







A Transformer-Based Multimodal Model for Urban–Rural Fringe Identification

Furong Jia, Quanhua Dong , Zhou Huang , Xiao-Jian Chen, Yi Wang , Xia Peng, Yuan Guo , Ruixian Ma , Fan Zhang, and Yu Liu 

Abstract—As the frontier of urbanization, urban–rural fringes (URFs) transitionally connect urban construction regions to the rural hinterland, and its identification is significant for the study of urbanization-related socioeconomic changes and human dynamics. Previous research on URF identification has predominantly relied on remote sensing data, which often provides a uniform overhead perspective with limited spatial resolution. As an additional data source, street view images (SVIs) offer a valuable human-related perspective, efficiently capturing intricate transitions from urban to rural areas. However, the abundant visual information offered by SVIs has often been overlooked and multimodal techniques have seldom been explored to integrate multisource data for delineating URFs. To address this gap, this study proposes a transformed-based multimodal methodology for identifying URFs, which includes a street view panorama classifier and a remote sensing classification model. In the study area of Beijing, the experimental results indicate that an URF with a total area of 731.24 km² surrounds urban cores, primarily located between the fourth and sixth ring roads. The effectiveness of the proposed method is demonstrated through comparative experiments with traditional URF identification methods. In addition, a series of ablation studies demonstrate the efficacy of incorporating multisource data. Based on the delineated URFs in Beijing, this research introduced points of interest data and commuting data to analyze the socioeconomic characteristics of URFs. The findings indicate that URFs are characterized by longer commuting distances and less diverse restaurant consumption patterns compared to more urbanized regions. This study enables the accurate identification of URFs through the transform-based multimodal approach integrating SVIs. Furthermore, it provides

a human-centric comprehension of URFs, which is essential for informing strategies of urban planning and development.

Index Terms—Deep learning, social sensing, street view images (SVIs), urban rural fringe (URF), urbanization.

I. INTRODUCTION

THE urban–rural fringe (URF) demarcates the boundary of urban expansion, serving as a pivotal interface between urban and rural areas, and integrating demographic, land use, economic, and environmental elements. Scholars commonly characterize the URF as a “semiurbanized” region, emphasizing its continuous sprawl and incomplete urban transformation, frequently functioning as suburban zones for major cities [1], [2], [3]. Given the limitations of the traditional dichotomy between urban and rural areas in various fields, the concept of URF as part of the rural–urban gradient, enriches our understanding by offering a more nuanced spatial representation of the interactions between rural areas, urban centers, and the broader rural–urban system [4]. The spatial identification of URFs holds significant importance for analyzing urbanization trends and promoting urban development, especially in countries that have undergone rapid urban growth in recent decades, such as China [5], [6]. However, the precise identification of URFs presents a challenge due to their geographically fragmented distribution and the complex, often disputed, nature of land use within these areas, where physical landscapes and social structures bear similarities to both urban and rural settings.

Remotely sensed data have become the predominant source for identifying URFs, as they can effectively capture detailed physical and socioeconomic features of the land surface, enabling the differentiation between urban and rural areas. Most methods utilizing remotely sensed data apply unsupervised learning to identify URFs. These methods typically identify regions that holds features found in both urban and rural areas, while also being distinct from both in terms of land use intensity and economic status [3], [5], [7]. For instance, impervious surface area data and DMSP/OLS nighttime light data were, respectively, utilized in the studies by the authors in [7] and [5] for identifying URFs. [3] integrated nighttime light data with land use data to create a comprehensive index of land development intensity. They employed a self-organizing feature map (SOFM) to cluster lands in Beijing into three categories for the identification of URFs. However, with limited spatial resolution in previous studies, such as kilometer-scale for nighttime light

Manuscript received 8 June 2024; revised 20 July 2024; accepted 25 July 2024. Date of publication 6 August 2024; date of current version 5 September 2024. This work was supported in part by the International Research Center of Big Data for Sustainable Development Goals under Grant CBAS2022GSP06 and in part by the National Natural Science Foundation of China under Grant 42401560, Grant 42371468, Grant 42201507, Grant U2344216, and Grant 42271471. (Corresponding author: Quanhua Dong.)

Furong Jia, Quanhua Dong, Zhou Huang, Xiao-Jian Chen, Yi Wang, and Fan Zhang are with the Institute of Remote Sensing and Geographical Information Systems, School of Earth and Space Sciences, Peking University, Beijing 100871, China (e-mail: jiafr1802@stu.pku.edu.cn; dqh@pku.edu.cn; huangzhou@pku.edu.cn; cxiaojian@pku.edu.cn; wang.yi@stu.pku.edu.cn; fanzhanggis@pku.edu.cn).

Yuan Guo is with the School of Resources and Environmental Engineering, Wuhan University of Technology, Wuhan 430070, China (e-mail: yuang@whut.edu.cn).

Xia Peng is with Tourism College, Beijing Union University, Beijing 100101, China (e-mail: ivy_px@163.com).

Ruixian Ma is with the Senseable City Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: ruixian@mit.edu).

Yu Liu is with the Institute of Remote Sensing and Geographical Information Systems, School of Earth and Space Sciences, Peking University, Beijing 100871, China, and also with the International Research Center of Big Data for Sustainable Development Goals, Beijing 100094, China (e-mail: liuyu@urban.pku.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3439429

data, there is insufficient spatial detail to effectively perform URF identification. Furthermore, URFs can hardly be efficiently recognized using only remotely sensed data sources, due to the constrained characteristics offered by an overhead perspective and the absence of a ground-based view.

Street view images (SVIs) generally refer to panoramic images of streetscapes from a human perspective [8], [9], providing additional visual modality data for URF identification [6]. These images record detailed information on human infrastructure associated with URFs, including architectural facades and the materials used for buildings and roadways. In addition to depicting physical space, SVIs unveil urban lifestyle and socioeconomic characteristics, emphasizing the importance of microscopic spatial scale of individuals [10]. They offering insights into neighborhood facilities and amenities, such as groceries, signs, and bus stations, which encode people's daily life experiences [11], [12], [13], [14], [15]. As a crucial supplement to remotely sensed data, SVIs provide a fine-grained description of urban built environment necessary for the precise identification of URFs. Given the wealth of detailed visual information offered by SVIs, incorporating SVIs with remote sensing (RS) imagery potentially enhances the accuracy of URF identification. However, previous studies commonly overlook the importance of combining these two data sources.

To facilitate the integration and processing of multisource data in urban studies, the adoption of multimodal learning has gained momentum, which can deliver more informative and precise outcomes. This approach, exemplified in [16], integrated diverse data, such as RS images and social sensing data, to tackle the challenge of recognizing urban region functions. Similarly, Srivastava et al. [17] developed a multimodal model that leverages visual information from both aerial and ground perspectives to predict land use patterns at the urban-object level. Furthermore, Huang et al. [18] constructed a comprehensive urban space representation through multimodal learning, combining SVIs, RS imagery, and social sensing data. This proposed framework demonstrated efficacy in identifying urban villages through experimentation. These studies show the potential of multimodal techniques in learning subtle difference between urban and rural areas.

In existing multimodal research, the Transformer-based models [19], [20] have demonstrated remarkable performance and scalability in visual studies. Transformers possess the capability to extract implicit urban features and enhance multimodal learning, contributing to a deeper understanding of urban spaces. The Swin Transformer [21], a vision Transformer model, employs a hierarchical structure with shifted windows to effectively model visual features across multiple scales. This architecture excels at capturing geographical information, enabling the detailed extraction of geospatial features from RS data [22]. Its application has extended to various tasks in geographical domain, including semantic segmentation of urban greenness and sea ice classification [23], [24]. Employing Transformer-based models, such as the Swin Transformer, to extract visual modality with detailed information might reveal and emphasize subtle traces of URFs. However, existing URF identification studies have

failed to capitalize on the advantages of visually intelligent deep learning methods due to their lack of inclusion of multisource, vast labeled visual data with fine-grained details and high spatial resolution.

To address the challenge of URF identification and develop a comprehensive understanding of URF using multisource data, we propose a Transformer-based multimodal approach integrating SVIs, RS imagery, and other remotely sensed data. First, we establish a framework for labeling a street view panorama dataset capable of distinguishing URF from urban and rural areas. Second, a street view panorama classifier and a Transformer-based multimodal RS classification model are built to identify URFs. In the multimodal classification model, multisource data are fused to accurately characterize URF and achieve reliable recognition results, including RS images, and auxiliary RS data, such as population, nighttime light, and normalized difference vegetation index (NDVI). Experimental results demonstrate the efficacy of fusing multimodal learning in the identification of URFs, yielding an overall accuracy (OA) of 72.70%. Based on the identified URFs, this study further analyzes their socioeconomic characteristics, with a focus on consumption characteristics and commuting patterns. By improving the accuracy of URF identification and shedding light on the characteristics of URFs, this research contributes valuable insights for urban planning and development.

The rest of this article is organized as follows. Section II first introduces the multimodal framework proposed in this research. Section III provides details regarding the study area, data sources, and implementation procedures. Section IV shows the results of URF identification and evaluates the effectiveness of the proposed multimodal framework. Section V further analyses the socioeconomic characteristics of URFs based on the identification results. Finally, Section VI concludes this article.

II. METHODS

A. Overall Framework

To delineate URFs in Beijing, this study proposed a deep-learning-based approach, using an integration of SVIs, RS and population data. The workflow is illustrated in Fig. 1, which included three major procedures: 1) SVI data acquisition and labeling, 2) SVI classification, and 3) Transformer-based multimodal RS classification.

First, the street view panoramas within this study area were collected, and a sample set was extracted and labeled with the proposed framework. Second, a street view panorama classifier was trained to classify SVIs into urban, URFs and rural areas, and automatically extended the classification results to each image in the entire Beijing SVI dataset. After mapping and rasterizing the classification results, a Transformer-based multimodal RS classification model was used to build the relationships between multisource remotely sensed data and street view category in the SVI-available area. This model was then utilized to predict and render the complete spatial distribution of URFs across Beijing.

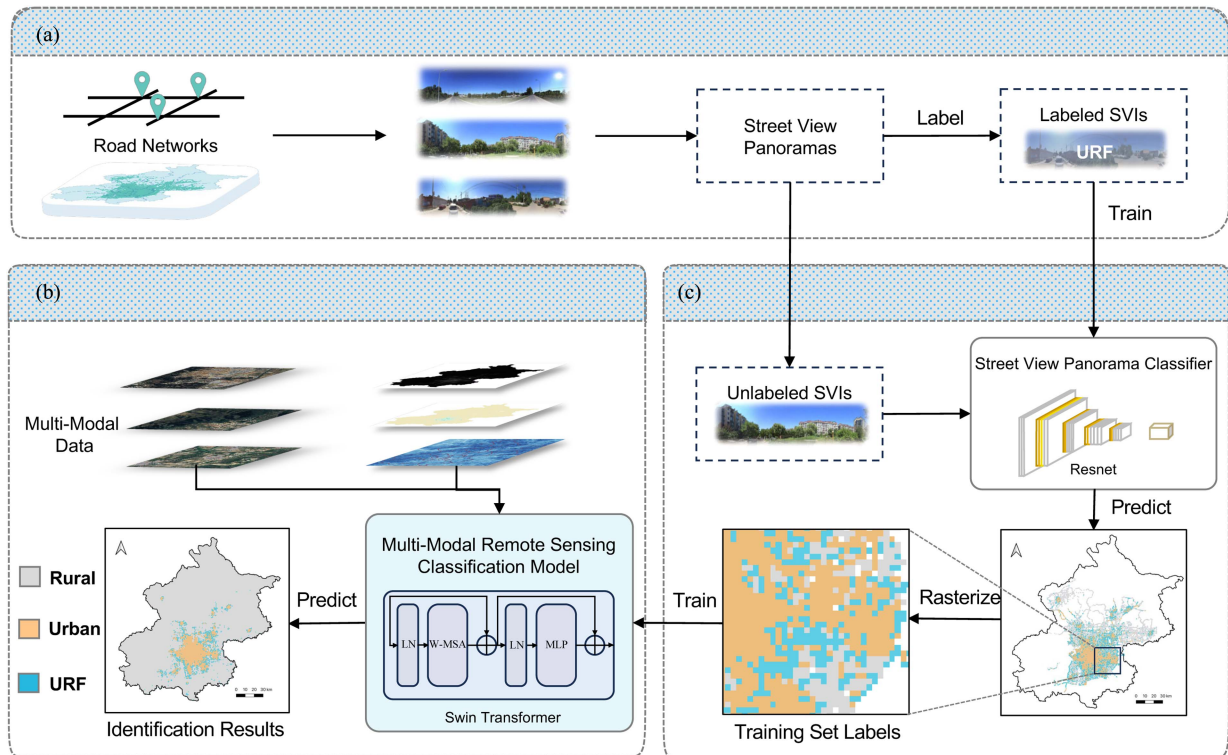


Fig. 1. Overall framework of the proposed method. (a) SVI data acquisition and labeling. (b) Street view panorama. (c) Transformer-based multimodal remote sensing classification.

B. Identifying URFs From Street View Panoramas

1) *Labeling the Street View Panorama Dataset:* Through delicate and human-centric streetscape visual information, SVIs offers opportunities for identifying URFs. Since there are no existing SVI datasets that label and distinguish the URFs from urban areas and rural areas, a labeled street view dataset was built for training and validating the street view panorama classifier in this study. Considering URF as a vague and complicated phenomenon [25], we build a multilevel street view-informed URF characterization framework for labeling SVIs upon [26], which has developed a generic ontological framework for image-based slum classification. This framework clarifies a set of visual indicators for URF identification at three levels: the environment, the settlement, and the object level. Combining the knowledge from authoritative definition, socioeconomic background, and professional experience, we can delineate a spectrum of typical streetscapes in URFs. The summarized description of the three levels is as follows and examples can be found in Fig. 2.

1) *Environment level:* URF is located at the intersection of highly urbanized areas and the rural hinterlands or less urbanized regions. This positioning means that SVIs from these areas often display a blend of urban and rural elements, particularly characteristics indicative of development, such as croplands interspersed with vacant lots and construction sites. The skyline is noticeably lower and less cluttered than in more urbanized locales, lacking skyscrapers and featuring only scattered buildings and structures.

- 2) *Settlement level:* The majority of settlements within URFs consist of informal settlements, often referred to as “urban villages” [27]. These are typically characterized by compact, overcrowded, self-constructed houses built by former villagers without official authorization or adherence to approved urban planning and building codes [28]. Consequently, SVIs often capture the high density and irregular layout of these settlements, including makeshift constructions [6], [29]. Many such settlements extend beyond standard regulations in terms of area and number of floors. Some feature additional floors and oversized balconies that encroach upon alleys. These architectural features, evident in SVIs, are adaptations by residents who modify and expand their properties to lease them out, which is driven by the high demand for affordable housing from the influx of low-income migrant workers.
- 3) *Object level:* The alleys depicted in SVIs are notably narrow due to the high-density construction of unauthorized buildings. Many roads lack urban greenery features such as trees along the sidewalks and are unsealed and uneven. The large influx of migrant workers residing in URF areas generates substantial amounts of waste, straining existing public facilities and services. The lack of timely waste removal results in accumulations of rubbish in these narrow alleys, visible in SVIs. Illegal building activities and renovations contribute to debris like bricks and concrete cluttering the laneways, which are also captured in street view imagery. The high population density of migrant workers places excessive demands on the already

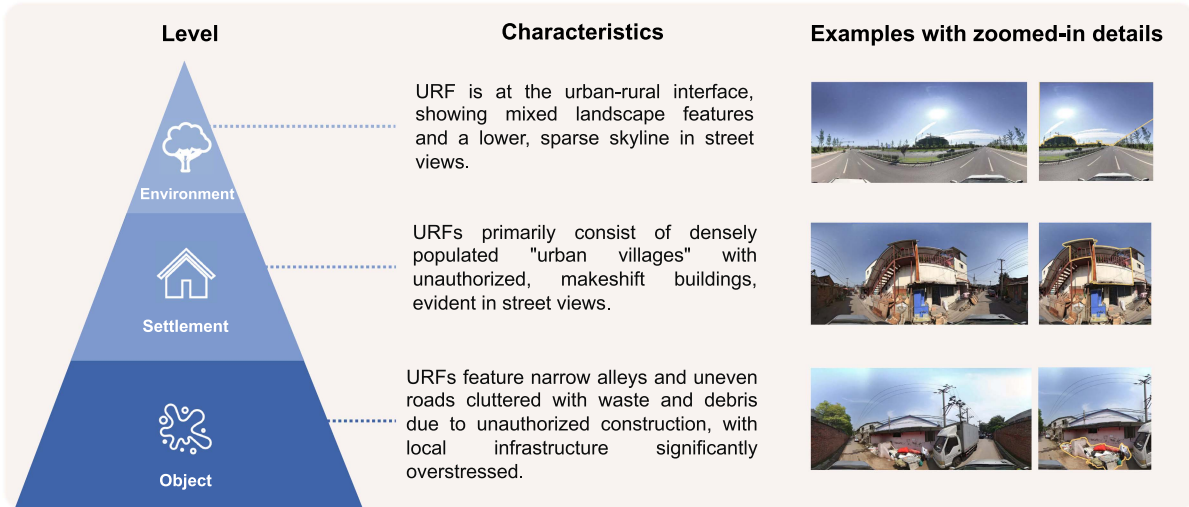


Fig. 2. Multilevel street view-informed URF characterization framework for labeling SVIs and street view examples.

inadequate infrastructure in URFs, leading to shortages in power and gas supplies. This has resulted in the proliferation of illegally installed wires crisscrossing overhead, further complicating the urban landscape.

Based on this framework, we randomly selected and labeled a total of 5800 street view panoramas from a dataset comprising 240 501 Beijing SVIs, effectively capturing streetscapes from the city center to the urban periphery. The labeled SVI dataset comprises 2255 urban images, 1830 village images, and 1715 URF images. Despite the uneven distribution of data points across different administrative districts—dense in the urban center and relatively sparse at the periphery—our stratified sampling method can balance the weights of each region in the sample by sampling a similar number of data points within each class in the rural–urban continuum. These images were initially labeled by volunteers and subsequently verified by experts, classified into three types, i.e., urban, rural areas, and URFs.

2) *Street View Panorama Classifier*: The labeled SVI dataset, derived from the Beijing SVI collection, was used to train the street view panorama classifier. Initially, it was divided into training and validation sets in an 8:2 ratio. This dataset was then employed for supervised learning to classify images into urban, URF, and rural categories. ResNet-34 [30] was adopted here due to its high effectiveness and relatively simple structure, considering the tradeoff between performance and cost.

The architecture of the street view panorama classifier consists of three main parts: the input, which includes street view panoramas scaled to (1024, 512, 3) for efficiency; the backbone, utilizing ResNet-34 to transform images into a (512, 16, 32) feature map; and the head, comprising an average pooling layer followed by a sequence of fully connected layers for class prediction.

After training and validation, all street view panoramas were processed by the classifier to automatically extend the classification results to the entire SVI dataset, which nearly cover the whole road network of Beijing. To transfer the classification result into spatially continuous, the classification results from

SVIs as discrete points were mapped into regular grids through rasterization, which can convert vector to rasters [see Fig. 1(b)]. This study employed a spatial resolution with the size of $1 \text{ km} \times 1 \text{ km}$ because it can effectively capture the characteristics of URF while also aligning with the resolution of multisource RS data. The category of each grid was determined based on the majority class of the street view sampling points located in the grid.

C. Transformer-Based Multimodal RS Classification Model

Due to the geographic constraints and the restriction of accessibility through road network, streetscape panoramas in some rural regions and urban rural fringes do not exist or are difficult to collect, resulting in the limited spatial coverage of the street view panoramas dataset and classification results through the street view panorama classifier, which can be completed through RS data due to the high coverage characteristics. Therefore, to obtain the land types in the whole city, a Transformer-based multimodal RS classification model was built with multisource remotely sensed data, including RS images, nighttime light intensity, NDVI, and population density, capturing features from natural environments to human socioeconomic behaviors. Using the geographic units in the street view reachable area as the training and validation set, the Transformer-based multimodal RS classification model was trained to classify and to obtain the land types for the entire city. The architecture of our proposed model is delineated in Fig. 3.

The process of multimodal RS classification model integrated two distinct modalities, A and B, processed through separate neural network architectures.

Modality A processes RS images using the Swin Transformer [21]. This architecture excels at capturing complex spatial hierarchies within the data, making it well-suited for our task where URFs vary in size, morphology, and environments. This module ultimately producing a 1024-D feature that represents the extracted characteristics of the RS imagery. Given a RGB

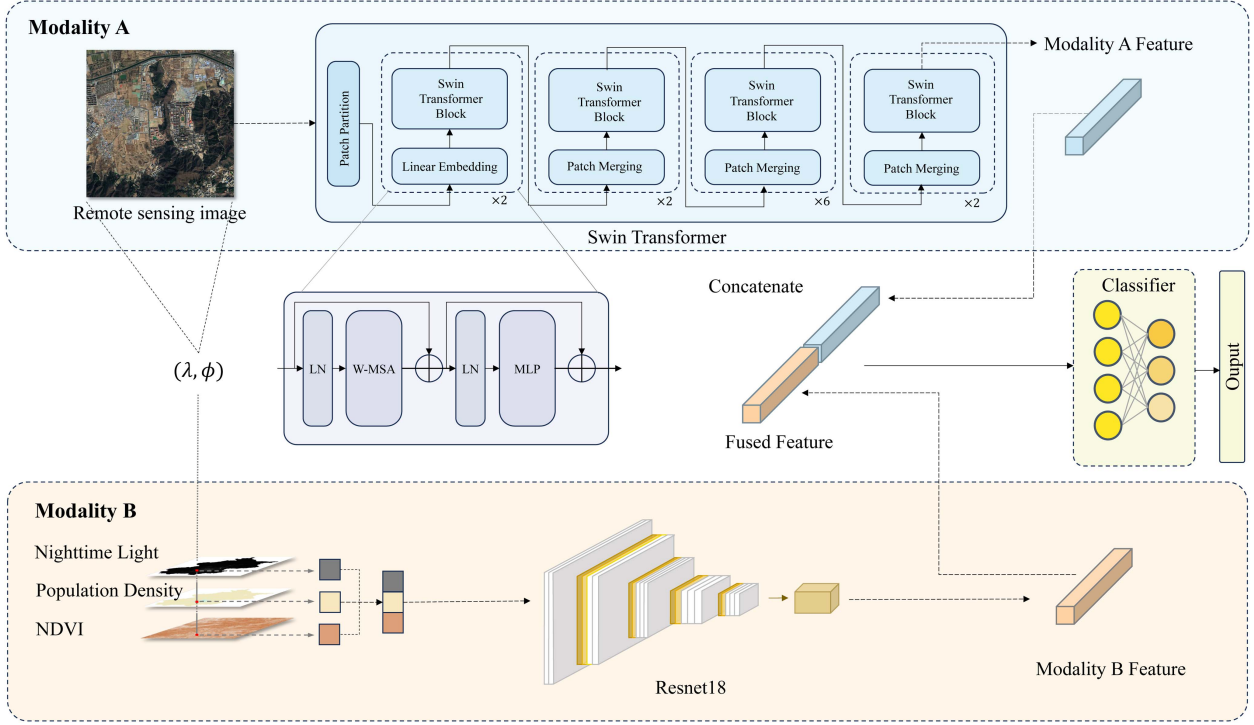


Fig. 3. Structure of Transformer-based multimodal RS classification model.

RS image of a $1 \text{ km} \times 1 \text{ km}$ grid within the city boundary, represented as $\mathbf{X}_A \in \mathbb{R}^{3 \times H \times W}$. We then employ Swin Transformer, denoted as \mathcal{F}_{ST} , for image feature extraction. The core of Swin Transformer are transformer blocks with modified self-attention computation. As illustrated in Fig. 3, a Swin Transformer block consists of a window-based multiheaded self-attention module (W-MSA), which can also be shifted window-based MSA (SW-MSA), followed by a two-layer MLP with GELU nonlinearity in between. A LayerNorm (LN) layer is applied before each MSA module and each MLP, and a residual connection is applied after each module. The Swin Transformer blocks are computed as

$$\begin{aligned} \hat{\mathbf{z}}^l &= \text{W-MSA}(\text{LN}(\mathbf{z}^{l-1})) + \mathbf{z}^{l-1} \\ \mathbf{z}^l &= \text{MLP}(\text{LN}(\hat{\mathbf{z}}^l)) + \hat{\mathbf{z}}^l \end{aligned} \quad (1)$$

where \mathbf{z}^{l-1} is the input feature of the Swin Transformer block, $\hat{\mathbf{z}}^l$ and \mathbf{z}^l denote the output features of the (S)W-MSA module and the MLP module for block l , respectively. The Modality A feature extracted through Swin Transformer can be expressed as

$$\mathbf{F}_A = \mathcal{F}_{\text{ST}}(\mathbf{X}_A) \in \mathbb{R}^d \quad (2)$$

where d is dimension of the output feature.

Modality B processes 3 channel auxiliary RS data including nighttime light intensity, population density, and NDVI, forming 3-D vector for an $1 \text{ km} \times 1 \text{ km}$ grid, represented as $\mathbf{X}_B \in \mathbb{R}^3$. Recent studies have embedded low-dimensional geographic data into high-dimensional neural representations, resulting in semantically rich, high-dimensional features that are well-suited for a wide range of geographic tasks [31]. Similarly,

the 3-D vector is then passed through a Resnet18 model, denoted as \mathcal{F}_{Res} , to get a 1024-D Modality B Feature $\mathbf{F}_B \in \mathbb{R}^d$

$$\mathbf{F}_B = \mathcal{F}_{\text{Res}}(\mathbf{X}_B \in \mathbb{R}^d). \quad (3)$$

The feature extracted from both modalities are then concatenated to form a fused RS feature \mathbf{F}_{RS} , which is given by

$$\mathbf{F}_{\text{RS}} = [\mathbf{F}_A, \mathbf{F}_B] \quad (4)$$

where $[\cdot]$ refers to the concatenating operation.

The fused RS feature encapsulates the comprehensive information captured from both the RS images and the auxiliary RS sources. Following the feeding of the fused feature into the fully connected layer classification module, the classification loss can be calculated using

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N_{\text{sup}}} \sum_{i=1}^{N_{\text{RS}}} y_i \log y'_i \quad (5)$$

where N_{RS} denotes the number of training samples with labels in each mini-batch during the training phase. y_i represents the ground truth of i th training sample and y'_i is predictive label.

III. MATERIALS AND EXPERIMENTS

A. Study Area and Data

1) *Study Area*: Beijing, the capital of the People's Republic of China, was selected as the subject of this study, which is located at the northwestern edge of the North China Plain ($39^\circ 28' - 41^\circ 05' \text{N}$, $115^\circ 25' - 117^\circ 30' \text{E}$) and covers an area of $16\,411 \text{ km}^2$. As of the end of 2010, Beijing consists of 16 municipal districts, which are divided into four layers according

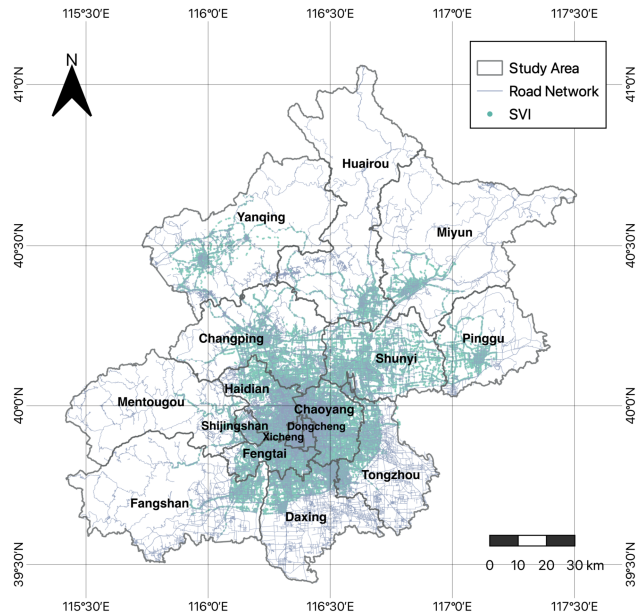


Fig. 4. Study area and spatial distribution of sample points of street view panorama.

to their functions from the inside to the outside: the core of the capital function (Dongcheng District and Xicheng District), the expanded urban function (Chaoyang District, Fengtai District, Haidian District, Shijingshan District), the new urban function (Changping District, Daxing District, Fangshan District, Shunyi District, Tongzhou District), and the Ecological Containment Zone (Huairou District, Mentougou District, Miyun District, Pinggu District, Yanqing District).

Since the 1990s, the built-up area of Beijing has been expanding as a result of rapid economic development and rapid urbanization. Urban expansion has also caused large-scale changes in urban land use. The central urban area has now crossed the two green belts initially designated by the municipal government to prevent the central urban area from being connected to the new development zone in the peripheral area [32], [33]. During urban sprawling, URFs with lots of illegal constructions and disorganized environments emerge in the green belts [28]. To increase the sustainability of urban development, the government has made policies to accelerate the urbanization of the URFs, in order to transfer the disorder into a more sustainable urban environment. Therefore, the quantitative identification of Beijing's URF is important for monitoring urbanization and optimizing urban governance.

2) *Datasets*: This study used multisource datasets to identify and analysis URFs in Beijing, including the street view panoramas, RS images, nighttime light data, NDVI, and population data. The street view panoramas came from Baidu Map open platform.¹ SVIs in panoramic form were obtained by setting the field of view (fov parameter in the Baidu API) to 360. As shown in Fig. 4, the street view panorama dataset, containing 240 501 images, was collected through the panoramic static view

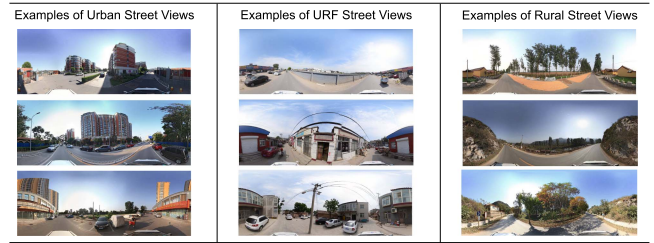


Fig. 5. Examples of classification results of the street view panorama classifier.

API,² covering photos taken along the road networks at 50 m intervals from 2013 to 2017. The 2017 RS imagery was acquired from Google Earth, encompassing the entirety of Beijing. The imagery boasts a fine spatial resolution of 2.39 m per pixel and includes three spectral bands: red, green, and blue. For the purposes of model integration, the imagery was uniformly divided into uniform segments, each measuring 1000 m \times 1000 m. The basic information of other datasets is shown in Table I.

B. Experimental Setting

1) *Implementation Details*: The street view panorama classifier was as implemented with a batch size of 16 and a learning rate initialized to 1e-4, and a training period of 30, with training performed on an NVIDIA GeForce RTX 2080 Ti GPU. The Swin Transformer model in the multimodal RS classification model was pretrained on ImageNet-21k [34], and the computational environment of the multimodal RS classification model is powered by a singular NVIDIA A800 GPU with 80G memory, with learning rate of 0.0005, batch size of 64 and 30 epoches of training. All codes are implemented with PyTorch framework [35].

2) *Evaluation Metrics*: In the experiments of the multimodal RS classification, four quantitative evaluation indices [i.e., accuracy for each category, OA, average accuracy (AA), and kappa coefficient (κ)] are applied to evaluate the classification performance. More specifically, OA is the ratio of correctly classified objects to the total objects, providing a global indication of the classification performance across all categories. AA is calculated by averaging the accuracies obtained for each individual class. The kappa coefficient is a statistical measure that compares an observed accuracy with an expected accuracy (random chance), which is particularly useful in situations where the classes are imbalanced.

IV. RESULTS

A. Identified URFs in Beijing

The street view panorama classifier, achieving an AA of 81.50% during the validation phase, effectively distinguishes URF streetscapes from urban and rural scenes. Examples of some classification results are illustrated in Fig. 5.

Through the Transformer-based multimodal RS classification model, the core urban area and the surrounding URF region can

¹[Online]. Available: <https://lbsyun.baidu.com>

²[Online]. Available: <https://lbsyun.baidu.com/index.php?title=viewstatic>

TABLE I
BASIC INFORMATION OF DATA USED IN THIS STUDY

Dataset	Time	Spatial resolution	Data source	Data description
Nighttime light data	2015	500 m × 500 m	https://www.ngdc.noaa.gov/eog/viirs	Global NPP-VIIRS nighttime light dataset
NDVI data	2013	500 m × 500 m	http://www.gscloud.cn	MODND1D datasets of the NDVI
Population data	2015	1000 m × 1000 m	https://www.worldpop.org	Global population distribution dataset from WorldPop
POI	2015	Vector data	Baidu	POIs of restaurants
Commuting data	2017	Vector data	China Unicom Smart Steps	Commuting flows extracted from the cell phone signaling data

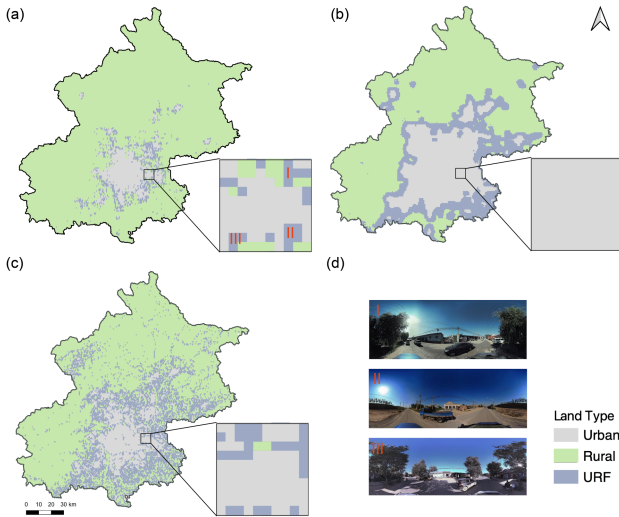


Fig. 6. Map comparison between the results of the (a) proposed model and those of, (b) K-means method, and (c) SOFM method. (d) SVI examples in the highlighted test region marked by Roman numerals in (a). All cells within the highlighted test region were identified as urban using the K-means method; however, image (d) clearly displays URF features in the area marked with Roman numerals.

be delineated can is shown in Fig. 6(a). The identified URFs generally surrounded the central city area, showing a ring-like pattern with a total area of about 731.24 km². The population of this area contained the smallest number of the three parts, accounting for 14.13% of the city's resident population. Most of the URFs were distributed in the new urban development zones, and a small part was located in the periphery of the urban function expansion areas, such as Haidian District, Shijingshan District, and Fengtai District. Meanwhile, some identified URFs were also distributed around satellite cities, such as Yanqing District. In addition, there were some URF areas in the ecological conservation zones that were not connected to urban areas. These URFs were generally evolved from rural areas, and their level of urbanization exceeded that of the rural areas but was still not up to that of the urban. In the future, with the continuous urbanization process, these areas have the possibility of developing into urban areas. The rural population constituted only 22.72% of the city's resident population, and the rural area comprised the majority of Beijing's land (87.83%) and was concentrated in the city's west and north, in accordance with the city's topographic features, which were elevated in the northwest and low in the southeast. Primarily situated within ecological conservation zones, these locations were predominantly characterized by mountainous and forested terrain, reaching elevations above 2000 m. Consequently, they were deemed unsuitable for urban development.

B. Ablation Studies

In order to measure the efficacy of using multimodal data and study the interpretability of different data sources in identifying URF, we trained unimodal (only RS imagery, referred as RI in the table later or only auxiliary data) and bimodal models (RS imagery integrated with different channels of auxiliary RS data) for comparison with the multimodal RS classification model. To ensure an equitable comparison, all models used identical training and validation datasets along with the same model hyperparameters. The results of this analysis on validation set are displayed in Table II, where the best results are denoted in bold.

The results demonstrate that the proposed model with all data source from two different modality outperforms other unimodal and bimodal models in terms of OA and kappa coefficient, achieving an OA of 72.7%, kappa coefficient of 0.57. This result demonstrates that integrating data from multiple sources across two modalities enhances the OA and provides comprehensive information for effectively recognizing URF. The bimodal models based on RS imagery and one channel RS auxiliary data also performs well, outperforming the proposed model in specific accuracy in terms of category-specific accuracy, which underscores the distinct characteristics and explanatory power of different RS data in recognizing specific land types. Combining RS imagery and nighttime light data can identify urban areas with highest accuracy, since nighttime light data can distinguish urban areas by detecting city lights from traffic, residential zones, and similar sources, clearly setting them apart from the dark backgrounds of rural areas. Combining RS imagery with NDVI data enables the detection of rural areas with the highest accuracy. This efficacy can be attributed to NDVI's capability to identify green vegetation, a predominant feature of rural landscapes characterized by cultivated land and forests. Combining RS imagery with population data achieves the highest accuracy in identifying URF areas. This effectiveness is likely because population density, a fundamental socioeconomic attribute, can significantly distinguish URFs, since URF areas often experience a marked decrease in population density as they transition into rural fringes notably. However, the unimodal models demonstrate limited accuracy both overall and in category-specific assessments, emphasizing the importance of integrating multiple data sources to enhance OA.

C. Comparative Experiments

To evaluate the effectiveness of our proposed method, we conducted a series of comparative experiments against existing URF identification techniques. Among the earlier methods, the K-means approach utilized in [5] employed nighttime light as

TABLE II
OA, AA, κ , AND THE ACCURACY OF PER CLASS FOR DIFFERENT INPUT DATA IN WHICH THE BEST RESULTS ARE DENOTED IN BOLD

Class	RI	NTL+NDVI+Pop	RI+NTL+NDVI	RI+NTL+Pop	RI+NDVI+Pop	RI+NTL	RI+NDVI	RI+Pop	Proposed
Urban	67.68	57.92	69.85	65.94	65.08	72.23	65.94	64.43	69.85
Rural	79.34	73.55	79.97	80.13	85.13	81.53	86.70	81.38	82.79
URF	56.23	60.14	54.09	62.28	49.1	42.70	45.20	63.70	54.45
OA	70.75	67.78	71.33	71.76	71.11	70.53	71.33	72.12	72.70
AA	67.75	65.10	67.97	69.45	66.44	65.49	65.95	69.83	69.03
$\kappa \times 100$	54	50	55	56	54	53	54	57	57

a data source to extract features, such as light intensity and fluctuation for clustering land in Beijing into urban, URF, and rural regions. Another method, the SOFM, used in research in Peng et al. [3], processed features based on land use data to delineate URFs through clustering, denoted as the SOFM method. These methods were chosen based on their relevance and proven effectiveness in similar urban–rural studies, their compatibility with our study area, and the availability of comprehensive datasets. We replicated these two previous studies to identify URFs in Beijing, using data from the same time period as our study to ensure a fair comparison. In addition, we labeled 25 388 SVIs within a specified test region to assess the accuracy of various URF identification methods. After rasterizing the SVI labels for land grid categories and calculating performance, our proposed method achieved an accuracy of 65.58%. In comparison, the K-means method and SOFM method yielded accuracies of 50.19% and 60.41%, respectively. These results indicated that our proposed method outperformed the others in URF identification. Fig. 6 maps the distinctions between the proposed method and the comparative methods.

Differences of identification within a specific square test region were zoomed in and highlighted. All cells within the highlighted test region were identified as urban according to K-means method, indicating an outward expansion trend of the urban region as recognized by this method. This suggests a more urbanized situation characterized by further urban sprawl, which may deviate from actual conditions. Examples of SVIs within the test region are displayed in the Fig. 6(d). Clear URF features are evident in the streetscapes, such as the coexistence of modern buildings and traditional rural single-story houses, low and sparse skyline, and roads that are unhardened and uneven. These characteristics epitomize typical URF landscapes, distinctly unlike those of urban regions, which were not detected by other comparative methods. The comparative experiments suggest that our approach identifies URFs in a more refined and accurate manner.

V. DISCUSSION

In this study, we identified the URFs in Beijing using a multimodal approach. Socioeconomic dimensions, as potential internal drivers shaping URFs, are crucial for understanding these areas. Investigating socioeconomic issues, such as disaster resilience, food insecurity, and the availability of educational and medical resources, within the rural–urban continuum is essential. Analyzing these factors at various points along the continuum can provide valuable insights for urban policy development and governance. Our quantitative identification of URFs facilitates the use of analytical tools to specify their

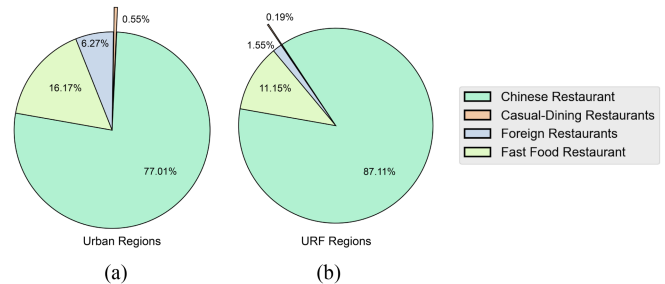


Fig. 7. Proportions of each restaurant type in the (a) urban and (b) URF region.

traits. Moreover, URFs serve as significant agglomeration zones for migrants, where the consumption patterns of housing and energy, as well as aspects like jobs-housing relationships, differ markedly from those in the urban center [36]. Based on the identification results, our study delves deeper into the human-related aspects of URFs. We employ social sensing data, including points of interest (POIs) and cell phone signaling data, to analyze restaurant consumption and commuting patterns. This approach provides a richer understanding of URF as well as the distinctions between URFs and urban regions.

A. Restaurant Consumption Structure in URFs

Based on the identification results, this study leverages daily relevant restaurant POI data from 2015 to explore the socioeconomic characteristics of URF regions. Recent research supports the use of such data as a predictive tool for socioeconomic attributes: Dong et al. [37] demonstrate that specific restaurant attributes can effectively predict socioeconomic characteristics of urban neighborhoods. Similarly, Liao et al. [38] find that the presence of western food restaurants is associated with higher levels of urbanization.

By analyzing the restaurant data, this study compares the spatial distribution and consumption structures between the URF and urban regions. The proportions of each restaurant type were calculated in the URF versus the urban region and found that there are significant differences (see Fig. 7). The findings show significant contrasts: Chinese restaurants are prevalent in both settings but dominate the URF at 87.11% compared to 77.01% in urban areas. This suggests less diversity in the URF, where fast food outlets are notably rarer and foreign restaurants much less common than in urban areas. The urban regions exhibit a higher diversity and balance of restaurant types, supporting the hypothesis that varied dining options correlate with urban socioeconomic trends.

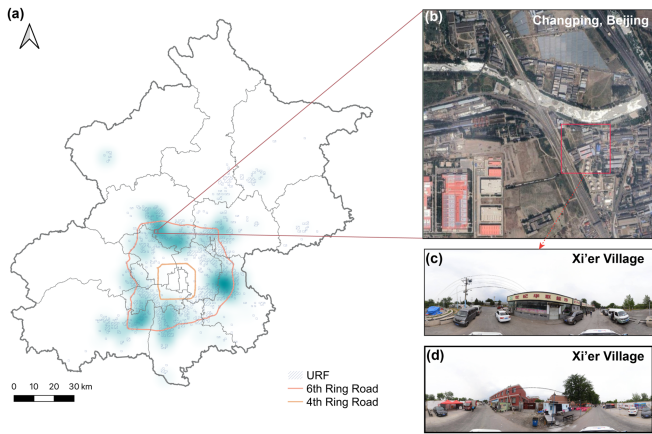


Fig. 8. Compare the heatmap of the commuting patterns to the spatial distribution of the URFs.

B. Commuting Pattern in URFs

In addition, the proposed method analyzed the identified URF and compared it with the urban region from a perspective of human mobility. Leveraging the commuting flows derived from the 2017 Beijing cell phone signaling data, which were aggregated via a 1 km grid, this study computed the average lengths of the outflows for each grid unit. This approach provided insight into the commuting patterns within the region. After excluding grids with fewer than 5000 commuters, units with an average commuting distance greater than 10 km were selected. A heatmap can then be created to illustrate the regions that feature extremely long-distance commuters. Comparing this heatmap to Fig. 8, it is found that the identified URF region is highly intersected with most of the highlighted regions, distributed between the 4th Ring Road and the 6th Ring Road of Beijing, showing the long-distance commuting pattern of URF. Fig. 8 also presents RS imagery and SVIs in a highlighted region, which depict the URF characteristics. Notably, the skyline is lower and less cluttered compared to more urbanized areas, featuring construction sites and predominantly low-rise buildings and bungalows. Research has indicated that low-income populations residing in URFs are more likely to undertake long-distance travel due to a lack of employment opportunities and transportation amenities [39], which is consistent with this study. However, without precise identification of URF regions based on geographic data, their unit divisions, including the category of urban fringe areas, were based solely on coarse-grained land planning schemes, which constrains the accuracy of the results. Our recognition provides new opportunities for further quantitative analysis. Overall, the socioeconomic characteristics extracted from both POIs and commuting data through quantitative analysis, give a further look into the human-related features of URF than previous URF identification studies.

VI. CONCLUSION

URFs are transitional regions between urban built-up areas and rural hinterlands. Given their status as focal points of intense social conflict and environmental protection challenges,

accurately identifying URFs is critically important for urban sustainability studies. Remotely sensed data have emerged as the primary data source for URF identification. However, previous studies have frequently ignored the abundant visual information provided by SVIs, and have not explored multimodal deep learning techniques to integrate multisource data for delineating URFs. Therefore, to accurately identify URFs, this study proposes a Transformer-based multimodal approach that integrates street view imagery with remotely sensed data.

Compared to traditional URF identification methods, the proposed multimodal approach integrates both macroscopic and microscopic perspectives by combining street view imagery with remotely sensed data. Evaluations conducted in Beijing demonstrate that this multimodal method outperforms previous methods, which relied primarily on RS data-based clustering method, by more than 5%. To validate the effectiveness of using multisource data, ablation experiments were conducted. These experiments highlight the efficacy and advantages of integrating different data sources for a more comprehensive understanding of URFs. Based on the identification results, our study delves deeper into the human-related aspects of URFs, revealing that extreme commuting is prevalent in these areas, and residents have less diverse restaurant options compared to those in more urbanized regions.

This study accurately identifies URFs by introducing a Transformer-based multimodal approach that leverages street view imagery to provide streetscape comprehension. This approach uniquely integrates microscale, human-centric perspectives, which have been largely absent in previous studies of URF identification. In the future, the proposed methodology can be applied to identify URFs in other cities, further enhancing the interpretation of urbanization dynamics and contributing significantly to the study of urban sustainability.

ACKNOWLEDGMENT

The author Furong Jia would like to thank Weiyi Jiang and Yunqiao Li for their valuable suggestions in visualization. The author would also like to thank Haoshen Li, Lanxin Liu, and Yajing Wang for their discussions on computing methodology and to Yingjing Huang for her insightful discussions about the framework.

REFERENCES

- [1] C. Bittner and M. Sofer, "Land use changes in the rural-urban fringe: An Israeli case study," *Land Use Policy*, vol. 33, pp. 11–19, 2013.
- [2] W. Ma, G. Jiang, W. Li, and T. Zhou, "How do population decline, urban sprawl and industrial transformation impact land use change in rural residential areas? A comparative regional analysis at the peri-urban interface," *J. Cleaner Prod.*, vol. 205, pp. 76–85, 2018.
- [3] J. Peng et al., "Integrating land development size, pattern, and density to identify urban-rural fringe in a metropolitan region," *Landscape Ecol.*, vol. 35, pp. 2045–2059, 2020.
- [4] A. Cattaneo, A. Nelson, and T. McMenomy, "Global mapping of urban-rural catchment areas reveals unequal access to services," *Proc. Nat. Acad. Sci.*, vol. 118, no. 2, 2021, Art. no. e2011990118.
- [5] Z. Feng, J. Peng, and J. Wu, "Using dmsp/ols nighttime light data and k-means method to identify urban-rural fringe of megacities," *Habitat Int.*, vol. 103, 2020, Art. no. 102227.

- [6] W.-S. Tang and H. Chung, "Rural-urban transition in China: Illegal land use and construction," *Asia Pacific Viewpoint*, vol. 43, no. 1, pp. 43-62, 2002.
- [7] J. Peng et al., "A new approach for urban-rural fringe identification: Integrating impervious surface area and spatial continuous wavelet transform," *Landscape Urban Plan.*, vol. 175, pp. 72-79, 2018.
- [8] Z. Fan, F. Zhang, B. P. Loo, and C. Ratti, "Urban visual intelligence: Uncovering hidden city profiles with street view images," *Proc. Nat. Acad. Sci.*, vol. 120, no. 27, 2023, Art. no. e2220417120.
- [9] F. Zhang et al., "Urban visual intelligence: Studying cities with artificial intelligence and street-level imagery," *Ann. Amer. Assoc. Geographers*, vol. 114, no. 5, pp. 876-897, 2024.
- [10] F. Zhang et al., "Uncovering inconspicuous places using social media check-ins and street view images," *Comput., Environ. Urban Syst.*, vol. 81, 2020, Art. no. 101478.
- [11] P. Sturges, K. Alahari, L. Ladicky, and P. H. Torr, "Combining appearance and structure from motion features for road scene understanding," in *Proc. BMVC-Brit. Mach. Vis. Conf.*, 2009, pp. 1-11.
- [12] F. Zhang, L. Wu, D. Zhu, and Y. Liu, "Social sensing from street-level imagery: A case study in learning spatio-temporal urban mobility patterns," *ISPRS J. Photogrammetry Remote Sens.*, vol. 153, pp. 48-58, 2019.
- [13] F. Zhang, D. Zhang, Y. Liu, and H. Lin, "Representing place locales using scene elements," *Comput., Environ. Urban Syst.*, vol. 71, pp. 153-164, 2018.
- [14] Y. Zhang, N. Chen, W. Du, Y. Li, and X. Zheng, "Multi-source sensor based urban habitat and resident health sensing: A case study of wuhan, China," *Building Environ.*, vol. 198, 2021, Art. no. 107883.
- [15] Y. Zhang, P. Liu, and F. Biljecki, "Knowledge and topology: A two layer spatially dependent graph neural networks to identify urban functions with time-series street view image," *ISPRS J. Photogrammetry Remote Sens.*, vol. 198, pp. 153-168, 2023.
- [16] R. Cao et al., "Deep learning-based remote and social sensing data fusion for urban region function recognition," *ISPRS J. Photogrammetry Remote Sens.*, vol. 163, pp. 82-97, 2020.
- [17] S. Srivastava, J. E. Vargas-Munoz, and D. Tuia, "Understanding urban land-use from the above and ground perspectives: A deep learning, multimodal solution," *Remote Sens. Environ.*, vol. 228, pp. 129-143, 2019.
- [18] Y. Huang et al., "Comprehensive urban space representation with varying numbers of street-level images," *Comput., Environ. Urban Syst.*, vol. 106, 2023, Art. no. 102043.
- [19] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, 2022, Art. no. 200.
- [20] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000-6010.
- [21] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012-10022.
- [22] Z. Li et al., "A large scale digital elevation model super-resolution transformer," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 124, 2023, Art. no. 103496.
- [23] D. H. Lee, H. Y. Park, and J. Lee, "A review on recent deep learning-based semantic segmentation for urban greenness measurement," *Sensors*, vol. 24, no. 7, 2024, Art. no. 2245.
- [24] W. Li, C.-Y. Hsu, and M. Tedesco, "Advancing arctic sea ice remote sensing with ai and deep learning: Now and future," *EGU sphere*, vol. 2024, pp. 1-36, 2024.
- [25] Y. Liu, Y. Yuan, and S. Gao, "Modeling the vagueness of areal geographic objects: A categorization system," *ISPRS Int. J. Geo-information*, vol. 8, no. 7, 2019, Art. no. 306.
- [26] D. Kohli, R. Sliuzas, N. Kerle, and A. Stein, "An ontology of slums for image-based classification," *Comput., Environ. Urban Syst.*, vol. 36, no. 2, pp. 154-163, 2012.
- [27] S. Zheng, F. Long, C. C. Fan, and Y. Gu, "Urban villages in China: A 2008 survey of migrant settlements in Beijing," *Eurasian Geogr. Econ.*, vol. 50, no. 4, pp. 425-446, 2009.
- [28] P. Zhao and M. Zhang, "Informal suburbanization in Beijing: An investigation of informal gated communities on the urban fringe," *Habitat Int.*, vol. 77, pp. 130-142, 2018.
- [29] F. Wu, "Housing in Chinese urban villages: The dwellers, conditions and tenancy informality," *Housing Stud.*, vol. 31, no. 7, pp. 852-870, 2016.
- [30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770-778.
- [31] V. Vivanco Cepeda, G. K. Nayak, and M. Shah, "Geoclip: Clipinspired alignment between locations and images for effective worldwide geolocalization," *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 8690-8701, 2024.
- [32] M. Zhang, X. Meng, L. Wang, and T. Xu, "Transit development shaping urbanization: Evidence from the housing market in Beijing," *Habitat Int.*, pp. 8690-8701, 2014.
- [33] P. Zhao, "Too complex to be managed? New trends in peri-urbanisation and its planning in Beijing," *Cities*, vol. 30, pp. 68-76, 2013.
- [34] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "Imagenet-21 k pretraining for the masses," in *Proc. 35th Int. Conf. Neural Inf. Process. Syst. (NeurIPS 2021) Track Datasets Benchmarks*, 2021.
- [35] A. Paszke et al., "Automatic differentiation in pytorch," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017.
- [36] R. Yang, "Space reconstruction process and internal driving mechanisms of taobao villages in metropolitan fringe areas: A case study of lirendong village in Guangzhou, China," *J. Geographical Sci.*, vol. 32, no. 12, pp. 2599-2623, 2022.
- [37] L. Dong, C. Ratti, and S. Zheng, "Predicting neighborhoods-socioeconomic attributes using restaurant data," *Proc. Nat. Acad. Sci.*, vol. 116, no. 31, pp. 15447-15452, 2019.
- [38] C. Liao et al., "City level of income and urbanization and availability of food stores and food service places in China," *PLoS One*, vol. 11, no. 3, 2016, Art. no. e0148745.
- [39] P. Zhao and J. Wan, "Land use and travel burden of residents in urban fringe and rural areas: An evaluation of urban-rural integration initiatives in Beijing," *Land Use Policy*, vol. 103, 2021, Art. no. 105309.



Furong Jia received the B.S. degree in computer science from Beijing Normal University, Beijing, China, in 2022. She is currently working toward the Ph.D. degree in cartography and geographical information systems with the School of Earth and Space Sciences, Peking University, Beijing.

Her research interests include urban science, geographical artificial intelligence, and spatial cognition.



Quanhua Dong received the Ph.D. degree in surveying, mapping and remote sensing from the State Key Laboratory of Information Engineering, Wuhan University, Wuhan, China, in 2021.

She is currently an Associate Research Fellow with the Institute of Remote Sensing and Geographical Information Systems, Peking University, Beijing, China. Her research interests include 3-D spatial analysis, urban science, and deep learning.



Zhou Huang received the B.Sc. degree in GIS and the Ph.D. degree in cartography and GIS from Peking University, Beijing, China, in 2004 and 2009, respectively.

He is a tenured Associate Professor of GIScience with the Institute of Remote Sensing and Geographic Information System, Peking University. In addition, he is the Deputy Director with the Institute of Remote Sensing and GIS, Peking University, and Deputy Director of Engineering Research Center of Earth Observation and Navigation, Ministry of Education, China.

He has authored or coauthored more than 100 academic papers in international journals or conferences. His main research interests include big geo-data, high-performance geocomputation, distributed geographic information processing, spatial data mining, and spatial database.

Dr. Huang was selected for the Youth Talent Innovation Plan in Remote Sensing Science and Technology, funded by the Ministry of Science and Technology of China.



Xiao-Jian Chen received the B.S. degree in mathematics from the School of Mathematics and Statistics, Wuhan University, in 2013, and the Ph.D. degree in surveying, mapping, and remote sensing from the State Key Laboratory of Information Engineering, Wuhan University, Wuhan, China, in 2021.

He is currently an Associate Researcher with the Institute of Remote Sensing and Geographical Information Systems, School of Earth and Space Sciences, Peking University, Beijing, China. His research interests include human mobility and spatial network.



Yi Wang received the B.S. and the M.S. degrees in surveying and mapping from the Beijing University of Civil Engineering and Architecture, Beijing, China, in 2019 and 2022, respectively. He is currently working toward the Ph.D. degree in cartography and geographical information systems with the School of Earth and Space Sciences, Peking University, Beijing.

His main research interests include GeoAI, spatiotemporal modeling, time-series forecasting, and social sensing.



Yuan Guo received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2021.

She is currently a Lecturer with the School of Resource and Environment Engineering, Wuhan University of Technology, Wuhan. Her research interests include high-definition map, 3-D spatial analysis, and computer vision.



Xia Peng received the B.S. degree in geographical information system from the China University of Geosciences, Wuhan, China, in 2004, the M.S. degree in cartography and geographical information system from Peking University, Beijing, China, in 2007, and the Ph.D. degree in urban and rural planning from Tsinghua University, Beijing, in 2013.

She is currently an Associate Professor with the Tourism College, Beijing Union University, Beijing. Her major research interests include spatio-temporal data mining, GIS, and tourism decision support system.



Ruixian Ma received the B.A. degree in graphic design from the University of the Arts London, London, U.K., in 2011.

In 2013, he studied Information Experience Design with the Royal College of Art, London. In 2016, he joined the MIT Senseable City Lab as a Research Fellow. He was an engineering Manager with Alibaba Group. Currently, he is the Visualization Manager with the MIT Senseable City Lab. His research interests include generative visualization tools and the urban aesthetic mining based on spatial data.



Fan Zhang received the B.S. degree in electronic engineering from Beijing Normal University, Zhuhai, China, in 2012, and the Ph.D. degree in Earth geoinformation science from the Chinese University of Hong Kong, Hong Kong, in 2017.

He is currently an Assistant Professor of GIScience with the Institute of Remote Sensing and Geographical Information Systems, School of Earth and Space Sciences, Peking University, Beijing, China. His research interests include Data-driven urban studies and geographical artificial intelligence.



Yu Liu received the B.S. and M.S. degrees in human geography from the School of Urban and Environment, Peking University, in 1994 and 1997, respectively, and the Ph.D. degree in software engineering from the School of Computer Science and Technology, Peking University, Beijing, China, in 2003.

He is currently a Boya Professor of GIScience with the Institute of Remote Sensing and Geographical Information Systems, School of Earth and Space Sciences, Peking University. His research interests include humanities and social science based on big

geodata.