Towards Task-Adaptive Indoor Localization via Progressive and Lightweight Learning

Haobing Guo[a], Wanxin Zeng[b], Kungui Zeng[a], Xiushuang Yi[a]*
[a] School of Computer Science and Engineering, Northeastern University, Shenyang, China
[b] College of Computer Science and Technology, Jilin University, Changchun, China
xsyi@mail.neu.edu.cn
*Corresponding Author

***Abstract*—With the development of Internet of Things(IoT) technology and the widespread use of mobile terminal devices, indoor localization technology has gained significant attention in areas such as smart healthcare and smart buildings. Although the localization method that uses WiFi Received Signal Strength (RSS) fingerprinting offers good deployability and cost advantages, the accuracy of indoor localization is impacted by the variability over time, high noise levels, and uneven distribution of RSS measurements in indoor environments. In recent years, deep learning methods have been widely used in indoor localization tasks. Although they have enhanced localization accuracy to some degree, they have struggled with effectively capturing local features and modeling global dependencies. To address these issues, this paper introduces PL-Loc, a lightweight and progressive indoor localization model designed to efficiently represent fingerprint features from local to global levels. The model consists of a Feature Encoder and a Two-Step Capture Module, which is based on the Criss-Cross Attention. The Feature Encoder utilizes Convolutional Neural Network (CNN) to capture local features of fingerprints, while the Two-Step Capture Module effectively models the global dependencies of fingerprints without significantly increasing computational demands. Additionally, for classification and regression localization tasks in various scenarios, PL-Loc designs a Task-Adaptive, scalable output module. It also introduces a Temperature-Scaled Softmax strategy in classification localization tasks to better calibrate the model's output distribution. Experimental results on two public datasets, UJIIndoorLoc and Tampere, show that PL-Loc has achieved excellent performance in different indoor localization scenarios. Notably, on the Tampere dataset, the Mean Localization Error (MLE) reaches 7.51 m, and the localization accuracy surpasses that of the current leading models.**

***Index Terms*—WiFi Fingerprinting, Indoor Localization, Convolutional Neural Network, Criss-Cross Attention**

I. Introduction

In recent years, the rapid development of IoT technology and the widespread use of personal mobile devices have made location-based services (LBS) increasingly important in various applications, such as smart buildings and smart healthcare. However, in indoor environments, the obstruction of wireless signals by the building's structures has led to low localization accuracy with satellite-based global localization systems (GNSS). This has made it challenging to provide continuous, stable, and reliable localization services indoors [1]. Therefore, there has been a growing interest in exploring high-precision and easy-to-deploy indoor localization methods.

Currently, various wireless signal-based technologies, including ultra-wideband (UWB) [2], Bluetooth [3], Zigbee [4], RFID [5], etc., are being explored to develop effective indoor localization methods. Among these, the indoor localization method that utilizes WiFi RSS fingerprints relies on the deployed WiFi network infrastructure (such as access points, AP). This approach offers significant advantages in terms of deployment cost and availability, as it doesn't require additional measurement equipment. The indoor localization method based on WiFi RSS fingerprint consists of two main stages: the offline stage and the online stage. In the offline stage, a location fingerprint database is constructed by collecting the RSS of multiple APs at different measurement locations (reference points, RP). In the online stage, when a user collects the RSS from multiple APs at a target location, the RSS vector is constructed for the target location. This vector is matched with the fingerprint stored in the database to find the best match, which estimates the user's location information [6].

However, in complex indoor environments, due to the influence of building occlusion, pedestrian movement, and multipath effects, etc., the RSS collected at different locations has the characteristics of time-varying and high levels of noise [7]. Additionally, RSS is often unevenly distributed indoors, leading to significant regional imbalances in the collected RSS fingerprints. To address these problems, many studies have adopted deep learning methods to establish the mapping relationship between RSS fingerprints and spatial locations, resulting in notable progress. Nevertheless, CNN, which is commonly used in existing studies, is mainly good at extracting local fingerprint features but has struggled to capture the global feature relationship in fingerprint data effectively. The emergence of the self-attention mechanism has addressed some challenges in handling fingerprint data. However, the computational overhead has been significant when dealing with large-scale fingerprint data, which has made deployment on resource-constrained terminal devices difficult. To tackle these issues, this paper proposes a task-adaptive, progressive, and lightweight indoor localization model, PL-Loc, which is designed for multiple tasks in indoor localization. PL-Loc progressively learns representations of fingerprint features,

This work was supported by the National Natural Science Foundation of China(No.62032013, No.62132004), and the National Key Research and Development Program of China(No.2022YFB4500800).

moving from local to global levels. It effectively encodes local features in fingerprints while efficiently capturing global dependencies in fingerprint features without significantly increasing computational demands. Specifically, we introduced a Feature Encoder to encode local features in fingerprints and designed a Two-Step Capture Module based on Criss-Cross Attention. Through cross-path calculation, efficient fingerprint global dependency modeling was achieved, allowing for high-precision indoor localization with low resource consumption. At the same time, PL-Loc designs a task-adaptive output module for localization tasks in different indoor scenarios and employs a Temperature-Scaled Softmax strategy in classification tasks to produce a sharper output distribution. The contributions of this paper are as follows:

1. This paper proposes a **p**rogressive **l**ightweight indoor **loc**alization model—**PL-Loc**. The model encodes the local features of fingerprints through a Feature Encoder and efficiently captures the global dependencies through a Two-Step Capture Module, maintaining low computational demand while ensuring localization accuracy. To the best of our knowledge, this is the first instance of employing the Criss-Cross path calculation method to capture the global dependencies in WiFi-based indoor localization methods.

2. We design a Two-Step Capture Module to model global dependencies in fingerprint data cost-effectively. This mechanism significantly enhances the model's ability to perceive global patterns in fingerprint data by aggregating feature information through multi-head Criss-Cross paths in both horizontal and vertical dimensions.

3. We design a task-adaptive output module for various localization tasks in different indoor environments, and introduce the Temperature-Scaled Softmax strategy in the classified indoor localization task, which improves the output accuracy of the model in the classified localization task by adjusting the temperature.

4. Experimental results on the well-known indoor localization datasets UJIIndoorLoc [8] and Tampere [9] indicate that compared with mainstream indoor localization models, PL-Loc achieves superior indoor localization accuracy, achieving 100% building prediction accuracy on UJIIndoorLoc and 7.51 m MLE on Tampere.

The remainder of the paper is organized as follows: Section II introduces related work, Section III outlines the overall structure and details of PL-Loc, Section IV evaluates our model, and Section V concludes the paper.

II. Related Work

Indoor localization methods can be broadly categorized into two groups: device-dependent and device-free localization methods. Device-dependent localization methods can be further divided into geometric localization methods and Inertial Measurement Unit (IMU) localization methods [10]. Geometric localization relies on specific measuring equipment to gather information, such as Time of Arrival (TOA) [11], Angle of Arival(AOA) [12], Time Difference of Arrival(TDOA) [13], etc., to calculate distance information to achieve indoor localization, but it requires precise time and phase synchronization, and can be expensive to implement in practice. The IMU localization method uses inertial sensors such as gyroscopes, accelerometers, and other devices to estimate the movement direction and distance of the localization target. It calculates location information based on the known initial location. This method relies on accurate knowledge of the initial location and the movement trajectory, which limits its practical applications [14]. Device-free localization methods don't require additional measurement equipment, allowing users to obtain all necessary information for localization from their own devices. Among these methods, fingerprint-based localization is the most popular. This approach involves creating location fingerprints through the obtained Channel State Information(CSI) or RSS, and establishing the mapping relationship between fingerprints and locations for localization. While CSI-based fingerprinting requires specific Network Interface Card(NIC) support, which can limit its practical applications, our focus is on indoor localization using RSS fingerprints [15].

The rapid development of deep learning has significantly advanced RSS fingerprint indoor localization methods. In indoor environments, factors such as crowd flow can interfere with the propagation of WiFi signals, leading to substantial noise in the collected fingerprint information. To address this issue, the GNN [7] method introduces graph neural networks to model stable feature associations between adjacent access points and proposes an access point selection strategy to enhance the feature representation capability. To tackle the reduced localization accuracy caused by fingerprint sparsity, SALLoc [16] integrates a Stacked Auto-Encoder(SAE), attention mechanism, and LSTM structure to improve the model's modeling ability for temporal fingerprint features. CNNLoc [17] enhances localization accuracy by combining 1D-CNN with SAE, particularly useful when the number of data samples is limited. The DNNBN [18] method introduces batch normalization(BN) in each layer of Multilayer Perceptron(MLP) to improve the stability and generalization ability of localization. Considering the differences in localization tasks in different scenarios, HADNN [19] employs hierarchical auxiliary information to enhance the model's scalability. Additionally, 2L-HELM [20] proposes a hierarchical extreme learning machine with dual labels to improve localization accuracy in complex multi-building and multi-floor environments. For devices with limited computing resources, CRSS [21] offers a lightweight CNN model combined with a preprocessing strategy, effectively reducing the consumption of computing resources while maintaining localization accuracy.

III. PL-Loc

A. Problem Formulation

The indoor localization method based on RSS fingerprinting consists of two stages: 1) offline stage, 2) online stage. In the offline stage, N measurement locations (referred to as reference points, or RPs) are selected within the indoor environment to collect the RSS values from M APs, and the RSS

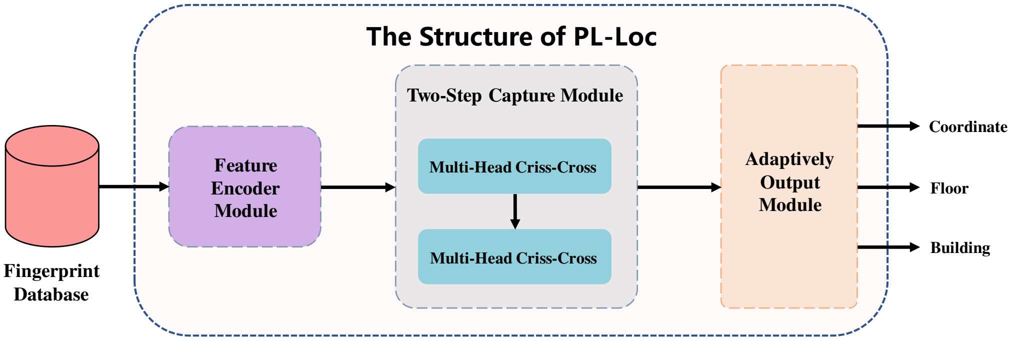

Fig. 1. The overall structure of PL-Loc. It first encodes the local features of fingerprints within the fingerprint database using a feature encoder, then the Two-Step Capture Module models the global dependencies among these features. Finally, the task-adaptive output model retrieves the localization information.

fingerprints $F_i, i = 1, 2, \cdots, N$ of different RPs are obtained. And F_i can be expressed as $F_i = [r_1^i, r_2^i, \cdots, r_j^i, \cdots, r_M^i]$, where r_j^i represents the RSS value of the j-th AP collected at the i-th reference point. In this paper, we set a minimum detection threshold of -110 dBm, meaning that any RSS values below -110 dBm are replaced with -110 dBm. The fingerprint database constructed in this stage comprises the location and fingerprint of the RP, defined as:

$$D = \{(F_1, L_1), \cdots, (F_i, L_i), \cdots, (F_N, L_N)\} \quad , \tag{1}$$

where L_i represents the location of the ith RP. For a multi-building and multi-floor environment, we define $L_i = [x_i, y_i, b_i, f_i]$, where x_i, y_i, b_i, f_i denote the horizontal and vertical coordinates, building, and floor of the ith RP, respectively. To ensure model learning stability, we normalize the horizontal and vertical coordinates of RP. The process is as follows:

$$x_i' = \frac{x_i - \mu_x}{\sigma_x}, \quad y_i' = \frac{y_i - \mu_y}{\sigma_y} \quad , \tag{2}$$

wherein, μ_x, μ_y represent the mean of all horizontal and vertical coordinates, respectively, while σ_x, σ_y denote the standard deviation of these coordinates. During the offline stage, the model needs to learn the mapping relationship between fingerprints and locations in the fingerprint database. Additionally, it should adjust its learning strategy according to the requirements of different indoor localization tasks. For coordinate prediction tasks, the Mean Squared Error(MSE) loss function is utilized, whereas for building and floor classification tasks, the Cross-Entropy loss function is employed. According to optimization theory, the optimization goal of this problem is to find the optimal parameter θ^*, which can be expressed using the following optimization formula:

$$arg \min_{\theta} \left(\alpha \cdot MSE(\hat{x}, \hat{y}, x, y) + \beta \cdot CE(\hat{b}, b) + \gamma \cdot CE(\hat{f}, f)\right), \tag{3}$$

wherein, MSE, CE represent the MSE loss function and the Cross-Entropy loss function, respectively. $\hat{x}, \hat{y}, \hat{b}, \hat{f}$ denote the predicted values of x, y, b, f respectively, while α, β, γ are the loss weights of different localization tasks.

In the online stage, the collected RSS vector at the target location is inputted into the model with optimal parameters to obtain the corresponding location information.

B. Overall Structure

The overall architecture of the model is illustrated in Fig. 1. It consists of three main components: the Feature Encoder Module, the Two-Step Capture Module, and the Task-Adaptive Output Module. Initially, the Feature Encoder Module encodes the local features of the fingerprint data stored in the database. Following this, the Two-Step Capture Module effectively models the global dependencies among the local fingerprint features. Finally, the local features, along with their global dependency relationships, are processed and predicted by the Task-Adaptive Output Module for various indoor localization tasks.

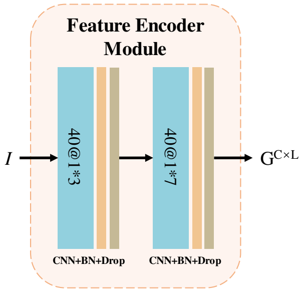

Fig. 2. The Structure of Feature Encoder. It consists of two convolutional layers with kernel sizes of 1*3 and 1*7, respectively. Additionally, BN and Dropout operations are applied after each convolutional layer.

C. Feature Encoder Module

Due to the complex WiFi signal propagation conditions in indoor environments, the distribution of RSS is often unbalanced and discontinuous, which in turn causes the problem of missing fingerprint data. To address these issues and improve indoor localization accuracy, it is necessary to extract the feature encoding information from the fingerprint data accurately. To this end, we propose the Feature Encoder Module. As shown in Figure 2, it extracts the local features of the fingerprint through two sets of convolutional layers. Each convolutional layer employs BN and Dropout to enhance the training stability of the model and prevent overfitting. The specific calculation process is as follows:

$$\begin{aligned} C &= \text{Dropout0.3}(\text{BN}(\text{Conv40@13}(I)) \quad , \\ G &= \text{Dropout0.3}(\text{BN}(\text{Conv40@17}(C)) \quad , \end{aligned} \tag{4}$$

where $\text{Dropout}_{0.3}$ indicates that the dropout rate of the Dropout layer is 0.3. The notation $40@1*3, 40@1*7$ represents 40 convolution kernels of size $1*3$ and 40 convolution kernels of size $1*7$, respectively.

D. Two-Step Capture Module

Due to noise interference and the imbalanced spatial distribution of RSS in the fingerprint database, it is challenging to ensure that the model effectively captures global dependencies in the fingerprint by relying solely on the Feature Encoder Module to extract the local features of the fingerprint, thus affecting the localization accuracy. To fully explore the global dependencies in the fingerprint data, we introduce the Two-Step Capture Module, which captures the global dependencies in the fingerprint through Multi-Head Criss-Cross Attention.

Criss-Cross Attention effectively captures contextual information on both horizontal and vertical paths, enabling the integration of global dependency with low computational complexity. Inspired by [22], we propose Multi-Head Criss-Cross Attention. Specifically, for the input feature map $H \in \mathbb{R}^{C\times H\times W}$, it is mapped to $Q \in \mathbb{R}^{C'\times W\times H}$, $K \in \mathbb{R}^{C'\times W\times H}$ and $V \in \mathbb{R}^{C\times W\times H}$ after convolution operations, where $C' < C$, that is, the feature map is reduced in dimension in the channel dimension:

$$Q = W_Q \cdot H, \quad K = W_K \cdot H, \quad V = W_V \cdot H, \tag{5}$$

For the location u in the feature map, its Criss-cross path is defined as Ω_u, which includes all other points in the same row and column where the point is located, see Fig. 3. The attention score calculation process is as follows:

$$d_{i,u} = Q_u \cdot \Omega_{i,u}^T \ , i = 1, \ldots, H + W - 1 \quad , \tag{6}$$

Next, apply the softmax function to normalize the values, then get the attention map:

$$A_{i,u} = \frac{\exp(d_{i,u})}{\sum_j \exp(d_{j,u})} \quad , \tag{7}$$

Finally, the output feature representation of specific location u is obtained as follows:

$$H'_u = \sum_{i\in\Omega_u} A_{i,u} \cdot V_{i,u} \quad . \tag{8}$$

To further improve the expressiveness and modeling efficiency of the model, we incorporate multiple attention heads on this basis to capture the feature information from various subspaces in parallel. Specifically, the local fingerprint feature $G \in \mathbb{R}^{C\times L}$, which is extracted in the Feature Encoder Module, is divided into h feature subspaces along the channel dimension, that is,

$$G = [G_1, G_2, \ldots, G_h], \quad G_j \in \mathbb{R}^{\frac{C}{h}\times L} \quad , \tag{9}$$

each feature subspace G_j models the global dependency of the fingerprint features in parallel. First, the original Q, K, V are divided into multiple feature subspaces according to the number of heads:

$$Q_i = W_Q^{(i)} K_i, \quad K_i = W_K^{(i)} K_i, \quad V_i = W_V^{(i)} G_i \quad , \tag{10}$$

wherein, $W_Q^{(i)}, W_K^{(i)} \in \mathbb{R}^{C'\times \frac{C}{h}}$, $W_V^{(i)} \in \mathbb{R}^{C\times \frac{C}{h}}$. Subsequently, the global dependencies in local fingerprint features are modeled independently in each feature subspace,

$$O_i = \text{CCNet}(Q_i, K_i, V_i) \in \mathbb{R}^{\frac{C}{h}\times L} \quad , \tag{11}$$

$$GM = \text{Concat}(O_1, \ldots, O_h) \in \mathbb{R}^{C\times L} \quad . \tag{12}$$

The results from each subspace are concatenated in the channel dimension, enabling effective modeling of global fingerprint dependencies.

After obtaining the global dependencies from the fingerprint data, the local features of the fingerprint are fused with the global dependencies, which fully ensures that the model effectively captures the fingerprint features.

E. Task-Adaptive Output Module

For various indoor localization tasks, we propose a multi-task output module that can be adaptively expanded. This module dynamically adjusts its output branch according to different localization tasks, and each output branch utilizes a different output module, as illustrated in Fig. 4.

In the classification tasks like floor prediction and building prediction, we introduce the Temperature-Scaled Softmax strategy in the output branch. It can adjust the smoothness of the classification output by adjusting the temperature, and effectively alleviate the problem of overconfidence of the model. The Softmax function is defined as follows:

$$P_i = \frac{\exp(z_i)}{\sum_{j=1}^{C} \exp(z_j)} \quad , \tag{13}$$

when the temperature parameter T is introduced, it is adjusted to:

$$P_i^{(T)} = \frac{\exp(z_i/T)}{\sum_{j=1}^{C} \exp(z_j/T)} \quad . \tag{14}$$

When $T > 1$, the probability distribution of the Softmax output becomes flatter, enhancing the model's fault tolerance. In contrast, when $T < 1$, the probability distribution of the Softmax output becomes sharper, which helps to calibrate the output distribution of the model. In this paper, we choose T to be 0.15.

For regression tasks such as coordinate prediction in indoor localization, the fully connected layer is used in combination with Dropout for prediction. The Task-Adaptive Output

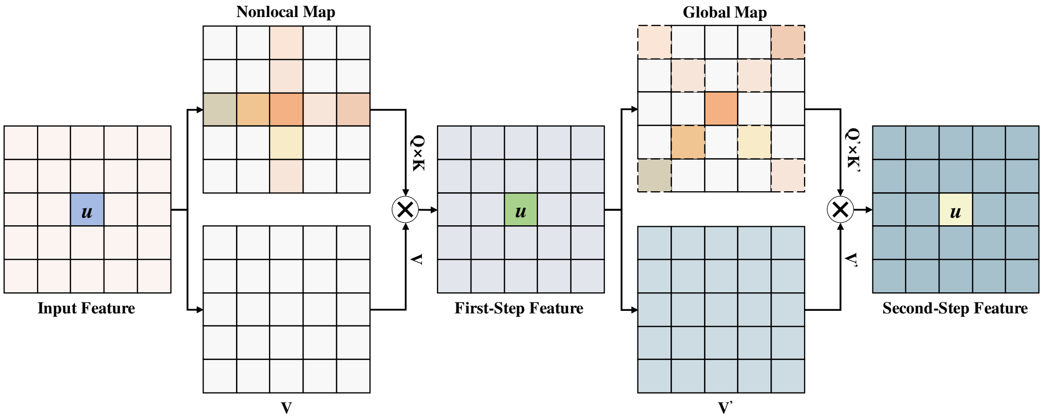

Fig. 3. The Two-step Criss-Cross Attention process. First, horizontal and vertical attention maps are calculated through the Criss-Cross path, allowing for the aggregation of non-local feature associations to form the first step feature map. Then, the attention map on the diagonal path is obtained, which is the dotted part in the Global Map, and finally, the global feature association relationship is obtained, which is the second step feature map.

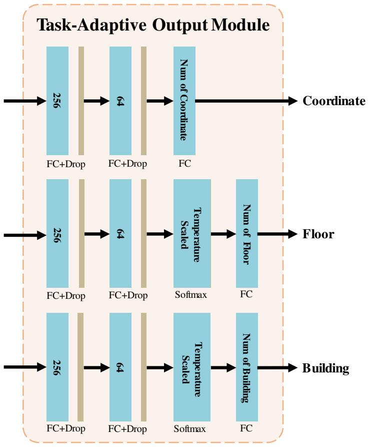

Fig. 4. The structure of Task-Adaptive Output Module.

Module can be easily expanded according to different indoor localization tasks and shows strong adaptability to different indoor environments.

IV. Experiments

A. Evaluation Metrics

The evaluation Metrics we use include accuracy (Acc) and mean localization error (MLE). Acc measures the accuracy of floor and building predictions, calculated using Eq. 15.

$$\mathrm{Acc} = \frac{N_{\text{correct}}}{N_{\text{total}}} \times 100\%, \tag{15}$$

where N_{correct} represents the number of samples for which the model correctly predicts the floor or building, and N_{total} represents the total number of test samples.

MLE measures the average Euclidean distance between predicted and actual coordinates of the model, calculated using Eq. 16:

$$\mathrm{MLE} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(x_i-\hat{x}_i)^2+(y_i-\hat{y}_i)^2}, \tag{16}$$

wherein, (x_i, y_i) and $(\hat{x}_i, \hat{y}_i)$ represent the actual coordinates and predicted coordinates of the ith RP, respectively. Additionally, N represents the total number of test samples.

B. Datasets and Settings

We choose the UJIIndoorLoc [8] and Tampere [9] datasets, both of which are widely utilized in indoor localization. The UJIIndoorLoc dataset includes three types of location information: coordinates, floors, and buildings. The Tampere dataset includes two types of location information: coordinates and floors. Detailed information is presented in Table I.

TABLE I
The details of datasets.

Number of	UJIIndoorLoc	Tampere
Total Data	21048	4648
Training Data	19937	697
Validation Data	1111	3951
AP	520	992
RP	933	4648
Buildings	3	1
Floors	4, 5	5

We train the model using the TensorFlow framework on NVIDIA GeForce RTX 4090 and evaluate its performance through experiments in the same experimental environment. For different datasets, we train the model for 100 epochs, using the Adam optimizer with a batch size of 256. We test the model performance with various hyperparameters, including learning rates of {0.001, 0.003, 0.01, 0.03, 0.05}, dropout of {0.1, 0.3, 0.5, 0.7}, convolution kernel of {(3,5), (3,7), (3,9), (5,7), (5,9)}, attention head of {2, 4, 8, 16}, softmax temperature of {0.1, 0.15, 0.3, 0.5, 0.75, 1} on two datasets. The best hyperparameter configuration is shown in Table II. It is important to note that building prediction tasks don't need to be performed on Tampere; therefore, the building prediction loss weight for Tampere is set to 0.

TABLE II
THE BEST SETTINGS.

Parameter	UJIIndoorLoc	Tampere
Epoch	100	100
Learning Rate	0.03	0.01
Optimizer	Adam	Adam
Drop Out	0.3	0.3
Batch Size	256	256
Loss Weight	1/3, 1/3, 1/3	1/3, 0, 1/3
Kernel	3, 7	3, 7
Head	4	4
Softmax Temperature	0.15	0.3

C. Experimental Results

We first compared the localization performance with the current mainstream indoor localization models using UJIIndoorLoc dataset, as summarized in Table III. In terms of floor prediction accuracy, the CNNLoc method achieved the highest at 96.03%, slightly exceeding PL-Loc's 95.23%, but performed poorly in the fine-grained coordinate prediction regression task, with MLE of only 11.78 m, and its building prediction accuracy was only 99.27%. Overall, PL-Loc outperformed the other models, especially in MLE, achieving 9.38 m. This was mainly since the Two-Step Capture Module efficiently modeled the global dependency relationship by capturing local fingerprint features through the UJIIndoorLoc dataset by the Feature Encoder Module. This approach effectively mined the correlation between fingerprint features and spatial locations, leading to the highest accuracy in the coordinate prediction regression task. Additionally, the temperature-scaled softmax was introduced on the floor and for building predictions, which enhanced the model's discrimination capabilities, thus achieving excellent performance.

TABLE III
THE RESULTS ON UJIINDOORLOC. BEST RESULTS IN **BOLD**, SECOND BEST IN *Italic.*

Model	Floor Acc(%)	Building Acc(%)	MLE(m)
CNNLoc [17]	**96.03**	*99.27*	11.78
HADNN [19]	93.15	**100**	11.59
GNN [7]	94.15	**100**	9.61
DNNBN [18]	93.97	**100**	*9.45*
PL-Loc	*95.23*	**100**	**9.38**

To further verify the model's generalization ability, we conducted an experimental evaluation using the Tampere dataset alongside mainstream indoor localization models. The results are presented in Table IV. There were two indoor localization tasks in Tampere: floor prediction and coordinate prediction. In the Task-Adaptive Output Module, only the output branches corresponding to these two tasks were used. The results in the table indicate that PL-Loc achieved the best performance in both the floor prediction classification task and the coordinate prediction regression task. This not only demonstrated that PL-Loc could achieve excellent localization performance according to different indoor localization tasks, but also showed that the Feature Encoder Module and Two-Step Capture Module of PL-Loc effectively captured the local fingerprint features of Tampere and modeled the global dependencies efficiently. This further confirmed that PL-Loc had excellent localization performance and robustness.

TABLE IV
THE RESULTS ON TAMPERE. BEST RESULTS IN **BOLD**, SECOND BEST IN *Italic.*

Model	Floor Acc(%)	MLE(m)
HADNN [19]	94.58	9.07
SALLoc [16]	-	9.52
CRSS [21]	91.32	-
CAE-CNNLoc [23]	88.90	10.24
2L-HELM [20]	*94.81*	*8.73*
PL-Loc	**94.96**	**7.51**

D. Ablation Experiment

To further verify the effectiveness of each module of PL-Loc, we conducted ablation experiments on the UJIIndoorLoc dataset by removing the Feature Encoder Module, Two-Step Capture Module, and Temperature-Scaled Softmax, respectively. The experimental results are listed in Table V.

TABLE V
THE RESULTS OF THE ABLATION EXPERIMENT. BEST RESULTS HIGHLIGHTED IN **BOLD**.

Methods	Floor Acc	Building Acc	MLE
w/o Feature Encoder	94.69	**100**	10.26
w/o Two-Step Capture	93.52	**100**	9.89
w/o Temperature-Scaled	93.16	99.91	9.64
PL-Loc	**95.23**	**100**	**9.38**

As can be seen from the table, when the Feature Encoder Module was removed, the performance of the model on MLE decreased the most, indicating that the capture of local fingerprint features by the Feature Encoder Module was crucial for coordinate prediction. Specifically, the accuracy of floor prediction declined by 0.54%. This could be attributed to the PL-Loc model's use of a multi-task joint training approach, where the task granularity for floor prediction was finer than that for building prediction. Consequently, it relied more heavily on the local features of fingerprints, underscoring the importance of the Feature Encoder Module in encoding these local features. When the Two-Step Capture Module was removed, the accuracy of floor prediction decreased by 1.71%, and the performance on MLE fell by 0.51 m. This finding showed that the module could establish global dependencies for different APs and spatial locations. Furthermore, when Temperature-Scaled Softmax was removed, the accuracy of floor prediction and building prediction classification tasks decreased by 2.07% and 99.91%, respectively. Correspondingly, due to multi-task joint training, MLE also decreased slightly by 0.26 m. This result demonstrated that Temperature-Scaled Softmax played a vital role in calibrating the output probability distribution for classification tasks, effectively utilizing the fingerprint features modeled by the preceding module. In summary, the ablation experiment clearly illustrated that each module within PL-Loc played an irreplaceable role in high-precision indoor localization.

V. Conclusion

This paper proposes a novel indoor localization method based on WiFi fingerprint, PL-Loc, for complex indoor environments with multiple buildings and multiple floors. PL-Loc employs lightweight progressive learning to effectively encode local features in fingerprint data while also efficiently modeling global dependencies through multi-head cross-path attention calculation. Additionally, it adaptively adjusts the output structure of prediction results for indoor localization tasks across various environments, achieving efficient and high-precision indoor localization. Comparative experiments with mainstream indoor localization methods were conducted on the UJIIndoorLoc and Tampere datasets. The results demonstrated that PL-Loc achieved 100% building prediction accuracy and 9.38 m MLE on UJIIndoorLoc, while achieving 94.96% accuracy in floor prediction on Tampere. The experimental results suggest that PL-Loc not only effectively and efficiently extracts location fingerprint features, but also achieves high-precision indoor localization in different indoor environments. PL-Loc offers a new perspective for research into indoor localization methods based on WiFi fingerprints.

References

[1] Z. Turgut and A. G. Kakisim, "An explainable hybrid deep learning architecture for wifi-based indoor localization in internet of things environment," *Future Generation Computer Systems*, vol. 151, pp. 196–213, 2024.

[2] T.-G. Sorescu, A.-V. Militaru, V.-M. Chiriac, C.-R. Comsa, and I.-E. Alecsandrescu, "Uwb indoor localization: Accuracy evaluation in a controlled environment," in *2024 16th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, 2024, pp. 1–6.

[3] Z. Luo, W. Li, Y. Wu, H. Dong, L. Bian, and W. Wang, "Accurate indoor localization for bluetooth low energy backscatter," *IEEE Internet of Things Journal*, 2024.

[4] B. Zhang, "Real-time localization of zigbee signals using graph neural network," *IEEE Transactions on Industrial Electronics*, pp. 1–8, 2025.

[5] S. Wang, S. Wang, Y. Feng, W. Huang, S. Jiang, and Y. Zhang, "Rp-fusion: Robust rfid indoor localization via fusion rssi and phase fingerprint," in *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE, 2024, pp. 145–150.

[6] X. Zhang, W. Sun, J. Zheng, A. Lin, J. Liu, and S. S. Ge, "Wi-fi-based indoor localization with interval random analysis and improved particle swarm optimization," *IEEE Transactions on Mobile Computing*, vol. 23, no. 10, pp. 9120–9134, 2024.

[7] S. Wang, S. Zhang, J. Ma, and O. A. Dobre, "Graph-neural-network-based wifi indoor localization system with access point selection," *IEEE Internet of Things Journal*, vol. 11, no. 20, pp. 33 550–33 564, 2024.

[8] J. Torres-Sospedra, R. Montoliu, A. Martínez-Usó, J. P. Avariento, T. J. Arnau, M. Benedito-Bordonau, and J. Huerta, "Ujiindoorloc: A new multi-building and multi-floor database for wlan fingerprint-based indoor localization problems," in *2014 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 2014, pp. 261–270.

[9] E. S. Lohan, J. Torres-Sospedra, H. Leppäkoski, P. Richter, Z. Peng, and J. Huerta, "Wi-fi crowdsourced fingerprinting dataset for indoor positioning," *Data*, vol. 2, no. 4, p. 32, 2017.

[10] F. Jiang, D. Caruso, A. Dhekne, Q. Qu, J. J. Engel, and J. Dong, "Robust indoor localization with ranging-imu fusion," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 11 963–11 969.

[11] R. A. Khalil and N. Saeed, "Hybrid toa/aoa localization for indoor multipath-assisted next-generation wireless networks," *Results in Engineering*, vol. 22, p. 102200, 2024.

[12] P. Yin, D. Zhang, T. Zhang, S. Yang, G. Wang, Y. Hu, and Y. Chen, "Autocali: Enhancing aoa-based indoor localization through automatic phase calibration," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 9046–9050.

[13] W. Zhao, A. Goudar, X. Qiao, and A. P. Schoellig, "Util: An ultra-wideband time-difference-of-arrival indoor localization dataset," *The International Journal of Robotics Research*, vol. 43, no. 10, pp. 1443–1456, 2024.

[14] X. Zhou, L. Chen, Y. Chen, H. Yin, X. Chen, and W. Wang, "Fusion of imu and probabilistic model for indoor localization based on bayesian framework," *IEEE Internet of Things Journal*, vol. 12, no. 11, pp. 17 080–17 094, 2025.

[15] X. Yang, Y. Zhuang, F. Gu, M. Shi, X. Cao, Y. Li, B. Zhou, and L. Chen, "Deepwipos: A deep learning-based wireless positioning framework to address fingerprint instability," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 6, pp. 8018–8034, 2023.

[16] S. L. Ayinla, A. A. Aziz, and M. Drieberg, "Salloc: An accurate target localization in wifi-enabled indoor environments via sae-alstm," *IEEE Access*, vol. 12, pp. 19 694–19 710, 2024.

[17] X. Song, X. Fan, X. He, C. Xiang, Q. Ye, X. Huang, G. Fang, L. L. Chen, J. Qin, and Z. Wang, "Cnnloc: Deep-learning based indoor localization with wifi fingerprinting," in *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, 2019, pp. 589–595.

[18] S. Lukman Ayinla, A. A. Aziz, M. Drieberg, M. Susanto, A. Tumian, and M. Yahya, "An enhanced deep neural network approach for wifi fingerprinting-based multi-floor indoor localization," *IEEE Open Journal of the Communications Society*, vol. 6, pp. 560–575, 2025.

[19] J. Cha and E. Lim, "A hierarchical auxiliary deep neural network architecture for large-scale indoor localization based on wi-fi fingerprinting," *Applied Soft Computing*, vol. 120, p. 108624, 2022.

[20] A. Alitaleshi, H. Jazayeriy, and J. Kazemitabar, "Affinity propagation clustering-aided two-label hierarchical extreme learning machine for wi-fi fingerprinting-based indoor positioning," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 6, pp. 3303–3317, 2022.

[21] I. Neupane, S. Shahrestani, and C. Ruan, "Indoor localization of resource-constrained iot devices using wi-fi fingerprinting and convolutional neural network," in *Proceedings of the 2024 Australasian Computer Science Week*, ser. ACSW '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 20–25.

[22] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 603–612.

[23] A. Kargar-Barzi, E. Farahmand, N. Taheri Chatrudi, A. Mahani, and M. Shafique, "An edge-based wifi fingerprinting indoor localization using convolutional neural network and convolutional auto-encoder," *IEEE Access*, vol. 12, pp. 85 050–85 060, 2024.