all volunteers moved slowly or paused between measurements, and the prediction performance in a dynamic situation has not been evaluated in this work. Second, scalability beyond a room-sized cell is untested. Large area trials of device-free localization have shown that coverage gaps grow rapidly once the monitored floor exceeds 100 m^2 unless access points are densified or partitioned into cooperative clusters, inflating both deployment cost and inference latency. Recent RTT field studies likewise observe meter-scale errors in wide corridors unless at least three FTM-capable anchors fall inside every ten-metre radius, underscoring the need for principled coverage planning. Also, prior CSI-based evaluations report a 10 to 20when subjects walk at everyday speeds through walls [67]. Third, the influence of heterogeneous client hardware is still unverified. Although the RTT protocol subtracts station-side processing delays, variations in chipset design,

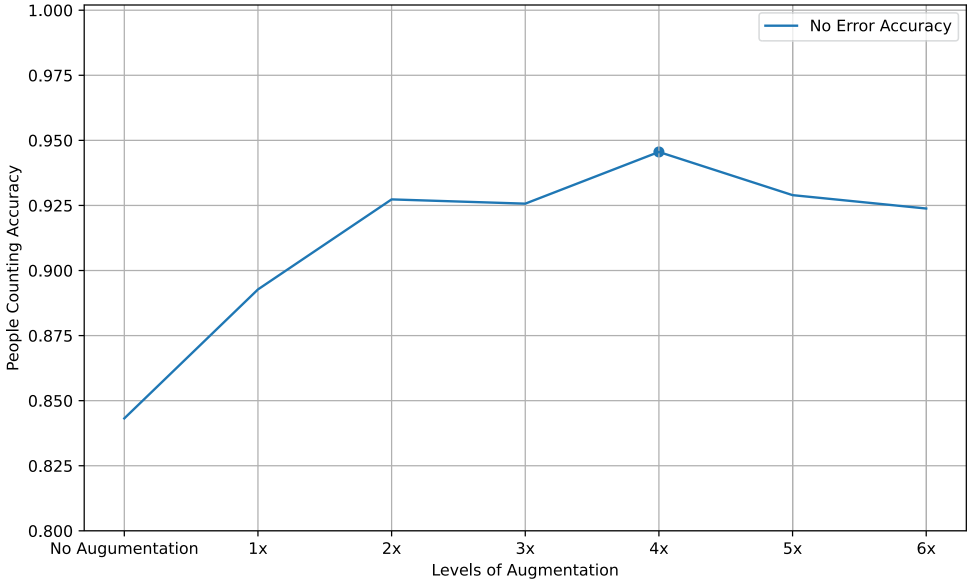


Fig. 17 Evaluation of performance by applying data augmentation

antenna gain, and firmware may alter the measured signals. Systematic cross-device experiments are therefore required to confirm robustness in mixed hardware environments.

9 Conclusion

In this paper, we introduced *Time4Count*, a novel people counting system that utilizes RTT data captured by commodity smartphones and access points. Specifically, *Time4Count* combines CNN and Transformer models to harness the unique advantages of each, effectively mitigating the impact of noisy measurements caused by cluttered environments or varying user distributions. We highlighted the critical importance of integrating time-series data into our methodology and detailed our approach for extracting robust representations using a transformer encoder. Notably, *Time4Count* achieved an accuracy of 92.7% in counting individuals. Additionally, we conducted comprehensive evaluations to assess the system's performance across various conditions, including changes in input sequence length, the number of individuals in the training dataset, and the density of access points and smartphones. Our results demonstrate the robustness of the prediction model, confirming its effectiveness in maintaining high accuracy despite fluctuations in the number of individuals present. This underscores the potential of *Time4Count* for practical deployment in diverse, real-world environments.

Acknowledgements This work was partially funded by JST, CREST Grant JPMJCR21M5, JST AIP Challenge 2024, JSPS KAKENHI Grant 22K12011, JST BOOST, Japan Grant Number JPMJBS2402, and NVIDIA award.

Author contributions Conceptualization and methodology were contributed by Haruki Yonekura and Hamada Rizk; Formal analysis, investigation, and writing–original draft preparation were involved by Haruki Yonekura; Writing–review and editing, funding acquision, and resources were performed by Hamada Rizk and Hirozumi Yamaguchi; Supervision was done by Hirozumi Yamaguchi

Funding Open Access funding provided by The University of Osaka. JST, CREST Grant JPMJCR21M5, JST AIP Challenge 2024, JSPS KAKENHI Grant 22K12011, JST BOOST, Japan Grant Number JPMJBS2402, and NVIDIA award.

Data availability Data, Materials, and Code availability: Data, Materials, and Codes used in this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical approval and consent to participate The authors declare that they have ethical and informed consent for the data used.

Consent for publication Consent for publication was obtained from the participants.

References

1. Boominathan L, Kruthiventi SSS, Babu RV (2016) Crowdnet: a deep convolutional network for dense crowd counting. In: Proceedings of the 2016 ACM on multimedia conference. MM '16. ACM, New York, NY, USA, pp 640–644

2. Zeng L, Xu X, Cai B, Qiu S, Zhang T (2017) Multi-scale convolutional neural networks for crowd counting. CoRR arXiv:abs/1702.02359
3. Zhang C, Li H, Wang X, Yang X (2015) Cross-scene crowd counting via deep convolutional neural networks. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR), pp 833–841
4. Li Z, Zhang L, Fang Y, Wang J, Xu H, Yin B, Lu H (2016) Deep people counting with faster r-cnn and correlation tracking. In: Proceedings of the international conference on internet multimedia computing and service. ICIMCS'16. ACM, New York, NY, USA, pp 57–60
5. Kaemarungsi K, Krishnamurthy P (2004) Properties of indoor received signal strength for wlan location fingerprinting. In: The first annual international conference on mobile and ubiquitous systems: networking and services, 2004. MOBIQUITOUS 2004. IEEE, pp 14–23
6. Youssef M, Agrawala A (2005) The Horus WLAN location determination system. In: the 3rd international conference on mobile systems, applications, and services. ACM, pp 205–218
7. Abbas M, Elhamshary M, Rizk H, Torki M, Youssef M (2019) Wideep: Wifi-based accurate and robust indoor localization system using deep learning. In: 2019 IEEE international conference on pervasive computing and communications (PerCom). IEEE, pp 1–10
8. Zheng H, Zhang Y, Zhang L, Xia H, Bai S, Shen G, He T, Li X (2021) Grafin: An applicable graph-based fingerprinting approach for robust indoor localization. In: 2021 IEEE 27th international conference on parallel and distributed systems (ICPADS). IEEE, pp 747–754
9. Youssef M, Mah M, Agrawala A (2007) Challenges: device-free passive localization for wireless environments. In: Proceedings of the 13th annual ACM international conference on mobile computing and networking, pp 222–229
10. Moussa M, Youssef M (2009) Smart cevices for smart environments: Device-free passive detection in real environments. In: 2009 IEEE international conference on pervasive computing and communications. IEEE, pp 1–6
11. Januszkiewicz Ł (2018) Analysis of human body shadowing effect on wireless sensor networks operating in the 2.4 ghz band. Sensors 18(10):3412
12. Feng X, Nguyen KA, Luo Z (2022) An analysis of the properties and the performance of wifi rtt for indoor positioning in non-line-of-sight environments. In: 17th international conference on location based services
13. Rizk H, Sakr A, Ghazal A, Hemdan M, Shaheen O, Sharara H, Yamaguchi H, Youssef M (2023) Indoor localization system for seamless tracking across buildings and network configurations. In: GLOBECOM 2023 - 2023 IEEE global communications conference, pp 776–782. https://doi.org/10.1109/GLOBECOM54140.2023.10437762
14. Rizk H, Uchiyama A, Yamaguch H (2024) Adaptability matters: Heterogeneous graphs for agile indoor positioning in cluttered environments. In: 2024 25th IEEE international conference on mobile data management (MDM), pp 103–108. https://doi.org/10.1109/MDM61037.2024.00033
15. Rizk H, Amano T, Yamaguchi H, Youssef M (2022) Cross-subject activity detection for covid-19 infection avoidance based on automatically annotated imu data. IEEE Sens J 22(13):13125–13135. https://doi.org/10.1109/JSEN.2022.3176291
16. Rizk H, Abbas M, Youssef M (2020) Omnicells: Cross-device cellular-based indoor location tracking using deep neural networks. In: 18th Annual IEEE conference on pervasive computing and communications (PerCom). IEEE
17. Wang J, Jiang H, Xiong J, Jamieson K, Chen X, Fang D, Xie B (2016) Lifs: Low human-effort, device-free localization with fine-grained subcarrier information. In: Proceedings of the 22nd annual international conference on mobile computing and networking, pp 243–256
18. Jiang W, Xue H, Miao C, Wang S, Lin S, Tian C, Murali S, Hu H, Sun Z, Su L (2020) Towards 3d human pose construction using wifi. In: Proceedings of the 26th annual international conference on mobile computing and networking, pp 1–14
19. Venkatnarayan RH, Shahzad M, Yun S, Vlachou C, Kim K-H (2020) Leveraging polarization of wifi signals to simultaneously track multiple people. Proc ACM Interact Mob Wear Ubiquitous Technol 4(2):1–24
20. Karanam CR, Korany B, Mostofi Y (2019) Tracking from one side: Multi-person passive tracking with wifi magnitude measurements. In: Proceedings of the 18th international conference on information processing in sensor networks, pp 181–192
21. Golden SA, Bateman SS (2007) Sensor measurements for wi-fi location with emphasis on time-of-arrival ranging. IEEE Trans Mob Comput 6(10):1185–1198
22. Chan Y-T, Tsui W-Y, So H-C, Ching P (2006) Time-of-arrival based localization under nlos conditions. IEEE Trans Veh Technol 55(1):17–24
23. Hashem O, Youssef M, Harras KA (2020) Winar: Rtt-based sub-meter indoor localization using commercial devices. In: 2020 IEEE international conference on pervasive computing and communications (PerCom), pp 1–10 IEEE
24. Ibrahim M, Liu H, Jawahar M, Nguyen V, Gruteser M, Howard R, Yu B, Bai F (2018) Verification: Accuracy evaluation of wifi fine time measurements on an open platform. In: Proceedings of the 24th annual international conference on mobile computing and networking, pp 417–427
25. Hashem O, Youssef M, Harras KA (2020) Winar: Rtt-based sub-meter indoor localization using commercial devices. In: 2020 IEEE international conference on pervasive computing and communications (PerCom), pp 1–10. https://doi.org/10.1109/PerCom45495.2020.9127363

26. Hashem O, Harras KA, Youssef M (2020) Deepnar: Robust time-based sub-meter indoor localization using deep learning. In: 2020 17th annual IEEE international conference on sensing, communication, and networking (SECON). IEEE, pp 1–9
27. Rizk H, Elmogy A, Rihan M, Yamaguchi H (2024) A precise and scalable indoor positioning system using cross-modal knowledge distillation. Sensors. https://doi.org/10.3390/s24227322
28. Rizk H, Elgokhy S, Sarhan A (2015) A hybrid outlier detection algorithm based on partitioning clustering and density measures. In: 2015 tenth international conference on computer engineering & systems (ICCES). IEEE, pp 175–181
29. Banin L, Schatzberg U, Amizur Y (2016) Wifi ftm and map information fusion for accurate positioning. In: 2016 international conference on indoor positioning and indoor navigation (IPIN)
30. Ciurana M, Barcelo-Arroyo F, Izquierdo F (2007) A ranging method with ieee 802.11 data frames for indoor localization. In: 2007 IEEE wireless communications and networking conference. IEEE, pp 2092–2096
31. Rizk H, Elmogy A, Yamaguchi H (2022) A robust and accurate indoor localization using learning-based fusion of wi-fi rtt and rssi. Sensors 22(7):2700
32. Youssef F, Elmogy S, Rizk H (2022) Magttloc: Decimeter indoor localization system using commercial devices. In: Proceedings of the 30th international conference on advances in geographic information systems
33. Zhou B, Wu Z, Chen Z, Liu X, Li Q (2023) Wi-fi rtt/encoder/ins-based robot indoor localization using smartphones. IEEE Trans Veh Technol 72(5):6683–6694
34. Mohsen M, Rizk H, Yamaguchi H, Youssef M (2023) Locfree: Wifi rtt-based device-free indoor localization system. In: Proceedings of the 2nd ACM SIGSPATIAL international workshop on spatial big data and AI for industrial applications, pp 32–40
35. Jurdi R, Chen H, Zhu Y, Ng BL, Dawar N, Zhang C, Han JK-H (2024) Whereartthou: A wifi-rtt-based indoor positioning system. IEEE Access 12:41084–41101
36. Han T, Bai L, Liu L, Ouyang W (2023) Steerer: Resolving scale variations for counting and localization via selective inheritance learning. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 21848–21859
37. Xu C, Li S, Liu G, Zhang Y, Miluzzo E, Chen Y-F, Li J, Firner B (2013) Crowd++ unsupervised speaker count with smartphones. In: Proceedings of the 2013 ACM international joint conference on pervasive and ubiquitous computing, pp 43–52
38. Ukyo R, Amano T, Rizk H, Yamaguchi H (2023) Pedestrian tracking using 3d lidars–case for proximity scenario. In: 2023 IEEE 26th international conference on intelligent transportation systems (ITSC). IEEE, pp 4683–4689
39. Zeng L, Xu X, Cai B, Qiu S, Zhang T (2017) Multi-scale convolutional neural networks for crowd counting. In: 2017 IEEE international conference on image processing (ICIP). IEEE, pp 465–469
40. Pham V-Q, Kozakaya T, Yamaguchi O, Okada R (2015) Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation. In: Proceedings of the IEEE international conference on computer vision, pp 3253–3261
41. Wang F, Feng Q, Chen Z, Zhao Q, Cheng Z, Zou J, Zhang Y, Mai J, Li Y, Reeve H (2017) Predictive control of indoor environment using occupant number detected by video data and co2 concentration. Energy Build 145:155–162
42. Zou H, Zhou Y, Yang J, Spanos CJ (2018) Device-free occupancy detection and crowd counting in smart buildings with wifi-enabled iot. Energy Build 174:309–322
43. Sobron I, Del Ser J, Eizmendi I, Velez M (2018) Device-free people counting in iot environments: New insights, results, and open challenges. IEEE Internet Things J 5(6):4396–4408
44. Mizutani M, Uchiyama A, Murakami T, Abeysekera H, Higashino T (2020) Towards people counting using wi-fi csi of mobile devices. In: 2020 IEEE international conference on pervasive computing and communications workshops (PerCom Workshops). IEEE Computer Society, pp 1–6
45. Choi H, Fujimoto M, Matsui T, Misaki S, Yasumoto K (2022) Wi-cal: Wifi sensing and machine learning based device-free crowd counting and localization. IEEE Access 10:24395–24410
46. Hou H, Bi S, Zheng L, Lin X, Wu Y, Quan Z (2022) Dasecount: Domain-agnostic sample-efficient wireless indoor crowd counting via few-shot learning. IEEE Internet Things J 10(8):7038–7050
47. Android API level 28 (WiFi location: ranging with RTT). https://developer.android.com/guide/topics/connectivity/wifi-rtt. Accessed 02 Aug 2024
48. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30
49. Nie Y, Nguyen NH, Sinthong P, Kalagnanam J (2023) A time series is worth 64 words: Long-term forecasting with transformers. In: The eleventh international conference on learning representations. https://openreview.net/forum?id=Jbdc0vTOcol
50. Yang R, Zha X, Liu K, Xu S (2021) A cnn model embedded with local feature knowledge and its application to time-varying signal classification. Neural Netw 142:564–572
51. Dong L, Xu S, Xu B (2018) Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 5884–5888

52. Zerveas G, Jayaraman S, Patel D, Bhamidipaty A, Eickhoff C (2021) A transformer-based framework for multivariate time series representation learning. In: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining, pp 2114–2124
53. Garnot VSF, Landrieu L, Giordano S, Chehata N (2020) Satellite image time series classification with pixel-set encoders and temporal self-attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12325–12334
54. Maaten L, Hinton G (2008) Visualizing data using t-sne. J Mach Learn Res 9(11)
55. Liu W, Wang H, Shen X, Tsang IW (2022) The emerging trends of multi-label learning. IEEE Trans Pattern Anal Mach Intell 44(11):7955–7974. https://doi.org/10.1109/TPAMI.2021.3119334
56. Jacot A, Gabriel F, Hongler C (2018) Neural tangent kernel: Convergence and generalization in neural networks. Adv Neural Inf Process Syst 31
57. Du SS, Zhai X, Poczos B, Singh A (2018) Gradient descent provably optimizes over-parameterized neural networks. In: International conference on learning representations
58. Cao Y, Gu Q (2019) Generalization bounds of stochastic gradient descent for wide and deep neural networks. Adv Neural Inf Process Syst 32
59. Wu Y, Liu F, Chrysos G, Cevher V (2023) On the convergence of encoder-only shallow transformers. Adv Neural Inf Process Syst 36:52197–52237
60. Karimi H, Nutini J, Schmidt M (2016) Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In: Joint European conference on machine learning and knowledge discovery in databases. Springer, pp 795–811
61. Tian L, Chen L, Xu Z, Chen ZD (2021) A people-counting and speed-estimation system using wi-fi signals. Sensors. https://doi.org/10.3390/s21103472
62. Cheng Y-K, Chang RY (2017) Device-free indoor people counting using wi-fi channel state information for internet of things. In: GLOBECOM 2017-2017 IEEE global communications conference. IEEE, pp 1–6
63. Liu S, Zhao Y, Xue F, Chen B, Chen X (2019) Deepcount: Crowd counting with wifi via deep learning. arXiv preprint arXiv:1903.05316
64. Loshchilov I, Hutter F Decoupled weight decay regularization. In: International conference on learning representations
65. Choi J-H, Kim J-E, Kim K-T (2021) Deep learning approach for radar-based people counting. IEEE Internet Things J 9(10):7715–7730
66. Zhang C, Jiao W (2023) Imgfi: A high accuracy and lightweight human activity recognition framework using csi image. IEEE Sens J
67. Abuhoureyah FS, Wong YC, Isira ASBM (2024) Wifi-based human activity recognition through wall using deep learning. Eng Appl Artif Intell 127:107171

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Haruki Yonekura[1] · **Hamada Rizk**[1,2,3] · **Hirozumi Yamaguchi**[1,3]

✉ Haruki Yonekura
h-yonekura@ist.osaka-u.ac.jp

Hamada Rizk
hamada_rizk@ist.osaka-u.ac.jp

Hirozumi Yamaguchi
h-yamagu@ist.osaka-u.ac.jp

[1] Information Science and Technology, The University of Osaka, Yamadaoka 1-5, Suita, Osaka 5650871, Japan

[2] Department of Computer and Control Engineering, Tanta University, Gharbia, Tanta 31733, Egypt