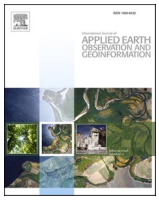
International Journal of Applied Earth Observation and Geoinformation 129 (2024) 103805

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

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



A multimodal data fusion model for accurate and interpretable urban land use mapping with uncertainty analysis

Xiaoqin Yan [a,b], Zhangwei Jiang [c], Peng Luo [d], Hao Wu [a], Anning Dong [a], Fengling Mao [c], Ziyin Wang [c], Hong Liu [c], Yao Yao [a,e,*]

[a] School of Geography and Information Engineering, China University of Geosciences, Wuhan 430078, Hubei Province, China
[b] Institute of Remote Sensing and Geographic Information Systems, School of Earth and Space Sciences, Peking University, Beijing 100871, China
[c] Alibaba Group, Hangzhou 311121, Zhejiang Province, China
[d] Chair of Cartography and Visual Analytics, Technical University of Munich, Munich, Germany
[e] Center for Spatial Information Science, The University of Tokyo, Chiba 277-8568, Japan

ARTICLE INFO

Keywords:
Urban land use mapping
Multimodal data fusion
Uncertainty analysis
Feature extraction
Deep learning

ABSTRACT

Urban land use patterns can be more accurately mapped by fusing multimodal data. However, many studies only consider socioeconomic and physical attributes within land parcels, neglecting spatial interaction and uncertainty caused by multimodal data. To address these issues, we constructed a multimodal data fusion model (MDFNet) to extract natural physical, socioeconomic, and spatial connectivity ancillary information from multimodal data. We also established an uncertainty analysis framework based on a generalized additive model and learnable weight module to explain data-driven uncertainty. Shenzhen was chosen as the demonstration area. The results demonstrated the effectiveness of the proposed method, with a test accuracy of 0.882 and a Kappa of 0.858. Uncertainty analysis indicated the contributions in overall task of 0.361, 0.308, and 0.232 for remote sensing, social sensing, and taxi trajectory data, respectively. The study also illuminates the collaborative mechanism of multimodal data in various land use categories, offering an accurate and interpretable method for mapping urban distribution patterns.

1. Introduction

Urban land use involves organizing and allocating land for various functions to support development and policy making of city (Duranton and Puga, 2015, Gong et al., 2020). Influenced by human activities, urban land use has evolved into various forms including residential, commercial, and public service facilities (Zhang et al., 2017, Xia et al., 2020,). As urbanization accelerates, land uses have become more diverse and complex, and their interactions become more intricate (Srivastava et al., 2019, Koroso et al., 2021). Thus, it becomes paramount to devise an efficient and automated methodology for identifying land use patterns, serving urban planning, and promoting sustainable development.

Deep learning techniques have been shown to recognize urban land use patterns efficiently and robustly by interpreting remote sensing imagery (Zhu et al., 2017, Alzubaidi et al., 2021). For example, Maggiori et al. (2016) constructed an end-to-end convolutional neural network (CNN) framework, which proved capable of handling large-scale remote sensing imagery and data classification at the micro-scale level. To improve the model feature extraction capabilities, Liu and Shi (2020) built a feature attention-based CNN for urban spatial distribution pattern recognition. Although deep learning effectively extracts image attributes for improved classification accuracy, the visual similarities of urban land use challenge accurate identification using only remote sensing imagery (Zhou et al., 2020). This challenge points to the necessity of integrating other attributes to improve accuracy.

The emergence of social sensing data, such as Points of Interest (Xu et al., 2023, Yao et al., 2023a, 2023b), social media data (Huang et al., 2018, Lyu and Zhang, 2019) and municipal service data (Guan et al., 2021), provides a novel perspective for studying urban spatial patterns. By reflecting socioeconomic attributes, social sensing data can compensate for the limitations of remote sensing imagery in identifying

* Corresponding author.
E-mail addresses: yanxiaoqin@stu.pku.edu.cn (X. Yan), zhangwei.jzw@alibaba-inc.com (Z. Jiang), peng.luo@tum.de (P. Luo), wuh@cug.edu.cn (H. Wu), donganning@cug.edu.cn (A. Dong), fengling.mfl@alibaba-inc.com (F. Mao), wangziyin.wzy@alibaba-inc.com (Z. Wang), liuhong.liu@alibaba-inc.com (H. Liu), yaoy@cug.edu.cn, yaoy@csis.u-tokyo.ac.jp (Y. Yao).

https://doi.org/10.1016/j.jag.2024.103805
Received 20 September 2023; Received in revised form 11 March 2024; Accepted 29 March 2024

urban land use (Yin et al., 2021, Xing et al., 2024). Yao et al. (2022) combined fine-grained temporal electricity data and remote sensing imagery for improved land use distribution patterns. Bai et al. (2023) further proposed an unsupervised multimodal geographic representation learning framework, enabling geographical mapping through effective representations of natural and socioeconomic aspects. These studies effectively identify land use by merging remote sensing and social sensing data. Nevertheless, they focus on fusing multiple attributes within land parcels while neglecting external spatial connections. Regions are interconnected through social activities and contribute significant information on spatial interactions (Liu et al., 2016). Understanding these interactions is essential for accurate land use identification, as it allows us to capture the underlying dynamics and interdependencies more effectively between regions.

The development of location acquisition technology has generated massive amounts of trajectory data, which can be used to understand urban distribution patterns by reflecting the spatial mobility of objects (Zheng, 2015, Park et al., 2020). With wide coverage and high collection frequency, taxi trajectory data has become a preferred data for analyzing urban spatial interaction patterns (Zhang et al., 2022, Hu et al., 2023). Moreover, graph convolutional network (GCN), combining graph learning structures and CNN, helps analyze spatial connections between land parcels (Xu et al., 2022, Yin et al., 2023, Guan et al., 2024). For example, Hu et al. (2021) built a road trajectory corpus to obtain road segment semantic embeddings and employed GCN models to recognize urban function. Zhu et al. (2020) introduced GCN to model locations as graphs and capture knowledge of geographic environments by optimizing network weights, which demonstrated the effectiveness of identifying land use by capturing spatial relationships. Taxi trajectory data can effectively characterize intricate mobility patterns and temporal dynamics. Fusing it with other multi-source data to construct a multimodal model for comprehensively identifying land use patterns is a worthwhile exploration.

Uncertainty in land use identification based on deep learning is caused by differences in model structure, study scale and data sources (Kwan and Weber, 2008, Zhong et al., 2020). Most studies focus on model structure or Modifiable Area Units Problem (MAUP), lacking quantitative analysis of data source uncertainty. Multimodal data represent different deep semantic features, such as remote sensing imagery identifying visually prominent categories, while social sensing data better distinguishes categories with significant social activities (Du et al., 2020, Feng et al., 2021). Thus, it is crucial to concentrate on data-driven uncertainty analysis. In deep learning, interpretability allows understanding a model decision-making process (Doshi-Velez and Kim, 2017). Ante-hoc methods involve constructing easily understandable models (Sokol and Flach, 2020). For example, Zhang et al. (2019) developed a generalized additive model (GAM) suitable for multiclass tasks to increase model interpretability. Post-hoc methods use interpretability techniques to explain trained models (Murdoch et al., 2019). Lu et al. (2022) used attention module visualization to quantify the contribution of multimodal data to the results. These interpretability techniques can provide technical support for analyzing uncertainty based on data-driven land use identification.

In conclusion, the reviewed literature reveals two outstanding issues yet to be addressed. Firstly, although multimodal data fusion has proven effective in identifying urban land use types, most research only considers attributes within land parcels, neglecting spatial connectivity. This oversight limits our understanding of dynamic urban interactions, highlighting the need for integrating trajectory data for a comprehensive perception of land use attributes. Secondly, investigating the mechanisms of data sources uncertainty in land use identification through a quantitative analysis, remains a research gap. Without a detailed quantification of these uncertainties, identification accuracy will be limited by the unavailability of formulating an effective data fusion solution. Interpretability technologies provide technical support for understanding the relationship between data and result uncertainty through quantification and visualization.

Therefore, we propose a deep learning-based land use identification framework that aims to overcome the barriers of previous research by comprehensively characterizing land attributes and exploring data-driven uncertainty mechanisms. This framework consists of two parts: land use mapping model and uncertainty analysis. In the first part, multimodal data fusion model (MDFNet) employs three network branches to extract physical information from high-resolution remote sensing imagery (HSR), socioeconomic information from real-time Tencent user density data (RTUD), and spatial interaction information from taxi trajectory data (TTD). MDFNet is designed to further integrate spatial connectivity attributes to improve identification accuracy. A learnable weight module is also incorporated to adaptively assign weights to these attributes for efficient fusion. In the second part, we employ GAM and learnable weight module to quantitatively analyze the uncertainty of multimodal data. This analysis provides a comprehensively understanding the significance of multimodal data in the overall task and in specific land use types, thereby enhancing the reliability and interpretability of model. This research contributions are as follows:

1) We construct a multimodal data fusion model (MDFNet) for land use identification. The model can adaptively learn and adjust multimodal data weights to achieve effective data fusion for efficient land use mapping, addressing the specific inadequacy in current methodologies regarding spatial interactions in data fusion.
2) By developing a framework for uncertainty analysis based on multimodal data, our study not only elucidates the contribution of different data, but also establishes a novel and interpretable mechanism to fill the previous research gap. This mechanism aids in precisely identifying the strengths and limitations of multisource data, enhancing data fusion techniques, and contributing to more accurate and reliable urban land use predictions.

The remainder of this paper is as follows: Section 2 describes the study area and data. Sections 3 and Section 4 describe the research methodology and experimental results, respectively. In Section 5, we discuss the results in depth and present the study limitations. Finally, the study is concluded in Section 6.

2. Study area and data

Shenzhen, Guangdong Province (Fig. 1) is the first special economic zone in China, and one of the four central cities in the Guangdong-Hong Kong-Macao Greater Bay Area (GBA). The city includes 10 districts with a total area of 1997.47 km^2, a resident population of 17,681,600 and a regional GDP of 3,460,640 million yuan (https://tjj.sz.gov.cn/). As a typical fast-growing city, Shenzhen's spatial pattern has changed greatly in the last fifty years, resulting in various land use categories such as residential, commercial, industrial and warehouse (Lei et al., 2021). Therefore, by taking Shenzhen as a typical study area and exploring the spatial structure and development patterns, providing guidance for optimizing urban structure and sustainable development.

As shown in Table 1, three types of datasets were applied as input data in this study, which includes HSR, RTUD and TTD. The study employs a 480 m × 480 m grid for experiments (Liu and Shi, 2020), cropping HSR and RTUD into equally sized image blocks for model input. For TTD, a graph network matrix is built based on taxi pick-up and drop-off locations, with interaction counts as edge weights. Average speed, pick-up, and drop-off counts within parcels serve as node attributes for model input.

Following the "Code for classification of urban land use and planning standards of development land (GB50137-2011)", we initially eliminated areas such as lakes, bare and cropland that were not relevant to our study. Then, the samples were preliminarily labeled based on the actual land use conditions of parcels in Shenzhen. Given the irregular sizes of the parcel, we validated and adjusted each sample using HSR

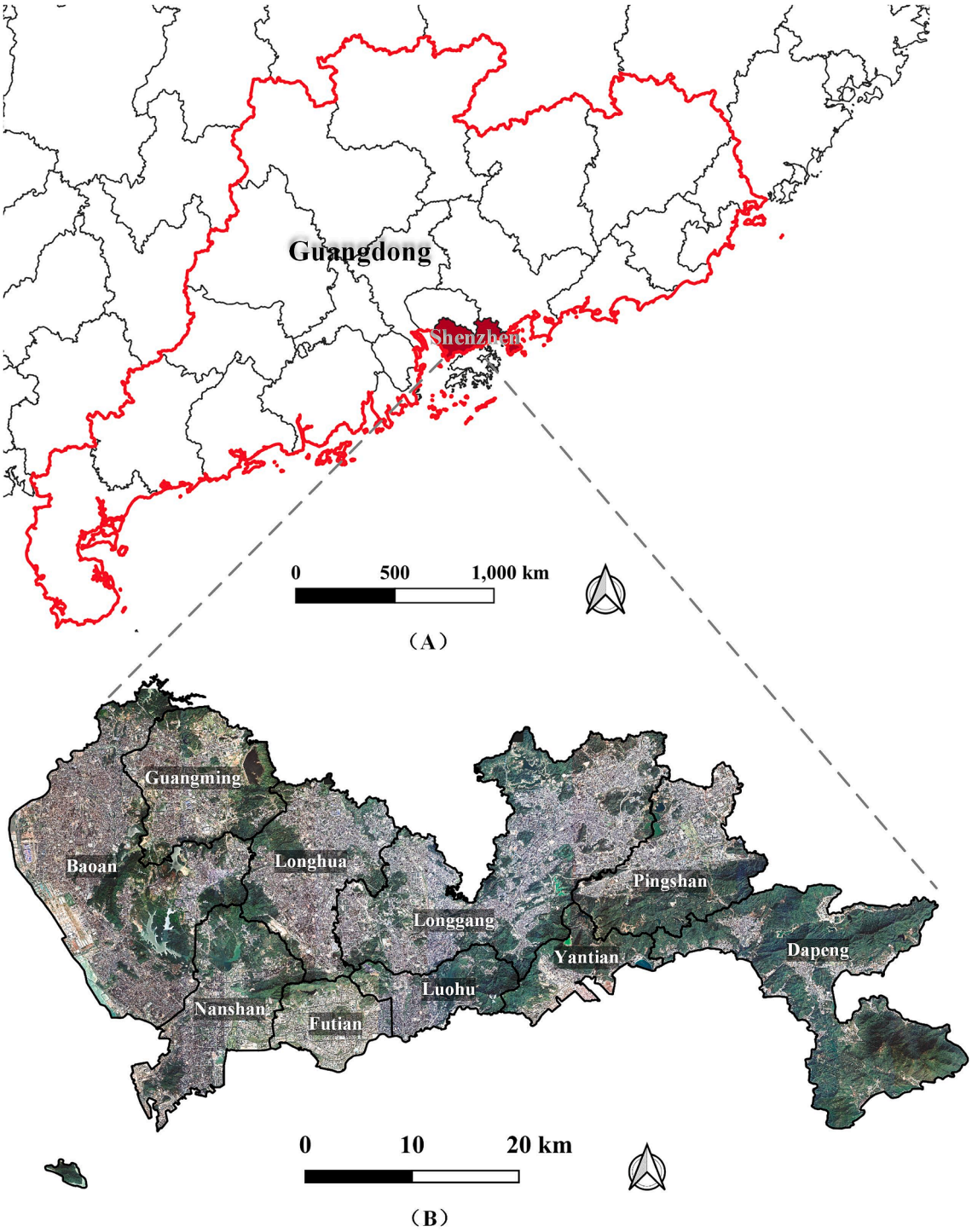


Fig. 1. Study area: (B) Shenzhen, located in (A) Guangdong Province.

and expert knowledge (Yao et al., 2022). Finally, we have processed and categorized 4464 samples across six categories: residential (1056), public service (800), commercial (316), industrial (656), warehouse (520), and green lands (1116). After sample labeling, we randomly assign 80 % of samples to training, and the remaining 20 % to testing.

3. Methodology

The flowchart of the proposed method is illustrated in Fig. 2. This method employs a deep learning-based multimodal feature fusion network for urban land use mapping. The research process includes three main parts: (1) constructing a multimodal data fusion model, selecting specific branch networks to extract multiple attributes from multimodal data for land use identification, and evaluating the model effectiveness; (2) quantitatively examining the influence mechanism of multimodal data on results using an uncertainty analysis framework; (3) mapping land use and analyzing spatial pattern of the study area based on the model results.

3.1. The construction of land use identification model

The overall structure of the proposed MDFNet model is shown in the

Table 1
An overview of datasets used for this study.

Category	Source	Resolution	Year	Description
HSR	Google Earth (https://earth.google.com)	2 *m*	2019	Includes RGB bands, used for extracting physical information.
RTUD	Tencent (https://heat.qq.com)	27.05 m, hourly	2019	Includes workday, weekend and holiday, and records user location, used for extracting socioeconomic information.
TTD	Transport Commission of Shenzhen	10 s	January 10–16, 2016	Includes ID, recording time, location, and passenger status, used for learning spatial interaction.

red dashed box in Fig. 2. Considering the heterogeneity of the multimodal data, we are focusing on using three deep learning representation learning to extract socioeconomic, building physical and spatial connectivity auxiliary information from the multimodal data of each grid. This information will be used to perform land use mapping tasks.

Taking public service land as an example, we use Temporal Convolutional Networks (TCN) and Bi-directional Long Short-Term Memory (BiLSTM) to extract socioeconomic information from RTUD, revealing the land patterns during different periods such as workday and weekend. ResNet-based model is employed to extract unique spatial distribution and physical building information of public service from HSR. Graph Convolutional Networks (GCN) are utilized to capture the spatial relationships between public service and others from TTD. Attributes from multimodal data are combined into a vector and updated via a learnable module, capturing salient public service features. These are processed through a fully connected and SoftMax layer for accurate results.

3.1.1. *Socioeconomic feature extraction based on TCN and BiLSTM*

Considering that RTUD includes multiple time periods, this study uses TCN and BiLSTM modules to extract temporal features in different view from RTUD. TCN is suitable for shorter sequence periods in RTUD by its robust feature extraction (Bai et al., 2018). As shown in Fig. 3, the 1-D convolutional network in TCN captures the hourly fluctuations in RTUD, and its dilated causal convolution ensures outputs depend only on past hours, preventing future information leakage. Furthermore, a structure like residual blocks is implemented to alleviate the issue of gradient vanishing. For the one-dimensional RTUD input data $x \in R^{24}$ and filter $f : \{0, \cdots, k-1\} \rightarrow R$, the convolution operation F pertaining to sequence s is as follows:

$$F(s) = (X *_{d} f)(s) = \sum_{i=0}^{k-1} f(i) * X_{s-di} \quad (1)$$

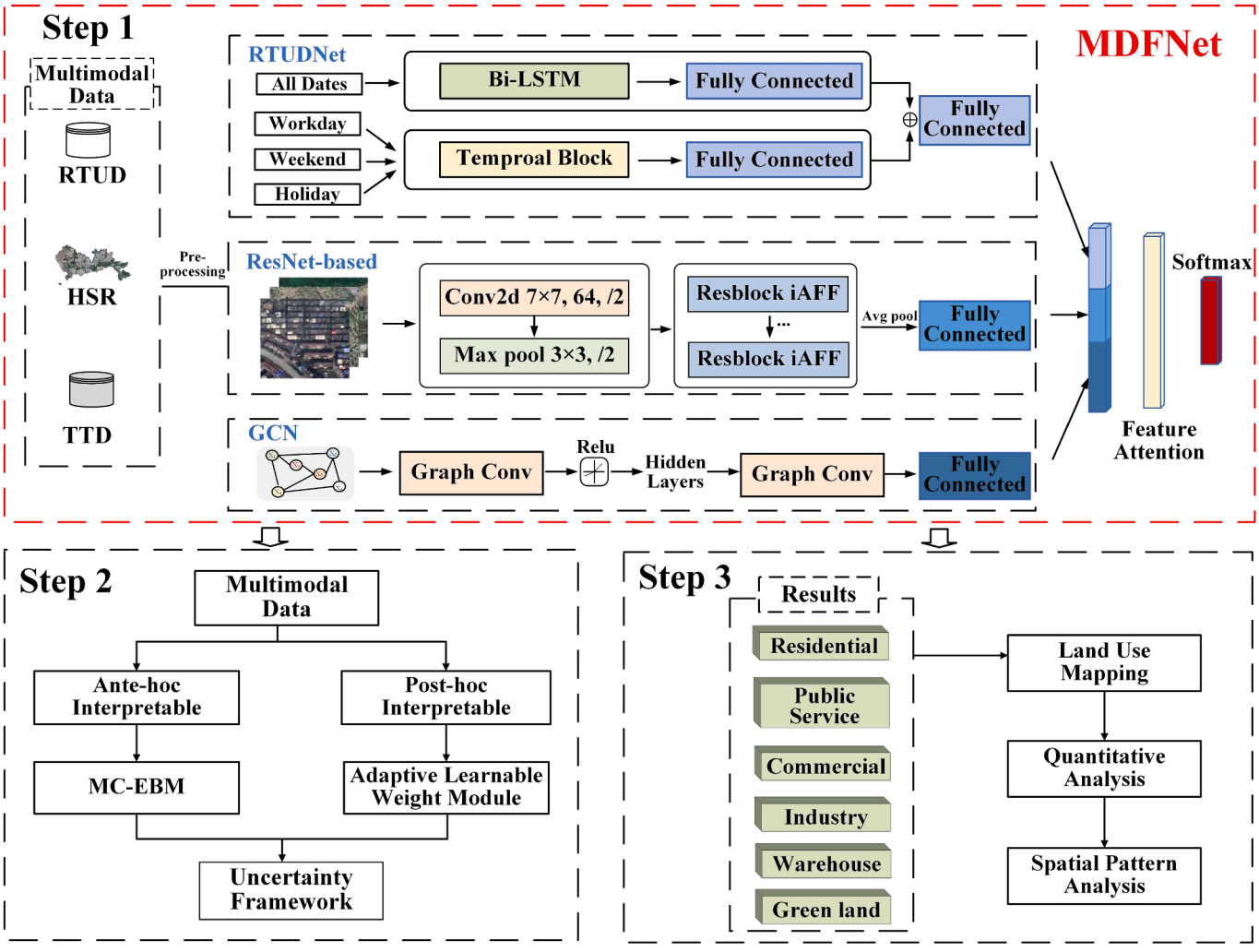


Fig. 2. The workflow of the proposed method.

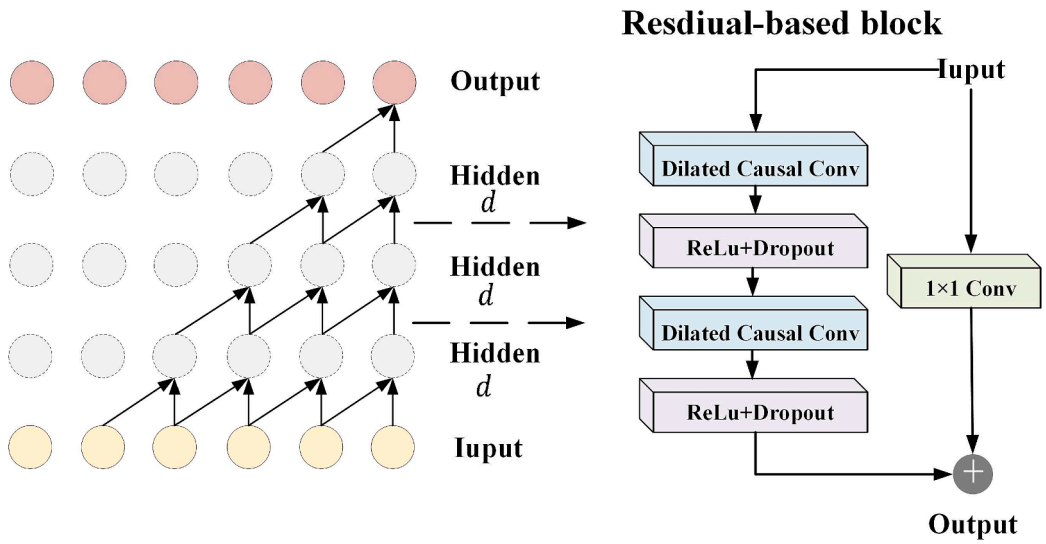


Fig. 3. Temporal Convolutional Neural Network Framework in MDFNet.

d denotes the dilation factor, k denotes the filter size, and $s-di$ indicates the past-oriented direction. By fine-tuning k and d control fields of perception, RTUD features can be extracted more accurately.

LSTM consists of input, forget and output gates, which can be used to reveal the long-term dependencies of RTUDs in different time periods by retaining crucial temporal information while discarding irrelevant noise (Wang et al., 2020). To further enhance the ability to capture context from both past and future time steps, we employ a BiLSTM architecture, which extends the standard LSTM by processing RTUD in both forward and backward directions. The RTUD is processed through both BiLSTM and TCN, and the fully connected layers from both models are combined. This approach leverages the strengths of both models to efficiently extract the temporal features of RTUD.

3.1.2. *Physical feature extraction based on ResNet and attention mechanism*

To extract high-dimensional building physical attributes from HSR, this study employs the ResNet model (He et al., 2016), which has effective performance and broad adoption, and embeds an attention mechanism. To prevent overfitting, this study employs a Basic block based ResNet for HSR attributes extraction. The core of ResNet lies in residual blocks, which facilitate connections between layers, ensuring that detailed features are effectively propagated to deeper layers, leading to the extraction of more comprehensive ancillary information of HSR.

To delve deeper into the architectural details of HSR, this study employs the Multi-Scale Channel Attention Module (MS-CAM) proposed by Dai et al. (2021). MS-CAM (Fig. 4) realizes multi-channel attention through varying spatial pooling sizes, and the iterative Attention Feature Fusion (iAFF) mechanism to address initial building physical attributes aggregation issues in HSR. Given the global channel context $G(X)$ and local context information $L(X)$ features obtained via MS-CAM are:

$$X' = X \otimes M(X) = X \otimes \sigma(L(X) \oplus G(X)) \tag{2}$$

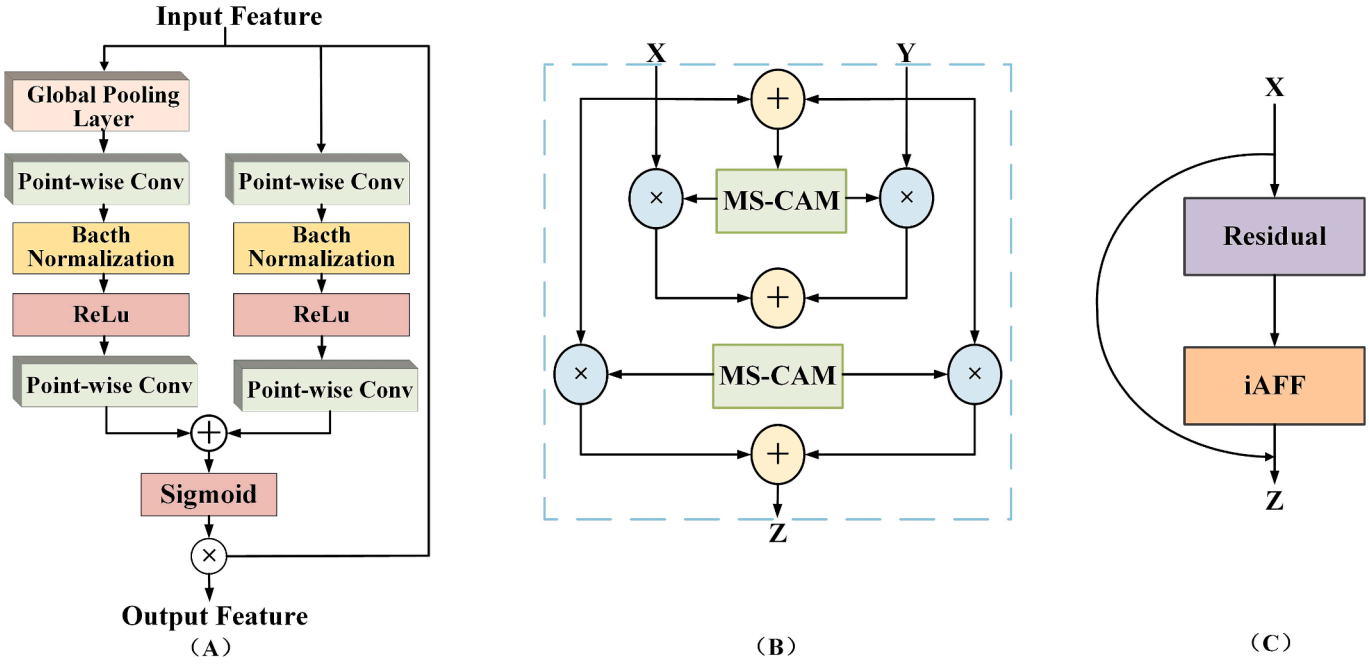


Fig. 4. The proposed (A) MC-CAM, (B) iAFF module, (C) iAFF embedded to ResNet model.

in ResNet, Given $X, Y \in R^{C\times H\times W}$ where X represents the identity mapping in ResNet and Y corresponds to the residual block. Based on MS-CAM, iAFF can be expressed as:

$$X \uplus Y = M(X+Y) \otimes X + (1 - M(X+Y)) \otimes Y \tag{3}$$

$Z \in R^{C\times H\times W}$ denotes fused features, $\uplus$ indicates initial feature fusion. The iAFF mechanism addresses discrepancies in semantics and scales among features, effectively extracting class-specific information from complex HSR.

3.1.3. Spatial connectivity feature extraction based on graph neural network

The ability of GCN to efficiently capture non-Euclidean spatial data attributes makes them an attractive tool for extracting and interpreting TTD for land use identification (Wu et al., 2020). This study designs a multilayer GCN to extract grid features and spatial connectivity in graph structures. In the urban context P, P_i denotes each parcel,C represents the spatial adjacency matrix. A graph network $G = (V, E^{(C)})$ is constructed based on TTD. Each P_i is represented by node $v_i \in V$ in C, while $e_{ij} = (v_i, v_j, a_{ij}) \in E^{(c)}$ denoting the edge between nodes, a_{ij} is the edge weight. During GCN training, let $\widetilde{X}$ be the node feature matrix. Regional node features and edge connections are represented by $\widetilde{E} \in R^{n\times n}$, the matrix form of spatial connection weights. The forward propagation between layer l and l+1 is:$\widetilde{X}^{l+1} = h\left(\widetilde{X}^{l}, \widetilde{E}\right) = \sigma\left(\widetilde{D}^{-\frac{1}{2}}\widetilde{E}\widetilde{D}^{-\frac{1}{2}}\widetilde{X}^{l}W^{l}\right)$(4)

where $\widetilde{D}_{ii} = \sum_j \widetilde{E}_{ij}$ represents the diagonal matrix, W^l the layer weight in the neural network, and σ the ReLu activation function for non-linear feature extraction. $\widetilde{D}^{-\frac{1}{2}}\widetilde{E}\widetilde{D}^{-\frac{1}{2}}$ constitutes the normalized Laplacian matrix, indicating spatial connectivity among parcels. By leveraging TTD through GCN, we extract spatial connectivity attributes for land use grid. This process demonstrates a reliable method for revealing complex spatial relationships to identify land use.

3.2. A framework for uncertainty analysis based on multimodal data

The deep semantic differences in multimodal data contribute variably to uncertainty in results, and the model provide results based on input without clarifying the decision-making mechanism. This study employs interpretability techniques, analyzing the complex relationship between input data and results from both ante-hoc and post-hoc perspectives, quantifying the influence of different data sources on land use identification. In ante-hoc, a multi-class explainable boosting machine (MC-EBM) calculates the contributions of different data sources. In post-hoc, updated feature vectors from learnable weight modules are extracted to further analyze contributions of various data sources across land use categories.

3.2.1. Multi classification generalized additive model for ante-hoc

Explainable boosting machine (EBM) is an interpretable generalized additive model capable of quantitatively analyzing the impact of input features on results (Nori et al., 2019). The formal representation is as follows:

$$g(E[y]) = \beta_0 + \sum f_j(x_j) + \sum f_{i,j}(x_i, x_j) \tag{5}$$

where g is a monotonically differentiable link function, β_0 is the intercept term, and f_j is the feature function for each feature x_j, the EBM enhances accuracy and interpretability by detecting feature interactions. In this study, we apply EBM to multi-class task (MC-EBM) and employ the cyclic gradient boosting algorithm to iteratively learn each feature function f_j. Each boosting step fits the base learner to the pseudo-residual space, and then adds it to the ensemble, effectively performing an approximate gradient computation in the function space. This results in the following MC-EBM formulation:

$$f_{ik}^{+} = f_{ik} + \eta \sum_{l\in[L]} \gamma_{ilk} 1_{x_i \in R_{il}} \tag{6}$$

where $i \in [d], k \in [k], l \in [L]$, R_{il} denotes the training point set for feature i at leaf node l, γ_{ilk} denotes pseudo-residuals in a multi-class setting. Using cyclic gradient boosting and maximum variance reduction, the base learner bagging quantity B is set, and leaf count L is constrained. This iterative process results in the multi-class cyclic boosting algorithm.

3.2.2. Adaptive learnable weight module for post-hoc

Post-hoc refers to the interpretability of a model after training, aids in understanding the model's inner workings (Du et al., 2019). Given the varying contributions of different data sources, it is critical to adaptively increase and visualize data weights. We propose a learnable weight module that uses concatenated feature vectors from three branch networks for updating. As shown in Fig. 5, these vectors are first reduced via a convolutional layer to eliminate redundancy. Average and max pooling are then applied to calculate feature weights $W_i \in [W_{RTUD}, W_{HSR}, W_{TTD}]$, symbolizing land use recognition capabilities. By multiplying these weights with the original features, we update the multimodal data features for output and further model optimization.

To analyze the collaborative mechanism of multimodal data sources in land use classification, we normalize feature importance from different sources using a learnable weight module. We use variance (V) and mean (E) of feature values as metrics to assess each data source's contribution to the results.

3.3. Analysis the land use distribution pattern

The results obtained from MDFNet were utilized to conduct land use mapping in the study area. Firstly, quantitative analysis was employed to calculate the proportion, quantity, and distribution of different types of land within the study area. Then, the mapping results were combined with the policy plans of the study area for spatial analysis, and to compare the distribution of land use with the policy. This allowed for an evaluation of whether the land use distribution pattern satisfies the demands of urban development.

3.4. Implementation and training detail

All experiments in this study were conducted using Pytorch framework in Python 3.8, with an NVIDIA GeForce RTX 3080 10G GPU for acceleration. The learning rate, iteration count, and batch size were consistently set to 0.0005, 200, and 32. The Adam optimizer (Kingma and Ba, 2014) and the CrossEntropy loss function (Ho and Wookey, 2019) are used to optimize objective. A maximum of 200 iterations was used, with early stopping to prevent overfitting.

4. Results

4.1. Validation of the proposed MDFNet model

4.1.1. Model accuracy assessment

To verify the effectiveness of the proposed model in fusing three data sources, we conducted comparative experiments with single data sources or two-source combinations. Table 2 presents the performance results for multiple models. Among the three individual data sources, the ResNet-based model using HSR achieved a test accuracy of 0.806 and a Kappa of 0.752, which outperformed RTUD (test accuracy: 0.654, Kappa: 0.565) and TTD (test accuracy: 0.571, Kappa: 0.412).

Compared to classification based on single data, data fusion further improves accuracy. Model 5 and model 6 combine HSR with RTUD or TTD, respectively, achieving test accuracies of 0.862 and 0.851, increasing by 6.9 % and 5.5 % compared to the HSR-only method. The three data sources are fused to form model 7 (MDFNet), which can

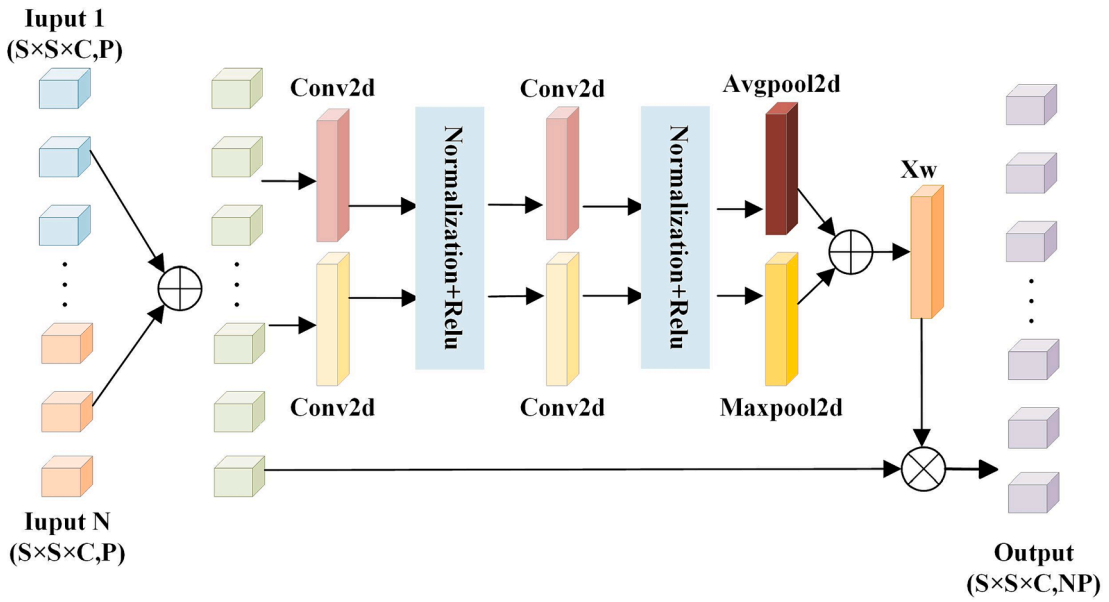


Fig. 5. Adaptive learnable weight module.

Table 2
Performance of all models on the dataset. Models 1–3: Single data; Models 4–6: Two-data fusion; Model 7: Three-data fusion (proposed method).

Model	HSR	RTUD	TTD	Test Accuracy	Kappa
1. ResNet-based	√			0.806	0.752
2. RTUDNet		√		0.654	0.565
3. GCN			√	0.571	0.412
4. RTUD&GCN		√	√	0.631	0.536
5. ResNet&RTUD	√	√		0.862	0.831
6. ResNet&GCN	√		√	0.851	0.818
7. MDFNet	√	√	√	0.882	0.858

obtain optimal performance compared other method. Moreover, comparing models 5 and 7 allows for analysis of spatial interaction effect. Results show that multimodal data fusion considering TTD increases classification accuracy by 2.3 % compared to using HSR and RTUD. Additionally, we noted that the accuracy of models using either a single RTUD or TTD is relatively low. In this situation, data fusion creates a complex feature space distribution, thereby increasing the misclassification probability (Ghamisi et al., 2016). This results in an accuracy of only 0.631 for model 4, a 3.6 % drop compared to using only RTUD.

Through the confusion matrix (Fig. 6), we observe that the accuracy of most categories significantly improves with data fusion. The residential and green land categories have achieved satisfactory results based on a single data source, and data fusion has further improved or maintained accuracy. Public service facilities and warehouse categories have shown poor performance with single data sources, with a maximum of only 54 % and 76 % respectively. Following multimodal data fusion, the accuracy of these two categories significantly improves,

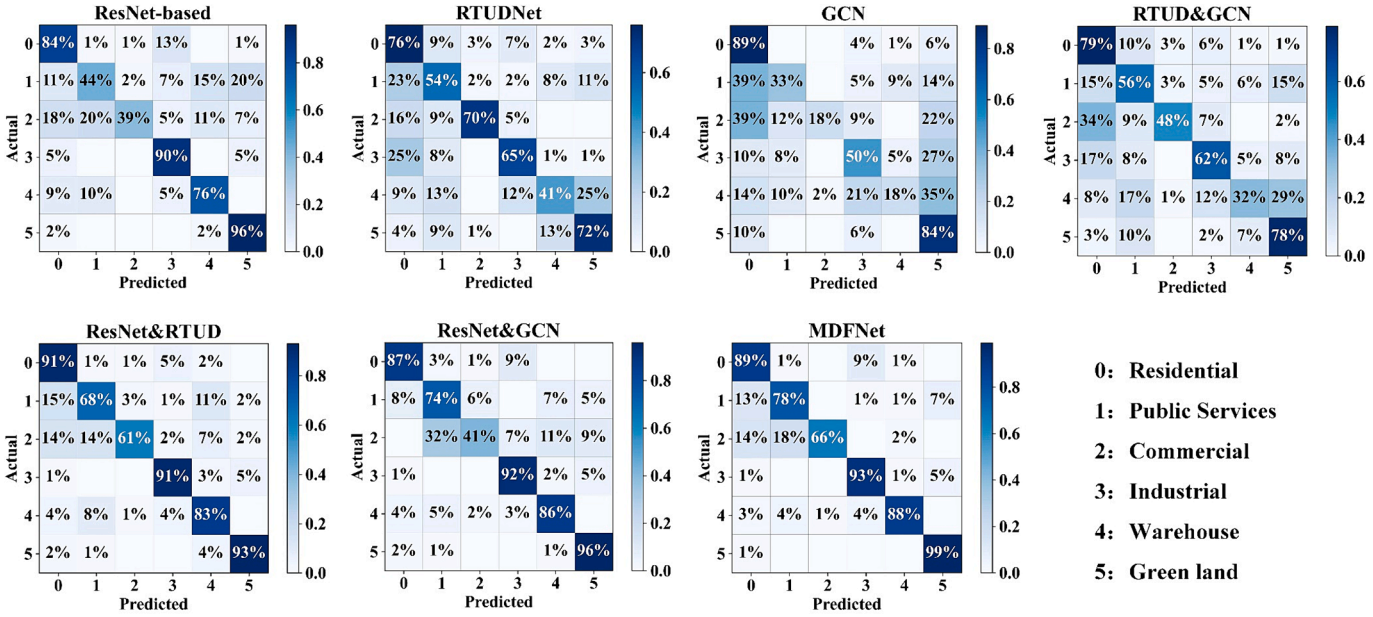


Fig. 6. Confusion matrix for all comparison models. Single data source, two data sources fusion, and three data sources fusion schemes proposed in this study are shown in order.

reaching 78 % and 88 %, respectively. This demonstrates that the proposed model can adaptively learn features from different data sources and improve model accuracy.

4.1.2. Model ablation analysis

MDFNet includes feature extraction and attention mechanisms, necessitating ablation analysis to verify each module importance. Ablation results in Table 3 show that removing BiLSTM or FCN (model 2 and model 3) leads to reduced performance, emphasizing the need for complementary long and short-term temporal features in efficiently capturing RTUD activities to represent land characteristics.

To verify the baseline model ResNet, this study excluded the iAFF of the HSR feature extraction module and replaced ResNet with VGG16, forming model 4 and model 5 respectively. The results show that ResNet uses the excellent feature extraction ability of residual blocks to obtain the optimal performance. Finally, by removing the learnable weight module to create model 6, performance decreased. This suggests that the module effectively extracts features for classification by weighing different data sources, thus enhancing accuracy. Through ablation analysis experiments, it was verified that each module of the MDFNet model is necessary to obtain optimal recognition results.

Moreover, we found that altering the model structure produced insignificant differences in accuracy, with maximum variation of 3.6 %. Conversely, as Table 2 illustrates, data fusion significantly impacts accuracy. When integrating two additional data sources with the HSR, accuracy improves by 9.4 %, surpassing model structure modifications. This highlights the necessity of data source fusion and offers valuable insights for future research.

4.1.3. Identification result error analysis

We selected several typical regions for error analysis. RTUDNet is difficult to identify regions without obvious temporal characteristic patterns. However, the ResNet-based model can classify distinctive building structure (such as regular buildings, athletic field, etc.) to avoid misclassification by RTUDNet. Conversely, areas with ambiguous natural physics but significant temporal patterns are accurately classified by RTUDNet. Fig. 7(C) is an amusement park, which shows a typical pattern of significantly higher foot traffic on holiday than on workday. However, the facility's abundant greenery and expansive layout make it misleadingly regarded as a public service facility. Fig. 7(D) shows an office building like a high-rise residence in HSR. While the natural physics of these samples are confused with others, their regular temporal patterns enable RTUDNet to classify accurate results.

In Fig. 7(E), the basketball court in the HSR has the typical characteristics of a public service facility. Coupled with the RTUD, which shows lower weekday populations and higher weekend populations, a pattern consistent with public facility usage, this could easily lead to a misclassification of the land use type. Fig. 7(F) has an architectural style like of the residential, and the RTUD also shows that the population density of weekend is more than weekdays, which is not consistent with the industrial area. Using HSR and RTUD in such scenarios is insufficient. Extracting spatial context information from TTD effectively identifies complex samples. Overall, integrating spatial context in multimodal data can reduce error rates and yields more reliable land use results.

Table 3
Ablation analysis.

No	HSR	RTUD	LW module	Test Accuracy	Kappa
1	ResNet-based	RTUDNet	√	0.882	0.858
2	ResNet-based	FCN	√	0.858	0.821
3	ResNet-based	BiLSTM	√	0.860	0.822
4	ResNet18	RTUDNet	√	0.863	0.824
5	VGG16	RTUDNet	√	0.851	0.816
6	ResNet-based	RTUDNet	/	0.857	0.818

4.2. Uncertainty analysis of multimodal data

MC-EBM is an interpretable additive model that visualizes each attribute's contribution. Given that the model only supports one-dimensional input, we have chosen the mean and variance of the data for dimensionality reduction, which can effectively represent the concentration and dispersion of high-dimensional data such as HSR and RTUD. The average speed, number of pick-ups and drop-offs from the TTD are also used as input data. The value of each data source reflects its importance in results. As shown in Table 4, the significance of HSR, RTUD, and TTD decreases sequentially, with values of 0.361, 0.308, and 0.232, respectively.

The most significant contributor is the variance of HSR pixel values, with an importance of 0.453. As illustrated in Fig. 8, lower variance allows for easier identify of categories like green land, which have consistent spectral information. While increased variance enhances identification of commercial and industrial land use, indicating that differential distribution patterns exist for physical architectural ancillary information in HSR across land use types, facilitating accurate classification of these types. The RTUD's average weekend population density attribute also made a great contribution, which can effectively identify green spaces, public service facilities, and warehouse categories when it is below 2800. Although the overall contribution of TTD was the lowest, Fig. 8 shows that off and up attributes of TTD (values within 3000) improve residential identification. This demonstrated that despite disparate contributions from diverse data sources, they retain essential roles within specific categories.

The contribution of multimodal data in different land use categories was analyzed based on the learnable weight module, as shown in Fig. 9. The HSR attribute W_{HSR} has the highest importance of about 0.4 for industrial, warehouse and green land. This may be due to the regularities of building facilities within these categories, such as large-scale green vegetation coverage in green land, which is easy to identify in HSR (Zhang et al., 2019). In public service facility land and commercial land, the W_{RTUD} is the highest, indicating that socioeconomic attributes perform better in most of these categories (He et al., 2020). The W_{TTD} has highest value in the residential, effectively identifying categories with frequent spatial interactions. For most land use categories, the collaborative mechanism of multimodal data fusion is characterized by one data source providing the primary contribution while the remaining data assist in identification.

4.3. Land use mapping results and analysis

Using the trained MDFNet for land use mapping (Fig. 10), we found that residential areas are mainly distributed in the center of Shenzhen, with commercial and public service facilities evenly distributed around residential areas. Industrial and warehouse regions are in the western and northern of Shenzhen, while green lands are mainly distributed in the eastern part of the research region. Green land dominated non-built-up area land accounts for 51.1 % of the total area, which is close to the conclusion expressed in the 2021 Urban Construction Statistical Yearbook (https://www.mohurd.gov.cn/). In urban built-up areas, residential land accounts for 15.5 %, while industrial and warehouse account for 12 % and 10.4 % respectively. Public service facility and commercial land account for 7.8 % and 3.2 % respectively, distributed in various regions to meet people daily living needs. In terms of built-up area proportion to the urban total area, Shenzhen ranks second nationwide (Lin et al., 2020). This may be due to that Shenzhen started to develop after China's reform and opening, resulting in a relatively smaller total area. However, the urban thriving economy and large population have rapidly propelled the development of built-up areas.

We further analyzed urban spatial patterns based on administrative planning. Futian and Luohu districts are Shenzhen's financial, cultural, and administration center, where the People's Government of Shenzhen Municipality and Central Business District are located. Nanshan District

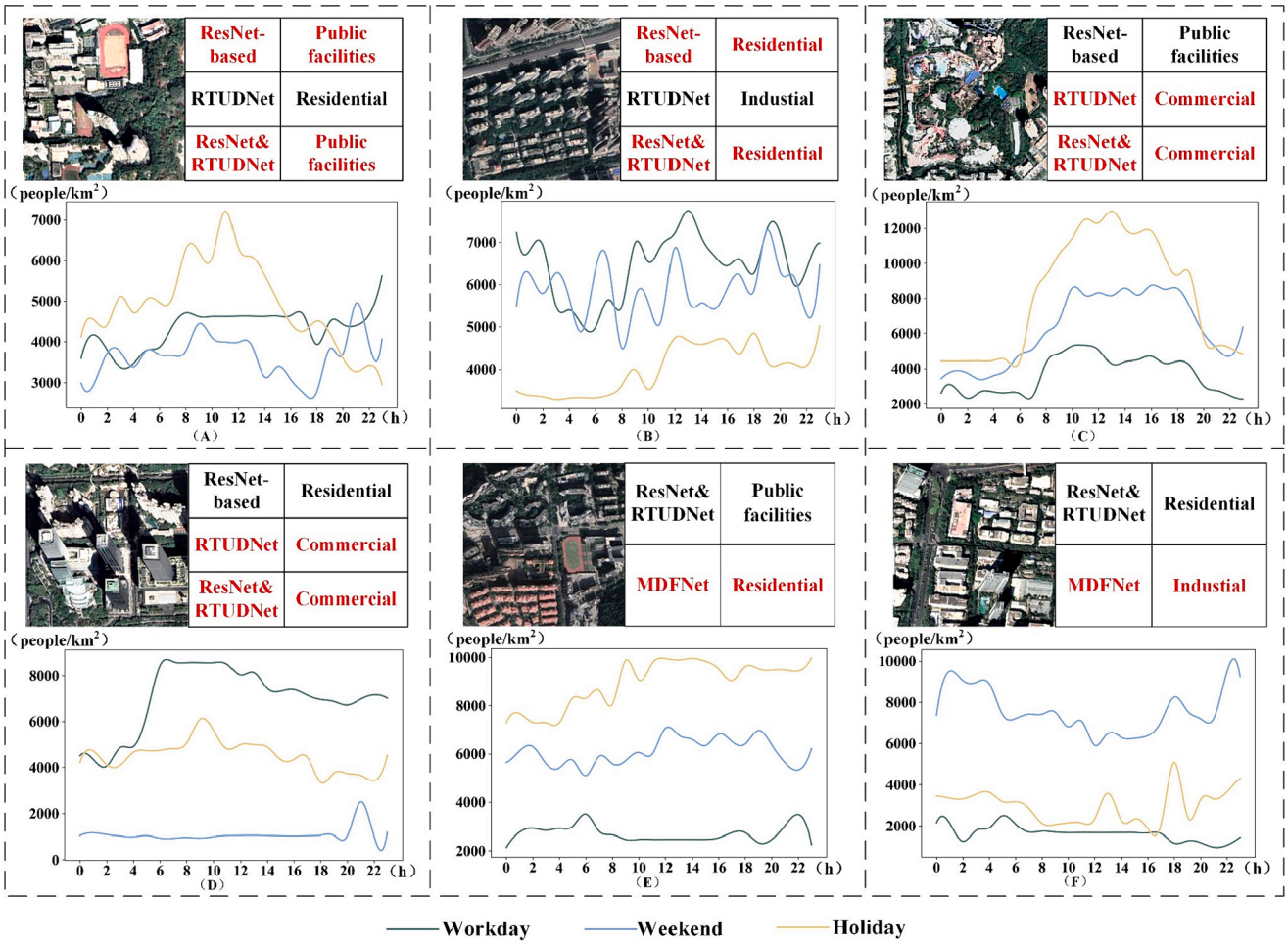


Fig. 7. Error analysis of randomly selected land use classification results. Red font indicates correct classification, black is incorrect classification. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
The importance of data sources and their attributes, data importance is obtained from the mean value of the attributes of different data.

Data source	Attributes	Attributes importance	Data importance
HSR	HSR_var	0.453	0.361
	HSR_avg	0.268	
RTUD	Workday_var	0.367	0.308
	Workday_avg	0.361	
	Weekend_var	0.284	
	Weekend_avg	0.448	
	Holiday_var	0.208	
	Holiday_avg	0.181	
TTD	Up	0.262	0.232
	Speed	0.218	
	Off	0.216	

is an education center, and Baoan and Guangming districts focus on high-tech and manufacturing. The study reveals industrial zones in western and northern suburbs due to lower rents. Eastern coastal regions support offshore trade and logistics, while remote areas promote tertiary industries like tourism. The land use distribution of the proposed method aligns with the "14th Five-Year Plan and 2035 Long-term Goals for National Economic and Social Development of Shenzhen", demonstrating the proposed model effectiveness in land use identification.

5. Discussion

5.1. Effectiveness of the proposed method

To address the limitations of previous studies that focus on socioeconomic and building physical attributes while neglecting spatial connectivity, this study proposes a MDFNet model for land use perception from multiple perspectives by extracting multimodal data attributes. The proposed MDFNet exhibits superior performance with a test accuracy of 0.882 and a Kappa value of 0.858. Compared to other studies that rely on socioeconomic and physical attributes (Cao et al., 2020, Ye et al., 2021), our method introduces spatial connectivity into the analysis, which improves the accuracy of MDFNet by 2.3 %. This underscores the importance of multimodal data fusion that considers spatial connectivity for effective land use identification.

The results demonstrate that MDFNet significantly improves the identification results of land use categories such as public service facilities and warehouse, improving the accuracy of difficult samples. The model adaptively finds efficient data fusion schemes to correctly classify complex samples, which is challenging in the case of a single data source. Ablation analysis reveals that changing the model structure has a lower impact on performance improvement (the highest is only 3.6 %), while data fusion breaks through the model identification bottleneck (9.4 % improvement). Data-centric AI has proven efficient (Chen, 2023), suggesting that future research should focus on multimodal data fusion methods.

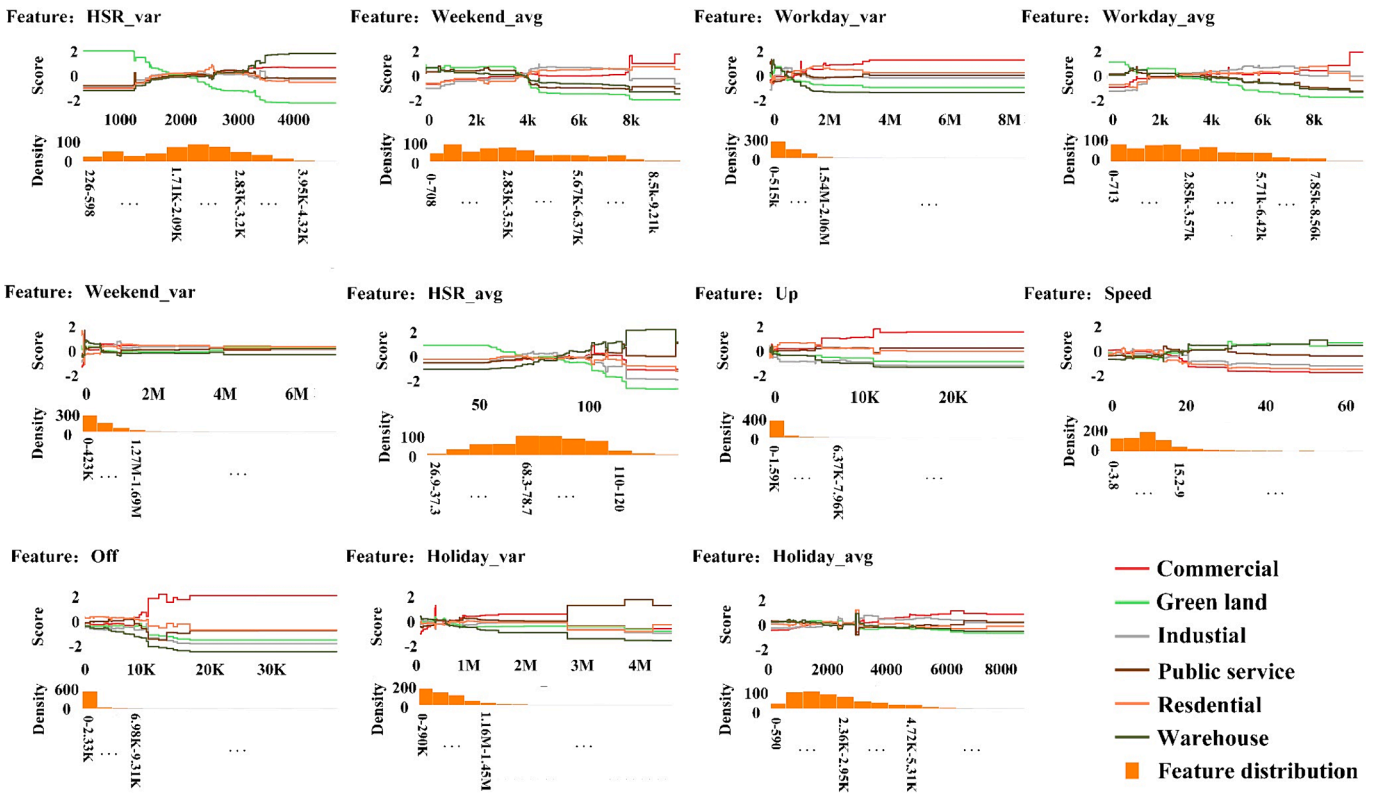


Fig. 8. Distribution of the contribution of each type of attributes to the results. The score indicates the degree of contribution of each attribute to the result, density represents the distribution space.

5.2. Mechanisms of data influence on land use identification

This study introduces a novel uncertainty analysis framework, quantitatively evaluating multimodal data and addressing the challenge of determining data contributions in complex urban land use classification. Unlike permutation-based method (Wu et al., 2023) and decision tree-based method (Zhang et al., 2019) for multimodal data analysis, which are struggled to delineate the importance of multimodal data in the overall task and in each specific category. In this study, ante-hoc demonstrates a descending order of contribution for HSR, RTUD, and TTD in the overall land use identification, with values of 0.361, 0.308, and 0.232, respectively. Post-hoc analysis reveals the impact of data sources on land use categories. HSR is important in most categories with prominent building attributes, with W_{HSR} value of about 0.4. RTUD significantly influences commercial and public service facilities (W_{RTUD} values of 0.408 and 0.404). TTD is crucial for residential areas (W_{TTD} of 0.4). The coordination mechanism of multimodal data fusion allows different data sources to lead in various categories, with others assisting in identification.

The uncertainty analysis confirms the significant contribution of HSR to the overall task and various land use categories, with the greatest impact seen in industrial, warehouse, and green land (W_{HSR} values of 0.418, 0.373, and 0.391, respectively). This underscores the essential role of HSR in land use identification, suggesting its prioritization in future research. Despite their lesser overall importance, other data sources still have notable roles in specific categories and can assist HSR to enhance identification accuracy. This study demonstrates the diverse contributions of different data sources to distinct land use categories. HSR is effective for visually salient categories, RTUD for areas with strong socioeconomic attributes, and TTD for regions with frequent spatial interactions. This implies future research can strategically use various data to accurately identify specific complex categories.

5.3. Land use development patterns of typical metropolis

Shenzhen is a metropolis and understand its land use spatial patterns provides valuable insights for the development of other cities. The study finds that green lands are predominantly located in the southeastern, where there is higher elevation and rich vegetation. Industrial and warehouse areas are interspersed in suburban and coastal regions, minimizing rent costs, and facilitating transportation. The mapping results indicate that Shenzhen exhibits a mixed distribution of residential, public service, and commercial lands, aligning with the “15-minute community life circle” planning scheme (Weng et al., 2019), promoting urban vitality and sustainable development. The results validate the proposed method effectiveness and provide technical support for efficient and automated urban mapping.

5.4. Limitations and future works

There are still limitations and improvements. First, this study examines multimodal data uncertainty, but spatial scale also introduces uncertainty in land use (Chen et al., 2019, Wu et al., 2019). Scale can affect detail capture, possibly causing mixed land use within units and creating errors. Future research could analyze scale sensitivity based on the Modifiable Areal Unit Problem (MAUP). In addition, identifying mixed land use also deserves further study. Second, recent studies have highlighted multimodal data fusion as promising research (Li et al., 2022; Liu et al., 2020). Future research will focus on algorithm optimization for data fusion to improve feature extraction, and comparison with existing state-of-the-art methods to further improve the classification accuracy. Finally, testing the model in diverse cities to evaluate its transferability and generalization capabilities is a valuable research direction (Rosier et al., 2022).

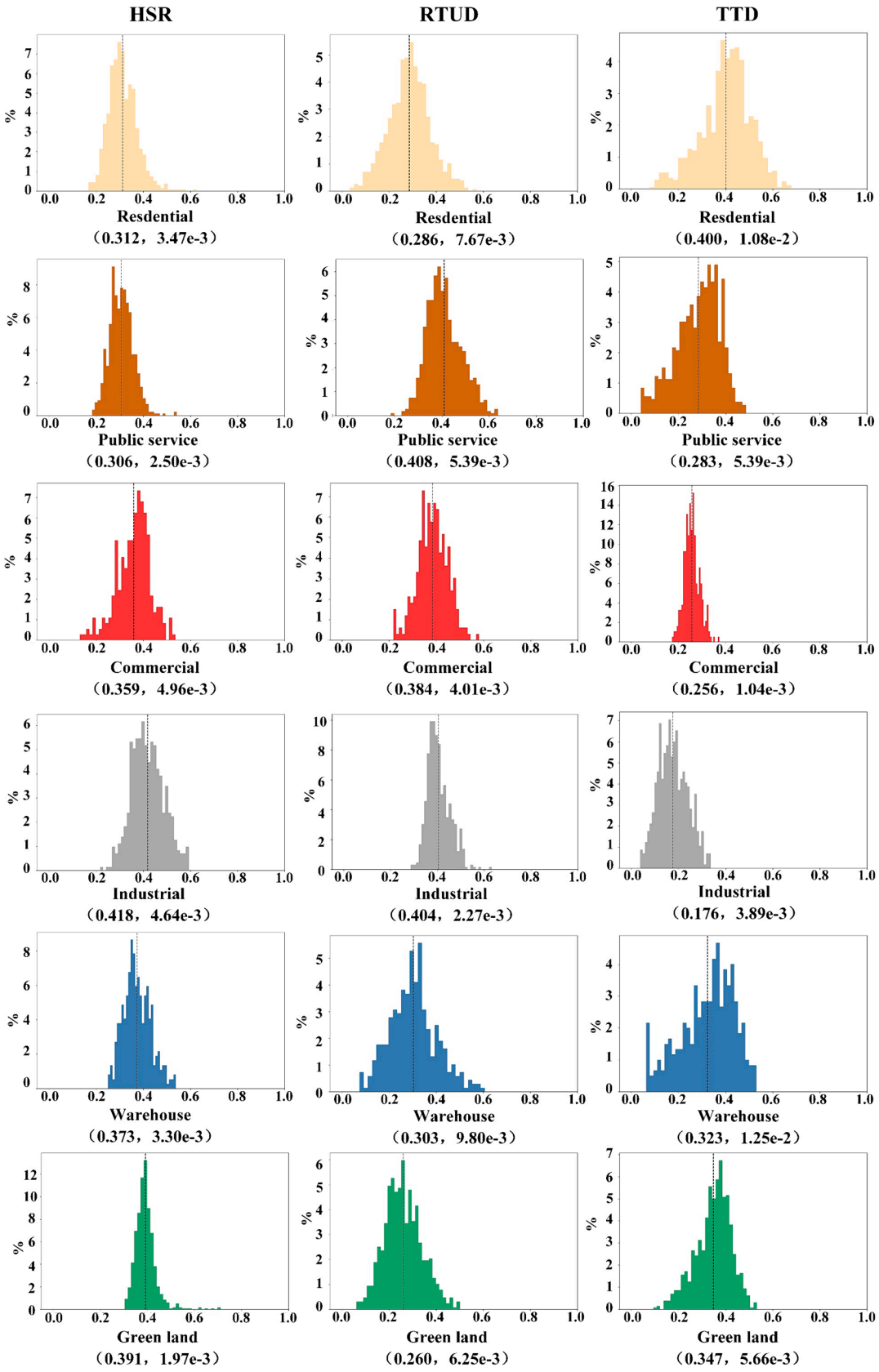


Fig. 9. Histogram of attribute contributions from land use. The horizontal axis denotes attribute weight, and the vertical axis represents the percentage. Mean and variance of weights are in parentheses.

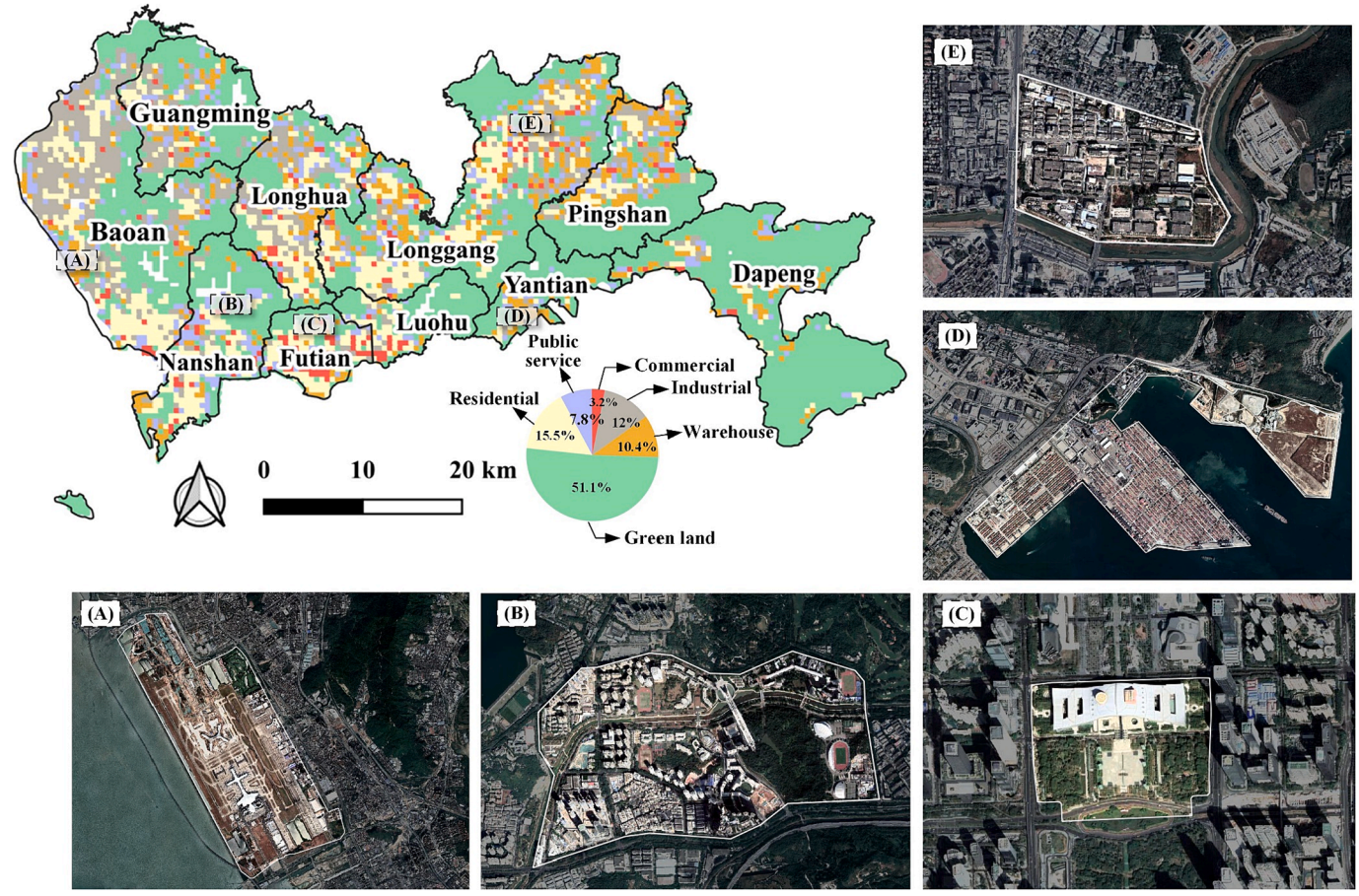


Fig. 10. Land use mapping results and typical regions. (A): Airport, (B): University town, (C): Government, (D): Industrial zone, (E): Trade Port. Each irregular subplot represents the core region of one or more grids, and most of the grid within these regions are correctly identified by MDFNet.

6. Conclusion

This study introduces a novel framework for land use mapping by integrating HSR, RTUD, and TTD to extract building physical, socio-economic, and spatial connectivity ancillary information, offering a comprehensive understanding of land use patterns. The method was validated in Shenzhen, and the results showed that the method outperformed other methods with a test accuracy of 0.882, demonstrating the necessary of multimodal data fusion. Another key contribution of this study is constructing an uncertainty analysis framework, which quantitatively examines the contribution and collaboration mechanisms of multimodal data in land use mapping. This framework clarifies the significance of multimodal data in land use studies and highlights the specific categories they effectively influence. Our research offers a practical framework for selecting appropriate data sources and combinations in future land use mapping tasks, improving urban spatial distribution pattern analyses.

CRediT authorship contribution statement

Xiaoqin Yan: Writing – review & editing, Writing – original draft, Methodology. **Zhangwei Jiang:** Writing – review & editing, Methodology, Data curation, Conceptualization. **Peng Luo:** Writing – review & editing, Writing – original draft, Data curation, Conceptualization. **Hao Wu:** Writing – review & editing, Writing – original draft, Validation, Methodology. **Anning Dong:** Writing – review & editing, Writing – original draft, Validation, Methodology, Data curation. **Fengling Mao:** Writing – review & editing, Writing – original draft, Validation, Data curation. **Ziyin Wang:** Writing – review & editing, Writing – original draft, Data curation, Conceptualization. **Hong Liu:** Writing – review & editing, Writing – original draft, Validation, Funding acquisition, Data curation, Conceptualization. **Yao Yao:** Writing – review & editing, Writing – original draft, Validation, Project administration, Methodology, Funding acquisition, Formal analysis, Data curation, 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.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by Alibaba Group through Alibaba Innovation Research Program [No.20228670], the National Key Research and Development Program of China [2019YFB2102903], the National Natural Science Foundation of China [41801306]; the “CUG Scholar” Scientific Research Funds at China University of Geosciences (Wuhan) [2022034] and a grant from State Key Laboratory of Resources and Environmental Information System.

References

Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M., Farhan, L., 2021. Review of deep Learning: concepts CNN architectures, challenges, applications, future directions. J BIG DATA-GER 8, 1–74.

Bai, S., Kolter, J.Z., Koltun, V., 2018. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling, arXiv preprint arXiv:1803.01271.

Bai, L., Huang, W., Zhang, X., Du, S., Cong, G., Wang, H., Liu, B., 2023. Geographic mapping with unsupervised multi-modal representation learning from VHR images and POIs. ISPRS J PHOTOGRAMM 201, 193–208.

Cao, R., Tu, W., Yang, C., Li, Q., Liu, J., Zhu, J., Zhang, Q., Li, Q., Qiu, G., 2020. Deep learning-based remote and social sensing data fusion for urban region function recognition. ISPRS J PHOTOGRAMM 163, 82–97.

Chen, M., 2023. Machine Learning empowered intelligent data center networking evolution. Challenges and Opportunities, MACH LEARN.

Chen, L., Gao, Y., Zhu, D., Yuan, Y., Liu, Y., 2019. Quantifying the scale effect in geospatial big data using semi-Variograms. PLoS One 14, e0225139.

Dai, Y., Gieseke, F., Oehmcke, S., Wu, Y., Barnard, K., 2021. Attentional feature fusion. Proce. IEEE/CVF Winter Conference on Applications of Comp. Vision 3560–3569.

Doshi-Velez, F., Kim, B., 2017. Towards a Rigorous Science of Interpretable Machine Learning, arXiv preprint arXiv:1702.08608.

Du, S., Du, S., Liu, B., Zhang, X., Zheng, Z., 2020. Large-scale urban functional zone mapping by integrating remote sensing images and open social data. GISCI REMOTE SENS 57, 411–430.

Du, M., Liu, N., Hu, X., 2019. Techniques for interpretable machine Learning. COMMUN ACM 63, 68–77.

Duranton, G., Puga, D., Urban Land Use, Handbook of Regional and Urban Economics, Elsevier2015. pp. 467-560.

Feng, Y., Huang, Z., Wang, Y., Wan, L., Liu, Y., Zhang, Y., Shan, X., 2021. An SOE-based Learning framework using Multisource big data for identifying urban functional zones. IEEE J-STARS 14, 7336–7348.

Ghamisi, P., Höfle, B., Zhu, X.X., 2016. Hyperspectral and Lidar data fusion using extinction profiles and deep convolutional neural network. IEEE J-STARS 10, 3011–3024.

Gong, P., Chen, B., Li, X., Liu, H., Wang, J., Bai, Y., Chen, J., Chen, X., Fang, L., Feng, S., 2020. Mapping essential urban land use categories in China (EULUC-China): Preliminary results for 2018. SCI BULL 65, 182–187.

Guan, Q., Cheng, S., Pan, Y., Yao, Y., Zeng, W., 2021. Sensing mixed urban land-use patterns using municipal water consumption time series. ANN AM ASSOC GEOGR 111, 68–86.

Guan, Q., Wang, J., Ren, S., Gao, H., Liang, Z., Wang, J., Yao, Y., 2024. Predicting short-term PM2. 5 concentrations at fine temporal resolutions using a multi-branch temporal graph convolutional neural network. INT J GEOGR INF SCI, pp. 1–24.

He, J., Li, X., Liu, P., Wu, X., Zhang, J., Zhang, D., Liu, X., Yao, Y., 2020. Accurate estimation of the proportion of mixed land use at the street-block level by integrating high spatial resolution images and geospatial big data. IEEE T GEOSCI REMOTE 59, 6357–6370.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual Learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Ho, Y., Wookey, S., 2019. The real-world-weight cross-entropy loss function: modeling the costs of mislabeling. IEEE Access 8, 4806–4813.

Hu, S., Gao, S., Wu, L., Xu, Y., Zhang, Z., Cui, H., Gong, X., 2021. Urban function classification at road segment level using taxi trajectory data: a graph convolutional neural network approach. Computers, Environ. Urban Systems 87, 101619.

Hu, S., Gao, S., Luo, W., Wu, L., Li, T., Xu, Y., Zhang, Z., 2023. Revealing intra-urban Hierarchical spatial structure through representation Learning by combining road network Abstraction model and taxi trajectory data. ANN GIS 29, 499–516.

Huang, R., Taubenböck, H., Mou, L., Zhu, X.X., 2018. Classification of settlement types from tweets using LDA and LSTM. IGARSS 2018–2018 IEEE Int. Geoscience and Remote Sensing Symposium, IEEE 6408–6411.

Kingma, D.P., Ba, J., 2014. Adam: A Method for Stochastic Optimization, arXiv preprint arXiv:1412.6980.

Koroso, N.H., Lengoiboni, M., Zevenbergen, J.A., 2021. Urbanization and urban land use efficiency: evidence from regional and addis ababa satellite cities, Ethiopia. HABITAT INT 117, 102437.

Kwan, M., Weber, J., 2008. Scale and accessibility: implications for the analysis of land use-travel Interaction. APPL GEOGR 28, 110–123.

Lei, Y., Flacke, J., Schwarz, N., 2021. Does urban planning affect urban growth pattern? a case study of Shenzhen, China. Land Use Policy 101, 105100.

Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., Chanussot, J., 2022. Deep Learning in multimodal remote sensing data fusion: a comprehensive review. INT J APPL EARTH OBS 112, 102926.

Lin, J., Wan, H., Cui, Y., 2020. Analyzing the spatial factors related to the distributions of building heights in urban areas: a comparative case study in guangzhou and shenzhen. SUSTAIN CITIES SOC 52, 101854.

Liu, X., Kang, C., Gong, L., Liu, Y., 2016. Incorporating spatial Interaction patterns in classifying and understanding urban land use. INT J GEOGR INF SCI 30, 334–350.

Liu, J., Li, T., Xie, P., Du, S., Teng, F., Yang, X., 2020. Urban big data fusion based on deep Learning: an overview. INFORM FUSION 53, 123–133.

Liu, S., Shi, Q., 2020. Local climate zone mapping as remote sensing scene classification using deep Learning: a case study of metropolitan China. ISPRS J PHOTOGRAMM 164, 229–242.

Lu, W., Tao, C., Li, H., Qi, J., Li, Y., 2022. A unified deep Learning framework for urban functional zone extraction based on multi-source heterogeneous data. REMOTE SENS ENVIRON 270, 112830.

Lyu, F., Zhang, L., 2019. Using multi-source big data to understand the factors affecting urban park use in Wuhan. URBAN FOR URBAN GREE 43, 126367.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2016. Convolutional neural networks for Large-scale remote-sensing image classification. IEEE T GEOSCI REMOTE 55, 645–657.

Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B., 2019. Definitions, methods, and applications in interpretable machine Learning. Proc. Natl. Acad. Sci. 116, 22071–22080.

Nori, H., Jenkins, S., Koch, P., Caruana, R., 2019. InterpretML: A Unified Framework for Machine Learning Interpretability, arXiv preprint arXiv:1909.09223.

Park, S., Xu, Y., Jiang, L., Chen, Z., Huang, S., 2020. Spatial structures of tourism destinations: a trajectory data mining approach leveraging mobile big data. ANN TOURISM RES 84, 102973.

Rosier, J.F., Taubenböck, H., Verburg, P.H., van Vliet, J., 2022. Fusing earth observation and socioeconomic data to increase the transferability of large-scale urban land use classification. REMOTE SENS ENVIRON 278, 113076.

Sokol, K., Flach, P., 2020. Explainability fact sheets: a framework for systematic assessment of explainable approaches. Proce. 2020 Conference on Fairness, Accountability, and Transparency 56–67.

Srivastava, S., Vargas-Munoz, J.E., Tuia, D., 2019. Understanding urban landuse from the above and ground perspectives: a deep learning multimodal solution. REMOTE SENS ENVIRON 228, 129–143.

Wang, S., Cao, J., Yu, P., 2020. Deep Learning for spatio-temporal data mining: a survey. IEEE T KNOWL DATA EN.

Weng, M., Ding, N., Li, J., Jin, X., Xiao, H., He, Z., Su, S., 2019. The 15-minute walkable neighborhoods: measurement social inequalities and implications for building healthy communities in urban China. J TRANSP HEALTH 13, 259–273.

Wu, H., Li, Z., Clarke, K.C., Shi, W., Fang, L., Lin, A., Zhou, J., 2019. Examining the sensitivity of spatial scale in Cellular automata Markov chain simulation of land use change. INT J GEOGR INF SCI 33, 1040–1061.

Wu, H., Luo, W., Lin, A., Hao, F., Olteanu-Raimond, A., Liu, L., Li, Y., 2023. SALT: a multifeature ensemble Learning framework for mapping urban functional zones from VGI data and VHR images, computers. Environ. Urban Systems 100, 101921.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Philip, S.Y., 2020. A comprehensive survey on graph neural networks. IEEE T NEUR NET LEAR 32, 4–24.

Xia, C., Yeh, A.G., Zhang, A., 2020. Analyzing spatial relationships between urban land use intensity and urban vitality at street block level: a case study of five chinese megacities. LANDSCAPE URBAN PLAN 193, 103669.

Xing, X., Yu, B., Kang, C., Huang, B., Gong, J., Liu, Y., 2024. The synergy between remote sensing and social sensing in urban studies. Review and Perspectives, IEEE GEOSC REM SEN M.

Xu, R., Huang, W., Zhao, J., Chen, M., Nie, L., 2023. A spatial and Adversarial representation Learning approach for land use classification with POIs. ACM T INTEL SYST TEC 14, 1–25.

Xu, Y., Zhou, B., Jin, S., Xie, X., Chen, Z., Hu, S., He, N., 2022. A framework for urban land use classification by integrating the spatial context of points of interest and graph convolutional neural network method. Computers, Environ. Urban Systems 95, 101807.

Yao, Y., Guo, Z., Dou, C., Jia, M., Hong, Y., Guan, Q., Luo, P., 2023a. Predicting mobile users' next location using the semantically enriched geo-embedding model and the multilayer attention mechanism. Environ. Urban Systems 104, 102009.

Yao, Y., Yan, X., Luo, P., Liang, Y., Ren, S., Hu, Y., Han, J., Guan, Q., 2022. Classifying land-use patterns by integrating time-series electricity data and high-spatial resolution remote sensing imagery. INT J APPL EARTH OBS 106, 102664.

Yao, Y., Zhu, Q., Guo, Z., Huang, W., Zhang, Y., Yan, X., Dong, A., Jiang, Z., Liu, H., Guan, Q., 2023b. Unsupervised land-use change detection using multi-temporal POI embedding. INT J GEOGR INF SCI 37, 2392–2415.

Ye, C., Zhang, F., Mu, L., Gao, Y., Liu, Y., 2021. Urban function recognition by integrating social media and street-level imagery. Environ. Planning B: Urban Analytics and City Sci. 48, 1430–1444.

Yin, J., Dong, J., Hamm, N.A., Li, Z., Wang, J., Xing, H., Fu, P., 2021. Integrating remote sensing and geospatial big data for urban land use mapping: a review. INT J APPL EARTH OBS 103, 102514.

Yin, G., Huang, Z., Bao, Y., Wang, H., Li, L., Ma, X., Zhang, Y., 2023. ConvGCN-RF: a hybrid learning model for commuting flow prediction considering geographical semantics and neighborhood effects. GeoInformatica 27, 137–157.

Zhang, X., Du, S., Wang, Q., 2017. Hierarchical semantic cognition for urban functional zones with VHR satellite images and POI data. ISPRS J PHOTOGRAMM 132, 170–184.

Zhang, Y., Li, Q., Tu, W., Mai, K., Yao, Y., Chen, Y., 2019. Functional urban land use recognition integrating multi-source geospatial data and cross-correlations. Computers, Environ. Urban Systems 78, 101374.

Zhang, X., Tan, S., Koch, P., Lou, Y., Chajewska, U., Caruana, R., 2019. Axiomatic interpretability for multiclass additive models. Proce. 25th ACM SIGKDD Int. Conference on Knowledge Discovery & Data Mining 226–234.

Zhang, Z., Zhang, Y., He, T., Xiao, R., 2022. Urban vitality and its influencing factors: comparative analysis based on taxi trajectory data. IEEE J-STARS 15, 5102–5114.

Zheng, Y., 2015. Trajectory data mining: an overview. ACM Trans. Intelligent Systems and Technol. (TIST) 6, 1–41.

Zhong, Y., Su, Y., Wu, S., Zheng, Z., Zhao, J., Ma, A., Zhu, Q., Ye, R., Li, X., Pellikka, P., 2020. Open-source data-driven urban land-use mapping integrating point-line-polygon semantic objects: a case study of chinese cities. REMOTE SENS ENVIRON 247, 111838.

Zhou, W., Ming, D., Lv, X., Zhou, K., Bao, H., Hong, Z., 2020. SO-CNN based urban functional zone fine division with VHR remote sensing image. REMOTE SENS ENVIRON 236, 111458.
Zhu, X.X., Tuia, D., Mou, L., Xia, G., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: a comprehensive review and list of resources. IEEE GEOSC REM SEN M 5, 8–36.
Zhu, D., Zhang, F., Wang, S., Wang, Y., Cheng, X., Huang, Z., Liu, Y., 2020. Understanding place characteristics in geographic contexts through graph convolutional neural networks. ANN AM ASSOC GEOGR 110, 408–420.

Mr. Xiaoqin Yan has graduated from the School of Geophysics and Information Engineering, China University of Geosciences (Wuhan) with a master's degree and is currently pursuing a doctoral degree at the School of Earth and Space Sciences, Peking University. His research interests are geospatial big data mining, and urban land mapping.

Mr. Zhangwei Jiang is a staff algorithm engineer at Alibaba Group. His research interests are LBS data mining and research&recommendation algorithm.

Mr. Peng Luo is a Ph.D. candidate at the Chair of Cartography and Visual Analytics at the Technical University of Munich, Germany. He is currently a visiting researcher at School of Geography and the Environment, the University of Oxford. His research interests include spatial association modelling, social sensing, and applied artificial intelligence.

Mr. Hao Wu is a graduate student at China University of Geosciences (Wuhan). His research interests are geospatial big data mining, data-centric urban modeling.

Mr. Anning Dong is a graduate student at China University of Geosciences (Wuhan). His research interests are Volunteer Geographic Information Data Analysis and GIS Engineering.

Mr. Hong Liu is a senior staff algorithm engineer at Alibaba Group. His research interests are data mining and research&recommendation algorithm.

Ms. Fengling Mao is an algorithm engineer at Alibaba Group. Her research interests are trajectory pattern mining and spatiotemporal data embedding.

Mr. Ziyin Wang is an algorithm engineer at Alibaba Group. His research interests are natural language processing, data-centric AI and Recommendation System.

Dr. Yao Yao is a Professor at China University of Geosciences (Wuhan), a researcher at the University of Tokyo and a senior algorithm engineer at Alibaba Group. His research interests are geospatial big data mining, analysis, and computational urban science.