Cities 169 (2026) 106557

Contents lists available at ScienceDirect

Cities

journal homepage: www.elsevier.com/locate/cities

# From human mobility to building functions: A deep learning approach for urban building classification in Megacity Tokyo

Zhihui Hu[a], Yao Yao[a,b,g,h,i,*], Qia Zhu[c], Zijin Guo[d], Guicheng Li[a], Peiran Li[e,h], Renhe Jiang[f], Junfang Gong[a], Qingfeng Guan[a,b], Ryosuke Shibasaki[e,g,i]

[a] *School of Geography and Information Engineering, China University of Geosciences, Wuhan, Hubei province, China*
[b] *National Engineering Research Center of Geographic Information System, China University of Geosciences, Wuhan, Hubei province, China*
[c] *School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, 430072, Hubei province, China*
[d] *Key Laboratory of Geotechnical Mechanics and Engineering of Ministry of Water Resources, Changjiang River Scientific Research Institute, Wuhan, 430010, China*
[e] *Interfaculty Initiative in Information Studies & Graduate School of Interdisciplinary Information Studies, The University of Tokyo, Tokyo, Japan*
[f] *Center for Spatial Information Science, The University of Tokyo, Chiba, Japan*
[g] *LocationMind Institution, LocationMind Inc., Chiyoda, Tokyo, Japan*
[h] *Hitotsubashi Institute for Advanced Study, Hitotsubashi University, Kunitachi, Tokyo, Japan*
[i] *Faculty of Engineering, Reitaku University, Kashiwa, Chiba, Japan*

ARTICLE INFO

*Keywords:*
Building function classification
Deep learning
Point of interest
Human trajectory
Multi-source data fusion

ABSTRACT

Accurately identifying building function is essential for urban management, urban renewal, and promoting sustainable city development. Previous studies on building function classification have primarily focused on extracting external physical characteristics from remote sensing imagery and socio-economic attributes from Points of Interest (POI) data. However, these studies often overlook the patterns of human mobility within buildings, making it challenging to identify building functions accurately. To address these issues, this study mines the latent features embedded in POI and human trajectory data, and constructs a deep learning model, which integrates POI category semantics and human mobility patterns to identify urban building functions. We evaluated the proposed model in the 23 districts of Tokyo, Japan. The results indicate that the proposed model is able to extract features obtained from diverse data sources to identify building functions, achieving a test accuracy of 90.27 % and a Kappa coefficient of 0.8858. The building function mapping results in Tokyo demonstrate that the proposed model can accurately classify building functions in megacities. This study finds that human mobility patterns within buildings significantly improve the identifying accuracy of residential and commercial buildings. The building function mapping results of this study can provide effective data support for urban planning in Tokyo.

## 1. Introduction

### 1.1. Background

Urban Buildings, as the fundamental units supporting urban functions, provide essential spaces and venues for residents' daily activities, such as living, working, studying, and leisure (Niu et al., 2017; Zhang et al., 2023). With the acceleration of urbanization and the rapid growth of urban populations, the number of buildings is continuously increasing, and the functional types of buildings are becoming more complex and diverse. The functional attributes of buildings, as core elements, are essential for understanding human activity intentions (Humtsoe, 2022; Yao et al., 2023), predicting urban traffic flow (Liu et al., 2022), and urban planning (Wang et al., 2020). In the context of dynamic urban management and renewal, accurate identification of building functions is crucial for optimizing urban resource allocation and formulating adaptive strategies. For example, unreasonable urban functional layout leads to increasingly serious problems such as traffic

* Corresponding author at: UrbanComp Lab, School of Geography and Information Engineering, China University of Geosciences, Wuhan, 430078, Hubei province, China.

*E-mail addresses:* zhihui.hu@cug.edu.cn (Z. Hu), yaoy@cug.edu.cn (Y. Yao), qiazhu@cug.edu.cn (Q. Zhu), gzj2017@cug.edu.cn (Z. Guo), liguicheng@cug.edu.cn (G. Li), lipeiran@iii.u-tokyo.ac.jp (P. Li), jiangrh@csis.u-tokyo.ac.jp (R. Jiang), jfgong@cug.edu.cn (J. Gong), guanqf@cug.edu.cn (Q. Guan), shiba@csis.u-tokyo.ac.jp (R. Shibasaki).

https://doi.org/10.1016/j.cities.2025.106557
Received 6 November 2024; Received in revised form 23 August 2025; Accepted 1 October 2025

congestion, energy waste and environmental pollution (Choi & Yoon, 2023; Shen et al., 2021). However, the identification of urban building functions still relies on national land surveys and official census data, which require substantial human and financial resources. This is particularly challenging in megacities with a vast number of buildings.

Previous research attempts to break through this bottleneck through remote sensing image interpretation, mainly using methods such as morphological feature-based and texture analysis to conduct land use classification at the scale of kilometer grids or traffic analysis zones (Liu & Shi, 2020; Tong et al., 2020; Wang et al., 2023). Although the external shape and textural characteristics of buildings remain mostly constant, their function might alter as a result of human activity (Zhong et al., 2014). For example, buildings with similar roof structures may have completely different functions, and functional changes may only be achieved through changes in indoor activities. This phenomenon of "physical-functional" decoupling leads to a theoretical ceiling for the classification accuracy that solely relies on remote sensing data (Feng et al., 2021; Zhang et al., 2018). Besides, this phenomenon is growing in rapidly evolving urban environments due to economic shifts, demographic changes, and the repurposing of existing buildings. For instance, an old factory might be converted into creative workshops, a residential building might house various small businesses. Thus, a research shift has occurred towards multi-source spatiotemporal data fusion methods, which integrate multimodal data to enable large-scale and fine-grained urban building function identification.

### 1.2. Related work

With the development of location-based services (LBS), a large amount of social sensing data with spatiotemporal attributes has been generated, such as social media check-in data, taxi trajectories, mobile signaling, points of interest, and street view images (Liu et al., 2015). The accessibility of this data allows for the identification of building functions by incorporating human mobility and socio-economic information (Cao et al., 2020). For instance, Zhong et al. (2014) classified the building functions by extracting interaction features between residents' trips and buildings from smart card data. Chen et al. (2020) identified building functions by calculating the text-similarity of POIs within buildings and the ratio of different types of POIs. Compared to remote sensing images, street view images and social media images can provide cross-sectional views and ground-level information about city spaces. Srivastava et al. (2018) used convolutional neural networks to perform multi-label classification on multi-view Google Street View images to infer building functions. Hoffmann et al. (2023) harnessed Google Street View and Flickr social media image data to construct a content-based automatic filtering pipeline and fine-tuned Convolutional Neural Network architectures, thereby achieving building function classification.

However, studies that categorize building functions relying on a single data source struggle with the problem of uneven data distribution. For example, POI data is unevenly distributed in cities, which is dense in commercial centers but sparse in residential areas and suburbs, making it challenging to classify buildings in these areas based on POI data. Similarly, street view data generally only covers buildings along major roads, which hampers the identification of building functions off these roads.

In recent years, researchers have tried to improve building function classification by integrating multi-source spatiotemporal data. For example, Niu et al. (2017) inferred the functions of buildings in Tianhe District, Guangzhou, China, by integrating real-time location records from WeChat users, taxi trajectories, and POI data using a density-based method. Liu et al. (2018) realized the building function classification by designing a probabilistic model that utilizes taxi trajectories, social media data, remote sensing imagery, and POI data. Zhuo et al. (2019) identified building functions by extracting spatiotemporal interaction features between different functional buildings based on taxi trajectory data and calculating the daily crowd density characteristics of buildings using Tencent Temporal Population data, employing an iterative clustering method. Deng et al. (2022) improve the accuracy of identifying building functions within residential areas by developing a hierarchical data mining model that integrates remote sensing imagery, street view images, and POI data. Nevertheless, the majority of these studies focus on extracting and calculating various indicator features from multi-source data, employing a conventional machine learning method to classify building function types. These methods often involve complex and cumbersome feature design steps, and their accuracy requires improvement.

The development of deep learning representation techniques has provided a technical foundation for deeply mining and representing socio-economic attributes and human mobility information in multi-source data (Moreira et al., 2019; Wei & Yu, 2024). In terms of POI data representation, Yao et al. (2017) pioneered the combination of POI data with the word2vec model, which embeds POIs into vector space to identify land use. Zhai et al. (2019) developed a Place2vec model to classify urban functional zones by analyzing the high-semantic information of POI data. Huang et al. (2022) used the word2vec model and manifold learning algorithm to extract spatial co-occurrence and category semantic information from POI data, conducting urban function classification for Xiamen Island.

For temporal data representation, early studies used methods like DTW to extract temporal features from time-series data (Chen et al., 2017), while this method is complex to model and has limited feature extraction capabilities. Recently, time-series representation models such as LSTM, ConvLSTM, and TCN have been widely used in temporal data mining and representation because of their effectiveness in capturing time dependencies in time-series data (Hao et al., 2023; Ismail Fawaz et al., 2019; Shi et al., 2015). For instance, Yao et al. (2022) achieved urban land use classification at the scene scale by extracting temporal features from time-series electricity data based on the TCN model. Nonetheless, how to represent and integrate temporal series data at the building scale to construct a high-accuracy urban building function classification model remains a pressing issue.

### 1.3. Research gaps

In summary, two significant issues remain unresolved, as revealed by the reviewed literature. Firstly, previous studies have primarily focused on extracting the natural attributes from remote sensing imagery and the static socio-economic features from POI, neglecting the dynamic human mobility information within buildings. This oversight limits our comprehensive understanding of building functions. Secondly, although the fusion of multi-source spatiotemporal data has proven effective in identifying building functions, most existing research on building function classification based on such data relies heavily on manual feature engineering and traditional machine learning models. These approaches struggle to effectively extract and integrate the deep features and latent correlations within the multi-source spatiotemporal data inside buildings, which limits the accuracy of building function classification.

### 1.4. Research questions

Based on the identified gaps in existing research, this study addresses the following key research questions:

1) How can dynamic human mobility information within buildings be effectively captured and utilized to overcome the limitations of relying solely on static POI and remote sensing data for building function classification?
2) How to mine and integrate potential associations within POI data and crowd trajectories to improve the accuracy and generalization performance of building function classification models?

These research questions directly address the two major limitations identified in existing literature: (1) the neglect of dynamic human mobility information within buildings, and (2) the inability of traditional methods to effectively extract and integrate deep features from multi-source spatiotemporal data. By answering these questions, this study aims to advance both the methodology and practical applications of urban building function classification.

## 2. Study area and data

### *2.1. Study area*

The study area of this study is the 23 districts of Tokyo, Japan, with a total area of 628 km$^2$ and a population of approximately 9.71 million. The 23 districts are widely regarded as the core area of Tokyo, accounting for about 70 % of Tokyo Metropolis' total population. Characterized by flat terrain and densely packed with over 1 million buildings and service facilities, the 23 districts of Tokyo present a highly dense urban landscape. We chose Tokyo as the study area, aiming to investigate the spatial distribution of urban function in megacities. Fig. 1 depicts the distribution of buildings across these districts.

### *2.2. Datasets*

#### *2.2.1. Building footprint data*

We collected a total of 911,332 building footprints within the study area from the OpenStreetMap. The original type tags from OSM were reclassified into six building type labels: residential, commercial, administrative, educational, public service, and industrial refer to Deng et al., 2022 and Zhang et al., 2023, as shown in Table 1. To ensure sufficient sample sizes for each category, we supplemented the label for categories with fewer samples, following the approach proposed by Liu et al. (2018). A total of 15,071 labeled building samples were obtained, representing 1.65 % of all buildings, distributed as follows: residential (1952, 12.95 %), commercial (4219, 28.00 %), administrative (1806, 11.98 %), educational (2358, 15.65 %), public service (2719, 18.04 %),

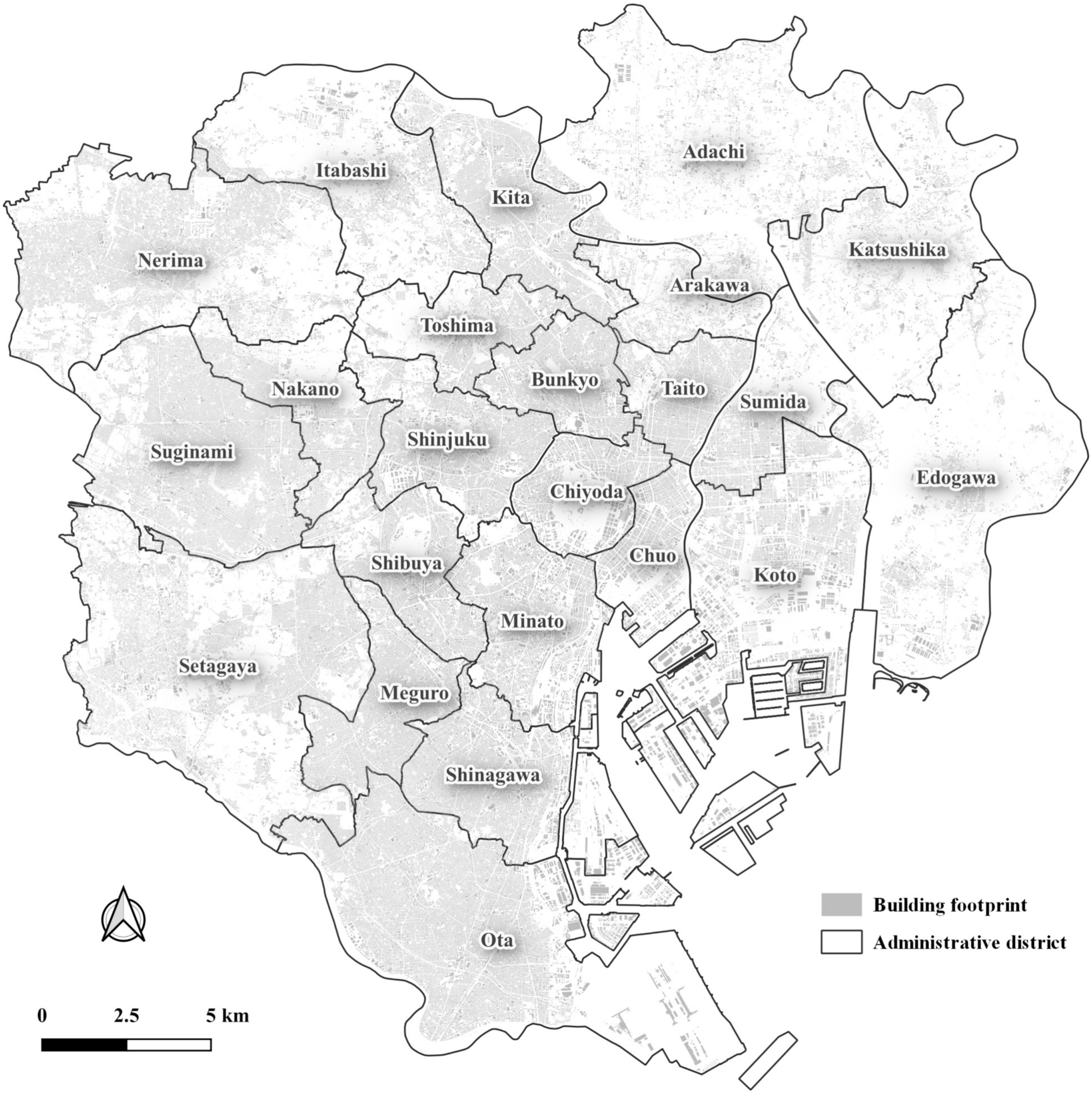


**Fig. 1.** The study area: Tokyo, Japan.

**Table 1**
The mapping of building function types to OSM platform tags.

| Buildings type label | Building type tags from OpenStreetMap |
|---|---|
| Residential (Res.) | Residential, Apartments, House, etc. |
| Commercial (Com.) | Retail, Supermarket, Cafe, Office, Hotel, Company, etc. |
| Administrative (Adm.) | Government, Fire Station, Police Station, etc. |
| Education (Edu.) | School, Kindergarten, College, University, etc. |
| Public service (Pub.) | Temple, Library, Museum, Sports Center, Hospital, etc. |
| Industrial (Ind.) | Industrial, Warehouse, Factory, etc. |

and industrial (2017, 13.38 %).

### 2.2.2. *Point of interest*

In order to train high-quality POI embedding vectors, we obtained POI data for all major cities in Japan from the “Telepoint Pack DB” provided by ZENRIN DataCom Co., Ltd. The dataset includes 39 primary categories and 731 secondary categories, totaling 5.6 million entries. Due to the presence of substantial redundant information and inappropriate categories for building function classification in the POI dataset, this study consolidated specific categories. For example, industrial types like “paper”, “manufacturing” and “petroleum coal” were grouped into the industrial category, while road-related categories were omitted. The final POI was reclassified into 25 categories: catering, shopping services, hotel services, commerce, leisure and entertainment, real estate, tourism and sightseeing, mining, aquaculture, agriculture and forestry, steel, professional technical services, finance and insurance, nonferrous metals, metal products, petroleum and coal products, educational services, culture and media, lifestyle services, sports facilities, communication and information services, medical welfare, government and public institutions, transportation and logistics, vehicle - related, and others, totaling 4.9 million entries.

### 2.2.3. *Trajectory data*

Human trajectory data was used in this study to obtain human mobility time series in buildings. The trajectory data utilized in this study is big human GPS trajectory data provided by Blogwatcher Inc. (Fan et al., 2019; Jin et al., 2023; Zhiwen et al., 2023). We collected mobile device location data within the study area from August 15 to August 21, 2022. The original GPS data exceeds 200 million records, completely covering Tokyo. Since its users cover 80 million of Japan's total population of 126 million, this estimate can achieve very high statistical accuracy. This study conducted 30-minute interval statistics and 5-meter resolution interpolation resampling on the mobile device location data within buildings, generating the time-series population raster data in the 23 districts of Tokyo. Then using “extract by mask” tool, the vector data of buildings is superimposed and analyzed with the obtained multi-band time series raster data of people flow, and the time series of people flow in buildings is extracted. Fig. 2 is a visualization of the human mobility flow time series for different building types.

## 3. Methodology

Fig. 3 shows the framework of STAF-Net, which consists of three main parts. The POI category semantic feature extraction module employs a semantic preservation algorithm and a Set2Set aggregation function to extract semantic features of POI categories within buildings. The time series feature extraction module utilizes the TimesNet model to capture the temporal change characteristics of human mobility flow in buildings. In the adaptive fusion module, a multi-head attention mechanism fully integrates these features, and a SoftMax function classifies the building functions.

### 3.1. *Semantics-preserved-based feature extraction from POI data*

The POI category semantic extraction module is designed to derive category semantic feature from POI data, which consists of two main components: POI category encoder and POI aggregation function. Specifically, the POI encoder is trained on POI data using a semantic-preserving POI embedding method to generate a dictionary of category embedding vectors for each POI category. For all POI data within a building, the corresponding POI embedding vectors are obtained based on this dictionary. Finally, the POI embedding vectors within the building are aggregated using the POI aggregation function to derive the feature vector that represents the POI semantic of the building.

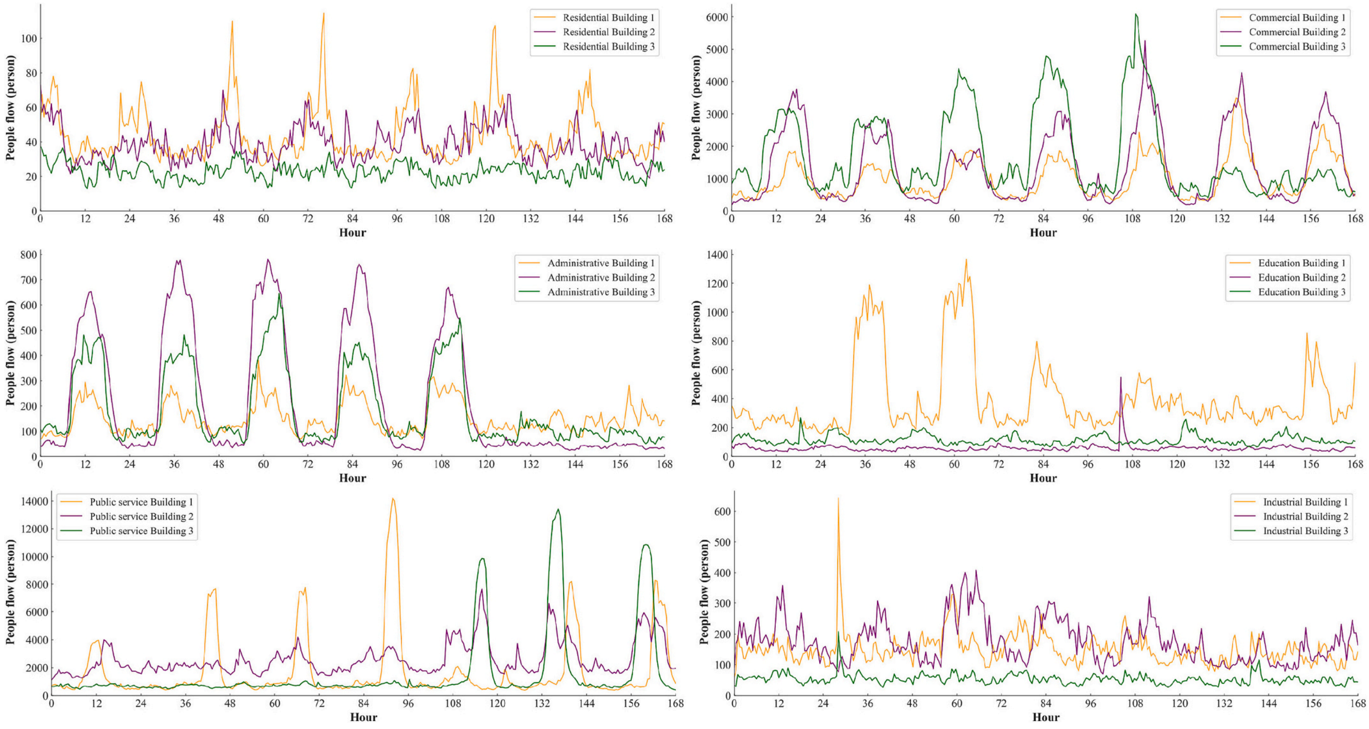


**Fig. 2.** The human mobility flow time series over a week for six types of building samples.

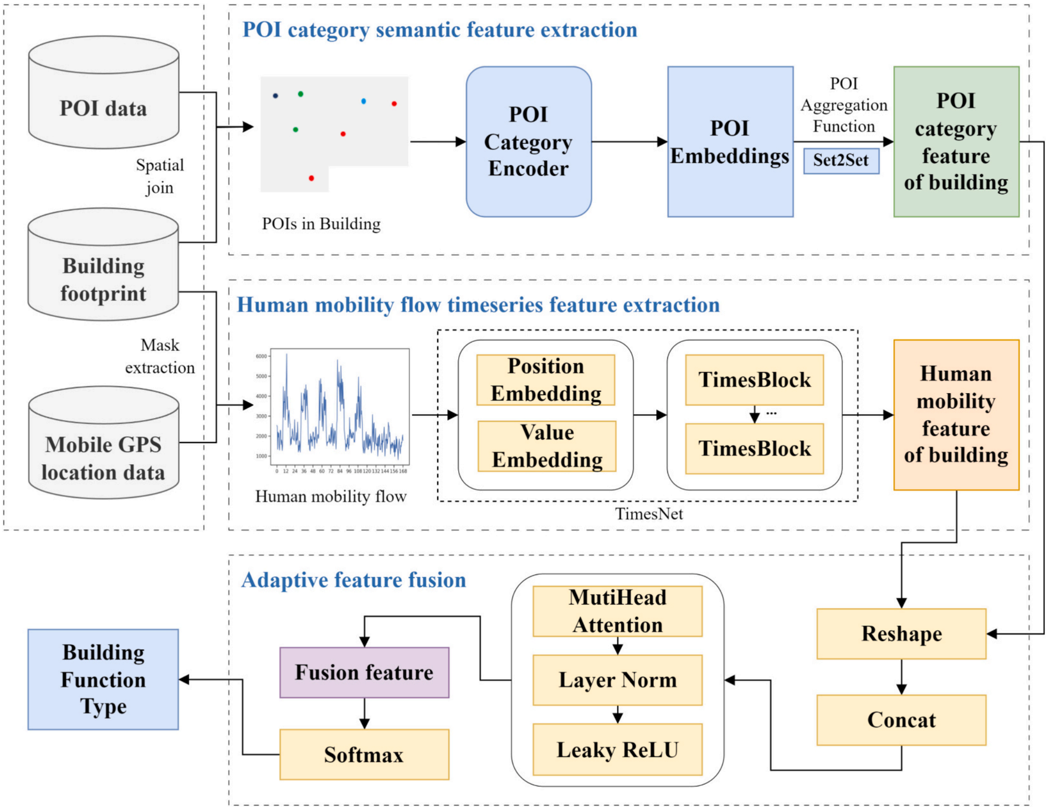


**Fig. 3.** The framework of the proposed model.

### 3.1.1. *POI category encoder*

This study employs the semantics preserved algorithm proposed by Huang et al. (2022) to embed the POI data. As shown in Fig. 4, the POI Category Encoder mainly consists of two components, which are utilized to extract the spatial co-occurrence and hierarchical structure information of POIs. Firstly, POI data from all over Japan is used to create a network of points of interest using the Delaunay triangulation. Each point of interest is represented as a vertex in the network, and the edges

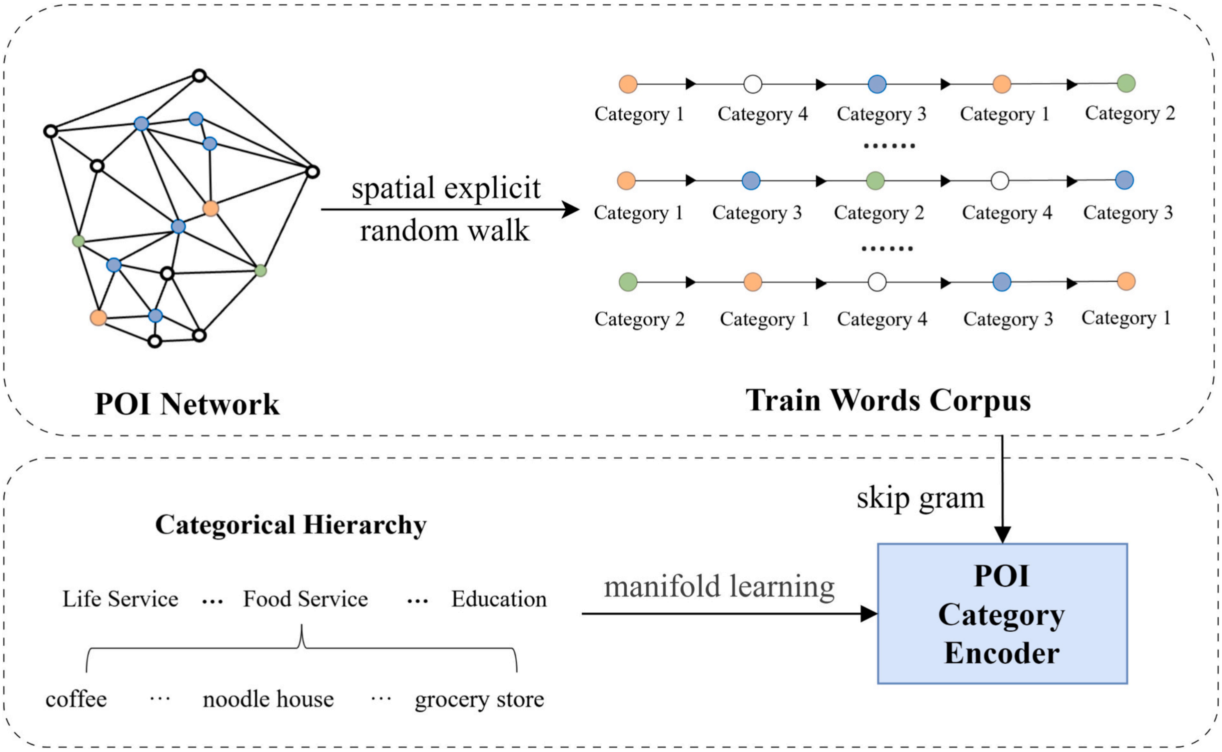


**Fig. 4.** The workflow of POI embedding based on the semantic preserved algorithm.

connecting two points of interest are considered edges. A random walk algorithm is used to traverse several times from each node to generate several POI sequences. In this study, each node is assigned three random walk paths, each with a length of 5. The skip-gram model was used to train the POI sequences (Mikolov et al., 2013).

Secondly, the manifold learning algorithm is used to capture category hierarchy information, which is based on the hypothesis that if two secondary POI categories are part of the same primary category, their embeddings (i.e., their embeddings in a vector space) should be close to each other. In order to obtain more detailed POI semantic information, this study employs secondary categories to embed every POI. Additionally, the negative sampling loss function was used to minimize the training loss, as shown in Eq. (1). Through this process, we can train a POI encoder that comprehensively learns POI space co-occurrence and hierarchical semantics. Consequently, POIs in the building can generate their corresponding embeddings through the POI encoder.

$$\mathscr{L}_{co-occurrence} = \sum_{t=1}^{P} \sum_{-w \le j \le w} - \left( \log\left(\sigma\left(e_t^T e_{t+j}\right)\right) - \sum_{i=1}^{Q} \log\left(\sigma\left(e_t^T e_n\right)\right) \right) \tag{1}$$

where $P$ denotes POI categories from the sequences obtained through random walks, $w$ is the context window size, $e_t$ and $e_{t+j}$ denote the vector embedding of the center word at position $t$ and the context word at position $t+j$, $\sigma$ is the sigmoid function, $e_N$ is the vector embedding of $n$-th negative sampling word, Q is the number of negative sampling words for each center word.

### 3.1.2. Set2Set

The number of POI within different buildings varies, necessitating the aggregation of POI vectors to facilitate subsequent feature fusion and model training. Unlike common average pooling or max pooling methods, this study adopted Set2Set to aggregate the collection of POI vectors within buildings. Set2Set is a neural network for processing graph data, which is based on the sequence-to-sequence framework and is designed specifically for processing set-type data. The core idea of Set2Set is to transform the set of nodes into a fixed-size vector representation, which can capture the order and structure information in the set of nodes (Vinyals et al., 2016). This method excels in handling the complex relationships and patterns among set data, thereby mitigating the potential information loss. Specifically, it utilizes LSTM and attention mechanisms to capture the optimal hidden sequence within a set of entities. In this study, POIs within a building generate embeddings, $E_{p_1}$, $E_{p_2} \ldots E_{p_n}$ according to the POI Encoder. These embeddings were then aggregated by the Set2Set model to obtain the feature vector $P_b$, which represents the semantic category of the POIs within the building.

$$P_b = set2set\left(E_{p_1}, E_{p_2} \ldots E_{p_n}\right) \tag{2}$$

where $n$ denotes the number of POI within a building, $E_{p_1}, E_{p_2} \ldots E_{p_n}$ denote the POI embeddings, $P_b$ represents the embeddings aggregated by Set2Set.

## 3.2. TimesNet-based feature extraction from human mobility time series data

Observing the sample data in Fig. 2, it is evident that the variation in human mobility flow around the building exhibits a distinct periodic pattern. Hence, capturing the periodic fluctuations in the human mobility flow time series is a crucial aspect of this module. Conventional time series models, which rely on a one-dimensional time axis, are limited to identifying changes between adjacent time points. However, TimesNet demonstrates its unique advantage by extending the 1D time series into 2D space, effectively extracting the periodic changes in the time series data. Currently, it holds a leading position in the field of time series classification.

TimesNet is composed of multiple TimesBlocks stacked and connected via residuals (He et al., 2016). In Times Block, there are three primary steps involved. Firstly, a 1D time series is transformed into a 2D space. Convolutional neural networks then extract features related to both intra-period and inter-period variations from these 2D time series images simultaneously. Finally, the extracted features undergo a dimension transformation and weighted summation. Specifically, for each time series $\boldsymbol{X}_{1D}$ of input length $N$, a Fast Fourier Transform is applied to transformed it from the time space to the frequency space. In the frequency domain, the amplitudes are used to extract the top $k$ significant periodic features. Subsequently, based on the number of periods, the time series is divided and folded to produce $k$ two-dimension time series images.

$$f_1, \ldots, f_k = Topk(Amp(fft(\boldsymbol{X}_{1D}))) \tag{3}$$

$$p_1, \ldots, p_k = \left\lceil \frac{N}{f_1} \right\rceil, \ldots, \left\lceil \frac{N}{f_k} \right\rceil \tag{4}$$

where, $\{f_1, \ldots, f_k\}$ represents the k components with the highest amplitude intensity in the frequency domain after transformation and $\{p_1, \ldots, p_k\}$ are their corresponding periods.

The two-dimensional time series images obtained exhibit two-dimensional locality because each column and row correspond respectively to adjacent moments and periods, where neighboring moments and periods often contain similar temporal changes. This characteristic allows for the extraction of feature information through convolutional kernels (Wu et al., 2023). Therefore, this study uses the Inception model to extract features from the constructed two-dimensional time series images. Subsequently, after dimension adjustment of the $k$ temporal feature maps obtained from the feature extraction, they are weighted and summed according to the intensity of their corresponding frequencies to produce the final output vector $E_t$.

Before using the TimesNet model to extract features from the human mobility flow time series of each building, it is necessary to compute the initial embeddings of the time series, which is represented as $B_t = \{t_1, t_2, \ldots, t_s\}$, where $s$ is the steps of the time series. Initially, the data undergoes processing through a value embedding layer and a positional embedding layer to compute the value embedding vector $E_v$ and the positional embedding vector $E_p$. These obtained embedding vectors are then summed to produce the initialized temporal vector $E_t$. Where $E_v$, $E_p$, $E_t \in R^{n \times D}$, $D$ is the dimensionality output by the embedding layers. Subsequently, this initial vector is input into TimesNet for feature extraction, resulting in the final feature vector $T_b$ This process can be expressed as follows:

$$E_v = ValueEmbed(B_t) \tag{5}$$

$$E_p = PositionEmbed(B_t) \tag{6}$$

$$T_b = TimesNet(E_t) \tag{7}$$

## 3.3. Adaptive fusion module

In this study, an adaptive feature fusion module was developed to integrate features from multisource data using multi-head attention. Attention mechanisms can adaptively allocate weights between two parts of feature vectors during model training, enhancing model classification performance. This method is widely used in feature fusion research involving multi-source data (Li et al., 2022). The module includes a normalization layer, a multi-head attention layer, and an activation function. Within this module, the POI semantic feature vector $P_b$ and the human mobility temporal feature vector $T_b$, extracted earlier, are first adjusted to a uniform dimension through a fully connected layer. Then, a multi-head attention layer adaptively learns the weights of different features. Finally, the fused feature vector is classified for building functions using the softmax function.

$$P^b_{atten} = MutiheadAttention(P_b) \odot P_b \tag{8}$$

$$T^b_{atten} = MutiheadAttention(T_b) \odot T_b \tag{9}$$

$$E_{final} = \left( \left[ layerNormal\left(P^b_{atten}\right), layerNormal\left(T^b_{atten}\right) \right] \right) \tag{10}$$

where $P^b_{atten}$ and $T^b_{atten}$ represent the building POI category semantic vector and the building human mobility pattern vector, respectively, both of which have been processed through multi-head attention calculations. $E_{final}$ denotes the vector that results from the layer normalization and concatenation of the vectors. This structured approach ensures that the model effectively integrates and utilizes the distinct characteristics of each data source, enhancing the overall predictive accuracy and relevance of the output.

### 3.4. Baseline models

In our experiments, three representative baseline models were chosen to compare with our model. The word2vec and LSTM models utilize POI features and temporal features for classification, respectively, while the Random Forest model directly uses both types of feature vectors for building function classification. This study uses test accuracy and kappa coefficient to evaluate the classification results of the models.

**Word2Vec:** Yao et al. (2017) first utilized this model to represent POI data and applied it to land use classification. This method treats each POI as a word, using a greedy algorithm to acquire the shortest paths among POIs within TAZs, considering them as sentences to create a corpus of POIs for all TAZs. Then, the POI corpus is trained by the word2vec algorithm proposed by (Mikolov et al., 2013) to obtain POI embeddings for each POI category. Finally, the POI embeddings are input into multilayer perceptron (MLP) for building function classification.

**LSTM:** LSTM (Long-Short-Term Memory Network) is a deep learning model commonly used to process sequence data (Hochreiter & Schmidhuber, 1997). It enhances the network's memory capability by introducing memory cells and gate mechanisms, which effectively capture the long-term dependencies in sequence. Besides, it is well-regarded for its generalization capabilities and ability to avoid overfitting and is widely used in time series data modeling (Hua et al., 2019).

**Random Forest:** It is a supervised learning algorithm designed to enhance the accuracy and robustness of classification or regression tasks (Breiman, 2001). An essential feature of random forest is that it can reduce the overfitting of decision trees due to overfitting data, thus improving the performance of the model (Biau, 2012).

### 3.5. Experiments setup

This study conducted a comparative analysis of the STAF-Net against a range of baseline models to evaluate and analyze its performance. The dataset mentioned in Section 2 was randomly divided into training data and test data, following an 8:2 ratio. All experiments utilized the PyTorch framework in Python 3.8, leveraging the acceleration capabilities of an NVIDIA GeForce 4090 24G GPU. To optimize the objectives, the model was trained using the Adam optimizer alongside the Cross-Entropy loss function (Ho & Wookey, 2020; Kingma & Ba, 2017). We also incorporated a learning rate decay strategy and an early stopping mechanism during the model training process to prevent overfitting.

To assess the performance of the models' classification outcomes, this study employs Test Accuracy and the Kappa Coefficient as evaluation metrics. The formulas for computing these metrics are provided below:

$$Test\ Accuracy = \frac{\sum_{i=1}^{n} x_{ii}}{N} \tag{11}$$

$$Kappa = \frac{\sum_{i=1}^{n} x_{ii} \Big/ N - \sum_{i=1}^{n} \left( \sum_{j=1}^{n} x_{ij} \sum_{j=1}^{n} x_{ji} \right) \Big/ N^2}{1 - \sum_{i=1}^{n} \left( \sum_{j=1}^{n} x_{ij} \sum_{j=1}^{n} x_{ji} \right) \Big/ N^2} \tag{12}$$

where $x_{ij}$ is the elements of the $i$-th row and $j$-th column of the confusion matrix, $x_{ii}$ is the correctly predicted samples, $n$ is the number of categories, and $N$ is the number of test samples.

## 4. Results

### 4.1. Model accuracy evaluation and model comparison

This study performed comparative experiments utilizing both individual data sources and pairs of combined data sources to evaluate the effectiveness and reliability of the STAF-Net. We also selected three baseline models for comparative analysis. For each baseline, we adjusted the corresponding hyperparameters and conducted the experiments ten times, taking the average as the final accuracy of the model. Table 2 shows the evaluation results of STAF-Net and baselines on the test data.

As shown in Table 2, experiments 2, 4, 5, and 6 demonstrate comparisons between models based on single data. When using only trajectory data, the model achieved higher test accuracy compared to the model using POI data. TimesNet demonstrated superior performance over LSTM in trajectory data analysis, achieving an 11.29 % higher test accuracy. Similarly, the Semantic-based model substantially outperformed Word2vec in POI data processing. The performance gap between temporal modeling approaches was particularly notable, with TimesNet's multi-periodic design showing clear advantages over conventional LSTM in capturing trajectory patterns. For POI data modeling, semantic-enhanced methods proved more effective than vectorization-based approaches.

Experiments 1 and 3 reflect comparisons between models utilizing two types of data. The results indicate that methods based on two types of data show significant improvements compared to methods using only one type of data. When these two types of data are utilized, the proposed model achieves an test accuracy of 90.27 % and a Kappa coefficient of 0.8858, with an test accuracy improvement of 7.2 % compared to the Random Forest model. These findings indicate that employing multi-source data can produce more precise building function classification results compared to using single data sources. Additionally, the proposed model in this study, based on deep learning, is superior in extracting complex data features compared to traditional machine learning models, significantly enhancing the effectiveness of building function classification.

### 4.2. Ablation study

To confirm the effectiveness of each component in our model, ablation experiments were conducted in this study. Table 3 shows the ablation study results, illustrating each component's impact on our model's performance. The TimesNet-based model utilizes trajectory data, while the Semantic-based model exclusively uses POI data. Both the STAF-Net (without AFM) and STAF-Net models incorporate both types of data. The difference is that the STAF-Net (without AFM) model

**Table 2**
Evaluation result of the proposed model and baselines.

| No | Model | Data source | | Test accuracy (%) | Kappa |
|---|---|---|---|---|---|
| | | POI | Trajectory | | |
| 1 | STAF-Net | √ | √ | 90.27 | 0.8858 |
| 2 | TimesNet | | √ | 85.05 | 0.8244 |
| 3 | Random Forest | √ | √ | 83.07 | 0.7921 |
| 4 | Semantic-based | √ | | 79.67 | 0.7644 |
| 5 | LSTM | | √ | 73.76 | 0.6799 |
| 6 | Word2Vec | √ | | 58.21 | 0.4877 |

**Table 3**
The ablation experiment results.

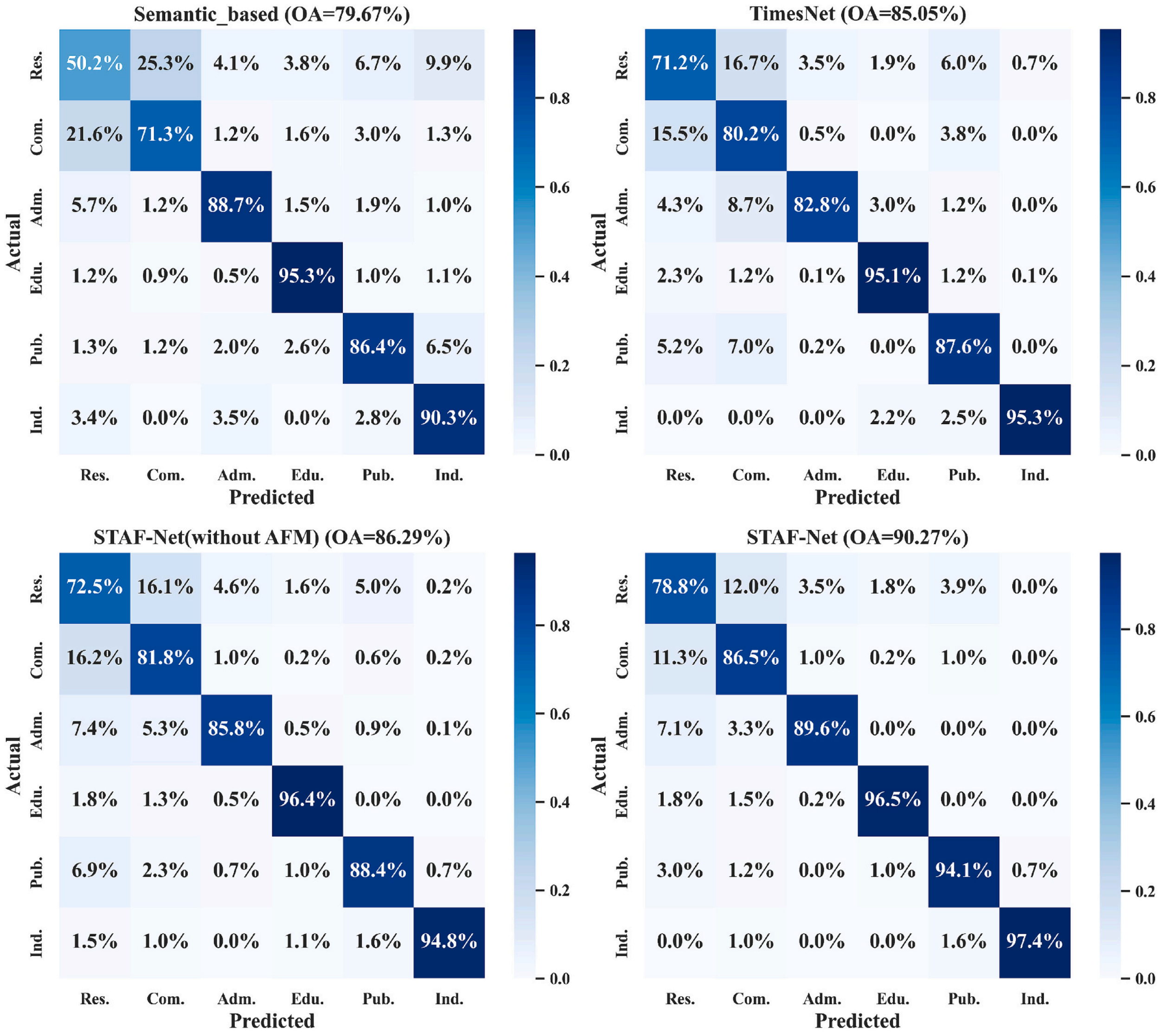
| Model | Input data | Test accuracy (%) | Kappa |
|---|---|---|---|
| STAF-Net | POI + Trajectory | 90.27 | 0.8858 |
| STAF-Net (without AFM) | POI + Trajectory | 86.29 | 0.8394 |
| TimesNet-based | Trajectory | 85.05 | 0.8244 |
| Semantic-based | POI | 79.67 | 0.7644 |

does not employ the adaptive fusion module for feature integration. The results indicate that STAF-Net achieves the best performance, followed by the STAF-Net (without AFM), the TimesNet-based model with trajectory data only, and the Semantic-based with POI data only. This indicates that the adaptive fusion module, POI semantic features, and human mobility pattern features all play a beneficial role in enhancing classification accuracies.

Specifically, when our model does not include the adaptive fusion module, the accuracy drops to 86.29 %, and the Kappa coefficient decreases to 0.8394. It confirms that the AFM plays an essential role in our model, effectively enhancing the accuracy of predictions. In comparison, the TimesNet-based model, which only inputs trajectory data, obtains an test accuracy of 85.05 % and a Kappa coefficient of 0.8244. When only using POI data, the accuracy drop s to 79.67 %, and the Kappa coefficient is 0.7644. This suggests that human mobility information from trajectory data is more effective for identifying building functions. Additionally, utilizing multisource data fusion outperforms methods that rely solely on a single data source.

In order to further analyze the degree of contribution of models in identifying different building functions, this study examines the confusion matrix of different models. The results, as shown in Fig. 5, indicate that show that STAF-Net significantly enhanced the accuracy of identifying different building types after data fusion. Specifically, the classification of industrial types showed the best performance, achieving an accuracy of 97.4 %. The classification results for educational and public service types were also excellent, reaching 96.5 % and 94.1 %, respectively. The classification accuracies for these three types of buildings were all above 90 %. Additionally, the classification results for administrative types and commercial types were also good, with accuracies of 89.6 % and 86.5 %, respectively. In contrast, the type with the poorest classification results was residential buildings, with an accuracy of 78.8 %.

A comparison of the confusion matrices for various models reveals that data fusion improved the classification accuracy for all six building types to differing extents. The most significant improvement was seen in residential types, where the accuracy increased from 50.2 % (Semantic_based) to 78.8 % (STAF-Net). Moreover, incorporating the human



Semantic_based (OA=79.67%)

| Actual \ Predicted | Res. | Com. | Adm. | Edu. | Pub. | Ind. |
|---|---|---|---|---|---|---|
| Res. | 50.2% | 25.3% | 4.1% | 3.8% | 6.7% | 9.9% |
| Com. | 21.6% | 71.3% | 1.2% | 1.6% | 3.0% | 1.3% |
| Adm. | 5.7% | 1.2% | 88.7% | 1.5% | 1.9% | 1.0% |
| Edu. | 1.2% | 0.9% | 0.5% | 95.3% | 1.0% | 1.1% |
| Pub. | 1.3% | 1.2% | 2.0% | 2.6% | 86.4% | 6.5% |
| Ind. | 3.4% | 0.0% | 3.5% | 0.0% | 2.8% | 90.3% |

TimesNet (OA=85.05%)

| Actual \ Predicted | Res. | Com. | Adm. | Edu. | Pub. | Ind. |
|---|---|---|---|---|---|---|
| Res. | 71.2% | 16.7% | 3.5% | 1.9% | 6.0% | 0.7% |
| Com. | 15.5% | 80.2% | 0.5% | 0.0% | 3.8% | 0.0% |
| Adm. | 4.3% | 8.7% | 82.8% | 3.0% | 1.2% | 0.0% |
| Edu. | 2.3% | 1.2% | 0.1% | 95.1% | 1.2% | 0.1% |
| Pub. | 5.2% | 7.0% | 0.2% | 0.0% | 87.6% | 0.0% |
| Ind. | 0.0% | 0.0% | 0.0% | 2.2% | 2.5% | 95.3% |

STAF-Net(without AFM) (OA=86.29%)

| Actual \ Predicted | Res. | Com. | Adm. | Edu. | Pub. | Ind. |
|---|---|---|---|---|---|---|
| Res. | 72.5% | 16.1% | 4.6% | 1.6% | 5.0% | 0.2% |
| Com. | 16.2% | 81.8% | 1.0% | 0.2% | 0.6% | 0.2% |
| Adm. | 7.4% | 5.3% | 85.8% | 0.5% | 0.9% | 0.1% |
| Edu. | 1.8% | 1.3% | 0.5% | 96.4% | 0.0% | 0.0% |
| Pub. | 6.9% | 2.3% | 0.7% | 1.0% | 88.4% | 0.7% |
| Ind. | 1.5% | 1.0% | 0.0% | 1.1% | 1.6% | 94.8% |

STAF-Net (OA=90.27%)

| Actual \ Predicted | Res. | Com. | Adm. | Edu. | Pub. | Ind. |
|---|---|---|---|---|---|---|
| Res. | 78.8% | 12.0% | 3.5% | 1.8% | 3.9% | 0.0% |
| Com. | 11.3% | 86.5% | 1.0% | 0.2% | 1.0% | 0.0% |
| Adm. | 7.1% | 3.3% | 89.6% | 0.0% | 0.0% | 0.0% |
| Edu. | 1.8% | 1.5% | 0.2% | 96.5% | 0.0% | 0.0% |
| Pub. | 3.0% | 1.2% | 0.0% | 1.0% | 94.1% | 0.7% |
| Ind. | 0.0% | 1.0% | 0.0% | 0.0% | 1.6% | 97.4% |

**Fig. 5.** Confusion matrix for the four models.

mobility flow time series led to the most significant improvement in classification accuracy. The classification results for commercial and public service types also showed significant improvements, with accuracies increasing from 71.3 % and 86.4 % to 86.5 % and 94.1 %, respectively. Overall, the proposed model effectively enhanced the classification results for different building function types after data fusion, particularly for residential, commercial, and public service types.

### *4.3. Urban building function mapping results*

In this study, the STAF-Net was applied to realize large-scale function classification mapping of buildings in 23 districts of Tokyo, over 93 thousand buildings were identified, as shown in Fig. 6. The results indicate that residential buildings are primarily concentrated on the western side of the 23 districts, particularly in Suginami and Nakano districts. Commercial areas are mainly located on the east side of Tokyo Station, spanning from Otemachi to Nihonbashi and Ginza areas, which serve as Tokyo's central business district where commercial activities are highly concentrated. Administrative buildings are predominantly found in the Kasumigaseki area of Chiyoda Ward and the western part of Shinjuku Station area. Both areas are critical political centers of Tokyo and Japan, hosting numerous government institutions and organizations. Educational buildings are mainly concentrated around point D, near the University of Tokyo in Bunkyo District, which is a renowned educational district of Tokyo. Additionally, point E, located near Shibuya Park, has a dense distribution of public service buildings due to the concentration of sports venues and tourist attractions. Point F in Koto District is characterized by industrial parks and logistics centers. The spatial distribution of residential and commercial buildings in the 23 districts of Tokyo highlights a clear separation of workplaces and residential areas.

To further understand the distribution of building functions in the 23 districts of Tokyo, this study conducted statistical analyses based on the number and area of buildings, as shown in Fig. 7. Based on the count of buildings, residential and commercial service types account for the majority in the 23 districts. Commercial business buildings come next. Public service buildings rank third, followed by education and research buildings, industrial buildings, and administrative buildings. When analyzed by area, residential and commercial service buildings still dominate. By comparing these two sets of data, we found that from the quantity ratio to the area ratio, only the land occupation of residential buildings decreased, while the proportion of other types increased. Notably, the proportion of industrial types increased the most, followed by commercial services. This is because residential buildings in Tokyo are predominantly single - family homes, occupying smaller land plots. In contrast, industrial buildings include factories and large logistics warehouses, and commercial buildings include various shopping centers and office buildings, all covering relatively large areas.

Whether looking at the results of the number of different types of buildings or the proportion of area, it is evident that residential and commercial buildings in 23 districts of Tokyo occupy a dominant position. This phenomenon not only highlights the high population density of the 23 districts but also reflects their significant status as an international financial city. The concurrent development of commercial and residential buildings not only meets the daily living needs of residents but also provides a solid foundation for the prosperity of commercial activities, driving economic and cultural development and progress in the 23 districts of Tokyo.

### *4.4. Urban functional structure mode in Tokyo's 23 districts*

To understand the core function of each district in the 23 districts of Tokyo, this study analyzed the proportion of different types of buildings based on their occupied area, with results displayed in Fig. 8.

In Suginami and Nakano districts, residential buildings occupy over 65 % of the building area, significantly higher than other areas, making them true residential districts. Chiyoda and Chuo districts have over 40 % of their area occupied by commercial buildings, indicating a significant concentration of commercial activities, positioning them as the commercial centers of the 23 districts. Chiyoda, being the administrative center of the 23 districts, has more than 9 % of its area occupied by administrative buildings, which is much higher than other areas. Bunkyo District leads in educational buildings, occupying 17.8 % of its area, which corresponds with its reputation as a famous school district in Tokyo and the district with the highest number of universities in Japan. Shibuya District has the highest proportion of public service buildings due to its many tourist attractions and recreational facilities, making it a popular destination for young people. Koto and Edogawa districts have higher proportions of industrial areas, 17.6 % and 12.4 %, respectively, because these districts are located near the numerous ports along Tokyo Bay, accommodating many logistics warehouses and factories.

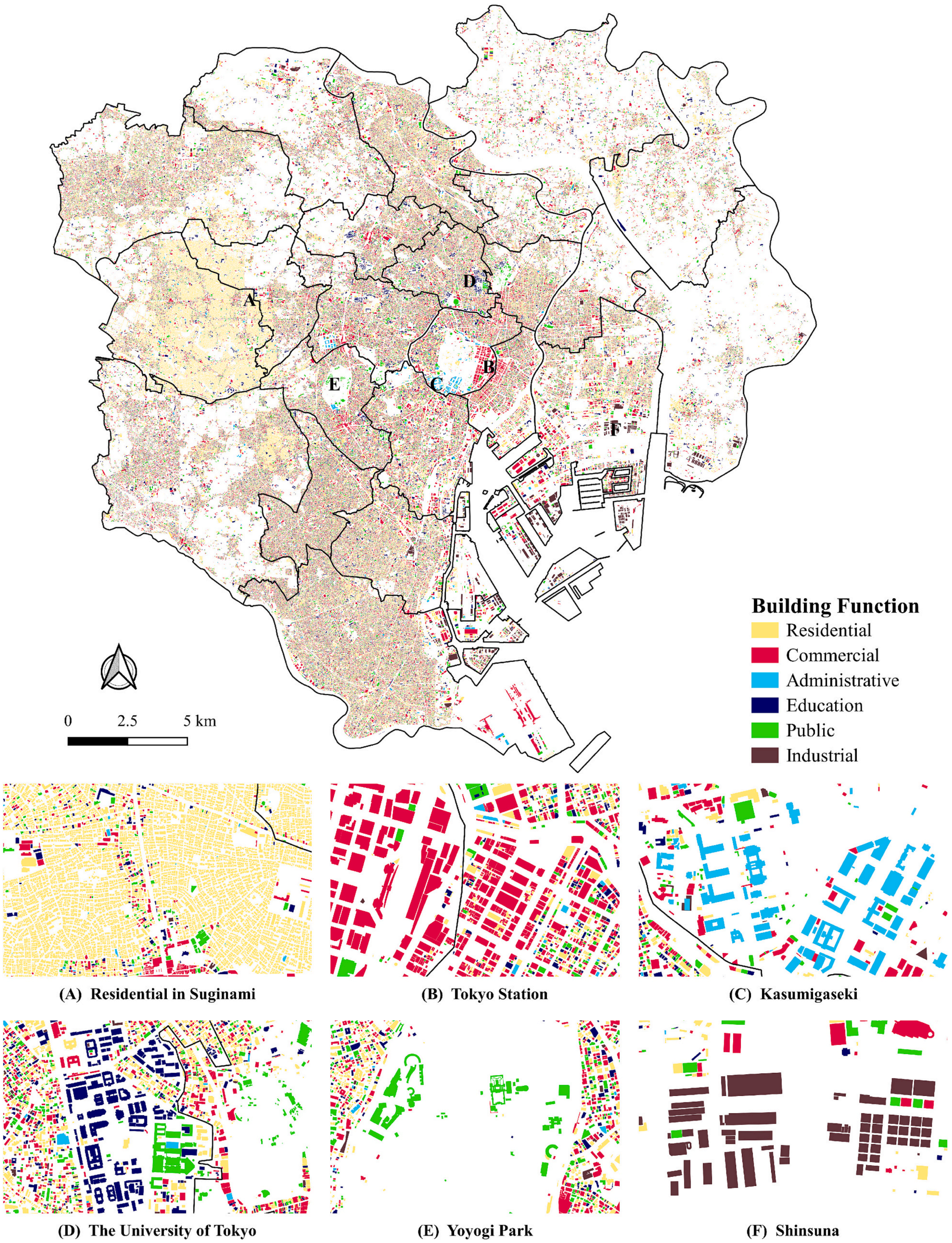
## 5. Discussion

### *5.1. Effectiveness of the proposed model*

In contrast to previous research, this study marks a significant breakthrough. Due to the fine granularity of building scales and the difficulties in extracting effective features from multi-source data, limited research has explored the relationships between human trajectory data and POI data for urban building function classification. To solve this problem, we delved into the structural differences between POI and human trajectory and introduced a deep learning model, STAF-Net. This model is capable of effectively extracting both the static POI semantics and the dynamic human mobility patterns within buildings. For the first time, we achieved large-scale building function mapping in the highly dense 23 wards of Tokyo, and the STAF-Net outperforms all existing models, boasting a test accuracy of 90.27 % and a Kappa of 0.8858. The building function mapping results from Tokyo show that our model can adapt to the accurate classification of massive building functions in megacities.

Based on the classification result of a single building, it can be observed that STAF-Net demonstrates a high accuracy in identifying buildings associated with industrial, educational, and public service categories. However, the classification accuracy for residential and commercial buildings is relatively lower despite having sufficient training samples. This discrepancy could stem from insufficient residential POI data and the presence of commercial POIs within residential areas, causing biases in model recognition. In summary, the results suggest that the human mobility patterns revealed by human trajectory data, along with the category semantics provided by POI data, can capture both the static and dynamic characteristics of buildings. This dual data source method addresses the shortcomings of using a single data source, thereby improving the precision of building function classification.

The necessity of each module in STAF-Net was verified through ablation analysis. Compared to the LSTM model, TimesNet more accurately captures both long and short-period patterns in human mobility flow, leading to higher classification accuracy. In the analysis of POI semantic features, semantic preservation-based POI embedding model significantly outperformed the earlier word2vec model, proving more effective for identifying building functions. Additionally, ablation analysis of the data sources revealed that different data sources have varying impacts on building function classification. POI data strongly indicate educational and administrative buildings, whereas human mobility flow time series significantly influence the classification of residential and commercial buildings. Altering the model structure yields minimal performance gains compared to data fusion. Therefore, future research should focus on extracting comprehensive features from multi-source data to enhance accuracy.



(A) Residential in Suginami

(B) Tokyo Station

(C) Kasumigaseki

(D) The University of Tokyo

(E) Yoyogi Park

(F) Shinsuna

**Fig. 6.** Building function classification results in 23 districts of Tokyo.

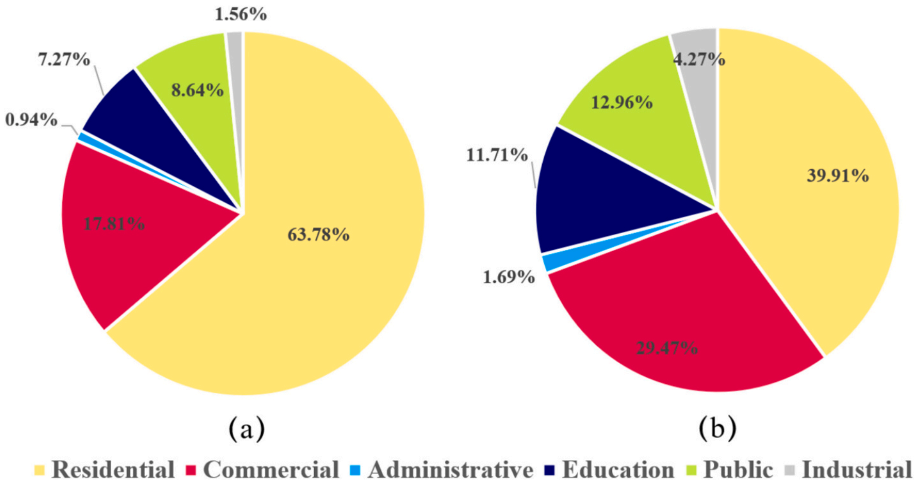


**Fig. 7.** Functional ratio of buildings in the 23 districts of Tokyo. (a) Number ratio (b) Area ratio.

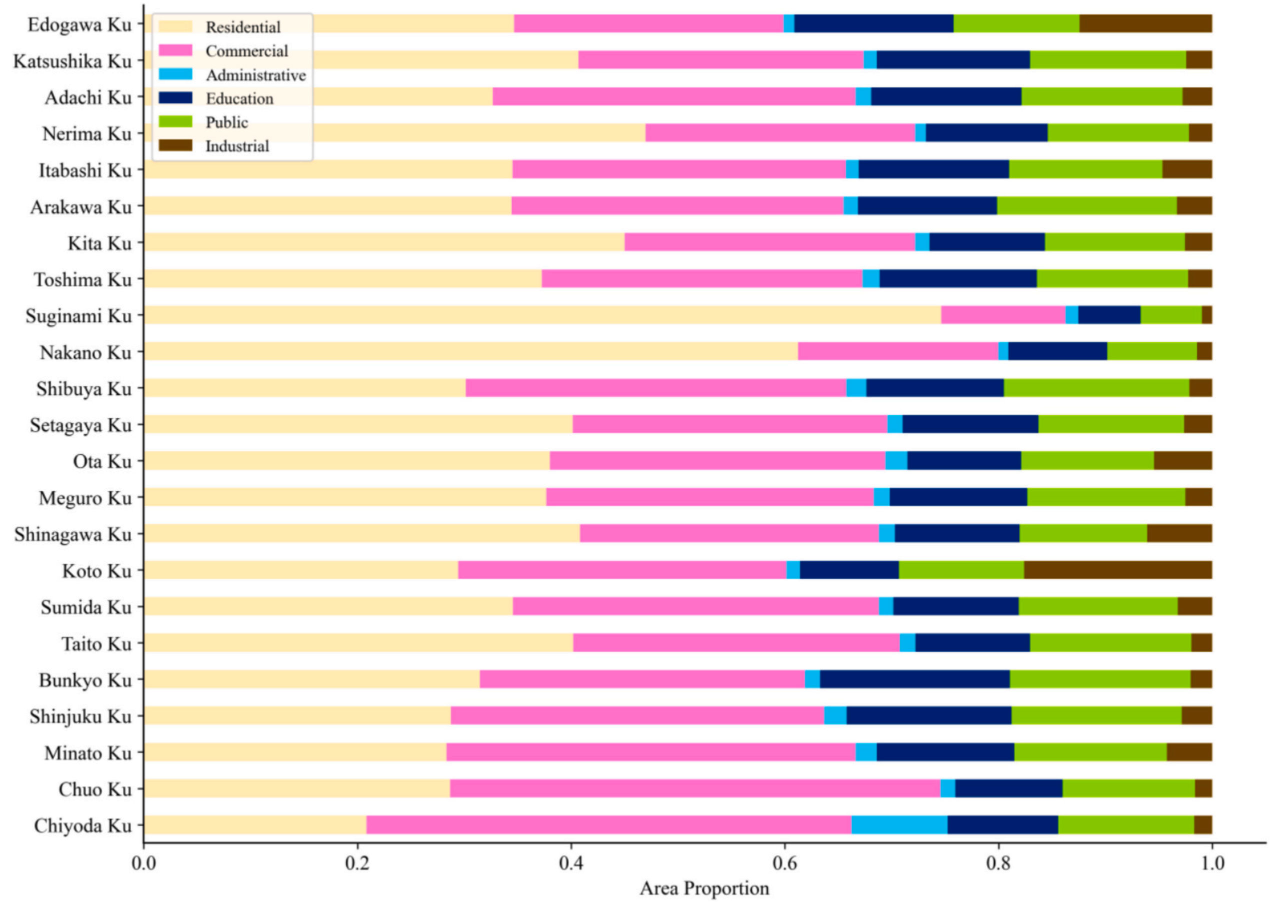


**Fig. 8.** Proportion of six types of building areas in 23 districts of Tokyo.

### 5.2. *Human trajectory data reflecting urban building function*

The time-series data of human mobility exhibits significant differences among various types of buildings, offering important clues for distinguishing building categories. As illustrated in Fig. 2, pedestrian traffic in residential buildings shows a pattern of high activity during the day and low activity at night, reflecting the regularity of residents' daily lives. In contrast, commercial buildings display an opposite trend, with dense foot traffic during working hours and sparse traffic outside these hours. This contrast not only reveals the fundamental differences in building usage but also provides valuable information for urban planning and management. Educational buildings exhibit unique short-term and long-term cyclical patterns, with three peaks occurring in the morning, midday, and evening, and a noticeable difference in trends between weekdays and weekends. This phenomenon reflects the rhythm of students' academic lives and suggests the temporal distribution characteristics of educational resource utilization. Public service buildings experience relatively lower pedestrian traffic on weekdays compared to weekends, highlighting the public's habits and needs related to the use of such facilities. For instance, venues like libraries and gyms typically attract more visitors on weekends, reflecting people's tendency to engage in self-improvement or leisure activities during their free time. Moreover, these variations in patterns provide a basis for assessing the service efficiency and social benefits of public facilities.

### *5.3. Policy discussion*

The approach proposed in this study offers transformative potential for urban planning. The STAF-Net model enables planning departments to accurately identify building functions at unprecedented scale and speed - a capability particularly valuable for Tokyo's complex urban fabric. The city-scale building function mapping result reveals the spatial structure characteristics of urban functions. The functions of urban buildings in Tokyo's 23 wards exhibit a clear separation pattern between occupation and residence: residential buildings are primarily concentrated in the outer ring districts of Suginami and Nakano, while commercial facilities are densely located in the central area from Otemachi to Ginza. Industrial buildings are predominantly situated in the port areas along the eastern coast. This phenomenon is mainly attributed to urban planning strategies that concentrate business districts in the city center to foster economic development, coupled with high housing prices and land costs that drive residential areas outward. Although the separation of work and residence benefits finance and manufacturing sectors (Lucas & Rossi-Hansberg, 2002), it also leads to long commutes, traffic congestion, and increased environmental burdens (Van Acker & Witlox, 2011; Zhao et al., 2011).

Based on these insights, our research can guide targeted planning interventions and inform urban renewal strategies. For areas with concentrated residential buildings, like Suginami and Nakano, our fine-grained mapping can identify specific communities lacking essential services. Planners can then use this information to conduct micro-level planning, such as strategically adding community parks and small-scale commercial complexes to meet residents' daily leisure and shopping needs. This approach improves the residents' quality of life and reduces their reliance on long-distance travel to central commercial areas. Similarly, in highly commercialized areas like the districts around Tokyo Station and Ginza, our model can highlight a lack of residential and public service buildings. This can prompt planners to consider adopting mixed-use development strategies, which is a core concept in urban renewal, by encouraging the integration of residential units, cultural facilities, and diverse services to promote a more balanced and vibrant urban life throughout the day and week.

### *5.4. Limitations and future works*

Several limitations of this should be mentioned. Firstly, in some regions, privacy laws and regulations may impose strict controls on the use of location data, which may affect the effectiveness of our proposed approach (e.g., reducing the classification accuracy of commercial and industrial types of buildings). Secondly, due to the challenges in acquiring building data labels, this study only classified six types of buildings. Further research should aim to refine classification labels to develop models capable of handling more detailed and diverse building functions. Recent research highlights the importance of interpretability in deep learning models. Consequently, future studies could focus on understanding the model's interpretability by examining how various data types specifically influence classification outcomes. This would contribute to increasing the model's transparency and credibility. In addition, exploring the relationship between geo-context features and building functions to improve the generalization of the model is another valuable research direction.

## 6. Conclusion

This study proposed a promising building function classification model (STAF-Net), which makes use of the semantic categories of POIs and human mobility patterns. Specifically, POIs and human mobility flow time series in each building are represented as vector embeddings that contain information about POI semantics and human mobility, respectively. The embeddings of the POIs and human mobility flow time series within a building are then fused to generate building embeddings using the multi-attention mechanism, which is aware of the different importance of the POIs and human mobility time series in a building. Finally, the building embedding is applied to identify building functions. The results in 23 districts of Tokyo show that the proposed model significantly outperforms the baseline, achieving a classification accuracy of 90.27 % and a Kappa coefficient of 0.8858.

The ablation experiment indicated that human trajectory data has advantages in classifying building functions by extracting the human mobility patterns within buildings. In this study, identifying residential and commercial buildings using human mobility data proved to be relatively straightforward. On the other hand, it turned out that using POI data made it easier to identify educational and administrative buildings. It is proved that data fusion can effectively improve the performance of building functional classification. Additionally, this study reveals significant correlations between human mobility patterns and urban building functions. Distinct temporal human mobility flows are exhibited by different types of buildings, providing the model with rich information on human activities that aid in identifying building functions.

This study offers a promising model for identifying the function of numerous buildings in megacities. Additionally, this study precisely identified the building functions in the 23 districts of Tokyo, Japan, which is valuable for the urban planning department in enhancing urban planning and promoting urban renewal. In the future, we will apply our approach to other cities in the world and aim to improve the model to identify more fine-grained or mixed types of buildings. Furthermore, we will introduce geo-contextual environmental features (e.g., distance from roads and distribution of buildings) to enhance the model's generalization performance in local areas.

## Data and codes availability statement

The codes and sample data to reproduce our work are publicly available at https://figshare.com/s/dae81f2d9bf7d14b6f42. We do not have the permission to share the human trajectory data and POI dataset used in the research. Readers can contact https://www.blogwatcher.co.jp/and https://www.zenrin-datacom.net/ for source data access. However, we have already provided a sample data used to reproduce our work are publicly available at https://figshare.com/s/dae81f2d9bf7d14b6f42.

## CRediT authorship contribution statement

**Zhihui Hu:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology. **Yao Yao:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Qia Zhu:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology. **Zijin Guo:** Writing – review & editing, Writing – original draft, Validation, Methodology, Formal analysis, Data curation. **Guicheng Li:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology. **Peiran Li:** Writing – review & editing, Validation, Methodology. **Renhe Jiang:** Writing – review & editing, Funding acquisition. **Junfang Gong:** Writing – review & editing, Supervision, Project administration. **Qingfeng Guan:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition. **Ryosuke Shibasaki:** Writing – review & editing, Validation, Conceptualization.

## Declaration of competing interest

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

## Acknowledgments

This work was supported by the National Natural Science Foundation of China [42471491, 42171466]; the Project Grant from the Co-creation Center for Disaster Resilience, IRIDeS, Tohoku University [ID: 2-QR001] and the "CUG Scholar" Scientific Research Funds at China University of Geosciences (Wuhan) [2022034].

## Data availability

I have already stated in the paper how to access the test data and code to facilitate replication by the readers.

## References

Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research, 13*(1), 1063–1095.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Cao, R., et al. (2020). Deep learning-based remote and social sensing data fusion for urban region function recognition. *ISPRS Journal of Photogrammetry and Remote Sensing, 163*, 82–97.

Chen, W., et al. (2020). Urban building type mapping using geospatial data: A case study of Beijing, China. *Remote Sensing, 12*(17), 2805.

Chen, Y., et al. (2017). Delineating urban functional areas with building-level social media data: A dynamic time warping (DTW) distance based k-medoids method. *Landscape and Urban Planning, 160*, 48–60.

Choi, S., & Yoon, S. (2023). Energy signature-based clustering using open data for urban building energy analysis toward carbon neutrality: A case study on electricity change under COVID-19. *Sustainable Cities and Society, 92*, Article 104471.

Deng, Y., et al. (2022). Identify urban building functions with multisource data: A case study in Guangzhou, China. *International Journal of Geographical Information Science, 36*(10), 2060–2085.

Fan, Z., et al. (2019). Decentralized attention-based personalized human mobility prediction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 3*(4), 1–26.

Feng, Y., et al. (2021). An SOE-based learning framework using multisource big data for identifying urban functional zones. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14*, 7336–7348.

Hao, H., et al. (2023). *Temporal convolutional attention-based network for sequence modeling*. arXiv preprint arXiv:2002.12530.

He, K., et al. (2016). Deep residual learning for image recognition. In *2016 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 770–778).

Ho, Y., & Wookey, S. (2020). The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE Access, 8*, 4806–4813.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*, 1735–1780.

Hoffmann, E. J., et al. (2023). Using social media images for building function classification. *Cities, 133*, Article 104107.

Hua, Y., et al. (2019). Deep learning with long short-term memory for time series prediction. *IEEE Communications Magazine, 57*, 114–119.

Huang, W., et al. (2022). Estimating urban functional distributions with semantics preserved POI embedding. *International Journal of Geographical Information Science, 36*(10), 1905–1930.

Humtsoe, T. Y. (2022). Travel mode choice in the North-eastern Indian City of Kohima: Lessons from empirical study. *Journal of Urbanism: International Research on Placemaking and Urban Sustainability*, 1–15.

Ismail Fawaz, H., et al. (2019). Deep learning for time series classification: A review. *Data Mining and Knowledge Discovery, 33*, 917–963.

Jin, Y., et al. (2023). Understanding railway usage behavior with ten million GPS records. *Cities, 133*, Article 104117.

Kingma, D. P., & Ba, J. (2017). *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980.

Li, J., et al. (2022). Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal of Applied Earth Observation and Geoinformation, 112*, Article 102926.

Liu, S., & Shi, Q. (2020). Local climate zone mapping as remote sensing scene classification using deep learning: A case study of metropolitan China. *ISPRS Journal of Photogrammetry and Remote Sensing, 164*, 229–242.

Liu, S., et al. (2022). Concordance between regional functions and mobility features using bike-sharing and land-use data near metro stations. *Sustainable Cities and Society, 84*, Article 104010.

Liu, X., et al. (2018). Characterizing mixed-use buildings based on multi-source big data. *International Journal of Geographical Information Science, 32*(4), 738–756.

Liu, Y., et al. (2015). Social sensing: A new approach to understanding our socio-economic environments. *Annals of the Association of American Geographers, 105*(3), 512–530.

Lucas, R. E., & Rossi-Hansberg, E. (2002). On the internal structure of cities. *Econometrica, 70*(4), 1445–1476.

Mikolov, T., et al. (2013). *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781.

Moreira, D., et al. (2019). Multimodal data fusion for sensitive scene localization. *Information Fusion., 45*, 307–323.

Niu, N., et al. (2017). Integrating multi-source big data to infer building functions. *International Journal of Geographical Information Science, 31*(9), 1–20.

Shen, P., Liu, J., & Wang, M. (2021). Fast generation of microclimate weather data for building simulation under heat island using map capturing and clustering technique. *Sustainable Cities and Society, 71*, Article 102954.

Shi, X., et al. (2015). *Convolutional LSTM network: A machine learning approach for precipitation nowcasting*. arXiv preprint arXiv:1506.04214.

Srivastava, S., et al. (2018). Fine-grained landuse characterization using ground-based pictures: A deep learning solution based on globally available data. *International Journal of Geographical Information Science, 34*(6), 1117–1136.

Tong, X.-Y., et al. (2020). Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment, 237*, Article 111322.

Van Acker, V., & Witlox, F. (2011). Commuting trips within tours: How is commuting related to land use? *Transportation, 38*, 465–486.

Vinyals, O., et al. (2016). *Order matters: Sequence to sequence for sets*. arXiv preprint arXiv: 1511.06391.

Wang, C., et al. (2020). Dynamic occupant density models of commercial buildings for urban energy simulation. *Building and Environment, 169*, Article 106549.

Wang, Y., et al. (2023). A review of regional and Global scale Land Use/Land Cover (LULC) mapping products generated from satellite remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing, 206*, 311–334.

Wei, C., & Yu, W. (2024). A spatial dependency based reinforcement learning model for selecting features in spatial classification. *GeoInformatica*, 1–29.

Wu, H., et al. (2023). *TimesNet: Temporal 2D-variation modeling for general time series analysis*. arXiv preprint arXiv:2210.02186.

Yao, Y., et al. (2017). Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. *International Journal of Geographical Information Science, 31*(4), 825–848.

Yao, Y., et al. (2022). Classifying land-use patterns by integrating time-series electricity data and high-spatial resolution remote sensing imagery. *International Journal of Applied Earth Observation and Geoinformation, 106*, Article 102664.

Yao, Y., et al. (2023). Predicting mobile users' next location using the semantically enriched geo-embedding model and the multilayer attention mechanism. *Computers, Environment and Urban Systems, 104*, Article 102009.

Zhai, W., et al. (2019). Beyond Word2vec: An approach for urban functional region extraction and identification by combining Place2vec and POIs. *Computers, Environment and Urban Systems, 74*, 1–12.

Zhang, X., et al. (2018). Integrating bottom-up classification and top-down feedback for improving urban land-cover and functional-zone mapping. *Remote Sensing of Environment, 212*, 231–248.

Zhang, X., et al. (2023). Inferring building function: A novel geo-aware neural network supporting building-level function classification. *Sustainable Cities and Society, 89*, Article 104349.

Zhao, P., Lü, B., & De Roo, G. (2011). Impact of the jobs-housing balance on urban commuting in Beijing in the transformation era. *Journal of Transport Geography, 19*(1), 59–69.

Zhiwen, Z., et al. (2023). Assessing the continuous causal responses of typhoon-related weather on human mobility: An empirical study in Japan. In *Proceedings of the 32nd ACM international conference on information and knowledge management* (pp. 3524–3533).

Zhong, C., et al. (2014). Inferring building functions from a probabilistic model using public transportation data. *Computers, Environment and Urban System, 48*, 124–137.

Zhuo, L., et al. (2019). Identifying building functions from the spatiotemporal population density and the interactions of people among buildings. *ISPRS International Journal of Geo-Information, 8*(6), 247.