Indoor Localization with Spatial and Temporal Representations of Signal Sequences

Tao He[1], Qun Niu[1], Suining He[2], Ning Liu[1],*
[1] *Sun Yat-sen University,* [2] *The Hong Kong University of Science and Technology*
{hetao23, niuqun}@mail2.sysu.edu.cn, sheaa@cse.ust.hk, liuning2@mail.sysu.edu.cn

***Abstract*—Indoor localization has attracted considerable attention lately, due to its large commercial and social values in smart cities. The existing indoor localization approaches mostly rely on fingerprint techniques, and many of those leverage either spatially discrete fingerprints or temporally consecutive ones for localization, which either suffers from large errors due to signal ambiguities or high time overhead with long sequences.**

To achieve high accuracy with low computational cost, we propose *ST-Loc*, a deep neural network that extracts features from multiple *representations* of a single signal sequence for localization, where each representation indicates a corresponding signal structure with underlying feature correlations. Taking geomagnetism as an example, we infer location features from two different representations, e.g., spatial and temporal. In spatial representation, a signal sequence is converted to a signal heatmap, where each pixel corresponds to a spatial location and the value indicates fingerprint. Temporal representation, on the other hand, is a signal sequence with ordered readings, which provides temporal correlations. Using these different representations, we employ convolutional and recurrent networks to extract location features and fuse them to generate more distinguishing features for localization. We have conducted extensive experiments in two different trial sites, a narrow office area and a spacious food plaza. Our experimental results show that ST-Loc achieves more than 43% average localization error reduction compared with state-of-the-art competing schemes in both trial sites.

I. Introduction

Indoor localization plays a fundamental role in a wide range of indoor location-based services, e.g., pedestrian localization [1], targeted advertising [2], and crowd monitoring [3], to name a few. The quality of these services, however, largely relies on the accuracy of underlying positioning algorithms.

To achieve sufficient accuracy, researchers have studied the indoor localization extensively. Of all techniques, fingerprint-based ones have attracted much attention. Recent fingerprint-based techniques are broadly divided into two categories: *spatial* based and *temporal* based approaches [4]. In the first category, the spatial clues refer to discrete measurements of input values at different locations (e.g., an image, a Wi-Fi/Bluetooth fingerprint at a location) [5]. Using these discrete inputs, state-of-the-art approaches compare them with a database and infer current location with the most similar geo-tagged fingerprint. These discrete inputs-based approaches are prone to feature ambiguity and noise, which may lead to large localization errors.

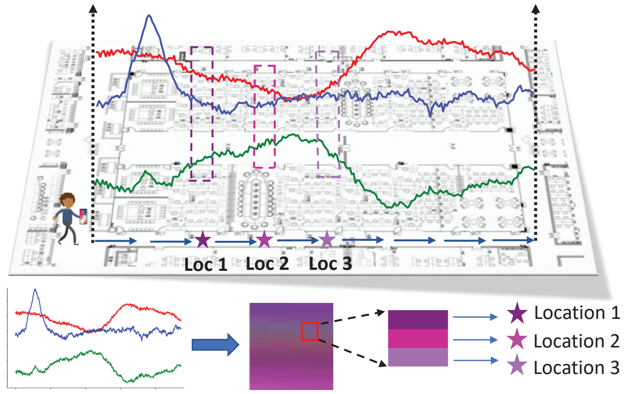


Fig. 1: We convert geomagnetic sequences to heatmaps (matrices), where each patch indicates spatially distributed readings.

In order to enhance the localization accuracy, other approaches begin to study temporal information, which indicates successive measurements of signal readings at a series of positions, e.g., a short video clip or a geomagnetic sequence [6], [7]. These approaches advance previous spatial input-based ones by considering *temporal correlations* between consecutive inputs. By incorporating these temporal correlations and continuity constraints, they reduce the impact of noisy inputs and smooth out erroneous estimations, thus achieving higher accuracy. However, the computational complexity usually increases with longer input signal sequences [8]. To achieve balance between computation cost and accuracy, one leverages *short* sequences for localization (e.g., a few frames of videos and a short geomagnetic sequence), leading to degraded distinctiveness of location features due a limited spatial coverage and large localization errors consequently.

To address above challenges, we propose a deep neural network that considers both spatial and temporal representations of signal sequences for indoor localization, termed ***ST-Loc***. We convert a single signal sequence to different structures, e.g., heatmaps (spatial representation) or sequences (temporal representations). Afterwards, ST-Loc extracts *spatial* and *temporal* location features and fuse them together to achieve high accuracy. We make the following contributions:

- *Converting signal to different representations for localization:* To facilitate distinguishing feature extraction from signal, we propose a network to convert a single signal sequence to different representations, e.g., spatial

[*] Corresponding author.

and temporal ones. Then, based on the dimension of corresponding representations, we use convolutional and recurrent networks to extract features and fuse them together to localize.

- *Inferring spatial features via visual approach*: Moreover, we convert a sequence of input into a heatmap. Then, we employ a modified ResNet [9], applying convolution to different patches of this heatmap, which correspond to spatially distributed samples at regular intervals (Fig. 1). Using convolution operations, we infer spatial location features from these readings that span a long range.
- *Extracting temporal features with recurrent models*: We design hierarchical bidirectional long short-term memory (LSTM) [10] model to capture the temporal correlations of ordered signal sequence. With hierarchical structure, we reduce the average computational complexity of LSTM units in the model. And we further enhance the extracted temporal features by a bidirectional LSTM scheme which considers both past and future contextual information in the sequence.

As an example, we evaluate the localization accuracy of the proposed network with geomagnetic sequences. We have conducted extensive experiments in two different trial sites: a narrow office area and a spacious food plaza. Evaluation results show that the proposed network reduces the localization error by more than 43% compared with competing approaches. In addition to geomagnetism, it is possible to adapt ST-Loc to other signal sequences, such as Wi-Fi [11], [12], Bluetooth [13] or visible light [14] sequences, for localization.

The remainder of the paper is structured as follows. We review existing approaches that are most similar to ours in Section II. Then, we elaborate our network design in Section III. We present illustrative experimental results in Section IV and conclude in Section V.

II. Related Work

We review related work as follows. Considering spatial features of localization signals, some researchers evaluate the measurement of the signals at different locations and use this pattern to pinpoint users. For example, SemanticSLAM [15] clusters geomagnetic signals and discovers landmarks to calibrate current location. Although efficient, these approaches utilize discrete readings, which lacks dimensionality and still insufficient for large-scale indoor localization. Recent researches [16], [17] adapt particle filter mechanism to localize with fingerprints. Furthermore, WAIPO [18] and Magicol [19] fuse other signals (images, Wi-Fi) to enhance the localization accuracy. Despite accuracy in specific sites, signal ambiguities indoors may lead to degraded distinctiveness of features and large localization errors consequently.

Recently, some researchers propose to leverage sequential measurements of signals as input (vectorize multiple successive observations) to enhance localization accuracy with temporal correlations. Travi-Navi [20] and NaviLight [21] both leverage dynamic time warping (DTW) for localization, which considers both stretching and squeezing sequences to align them. However, the comparison of two sequences is computationally expensive and may lead to high computational cost. And some approaches employ neural networks to process sequential inputs for localization [22]. The work in [23] proposes to use a basic recurrent neural network (RNN) unit to localize with geomagnetic sequence, while it's still hard to extract sufficient features with only a simple basic RNN especially in wide open space.

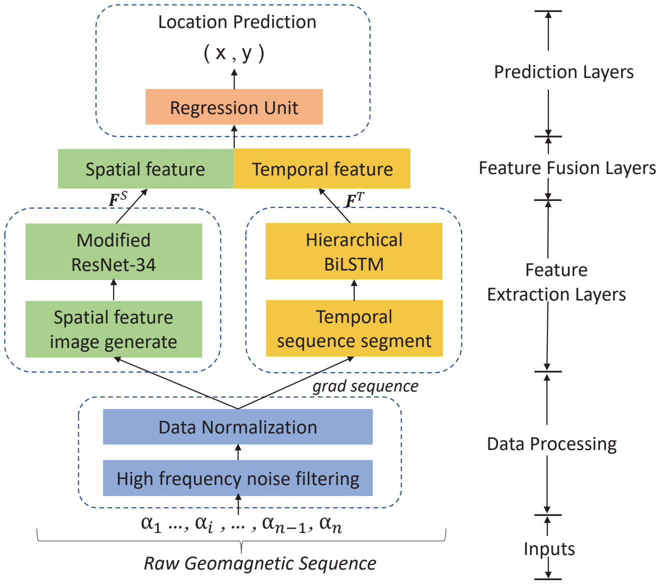


Fig. 2: Overall framework of ST-Loc.

III. Design of ST-Loc

In this section, we present the design of proposed ST-Loc. We elaborate the overall structure of the network in Section III-A, followed by the elaboration of spatial and temporal feature extraction in Section III-B and Section III-C, respectively.

A. Overall Structure of ST-Loc

The overall framework of proposed ST-Loc is shown in Fig. 2, which consists of four main modules: 1) Data preprocessing; 2) Multi-scale spatial feature extraction; 3) Hierarchical temporal feature extraction and 4) Feature fusion and location prediction. We overview each module as follows:

1) *Data preprocessing*. In this part, we first employ empirical mode decomposition (EMD) [24] technique to filter high frequency noise caused by user motion. And for device heterogeneity (different devices may have different calibrations for magnitude of geomagnetic field intensity), we calculate the gradient of raw sequence as input instead of using raw data directly, knowing that the distortions of geomagnetic sequences collected by different devices at same location are the same. Through above data preprocessing operations, we can effectively reduce the impact of external noise.

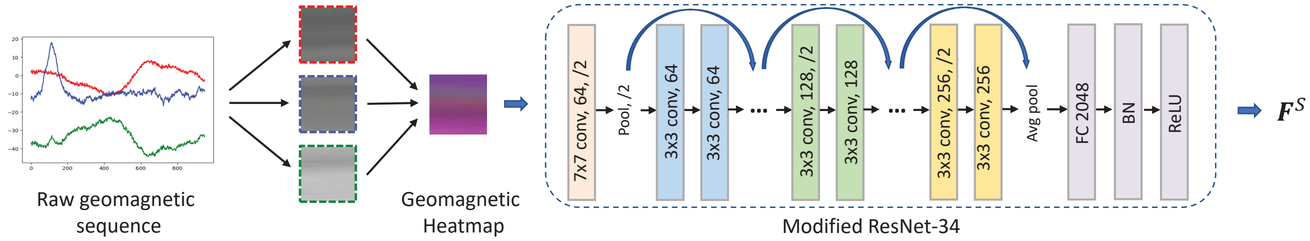


Fig. 3: Spatial representation and feature extraction.

2) *Multi-scale spatial feature extraction*. Noticing the spatial correlations, we consider the geomagnetic sequence from computer vision angle and propose a multi-scale spatial feature extraction (MSFE) model. Instead of processing low dimensional sequence directly, we convert the sequence to a geomagnetic heatmap, a spatial representation, where each pixel corresponds to a spatial location and the value denotes a signal reading. Then, we employ a modified ResNet [9] to extract multi-scale spatial features from resulted heatmaps. Details of MSFE will be discussed in Section III-B.

3) *Hierarchical temporal feature extraction*. Considering the temporal correlations, we employ start-of-the-art LSTM model. In practice, it is resource and time consuming to extracted temporal features from a long input sequence by feeding the sequence to a LSTM model directly. Therefore, we design a hierarchical structure, utilizing sequence segmenting scheme and multiple-level LSTM to extracted temporal features. Furthermore, considering both past and future contextual information at a timestep, we apply a bidirectional structure in LSTM (BiLSTM) to further enhance the extracted temporal features. Details of the hierarchical BiLSTM scheme are illustrated in Section III-C.

4) *Feature fusion and Location prediction*. Finally, we concatenate the extracted spatial and temporal features. Then, based on the fused spatial-temporal features which are more comprehensive and distinguishing, we predict user's location with a regression unit which is mainly composed of fully connected layers (FC layers) and non-linear mapping functions.

B. Multi-Scale Spatial Feature Extraction

1) Spatial representation of the signal: Considering from a computer vision angle, we propose a spatial representation of the geomagnetic sequence by converting the sequence to a heatmap. Each pixel denotes a single geomagnetic observation. As shown in Fig. 1, we assume a small window (denoted by red block) in the geomagnetic heatmap, the rows of which are actually sub-portions in original geomagnetic sequence at regular interval and correspond to spatially distributed locations. Applying convolution to different patch of the heatmap, we can extract the features that reflect the spatial correlations of original geomagnetic sequence. As shown in Fig. 3, a single geomagnetic observation collected by device consists of values on three axes of X, Y and Z. For a geomagnetic sequence, we first reshape it to a three-dimensional rectangular matrix (shape of *width***height**3), then normalize all elements of the rectangular matrix to RGB color space $[0, 255]$ and convert it to a 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 a single geomagnetic observation on three axes.

2) Modified residual neural network: As ResNet has achieved superior accuracy due to its residual structure in various vision tasks, e.g., image identification, we leverage ResNet to extract spatial features from geomagnetic heatmap. However, original ResNet primarily processes natural image, which is fundamentally different from our geomagnetic heatmaps in resolution, pixel densities and spatial correlations of nearby pixels. According to research in [25], to achieve high accuracy in transfer learning models, it is essential that having the FC layers in the source domain pre-trained model when task objective or image properties in the source domain are far different from those in target domain. Therefore, we use ResNet (pre-trained on ImageNet [26]) as a basis, then add FC layers, normalization layer and activate functions after removing the final classification layer. Finally, we fine-tune the modified ResNet by training it with our geomagnetic heatmaps.

The modified ResNet can extract lower level features (signals in small window) in the previous stage of ResNet while higher level features (signals in much larger window) in the latter stage of the network. In summary, ResNet is able to extract multi-scale spatial features, including low-level and high-level features which corresponding to short-range and long-range signal fluctuations, respectively.

More specifically, as presented in Fig. 3, we first remove final classification layers of original ResNet. In this case, the network outputs a 512-D spatial feature vector $\mathbf{f}^S$. Then we insert a 2048-D FC layer that maps original 512-D feature to higher dimensional vector [25], followed by a batch normalization layer. In the meantime, we insert a rectified linear unit (ReLU) as activate function. The equations are as follow:

$$\tilde{\mathbf{f}}^S = W\mathbf{f}^S + b, \tag{1}$$

$$\mathbf{f}^S_{norm} = \frac{\gamma}{\sqrt{Var[\tilde{\mathbf{f}}^S] + \epsilon}} \cdot \tilde{\mathbf{f}}^S + (\beta - \frac{\gamma E[\tilde{\mathbf{f}}^S]}{Var[\tilde{\mathbf{f}}^S] + \epsilon}), \tag{2}$$

$$\mathbf{F}^S = v \cdot ReLU(\mathbf{f}^S_{norm}), \tag{3}$$

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

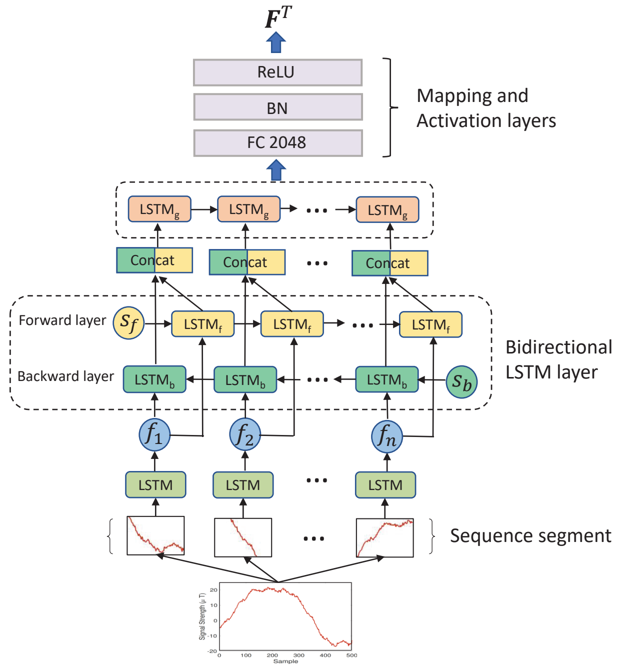


Fig. 4: Temporal representation and feature extraction.

Finally, we obtain multi-scale spatial feature $\mathbf{F}^S$ extracted from the heatmap with fine-tune modified ResNet.

C. Hierarchical Temporal Feature Extraction

1) Temporal representation of the signal: In many fingerprint-based approaches, the location prediction is entirely independently based on each fingerprint (a single observation). However, geomagnetic observations collected by device are actually a geomagnetic-stream with temporal continuity, which mean that a lot of feature information can be extracted by employing temporal dependencies. Intuitively, we propose to make use of ordered geomagnetic readings (temporal representation) as input of the network.

2) Hierarchical bidirectional LSTM: To capture these temporal dependencies, we take advantage of the LSTM model in our network. The LSTM makes an improvement on standard RNN, overcoming the vanishing gradient problem. And LSTM applies following operations at each timestep:

$$f_t = \sigma_g(W_f \mathbf{x}_t + U_f \mathbf{h}_{t-1} + b_f), \tag{4}$$

$$i_t = \sigma_g(W_i \mathbf{x}_t + U_i \mathbf{h}_{t-1} + b_i), \tag{5}$$

$$o_t = \sigma_g(W_o \mathbf{x}_t + U_o \mathbf{h}_{t-1} + b_o), \tag{6}$$

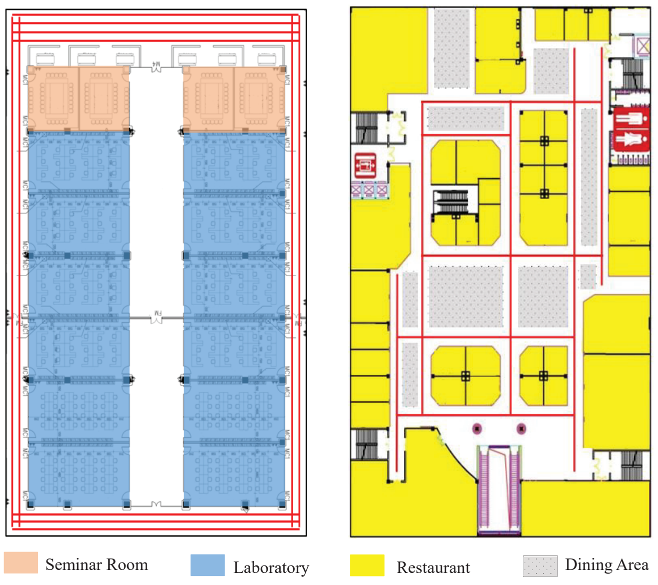
$$\tilde{c}_t = \sigma_c(W_c \mathbf{x}_t + U_c \mathbf{h}_{t-1} + b_c), \tag{7}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t, \tag{8}$$

$$\mathbf{h}_t = o_t \circ \sigma_h(c_t), \tag{9}$$

$$\mathbf{y}_t = \sigma_o(W_y \mathbf{h}_t + b_y), \tag{10}$$

where $\mathbf{x}_t$ and $\mathbf{h}_t$ denote the input and hidden state at time t. W, U and b are the learnable parameters, σ is the non-linear activation function. And f, i, o denote *forget gate*, input and output *reset gates* respectively, and c is a *memory cell state*.



(a) Office area.

(b) Food plaza.

Fig. 5: Floorplans of our test sites.

However, in practice, it is still hard to efficiently correlate from the first to last input for a long input sequence by feeding the sequence to a LSTM model directly, which is also resource and time consuming for a long input sequence. Therefore, we design a hierarchical LSTM structure, as shown in Fig. 4, we first segment the geomagnetic sequence with specific scale and obtain subsequence set $\mathbf{s} : \{s_1, s_2, ..., s_n\}$. Then we extract the temporal features of these local subsequences with low-level LSTM respectively. For a local subsequence s_i, we obtain corresponding local temporal feature $\mathbf{f}_i : \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_m\}$ (m is the length of input subsequence s_i), which will be taken as input of a high-level LSTM. With this hierarchical structure, each LSTM unit in the network processes the shorter subsequence, which reduces the average computational complexity of the network.

Furthermore, we employ a bidirectional LSTM scheme to enhance these local temporal features $\{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_n\}$. As shown in Fig. 4, BiLSTM takes this ordered feature sequence as input, making use of both past and future contextual information for each instance in the sequence. Then we obtain the enhanced local temporal features $\{\tilde{\mathbf{f}}_1, \tilde{\mathbf{f}}_2, ..., \tilde{\mathbf{f}}_n\}$:

$$\tilde{\mathbf{f}}_i = BiLSTM([\mathbf{h}_i^f, \mathbf{h}_{n-i}^b], \mathbf{f}_i), \tag{11}$$

where $\mathbf{h}^f$ and $\mathbf{h}^b$ represent the forward and backward hidden states, respectively.

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

IV. Illustrative Experimental Results

We present detailed experimental settings and comparison schemes in Section IV-A. Then we illustrative experimental results in Section IV-B, followed by the overhead analysis in Section IV-C.

A. Experimental Settings and Comparison Schemes

1) Dataset and training settings: We conduct experiments in two typical trial sites, a narrow office area in our university and a more spacious food court in a mall (shown in Fig. 5). The narrow office area covers around 2,800 m^2 and the food plaza is more spacious which covers around 3,500 m^2. To construct datasets, we develop an Android application to collect signals including geomagnetic signal strength and IMU sensor data (Inertial Measurement Unit). While surveyors walk though the survey path, the application will record various signals along path, then we get the signal value sequence corresponding to the path: $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, ...\}$ where $\mathbf{v}_i = \{\mathbf{m}_i, \mathbf{a}_i, \mathbf{g}_i, \mathbf{o}_i\}$, and $\mathbf{m}_i = \{m_x, m_y, m_z\}$ indicates geomagnetic signal strength in three axes. $\mathbf{a}_i, \mathbf{g}_i, \mathbf{o}_i$ denote corresponding acceleration vector, gyroscope angles and orientation angles respectively. The sampling frequency of signals is 50 Hz and the length of collected sequence is 500.

For training dataset, we designed dense survey paths (denoted by red solid lines in Fig. 5) in public areas of these sites and we have collected 2,390 signal sequences in the office area and 1,952 signal sequences in food plaza for training. For testing dataset, volunteers are asked to walk though some randomly chosen paths in trial sites, and we collected 770 and 482 signal sequences in two trial sites for evaluating, respectively.

We train proposed ST-Loc separately with collected geomagnetic sequences in each trial site and evaluate its performance with test ones in corresponding site, respectively. Baseline training parameters in our experiments are shown in Table I. We choose *Pytorch* as deep learning framework in experiments and employ *Adam* as network's optimizer and the loss function is *MSELoss*. For biases and weights in each layer of the network, we initialize them with a standard Gaussian distribution. All experiments are performed on Ubuntu 16.04 server with two Nvidia 1080ti GPU cards, an Intel i7-6700 CPU and 48GB memory.

2) Comparison schemes and evaluating metric: We compare proposed approach with the following state-of-the-art geomagnetic localization methods:

- MaLoc [17] utilizes an enhanced particle filter to estimate user's position based on fingerprint comparison. Then it proposes an adaptive sampling algorithm to reduce the number of particles and increase tracking efficiency.
- Magicol [19] vectorizes collected sequential geomagnetism based on user's steps and employs DTW to compare them with a database to infer current location. It calibrates user traces with an enhanced bi-directional particle filter.
- *RNN-4* [23]. Jang et al. train a standard RNN network to predict the position of user, using geomagnetic sequence as input. In our experiment, we build a 4-layer standard RNN network as a comparison scheme.

TABLE I: Baseline parameters in experiments

Parameters	Food Plaza	Office Area
Sequence Length	500	500
Iterations	500	500
Initial Learning Rate	0.0001	0.0001
Mini-batch	125	125

In addition, to evaluate the effectiveness of each network component, we also compare with following model's variants:

- *ST-Loc-ns*: There is no spatial feature extraction module in the network, by which we can validate the effectiveness of spatial features.
- *ST-Loc-nt*: We remove the temporal feature extraction module to validate the effectiveness of temporal features.

We use the overall mean localization error e as evaluation metric. Suppose we have N trial cases, where ground truth location corresponding to each one is $\mathbf{x}_n$ while estimated position is $\hat{\mathbf{x}}_n$. Then the overall mean localization error e is determined as:

$$e = \frac{1}{N} \sum_{n=1}^{N} ||\hat{\mathbf{x}}_n - \mathbf{x}_n||_2, \tag{12}$$

where $|| \cdot ||_2$ is an L_2 norm.

B. Experimental Results

We compare the performance of ST-Loc with state-of-the-art competing approaches. Fig. 6 illustrates the CDF of localization errors in office area. It demonstrates that proposed ST-Loc is able to achieve higher accuracy than competing schemes. This is because ST-Loc considers both spatial and temporal correlations of inputs and extracts more comprehensive, distinguishing spatial-temporal features of original geomagnetic sequence, thus is able to achieve higher overall accuracy. Meanwhile, we do not employ noisy motion sensors of smartphones to localize, thus reducing the impact of complicated user's behaviors and random noise on motion sensors.

Fig. 7 shows the localization error in the food plaza. ST-Loc is also able to achieve sufficient localization accuracy. However, Fig. 7 has long tails compared with the results in office area. This is because the food plaza is more spacious and has fewer local disturbances, which incurs signal ambiguity. Thus, the localization error in some cases is larger compared to constrained office environment.

Fig. 8 presents CDF of location error with different devices in office area. It shows that ST-Loc achieves high localization accuracy with different devices which have different calibrations for magnitude of geomagnetic field intensity. It demonstrates that ST-Loc could effectively handle the device heterogeneity problem. The reason is that ST-Loc takes *gradient sequence* as input instead of raw data.

Fig. 9 evaluates the mean localization error with different layers of ResNet in *ST-Loc-nt*. It shows that the overall localization error decreases with more layers (or deeper network).

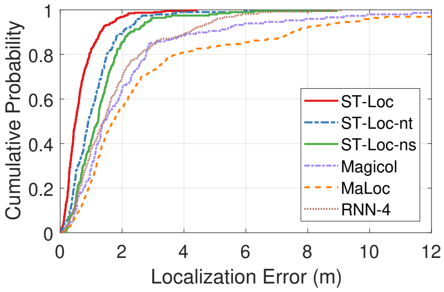


Fig. 6: CDF of location error in the office area.

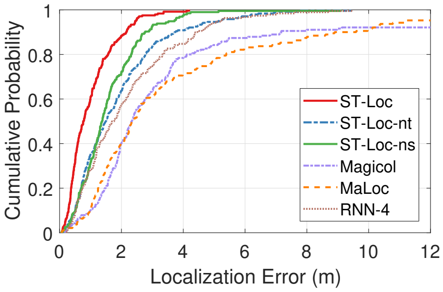


Fig. 7: CDF of location error in food plaza.

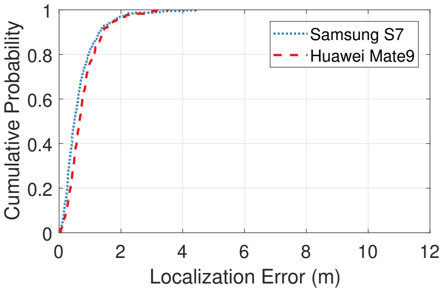


Fig. 8: CDF of location error of ST-Loc with different devices in office area.

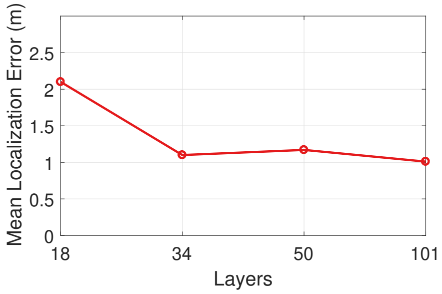


Fig. 9: Mean localization error with different layers of ResNet in *ST-Loc-nt*.

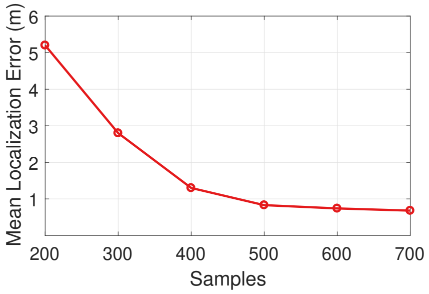


Fig. 10: Mean localization error with different numbers of geomagnetic samples.

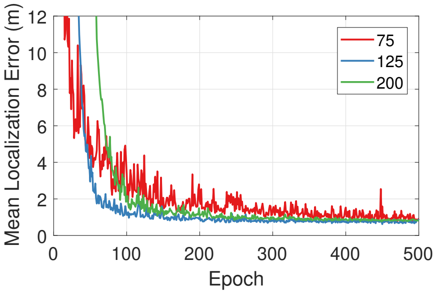


Fig. 11: Mean localization error with different mini-batch sizes during training.

This is because deeper network is more capable of learning a robust feature from geomagnetic heatmap. To achieve trade-off between time, training effort and accuracy, we use ResNet-34 in our experiment.

Fig. 10 illustrates the localization accuracy with number of geomagnetic readings. It shows that ST-Loc is able to achieve higher accuracy with more readings. This is because more readings cover longer path with more local unique disturbances. Thus the neural network is able to learn more location clues from input. As the number of samples is larger than 500, the decrease slows down. This is because we have sufficient information with 500 samples. However, more samples mean that it takes more time to collect and calculate. To achieve trade-off between localization accuracy and system response time, we use the ordered past 500 readings as input.

Fig. 11 shows the changes of localization error during training process with different mini-batch sizes. It shows that the error decreases quickly in the first 100 epochs. Then the decrease slows down. Finally it converges after 500 epochs. Meanwhile, with small mini-batch size, ST-Loc runs more iterations in each epoch. Therefore, it learns to adapt our training data through more forward and backward propagations, thus achieving smaller localization error in initial epochs but with more fluctuations during training. However, with larger mini-batch size, the number of iterations is fewer which means fewer propagations, leading to larger localization error in initial epochs but with fewer fluctuations. In our experiment, we train our network with 500 epochs and set mini-batch size to 125, thus achieving trade-off.

C. System Overhead

The ST-Loc works in client-server mode. In our experiment, the client collects and sends 50 signal samples (less than 2KB) to server every second, and the server sends back localization results after analyzing the signals. The average network transmission time is less than 0.0033s via a 100Mbps Wi-Fi router, and we evaluate the average responding time of ST-Loc and competing schemes for location predicting or calculating with more than 1000 test cases. As shown in Table II, ST-Loc outperforms competing schemes and ST-Loc is able to achieve real-time service with only 0.038s average responding time.

TABLE II: The average responding time for localization.

Approach	ST-Loc	RNN-4	MaLoc	Magicol
Responding time (s)	0.038	0.061	0.238	1.791

For energy consumption, we recorded the measurements of system power consumption, where all computations happened on the experiment device. The current version of our client application has not yet been optimized well for energy efficiency. After the simulation localization around 30 minutes, we notice 6% drop in the battery life of our test phone. The total energy consumption of the client is 243 mAh, due to high sampling frequency. We can reduce the energy consumption by reducing the sampling frequency of signal when localization info sufficiently meet actual needs.

V. Conclusion

Indoor localization with either spatial or temporal clues are prone to signal ambiguities or high time overhead, which hinders its wide deployment. To address above, we propose to convert a single signal sequence to different representations, and then extract features from each representation to form distinctive location features. More specifically, we convert sequential signal inputs to a heatmap, where we use convolutional operations to find spatial correlations of inputs. In the meantime, we use hierarchical bidirectional LSTM to extract temporal correlations with both past and future context. Then, we fuse these spatial and temporal features together to enhance the distinctiveness of features. We have conducted extensive experiments in two different trial sites, the fifth floor of a narrow office building and the third floor of a mall. Experimental results in these sites show that our model reduces localization error by more than 43% compared with other state-of-the-art competing schemes.

Acknowledgment

This work was supported, in part, by Guangxi Special Fund Project for Innovation-Driven Development (Grant No. AA18118039), the Fundamental Research Funds for the Central Universities, Guangzhou Science Technology and Innovation Commission (GZSTI16EG14/201704030079) and Opening Project of Guangdong Province Key Laboratory of Information Security Technology (Grant No.2017B030314131).

References

[1] T. Li, Y. Chen, R. Zhang, Y. Zhang, and T. Hedgpeth, "Secure crowdsourced indoor positioning systems," in *Proc. IEEE INFOCOM*, April 2018, pp. 1034–1042.

[2] X. Liu, Y. Jiang, P. Jain, and K.-H. Kim, "TAR: Enabling fine-grained targeted advertising in retail stores," in *Proc. ACM MobiSys*, 2018, pp. 323–336.

[3] H. Hong, G. D. De Silva, and M. C. Chan, "CrowdProbe: Non-invasive crowd monitoring with Wi-Fi probe," *Proc. ACM IMWUT.*, vol. 2, no. 3, pp. 115:1–115:23, Sep. 2018.

[4] S. He and K. G. Shin, "Geomagnetism for smartphone-based indoor localization: Challenges, advances, and comparisons," *ACM CSUR*, vol. 50, no. 6, pp. 97:1 – 97:37, 2018.

[5] M. Li, N. Liu, Q. Niu, C. Liu, S.-H. G. Chan, and C. Gao, "Sweeploc: Automatic video-based indoor localization by camera sweeping," *Proc. ACM IMWUT.*, vol. 2, no. 3, pp. 120:1 – 120:25, 2018.

[6] H. Wu, S. He, and S. G. Chan, "A graphical model approach for efficient geomagnetism-pedometer indoor localization," in *Proc. IEEE MASS*, 2017, pp. 371–379.

[7] Q. Niu, N. Liu, J. Huang, Y. Luo, S. He, T. He, S.-H. G. Chan, and X. Luo, "DeepNavi: A deep signal-fusion framework for accurate and applicable indoor navigation," *ACM IMWUT*, 2019, to appear.

[8] 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," *Proc. ACM IMWUT.*, vol. 2, no. 1, pp. 17:1–17:28, Mar. 2018.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, June 2016, pp. 770–778.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] Z. Yin, C. Wu, Z. Yang, and Y. Liu, "Peer-to-peer indoor navigation using smartphones," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1141–1153, May 2017.

[12] Q. Niu, Y. Nie, S. He, N. Liu, and X. Luo, "RecNet: A convolutional network for efficient radiomap reconstruction," in *Proc. IEEE ICC*, 2018, pp. 1–7.

[13] 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," *Proc. ACM IMWUT.*, vol. 2, no. 1, pp. 9:1–9:25, Mar. 2018.

[14] S. Shao, A. Khreishah, and I. K. and, "RETRO: Retrorefelctor based visible light indoor localization for real-time tracking of IoT devices," in *Proc. IEEE INFOCOM*, April 2018, pp. 1–9.

[15] 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 Trans. Mobile Comput.*, vol. 15, no. 7, pp. 1770–1782, July 2016.

[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," *Proc. ACM IMWUT.*, vol. 2, no. 1, pp. 17:1–17:28, Mar. 2018.

[17] 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 Trans. Mobile Comput.*, vol. 15, no. 8, pp. 1877–1892, Aug 2016.

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

[19] 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 J. Sel. Areas Commun.*, vol. 33, no. 7, pp. 1443–1457, July 2015.

[20] Y. Zheng, G. Shen, L. Li, C. Zhao, M. Li, and F. Zhao, "Travi-Navi: Self-deployable indoor navigation system," *IEEE/ACM Trans. Netw.*, vol. 25, no. 5, Oct 2017.

[21] Z. Zhao, J. Wang, X. Zhao, C. Peng, Q. Guo, and B. Wu, "NaviLight: Indoor localization and navigation under arbitrary lights," in *Proc. IEEE INFOCOM*, May 2017, pp. 1–9.

[22] X. Wang, Z. Yu, , and S. Mao, "DeepML: Deep LSTM for indoor localization with smartphone magnetic and light sensors," in *Proc. IEEE ICC*, May 2018, pp. 1–6.

[23] H. J. Jang, J. M. Shin, and L. Choi, "Geomagnetic field based indoor localization using recurrent neural networks," in *Proc. IEEE GLOBECOM*, Dec 2017, pp. 1–6.

[24] P. Flandrin, G. Rilling, and P. Goncalves, "Empirical mode decomposition as a filter bank," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 112–114, Feb 2004.

[25] C.-L. Zhang, J.-H. Luo, X.-S. Wei, and J. Wu, "In defense of fully connected layers in visual representation transfer," in *Proc. Springer PCM*, 2017, pp. 807–817.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE CVPR*, 2009, pp. 248–255.