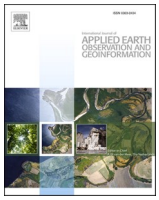
International Journal of Applied Earth Observations and Geoinformation 106 (2022) 102661

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

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



A hierarchical approach for fine-grained urban villages recognition fusing remote and social sensing data

Dongsheng Chen [a,b,c], Wei Tu [a,b,*], Rui Cao [d], Yatao Zhang [e], Biao He [a,b], Chisheng Wang [a,b], Tiezhu Shi [a,b], Qingquan Li [a,b]

[a] Guangdong Key Laboratory of Urban Informatics, and Shenzhen Key Laboratory of Spatial Smart Sensing and Service, School of Architecture and Urban Planning, Shenzhen University, Shenzhen 518060, China
[b] Ministry of Natural Resources (MNR) Key Laboratory for Geo-Environmental Monitoring of Great Bay Area, Shenzhen University, Shenzhen 518060, China
[c] China Regional Coordinated Development and Rural Construction Institute, Sun Yat-sen University, Guangzhou 510275, China
[d] Smart Cities Research Institute, Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Kowloon, Hong Kong Special Administrative Region
[e] Future Resilient Systems, Singapore-ETH Centre, ETH Zurich, Singapore 138602, Singapore

ARTICLE INFO

Keywords:
Informal settlement
Urban villages
Hierarchical recognition
Remote sensing
Social sensing

ABSTRACT

Timely and accurate maps of fine-grained urban villages (UVs) are essential for rational urban planning, which highlights the importance for automatic recognition methods as alternative to labor-intensive land survey, especially for large cities with high-density urban areas where UV maps cannot be updated frequently. However, it is challenging to simultaneously achieve accurate and fine-grained recognition of UVs from remote sensing images in high-density cities, due to the problem of low discrimination of remote sensing features showed in UVs. To address this issue, in this paper, we have proposed a hierarchical recognition framework which can integrate remote and social sensing data to recognize fine-grained UVs. The hierarchical framework follows the human cognition processes and has explicit geographical meaning for each step, which ensures its interpretability. Besides, remote and social sensing data can be fused easily in this framework so that the abstract concept of UV can be sufficiently characterized in both coarse and fine scales. To validate the effectiveness of the proposed approach, extensive experiments in Shenzhen, a typical high-density megacity in China with complicated UVs, have been conducted and a fine-grained map with spatial resolution of 2.5 m was obtained. The results show that the proposed approach achieved an impressive performance, with overall accuracy and Kappa of 96.23% and 0.920 respectively. Furthermore, comparative assessments and ablation studies were performed to demonstrate the effectiveness of the hierarchical recognition framework as well as the fusion of remote and social sensing data.

1. Introduction

Rapid global urbanization in the past fifty years has led to many informal settlements (Gallaher et al., 2013), such as the Dharavi Slum (Mumbai, India), the Rocinha Favela (Rio De Janeiro, Brazil), and the Kibera Slum (Nairobi, Kenya) (Gallaher et al., 2013; Handzic, 2010; Sharma, 2000). In China, many informal settlements have emerged in large cities, which manifest themselves as urban villages (UVs) (Hao et al., 2013). From an ecological landscape perspective, UVs usually suffer from poor livability (e.g., overcrowding of migrant workers, lack of public facilities, security risks, etc.), and it is always associated with problems such as traffic congestion, social segregation, and socio-economic inequalities (Wang et al., 2009; Friesen et al., 2018). Therefore, it is necessary to well manage UVs, especially for large high-density urban areas where it's difficult to manage these UVs due to the over-crowded land use distribution and rapid development (Guan et al., 2018). Timely and accurate maps of fine-grained UVs are essential for efficient urban management, which however are usually unavailable to the public. Hence, it is necessary and valuable to develop automatic recognition approaches to obtain up-to-date fine-grained maps of UVs from accessible data sources.

The fine-grained UV recognition problem cannot be solved well by conventional remote sensing (RS)-based methods. RS has demonstrated to be an effective tool for urban land recognition (Kuffer et al., 2016a;

* Corresponding author at: Huixinglou 1406, Shenzhen University, Nanhai Avenue 3688, Shenzhen 518060, China
E-mail address: tuwei@szu.edu.cn (W. Tu).

https://doi.org/10.1016/j.jag.2021.102661
Received 18 August 2021; Received in revised form 21 November 2021; Accepted 20 December 2021
Available online 29 December 2021

Wurm et al., 2017a). However, unlike the concept of land use with distinct spectral features (e.g., forest), the concept of UV is more abstract and cannot be easily recognized. On the one hand, UVs show diverse forms within the same city due to differences in architectural styles or building policies in different regions, including high-rise UVs, mid-rise UVs and low-rise UVs, as shown in Fig. 1 (a). On the other hand, UVs have similar features to other types of urban land, such as commercial zones, residential districts, high-class residential areas, etc., as shown in Fig. 1 (b). Because they all consist mainly of buildings and impervious surfaces, from the perspective of geographical objects. This leads to the problem of UVs' low-discrimination rate, which limits the achievable accuracy from single RS-based interpretation (Huang et al., 2015; Taubenböck et al., 2018; Wurm et al., 2019).

Additionally, current context-aware methods are difficult to meet the need for fine-grained UV recognition. Context-aware methods are normally used to infer the theme of a geographical area from the combination of elements in that area, for example, object-oriented methods (Blaschke et al., 2014; Chen et al., 2018; Mboga et al., 2019) and spatial scene understanding (Li et al., 2017; Zhang and Du, 2015; Zhang et al., 2019). They often have good performance at coarse scale because the regional elements contain redundant spatial context information. However, it's difficult for these methods to perform well on fine-grain tasks. Besides, unlike regional urban function recognition, recognizing urban villages need to be performed at fine-grained scale without redundant spatial context information. It further increases the difficulty of UV recognition. Thus, previous studies of UV detection often cannot simultaneously achieve excellent recognition accuracy and fine-grained results (Huang et al., 2015; Wurm et al., 2017b; Wurm et al., 2019).

The hierarchical recognition (HR) framework provides a new idea for the UVs recognition methods. In the perspective of the human perceptual system, perceptual and cognitive processes are relative to the decomposition of visual information in different spatial scales (Peyrin

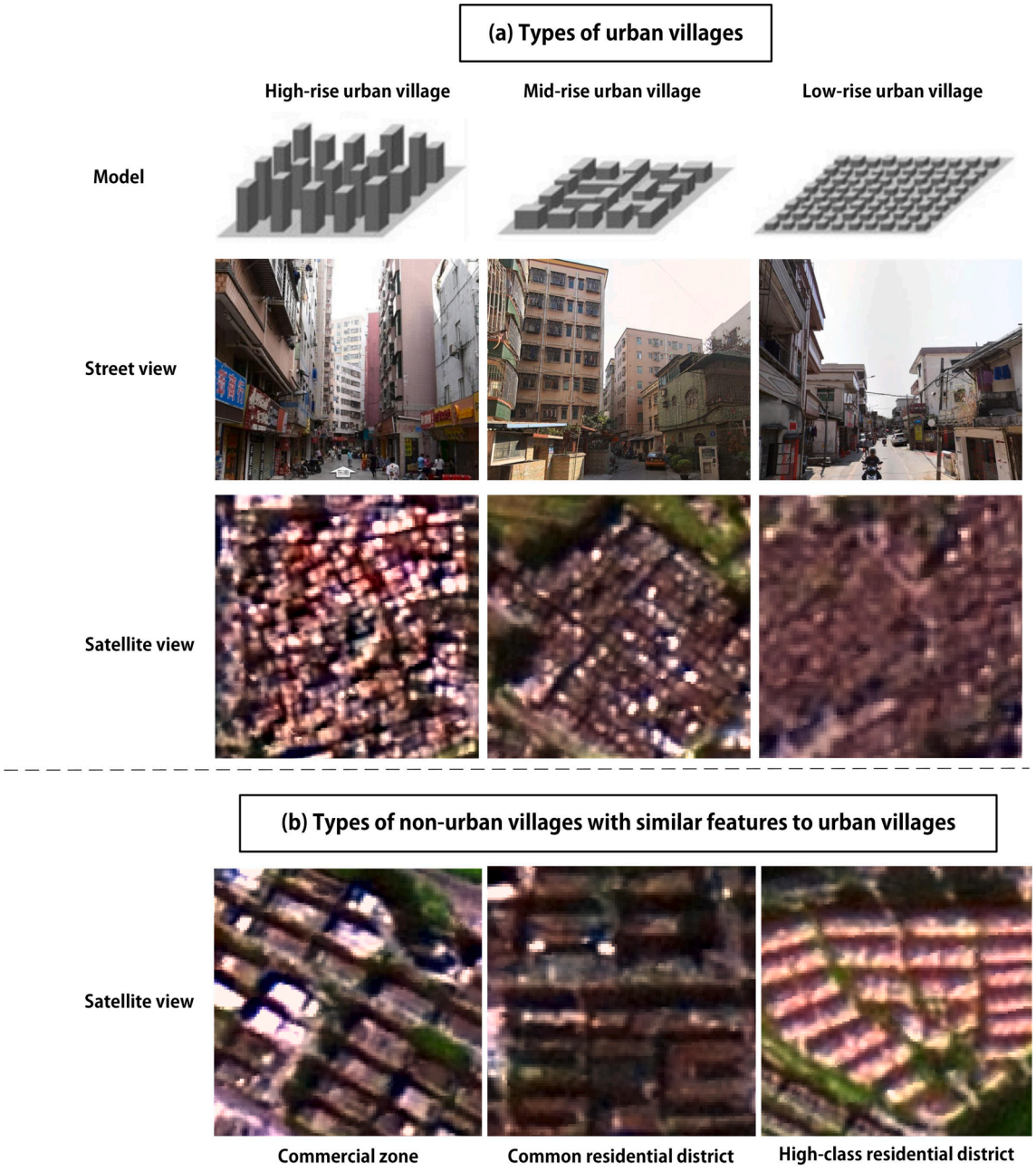


Fig. 1. Comparison of remote sensing images of diverse forms of urban villages with the types of non-urban villages.

et al., 2010). And the HR framework can discover the multi-scale spatial information and promote the high-level visual categorization during the human cognitive processes (Peyrin et al., 2010). Zhang et al. (2017) proposed that considering hierarchical relations can improve the performance of geographic recognition and applied it on functional zone recognition. Inspired by the processes, this paper expands this philosophy to the application of machine recognition of the UVs. UVs recognition is equivalent to a high-level process of machine recognition as UVs are the geographical objects with low discrimination of RS features. With the advantages of simple structure and strong interpretability, the HR can enhance the understanding of UVs by uniting the coarse-scale context information with fine-scale local information (Fig. 2) (Zhang et al., 2018). Thus, the HR process can reduce the difficulty of sensing geo-objects in high-density cities like Shenzhen and Shanghai. In this study, we develop a new approach based on the HR framework to recognize UVs in high-density cities with good recognition accuracy and high spatial resolution.

Social sensing (SS) data is a valid complement to RS images (Liu et al., 2015b; Zhu et al., 2019). Most previous studies only relied on RS images in the field of UVs recognition (Kuffer et al., 2016a; Wurm et al., 2019). Low-level features (e.g., spectral and textural features) and high-level features (e.g., vegetation indexes, deep convolutional features) extracted from RS images are the most widely used for geo-objects recognition, like buildings, roads, etc. (Zhang et al., 2019). But it is still difficult to recognize geo-objects with similar physical features but different social functions. Fortunately, recent development of smart cities enables us to acquire massive SS data, i.e., points of interest (POIs), vehicle trajectories, social media check-in records, etc., which capture human mobility and activities and thus contain rich socioeconomic information about land functions (Tu et al., 2020). Fusing RS images and SS data is a promising method to understand urban scenes in high-density cities (Cao et al., 2020; Tu et al., 2020; Zhu et al., 2019).

In this paper, we present a **h**ierarchical **r**ecognition framework with **r**emote and **s**ocial sensing **f**usion for fine-grained **UV** recognition (we call it HR-RSF-UV framework). First, the HR framework following the human cognition processes is integrated into the proposed HR-RSF-UV framework, which is both flexible and interpretable. It enables the use of redundant spatial context information on the fine-grained recognition task by uniting coarse-scale context information and fine-scale local information. Also, the RS and SS data are fused to characterize the abstract concept of UVs in both coarse and fine scales to capture UVs' distinct environmental and socioeconomic information. These sufficient information can be applied to solve the problem of UVs' low-discrimination rate. Hence, the proposed HR-RSF-UV framework is equipped with the ability to recognize UVs accurately at fine-grained scale. The major contributions of this paper are highlighted as follows:

- **The hierarchical recognition framework** is integrated into our proposed approach and its improved effect was demonstrated.
- **Remote sensing and social sensing data** are effectively fused to recognize fine-grained urban villages in high-density cities.
- **A case study** in Shenzhen, a typical high-density megacity in China with complicated UVs, has been implemented to evaluate the performance of the presented approach. The results demonstrate that the presented approach achieves high accuracy and outperforms traditional methods.
- **The zone-based training strategy and parameter sensitivity analysis** are also presented, which provide insights for urban scene understanding.

The remainder of this paper is organized as follows. Section 2 introduces the study area and datasets. Section 3 describes the details of the methodology. Section 4 reports and analyzes the results. Section 5 discusses the contributions of the HR framework and RS and SS data integration. Section 6 concludes this study.

2. Study area and data

This study is conducted in Shenzhen, which is located in the northern part of the Pearl River Delta in southern China (113°46′ - 114°37′*E*, 22°27′ *to* 22°52′*N*) (Fig. 3). Shenzhen, with a total area of 1,996.850 km^2, governs 10 administrative districts. Of these districts, Futian, Luohu, Nanshan and Yantian form the special economic zone (SEZ) in Shenzhen, while the remaining administrative districts are collectively known as the non-special economic zone (non-SEZ) (Fig. 3 (b)). The SEZ area is the most populous and developed region in Shenzhen, accounting for 52.48% of the city's total GDP (Fig. 3 (a)). With a huge population of more than 13 million and a complex urban land use pattern, Shenzhen need to alleviate the problems of UVs that bring urban poverty and economic inequality (Wang et al., 2009).

The multi-source geospatial data used consist of two main categories, i.e., RS data and SS data. The former contains the RS image and the nighttime light image, while the latter include POIs and taxi trajectory data.

- The RS image used is synthesized from SPOT-5's HSR images and panchromatic images covering Shenzhen using the pan-sharpening fusion method (Rahmani et al., 2010). The SPOT-5 images were obtained on November 30, 2013. The RS image has 37,368 × 19,440 pixels with a spatial resolution of 2.5 m per pixel, and four spectral channels, including near infrared, red, green and blue channel (Zhang et al., 2019).
- The nighttime light image is obtained from the Version 4 DMSP-OLS Nighttime Lights Time Series (http://ngdc.noaa.gov/eog). The time span of this dataset covers from 1992–2013. The night light shows the light intensity of built-up areas, thereby having the potential to map the urban development areas. Thus, we collected the nighttime light image obtained in 2013, and then extracted Shenzhen's urban

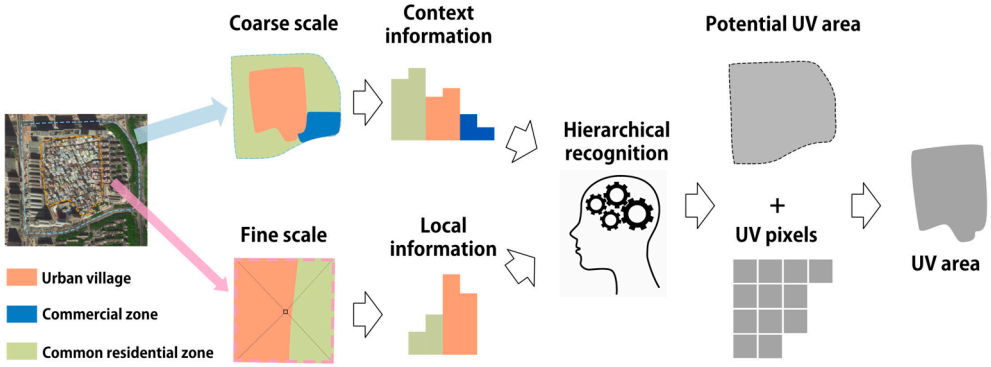


Fig. 2. The concept of the hierarchical recognition framework.

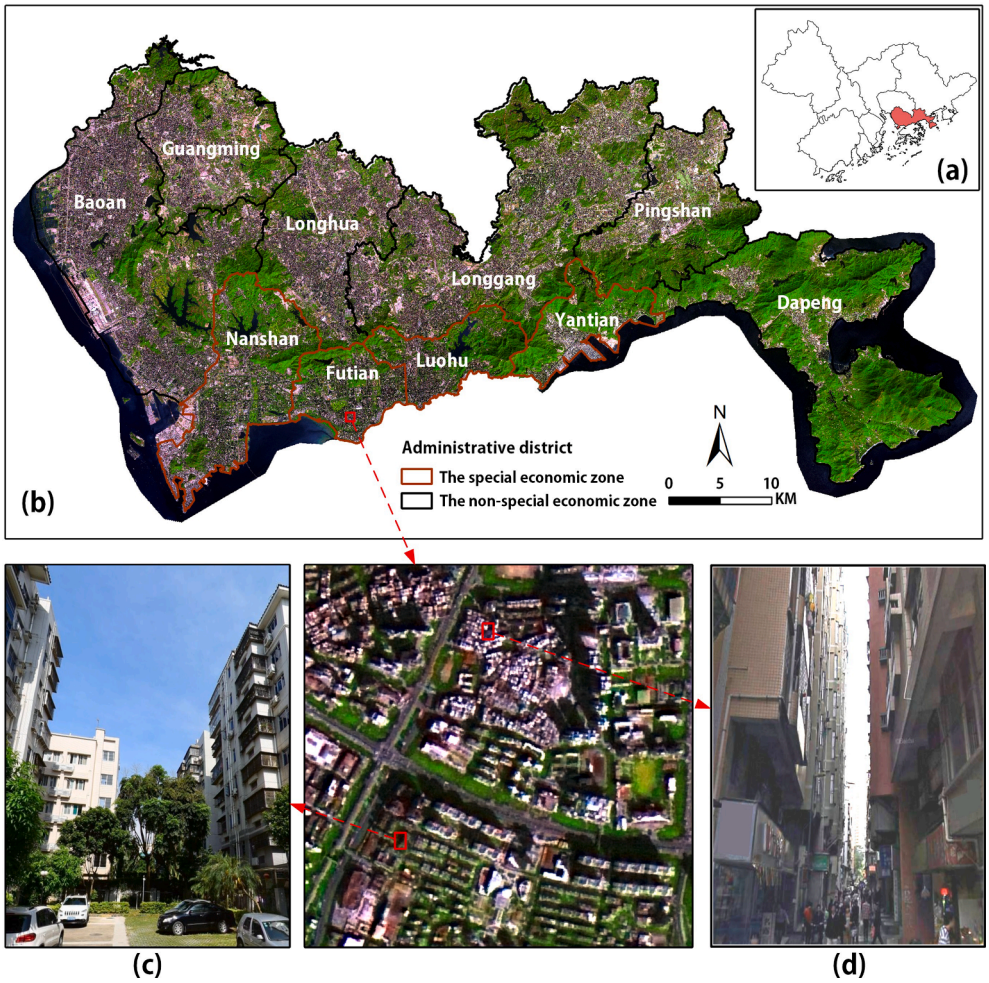


Fig. 3. Study area and examples of formal settlements and urban villages. (a) Shenzhen located in the GBA, (b) the administrative districts in Shenzhen, (c) a formal settlement, and (d) an urban village.

development extent by using the algorithm of urban extent (UE) mask (Yao et al., 2018).

- POI data refer to geographic points recorded in the actual geographic places. Each POI contains several aspects of property: place name, type of place function, longitude and latitude, etc., which has been widely used in urban studies (Mou et al., 2019; Zhang et al., 2019). The POI data used are obtained from a China's online map site. Through web crawlers and the application programming interface of the map sites, we crawled 211,076 POI records within Shenzhen in 2015.
- Taxi trajectory data were obtained from the smart GPS receiver installed inside taxis, which record data concerning the vehicle identification, time, position, speed, and working status. Taxi trajectory data are widely used in urban travel analysis (Mou et al., 2019). This paper employed the taxi trajectory dataset with 35,546,005 records for a 181-day period from January 1st, 2016 to June 30th, 2016.

Despite the time differences of the multi-source data, the recognition experiments are reasonable and do not significantly affected, since both the built-up areas and planned demolitions of Shenzhen were maintained with no significant changes during 2013–2016 (Dou and Chen, 2017; Lai et al., 2021). However, it should be noted that it is ideal to collect all sources of data from the same period, otherwise there should be no significant changes among the time gap of the collection of different data sources.

The ground-truth data of UVs are collected within the urban development areas of Shenzhen. Here, it should be noted that the extent of urban development areas is different from the administrative extent (as illustrated in Fig. 4 (a)), since the conceptual spatial locations of UVs are within highly developed urban built-up areas. Specifically, in this study, Shenzhen's urban development extent is extracted from the nighttime light image (Yao et al., 2018), as shown in Fig. 4 (b). Within the UE mask, high-rise UVs, mid-rise UVs, and low-rises UVs (Fig. 1) are labeled as the ground-truth data by experts through visual interpretation based on RS images, street view images, and urban planning documents.

3. Methodology

The details of the proposed HR-RSF-UV approach are described in this section. As Fig. 5 shows, the HR-RSF-UV approach is composed of three major steps. First, multi-source geospatial data are pre-processed to produce spatial and social features. Second, coarse-scale features that consider regional contextual information are obtained to detect potential UV areas from top to down at the coarse scale. Finally, within the extents of potential UV areas, fine-scale features of local details are derived for fine-scale UV areas recognition from bottom to up. Here, the "coarse scale" refers to the geospatial units that include the surrounding areas of the UVs and thus contain the context information (i.e., neighborhood). While the "fine scale" denotes the geospatial units with area smaller than urban villages, which are used to recognize the local detail information (e.g., edges, points, etc.) of UVs. In the proposed HR-RSF-UV framework, a geospatial unit at the coarse scale is represented by a homogeneous patch (a.k.a. a geo-object) and a unit at the fine scale is

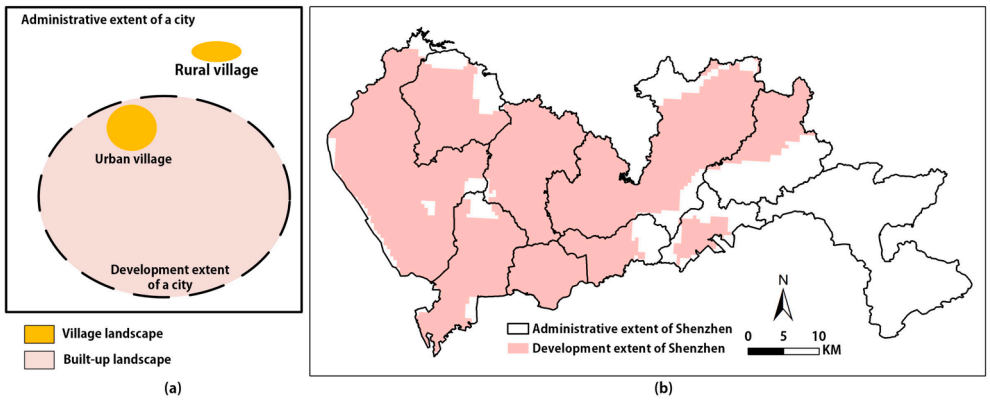


Fig. 4. The relationship between administrative urban extent and urban development extent. (a) A conceptual city, (b) Shenzhen City.

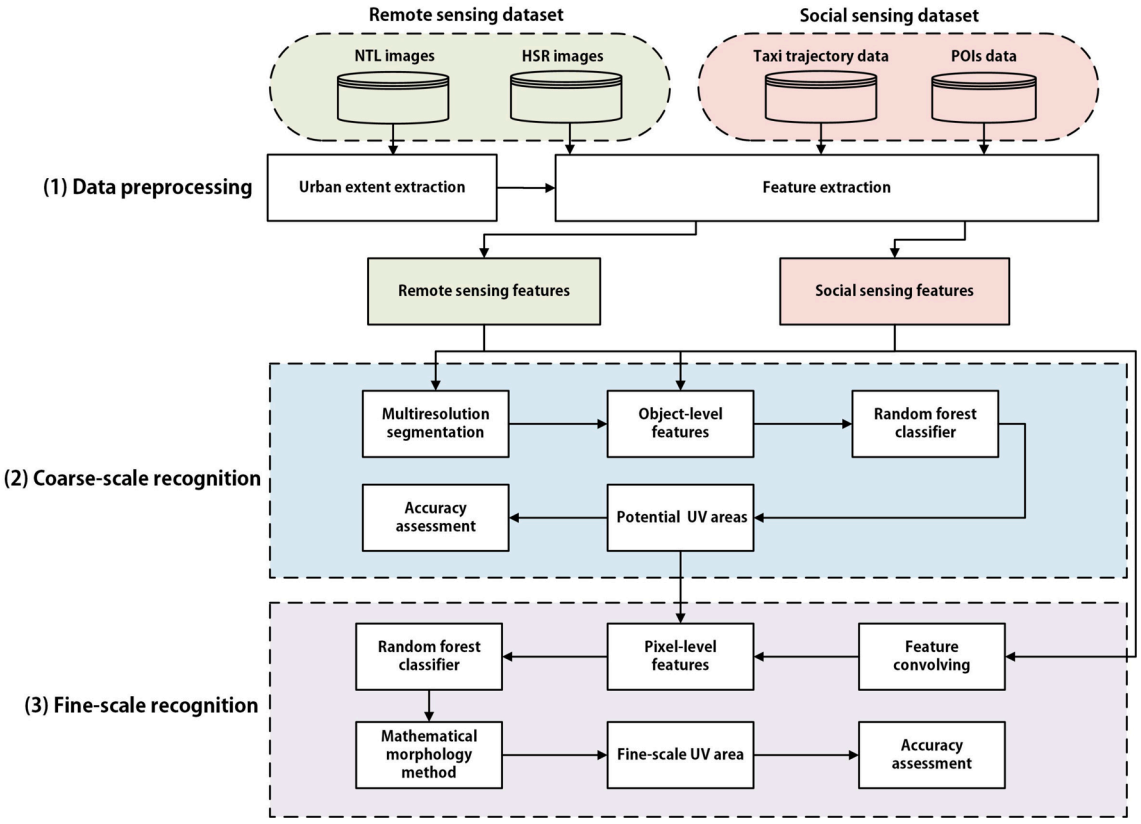


Fig. 5. Workflow of the HR-RSF-UV approach.

expressed in the form of the center pixel of a sliding window (Fig. 2).

3.1. Data preprocessing

The goal of this step is to pre-process multi-source geospatial data to extract features characterizing UVs. Specifically, RS features, and SS features are extracted from RS images, and SS data, respectively. Depending on the recognition scale, we will set geo-object masks or sliding windows below to extract the corresponding features of study units (geo-objects or pixels).

1) HSR images: Basic RS landscape features, including spectral, textural, and structural features are widely applied in land use classification and object recognition (Zhang et al., 2019). Given a RS image with n channels, we calculate the spectral descriptors of the image in a window, including the mean and standard deviation. The spectral features can be described as: $spectral\ features = \{mean_1, std_1, \cdots mean_n, std_n\}$. The RS image that we used contains 4 channels, so the spectral features are 8 dimensions in total.

This paper applies the gray-level co-occurrence matrix (GLCM) to describe the textural features. Four commonly used Haralick's GLCM statistics are calculated in a window, including contrast, energy, correlation, and homogeneity (Haralick et al., 1987). And the textural features can be described as: $texture\ features = \{con_1, ene_1, cor_1, hom_1, \cdots, con_n, ene_n, cor_n, hom_n\}$. The features have 16 dimensions in total.

Scale-invariant feature transform (SIFT) descriptor is widely used to express the structural features of an image (Farabet et al., 2013). SIFT algorithm identifies key points in the image and generates 128-dimensional feature vectors. While the Dense-SIFT method divides the target image into rectangular blocks of the same size and then calculates SIFT features for each raster, so the SIFT features are extracted at the fine

scale (Liu et al., 2015a). Thus, this paper uses the Dense-SIFT to describe the structural features: *structural features* = $\{sift_1, \cdots, sift_{128}\}$. All the above features are common RS features.

Deep convolution features have been proven to be effective in object recognition (Li et al., 2017). VGG-Net is one of the most popular deep convolutional neural networks which has a simple structure and has been shown to perform well in both image classification and object recognition tasks (Qassim et al., 2018). Specifically, VGG16 pretrained on ImageNet (a large image dataset with more than 1 million annotated images (Jia et al., 2009)) has been exploited to extract the deep convolution features from the RS imagery in our experiments, since previous works show that the pretrained VGG16 shows a strong capability of transfer learning and can be applied directly for image feature extraction for downstream tasks (Ma et al., 2020; Qassim et al., 2018). 13 convolutional layers of the model are used to process the image in a moving window. After global average pooling, features with 512 dimensions are extracted, and can be described as: *deep convolution features* = $\{vgg_1, \cdots, vgg_{512}\}$.

2) POIs: POI data contain semantic information of places, which helps to understand place functions. We merge the semantic information into 20 types. The densities of these POI types constitute the place semantic features of a region. Through point density analysis and regional average calculation, a 20-dimensional numeric vector is obtained with each entry representing the density of a certain POI type, which can be formulated as *place semantic features* = $\{poi_1, \cdots, poi_{20}\}$.

3) Taxi OD data: For a taxi trip, the time and location at which a passenger is picked up and dropped off are regarded as the origin and destination (OD) of the taxi trip. The sequence of OD frequencies in one place reveal the taxi travel activities of this place, which is important for studying urban mobility patterns and land use classification (Mou et al., 2019). Given any taxi trip T_r, when the taxi status changes from vacant to occupied at (x_o, y_o) in time t_o, the triplet $< t\_o, x\_o, y\_o >$ can be defined as the space–time point of origin O. Conversely, when the status changes from occupied to vacant at (x_d, y_d) in time t_d, the space–time point $D =< t\_d, x\_d, y\_d >$ is treated as the destination. Next, the OD points are divided into two coarse groups in terms of weekday and weekend. And then the OD points are categorized into 48 fine group at hourly intervals based on time stamps. After the point density analysis, a series of the hourly average maps of OD densities at the fine scale are obtained, which can be described as *taxi travel features* = $\{\{O\_w, 1, \cdots, O\_w, 24, \cdots, O\_r, 1, \cdots, O\_r, 24\}, \{D\_w, 1, \cdots, D\_w, 24, \cdots, D\_r, 1, \cdots, D\_r, 24\}\}$.

3.2. Coarse-scale UV area recognition

This step aims to detect potential UV areas at the coarse scale. Specifically, coarse-scale recognition mainly includes two steps. (1) Regional segmentation: divide the urban space into homogeneous patches. (2) Potential UV areas recognition: extract coarse-scale features and then use a classifier to classify the homogeneous objects.

To segment regions, the multi-resolution segmentation (MRS) algorithm is used, which is a classic segmentation algorithm widely used in RS image segmentation (Chen et al., 2019). We use eCognition to implement MRS to segment the urban area into regions. The scale parameter of this algorithm is suggested to be set to guarantee that homogeneous patches cover the areas slightly larger than regions of urban villages and their surrounding area. So sufficient contextual information can be collected to perform effective recognition of potential UV areas.

Next, potential UV area recognition is performed based on the objects derived from regional segmentation. This step entails recognizing objects that contain UVs, i.e., potential UV areas. Fig. 6 shows several examples of UV and non-UV areas. The regional context information of potential UV areas presents noticeable differences from non-UV areas. It is worth noting that Fig. 7 (a) (b) show that the numbers of catering and life POIs, as well as Taxi ODs in urban villages far exceeds those of other negative samples. It is consistent with the general pattern in urban villages that the social activities in urban villages are often different from those in ordinary residential communities.

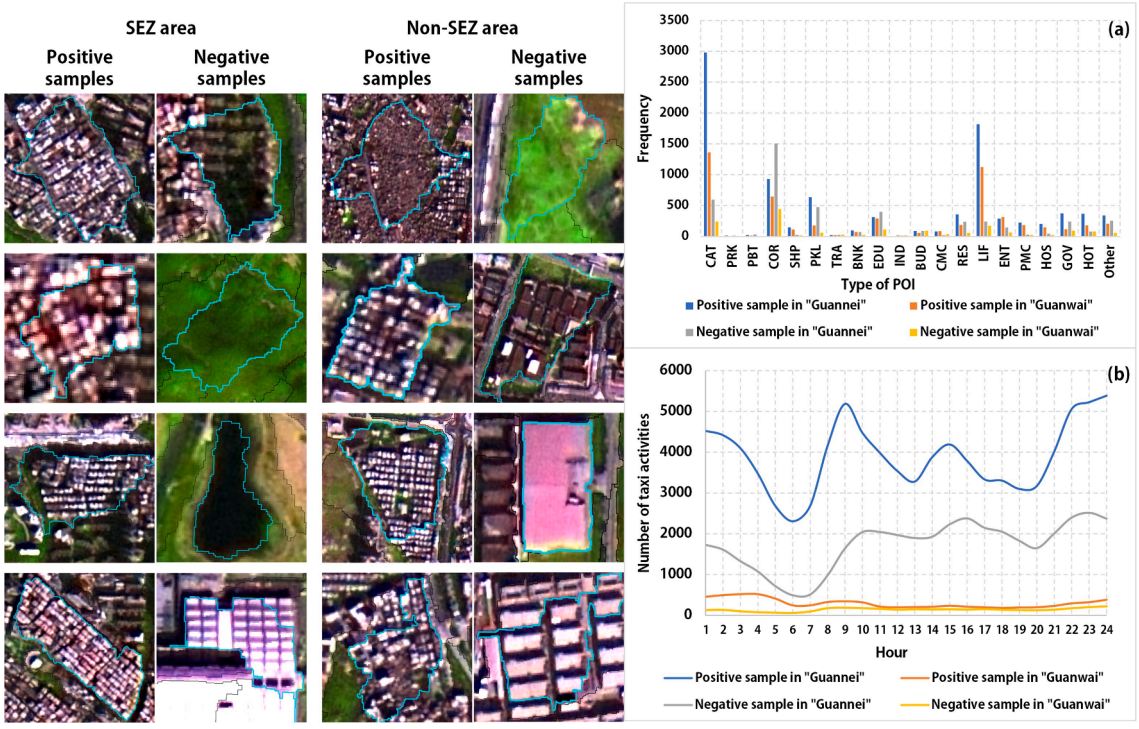


Fig. 6. Example remote sensing images of positive and negative samples and their social sensing data. (a) POIs data, (b) taxi OD data. Types of POIs include cafeteria (CAT), park (PRK), public toilet (PBT), corporation (COR), shopping (SHP), parking lot (PKL), transportation (TRA), bank (BNK), education (EDU), industry (IND), building (BUD), commercial area (CMC), residential area (RES), life-related services (LIF), entertainment (ENT), pharmacies (PMC), hospital (HOS), government (GOV), hotel (HOT).

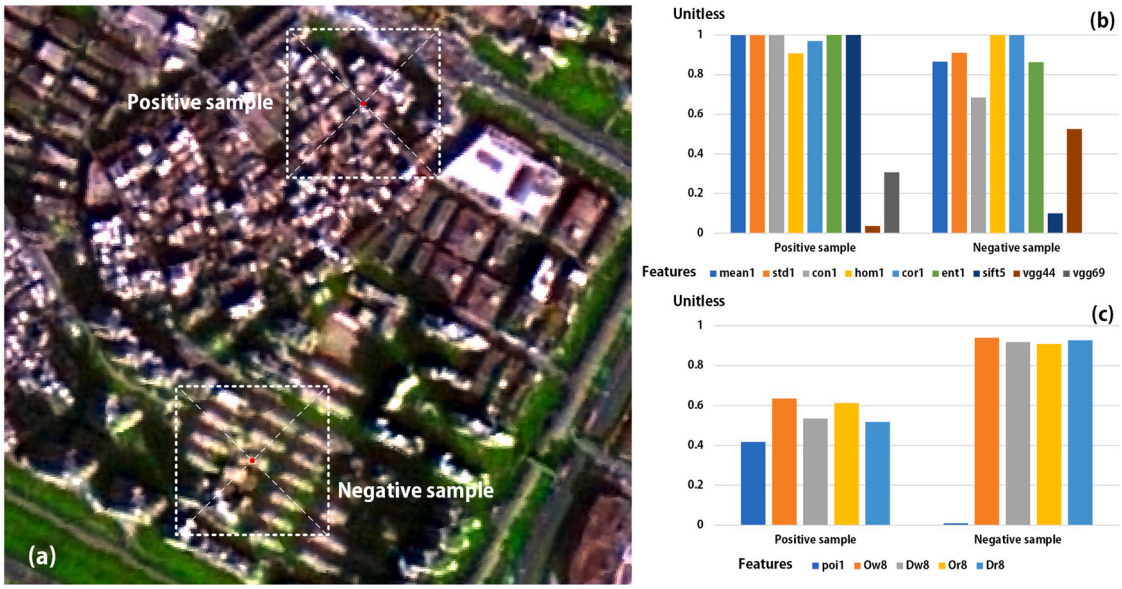


Fig. 7. Example pixels of UVs and non-UVs and some corresponding fine-scale features. (a) The locations of the positive sample and negative sample, (b) some corresponding remote sensing features and (c) social sensing features.

For coarse-scale recognition, the regional context information needs to be transformed into coarse-scale features to determine whether one area contains UVs. First, those features containing a large number of dimensions, such as SIFT features (128), taxi travel features (96) and deep convolution features (512), need to be dimensionally reduced. Next, a Random Forest (RF) classifier is applied to classify the objects. RF classifiers have been widely used in land use and urban planning studies (ÇÖmert et al., 2019). RF algorithm increases the weights of features with high contribution, thereby excelling in dealing with high dimensional data and avoiding overfitting (Fernández-Delgado et al., 2014). Thus, we adopt the RF classifier to recognize potential UV areas. In addition, UVs are located within urban extents. Thus, villages located outside the UE mask need to be eliminated.

3.3. Fine-scale UV area recognition

This step aims to identify fine-scale UV areas from bottom to up using fine-scale local information. Specifically, fine-scale recognition consists of two steps. (1) Pixel-based fine-scale local features are extracted from multi-source geospatial data, and a classifier is further applied to identify UV pixels. (2) The detected UV areas are refined via mathematical morphological methods. According to simple sample statistics, the local features of the UV pixels are distinguishable from those of the nearby non-UV pixels. For example, Fig. 7 shows that some of the RS features and SS features of positive pixels are different from negative pixels. With the RF classifier, we can further explore the differences between the two and separate them.

For UV pixels recognition, we use fine-scale local information to recognize UV pixels through pixel-based classification. First, for each pixel, we use a moving window that centered on the pixel to extract the representative features of the pixel from multi-source geospatial data. The features can describe the characteristics of the pixel as well as local spatial contextual information of the pixel. Next, pixel-based classification is conducted to classify urban pixels as UVs or non-UVs based on the local information. Similar to the step of coarse-scale UV area recognition, the RF classifier is applied to categorize pixels within potential UV areas.

Simple pixel-based classification will easily lead to noticeable internal noise and uneven edges. To alleviate this issue and obtain regular-shaped UV areas, the pixel results should be further refined. We apply mathematical morphological methods, which are mainly used to extract image components from an image, which can capture the most discriminative shape features of the object (Bhateja et al., 2019). Specifically, opening and closing operations are applied. Their equations are presented as follows:

$$I \circ SE = (I \ominus SE) \oplus SE \tag{1}$$

$$I \cdot SE = (I \oplus SE) \ominus SE \tag{2}$$

where I and SE indicate a binary image and a morphological structural element, respectively. The symbols of $\circ, \cdot, \ominus$ and $\oplus$ donate opening, closing, dilation, and erosion operation, respectively. Finally, regular fine-scale UV areas can be obtained,which can be used as a reference for urban planning studies.

4. Experiments and results

Four experimental groups are set up: Set (1) of experiments shows the overall result for each step; Set 2 presents the reasonableness of the approach by comparing with baseline methods; Set 3 shows the effect of using zonal training samples or not on the UVs recognition; and Set 4 examine the optimal parameters of the model. The specific results are presented in the following subsections.

4.1. Experiment setup

To obtain reasonable and fair results, the training and testing datasets are sampled following the procedures proposed by (Huang et al., 2015). The representative UV samples are manually selected for training, while testing samples are randomly selected from the whole UV sample set (excluding the training samples). Specifically, at the coarse scale, positive samples are randomly selected from the regions where urban villages occupy more than 90% of the area, while the negative samples are the regions without any urban village area. At the fine scale, positive samples are randomly selected from the pixels within urban villages, while negative samples are randomly chosen from the pixels outside urban villages. The configuration is shown in Table 1. As can be seen, at the fine scale, the negative samples contain two sets, i.e., A and B. Set A consists of pixels within the potential UV areas, while Set B is composed of pixels outside the potential UV areas. The ratio of sample numbers is about 1:1. The performance of the fine-scale step is evaluated

Table 1
Configuration of sample numbers.

	Sample	Positive sample	Negative sample
Coarse scale	Training	202	197
	Testing	230	600
Fine scale	Training	8753	7808
	Testing	104830	102207 (A: 45586, B: 56621)

only using the Set A, while the overall performance of the HR-RSF-UV approach is evaluated using both test sets.

In addition, we select three 11 km × 11 km typical regions covering the core urban area of Shenzhen for a case study (Fig. 8). Overall accuracy (OA), Kappa coefficient, omission error (OE), commission error (CE) are adopted to evaluate the results. Given *TP, TN, FP, FN* denotes the number of true positive, true negative, false positive, false negative samples of the confusion matrix. And *n, N* denotes the number of classes, the total number of all the samples, respectively. Then the evaluation metrics can be formulated as follows:

$$\text{Overall accuracy}: \quad p\_0 = TP/N \tag{3}$$

$$\text{Kappa coefficient}: \quad Kappa = \frac{p\_0 - p\_e}{1 - p\_e} \tag{4}$$

where $p\_e = \frac{(TP+TN)*(TP+FN)+(FN+TN)*(FP+TN)}{N^2}$.

$$\text{Omission error}: \quad OE = 1 - \frac{TP}{(TP+FP)} \tag{5}$$

$$\text{Commission error}: \quad CE = 1 - \frac{TP}{(TP+FN)} \tag{6}$$

4.2. Overall results

The overall results of applying the proposed approach are shown in Table 2. We can see that the overall performance is excellent, with the Kappa and OA reaching 0.920 and 96.23%, respectively. The results

Table 2
Recognition results of the HR-RSF-UV approach.

	Kappa	OA	Omission error	Commission error
Coarse-scale step	0.929	98.04%	0.030	0.086
Fine-scale step	0.892	94.64%	0.071	0.028
Overall	0.920	96.23%	0.071	0.028

show that the overall classification performance is excellent since the Kappa of 0.8 is normally considered as well enough (Kim et al., 2020).

In the coarse-scale step, the Kappa and OA reach 0.929 and 98.04% respectively, which demonstrates the effectiveness of the classification method. It should be noted that the OE is controlled within a very low value (0.030), which means that very limited objects that contain UV areas are omitted in this step.

At the fine-scale step, the Kappa and OA also achieve a performance of 0.892 and 94.64%, respectively. The performance is impressive because of the huge number of pixels. And the Kappa and OA reach an impressive performance of 0.920 and 96.23%. The OE and CE indicate that the accuracy is improved, because the coarse-scale step excludes most of non-UV pixels outside the potential UV areas.

A map of the predicted UV areas in Shenzhen using our proposed HR-RSF-UV approach is shown in Fig. 8. Three typical areas, i.e., (a) Bao'an-Guangming, (b) Futian-Luohu, and (c) Longgang-Pingshan areas, are selected for case studies. The yellow areas, the orange parts in the typical areas, the areas with blue borders are the potential UV areas, the predicted UV areas, the ground truth, respectively. The proposed approach can achieve high performance in all three typical areas, with OA all higher than 95% and Kappa higher than 0.7. Specifically, the approach obtains the best performance in terms of Kappa (more than 0.9) in the Area (b). Because the area is mainly distributed in the SEZ area, the core economic area in Shenzhen. While Area (a) and Area (c) are located in the non-SEZ area which is covered by many mixed functional areas. UV buildings in these areas are much similar to surrounding factory buildings, which increases the difficulty in distinguishing them.

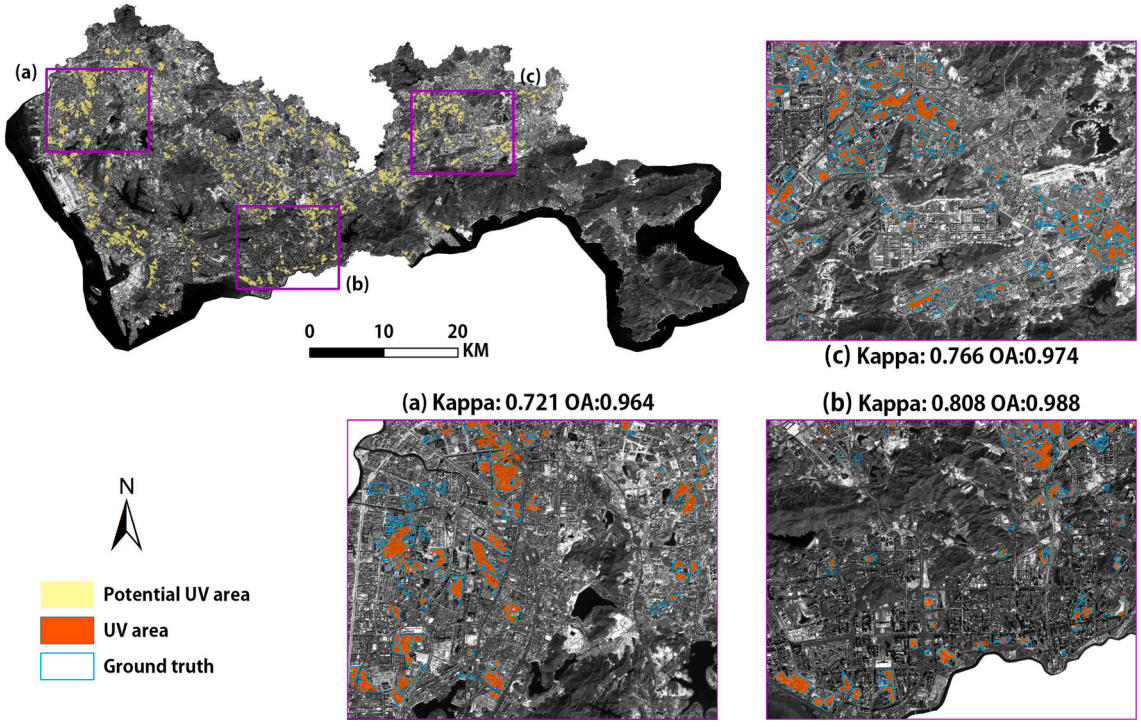


Fig. 8. Predicted UV area map using the HR-RSF-UV approach. (a) Bao'an-Guangming area, (b) Futian-Luohu area, (c) Longgang-Pingshan area.

4.3. Comparison with baseline methods

To analyze the performance of HR-RSF-UV approach at different steps, three sets of baseline experiments following the HR framework were set up. The first baseline experiment, SVM-MS, uses multi-source (MS) spatial features and the support vector machine (SVM) method, which is one of the most effective general machine learning classification methods. The second baseline experiment (RF-RS) applied common RS features and the RF classifier, which represents traditional RS classification methods. The third baseline method (U-Net) is a widely used convolutional neural network model for semantic segmentation. To make a fair comparison, 10400 images with size of 128 × 128 pixels are used and data augmentation (e.g., flip, rotation) is also applied to finetune the U-Net.

Table 3 shows that in the coarse-scale recognition step, the Kappa of the HR-RSF-UV approach increases by 0.097 and 0.112 compared to the SVM-MS model and RF-RS model, respectively. The OA improves by 2.80% for both. And the best Kappa of different previous studies vary widely, from 0.620 to 0.968 (Li et al., 2017; Wang et al., 2019; Wurm et al., 2019), due to different study areas and samples. However, to some degree, the accuracies can be compared by considering the study scale and testing sample size. For example, Li et al. (2017), which had the highest Kappa of 0.962, used a scene size of 144 m and about 800 test samples. While the study area uses 830 test samples, with an average edge length of the geo-object of about 180 m. The accuracy of the proposed method falls in the upper range, presenting good performance in coarse-scale recognition.

In the fine-scale recognition, the proposed approach is also more accurate than the SVM-MS model and RF-RS model, with Kappa improving by 0.049 and 0.106, and OA improving by 2.50% and 5.33%. Previous studies showed that the best Kappa of previous studies in the same field for pixel-based and superpixel-based classification vary from 0.600 to 0.910 (Kuffer et al., 2016b; Verma et al., 2019; Wurm et al., 2017b). And the accuracy of the approach is high in the accuracy range of the previous studies in the same field.

The overall performance of HR-RSF-UV approach outperforms those of the baseline methods. As shown in Table 3, the fine-tuned U-Net achieves 0.884 and 94.44%, which exceeds the SVM-MS model and RF-RS model. In contrast, the Kappa and OA of the HR-RSF-UV model are slightly higher than the results of U-Net. Fig. 9 shows sample performances of the above methods for different areas of test data. First, for the UVs (Fig. 9 (a) (b)), the two baseline methods, SVM-MS and RF-RS, show more misclassification errors, while U-Net shows more over-classification errors. And the results of UVs recognized via the HR-RSF-UV approach are closest to the labels. Second, for non-built-up area (e.g., roads, water, grass), all models correctly predicted the results (Fig. 9 (e) (f)). Derived from the HR framework, SVM-MS, RF-FS and HR-RSF-UV approach effectively exclude non-built-up areas, with the same performance as U-Net. Last, faced with built-up areas with features similar to those of UVs, different models behave differently. SVM-MS and RF-RS exhibit misclassification at some buildings in the common residential area (Fig. 9 (c)). While U-Net misclassifies a portion of industrial areas as UVs (Fig. 9 (d)). And the HR-RSF-UV approach has the best performance for built-up areas.

Table 3
Comparison of accuracies produced by different methods. SVM-MS: a baseline method using SVM and multi-source data. RF-RS: a baseline method using RF and common RS features (spectral, textural, and structural features).

	Method	Kappa	OA
Coarse-scale step	SVM-MS	0.832	95.24%
	RF-RS	0.817	95.24%
	HR-RSF-UV	0.929	98.04%
Fine-scale step	SVM-MS	0.843	92.14%
	RF-RS	0.786	89.31%
	HR-RSF-UV	0.892	94.64%
Overall	SVM-MS	0.844	93.29%
	RF-RS	0.797	92.11%
	U-Net	0.884	94.44%
	HR-RSF-UV	0.920	96.23%

4.4. Effect of training sample on the UVs recognition

This section discusses the effect of setting up regional training samples. Fig. 3 presents the differences between the samples distributed in the SEZ area and those in the non-SEZ area. In the SEZ area, UV areas show a more regular and neater pattern. The negative sample mainly consists of vegetation, water, high-rise buildings and ordinary residential areas, which are obviously different from the positive sample. In contrast, the positive samples in the non-SEZ area are more diverse in their morphology. And the negative samples contain many industries with a smaller difference with the positive samples.

To verify the effect of differences in training samples on the results, we set up 2 groups using different training samples. Table 4 shows that in both steps, the results using the region training samples are better than the results using the mixed samples. Whereas in the coarse-scale recognition, Kappa and OA improve by 0.028 and 0.07%, respectively. While in fine-scale recognition, Kappa and OA improve by 0.018 and 0.93%, respectively. A limited difference are shown between whether or not to differentiate between SEZs and non-SEZs.Thus, it suggests that the proposed HR-RSF-UV framework generally works well in different regions.

4.5. Parameter sensitivity analysis

Several parameters may affect the result, such as the window size and mathematical morphology methods. In the fine-scale recognition step, a sliding window needs to be set around the center of the pixel to extract its surrounding features. Its size determines how large a range of local detail information is considered by the classifier. Multiple sets of experiments are set up to analyze the effect of the window size on the results. The size can be freely adjusted for common RS and SS features. Thus, the differences between the results obtained from a size of 100–180 m are analyzed according to the set by previous studies. However, deep convolution features are extracted using VGG-net, and the window size setting is limited. So, the experiment for testing deep convolution features only considers 3 groups, including 64 × 64, 128 × 128, and 256 × 256 pixels.

As shown in Fig. 10 (a), both Kappas and OAs initially improve in accuracy as the window size increases. However, when the size is larger than 150 m, the accuracy starts to decline. Thus, setting the size to 150 m is the optimal choice for common RS and SS features extraction. And Fig. 10 (b) shows that the recognition has the best performance when the size is set to 64 × 64 pixels to extract deep convolution features. So, the window size setting should be a combination of 150 × 150 m for non-deep convolution features and 64 × 64 pixels for deep convolution features.

In addition, the combination and order of the morphology methods affect the effect of removing noise and trimming edges. The morphological methods used include the opening operator and the closing operator. Four sets of experiments are set up to test the effect of different combinations and orders of operations on the accuracy, including Set 1 using open operator, Set 2 using closing operator, Set 3 using the opening operator first followed by the closing operator, and Set 4 using closing operator first followed by opening operator.

Fig. 11 shows that the sets that use the closing operator first all have higher accuracy than the sets that use the opening operator first. Kappa and OA of Set 2 are 0.013 and 0.70% higher than those of Set 1. While Kappa and OA of Set 4 are 0.015 and 0.70% higher than those of Set 3. Because the effect of the closing operator fills in fine gaps, making the target pixels connected into regular surface plots, which is consistent with the general shape of UV areas. The accuracy of Set 4 is higher than

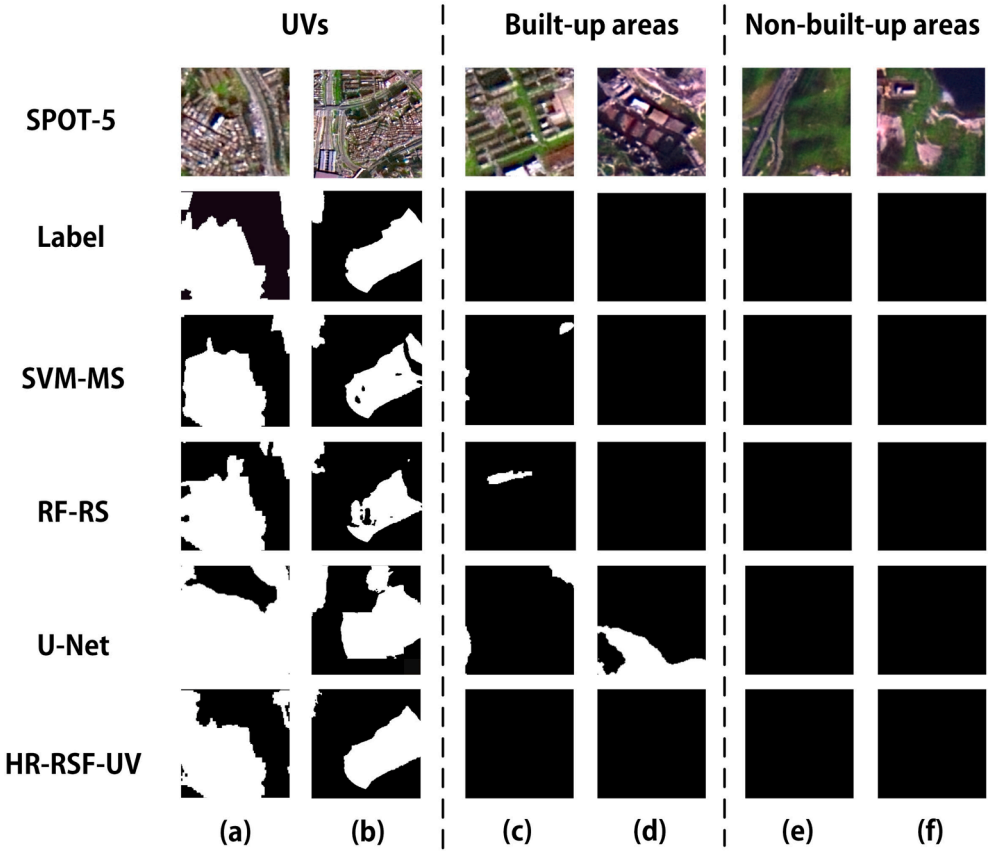


Fig. 9. Comparison of examples of predicted results produced by different deep methods.

Table 4
Comparison of the accuracies of inputting regional training samples and mixed training samples.

	Training Sample	Kappa	OA
Coarse-scale recognition	Regional samples	0.929	98.01%
	Mixed samples	0.901	97.32%
Fine-scale recognition	Regional samples	0.892	94.64%
	Mixed samples	0.874	93.71%

that of Set 2. Kappa and OA improve by 0.005 and 0.20%, respectively. This is because the opening operator removes convex or free target pixels, and the operation of using the closing operator first followed by the opening operator can remove the free noise based on the result of using the closing operator.

5. Discussion

This study proposes the HR-RSF-UV method for recognizing fine-grained UV areas using the HR framework and RS and SS data fusion. It is a new attempt to fully combine the use of different scales and dimensions of information to more comprehensively describe such complex geo-objects like UV. The results prove that the method has a good performance. However, it is still not known whether this is due to the application of the HR framework or RS and SS data fusion. The next two subsections discuss the gaining effects of the two application on the UVs

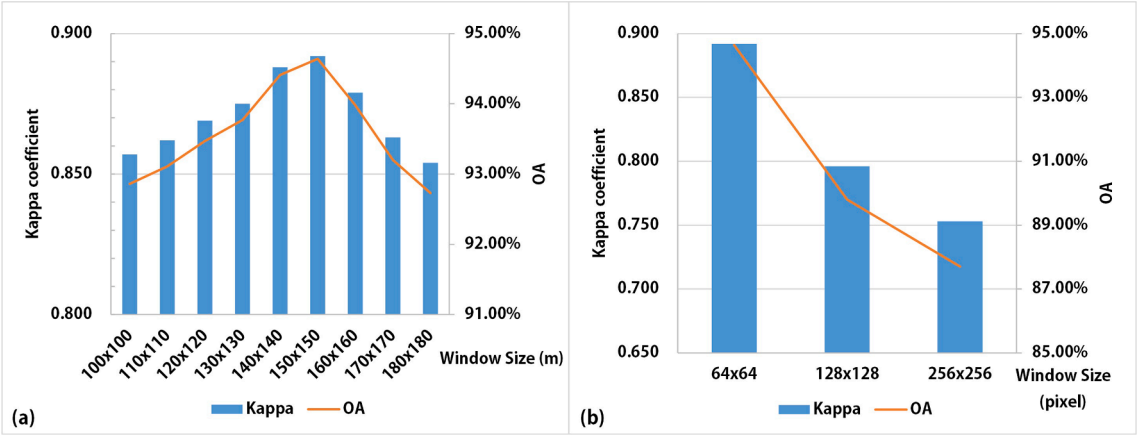


Fig. 10. Comparison of the accuracies produced by different window sizes. (a) Windows for extracting non-deep convolution features. (b) Windows for extracting deep convolution features.

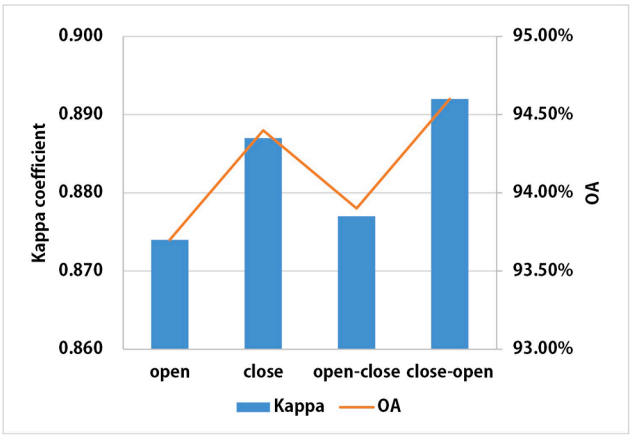


Fig. 11. Comparison of the accuracies produced by different combinations and orders of morphology operators.

recognition through experiments, respectively. And then the pros and cons of the HR-RSF-UV method are discussed.

5.1. Ablation study of proposed hierarchical recognition framework

The HR framework emphasises the importance of hierarchical information for recognizing geo-objects. Previous geo-object classification methods can be divided into two categories, i.e., the bottom-up and the top-down recognition approaches. The bottom-up approach, such as the pixel-based and superpixel-based methods (Zhao et al., 2017), analyzes local information of small areas and classify pixels. This approach is simple but easily fails, especially by using HSR RS images because of the increase of the spectral diversity. While the top-down approach determine whether one area contain UVs by considering regional contexts, i.e., object-oriented and scene understanding methods (Li et al., 2017). Different from the bottom-up approach, it iteratively divides the UVs from their spatial backgrounds. It is with the advantage to locate the object fast but cannot well outline the fine-grained boundaries. Both the bottom-up and the top-down approaches have their own shortcomings. Thus, the HR framework is necessary to be used to integrate both and avoid their shortcomings (Zhang et al., 2018).

A set of experiments is set up under the same input condition to test the effect of applying the HR framework. One experiment uses the proposed method following the HR framework, while the other applies the multi-source geospatial features and the RF algorithm for the one-stage pixel-based classification as the baseline method (RF-MS (One stage)). Table 5 shows that the accuracy using the proposed method is higher than that of the one-stage method. The Kappa improves by 0.162, while OA improves by 8.06%. On one hand, the proposed method reduced the CE by 0.176. This indicates that the one-stage method predicts more wrong cases for negative samples in non-potential UV areas. While the coarse-scale recognition is able to lower the errors in this case very well. On the other hand, the OEs of the two methods are almost identical. It means that the coarse-scale recognition with lower OE helps to control the OE of the final result.

The HR framework can bring considerable improvements in geographic recognition. This is because the HR framework combines a large range of background information and a small range of detailed information. On one hand, it helps to exclude areas that are absolutely impossible to be the target features at the coarse scale. On the other hand, it allows the target areas to be efficiently distinguished from the surrounding features at the fine scale.

Table 5
Comparison of the accuracies using the HR framework or not.

Method	Kappa	OA	Omission error	Commission error
HR-RSF-UV (Two stage)	0.920	96.23%	0.071	0.028
RF-MS (One stage)	0.758	88.17%	0.070	0.204

5.2. Effect of remote and social sensing data on the UVs recognition

This section tests the gaining effect of fusing the RS and SS data on the UV recognition method. Different features from different data sources are grouped and the accuracy of their results are verified. A total of four groups are divided: Group 1 is RS physical features, including common-used spectral, textural, structural features, and deep convolution feature. Group 2 is SS features extracted from SS data, including local semantic features of POIs data and taxi trip activity features of taxi trajectory data. Group 3 combines RS features and SS features from multi-source data. Group 4 is a combination of RS features, SS features, and the UE mask. Group 4 is not performed in fine-scale recognition since the UE mask with a coarse spatial resolution (500 m) cannot be applied to a 2.5 m pixel.

Table 6 presents the effect of inputting multi-source geospatial data into the same method. In the coarse-scale step, the Kappa and OA of the result using SS data only are both lower than those of the group using RS data only. This indicates that the RS physical features of UV areas are more discriminating than the SS features. However, the accuracy of the third set using a combination of RS and SS data is higher than that of the RS group. The Kappa and OA improved by 0.027 and 0.54%, respectively. This indicates that SS is suitable as a supplement to RS to further characterize UV areas. Whereas Group 4 excluded non-UV areas using the UE mask based on Group 3. The results showed an improvement of 0.030 for Kappa and 0.87% for OA. This implies that Group 3 failed to effectively distinguish between UV and common village, due to the high similarity in physical morphology. Thus, the UE mask can be used as an auxiliary perspective in the coarse-scale UV recognition to further improve accuracy.

In the fine-scale recognition step, the set of using SS data only gives the lowest-precision results (Kappa: 0.503, OA: 75.16%). This may be because SS data including POIs and taxi track data behaves sparsely at the fine-scale. The accuracy of the set using RS only, on the other hand, remains very good (Kappa: 0.874, OA: 93.71%). Thus, RS data have good performance at both the coarse scale and the fine scale. And the third set combining RS and SS data has the best performance, with the Kappa and OA improved by 0.018 and 0.93%, respectively. Therefore, SS data at the fine scale can be used for RS data as another perspective to complement and improve the description of UV. In conclusion, multi-source data fusion can effectively exclude confusing target features such as vegetation and common residential neighborhoods, and further characterize the UV.

In summary, both the application of the HR framework and the RS and SS data fusion have a beneficial effect on the accuracy of UVs recognition.

5.3. Pros and cons of HR-RSF-UV

The task of UV recognition has witnessed increasing accuracy in

Table 6
Comparison of the accuracies produced by inputting different combinations of multi-source geospatial data.

	Dataset	Kappa	OA
Coarse-scale recognition	RS	0.872	96.57%
	SS	0.810	95.17%
	RS + SS	0.899	97.13%
	RS + SS + UE mask	0.929	98.04%
Fine-scale recognition	RS	0.874	93.71%
	S	0.503	75.16%
	RS + SS	0.892	94.64%

recent years, especially with the development of deep learning (Mast et al., 2020). Among the state-of-the-art methods, the proposed HR-RSF-UV approach is competitive due to three major reasons. Firstly, the HR-RSF-UV approach is interpretable by using a coarse-to-fine hierarchical recognition structure. It has explicit geographical meaning for each step, which ensures the interpretability of the model as well as the geographic understanding (Zhou et al., 2016). This is essential for real-world decision and policy making, which many black-box models lack (e.g., end-to-end deep learning methods like U-Net). Secondly, the HR-RSF-UV framework is general and transferable (Zhang et al., 2017). It doesn't involve any specific settings of the study area. The components of the HR-RSF-UV framework, e.g., the target, the classifier, and the data used, can all be easily replaced as demanded when applying to other cities. Thus, the HR-RSF-UV model is extensible for diverse conditions and applications. Thirdly, the proposed framework trained by small amount of representative samples can show a good performance. Compared with deep learning models like U-Net, the proposed approach does not require a large number of input training samples due to the use of traditional machine learning classifier. This advantage is significant and make our approach more scalable, especially for the UV application, since the labelled samples of urban villages are hard to obtain.

HR-RSF-UV also has some drawbacks. Firstly, the model maps the fine-grained UVs in the order of the coarse-scale recognition first, and then fine-scale recognition. So the accuracy error in coarse scale can severely affect the whole recognition. More attention should be paid to the usability of the coarse-scale recognition's results in practical applications. Secondly, HR-RSF-UV ignores the cross-correlation between multi-source spatial features, which is important for further data fusion (Zhang et al., 2019). Thus, the HR-RSF-UV model needs further study to address such issues in the future.

6. Conclusion

Urban villages is essential for urban renewal. In the recent years, RS-based UV recognition methods have been widely used in this field. However, due to the problem of low discrimination of RS features showed in UVs, previous methods cannot simultaneously perform excellent recognition accuracy and high spatial resolution in high-density cities. The proposed HR-RSF-UV approach apples the HR framework, taking the advantages of hierarchical information at the coarse and fine scales to obtain fine-grained UV maps. At the same time, it comprehensively characterize UVs by fusing the two perspectives of RS and SS. The case study in Shenzhen demonstrated the excellent performance of HR-RSF-UV. Also, the gaining effects of the HR framework, RS and SS data fusion, the zone-based training strategy on HR-RSF-UV are also demonstrated, providing insights for improving urban village recognition methods. We believe that the urban village map with a spatial resolution of 2.5 m can help provide urban planning researchers with more comprehensive and detailed information and provide a useful reference for future urban renewal decisions.

In the future, this paper will further improve the method for more applications. First, we will improve the method of setting up a regional training set, to further refine the regional differences of urban villages. The UV training set in Shenzhen City is only empirically differentiated into two regions (SEZ area and non-SEZ area). However, the urban village differences can be further refined. Next, we will try to apply this urban village recognition method to detect other types of informal settlements around the world, such as slums in India and Brazil, and compare their accuracies. In this paper, limited by the data from the study area, we only discuss the accuracy of applying this method to UV types of informal settlements in Shenzhen. Future research needs to collect more data to explore the differences in recognition methods for more types of informal settlements.

Declaration of Competing Interest

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

References

Bhateja, V., Nigam, M., Bhadauria, A.S., Arya, A., Zhang, E.Y.D., 2019. Human visual system based optimized mathematical morphology approach for enhancement of brain MR images. J. Ambient Intell. Hum. Comput. 1–9. https://doi.org/10.1007/s12652-019-01386-z.

Blaschke, T., Hay, G.J., Kelly, M., Lang, S., Hofmann, P., Addink, E., Feitosa, R.Q., van der Meer, F., van der Werff, H., van Coillie, F., Tiede, D., 2014. Geographic object-based image analysis – towards a new paradigm. ISPRS J. Photogramm. Remote Sens. 87, 180–191. https://doi.org/10.1016/j.isprsjprs.2013.09.014.

Cao, R., Tu, W., Yang, C., Li, Q., Liu, J., Zhu, J., Zhang, Q., Li, Q., Qiu, G., 2020. Deep learning-based remote and social sensing data fusion for urban region function recognition. ISPRS J. Photogramm. Remote Sens. 163, 82–97. https://doi.org/10.1016/j.isprsjprs.2020.02.014.

ÇÖmert, R., Matci, D.K., Avdan, U., 2019. Object based burned area with random forest algorithm. Int. J. Eng. Geosci. 4, 78–87. https://doi.org/10.26833/ijeg.455595.

Chen, G., Weng, Q., Hay, G.J., He, Y., 2018. Geographic object-based image analysis (GEOBIA): emerging trends and future opportunities. GISci. Remote Sens. 55, 159–182. https://doi.org/10.1080/15481603.2018.1426092.

Chen, Y., Chen, Q., Jing, C., 2019. Multi-resolution segmentation parameters optimization and evaluation for VHR remote sensing image based on meanNSQI and discrepancy measure. J. Spat. Sci. 66, 253–278. https://doi.org/10.1080/14498596.2019.1615011.

Dou, P., Chen, Y., 2017. Dynamic monitoring of land-use/land-cover change and urban expansion in shenzhen using landsat imagery from 1988 to 2015. Int. J. Remote Sens. 38, 5388–5407. https://doi.org/10.1080/01431161.2017.1339926.

Farabet, C., Couprie, C., Najman, L., LeCun, Y., 2013. Learning hierarchical features for scene labeling. IEEE Trans. Pattern Anal. Mach. Intell. 35, 1915–1929. https://doi.org/10.1109/tpami.2012.231.

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems? J. Mach. Learn. Res. 15, 3133–3181.

Friesen, J., Taubenböck, H., Wurm, M., Pelz, P.F., 2018. The similar size of slums. Habitat Int. 73, 79–88. https://doi.org/10.1016/j.habitatint.2018.02.002.

Gallaher, C.M., Kerr, J.M., Njenga, M., Karanja, N.K., WinklerPrins, A.M.G.A., 2013. Urban agriculture, social capital, and food security in the kibera slums of nairobi, kenya. Agric. Hum. Values 30, 389–404. https://doi.org/10.1007/s10460-013-9425-y.

Guan, X., Wei, H., Lu, S., Dai, Q., Su, H., 2018. Assessment on the urbanization strategy in china: Achievements, challenges and reflections. Habitat Int. 71, 97–109. https://doi.org/10.1016/j.habitatint.2017.11.009.

Handzic, K., 2010. Is legalized land tenure necessary in slum upgrading? Learning from Rio's land tenure policies in the Favela Bairro Program. Habitat Int. 34, 11–17. https://doi.org/10.1016/j.habitatint.2009.04.001.

Hao, P., Hooimeijer, P., Sliuzas, R., Geertman, S., 2013. What drives the spatial development of urban villages in china? Urban Stud. 50, 3394–3411. https://doi.org/10.1177/0042098013484534.

Haralick, R.M., Sternberg, S.R., Zhuang, X., 1987. Image analysis using mathematical morphology. IEEE Trans. Pattern Anal. Mach. Intell. PAMI-9, 532–550. https://doi.org/10.1109/tpami.1987.4767941.

Huang, X., Liu, H., Zhang, L., 2015. Spatiotemporal detection and analysis of urban villages in mega city regions of china using high-resolution remotely sensed imagery. IEEE Trans. Geosci. Remote Sens. 53, 3639–3657. https://doi.org/10.1109/TGRS.2014.2380779.

Jia, D., Wei, D., Socher, R., Li, L.J., Kai, L., Li, F.F., 2009. Imagenet: A large-scale hierarchical image database. In: Proc of IEEE Computer Vision & Pattern Recognition, pp. 248–255.

Kim, J.O., Shin, J.Y., Kim, S.R., Shin, K.S., Kim, J., Kim, M.Y., Lee, M.R., Kim, Y., Kim, M., Hong, S.H., Kang, J.H., 2020. Evaluation of two EGFR mutation tests on tumor and plasma from patients with non-small cell lung cancer. Cancers 12, 785. https://doi.org/10.3390/cancers12040785.

Kuffer, M., Pfeffer, K., Sliuzas, R., 2016a. Slums from space—15 years of slum mapping using remote sensing. Remote Sens. 8, 455. https://doi.org/10.3390/rs8060455.

Kuffer, M., Pfeffer, K., Sliuzas, R., Baud, I., 2016b. Extraction of slum areas from VHR imagery using GLCM variance. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 9, 1830–1840. https://doi.org/10.1109/jstars.2016.2538563.

Lai, Y., Jiang, L., Xu, X., 2021. Exploring spatio-temporal patterns of urban village redevelopment: The case of Shenzhen, China. Land 10.

Li, Y., Huang, X., Liu, H., 2017. Unsupervised deep feature learning for urban village detection from high-resolution remote sensing images. Photogramm. Eng. Remote Sens. 83, 567–579. https://doi.org/10.14358/pers.83.8.567.

Liu, Y., Liu, S., Wang, Z., 2015a. Multi-focus image fusion with dense SIFT. Inf. Fusion 23, 139–155. https://doi.org/10.1016/j.inffus.2014.05.004.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., Chi, G., Shi, L., 2015b. Social sensing: A new approach to understanding our socioeconomic environments. Ann. Assoc. Am. Geogr. 105, 512–530. https://doi.org/10.1080/00045608.2015.1018773.

Ma, W., Wu, Y., Cen, F., Wang, G., 2020. MDFN: Multi-scale deep feature learning network for object detection. Lect. Notes Comput. Sci. 100, 107149. https://doi.org/10.1016/j.patcog.2019.107149.

Mast, J., Wei, C., Wurm, M., 2020. Mapping urban villages using fully convolutional neural networks. Remote Sens. Lett. 11, 630–639. https://doi.org/10.1080/2150704x.2020.1746857.

Mboga, N., Georganos, S., Grippa, T., Lennert, M., Vanhuysse, S., Wolff, E., 2019. Fully convolutional networks and geographic object-based image analysis for the classification of VHR imagery. Remote Sens. 11, 597. https://doi.org/10.3390/rs11050597.

Mou, X., Cai, F., Zhang, X., Chen, J., Zhu, R., 2019. Urban function identification based on POI and taxi trajectory data. In: Proceedings 2019 3rd Int. Conf. Big Data Res., pp. 152–156. https://doi.org/10.1145/3372454.3372468.

Peyrin, C., Michel, C.M., Schwartz, S., Thut, G., Seghier, M., Landis, T., Marendaz, C., Vuilleumier, P., 2010. The neural substrates and timing of top–down processes during coarse-to-fine categorization of visual scenes: A combined fMRI and ERP study. J. Cognitive Neurosci. 22, 2768–2780. https://doi.org/10.1162/jocn.2010.21424.

Qassim, H., Verma, A., Feinzimer, D., 2018. Compressed residual-VGG16 CNN model for big data places image recognition. In: 2018 IEEE 8th Annu. Comput. Commun. Workshop Conf., pp. 169–175. https://doi.org/10.1109/ccwc.2018.8301729.

Rahmani, S., Strait, M., Merkurjev, D., Moeller, M., Wittman, T., 2010. An adaptive IHS pan-sharpening method. IEEE Geosci. Remote Sens. Lett. 7, 746–750. https://doi.org/10.1109/lgrs.2010.2046715.

Sharma, K., 2000. Rediscovering Dharavi: Stories from Asia's largest slum.

Taubenböck, H., Kraff, N., Wurm, M., 2018. The morphology of the arrival city - a global categorization based on literature surveys and remotely sensed data. Appl. Geography 92, 150–167. https://doi.org/10.1016/j.apgeog.2018.02.002.

Tu, W., Zhang, Y., Li, Q., Mai, K., Cao, J., 2020. Scale effect on fusing remote sensing and human sensing to portray urban functions. IEEE Geosci. Remote Sens. Lett. 18, 38–42. https://doi.org/10.1109/lgrs.2020.2965247.

Verma, D., Jana, A., Ramamritham, K., 2019. Transfer learning approach to map urban slums using high and medium resolution satellite imagery. Habitat Int. 88, 101981. https://doi.org/10.1016/j.habitatint.2019.04.008.

Wang, Y., Qi, Q., Liu, Y., Jiang, L., Wang, J., 2019. Unsupervised segmentation parameter selection using the local spatial statistics for remote sensing image segmentation. Int. J. Appl. Earth Obs. Geoinf. 81, 98–109. https://doi.org/10.1016/j.jag.2019.05.004.

Wang, Y., Wang, Y., Wu, J., 2009. Urbanization and informal development in china: Urban villages in shenzhen. Int. J. Urban Regional. 33, 957–973. https://doi.org/10.1111/j.1468-2427.2009.00891.x.

Wurm, M., Stark, T., Zhu, X.X., Weigand, M., Taubenböck, H., 2019. Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. ISPRS J. Photogramm. Remote Sens. 150, 59–69. https://doi.org/10.1016/j.isprsjprs.2019.02.006.

Wurm, M., Taubenböck, H., Weigand, M., Schmitt, A., 2017a. Slum mapping in polarimetric SAR data using spatial features. Remote Sens. Environ. 194, 190–204. https://doi.org/10.1016/j.rse.2017.03.030.

Wurm, M., Weigand, M., Schmitt, A., Geiss, C., Taubenbock, H., 2017b. Exploitation of textural and morphological image features in sentinel-2a data for slum mapping. In: 2017 Jt. Urban Remote Sens. Event, pp. 1–4. https://doi.org/10.1109/jurse.2017.7924586.

Yao, Y., Chen, D., Chen, L., Wang, H., Guan, Q., 2018. A time series of urban extent in china using DSMP/OLS nighttime light data. PloS One 13, e0198189. https://doi.org/10.1371/journal.pone.0198189.

Zhang, X., Du, S., 2015. A linear dirichlet mixture model for decomposing scenes: Application to analyzing urban functional zonings. Remote Sens. Environ. 169, 37–49. https://doi.org/10.1016/j.rse.2015.07.017.

Zhang, X., Du, S., Wang, Q., 2017. Hierarchical semantic cognition for urban functional zones with VHR satellite images and POI data. ISPRS J. Photogramm. Remote Sens. 132, 170–184. https://doi.org/10.1016/j.isprsjprs.2017.09.007.

Zhang, X., Du, S., Wang, Q., 2018. Integrating bottom-up classification and top-down feedback for improving urban land-cover and functional-zone mapping. Remote Sens. Environ. 212, 231–248. https://doi.org/10.1016/j.rse.2018.05.006.

Zhang, Y., Li, Q., Tu, W., Mai, K., Yao, Y., Chen, Y., 2019. Functional urban land use recognition integrating multi-source geospatial data and cross-correlations. Comput. Environ. Urban Syst. 78, 101374. https://doi.org/10.1016/j.compenvurbsys.2019.101374.

Zhao, W., Jiao, L., Ma, W., Zhao, J., Zhao, J., Liu, H., Cao, X., Yang, S., 2017. Superpixel-based multiple local CNN for panchromatic and multispectral image classification. IEEE Trans. Geosci. Remote Sens. 55, 4141–4156. https://doi.org/10.1109/tgrs.2017.2689018.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. In: 2016 IEEE Conf. Comput. Vis. Pattern Recognit., pp. 2921–2929. https://doi.org/10.1109/cvpr.2016.319.

Zhu, Z., Zhou, Y., Seto, K.C., Stokes, E.C., Deng, C., Pickett, S.T., Taubenböck, H., 2019. Understanding an urbanizing planet: Strategic directions for remote sensing. Remote Sens. Environ. 228, 164–182. https://doi.org/10.1016/j.rse.2019.04.020.