Indoor Localization with Adaptive Signal Sequence Representations

Ning Liu, Tao He, *Student Member, IEEE,* Suining He, *Member, IEEE,* and Qun Niu, *Member, IEEE*

***Abstract*—Indoor location-based services (LBS) have exhibited large commercial and social values in smart cities, and urgent demands of which have spurred many localization techniques. Existing indoor localization approaches mostly rely on fingerprint techniques, leveraging either spatially discrete fingerprints or temporally consecutive ones for localization. However, these approaches often suffer from large errors or high time overhead in practice due to signal ambiguities or long input sequences. To overcome these drawbacks, this paper proposes a framework utilizing multiple adaptive *representations* of signal sequences for localization, where each representation indicates a corresponding signal structure with underlying location clues. As an example, the proposed approach takes geomagnetic signal sequences as input and infers location features from two intuitive representations, e.g., spatial and temporal ones. With adaptive signal representations, the proposed approach takes specifically optimized neural networks to extract corresponding location clues respectively and fuses them to generate more distinguishing features for more accurate localization. Furthermore, the ensemble learning mechanism is adopted in the approach and a weighted k-NN based location estimation algorithm is devised to enhance the robustness. Extensive experiments in three different trial sites demonstrate that the proposed approach outperforms state-of-the-art competing schemes by a wide margin, reducing mean localization error by more than 46%.**

***Index Terms*—indoor localization, geomagnetism, signal representations, neural networks, ensemble learning**

I. Introduction

The growing demands for indoor location-based services (LBS) and the popularization of smart mobile devices have spurred rapid development of indoor localization techniques. Indoor localization plays a more and more critical role in empowering Internet of Things for a wide range of applications, e.g., pedestrian or robot localization [1], [2], crowd monitoring [3] and targeted advertising [4], to name a few. Traditional satellite-based positioning and navigation systems (such as GPS) cannot meet the requirements of accurate indoor positioning due to signal attenuation caused by poor connectivity between end devices and satellites in indoor environment, which triggers researchers to bend their energies to further explorations on indoor localization.

With rapid popularization and ubiquitous nature of various sensors in mobile devices, e.g., radio-frequency sensor, imaging sensor and magnetometer, various signals captured by these sensors are employed for indoor localization, such as Wi-Fi [5], [6], Bluetooth Low Energy (BLE) [7], vision [8], and geomagnetism [9], [10]. Among all those signals explored, geomagnetic signal shows great application prospect due to its omnipresence, which means there is no need for any extra infrastructure deployment for localization. Moreover, geomagnetic signal which mainly generates from natural earth's magnetic field exhibits high global stability over time and it also has strong local variations due to metamorphic nature of indoor environment caused by nearby ferromagnetic objects, e.g., electrical appliances, steel-based building materials, which provide much promise for accurate localization [11]. And the impact of pedestrians on geomagnetic field is also marginal compared with that on other signals (vision, Wi-Fi, BLE). In that respect, geomagnetic signal is more adaptable especially in the scenes with large human traffic.

On the other hand, the underlying positioning algorithms are the key to indoor localization. Reviewing the existing positioning algorithms, fingerprint-based ones have drawn much attention [12], [13]. Most fingerprint-based positioning techniques can be broadly divided into two categories: *spatial* based and *temporal* based approaches. In the first category, the spatial location clues refer to discrete measurements of input signals at different indoor locations (e.g., a Wi-Fi/Bluetooth fingerprint, a geomagnetic measurement or an image at a fixed location). Ferromagnetic objects, such as doors, iron cabinets, escalators or lifts usually fluctuate nearby geomagnetic fields, posing distinguishing spatial patterns for localization [14]. Based on these discrete signal measurements with spatial location clues, existing approaches infer current position with the most similar geo-tagged signal fingerprint by comparing it with a pre-established database. However, suffering from poor distinctiveness of discrete signal measurements (locations far away may have very similar signal fingerprint), these approaches are usually incapable of achieving sufficient accuracy and robustness, especially in large spacious sites. The approaches that utilize discrete signal measurements are prone to location feature ambiguity and be easily impacted by random noise, which may cause large localization errors consequently.

To revamp the localization performance, some researches switch from discrete signal observations to explore context temporal correlations [15], in which temporal successive mea-

N. Liu is with School of Computer Science and Engineering, Sun Yat-sen University, and Guangdong Key Laboratory of Information Security Technology, Guangzhou 510006, China (e-mail: liuning2@mail.sysu.edu.cn).

T. He and Q. Niu are with School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China (e-mail: hetao23@mail2.sysu.edu.cn; niuq3@mail.sysu.edu.cn).

S. He is with Department of Computer Science and Engineering, The University of Connecticut, Storrs, CT 06269, USA (e-mail: suining.he@uconn.edu).

Manuscript received July 28, 2020; revised December 26, 2020 and March 31, 2021; accepted August 30, 2021. *(Corresponding author: Qun Niu.)*

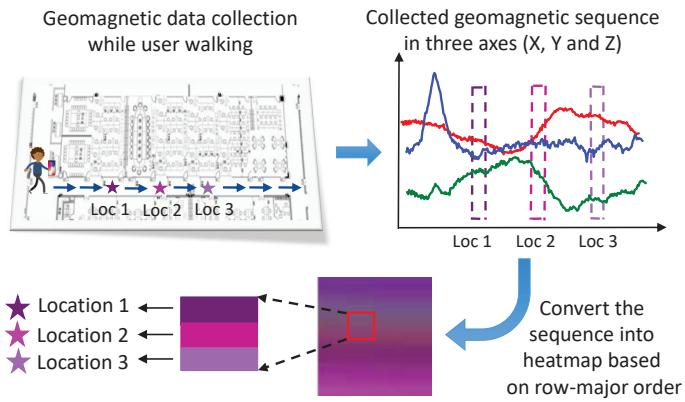


Fig. 1: For a geomagnetic signal sequence collected while user walking, we convert it into a geomagnetic heatmap. (Red/blue/green lines denote the components of geomagnetic sequence in three axes of X, Y and Z respectively.)

surements of signal are employed to indicate location clues, e.g., a signal RSSI (Received Signal Strength Indicator) vector, a video clip or a geomagnetic measurement sequence. These approaches evaluate the specific fluctuations of successive signal sequence in indoor environment and take advantage of this pattern which implies temporal clues to pinpoint current position. Taking temporal correlations between consecutive signal measurements into consideration, these temporal based approaches transcend previous spatial based ones which use discrete inputs. And the impact of signal random noise can be effectively reduced by means of employing such continuity constraints and temporal correlations, thus eliminating erroneous position estimations. However, employing long input signal sequences usually causes the increasing of computational complexity [16], which leads to large time consuming. To achieve lower time overhead, some approaches utilize short signal sequences as input (e.g., a short RSSI vector or a few frames of videos). Suffering from a limited spatial coverage of short sequences, it leads to degenerate distinctiveness of location clues and large localization errors consequently.

In this paper, we propose to utilize multiple adaptive representations of a signal sequence for accurate localization. More specifically, we devise a location estimation framework that considers both **S**patial and **T**emporal location features of signal for **Loc**alization, termed ***ST-Loc***. And the crux of ST-Loc is how to effectively extract distinctive location features under different dimensions. Firstly, instead of using raw data directly, we convert the input sequence into multiple adaptive representations, e.g., spatial and temporal ones in this paper, where each representation indicates a corresponding signal structure with underlying specific location correlations. Then, we extract corresponding location features on the basis of these signal representations respectively and fuse them together to generate more comprehensive and distinguishing fusion features for localization. To reduce the impact of random noise, we further improve its performance by employing ensemble learning mechanism in final location estimation, using selective ensemble method among multiple independent trained localization models to achieve higher robustness.

In summary, we make the following contributions:

- *Utilizing multiple adaptive representations of signal for more distinctive location features:* As discussed above, the distinctiveness of location features is critical for revamping localization accuracy. To facilitate distinctive feature extraction from signal, we propose to convert the original signal sequence into multiple adaptive representations with underlying correlations, e.g., spatial and temporal ones in this paper. Then, we adopt specifically optimized networks to extract features respectively and fuse them together for accurate localization.
- *Inferring spatial features through image processing methods*: A collected signal sequence is firstly converted into a heatmap, a spatial representation where each pixel corresponds to a spatial location and the pixel value denotes signal reading. Then, employing specifically optimized ResNet [17], we apply convolution to different patches of this heatmap, which correspond to spatially distributed signal observations at regular intervals in the local region, as shown in Fig. 1. In this way, we are able to infer spatial location features from these signal readings that span a long range, which reflect the regional correlation.
- *Extracting temporal features with hierarchical recurrent network*: For an ordered signal sequence, LSTM model is employed as basic unit in ST-Loc to extract underlying temporal features of signal sequence. Since it's usually time-consuming to extract temporal features with conventional LSTM from a long sequence directly, we devise a hierarchical structure to reduce the overall time overhead of feature extraction by means of sequence segmentation and parallel processing mechanism. Furthermore, we enhance the extracted features by employing a bidirectional scheme, considering both past and future contextual correlations in the sequence.
- *Employing ensemble learning for robust location estimation*: To reduce the impact of random noise and outliers, we apply ensemble learning mechanism in ST-Loc to improve robustness. We first construct multiple location estimation models with different initial training settings. Then by integrating these models to make a vote on localization results, ST-Loc overcomes the problem that the individual model is prone to random errors. On this basis, we further design a weighted k-NN based joint location estimation strategy for better adaptability facing complex indoor scenes. Through the mechanism above, we further improve localization accuracy and robustness of ST-Loc.

As mentioned earlier, geomagnetic signal benefits from its omnipresence, high global stability over time and strong local variations in indoor environment. In this paper, as an example, we take geomagnetic sequences as input to evaluate the localization performance of ST-Loc. We have conducted extensive experiments in three different trial sites and experimental results demonstrate that ST-Loc reduces mean localization error by more than 46% and achieves lower time overhead during localization compared with state-of-the-art competing approaches. Besides, ST-Loc can be also theoretically adapted

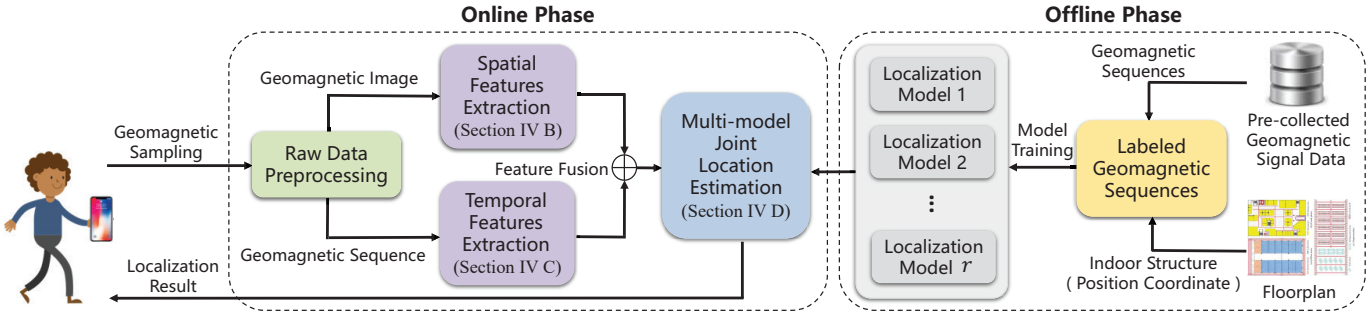


Fig. 2: The workflow of the proposed localization system.

to other signal sequences for indoor localization, such as Wi-Fi [5], Bluetooth [7] and vision [8] sequences.

The remainder of this paper is organized as following. We review related works in Section II. The workflow of ST-Loc is presented in Section III and the design of ST-Loc is elaborated in Section IV. Then we illustrate experimental results in Section V and finally conclude in Section VI.

II. Related work

In this section, we review the related works. To achieve higher accuracy and efficiency, researchers have explored various ambient signals for localization. e.g., Wi-Fi, Bluetooth Low Energy (BLE), visible light and so on. While Wi-Fi [18], BLE [19] and visible light-based [20] approaches are able to achieve sufficient accuracy in some specific scenes, they usually need to deploy extra infrastructures, which increases deployment and maintenance cost significantly. Even though pure vision-based approach [21] does not need external infrastructures, it is limited to specific scenes with rich textures. Moreover, the work in [22] proposes to utilize a short video clip to recognize and calibrate landmarks, then calculate the user's position based on triangulation technology. Although efficient, this approach need user's extra operations (record video in a specified manner), which is not user-friendly.

Compared with signals mentioned above, geomagnetism has attracted much attention lately. Considering spatial features of geomagnetic signals, some researchers evaluate the measurement of geomagnetic signal at different indoor locations and use this pattern to pinpoint users, which is inspired by that ferromagnetic objects, such as doors, iron cabinets, escalators or lift usually fluctuate nearby geomagnetic field, posing distinguishing spatial patterns. For example, LMDD [23] employs the geomagnetic pattern of door opening to discover doors. And SemanticSLAM [24] proposes to cluster geomagnetic signal observations so as to find landmarks for calibrating current position. However, such discrete signal observations still have a very limited discernibility, and single observation collected at different positions could be similar. This signal ambiguities may result in degraded distinctiveness of location features. As a result, discrete signal observation is not sufficient to be used as a unique location signature especially in large-scale indoor scene.

Noticing the temporal correlations of signal sequence, some researchers propose to leverage sequential measurements of signal as input. By vectorizing multiple successive signal observations to obtain higher dimensional temporal signature, these approaches enhance localization accuracy with such temporal location correlations. For instance, NaviLight [25] and Travi-Navi [26] both take signal sequence as input and employ dynamic time warping (DTW) algorithm for localization. However, the comparison of two sequences is usually computationally expensive and may result in high computational cost and time overhead especially when it needs to use relatively long signal sequences as input for sufficient accuracy. In addition, some researchers employ motion sensor assistance to achieve higher accuracy. The works in [16], [27] adapt particle filter mechanism to help positioning with signal fingerprints. Furthermore, WAIPO [28] and Magicol [29] fuse other signals (images, Wi-Fi) to enhance the localization accuracy. Although effective, those approaches still have some drawbacks. First, multiple signal-fusion localization means multiple signal data collection, which incurs higher cost of site survey. Second, the noise of inertial measurement unit (IMU) has a large impact on particle filter mechanism, which leads to potentially large localization error. Third, to achieve sufficient accuracy, they have to generate a large number of particles in particle filter mechanism, which also incurs large computational cost and causes high time overhead.

Recently, inspired by the success of deep learning algorithm, some approaches [30]–[33] propose to utilize neural networks to process signal sequences for predicting position. For example, DeepML [31] fuses magnetic field data and light intensity data and devises a long short-term memory (LSTM) based system for localization. The work in [33] proposes to utilize

TABLE I: Major symbols used in the paper.

Notation	Definition
$\mathbf{m}$	A single geomagnetic observation
$\mathbf{S}$	Geomagnetic observation sequence
$\mathbf{S}_{matrix}$	Geomagnetic matrix converted from $\mathbf{S}$
u	Width of the matrix $\mathbf{S}_{matrix}$
v	Height of the matrix $\mathbf{S}_{matrix}$
$\mathbf{F}^S$	Spatial feature extracted from geomagnetic heatmap
$\mathbf{F}^T$	Global extracted temporal feature
r	Preset number of trained models in ensemble learning
$\mathbf{L}$	An estimation location for input sequence

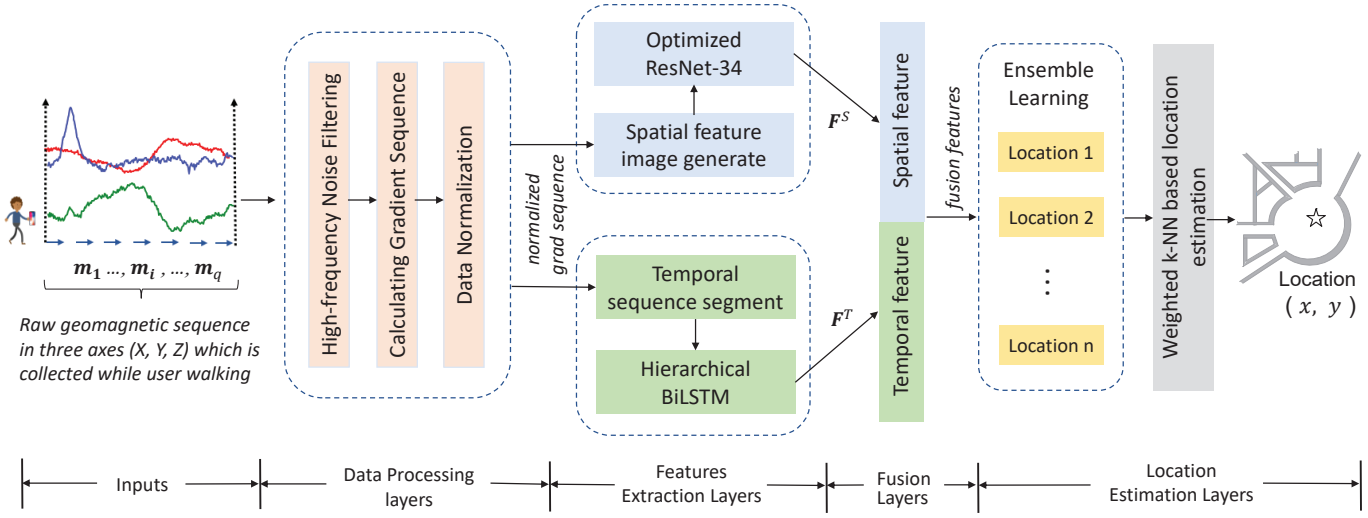


Fig. 3: Overall framework of ST-Loc.

the ordered geomagnetic sequence as input instead of discrete observations and employs a basic recurrent neural network (RNN) to extract the location features for localization. With state-of-the-art deep learning techniques, those approaches can achieve good performance in some typical scenes. However, facing more complex indoor scenes, there are few specifically optimized neural networks for indoor localization. And it's still hard to effectively extract valid location feature for sufficient localization accuracy just with a simple LSTM or a basic RNN, especially in large indoor site.

A preliminary version of ST-Loc has been reported in [34]. While it typically works well for localization with geomagnetic signal, its performance can be further improved. In this paper, we advance it as follows: 1) We first further analyze the characteristics of geomagnetic signal in detail, which provide a more comprehensive guidance for pre-processing of geomagnetic signal; 2) Considering the impact of statistical noises and outliers, we employ ensemble learning mechanism to enhance the robustness and further devise a weighted k-NN based location estimation algorithm for higher accuracy and robustness. 3) We also conduct more complete and extensive experiments in real trial sites to evaluate the performance of proposed network, and experimental results show that ST-Loc with further enhancements is able to achieve higher robustness and further reduce mean localization error, compared with previous version.

III. System Workflow

In this section, we present the workflow of proposed localization system in Fig. 2. For demonstration, we make use of geomagnetic sequences and consider both spatial and temporal features of which to locate targets.

The system workflow consists of two phases, an *offline* phase and an *online* phase. In the offline phase, we collect geomagnetic data in the trial site and use the labeled data to train the localization models based on the designed network. Specifically, we first design dense survey paths in the public area of the trial site according to its floorplan. Then surveyors walk along these designed paths, carrying a client device which records the sensor data including geomagnetic signal along the paths. Combining the indoor structure information (location coordinate) from floorplan, we label each collected geomagnetic sequence with the location coordinate where the last geomagnetic sample is collected. Then we store these labeled geomagnetic sequences in a database. Based on the constructed database, we take advantage of ensemble learning mechanism to train multiple localization models which will be used in online phase for location estimation.

In the online phase, each user carries a client device and walks in the trial site. The client program records the sensor data along the path including geomagnetic measurements automatically. To reduce the impact of the external factors, we first process the collected geomagnetic sequences to resolve the random noise and heterogeneity problem (Section IV-A). Then we convert the processed geomagnetic sequences into spatial and temporal representations respectively and extract corresponding features for localization (Section IV-B and Section IV-C). With multiple localization models trained in offline phase, we finally employ ensemble learning technique and devise an adaptive combination strategy for multiple trained models to enhance the accuracy and robustness of final location estimation (Section IV-D).

IV. Detailed Design of ST-Loc

In this section, we elaborate the design of proposed ST-Loc which employs both the temporal and spatial clues of signal sequences for localization. We first illustrate the overall structure of ST-Loc in Section IV-A. Then, we present the process of spatial and temporal feature extraction in Section IV-B and Section IV-C respectively, followed by the details of ensemble learning based position estimation in Section IV-D. Finally, a brief time complexity analysis is presented in Section IV-E. And Table I lists the major symbols used in this paper.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2021.3113333

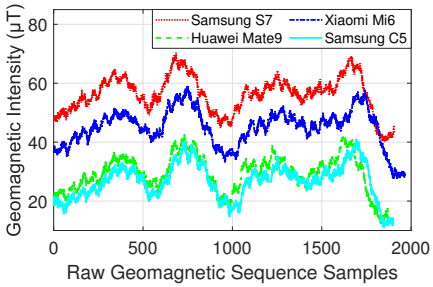


Fig. 4: Raw geomagnetic sequence collected with different devices along a same indoor trajectory.

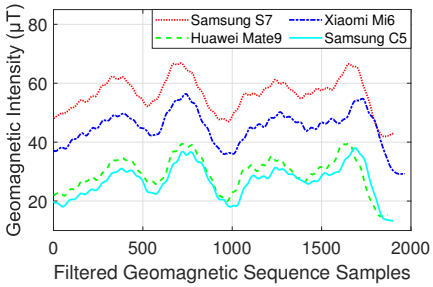


Fig. 5: Filtered geomagnetic sequence collected with different devices along a same indoor trajectory.

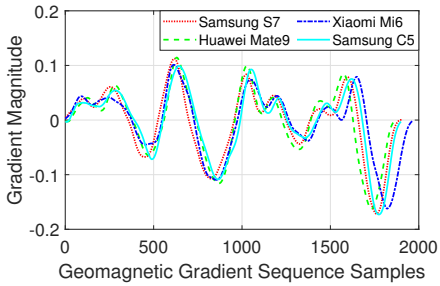


Fig. 6: Gradient magnitude of geomagnetic sequence collected with different devices along a same indoor trajectory.

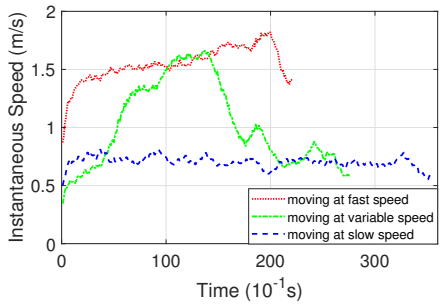


Fig. 7: Moving speed under different walking modes along a same trajectory (fast, slow and variable speed).

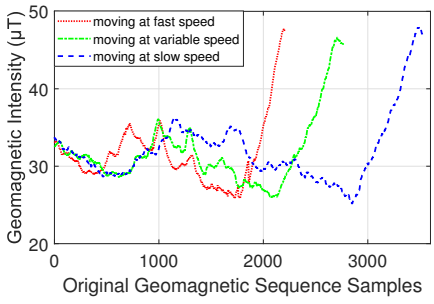


Fig. 8: Original geomagnetic sequence under different moving speed along a same indoor trajectory.

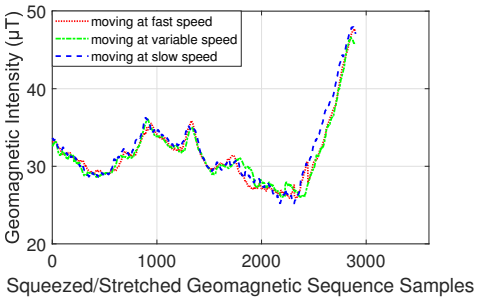


Fig. 9: Squeezed/Stretched geomagnetic sequence under different moving speed along a same indoor trajectory.

A. *Overall structure of ST-Loc*

Knowing the limited feature discernibility of single geomagnetic measurement, ST-Loc takes the consecutive geomagnetic sequences as input. Moreover, ST-Loc considers both temporal and spatial location features of geomagnetic sequences for localization. The overall framework of ST-Loc is presented in Fig. 3. Noticing the noise of device sensor, we first filter high frequency noise of the sequence, and employ the gradient of sequence as input instead of raw sequence to deal with the device heterogeneity problem. With processed sequence, we design a hierarchical BiLSTM to capture the corresponding temporal features. Meanwhile, we convert the geomagnetic sequence into a heatmap, and apply a specially optimized ResNet to extract spatial features from resulted heatmap. Then we fuse extracted temporal and spatial features to generate more comprehensive and distinctive fusion features. Finally, we employ ensemble learning mechanism on ST-Loc and devise a weighted k-NN based multi-model joint position estimation strategy for final position estimation. More specifically, ST-Loc consists of four major modules: 1) Data preprocessing; 2) Multi-scale spatial feature extraction; 3) Hierarchical temporal feature extraction and 4) Ensemble learning based position estimation. We overview each module as following:

1) *Data pre-processing*. For raw geomagnetic sequence, as shown in Fig. 4, it is inevitably mixed with random noise caused by user motion and other factors. Therefore, we first apply empirical mode decomposition (EMD) [35] algorithm to filter out high frequency noise of the raw sequence to obtain filtered geomagnetic sequence as presented in Fig. 5. Then for device heterogeneity (various devices or sensors may have different calibrations for magnitude of geomagnetic field intensity as illustrated in Fig. 4), we calculate the gradient of filtered geomagnetic sequence as input instead of using geomagnetic sequence directly, in view of that the distortions of geomagnetic sequences collected by different devices at the same position are the same. Utilizing gradient sequence, as shown in Fig. 6, we don't have to put extra effort to calibrate different devices to a uniform standard.

In addition, user (speed) heterogeneity problem also needs to be considered in practice. Even using a same device along a same trajectory, different user moving speed usually leads to distinct geomagnetic sequence with different scales as shown in Fig. 8. To address this, we first estimate the speed of users based on the inertial measurement unit of mobile devices with state-of-the-art techniques [36]. And the estimated speed information is presented in Fig. 7. Then we segment original geomagnetic sequence into many small subsequences. Based on estimated user speed in each segment, we stretch or squeeze the subsequences to a standard length that corresponds to the reference speed. Finally, we concatenate these stretched or squeezed subsequences together so that geomagnetic sequences with different scales are mapping to a uniform standard as shown in Fig. 9.

2) *Multi-scale spatial feature extraction*. To enhance the

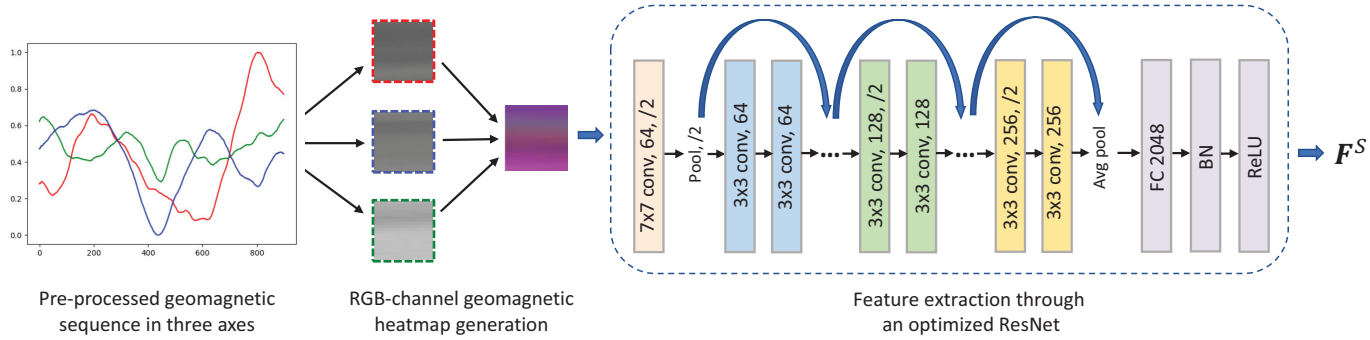


Fig. 10: Spatial representation and corresponding feature extraction.

distinctiveness of spatial location clues, we propose a spatial representation of magnetic sequence and design a multi-scale spatial feature extraction (MSFE) model. We first convert input geomagnetic sequence into a high dimensional geomagnetic heatmap, a spatial representation where each pixel is corresponding to a spatial position. The resulted heatmap is able to provides more regional correlations of distributed signal observations compared with low dimensional sequence. Then we utilize a specially optimized ResNet to extract more distinctive spatial feature from the heatmap for localization. Details of MSFE is presented in Section IV-B.

3) *Hierarchical temporal feature extraction*. Intuitively, we take advantage of state-of-the-art LSTM model to extract temporal features. Nevertheless, it is still hard in practice to efficiently correlate from the first to last instance for a long sequence by feeding the sequence to a LSTM model directly. And the serial processing for a long sequence is also time consuming. Therefore, we devise a hierarchical structure, employing sequence segmentation and parallel processing mechanism to efficiently extract temporal features while reducing overall time overhead. Furthermore, we adopt a bidirectional LSTM (BiLSTM) to enhance the extracted features. And Details are illustrated in Section IV-C.

4) *Ensemble learning based position estimation*. With extracted spatial and temporal features, we concatenate them to generate more distinctive fusion features for localization. Moreover, we employ ensemble learning mechanism to further improve the robustness and accuracy of ST-Loc. To be specific, we train multiple independent models based on proposed network with different initial settings, from which we can obtain multiple predictions of position. Furthermore, we design a weighted k-NN (k-nearest neighbors) based algorithm to integrate those predictions to get the final estimation position. Details of ensemble learning based position estimation will be elaborated in Section IV-D.

B. Multi-scale spatial feature extraction

1) Spatial representation of the geomagnetic sequence: Considering that discrete and low dimensional signal fingerprint lacks sufficient spatial distinctiveness, we propose a spatial representation of input signal, converting raw signal sequence into a high dimensional heatmap so as to obtain more regional correlations of distributed signal observation. In this paper, we convert a raw geomagnetic sequence into a geomagnetic heatmap, where each pixel denotes a single geomagnetic measurement in the sequence and pixel value corresponds to intensity of the measurement. As shown in Fig. 1, a geomagnetic sequence collected while the user walking is converted into a geomagnetic heatmap. And we conceive of a local window (denoted by red block in Fig. 1) in the resulted heatmap, rows of which are actually sub-portions of original geomagnetic sequence and corresponding to a sequence of spatially distributed locations in a local region. So employing convolution to patches of the heatmap, we are able to extract features that represent spatial clues of geomagnetic measurements distributed in a local region. And these regional correlations will provide more distinctiveness to construct spatial features for localization.

More specifically, as shown in Fig. 10, a single geomagnetic measurement usually consists of values in three axes (X, Y and Z). For a collected geomagnetic sequence $\mathbf{S}$ of length q, which has been pre-processed as elaborated in Section IV-A:

$$\mathbf{S} = \{\mathbf{m}_1\, \mathbf{m}_2, \mathbf{m}_3, ..., \mathbf{m}_q\}, \tag{1}$$

where $\mathbf{m}_i = (x_i, y_i, z_i)$ denotes a processed single geomagnetic observation including signal reading values in three axes. We first reshape it to a three-dimensional rectangular matrix, then normalize all elements of the matrix to RGB color space $[0, 255]$ as following:

$$\mathbf{S}_{matrix} = \begin{bmatrix} \mathbf{m}_1 & \mathbf{m}_2 & \cdots & \mathbf{m}_u \\ \mathbf{m}_{u+1} & \mathbf{m}_{u+2} & \cdots & \mathbf{m}_{u+u} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{m}_{u(v-1)+1} & \mathbf{m}_{u(v-1)+2} & \cdots & \mathbf{m}_{uv} \end{bmatrix}, \tag{2}$$

$$\tilde{\mathbf{S}}_{matrix} = \frac{255 \cdot (\mathbf{S}_{matrix} - Min(\mathbf{S}_{matrix}))}{Max(\mathbf{S}_{matrix})) - Min(\mathbf{S}_{matrix})}, \tag{3}$$

where u, v denote the *width* and *height* of the matrix ($uv = q$), and $Max(\cdot)$, $Min(\cdot)$ denote *Maximum function* and *Minimum function* respectively.

With normalized geomagnetic matrix $\tilde{\mathbf{s}}_{matrix}$, we convert it into an RGB-channels image, in which the values (r, g, b) of a pixel in three channels are corresponding to the values (x, y, z) of an element (a single geomagnetic measurement) in the matrix.

2) Optimized residual neural network: ST-Loc employs an optimized ResNet to extract features from geomagnetic heatmaps following transfer learning mechanism. ResNet tries to address the notorious vanishing gradients problem where gradients decrease slowly, preventing weights from changing their values. To address this, researchers realize ResNet with *shortcut connections*, where they skip one or several layers to achieve higher accuracy with more layers. We refer interested readers to [17] for more details.

Benefiting from the residual structure, ResNet has achieved superior accuracy in various image processing tasks. However, original ResNet is primarily adopted to processes natural images, which are fundamentally different from geomagnetic heatmaps in terms of image properties and task objectives. According to the empirical study in transfer learning [37], the activations in ResNet are too concentrated around the object (i.e., only features tightly related to the source domain have strong responses) when FC (fully connected) layers are missing. This relationship makes it inappropriate to transfer to a target domain if the source and target are distant from each other. But the models with FC layers show a different property. The distributed activations enable them to capture useful image features in the target domain when the target is dissimilar to the source domain. So the research [37] concludes that when image properties or task objectives in the source domain are far different from those in the target domain, it is essential to add FC layers in pre-trained model of the source domain. Hence, ST-Loc takes original ResNet which is pre-trained on ImageNet [38] as foundation, then replaces original classification layer with extra FC layers, normalization layers and non-linear activate function layers. Finally, using geomagnetic heatmaps as training data, we fine-tune the reconstructed ResNet for feature extraction.

More specifically, as illustrated in Fig. 10, the final classification layers of original ResNet are removed firstly, then the remain of the network will output a 512-D feature vector $\mathbf{f}^S$. Subsequently, we map this 512-D feature vector to higher dimensional vector by inserting a 2048-D FC layer [37] followed by a batch normalization layer. Finally, we add a rectified linear unit (ReLU) as non-linear activate function. The process is as following:

$$\tilde{\mathbf{f}}^S = W\mathbf{f}^S + b, \tag{4}$$

$$\mathbf{f}^S_{norm} = \frac{\gamma}{\sqrt{Var[\tilde{\mathbf{f}}^S] + \epsilon}} \cdot \tilde{\mathbf{f}}^S + (\beta - \frac{\gamma E[\tilde{\mathbf{f}}^S]}{Var[\tilde{\mathbf{f}}^S] + \epsilon}), \tag{5}$$

$$\mathbf{F}^S = \eta \cdot ReLU(\mathbf{f}^S_{norm}), \tag{6}$$

where $E[\cdot]$ and $Var[\cdot]$ denote *mean* and *standard deviation* respectively, and $W, b, \beta, \gamma, \epsilon$ and η are learnable parameters.

Finally, with fine-tuned ResNet, we extract multi-scale spatial feature $\mathbf{F}^S$ from the geomagnetic heatmap.

C. Hierarchical temporal feature extraction

1) Temporal representation of the geomagnetic sequence: In a plenty of fingerprint-based approaches, position estimation is entirely independently based on a single signal fingerprint (a signal measurement). Unfortunately, those approaches are usually prone to feature ambiguity and easily impacted by random noise, especially in large indoor scene. In practice, geomagnetic signal collected by devices is actually an ordered geomagnetic signal observation sequence over time dimension, in which each geomagnetic signal measurement is associated with other adjacent measurements. As shown in Fig. 4, such correlations (sequential fluctuation trend) of geomagnetic signal sequence is especially obvious and distinctive in complex indoor environment. Extracting and applying these temporal correlations and continuity constraints in geomagnetic signal sequence can effectively improve the distinguishability of location features. Intuitively, we propose to take advantage of consecutive geomagnetic sequence (a temporal representation) as input instead of discrete signal measurements. Then we are able to extract such temporal correlations and continuity constraints which provide more distinctive location clues in temporal dimension.

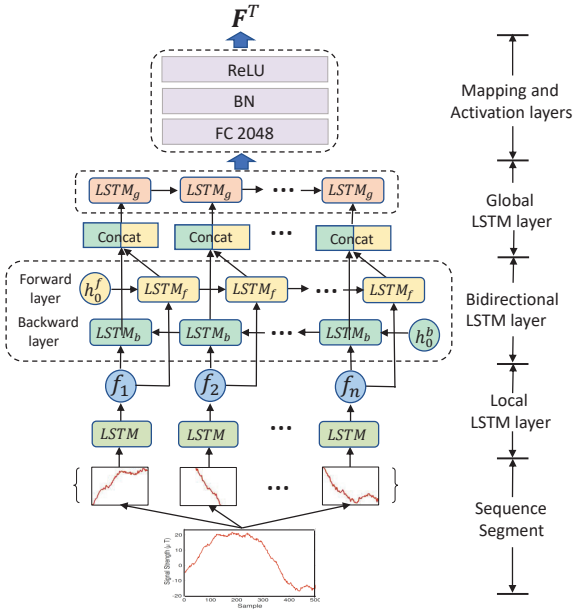


Fig. 11: Temporal representation and corresponding feature extraction.

2) Hierarchical bidirectional LSTM: In order to capture these temporal correlations and continuity constraints, we employ state-of-the-art LSTM model in ST-Loc. Bringing in *gate mechanism*, LSTM address the vanishing gradient problem and makes an improvement on standard RNN. Benefiting from this mechanism, LSTM model is able to learn long-term dependencies of input sequence, which usually applies following operations at each timestep:

$$f_t = \sigma_g(W_f\mathbf{x}_t + U_f\mathbf{h}_{t-1} + b_f), \tag{7}$$

$$i_t = \sigma_g(W_i\mathbf{x}_t + U_i\mathbf{h}_{t-1} + b_i), \tag{8}$$

$$o_t = \sigma_g(W_o\mathbf{x}_t + U_o\mathbf{h}_{t-1} + b_o), \tag{9}$$

$$\tilde{c}_t = \sigma_c(W_c\mathbf{x}_t + U_c\mathbf{h}_{t-1} + b_c), \tag{10}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t, \tag{11}$$

$$\mathbf{h}_t = o_t \circ \sigma_h(c_t), \tag{12}$$

$$\mathbf{y}_t = \sigma_o(W_y\mathbf{h}_t + b_y), \tag{13}$$

where $\mathbf{x}_t$ and $\mathbf{h}_t$ represent the input and hidden state at time t. And σ denotes the non-linear activation function. W, U and b are the learnable parameters. f, i, o represent *forget gate*, input and output *reset gates* respectively, and c is a *memory cell state*. The cell in LSTM is able to keep, update or forget feature information over time by these gates.

However, for a long input sequence, it is still hard in practice to efficiently correlate from the first to last instance by using a LSTM model directly. At same time, the serial processing for a long sequence is also time consuming. Therefore, we devise a hierarchical structure, as presented in Fig. 11, we first segment a long geomagnetic sequence into many subsequences and extract the temporal features of these local subsequences with low-level LSTM respectively, then we take these extracted features as input of a high-level LSTM to extract global temporal features in higher dimensions. More specifically, we segment input sequence with specific scale to get subsequence set $\{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n\}$. Then for these local subsequences, we extract the temporal features respectively, e.g., in the case of a local subsequence $\mathbf{s}_i$, we extract corresponding local temporal feature $\mathbf{f}_i$. Then extracted local features $\{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_n\}$ will be taken as input of a high-level LSTM to extract global temporal features. Applying this hierarchical structure, each LSTM unit in the network processes shorter subsequence, which means that it is able to keep more detail features and reduce loss of information that easily occurs in long sequence processing. At same time, with this sequence segmentation and parallel processing mechanism, we can effectively reduce the average time overhead of temporal feature extraction.

Furthermore, considering both past and future contextual temporal correlations of signal sequence, we make use of a bidirectional LSTM (BiLSTM) scheme to enhance the extracted local temporal features $\{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_n\}$. As illustrated in Fig. 11, BiLSTM takes the ordered feature sequence $\{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_n\}$ as input. Then we get the enhanced local temporal features $\{\tilde{\mathbf{f}}_1, \tilde{\mathbf{f}}_2, ..., \tilde{\mathbf{f}}_n\}$ which involve both past and future contextual temporal dependencies. The bidirectional LSTM has same state equations as Equation 7-13, but uses both forward and backward hidden states at each timestep as following:

$$\tilde{\mathbf{f}}_i = BiLSTM([\mathbf{h}_i^f, \mathbf{h}_{n-i}^b], \mathbf{f}_i), \tag{14}$$

$$h_0^f(x) = h_0^b(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \tag{15}$$

where $\mathbf{h}^f$ and $\mathbf{h}^b$ denote the forward and backward hidden states respectively. And we initialize the hidden state with a standard Gaussian Function.

Finally, we take enhanced local temporal features as input and utilize a high-level LSTM to extract global temporal feature $\mathbf{F}^T$, then map $\mathbf{F}^T$ to fixed size for feature fusion.

D. Ensemble learning based position estimation

1) Motivations for applying ensemble learning mechanism: In the supervised learning algorithm of machine learning, the goal is to learn a stable model with good performance. However, suffering from random noise or outliers of training data, sometimes we can find several different models with distinct predilection after training with different initial settings, which almost give same accuracy on the training dataset. From a statistical point of view, the effect of those factors can be reduced when the amount of training data is large enough. For indoor localization, considering inevitable existence of various random noises and outliers in collected signals, we can repeat site survey enough times to obtain more sufficient training data. So there will be multiple redundant signal measurements for each survey path in the trial site, and the effect caused by noise and outliers can be effectively reduced though statistical method.

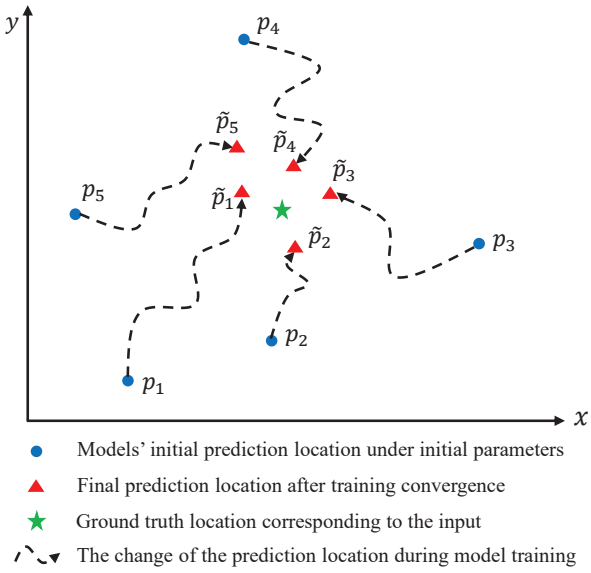


Fig. 12: Schematic diagram of ensemble learning based localization that training multiple independent models and integrating their prediction results for joint position estimation.

While the training data is sufficient enough (so that the effect mentioned above is absent), however, it has to face another challenge that computational cost is usually unacceptable when dataset is enormous as the optimal training of neural networks is proved to be NP-hard [39]. At same time, it's also labor-intensive and time-consuming to do vast repeated site surveys to obtain sufficient enough training data which is usually redundant.

To address above, we employ ensemble learning mechanism. With different initial training setting, we train proposed network to get multiple independent models, which have different degree of sensitivity to signal noise and outliers but almost share same accuracy on training dataset after convergence. Then the estimation algorithm will take all these models' predictions into consideration and make an average among prediction results so as to reduce the risk of choosing single prediction with large random error. The underlying idea is that even if one single sub-model gets prediction with large error, other sub-models can correct the error back. Thus we can efficient reduce large deviations which usually caused by signal noise and outliers.

In this paper, we set different initial parameters for training multiple independent models based on Gaussian distribution.

Algorithm 1: Find the optimal k

Input:
S_{train} representing the train dataset;
S_{val} representing the validation dataset;
r representing preset number of trained models;

Output:
Optimal k representing the number of selected models in weighted k-NN based algorithm;

Train the proposed ST-Loc on dataset S_{train} with different initial parameters to obtain r independent models $\{model_1, model_2, ..., model_r\}$;
for *each* $model_i$ $(1 \leq i \leq r)$ **do**
  for *each testcase in* S_{val} **do**
    Obtain the corresoding prediction location though $model_i$;
  Calculate the mean prediction error e_i of $model_i$ on whole S_{val} as Equation 24;
for j *in* $range(1, r+1)$ **do**
  Find j least prediction errors in $\{e_1, e_2, ..., e_r\}$ and obtain correaponding j models;
  Calculate weights of j models obtained above as Equation 18 and 19 ;
  for *each testcase in* S_{val} **do**
    Obtain the prediction locations of j selected models above, respectively;
    Calculate the weighted average prediction location for the testcase as Equation 20 ;
  Calculate mean prediction error on whole S_{val} with j selected models as Equation 24;
Find the optimal k with which ST-Loc could achieve minimum mean localization error on S_{val};
final;
return optimal k;

More specifically, for each layer of a training model, we set initial parameters based on a Gaussian distribution as follows:

$$Paras \sim N(\mu, \sigma^2), \tag{16}$$

and its probability density function is defined as:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{17}$$

where $\mu \in [-1, 1]$ and $\sigma^2 \in (0, 10]$, which are chosen at random from the corresponding range for each different training model.

Fig. 12 illustrates a schematic example of position estimation in this situation. For an input signal sequence, we can first obtain the corresponding initial prediction starting location $\{p_1, p_2, ..., p_5\}$ of multiple independent models $\{model_1, model_2, ..., model_5\}$ with different initial parameters. And we use dotted line to indicate the convergence process of model's predicted position during iterative training, adopting gradient descent optimization method. $\{\tilde{p1}, \tilde{p2}, ..., \tilde{p5}\}$ denote the final prediction locations of those trained models after model convergence and green star represents the corresponding ground truth location. As we can see in Fig. 12, these models' final predictions more or less deviate from ground truth location which mainly suffers from random noise and signal outliers. Intuitively, integrating these independent location prediction models, we can obtain a more accurate prediction of target location by taking an average among their prediction results.

Algorithm 2: Weighted k-NN based voting for final position estimation

Input:
S_{test} representing the test dataset;
$\{model_i | i = 1, 2, ..., r\}$ representing r models trained with different initial parameters;
$k (k \leq r)$ representing the optimal number of selected models for joint position estimation;

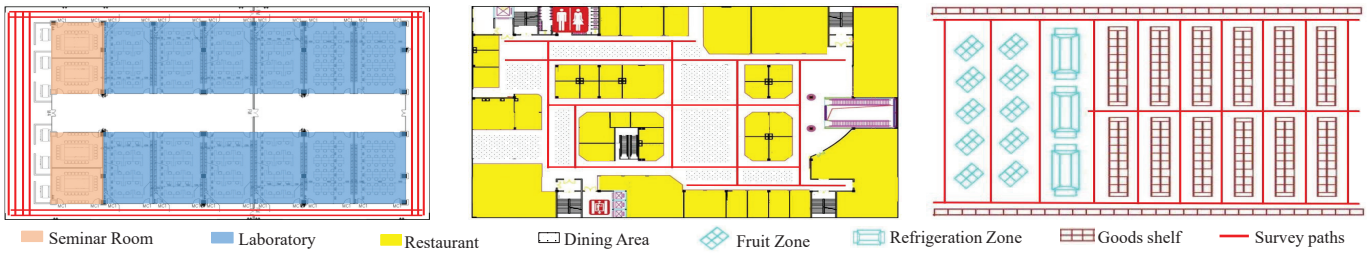
Output:
Position estimation for every testcase in S_{test}

Find k models with least mean validation error in r independent trained models $\{model_i | i = 1, 2, ..., r\}$;
Calculate corresponding weights of k selected models above as Equation 18 and 19 ;
for *each testcase in* S_{test} **do**
  Obtain the prediction locations of k selected models above, respectively;
  Calculate weighted average prediction location for the testcase as Equation 20;
final;
return Position estimation for every testcase in S_{test};

In summary, we apply ensemble learning mechanism in final position estimation, and we train multiple different models independently and combine them to generate a more robust and comprehensive ensemble model, which is able to effectively reduce the impact of noise and outliers, achieving higher accuracy and robustness without large amount of redundant training data collection.

2) Weighted k-NN based voting for position estimation: Based on ensemble learning method, we train multiple models independently with different initial settings. So for each input, we can obtain multiple corresponding prediction locations. To further improve the accuracy and robustness of prediction, we design a weighted k-nearest neighbors (k-NN) based algorithm on those prediction results to calculate the final estimation location for the input. First of all, we need to find a heuristically optimal number k of nearest neighbors, and the details are shown in Algorithm 1. Then we select k models with the smallest average validation error from all trained models and calculate final prediction location among those selected models' prediction results.

More specifically, as shown in Algorithm 2, suppose we have selected k models out of r trained models (r denotes the preset number of trained models), then for an input sequence, we get the corresponding prediction locations of the selected models: $\{\mathbf{L}_1, \mathbf{L}_2, ..., \mathbf{L}_k\}$. Furthermore, we set



(a) Office area. (b) Food plaza. (c) Supermarket.

Fig. 13: Floorplans of the trial sites.

adaptive weight for each selected model's prediction according to model's average validating error. And we choose Gaussian function as Error-Weight transition function and then make a normalization as follows:

$$\tilde{w_i} = \frac{1}{\sqrt{2\pi}} e^{-\frac{d_i{}^2}{2}}, \tag{18}$$

$$w_i = \frac{\tilde{w_i}}{\sum_{j=1}^{k} \tilde{w_j}}, \tag{19}$$

where w_i is the weight of the corresponding prediction and d_i is the average validation error of the corresponding model, and $i \in \{1, 2, ..., k\}$. Then, we calculate weighted average of k selected models' prediction locations:

$$\mathbf{L} = \sum_{i=1}^{k} w_i \mathbf{L}_i, \tag{20}$$

Finally, we get a more accurate final prediction location $\mathbf{L}$ for the input geomagnetic sequence.

E. Time complexity analysis

As an important requirement in indoor localization application, real-time performance is always need to be paid more attention. Traditional fingerprint-based approaches mostly rely on signal matching strategy utilizing discrete signal observation or successive signal sequence. And the time overhead for localization largely depends on the matching algorithm and the size of database used for matching.

Dynamic Time Warping (DTW) is a widely used technique to measure the similarity between two sequences, which considers both stretching and squeezing the sequences to align them. Making use of signal sequence as input, some approaches [26], [29], [40] leverage DTW algorithm as signal matching strategy for indoor localization. DTW algorithm uses dynamic programming to calculate the similarity between two time sequences, and the time complexity can be expressed as:

$$\mathcal{O}(n^2), \tag{21}$$

where n denotes the length of the input sequence. For indoor localization, it needs to compare observed signal sequence with each signal fingerprint in pre-established database though iteration and infer current location with the most similar geo-tagged fingerprint. Suppose the size of the database is m (the number of fingerprints that need to be matched with), then the time complexity for localization based on this mechanism:

$$\mathcal{O}(m \cdot n^2), \tag{22}$$

As discussed above, those iterative matching-based approaches usually incur high time overhead during localization, especially when it needs to use relatively long signal sequences for sufficient accuracy or the database for matching is enormous in the large indoor scene.

In this paper, ST-Loc employs deep learning technique to establish an end-to-end system and realize offline learning and online calculation of proposed localization model. Although deep learning methods is data-hungry which usually need large number of trainings to learn sufficient knowledge, time-consuming offline training only need to be conducted once generally. Then the online localization can continuously proceed without extra time-consuming model training. That's once and for all. On the other hand, we always pay more attention to online localization performance in actual application. And the trained model contains only simple linear and nonlinear transformation units, and the computational complexity of these operations is also lower. In addition, the computational complexity of employing trained model is also independent of the size of the database. Therefore, ST-Loc is able to greatly reduce the time overhead during online localization compared with matching-based approach mentioned above, providing a guarantee for real-time localization service.

Compared with other approaches [31], [33] which utilize deep learning technique, ST-Loc further devise a hierarchical structure for temporal feature extraction as elaborated in Section IV-C. Through the sequence segmentation and parallel processing mechanism, each basic LSTM unit in the network processes a relatively shorter sequence, and the unit at lower level can pay more attention to the local temporal feature of the corresponding subsequence. At same time, parallel processing can also effectively reduce the time overhead of the network.

V. Illustrative Experimental Results

In this section, we evaluate the performance of ST-Loc and state-of-the-art comparison schemes. We present dataset and experimental settings in Section V-A, followed by comparison schemes and evaluating metrics in Section V-B. Then the experimental results are illustrated in Section V-C and we finally analyze system overhead in Section V-D.

A. Dataset and experimental settings

We have conducted extensive experiments in three trial sites including an office area in our university, a spacious food court and a supermarket. The site plans are shown in Fig. 13.

TABLE II: Datasets established in three trial sites

Parameter \ Dataset	Office Area		Food Plaza		Supermarket	
	train	test	train	test	train	test
Length of a Sequence	500	500	500	500	500	500
Number of Sequences	2390	770	1952	482	496	244

The supermarket area covers around 720 m^2, the office area covers around 2,800 m^2 and the food plaza is more spacious which covers around 3,500 m^2. Without loss of generality, we also conduct our experiments on a variety of mobile devices (including Samsung Galaxy S7, Samsung Galaxy C5, Huawei Mate 9 and Xiaomi Mi 6). Meanwhile, we also invite multiple volunteers to participate in our experiments to evaluate the performance of ST-Loc.

Suffering from signal instability and multipath fading effect, RSS (Received Signal Strength) usually needs to be measured several time. Compared with RSS, geomagnetic signal is much more easier to collect, benefiting from stable geomagnetic signal distribution. And most mobile devices are equipped with high sensitivity magnetic field sensors which are able to achieve sampling frequency up to 100 Hz. In experiments, the signal sampling frequency of all devices is set to 50 Hz to achieve trade-off.

In order to build dataset for experiments, we develop a signal collection application based on Android system. The application collects various signals including geomagnetic signal strength and IMU (Inertial Measurement Unit) sensor data (IMU data is only needed for some comparison schemes). When surveyors walk though a preset survey path, the application will collect various signals along the path, then we can obtain a mixed signal sequence which corresponds to the survey path:

$$\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, ...\}, \tag{23}$$

where $\mathbf{v}_i = \langle \mathbf{m}_i, \mathbf{a}_i, \mathbf{g}_i, \mathbf{o}_i \rangle$ in which $\mathbf{m}_i$ denotes geomagnetic signal strength and $\mathbf{a}_i, \mathbf{g}_i, \mathbf{o}_i$ represent acceleration vector, gyroscope angles and orientation angles respectively.

In terms of dataset annotation for experiments, the collected geomagnetic sequence covers a path, so we label the sequence with the ground truth location coordinate where the last geomagnetic sample is collected. More specifically, we mark the locations of starting and ending points of each survey path based on nearby landmarks, e.g., doors, corners. Then, we label the ending locations of each signal segment based on distances to nearest landmarks. Note that in indoor trial sites, the number of these landmarks are usually large. Our data annotation benefits from this with higher accuracy. For training dataset, we have designed dense survey paths (denoted by red solid lines in Fig. 13) in public areas according to the floorplans of three trial sites and we collect signal sequences along those survey paths in three trial sites for training, respectively. As for testing dataset, experiment participants are requested to walk though some randomly chosen paths with mobile device in trial sites, then we use signal sequences collected in three trial sites for evaluation, respectively. Table II presents the detailed information of the datasets in three trial sites.

TABLE III: Baseline training parameters in experiments

Parameter \ Trial Site	Office Area	Food Plaza	Supermarket
Iterations	500	300	300
Mini-batch	125	125	125
Initial Learning Rate	0.005	0.002	0.001
Weight Decay	2e-4	2e-4	2e-4

We train proposed ST-Loc separately with training dataset built for each trial site and evaluate its performance in corresponding site, respectively. And baseline training parameters are presented in Table III. We choose *PyTorch* as deep learning framework in experiments and use *Adam* as deep network's optimizer. *MSELoss* is set as loss function. All experiments are performed on a simulation server installing Ubuntu 16.04 system with four Nvidia RTX 2080Ti GPU cards, an Intel Xeon E5-2640 CPU and 128 GB memory.

B. Comparison schemes and evaluating metrics

We compare ST-Loc with following state-of-the-art indoor localization methods which use geomagnetic signal as input:

- MaLoc [27] devises a reliability-augmented particle filter to improve the accuracy and robustness of position estimation. Furthermore, it proposes an adaptive sampling algorithm to reduce computation overhead so as to improve the overall usability.
- Magicol [29] overcomes the low discernibility of the geomagnetic signal by vectorizing consecutive geomagnetic measurements. It calibrates user positions with a bi-directional particle filter and uses the vectors to shape the particle distribution in position estimation process.
- *RNN-4* [33] also takes geomagnetic sequence as input. And it trains a standard RNN to predict user position. In our experiment, we build a deeper network with 4-layer RNN as a comparison scheme.

Moreover, in order to evaluate the effectiveness of core function components in proposed model including spatial and temporal feature extraction modules, we also take the following variants of the model into comparison:

- *ST-Loc-ns*: We remove the spatial feature extraction module from the network and only use temporal features for localization, by which we can validate the effectiveness of extracted spatial features.
- *ST-Loc-nt*: On the contrary, to validate the effectiveness of extracted temporal features, We remove the temporal feature extraction module from the network.

We make use of overall mean localization error e as uniform evaluation metric in experiments. Suppose we have N test cases, ground truth locations of which are $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$. And the estimation position of each corresponding test case is denoted by $\hat{\mathbf{x}}_n (1 \leq n \leq N)$. Then the overall mean localization error e is defined as following:

$$e = \frac{1}{N} \sum_{n=1}^{N} ||\hat{\mathbf{x}}_n - \mathbf{x}_n||_2, \tag{24}$$

where $|| \cdot ||_2$ denotes L_2 norm.

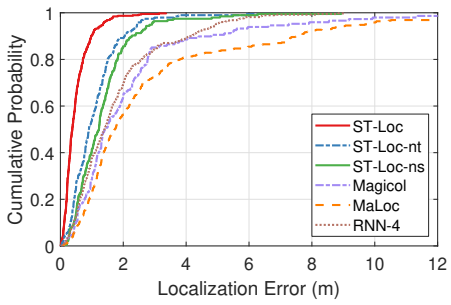


Fig. 14: Cumulative distribution function of indoor localization error in the office area.

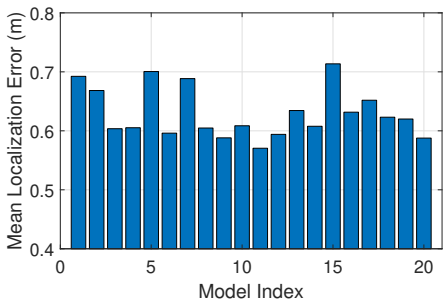


Fig. 15: Mean localization error versus models trained with different initial parameters in the office area.

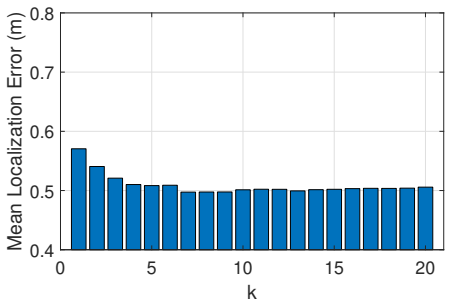


Fig. 16: Mean localization error versus different k in weighted k-NN based location estimation in the office area.

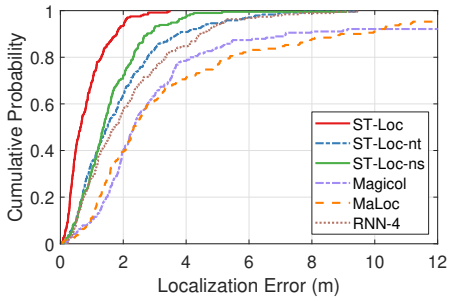


Fig. 17: Cumulative distribution function of indoor localization error in the food plaza.

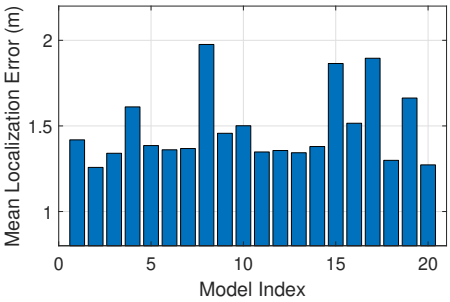


Fig. 18: Mean localization error versus models trained with different initial parameters in the food plaza.

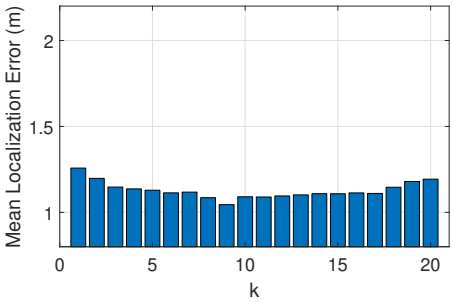


Fig. 19: Mean localization error versus different k in weighted k-NN based location estimation in the food plaza.

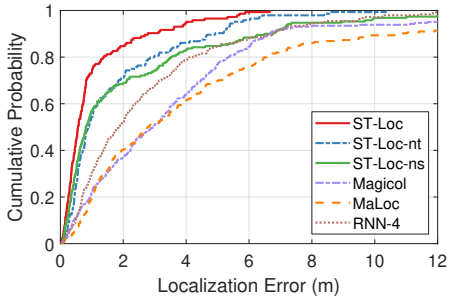


Fig. 20: Cumulative distribution function of indoor localization error in the supermarket area.

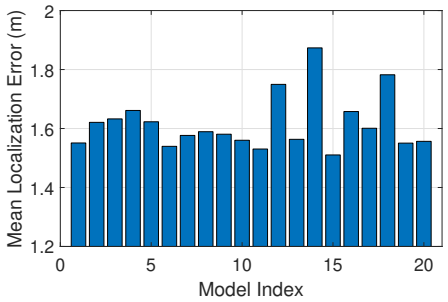


Fig. 21: Mean localization error versus models trained with different initial parameters in the supermarket area.

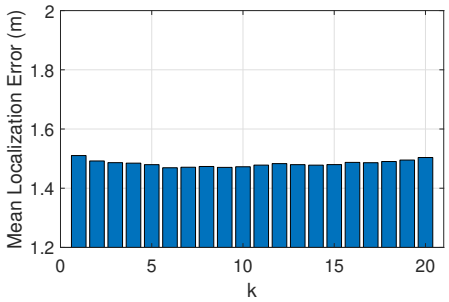
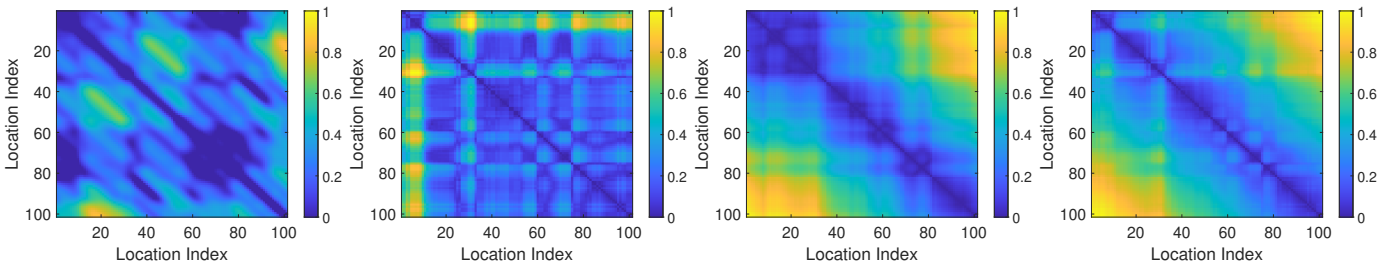


Fig. 22: Mean localization error with different k in weighted k-NN based location estimation in the supermarket area.

C. Experimental results

To evaluate the performance of ST-Loc and state-of-the-art competing approaches, we conduct extensive experiments in three typical trial sites. Fig. 14, Fig. 17 and Fig. 20 illustrate the CDF of localization errors in office area, food plaza and supermarket area respectively. And Table IV presents mean localization error of different approaches (including preliminary version [34] of ST-Loc) in three trial sites. The results demonstrate that proposed ST-Loc is able to achieve higher localization accuracy than competing schemes in all three trial sites. This is mainly because ST-Loc converts original geomagnetic sequence into spatial and temporal representations and considers both corresponding location correlations. Based on this, ST-Loc extracts and generates more comprehensive and distinctive spatial-temporal fusion features for localization, thus is able to achieve higher overall accuracy. Meanwhile, ST-Loc does not make use of the motion sensors of mobile devices, so there's no need to consider the impact of complicated user's behaviors on the motion sensors, thus avoiding accumulative error.

However, comparing the localization results in three sites which are shown in Fig. 14, Fig. 17 and Fig. 20, we find that ST-Loc performs better in office area. This is because office area has many narrow corridors and partitions, and a narrow indoor environment like that will cause strong signal variations locally, which provides much promise for more accurate localization. Meanwhile, we notice that the result in supermarket (as shown in Fig. 20) has long tails compared with the results in other two trial sites. It is because that the supermarket area is not only more spacious, but its indoor



(a) Raw geomagnetic signal. (b) Extracted spatial feature. (c) Extracted temporal feature. (d) Fusion feature.

Fig. 23: Signal/Feature differences between the locations which are uniformly selected in an office area.

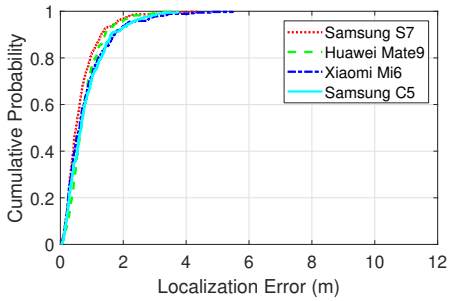


Fig. 24: CDF of localization error with different devices in office area.

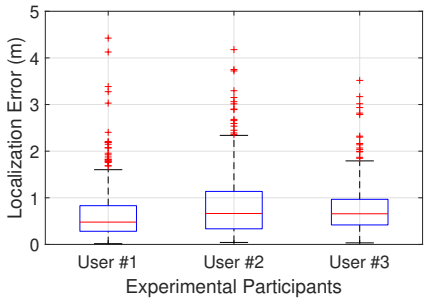


Fig. 25: The distribution of localization errors with different users in office area.

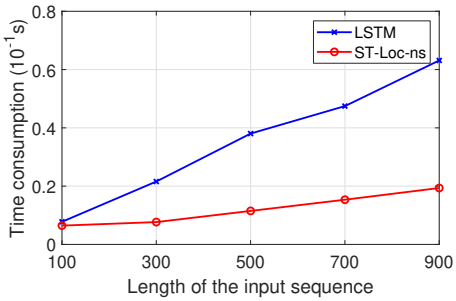


Fig. 26: Time cost for temporal feature extraction with different sequence length.

environment is highly similar as shown in Fig. 13 (c). For that reason, it more likely has similar geomagnetic disturbances in different local areas, which incurs signal ambiguity. Thus, the localization error in some cases is larger compared to constrained area like office environment.

To further enhancing the accuracy and robustness of ST-Loc, we apply ensemble learning mechanism in proposed network. As shown in Fig. 15, Fig. 18 and Fig. 21, we trained 20 models with different initial parameters for each trial site respectively as we discussed in Section IV-D. For a single model, different initial parameters lead to distinguishing results, which mainly suffers from input signal noise and outliers. Therefore, we take multiple models' prediction result into consideration. Furthermore, We devise weighted k-NN based voting among those models, and different values of k lead to different results in three trial sites as illustrating in Fig. 16, Fig.19 and Fig. 22, which mainly depends on corresponding indoor environment. Looked from the overall, the experimental results above demonstrate that ST-Loc is able to achieve higher accurate and more robust localization results with ensemble learning mechanism compared to preliminary version of ST-Loc. However, we notice that the diversity between learners is smaller with the increase in individual learner, and ensemble learning accuracy is worse, which mainly suffers from the limitation of dataset [41]. Therefore, we can choose a optimal value k for each specific trial site according to experimental validation results. For example, optimal k is set to 9 for the food plaza according to results shown in Fig.19.

TABLE IV: Mean localization error versus different approach in three trial sites (m).

Approach \ Trial site	Office Area	Food Plaza	Supermarket
ST-Loc	**0.49**	**1.04**	**1.45**
Preliminary ST-Loc	0.65	1.26	1.55
RNN-4	1.81	2.19	2.73
Magicol	2.19	5.64	3.76
MaLoc	3.34	5.05	4.56

In order to evaluate the effectiveness of extracted features and feature fusion, we collect geomagnetic measurements at 101 positions which are uniformly distributed in the office area. And we take advantage of pairwise matrix to evaluate the differences of raw signal or extracted features between these 101 selected location. The results are illustrated in Fig. 23 respectively, in which dark color indicates low degree of difference and light color indicates high degree of difference. And the average values of the difference pairwise matrix for raw signal, extracted spatial feature, extracted temporal feature and fusion feature are $0.18, 0.29, 0.38, 0.43$, respectively. As shown in Fig. 23 (a), raw geomagnetic signal at some distant locations can be similar (around location 60 to 80), leading to large localization error in the approach based on signal fingerprint matching. In ST-Loc, we convert raw geomagnetic signal into spatial and temporal representations and extract the corresponding features respectively. And as shown in Fig. 23 (b) and (c), the degree of difference between extracted spatial/temporal features in these locations has improved significantly, compared with the result in Fig. 23 (a). On the other hand, comparing Fig. 23 (b) and (c), we can find that spatial features imply more regional differences between the locations and temporal features reflect more consecutive differences between the locations, which confirms our previous analysis in

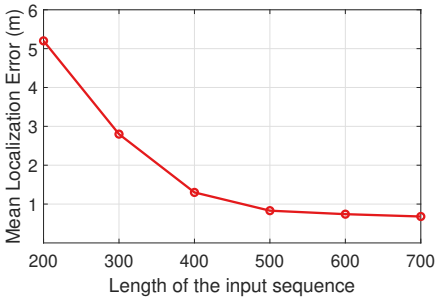


Fig. 27: Mean localization error with different length of input magnetic sequence.

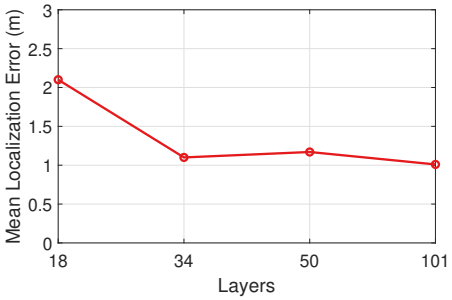


Fig. 28: Mean localization error with different layers of ResNet in *ST-Loc-nt*.

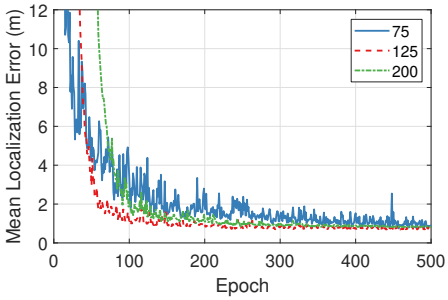


Fig. 29: Mean localization error with different mini-batch sizes during training.

Section IV. Finally, as presented in Fig. 23 (d), we are able to further increase the distinctiveness of features and enlarge the differences between distant locations by fusing the spatial and temporal features together (many dark areas are turned into light in the figure after feature fusion). So the results prove that our framework is able to enhance the distinctiveness of features for localization.

Fig. 24 illustrates CDF of localization error when using different devices for localization in office area. As discussed in Section IV-A (1), different devices usually have different calibrations for magnitude of geomagnetic field intensity (presented in Fig. 4). However, the experimental result shows that ST-Loc is able to achieve high localization accuracy even with those different devices. It demonstrates that ST-Loc is able to effectively solve the device heterogeneity problem. The reason is that ST-Loc uses *gradient sequence* as input instead of raw geomagnetic sequence, using the fluctuation trend of signal sequence as location clues to avoid the extra effort to calibrate different device or sensors to a uniform standard.

Fig. 25 shows the distribution of the localization errors with different experimental participants (include both male and female, height from 163 cm to 181 cm). And the results show that proposed ST-Loc achieves almost comparable localization accuracy even with different users while the mean localization errors are all less than 1 m. This is because ST-Loc conducts data pre-processing for collected signal sequences before localization to resolve the user heterogeneity problem, which has been elaborated in detail in Section IV-A (1). Thus, ST-Loc is able to achieve high applicability with different users.

Fig. 26 presents the time consumption for temporal feature extraction of *ST-Loc-ns* and the basic LSTM model. It demonstrates that proposed framework (illustrated in Fig. 11) achieves lower time consumption in temporal feature extraction, compared with the basic LSTM model. Meanwhile, as we use longer input geomagnetic sequence, the time consumption of the LSTM model increases while the time cost of *ST-Loc-ns* increases more slowly. This is because we devise a hierarchical structure and employ a segmentation and parallel processing mechanism. Therefore, the time overhead with the segmented subsequences is relatively lower compared with the original long sequence. On the other hand, parallel processing of the shorter subsequences also reduces the loss of feature information that easily occurs in serial processing of the long sequence.

Fig. 27 shows the localization accuracy when use different number of geomagnetic readings (the length of input geomagnetic sequence). As we can see, ST-Loc is able to achieve higher accuracy with more geomagnetic readings. The reason is that more signal readings (longer signal sequence) usually cover longer path with more local unique signal fluctuation. Hence the neural network is able to learn more location clues from those unique fluctuation. However, when the number of geomagnetic readings exceeds 500, the decrease of error slows down, which means that we have sufficient information to extract location clues with 500 geomagnetic readings. But more readings or longer sequences mean that it will take more time to collect and calculate for localization. Therefore, to achieve trade-off between time overhead and accuracy, we take the 500 signal readings as input (the length of input geomagnetic sequence is set to 500).

Fig. 28 presents the mean localization error when making use of different depths of RseNet in *ST-Loc-nt*. The result shows that the overall localization error decreases when use deeper ResNet. This is mainly because deeper network which have more layers is more capable to learn a robust feature from the input. However, we notice that the average localization error increases slightly when the number of the network layers reaches 50, which indicates that the network may have overfitting. At the same time, deeper network also requires more time for training to convergence. Therefore, to achieve trade-off between training time, training effort and accuracy, we take advantage of ResNet-34 in proposed model.

Fig. 29 illustrates the change process of localization error during training when applying different mini-batch sizes. As shown in the figure, the error decreases quickly in the first 100 epochs, then the decrease slows down. Finally model training converges after 500 epochs. Applying small mini-batch size means that ST-Loc will run more iterations in each epoch. So the model learns to adapt training data via more forward and backward propagations in each epoch, thus achieving smaller localization error in initial epochs. But small mini-batch size also leads to less local training data in each iteration, causing more fluctuations during training consequently. On the contrary, the number of iterations in each epoch is fewer when employing larger mini-batch size, leading to larger localization error in initial epochs but with fewer fluctuations. Therefore, to achieve trade-off, we set mini-batch size to 125 and the total number of epochs is 500 in our experiments.

D. System overhead

To evaluate the system overhead of ST-Loc, we have implemented it under client-server mode. The client is developed on Android system. And in experiments, the signal sampling frequency of the client is set to 50 Hz, which means that 50 signal samples (less than 2 KB) are sent to server every second. Correspondingly, the server will apply pre-trained localization model to estimate the current position after receive sufficient signal data from the client, then send back the results.

We use a 100 Mbps Wi-Fi router to provide the network connection, via which the average network transmission time is less than 0.0033s in experiments. Based on the above settings, we evaluate the localization responding time of ST-Loc and state-of-the-art competing approaches. Specifically, we take more than 1000 random chosen test cases to calculate the the average responding time as evaluation indicator. And the results are shown in Table V, which demonstrates that ST-Loc outperforms competing approaches with only 0.036s average responding time and it's enough to achieve real-time localization services.

TABLE V: Time overhead versus localization approaches.

Approach	ST-Loc	RNN-4	MaLoc	Magicol
Average responding time (s)	0.036	0.061	0.232	1.357

In term of energy consumption, we use the application power consumption records of the operating system as evaluation criterion. In experiments, we record the system power consumption of the client, which is most related to users. After 30 minutes of the simulated localization with ST-Loc, we record 6% drop in the battery status of test phone (Samsung Galaxy S7e with battery capacity of 3600 mAh). So the total energy consumption is around 216 mAh and the average power consumption per minute is 7.2 mAh. Therefore, using the client implemented based on ST-Loc will not consume too much power for location query. On the other hand, the current version of the client application has not yet been optimized well for energy efficiency and the signal sampling frequency is relatively high. So we can reduce the energy consumption by reducing the signal sampling frequency when collected signal data sufficiently meet actual needs of localization.

VI. Conclusion

Fingerprint-based indoor localization with either spatial or temporal location clues are prone to signal ambiguities or high time overhead, which hinders its widespread application. In view of the above, we propose to convert a single signal sequence into multiple adaptive representations, and extract features from each representation to form distinctive location features for accurate localization. In this paper, we use geomagnetic sequence as inuput. Firstly, we convert sequential geomagnetic inputs into a heatmap, where we use convolutional operations to find spatial correlations. At same time, we devise a hierarchical bidirectional structure to extract temporal correlations with both past and future context, achieving lower time overhead. Then, we fuse the spatial and temporal features together to enhance the distinctiveness of location features. Finally, we employ ensemble learning mechanism and design a weighted k-NN based location estimation algorithm to further enhance the accuracy and robustness. We have conducted extensive experiments in three different trial sites, the fifth floor of a narrow office building, the third floor of a mall and the second floor of a supermarket. Experimental results in these sites show that our model reduces localization error by a wide margin and achieves lower time overhead compared with other state-of-the-art competing schemes.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant 61972433 and 61772567, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2021A1515012242.

References

[1] K. Wen, K. Yu, Y. Li, S. Zhang, and W. Zhang, "A new quaternion kalman filter based foot-mounted IMU and UWB tightly-coupled method for indoor pedestrian navigation," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 4, pp. 4340–4352, 2020.

[2] Y. Zhuang, Q. Wang, M. Shi, P. Cao, L. Qi, and J. Yang, "Low-power centimeter-level localization for indoor mobile robots based on ensemble kalman smoother using received signal strength," *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 6513–6522, 2019.

[3] T. Kulshrestha, D. Saxena, R. Niyogi, and J. Cao, "Real-time crowd monitoring using seamless indoor-outdoor localization," *IEEE Transactions on Mobile Computing*, vol. 19, no. 3, pp. 664–679, 2019.

[4] X. Liu, Y. Jiang, P. Jain, and K.-H. Kim, "TAR: Enabling fine-grained targeted advertising in retail stores," in *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. Association for Computing Machinery, 2018, pp. 323–336.

[5] P. Chen, J. Shang, and F. Gu, "Learning RSSI feature via ranking model for Wi-Fi fingerprinting localization," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 2, pp. 1695–1705, 2020.

[6] M. Zhou, Y. Wang, Y. Liu, and Z. Tian, "An information-theoretic view of WLAN localization error bound in GPS-denied environment," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 4089–4093, 2019.

[7] S. Tomažič, D. Dovžan, and I. Škrjanc, "Confidence-interval-fuzzy-model-based indoor localization," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 3, pp. 2015–2024, 2019.

[8] J. Dong, M. Noreikis, Y. Xiao, and A. Ylä-Jääski, "ViNav: A vision-based indoor navigation system for smartphones," *IEEE Transactions on Mobile Computing*, vol. 18, no. 6, pp. 1461–1475, 2019.

[9] D. Liu, S. Guo, Y. Yang, Y. Shi, and M. Chen, "Geomagnetism-based indoor navigation by offloading strategy in NB-IoT," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4074–4084, 2019.

[10] H. Yuan, J. Wang, Z. Zhao, J. Cui, M. Yan, and S. Wei, "MagWi: Practical indoor localization with smartphone magnetic and WiFi sensors," in *2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS)*. IEEE, 2019, pp. 814–821.

[11] S. He and K. G. Shin, "Geomagnetism for smartphone-based indoor localization: Challenges, advances, and comparisons," *ACM Computing Surveys (CSUR)*, vol. 50, no. 6, pp. 1–37, 2017.

[12] Q. Niu, T. He, N. Liu, S. He, X. Luo, and F. Zhou, "MAIL: Multi-scale attention-guided indoor localization using geomagnetic sequences," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 2, pp. 1–23, 2020.

[13] C. Kumar and K. Rajawat, "Dictionary-based statistical fingerprinting for indoor localization," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 9, pp. 8827–8841, 2019.

[14] Y. Zou, G. Wang, K. Wu, and L. M. Ni, "Smartscanner: Know more in walls with your smartphone!" *IEEE Transactions on Mobile Computing*, vol. 15, no. 11, pp. 2865–2877, 2015.

[15] Y. Shu, K. G. Shin, T. He, and J. Chen, "Last-mile navigation using smartphones," in *Proceedings of the 21st annual international conference on mobile computing and networking*. ACM, 2015, pp. 512–524.

[16] M. Kwak, Y. Park, J. Kim, J. Han, and T. Kwon, "An energy-efficient and lightweight indoor localization system for Internet-of-Things (IoT) environments," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 1–28, 2018.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, 2016, pp. 770–778.

[18] R. C. Luo and T.-J. Hsiao, "Indoor localization system based on hybrid Wi-Fi/BLE and hierarchical topological fingerprinting approach," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 11, pp. 10 791–10 806, 2019.

[19] C. Gleason, D. Ahmetovic, S. Savage, C. Toxtli, C. Posthuma, C. Asakawa, K. M. Kitani, and J. P. Bigham, "Crowdsourcing the installation and maintenance of indoor localization infrastructure to support blind navigation," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 1–25, 2018.

[20] X. Liu, X. Wei, and L. Guo, "DIMLOC: Enabling high-precision visible light localization under dimmable LEDs in smart buildings," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 3912–3924, 2019.

[21] Q. Niu, M. Li, S. He, C. Gao, S.-H. Gary Chan, and X. Luo, "Resource-efficient and automated image-based indoor localization," *ACM Transactions on Sensor Networks (TOSN)*, vol. 15, no. 2, p. 19, 2019.

[22] M. Li, N. Liu, Q. Niu, C. Liu, S.-H. G. Chan, and C. Gao, "SweepLoc: Automatic video-based indoor localization by camera sweeping," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, p. 120, 2018.

[23] Y. Zhao, C. Qian, L. Gong, Z. Li, and Y. Liu, "LMDD: Light-weight magnetic-based door detection with your smartphone," in *2015 44th International Conference on Parallel Processing*. IEEE, 2015, pp. 919–928.

[24] H. Abdelnasser, R. Mohamed, A. Elgohary, M. F. Alzantot, H. Wang, S. Sen, R. R. Choudhury, and M. Youssef, "SemanticSLAM: Using environment landmarks for unsupervised indoor localization," *IEEE Transactions on Mobile Computing*, vol. 15, no. 7, pp. 1770–1782, 2016.

[25] Z. Zhao, J. Wang, X. Zhao, C. Peng, Q. Guo, and B. Wu, "NaviLight: Indoor localization and navigation under arbitrary lights," in *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE, 2017, pp. 1–9.

[26] Y. Zheng, G. Shen, L. Li, C. Zhao, M. Li, and F. Zhao, "Travi-navi: Self-deployable indoor navigation system," *IEEE/ACM transactions on networking*, vol. 25, no. 5, pp. 2655–2669, 2017.

[27] H. Xie, T. Gu, X. Tao, H. Ye, and J. Lu, "A reliability-augmented particle filter for magnetic fingerprinting based indoor localization on smartphone," *IEEE Transactions on Mobile Computing*, vol. 15, no. 8, pp. 1877–1892, 2016.

[28] F. Gu, J. Niu, and L. Duan, "WAIPO: A fusion-based collaborative indoor localization system on smartphones," *IEEE/ACM Transactions on Networking*, vol. 25, no. 4, pp. 2267–2280, 2017.

[29] Y. Shu, C. Bo, G. Shen, C. Zhao, L. Li, and F. Zhao, "Magicol: Indoor localization using pervasive magnetic field and opportunistic WiFi sensing," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 7, pp. 1443–1457, 2015.

[30] X. Sun, C. Wu, X. Gao, and G. Y. Li, "Fingerprint-based localization for massive MIMO-OFDM system with deep convolutional neural networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 11, pp. 10 846–10 857, 2019.

[31] X. Wang, Z. Yu, and S. Mao, "Indoor localization using smartphone magnetic and light sensors: A deep LSTM approach," *Mobile Networks and Applications*, pp. 1–14, 2019.

[32] H. J. Bae and L. Choi, "Large-scale indoor positioning using geomagnetic field with deep neural networks," in *2019 IEEE International Conference on Communications (ICC)*. IEEE, 2019, pp. 1–6.

[33] H. J. Jang, J. M. Shin, and L. Choi, "Geomagnetic field based indoor localization using recurrent neural networks," in *GLOBECOM 2017-2017 IEEE Global Communications Conference*. IEEE, 2017, pp. 1–6.

[34] T. He, Q. Niu, S. He, and N. Liu, "Indoor localization with spatial and temporal representations of signal sequences," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–7.

[35] P. Flandrin, G. Rilling, and P. Goncalves, "Empirical mode decomposition as a filter bank," *IEEE signal processing letters*, vol. 11, no. 2, pp. 112–114, 2004.

[36] J.-g. Park, A. Patel, D. Curtis, S. Teller, and J. Ledlie, "Online pose classification and walking speed estimation using handheld devices," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 2012, pp. 113–122.

[37] C.-L. Zhang, J.-H. Luo, X.-S. Wei, and J. Wu, "In defense of fully connected layers in visual representation transfer," in *Pacific Rim Conference on Multimedia*. Springer, 2017, pp. 807–817.

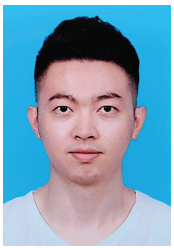
[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 248–255.

[39] A. L. Blum and R. L. Rivest, "Training a 3-node neural network is NP-complete," *Neural Networks*, vol. 5, no. 1, pp. 117–127, 1992.

[40] G. Wang, X. Wang, J. Nie, and L. Lin, "Magnetic-based indoor localization using smartphone via a fusion algorithm," *IEEE Sensors Journal*, vol. 19, no. 15, pp. 6477–6485, 2019.

[41] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018.

Ning Liu received the B.S. degree in computational mathematics from Northwest University, Xian, China, and the PhD degree in computer science from Sun Yat-sen University (SYSU), Guangzhou, China, in 1996 and 2004, respectively. He is currently a professor with School of Computer Science and Engineering, Sun Yat-sen University. He has served as a reviewer for several important conferences and journals. His current research interests include computer vision, indoor localization, and machine learning algorithms.



Tao He received the B.S. and M.S. degree in computer science and technology from Sun Yat-sen University, Guangzhou, China, in 2017 and 2020, respectively. He is currently working toward the PhD degree with School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China. He is a student member of the IEEE. His research interests mainly lie in indoor positioning and navigation algorithms and location-based data analysis.

Suining He is currently working as an assistant professor at the Department of Computer Science and Engineering, the University of Connecticut. He received his PhD degree in Computer Science and Engineering at the Hong Kong University of Science and Technology, and worked as a postdoctoral research fellow at the Real-Time Computing Lab (RTCL), the University of Michigan, Ann Arbor. His research interests include ubiquitous and mobile computing, crowdsourcing and big data analytics.

Qun Niu received the BEng and MEng degree from the School of Software, Sun Yat-sen University in 2013 and 2015, respectively. He received the PhD degree at the School of Data and Computer Science, Sun Yat-sen University in 2019. He is currently working as a postdoctoral research fellow at the School of Computer Science and Engineering, Sun Yat-sen University. His research interest includes indoor spatial sensing, Internet of Things (IoT) and mobile computing. He is a member of the IEEE.