This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3326160

MDFF: A method for fine-grained UFZ mapping with multimodal geographic data and deep network

Song Ouyang, Shihong Du, Xiuyuan Zhang*, Shouhang Du, and Lubin Bai

***Abstract*—As basic units of urban areas, urban functional zones (UFZs) are fundamental to urban planning, management, and renewal. UFZs are mainly determined by human activities, economic behaviors, and geographical factors, but existing methods 1) do not fully use multimodal geographic data owing to a lack of semantic modeling and feature fusion of geographic objects and 2) are composed of multiple stages, which leads to the accumulation of errors through multiple stages and increases the mapping complexity. Accordingly, this study designs a multimodal data fusion framework (MDFF) to map fine-grained UFZs end-to-end, which effectively integrates very-high-resolution remote sensing images and social sensing data. The MDFF extracts physical attributes from remote sensing images and models socioeconomic semantics of geographic objects from social sensing data, and then fuses multimodal information to classify UFZs where object semantics guide the fine-grained classification. Experimental results in Beijing and Shanghai, two major cities of China, show that the MDFF greatly improves the quality of UFZ mapping with the accuracy about 5% higher than state-of-the-art methods. The proposed method significantly reduces the complexity of UFZ mapping to complete the urban structure analysis conveniently.**

***Index Terms*—UFZ mapping, multimodal geographic data fusion, image classification, remote sensing, deep learning**

I. Introduction

URBAN functional zones (UFZs) are spatially represented by zones with the same social functions, such as commercial, residential, and industrial zones [1]. UFZs are often employed as basic units to analyze urban spatial structures and social characteristics [2], which have become increasingly crucial for urban planning [3][4], e.g., transportation planning, factory relocation, environmental protection, and sustainable development [5][7]. However, fine-grained UFZ maps with large areas are hardly available [8] due to 1) the lack of methods for fine-grained UFZ mapping and 2) unreliable interpretation of UFZs. The rapid development of urbanization and economy has brought a huge demand for timely updating UFZ maps. Therefore, urban studies urgently need to break through the existing difficulties to obtain large-scale and fine-grained UFZ maps that keep up with the pace of urban development in a timely and accurate manner [9].

As functions of urban zones are mainly determined by human activities, economic behaviors, and geographical factors, UFZs present complex physical structures and socioeconomic properties, which makes the UFZ mapping full of challenges. With the development of earth observation technologies, lots of high-quality remote sensing (RS) images are produced, especially for very-high-resolution satellite images (VHR) which hold advantages in representing UFZs, because of the large spatial coverage, detailed information, and wide availability [10][11]. However, it is challenge to accurately and reliably interpret UFZs relying on a single source data. For example, with only the satellite image (Fig.1), it is difficult to decide whether the blue area in (a) is a residential or a commercial area and which category the shadow area of (b) belongs to, because UFZs are not only related to physical attributes, but also to socioeconomic attributes.

There have been substantial multimodal geographic data to represent UFZs, such as image data (VHR, synthetic aperture radar images, nightlight images, etc.) and social sensing data (e.g., points of interest, mobile signal data, social media data), which constitute complementary information sources [12]. RS images provide pixel-level characteristics but lack information on the characterization of geographic objects, while social sensing data contain socioeconomic attributes and correlations of geographic objects. Therefore, these two types of data are complementary. Comprehensive utilization of multimodal information can help to reduce the uncertainty of interpretation. With socioeconomic properties of points of interest (POIs) attached to the satellite image in Fig.1, the blue area in (a) and the shadow area in (b) are easily assigned as the commercial and the transport, respectively. However, existing studies do not full fuse multimodal semantics to assist the classification and are complex for multi-stage processes, which significantly limits the quality and reliability of UFZ mapping. To map fine-grained UFZs end-to-end with multi-modal data, three challenging issues must be resolved, i.e., multimodal feature representation, multimodal feature fusion, and urban-function classification.

This work was supported by the Chinese National Nature Science Foundation (No. 42001327) (*Corresponding author*: *Xiuyuan Zhang*)

S. Ouyang, S. Du, X. Zhang, S. Du and L. Bai are with the Institution of Remote Sensing and Geographic Information System, School of Earth and Space Sciences, Peking University, Beijing 100871, China. (e-mail: song_ouyang@outlook.com; shdu@pku.edu.cn.com; xy_zhang@pku.edu.cn; dsjcug@163.com; lbbai@stu.pku.edu.cn)

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3326160

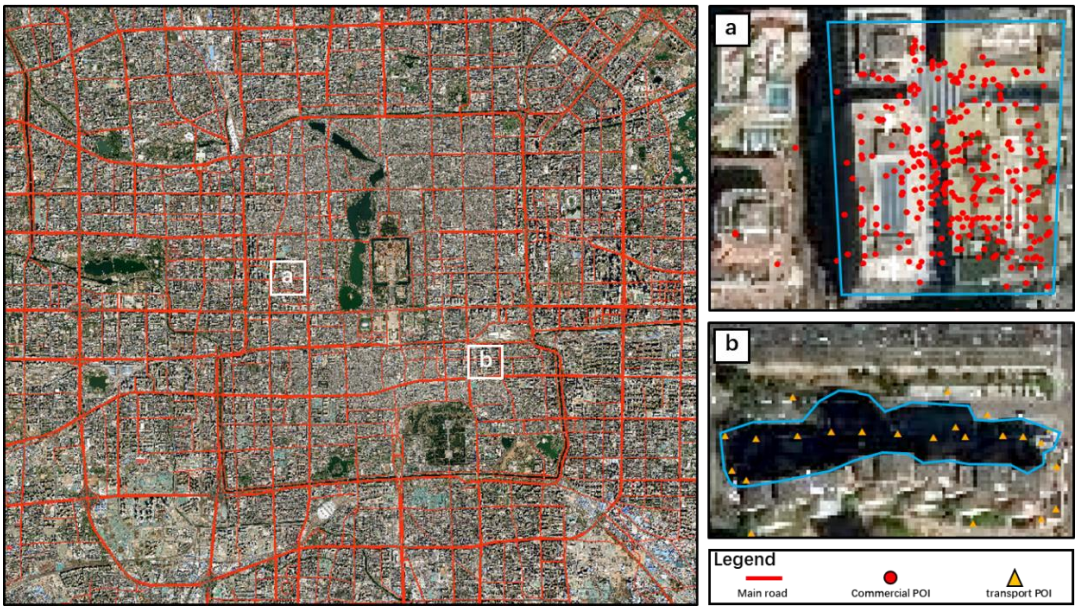


Fig. 1. Examples of interpreting UFZs in satellite images.

A. Multimodal feature representation

Appearance features extracted from images, such as spectrums, textures, and geometry, are adopted in the early land cover and urban function classification [13]-[15]. These features are shallow-level features because of their weak expressiveness and robustness. For better discrimination, middle-level cues are generated from shallow-level features by the developed methods such as the principal component analysis, the probability latent semantic analysis, the support vector machine, and the latent Dirichlet allocation [16][19]. With the rapid progress of deep learning, convolutional neural networks (CNNs) have great advantages in extracting deep-level features from images for their powerful learning ability [20]-[22]. Zhou et al. (2020) employed a trained CNN to assign functional attributes with RS images [23]. Therefore, deep features from CNNs can effectively characterize UFZs' physical properties. Unlike RS images, social sensing data contain socioeconomic semantics which are important for UFZ interpretation. However, socioeconomic semantics are difficult to extract due to discrete arrangement of social sensing data and the lack of modeling methods. Existing studies mainly used social sensing data as training labels of classifiers, samples for accuracy evaluation, or shallow-level features [24][25]. It not only failed to mine the high-level information of social sensing data, but also discarded semantic relationships between objects. Thus, there is an urgent need for a modeling approach within social sensing data to extract UFZs' socioeconomic semantics.

B. Multimodal feature fusion

It is challenge to fusing multimodal features especially for RS images and social sensing data as these two types of data are arranged in completely different ways. The fusion methods are divided into two types: the pixel-wise overlay approaches and the object-based methods [27]. The former directly concatenates images and the density maps of social sensing data pixel by pixel. Some studies follow this way [28], but these pixel-wise approaches only fuse at the pixel level and lose the object semantics [29][30]. The latter extracts attributes of social sensing data to characterize objects for classifiers [31][32]. Based on the units of road blocks, Zhang et al. (2017b) calculated discrete features from POIs and social media data, then combined them with spectral attributes to map land uses [12]. Du et al. (2020) merged spectral features and semantic features at the object level for large-scale UFZ mapping [24]. However, existing methods mainly fuse features at a single scale (pixel scale or object scale), and shallow fusion at the attribute level rather than semantic level cannot mine objects' semantic information from multimodal geographic data. The pixel-wise approaches provide detailed features in pixel space, and the object-based methods are suitable for expressing object semantics [33]. Therefore, integrating information at both pixel and object levels to characterize UFZs is critical for UFZ mapping. Therefore, it is urgent to develop fusion methods that consider joint features and the high-level semantics of multimodal geographic data under the two levels [34]-[36].

C. Urban-function classification

Due to its importance to urban space planning, urban-function classification has attracted considerable attention in the past decade [9]. On the one hand, RS images and social sensing data are employed to map UFZs in early attempts. However, multimodal geographic data were not fully utilized in UFZ mapping due to the insufficient feature expression. On the other hand, each UFZ is composed of geographic objects [37], such as the commercial zone in (a) of Fig.1 including buildings and roads. Therefore, it has become the mainstream method that segments land cover objects at first and then aggregates them into UFZs. In the hierarchical semantic cognition model [12], objects in the

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3326160

model are segmented with the multiresolution segmentation method [38], and then developed the inverse hierarchical semantic cognition model for refining the results at the land-cover layer [39]. Du et al. (2021) designed a CNN to extract pixel-wised predictions for initializing UFZ categories of geographic objects, then used the conditional random field with road block constraints to regroup the objects into the fine-grained UFZ map [9]. In general, these methods are multi-stage. The lack of the end-to-end mechanism leads to the accumulation of errors from different stages and increases the complexity of UFZ mapping. Therefore, it is necessary to develop an end-to-end method to interpret UFZs with multimodal geographic data.

In summary, existing studies have three limitations. First, making no full use of multimodal geographic data, i.e., insufficient expression of image features and the lack of socioeconomic semantic modeling of social sensing data, results in failing to interpret UFZs accurately and reliably. Second, the feature fusion of multimodal geographic data is still at the shallow level rather than the semantic level, which limits the semantic representation of UFZs. Third, existing UFZ mapping methods are multi-stage, e.g., CNN + post-processing. There is a lack of methods fusing multimodal geographic data to map UFZs end-to-end. To solve these issues, this study designs a multimodal data fusion framework (MDFF) to map fine-grained UFZs end-to-end, which organically integrates multimodal geographic data, i.e., RS images and social sensing data. The MDFF learns the deep features from images at the pixel scale and models socioeconomic semantics of UFZs with social sensing data at the object scale. With the designed deep feature fusion module, the MDFF fuses multimodal features at both pixel- and object scales to take advantage of the two scales in expressing detailed features and object semantics for UFZs. The training or operation of the MDFF is completed in one step. Three contributions have been made in this study:

1) A method (MDFF) fusing multimodal geographic data is proposed for mapping fine-grained UFZs end-to-end. It automatically learns physical attributes from RS images and models socioeconomic semantics from social sensing data;

2) The MDFF fuses multi-modal information both at pixel- and object scale with two fusion modules (i.e., the IF2GF and the GF2IF in Sec.2.2) for learning urban semantics where object semantics guide the fine-grained UFZ mapping;

3) Experiments conducting on two typical modal data, i.e., RS images and social sensing data, show that the MDFF can significantly improve the quality of UFZ mapping and reduce the complexity of UFZ mapping.

II. Methodology

A. Overview

Fig. 2 illustrates the workflow of the proposed MDFF.

1) Multimodal feature extraction. In the MDFF, the pixel-based module extracts deep features from RS images, which provides detailed information for fine-grained mapping, while the object-based module generates characteristics of objects from social sensing data, which models semantics relationships between objects within UFZs.

2) Multimodal feature fusion. With the proposed feature fusion modules, i.e., the Image Feature to Graph Feature (IF2GF) and the Graph Feature to Image Feature (GF2IF), the pixel-wise deep features and the object-wise semantic characteristics are fused at pixel and object scales. Integrating multimodal information under the "pixel-object-UFZ" semantic hierarchy is critical for mapping fine-grained UFZs and improving the reliability of the interpretation.

3) UFZ mapping. The MDFF is trained end-to-end with pixel samples and object samples, where the pixel samples are collected manually in the form of polygons, and the object samples are generated with the first category of POI data and the pixel samples. After the training, the MDFF is exploited to map UFZs with multimodal geographic data in the study area.

B. Multimodal feature extraction

In this big data era, image and social sensing data are important components of multimodal geographic data. They bring rich information for eliminating the uncertainty of the interpretation. The former provides pixel-level characteristics, while the latter contains socioeconomic

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3326160

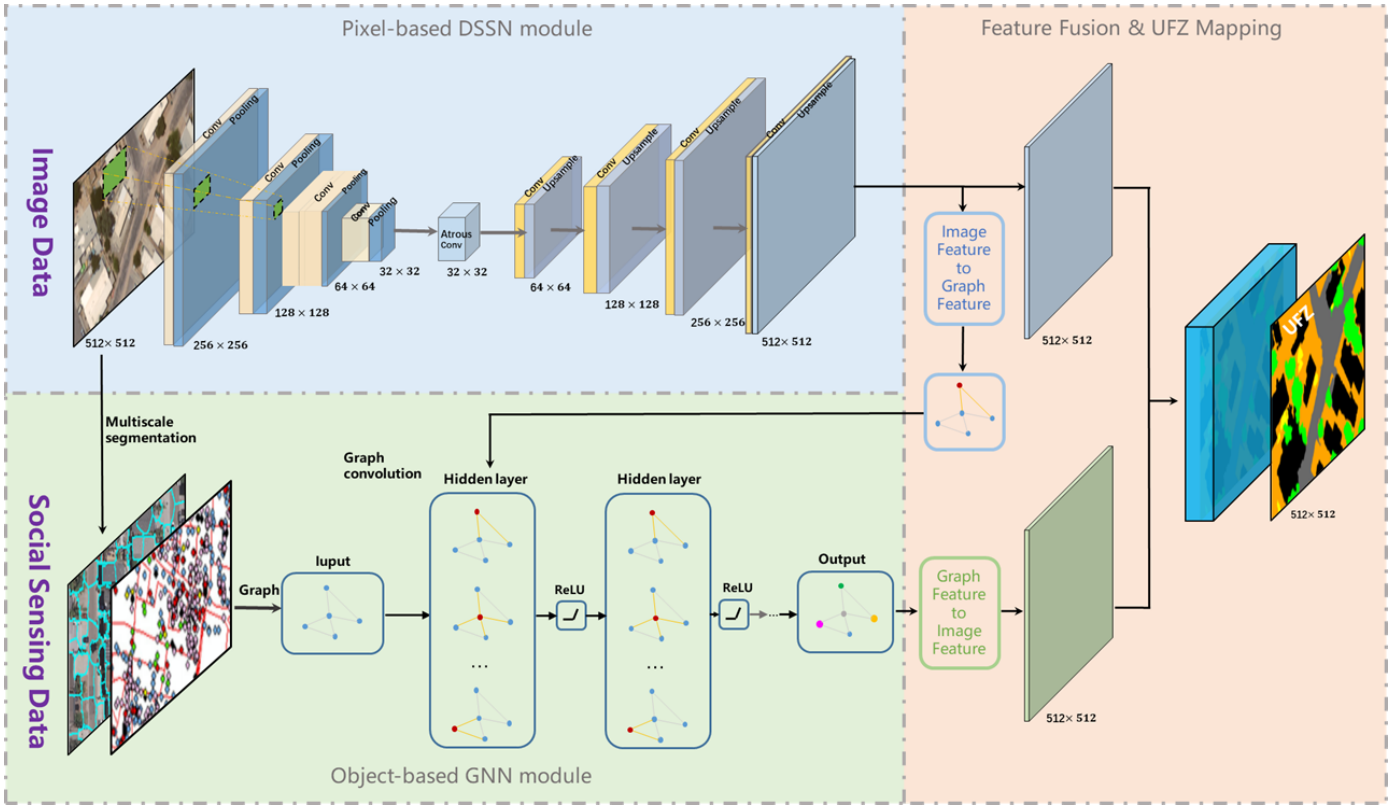


Fig. 2. The workflow of the proposed MDFF

properties of and mutual relationships between objects. Images and social sensing data need to be processed separately to obtain representative features because of their different characteristics. The MDFF is proposed to take full advantage of RS images and social sensing data. Therefore, the pixel-based module and the object-based module are designed in the MDFF to extract multimodal features. Each UFZ is composed of objects and pixels constitutes each object. Pixel-scale features and object-scale features are exploited to characterize UFZs in physical and socioeconomic terms.

1) The pixel-based module

The pixel-based module is designed to learn image features, which consists of an encoder and a decoder. The encoder extracts deep features from RS images with the stacked layers. Each layer is composed of the convolution and the pooling operations. The convolution scans the image step by step and generates feature representations of a local region in each step. The pooling expands the receptive field of the network to reduce information redundancy and obtain global information. The deep features with abstract expression are generated through the stacked layers. f_i^e denotes the *i*-th layer feature from the encoder function φ_i^e which is composed of the convolution $Conv(\cdot)$ and the pooling $Pool(\cdot)$. w_i^e and b_i^e, which are learnable parameters, represent weights and bias of the convolution kernel of φ_i^e.

$$f_{i+1}^e = \varphi_i^e(w_i^e \cdot f_i^e + b_i^e) \quad (1)$$

$$\varphi_i^e = Pool(Conv(\cdot)) \quad (2)$$

The encoder reduces the size of the features with the convolution and the pooling operations. The features are too small to hold fine-grained information, which hardly meets the requirements of intensive prediction tasks (e.g. UFZ mapping). Therefore, the features are enlarged to the original size with the upsampling of the decoder. The decoder function φ_j^d generates the *j*-th layer feature f_j^d. φ_j^d consists of the convolution $Conv(\cdot)$ and the upsampling $Ups(\cdot)$ with the learnable parameters w_j^d and b_j^d.

$$f_{j+1}^d = \varphi_j^d(w_j^d \cdot f_j^d + b_j^d) \quad (3)$$

$$\varphi_j^d = Conv(Ups(\cdot)) \quad (4)$$

2) The object-based module

UFZs are aggregated from objects and the compositional patterns of objects affect the categories of UFZs [40]. To construct the graph, image objects are regarded as the graph nodes and the spatial relationships between objects as the graph edges. In this way, the object compositional patterns of UFZs are modeled in the form of the graph. The objects are characterized with object features ($\boldsymbol{f}^o = [\boldsymbol{f}^{sp}, \boldsymbol{f}^t, \boldsymbol{f}^g, \boldsymbol{f}^s]$) in TABLE I. Spectral features ($\boldsymbol{f}^{sp}$), textural features ($\boldsymbol{f}^t$), and geometrical features ($\boldsymbol{f}^g$) have been widely used to measure objects from different aspects in most object-based image analyses [25]. Features from social sensing data present the composition of socioeconomic attributes of UFZs, as social sensing data are closely related to socioeconomic activities. The vector $\boldsymbol{f}^s = [\boldsymbol{f}^s_{poi}, \boldsymbol{f}^s_{sm}, \boldsymbol{f}^s_{ms}]$ is used to express primary semantics of each object, where $\boldsymbol{f}^s_{poi}$, $\boldsymbol{f}^s_{sm}$, and $\boldsymbol{f}^s_{ms}$ denote features from

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3326160

POIs, social media data, and mobile signal data, respectively. POIs and social media data are generated by socioeconomic activities, which contain socioeconomic semantics. Mobile signal data reflects the flow of urban population and is closely related to socioeconomic activities. Therefore, these three kinds of social sensing data are selected to extract socioeconomic semantics for UFZ mapping. $\boldsymbol{f}_{poi}^{s} = [f_{density}^{s}, f_1^s, f_2^s, \dots, f_c^s]$, where c denotes the second category of POIs and f_i^s represents the proportion of *i*-th category within an object. Similarly, $\boldsymbol{f}_{sm}^{s}$ can be obtained from social media data. The average population mobility of objects constitutes $\boldsymbol{f}_{ms}^{s}$.

We first segment RS images into image objects using multiresolution segmentation method [38]. Then, object features f_i^o in TABLE I and deep features f_j^d are adopted to characterize objects. The feature vector of each node is $\boldsymbol{x}_n = [f_1^o, f_2^o, \dots, f_m^o, f_j^d]$ and the graph feature is denoted as $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N]^T \in \mathbb{R}^{K \times D}$, where N is the number of the nodes and D is the dimension of $\boldsymbol{x}_k$. Finally, the first-order adjacency relationships (with common edge) between objects are adopted as graph edges to take the topological spatial relationship into consideration. The adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ quantifies the graph edges where the strength of each edge is determined by the spectral similarity and the spatial location of objects. The connecting edge between node v_i and node v_j has the weight $a_{ij} \in \boldsymbol{A}$. if v_i is adjacent to v_j, $a_{ij} = exp\left(-\frac{\|lab_i - lab_j\|}{\sigma_w^2}\right)$, otherwise, $a_{ij} = 0$. lab denotes the average color value of the node in the CIE LAB color space and σ_w^2 controls the range of the weight.

TABLE I
OBJECT FEATURES FROM IMAGES AND POI

Types	Names	Meanings
Socioeconomic	POI Semantics	The proportion of each semantic category of POIs
		The density of POIs in an object
	Population Mobility	Within an object, the population mobility average over time
	Semantics of Social Media Data	The proportion of each semantic category of social media data within an object
Spectral	Mean	Average spectrum of pixels
	Std. Dev	Gray standard deviation of pixels in an object
	Skewness	Skewness of spectral histogram
	Border Contrast	The average differences between border pixels and their neighborhood
Textural	GLCM Homogeneity	The homogeneity derived from GLCM
	GLCM Dissimilarity	The heterogeneity parameters derived from GLCM
	GLCM Entropy	Information entropy derived from GLCM
	GLCM Correlation	Correlation of pixels which is derived from GLCM
	GLDV	The vector composed of diagonal elements of GLCM
	GLDV Mean	Average value of GLCM
	GLDV Entropy	Information entropy derived from GLDV
	GLDV Contrast	The Contrast of GLDV
	GLDV Ang. 2nd moment	The Ang. 2nd moment of GLDV
Geometrical	Area	The number of pixels within image objects
	Length/Width	Length-width ratio of the object's MBR
	Eclipse Fit	The fitting degree of eclipse fit
	Main Direction	Eigenvectors of covariance matrix
	Shape Index	The ratio of perimeter to four times side length

Graph convolution has powerful ability in modeling semantic relationships of discrete objects. Thus, the object-based module is designed with the graph convolution to extract socioeconomic semantics from social sensing data and its structure is shown in TABLE II. This module is a hierarchical network with L layers ($L = 3$). Each layer consists of a graph convolution function and a nonlinear activation function.

TABLE II
THE LAYER STRUCTURE OF THE OBJECT-BASED MODULE. N REPRESENTS THE NUMBER OF OBJECTS IN THE GRAPH

Layers	Components	Input Size	Output Size
Input	input	$N \times 100$	-
Layer-1	graph convolution ReLU	$N \times 100$	$N \times 18$
Layer-2	graph convolution ReLU	$N \times 18$	$N \times 18$
Layer-3	graph convolution ReLU	$N \times 18$	$N \times 9$
Output	output	-	$N \times 9$

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3326160

The object-based module implements the graph convolution [41] to extract objects' features $\boldsymbol{X}$. $\boldsymbol{X}^{(l)}$ is the feature from the l-th layer of the graphic neural network (GNN). After the graph convolution, new features $\boldsymbol{X}^{(l+1)}$ are activated by the nonlinear activation function σ. The core idea of the graph convolution is that graph nodes with edge connections are aggregated for generating new representations of the nodes where semantic information flows between the nodes with edge connections, so the object-based module can model the semantic relationships between image objects and build the spatial object patterns of UFZs. On the other hand, graph nodes have the socioeconomic features from POI data, thus the object-based module can learn both the socioeconomic attributes and relationships within $\boldsymbol{X}^{(l+1)}$ to restore the real state of UFZs.

$$\boldsymbol{X}^{(l+1)} = \sigma\left(\boldsymbol{L}_G \boldsymbol{X}^{(l)} \boldsymbol{W}^{(l)}\right) \tag{5}$$

$$\widetilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I} \tag{6}$$

$$\boldsymbol{L}_G = \widetilde{\boldsymbol{D}}^{-\frac{1}{2}} \widetilde{\boldsymbol{A}} \widetilde{\boldsymbol{D}}^{-\frac{1}{2}} \tag{7}$$

where $\boldsymbol{W}$ and $\boldsymbol{L}_G$ are the Laplacian matrix and the learnable parameters, respectively. $\boldsymbol{I}$ is the identity matrix and $\widetilde{\boldsymbol{D}}$ is the degree matrix of $\widetilde{\boldsymbol{A}}$.

C. Multimodal feature fusion

Features from images are pixel-regular and vision-related, while graph features from social sensing data are discrete and semantic-related, thus different modalities of features make it difficult to fuse these two kinds of features. One solution is to transform the form of the two kinds of features. Therefore, the Image Feature to Graph Feature (IF2GF) and the Graph Feature to Image Feature (GF2IF) are designed to conduct this task in the MDFF (Fig.3).

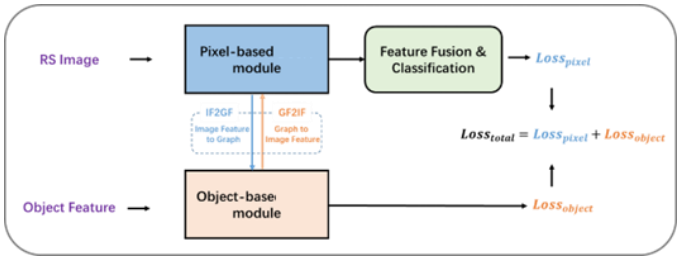


Fig. 3. The multimodal feature fusion and network training of the proposed MDFF.

Due to the gradient vanishing [42], GNNs are usually limited to shallow layers, which restricts GNN's performance in feature learning. Therefore, the deep features from the pixel-based module's decoder are transformed by the IF2GF and then added to the GNN-base module (Fig.3), which helps GNN focus on modeling semantic relationships. In the IF2GF, deep features $\boldsymbol{f}_1^d$ are transformed to graph features $\boldsymbol{f}_1^g$ with the transforming function $\boldsymbol{T_1}$. Within each object (the addressing function $\boldsymbol{A}$), the average value on each channel of the deep features is taken as the graph feature (the assignment function $\boldsymbol{S}$). In the object-based module, the graph features and the attribute features presented in TABLE II are concatenated together at object scale, and are used as the features of graph nodes.

$$\boldsymbol{f}_1^g = \boldsymbol{T_1}(\boldsymbol{f}_1^d) \tag{8}$$

$$\boldsymbol{T_1} = \boldsymbol{S}(\boldsymbol{A}(\boldsymbol{Mean}(\cdot))) \tag{9}$$

The object-based module outputs graph features $\boldsymbol{f}_2^g$. In the GF2IF, the graph features are transformed to image features $\boldsymbol{f}_2^d$ with the transforming function $\boldsymbol{T_2}$, where the pixels in each object are assigned the same features from the graph feature of the object. Then, the GF2IF aligns the graph features with the deep features pixel to pixel. To adjust value distribution of features, the transformed features are input into a module $\boldsymbol{Con}$ with a convolution layer (3×3 kernel, 1 stride, and 1 padding), a BatchNorm layer and a ReLU activation function. Then, both the transformed features and the deep features are concatenated together at the pixel scale, where the graph features provide additional semantics of socioeconomic relationships for the UFZ mapping.

$$\boldsymbol{f}_2^d = \boldsymbol{T_2}(\boldsymbol{f}_2^g) \tag{10}$$

$$\boldsymbol{T_2} = \boldsymbol{Con}(\boldsymbol{S}(\boldsymbol{A}(\cdot))) \tag{11}$$

The deep features provide detailed features in pixel space, while the graph features contain object semantics and relationships. With the feature fusion modules, i.e., the IF2GF and the GF2IF, the deep feature and the graph feature are fused at both the object and the pixel scales successively, which further help to improve features' expressiveness in depth or semantics and generate richer and more comprehensive features for UFZs. Therefore, multimodal information is integrated for better details and semantics to interpret UFZs.

D. UFZ mapping

1) Study area and data collection

The two largest and most famous cities in China, Beijing and Shanghai, are chosen in this study. Beijing (Fig.4) is the capital of China and Shanghai is the economic center of China. Beijing and Shanghai are rapidly developing at a high urbanization level and contains diverse archaic and modern urban zones. The evolution of UFZs is crucial for revealing urban sustainable development. UFZ maps are produced in the study area (within the fifth ring road of Beijing and the outer ring road of Shanghai, main urban area of Beijing and Shanghai) with the proposed method. For image data, the level-3 VHR satellite image (Fig.4) from ArcGIS World Imagery in 2019 is used as the experimental area, which cover 667 km^2 of Beijing and 664 km^2 of Shanghai with the spatial resolution of 2.4m and the RGB band. As shown in TABLE III, the class system proposed in the paper [9] is adopted to map UFZs in this study. As there are no woodlands and farmlands in the main urban area of Beijing and Shanghai, the two categories are removed. The Residential-1 and the Residential-2 zones are merged into one class, Residential area, because of their similar functions.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3326160

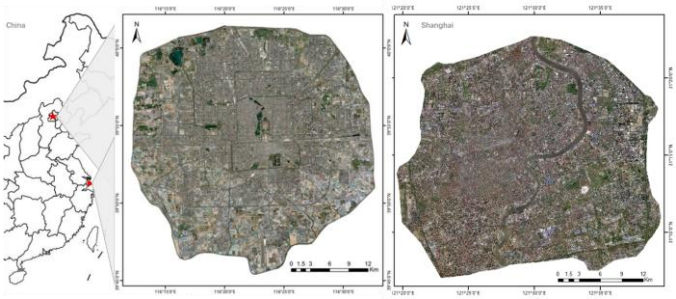


Fig. 4. The study areas in Beijing and Shanghai.

TABLE III
THE CATEGORY SYSTEM

Du's category system	Ours category system
Commercial	Commercial
Residential-1 Residential-2	Residential
Residential-3	Shantytown
Institutional	Institutional
Industrial	Industrial
Transport	Transport
Urban green	Urban green
Undeveloped	Undeveloped
Woodland	-
Farmland	-
Water	Water

Social sensing data is common in cities and easily accessible. It reflects socioeconomic attributes which are closely related to UFZs. POIs, mobile signal data, and social media data are typical social sensing data. Take POIs as an example, it is discrete and irregularly arranged in the geospatial space (Fig.5 and Fig.6).

POIs: about 267,000 POIs of Beijing and 356,000 POIs of Shanghai are adopted in this study. The categories of POIs are divided into two levels: the first category and the second category (TABLE IV). The POIs are labeled with six categories, i.e., commercial points (55.8% of Beijing), residential points (20.6%), shantytown points (3.1%), institutional points (11.4%), industrial points (8.6%), and urban green points (0.5%). There is a large class imbalance within the POI categories because of the POI's correlation with socioeconomic activities.

Mobile signal data: the data records the spatial locations of the mobile phone users when exchanging information with the base station. It presents the spatial and temporal flow of people's movement on a large scale. The base stations are distributed in the grid of 500m*300m. Within a base station, local population, working population, residential population, flow population, and transit population are recorded, which are closely related to socioeconomic activities. This study adopts mobile signal data in Beijing and Shanghai from January to February 2020 (33 days in total).

Social media data: the data is obtained from the web of Dianping, showing commercial activities of merchants which are divided into 15 categories (TABLE IV). About 268,000 records of Beijing and 688,000 records of Shanghai are used to extract socioeconomic attributes of UFZs.

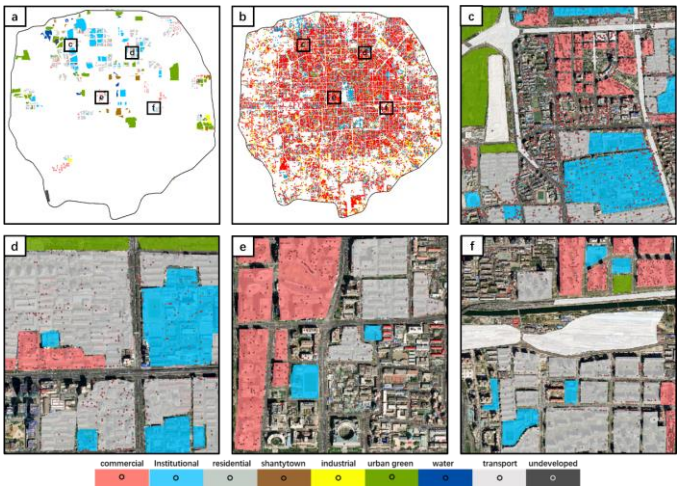


Fig. 5. UFZ samples (a) and POIs (b) of Beijing. (c) ~ (f): detailed illustrations of four sub-regions.

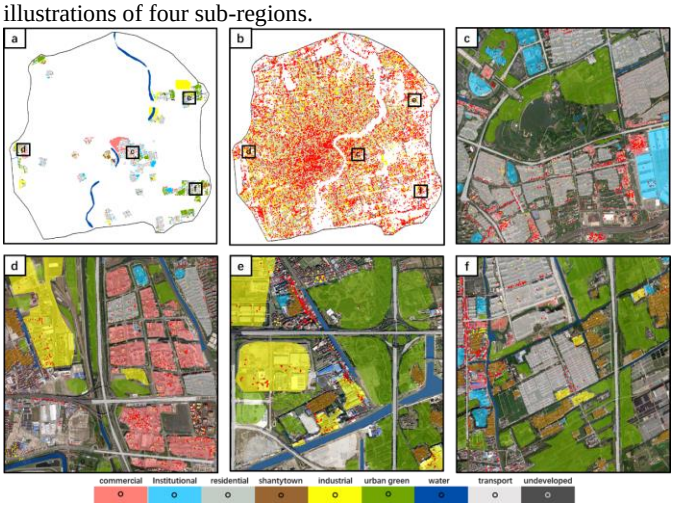


Fig. 6. UFZ samples (a) and POIs (b) of Shanghai. (c) ~ (f): detailed illustrations of four sub-regions.

The MDFF is trained end-to-end with both pixel and object samples. In the MDFF, pixel samples guide the pixel-based module to extract deep vision features and the object samples help the object-based module to learn semantic features for reliable UFZ mapping. Pixel samples of UFZs are manually selected in polygonal form, and polygons are labeled according to the new class system (Fig.5 and Fig.6). These samples are randomly divided into three parts with the ratio of 8:1:1. Then, 80% samples are exploited for training the MDFF, 10% for evaluation and 10% for test. By the proposed method in Sec.2.3.2, the object samples are automatically generated with the first category of POI data and the pixel samples.

TABLE IV
THE CATEGORIES OF POIS AND SOCIAL MEDIA DATA

POI		**Social media data**
First category	Second category	
Commercial	Car shop and parking lot	Gourmet shop
Residential	Company	Shopping store
Shantytown	Restaurant	Leisure and entertainment
Institutional	Government agency	Scenic spot
Industrial	Hospital	Sports venue

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3326160

Urban green	Public place	Barber and beauty shop
-	Hotel and residential	Wedding spot
-	Shop	Hotel
-	Sport center	Car shop
-	Scenic spot	Parent-child spot
-	-	Business education institution
-	-	Life service
-	-	Medical health
-	-	Home decoration
-	-	Pet care store

2) Generating object samples

Categories of objects are required for supervising the MDFF to learn graph features. Objects (graph nodes) can be described by its main attributes. Therefore, we count POIs in each object and assign the most frequent class of POIs as the category of the object. However, POIs are distributed unevenly in space as shown in Fig.5 and Fig.6. Categories of objects without POIs are unknown. In addition, categories of objects may be incorrect as there is a large class imbalance in POI data. Incorrect samples will seriously affect the learning process of deep networks. Thus, object samples must be optimized. Pixel samples are the true label of the real world and thus can be used to optimize the object samples. If a graph object intersects with a sample polygon by more than 50%, its category will be changed to the category of the polygon; otherwise, its category will be maintained.

As described above, object samples need to be optimized by pixel samples. As the example shown in Fig.7, the white circle areas in (b) are corrected in (c) with true labels of (a) and object samples become better in (c), which means the object samples have been optimized. The better samples will facilitate the network to learn more discriminative feature representations.

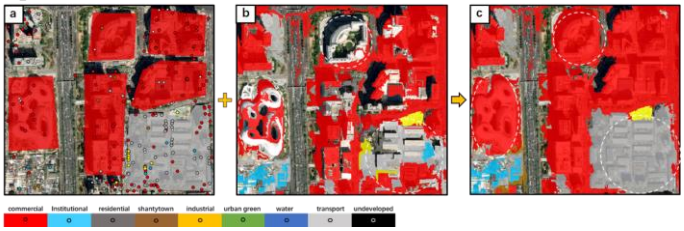


Fig. 7. Examples of optimizing object samples with pixel samples. (a) the pixel samples and POIs; (b) the object samples generated from POIs where the areas of each object is shown with its class; and (c) is object samples optimized from (b) with the image polygon sample of (a).

3) Generating UFZ maps end-to-end

UFZs are the results of urbanization and are closely related to human activities, economic behaviors, and geographical factors. Meanwhile, land covers constitute UFZs in RS images. As a result, physical features and semantic cues are critical for mapping UFZs. The multimodal features from images and social sensing data are more helpful for the interpretation. In the MDFF, the fused multimodal features from Section II.C are input into the last convolution layer (3 × 3 kernel, 1 stride, and 1 padding) to output the confidence map. The UFZ map is generated from the confidence map according to the maximum confidence. $Loss_{pixel}$ and $Loss_{object}$ are calculated by the cross-entropy loss function with pixel samples and object samples, respectively. The MDFF is trained end-to-end with the total loss $Loss_{total}$ which is composed of the pixel- and the object-based losses (Fig.3).

$$Loss_{total} = Loss_{pixel} + Loss_{object} \quad (12)$$

$$Loss_{pixel} = -\sum_{i=1}^{w}\sum_{j=1}^{h}\sum_{c=1}^{n} y_{ij}^{c} log\ (p_{ij}^{c}) \quad (13)$$

$$Loss_{object} = -\sum_{k=1}^{k}\sum_{c=1}^{n} y_{k}^{c} log\ (p_{k}^{c}) \quad (14)$$

The feature map has width of w and height of h. There are c categories that the pixel (i, j) or the object k belongs to. prediction from the forward propagation of network is $Y_{ij} \in \boldsymbol{Y}$, if $Y_{ij} = c$, then $y_{ij}^{c} = 1$, otherwise, $y_{ij}^{c} = 0$.

III. Experiments

A. Implementation details and evaluation metrics

As the nodes in the graph, image objects are segmented from RS images using the multiresolution segmentation method [38] with the multiple scale of 60. The multiresolution segmentation algorithm is a commonly used image processing technique that can obtain more comprehensive and accurate segmentation results by processing images at different scales. The basic principle is achieved through steps of scale transformation, segmentation algorithms, and scale fusion. For multimodal feature representation, the advanced deep network DeepLab v3+ [43] is used as the baseline and its backbone is adopted as the backbone of the pixel-based module to extract deep features from images, and the object-based module with 3 layers ($L = 3$) is designed for social sensing data modeling. The RS image and training polygon labels are clipped into samples with size of 512 × 512 in the step of 256 pixels, where each sample has a clipped image and the corresponding label. After preparation of data, the proposed MDFF is trained with the learning rate, the batch size, and the epoch being 0.0001, 4 and 150, respectively. The stochastic gradient descent method (SGD) and the cross entropy are adopted as the optimizer and the loss function respectively.

In this paper, the overall accuracy (OA), the intersection over union (IoU), the mean intersection over union ($MIoU$) and the frequency weighted intersection over union ($FWIoU$) are adopted as the evaluation metrics. IoU, $MIoU$ and $FWIoU$ are classic metrics, which are often used to validate intensive predictions such as semantic segmentation.

$$OA = (TP + TN)/(TP + FP + TN + FN) \quad (15)$$

$$IoU_i = \frac{TP_i}{TP_i + FP_i + FN_i}, i = 1,2, \dots, n \quad (16)$$

$$MIoU = \frac{1}{n}\sum_{1}^{n} IoU_i\ , i = 1,2, \dots, n \quad (17)$$

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3326160

$$FWIoU = \sum_{1}^{n}(IoU_i \cdot \frac{TP_i + FN_i}{TP_i + FP_i + TN_i + FN_i}) \quad (18)$$

where TP, TN, FP, and FN refer to the number of true positive points, true negative points, false positive points, and false negative points respectively. n is the number of classes.

B. Classification accuracy of UFZ

The comparative experiments are set up to verify the effectiveness of the proposed MDFF. As the MDFF makes full use of UFZs' physical and socioeconomic information from multimodal geographic data, thus the MDFF is significantly better than the baseline (TABLE V, TABLE VI, TABLE VII and TABLE VIII), which demonstrates the advance of the proposed method. The MDFF has the reliable performance in classifying complex UFZs, e.g., commercial, institutional, and residential zones, as there are many POIs in these UFZs. It illustrates that the representation and the fusion of multimodal information improve the discriminative ability of the network. From the results of the third and the fourth rows, the MDFF performs better as it has learned the better features with the optimized labels. In addition, when trained with the Graph Loss, the MDFF achieves the better accuracy, which shows that the Graph Loss assists the MDFF to make full use of social sensing data. Therefore, the accurate label and the loss constraint guides the learning process of the graph model. With the Graph Label-OP and the Graph Loss, the MDFF obtains the best accuracy in OA, MIoU and FWIoU.

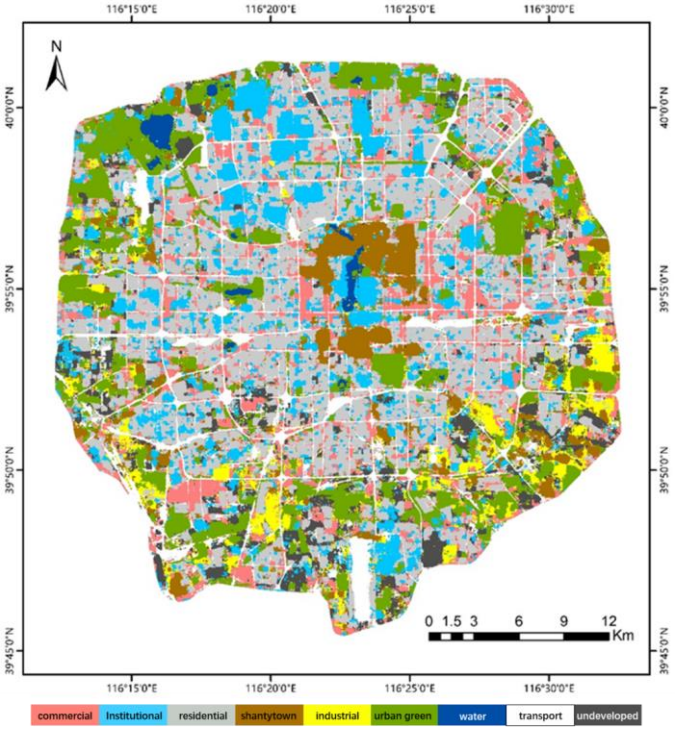


Fig. 8. The UFZ map of Beijing.

C. UFZ mapping results

Fig.8 and Fig.9 shows the final UFZ maps of Beijing and Shanghai, respectively, which is produced by the proposed MDFF with the Graph Label-OP and the Graph Loss. There are different UFZ categories: commercial, institutional, urban green, industrial, residential, shantytown, undeveloped, and transport.

In Beijing, it is shown in the UFZ map those large areas of commercial zones are in the city center. Institutional zones are mainly clustered in the center and northwest of the city for the distribution of government departments and campuses. Industrial zones and urban-green zones are located on the urban outskirts, which reflects the distribution of highly urbanized areas. In addition, shantytown zones cluster in the urban center because of the historical legacy (Hutong). Hutong is the unique building of Beijing.

Shanghai is the economic center of China, with a highly developed economy. Commercial zones are mainly clustered in the urban center in Shanghai. In addition, Shanghai's industry is highly developed, especially the automobile industry and the semiconductor industry. It is shown in Fig.9 that there are many industrial zones in Shanghai. Same with Beijing, shantytown zones in Shanghai cluster in the urban center for the historical legacy.

To evaluate the fine-grained UFZ mapping of the MDFF, the detailed UFZ maps of eight areas in Beijing and Shanghai are shown in Fig.10 (a)~(h) and Fig.11 (a)~(h), respectively. These maps overlap on the original RS images with 40% transparency. The proposed MDFF interprets multimodal geographic data accurately and achieves good performance in classify UFZs, especially for commercial, residential, institutional, industrial, shantytown, urban green, water. The classification results match well the real status of UFZs.

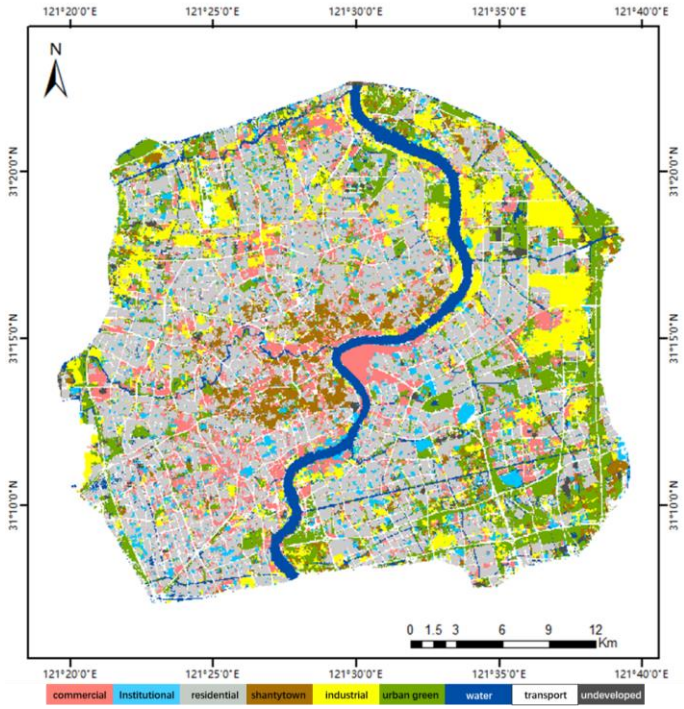


Fig. 9. The UFZ map of Shanghai.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3326160

TABLE V

THE OA (%) OF THE UFZ MAPPING OF BEIJING. THE GRAPH LABEL-OP MEANS THAT THE OBJECT SAMPLES HAVE BEEN OPTIMIZED WITH THE OBJECT. THE GRAPH LOSS REPRESENTS THAT THE GRAPH LOSS IS USED AS A PART OF THE TOTAL LOSS. (COM: COMMERCIAL, INSTI: INSTITUTIONAL, URGR: URBAN GREEN, INDU: INDUSTRIAL, RESI: RESIDENTIAL, SHTO: SHANTYTOWN, UNDE: UNDEVELOPED, TRAN: TRANSPORT.)

Model	Graph Label-OP	Graph Loss	Com.	Insti.	Urgr.	Indu.	Resi.	Shto.	Water	Unde.	Tran.	Overall (OA)
Baseline	×	×	87.63	92.93	98.37	95.39	96.05	98.56	99.90	83.27	95.87	95.33
MDFF			91.37	95.57	98.78	97.93	97.88	98.37	99.79	85.55	97.23	96.95
MDFF		✓	91.77	95.25	99.15	96.80	98.09	**99.19**	**99.98**	**87.81**	**97.76**	97.18
MDFF	✓	✓	**95.65**	**96.77**	**99.23**	**99.34**	**98.22**	98.05	99.91	87.47	97.50	**97.79**

TABLE VI

THE OA (%) OF THE UFZ MAPPING OF SHANGHAI. THE GRAPH LABEL-OP MEANS THAT THE OBJECT SAMPLES HAVE BEEN OPTIMIZED WITH THE OBJECT. THE GRAPH LOSS REPRESENTS THAT THE GRAPH LOSS IS USED AS A PART OF THE TOTAL LOSS. (COM: COMMERCIAL, INSTI: INSTITUTIONAL, URGR: URBAN GREEN, INDU: INDUSTRIAL, RESI: RESIDENTIAL, SHTO: SHANTYTOWN, UNDE: UNDEVELOPED, TRAN: TRANSPORT.)

Model	Graph Label-OP	Graph Loss	Com.	Insti.	Urgr.	Indu.	Resi.	Shto.	Water	Unde.	Tran.	Overall (OA)
Baseline	×	×	88.19	80.43	93.63	87.27	95.69	89.95	91.94	80.93	92.64	92.01
MDFF			88.61	82.86	95.02	92.38	96.37	**96.27**	91.81	79.23	93.85	93.51
MDFF		✓	89.88	82.27	94.66	**92.43**	96.78	96.20	**92.64**	81.65	93.43	93.69
MDFF	✓	✓	**91.94**	**87.04**	**95.30**	92.02	**97.29**	93.41	92.26	**86.59**	**94.36**	**94.45**

TABLE VII

THE IOU (%) OF THE UFZ MAPPING OF BEIJING

Model	Graph Label-OP	Graph Loss	Com.	Insti.	Urgr.	Indu.	Resi.	Shto.	Water	Unde.	Tran.	MIoU	FWIoU
Baseline	×	×	79.56	88.20	97.13	82.75	90.57	95.36	99.68	76.78	94.36	89.38	91.22
MDFF			87.05	92.10	97.98	88.29	93.66	96.59	99.75	81.21	96.56	92.58	94.14
MDFF		✓	87.70	92.66	98.23	**90.12**	94.14	96.43	**99.89**	**83.16**	96.77	93.24	94.57
MDFF	✓	✓	**91.56**	**9510**	**98.61**	87.19	**95.59**	**96.65**	99.68	82.84	**96.86**	**93.79**	**95.73**

TABLE VIII

THE IOU (%) OF THE UFZ MAPPING OF SHANGHAI

Model	Graph Label-OP	Graph Loss	Com.	Insti.	Urgr.	Indu.	Resi.	Shto.	Water	Unde.	Tran.	MIoU	FWIoU
Baseline	×	×	80.69	70.83	89.50	82.80	89.35	76.57	87.92	65.70	84.91	80.92	85.34
MDFF			83.59	74.16	90.66	89.86	91.02	78.31	89.41	70.07	88.34	83.94	87.94
MDFF		✓	83.61	75.53	91.27	89.44	91.05	81.46	**89.61**	67.74	**88.51**	84.25	88.23
MDFF	✓	✓	**86.74**	**82.03**	**91.47**	**89.91**	**92.85**	**84.87**	88.16	**69.60**	88.29	**85.99**	**89.56**

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3326160

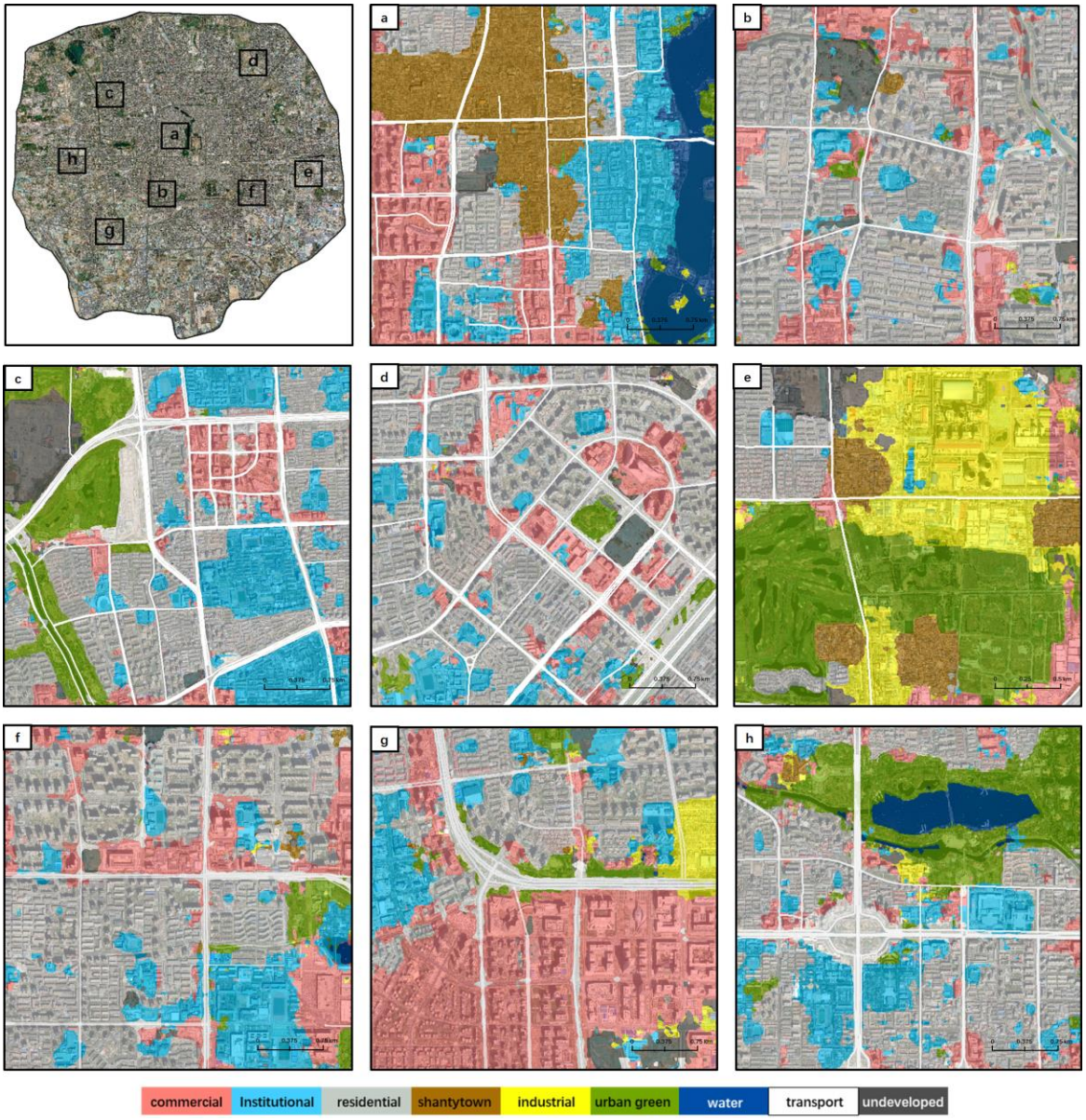


Fig. 10. The detailed UFZ maps of eight areas in Beijing.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3326160

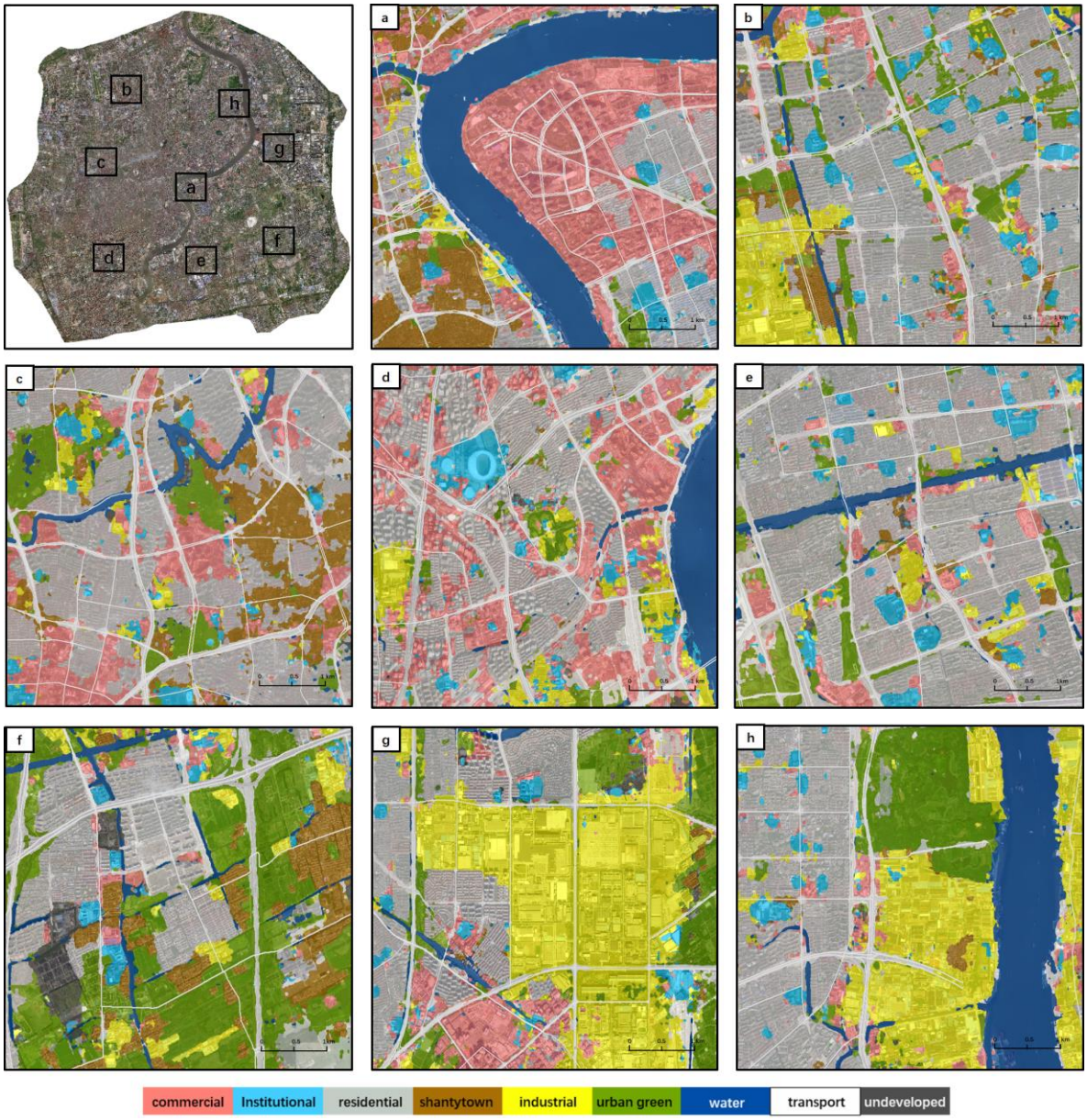


Fig. 11. The detailed UFZ maps of eight areas in Shanghai.

D. Urban functional structure analysis

The urban functional structures reflect the development status of cities [44]. Proportions of diverse UFZs in the main urban area of Beijing and Shanghai are calculated from the UFZ results of the proposed MDFF. As show in Fig.12, residential districts of the two city have the largest proportion about 30%. This is a normal phenomenon in big cities, because the population brought by urbanization requires many dwellings to carry. As the capital of China, Beijing is the political and cultural center of China, where there are many government agencies and research institutes (12% institutional zones). As the economic center of China, Shanghai has the thriving economy reflected by the commercial zones with 11%. The proportion of transport shows that ground transportation of the two city is well developed. However, urban greening rate (15% of Beijing and 13% of Shanghai) needs to be improved. In addition, there is 7% undeveloped area left in Beijing for further construction. Industrial zones are mainly distributed in the suburbs, so there are few industrial zones in the urban area. Due to the developed automobile and semiconductor industries in Shanghai, Shanghai has more industrial zones than Beijing in downtown. As a legacy of history, shanty towns still account for 5% in Beijing and Shanghai.

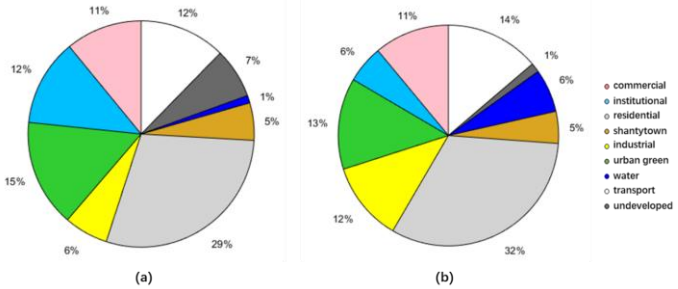


Fig. 12. Proportions of diverse UFZs. (a) for Beijing and (b) for Shanghai.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3326160

IV. Discussions

A. Comparing with existing methods

Some representative studies for mapping UFZs are listed in TABLE IX, such as traditional methods based on image tiles and blocks [8][44][45], hierarchical semantic cognition models [25][39], and CNN-based approaches [9][46]. Although the adopted evaluation methods are different, these studies are more likely consistent in the study area for comparison. Traditional methods are often limited because of insufficient feature representation. The CNN-based methods have obvious advantages in UFZ mapping. Du et al. (2021) designed a CNN for feature extraction and adopted the conditional random field to regroup objects for the large-scale and fine-grained UFZ maps [9]. These methods achieve good performance but still have limitations: 1) multimodal geographic data are not fused to eliminate uncertainty of the interpretation, and 2) the two-stage methods (CNN + post-processing) are not the end-to-end mechanism, which increases the complexity of interpretation and can lead to accumulation of classification errors. The proposed MDFF overcomes the above limitations and improves the accuracy (about 5% in OA) and reliability of UFZ mapping.

TABLE IX

Comparisons with Existing Studies. (MRI: Medium-Resolution Images; SHRI: Submeter-Level High-Resolution Images; MHRI: Meter-Level High-Resolution Images; MPPD: Mobile Phone Positioning Data; POI: Point Of Interest)

Model	Method	Data	Study Area	Mapping unit	Reported Accuracy
Hu et al. (2016)	Traditional method	MRI POI	Beijing (16,410 km^2)	road blocks	81.0%
Simwanda and Murayama (2017)	Traditional method	MHRI	Lusaka (420 km^2)	road blocks	85.0%
Zhang et al. (2017)	HSC	MHRI POI	Beijing (67.1 km^2)	road blocks	90.8%
Huang et al. (2018)	Traditional method	MHRI	Hong Kong (170.8 km^2)	road blocks	91.3%
Du et al. (2021)	CNN+CRF	MHRI	Beijing (2267 km^2), Shanghai (2967 km^2)	Objects and road blocks	91.6% (Beijing) 89.1% (Shanghai)
Rosier et al. (2022)	CNN	MHRI POI Road network	Netherlands	Pixels	85.0%
The proposed MDFF	CNN	MHRI POI Mobile signal data Social media data	Beijing (667 km^2), Shanghai (664 km^2)	Pixels and objects	97.8% (Beijing) 94.5% (Shanghai)

B. Advantages of fusing multimodal geographic data

Data-driven learning strategies are often unstable, which makes deep networks susceptible to noise, especially for RS images. The complex spectrum and spatial structure greatly increase the uncertainty of classification. Multimodal geographic data contains information (e.g., physical, and socioeconomic information) obtained from different perspectives, which can give the classifier more discriminative basis to enhance the reliability of the classification. With all the multimodal geographic data shown in TABLE X and TABLE XI, the best classification accuracy is achieved (about 3% higher than that of single modality with images, TABLE X and TABLE XI), and POIs are more important than mobile signal data and social media data in UFZ mapping, because POIs are more closely related to socioeconomic semantics of UFZs. In addition, mobile signal data and social media data are sparser, resulting in local lack of semantics.

TABLE X

The UFZ Mapping Accuracy (%) of the MDFF Under Different Data Configurations in Beijing

POIs	social media data	mobile signal data	images	OA (%)
✓				26.58
			✓	95.33
		✓	✓	96.87
	✓		✓	96.90
✓			✓	97.77
✓	✓	✓	✓	**97.79**

TABLE XI

The UFZ Mapping Accuracy (%) of the MDFF Under Different Data Configurations in Shanghai

POIs	social media data	mobile signal data	images	OA (%)
✓				26.60
			✓	92.01
		✓	✓	93.74

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3326160

	✓		✓	93.80
✓			✓	94.32
✓	✓	✓	✓	**94.45**

However, effectively fusing multimodal geographic data and features at the high level is difficult due to the different distributions and structures of the data. In the MDFF, two feature fusion structures are designed to fuse image data and social sensing data at pixel scale and object scale. The MDFF effectively fuses multimodal geographic data and generates accurate and reliable classification results. Compared with (a) and (b) in Fig.13, the UFZ results from the MDFF with images and social sensing data are more refined and accurate. The detailed results of (c) and (d) shows that the MDFF accurately identifies UFZs, especially in commercial, shantytown, industrial, water, transport. As shown in the red circles of (c) and (d), the MDFF corrects misclassifications with the help of socioeconomic semantics of UFZs from POI data, which demonstrates the effectiveness of multimodal geographic data fusion.

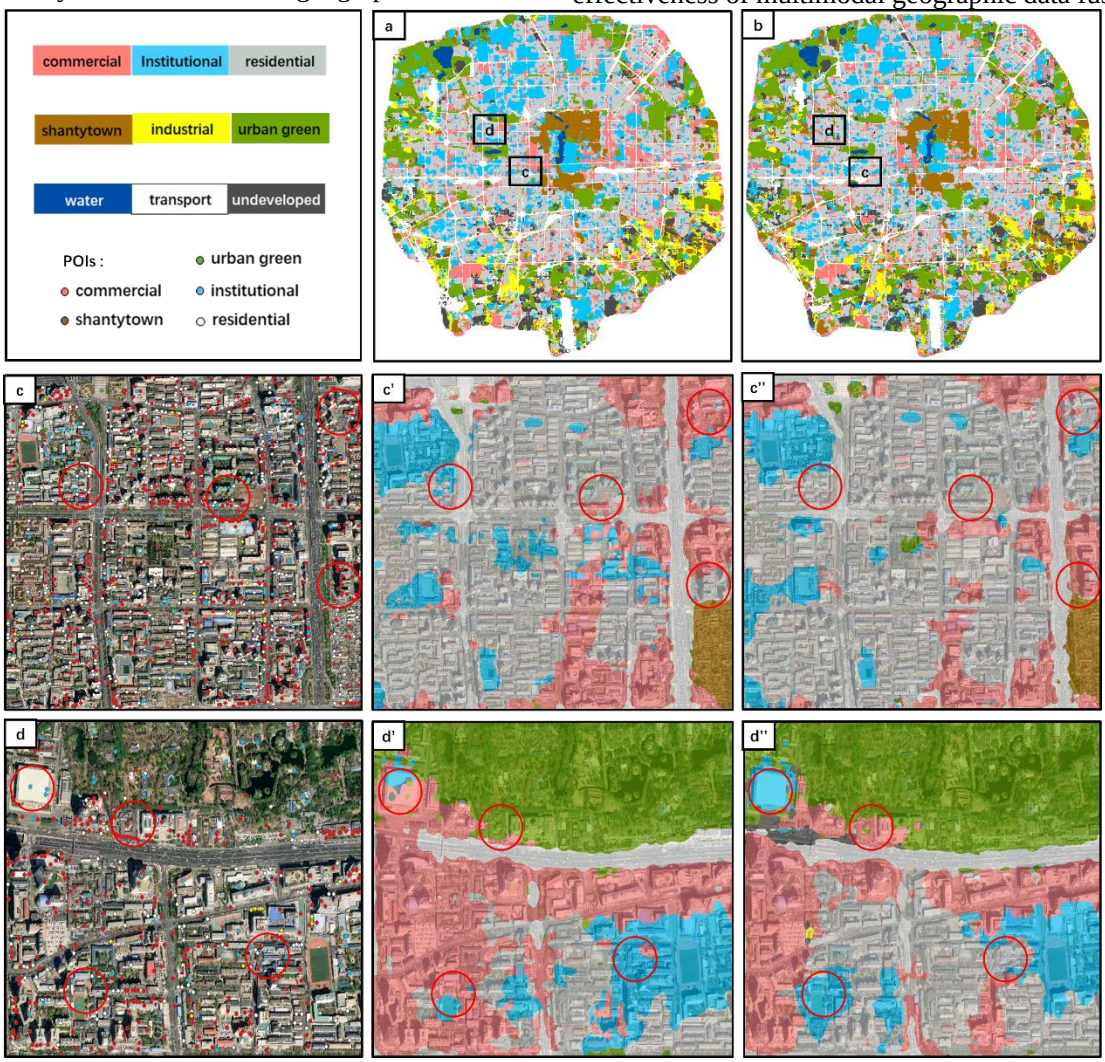


Fig. 13. Visual comparison of classification results. (a) is the UFZ map from the MDFF only with images, (b) is generated by the MDFF with images and social sensing data. (c) and (d) are sub areas of (a) and (b). (c') and (d') are UFZ results from (a). (c'') and (d'') are UFZ results from (b).

C. How to select multimodal geographic data for UFZ mapping

RS images carry the spectral and spatial properties of ground objects, which characterize UFZs in detail. In addition, RS images hold advantages in representing UFZs, because of the large spatial coverage and wide availability. UFZs are closely related to socioeconomic activities, thus socioeconomic attributes are very important for UFZ mapping. Socioeconomic semantics from social sensing data are helpful for interpreting UFZs. Three typical social sensing data, i.e., POIs, mobile signal data, and social media data, have different characteristics. POIs record the semantics of various subjects in socioeconomic activities, while social media data generally provide static information for a single subject and mobile signal data present the spatiotemporal flow of people's movement. Therefore, POIs provide richer and more representative socioeconomic semantics. RS images and the three typical social sensing data are selected for experiments. Experimental results in TABLE X and TABLE XI show that RS images contribute the most to the interpretation of UFZs, which demonstrates that RS images are essential to the interpretation. In addition, it can be seen from TABLE X and TABLE XI that POIs are

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3326160

more important than mobile signal data and social media data for UFZ mapping. In general, RS images and POIs are vital to the high-precision mapping of UFZs.

D. MDFF's extensibility

As demonstrated above, the MDFF serves as a general framework for mapping UFZs, and we have evaluated the framework on two typical modal data, i.e., RS images and POI, but it has different forms according to different modal data. Lots of different modal data are available in cities, such as images (VHR, SAR, DEM, nightlight images, etc.) and social sensing data (POIs, GNSS trajectories, mobile signal data, social media data, etc.). These data characterize the physical and socioeconomic attributes of urban zones from different perspectives. For example, nightlight images provide strength of nightlight related to economic development, and mobile signal data reflects population mobility. Therefore, the MDFF can be extended with more data for better performance. There are two approaches: 1) using another data, such as replacing POIs with mobile signal data, and 2) integrating all data (Fig.14) by adding branches (pixel-based module or object-based module) to extract features from data (image data or social sensing data). Then training the MDFF according to different types of input data. Thus, the MDFF is a flexible framework which can be easily extended according to the input data.

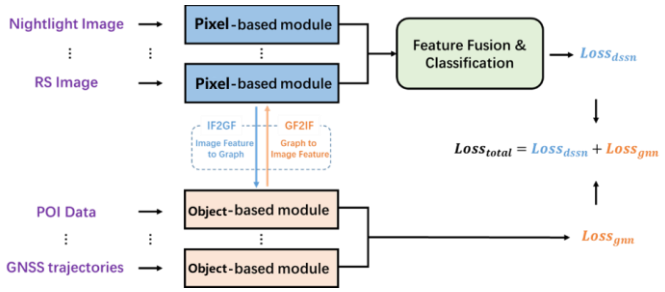


Fig. 14. The extended MDFF for integrating more data.

E. Pros and cons of the MDFF

The MDFF achieves good performance in interpreting UFZs due to the following three aspects. First, the MDFF effetely integrates multimodal geographic data and fuses multimodal features, which makes full use of physical compositions and socioeconomic characteristics of UFZs to reduce the uncertainty of interpretation. Second, unlike two-stage methods, the MDFF has only one stage and works end-to-end. It avoids error accumulation in stages and reduces mapping complexity, which facilitates fine-grained mapping in large urban areas. Third, the MDFF is a flexible framework which can be easily extended according to the input data, which is beneficial for dealing with unstable supply of urban data. However, the MDFF still has some limitations. First, objects generated by the multiresolution segmentation method are inconsistent with objects of the real world, which reduces the integrity of the mapping results. Second, UFZs are usually composed of land-cover objects, while the MDFF does not take semantics of land-covers into consideration. Consequently, these two issues need to be further studied.

V. Conclusions

UFZ maps are essential for urban studies and applications. It is a major open problem in the field of urban geography and survey to map fine-grained UFZs accurately and reliably. In order to solve the problem, the MDFF is proposed to fuse multimodal geographic data with deep network to map fine-grained UFZs end-to-end, which effectively integrates image data and social sensing data. With two designed fusion modules (i.e., the IF2GF and the GF2IF), the MDFF fuses multi-modal features to learn urban semantics. The semantic features of geographic objects assist the classification of UFZs. Experiments are conducted on two typical modal data, i.e., VHR satellite images and social sensing data. First, the MDFF significantly improve the quality of UFZ mapping with the accuracy about 5% higher than the state-of-the-art method, which shows advantages of the end-to-end mechanism and multimodal geographic data fusion of the MDFF. Second, socioeconomic semantics from social sensing data effectively enhance the interpretation's reliability. Third, the MDFF is a flexible framework which can be easily extended according to the input data, which is beneficial for dealing with unstable supply of urban data. Fourth, RS images and POIs are vital to the high-precision mapping of UFZs. Finally, the MDFF significantly reduces the complexity of UFZ mapping and fine-grained UFZ maps from the MDFF has effectively contributed to urban structure analysis.

Acknowledgments

The work is funded by the Chinese National Nature Science Foundation (No. 42001327).

References

[1] X. Zhang, S. Du, "A linear Dirichlet mixture model for decomposing scenes: application to analyzing urban functional zonings." *Rem. Sens. Environ*. 169, pp.37-49, 2015.

[2] X. Zhang, S. Du, S. Du, B. Liu, "How do land-use patterns influence residential environment quality? a multiscale geographic survey in Beijing." *Rem. Sens. Environ*. pp.249, 2020.

[3] R.H. Matsuoka, R. Kaplan, "People needs in the urban landscape: analysis of landscape and urban planning contributions." *Landscape Urban Planning*. 84 (1), pp.7–19, 2008.

[4] A.P. Montanges, G. Moser, H. Taubenböck, M. Wurm, D. Tuia, "Classification of urban structural types with multisource data and structured models." In: *2015 Joint Urban Remote Sensing Event (JURSE)*. pp. 1–4, 2015.

[5] J.C. Castella, S.P. Kam, D.D. Quang, P.H. Verburg, C.T. Hoanh, "Combining top-down and bottom-up modelling approaches of land use/cover change to support public policies: application to sustainable management of natural resources in northern Vietnam." *Land Use Policy*. 24 (3), pp.531–545, 2007.

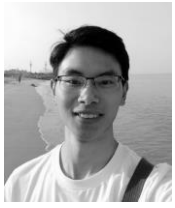
[6] U. Heiden, W. Heldens, S. Roessner, K. Segl, T. Esch, A. Mueller, "Urban structure type characterization using hyperspectral remote sensing and height information." *Landscape Urban Planning*. 105 (4), pp.361–375, 2012.

[7] H.B. Shin, "Residential redevelopment and the entrepreneurial local state: the implications of Beijing's shifting emphasis on urban redevelopment

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3326160

policies." *Urban Stud*. 46 (13), pp.2815–2839, 2009.
[8] T. Hu, J. Yang, X. Li, P. Gong, "Mapping urban land use by using Landsat images and open social data." *Rem. Sens.* 8 (2), pp.151, 2016.
[9] S. Du, S. Du, B. Liu, X. Zhang, "Mapping large-scale and fine-grained urban UFZs from VHR images using a multi-scale semantic segmentation network and object-based approach." *Rem. Sens. Environ*. 261(2), pp.112480, 2021.
[10] M.M. Nielsen, "Remote sensing for urban planning and management: the use of window-independent context segmentation to extract urban features in Stockholm." *Comput. Environ. Urban. Syst*. 52, pp.1–9, 2015.
[11] M. Amani, A. Ghorbanian, M. Asgarimehr, B. Yekkehkhany, A. Moghimi, J. Shuanggen, A. Naboureh, F. Mohseni, S. Mahdavi, NF. Layegh, "Remote Sensing Systems for Ocean: A Review (Part 1: Passive Systems)." *IEEE J. Sel. Top. Appl. Earth Observations Rem. Sens*. 15, pp. 210-234, 2022.
[12] Y. Zhang, Q. Li, H. Huang, W. Wu, X. Du, H. Wang, "The combined use of remote sensing and social sensing data in fine-grained urban land use mapping: a case study in Beijing, China." *Remote Sens*. 9, pp.865, 2017b.
[13] J.K. Steeves, G.K. Humphrey, J.C. Culham, R.S. Menon, A.D. Milner, M.A. Goodale, "Behavioral and neuroimaging evidence for a contribution of color and texture information to scene classification in a patient with visual form agnosia." *J. Cogn. Neurosci*. 16 (6), pp.955–965, 2004.
[14] M. Voltersen, C. Berger, S. Hese, C. Schmullius, "Object-based land cover mapping and comprehensive feature calculation for an automated derivation of urban structure types at block level." *Remote Sens. Environ*. 154, pp.192–201, 2014.
[15] X. Zhang, S. Du, Y.-C. Wang, "Semantic classification of heterogeneous urban scenes using intrascene feature similarity and interscene semantic dependency." *IEEE J. Sel. Top. Appl. Earth Observations Rem. Sens*. 8, pp.2005–2014, 2015b.
[16] R. Kusumaningrum, H. Wei, R. Manurung, A. Murni, "Integrated visual vocabulary in latent Dirichlet allocation–based scene classification for IKONOS image." *J. Appl. Remote. Sens*. 8, pp.083690, 2014.
[17] M. Lienou, H. Maître, M. Datcu, "Semantic annotation of satellite images using latent Dirichlet allocation." *IEEE Geosci. Remote Sens. Lett*. 7 (1), pp.28–32, 2010.
[18] G. Moser, S.B. Serpico, "Combining support vector machines and Markov random fields in an integrated framework for contextual image classification." *IEEE Trans. Geosci. Remote Sens*. 51 (5), pp.2734–2752, 2013.
[19] B. Zhao, Y. Zhong, G.-S. Xia, L. Zhang, "Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery." *IEEE Trans. Geosci. Remote Sens*. 54, pp.2108–2123, 2015.
[20] Y. LeCun, Y. Bengio, G. Hinton, "Deep learning." *Nature*. 521, pp.436-444, 2015.
[21] L. Ma, Y. Liu, X. Zhang, Y. Ye, "Deep learning in remote sensing applications: A meta-analysis and review." *J Photogramm Remote Sens*. 152, pp.166–177, 2019.
[22] G. Mountrakis, J. Li, X. Lu, O. Hellwich, "Deep learning for remotely sensed data." *J Photogramm Remote Sens*. 145, pp.1–2, 2018.
[23] W. Zhou, D. Ming, X. Lv, K. Zhou, H. Bao, Z. Hong, "SO–CNN based urban functional zone fine division with VHR remote sensing image." *Remote Sens. Environ*. 236, pp.111458, 2020.
[24] S. Du, S. Du, B. Liu, X. Zhang, Z. Zheng, "Large-scale urban functional zone mapping by integrating remote sensing images and open social data." *GISci. & Rem. Sens*. 9, pp.1-20, 2020.
[25] X. Zhang, S. Du, Q. Wang, "Hierarchical semantic cognition for urban UFZs with VHR satellite images and POI data." *J Photogramm Remote Sens*. 132, pp.170-184, 2017.
[26] S. Ouyang, Y. Li, "Combining deep semantic segmentation network and graph convolutional neural network for semantic segmentation of remote sensing imagery." *Rem. Sens*. 13(1), pp.119, 2020.
[27] S.W. Myint, P. Gober, A. Brazel, S. Grossman-Clarke, Q.Weng, "Per-pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery." *Remote Sens. Environ*. 115, pp.1145–1161, 2011.
[28] H. Bao, D. Ming, Y. Guo, K. Zhang, S. Du, "DFCNN-based semantic recognition of urban UFZs by integrating remote sensing data and POI data." *Rem. Sens*. 12(7), pp.1088, 2020.
[29] T. Blaschke, "Object based image analysis for remote sensing. ISPRS J. Photogramm. Remote Sens. 65 (1), pp.2–16, 2010.
[30] C. Cleve, M. Kelly, F.R. Kearns, M. Moritz, "Classification of the wildland–urban interface: a comparison of pixel-and object-based classifications using high-resolution aerial photography." *Comput. Environ. Urban. Syst*. 32, pp.317–326, 2008.
[31] J.M. Peña-Barragán, M.K. Ngugi, R.E. Plant, J. Six, "Object-based crop identification using multiple vegetation indices, textural features and crop phenology." *Remote Sens. Environ*. 115 (6), pp.1301–1316, 2011.
[32] S.C. Zhu, R. Zhang, Z. Tu, "Integrating bottom-up/top-down for object recognition by data driven Markov chain Monte Carlo." In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. vol. 1. IEEE, pp. 738–745, 2000.
[33] H. Rhinane, A. Hilali, A. Berrada, M. Hakdaoui, "Detecting slums from SPOT data in Casablanca Morocco using an object-based approach." *J. Geogr. Inform. Syst*. 3 (03), pp.217, 2011.
[34] E. Farahzadeh, T.J. Cham, W. Li, "Semantic and spatial content fusion for scene recognition." In: *New Development in Robot Vision*. Springer, Berlin Heidelberg, pp.33–53, 2015.
[35] Y. Jiaqi, W. Chen, D. Bo, Z. Liangpei, "Enhanced Multiscale Feature Fusion Network for HSI Classification." *IEEE Trans. Geosci. Remote Sens*. 59(12), pp.10328-10347, 2021.
[36] J. Stuckens, P.R. Coppin, M.E. Bauer, "Integrating contextual information with per-pixel classification for improved land cover classification." *Remote Sens. Environ*. 71 (3), pp.282–296, 2000.
[37] S. Du, S. Du, B. Liu, X. Zhang, "Context-enabled extraction of large-scale urban UFZs from very-high-resolution images: a multiscale segmentation approach." *Remote Sens.* 11, pp.1902, 2019.
[38] M. Baatz, A. Schape, "Multiresolution segmentation: an optimization approach for high quality multi-scale image segmentation." *Angew. Geogr. Information sverarbeitung*. 12, pp.12–23, 2000.
[39] X. Zhang, S. Du, Q. Wang, "Integrating bottom-up classification and top-down feedback for improving urban land-cover and functional-zone mapping." *Remote Sens. Environ*. 212, pp.231-248, 2018.
[40] G.J. Hay, T. Blaschke, D.J. Marceau, A. Bouchard, "A comparison of three image-object methods for the multiscale analysis of landscape structure." *ISPRS J. Photogramm. Remote Sens*. 57, pp.327–345, 2003.
[41] T. Kipf, M. Welling, "Semi-supervised classification with graph convolutional networks." *Proceedings of the international conference on learning representations*, 2017.
[42] P. Veličković, G. Cucurull, A. Casanova, "Graph Attention Networks." *In Proceedings of the International Conference on Learning Representations*, Vancouver, BC, Canada, 30 April–3 May 2018.
[43] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation." *Proceedings of the European Conference on Computer Vision*, 2018.
[44] B. Huang, B. Zhao, Y. Song, "Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery." *Remote Sens. Environ*. 214, pp.73–86, 2018.
[45] M. Simwanda, Y. Murayama, "Integrating geospatial techniques for urban land use classification in the developing sub-Saharan African city of Lusaka, Zambia." ISPRS Int. J. Geo Inf. 2017.
[46] J. Rosier, H. Taubenbock, P. Verburg, J. Vliet, "Fusing Earth observation and socioeconomic data to increase the transferability of large-scale urban land use classification." *Remote Sens. Environ*. 278(113076), 2022.



Song Ouyang received the B.S. degree and the M.S. in the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China. He is currently pursuing the Ph.D. degree in the Institute of Remote Sensing and Geographic Information System, Peking University, Beijing, China. His current research interests include knowledge reasoning, urban landscape modeling, remote sensing, GIS, and ontology modeling.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

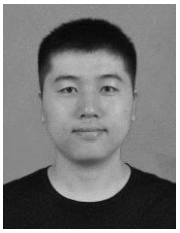
This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3326160



Shihong Du received the B.S. and M.S. degrees in cartography and geographic information system from Wuhan University, Hubei, China, in 1998 and 2001, respectively, and the Ph.D. degree in cartography and geographic information system from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China, in 2004. He is currently a Professor with Peking University, Beijing. His specific interests include urban remote sensing, machine learning, object-based image analysis, time series analysis, change detection and dynamic analysis of land cover/use, urban environmental information extraction, and applications of big earth data in support of landscape change analysis and urban sustainable development.



Xiuyuan Zhang received the B.S. degree in remote sensing from China University of Geosciences (Wuhan) in 2014, and the Ph. D degree in geographic information system from Peking University in 2019. He is currently a Research Associate at the Institute of Remote Sensing and GIS, Peking University. His research interests include remote sensing and GIS for urban landscape modeling and urban sustainable development monitoring.



Shouhang Du received the B.S. degree in surveying and mapping engineering from the China University of Geosciences (Wuhan), Wuhan, China, the M.S. degree in photogrammetry and remote sensing from the Central South University, Changsha, China and the Ph.D. degree in cartography and geographic information system from the Peking University, Beijing, China, in 2014, 2017, and 2021, respectively.

He is currently a lecturer with China University of Mining and Technology (Beijing), Beijing, China. His research interests include intelligent understanding of spatial data including GIS and remote sensing data.



Lubin Bai received the bachelor's degree in survey and mapping engineering from Wuhan University, Wuhan, China, in 2020. He is currently pursuing the master's degree in cartography and geographical information system with Peking University, Beijing, China. His research interests include remote sensing image processing, self-supervised learning, and representation learning.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/