WH-MAVS: A Novel Dataset and Deep Learning Benchmark for Multiple Land Use and Land Cover Applications

Jingwen Yuan, *Student Member, IEEE*, Lixiang Ru, *Student Member, IEEE*, Shugen Wang, *Member, IEEE*, and Chen Wu, *Member, IEEE*

***Abstract*—Over the past decade, many excellent data sharing efforts have enriched the remote sensing scene classification (SC) methods. These datasets have achieved great success in complex high-level semantic information interpretation. However, most existing datasets are collected from standard and ungeoreferenced image patches for algorithm training and evaluation. These datasets do not fit for practical applications and cannot be directly applied in further geographical study. Accordingly, we provide a large range high-resolution SC dataset with multiple time phases, called "Wuhan Multiapplication VHR Scene classification dataset (WH-MAVS)." It facilitates the study of SC and scene change detection (SCD) algorithms. Moreover, it can also be directly employed to perform a variety of real-life land use application tasks. To the best of our knowledge, this is the first free, publicly available, georeferenced, and annotated dataset to cover almost an entire megacity. The WH-MAVS was collected and annotated from Google Earth imagery with the same spatial resolution and uniform nonoverlapping patch size, covering the central area of Wuhan, China. The total number of scene samples is 47 137, which belong to 14 classes with 23 567 labeled patches for each time phase in 2014 and 2016, respectively. The geographic coordinates of all samples in both time phases exhibit one-to-one correspondence with 23 202 unchanged image patches of scene categories and 365 changed ones. The distribution of the number of samples in each class is highly imbalanced; moreover, there are large intraclass differences and indistinguishable interclass variances. These characteristics are closer to the real land use/land cover application tasks and introduce further challenges to the related algorithm research. In addition, we conducted benchmark experiments on SC and SCD based on the WH-MAVS dataset with widely used deep learning models. DenseNet169 was found to achieve the best performance. The overall accuracies are 91.07% and 92.09%, respectively, in the 2014 and 2016 validation sets of WH-MAVS. Furthermore, SCD obtained by DenseNet169 has a binary change detection accuracy of 89.56% and a multiple (from–to) change detection accuracy of 86.70%. Over and above the research value of the algorithm, it is also proven to have practical applications in fields such as urban planning, landscape pattern analysis, and urban dynamic monitoring and analysis.**

***Index Terms*—Benchmark dataset, deep learning, land use/land cover (LULC), multiapplication, multitemporal, scene change detection (SCD), scene classification (SC).**

Manuscript received November 24, 2021; revised December 20, 2021; accepted December 30, 2021. Date of publication January 13, 2022; date of current version February 10, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant T2122014, Grant 61971317, and Grant 41801285, in part by the Natural Science Foundation of Hubei Province under Grant 2020CFB594, and in part by the Science and Technology Major Project of Hubei Province (Next-Generation AI Technologies) under Grant 2019AEA170. *(Corresponding authors: Shugen Wang; Chen Wu.)*

Jingwen Yuan and Shugen Wang are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: jingwenyuan@whu.edu.cn; wangsg@whu.edu.cn).

Lixiang Ru is with the National Engineering Research Center for Multimedia Software, School of Computer Science, Institute of Artificial Intelligence, Wuhan University, Wuhan 430072, China (e-mail: rulixiang@whu.edu.cn).

Chen Wu is with the Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan 430072, China, and also with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430072, China (e-mail: chen.wu@whu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3142898

I. Introduction

In RECENT years, given the dramatic development of remote sensing (RS) satellites, image spatial resolution has been increased to the submeter level, enabling RS images to obtain detailed textures and plentiful structural information [1]–[6]. Very high-resolution (VHR) RS images are in tremendous demand in land use/land cover (LULC)-related applications such as urban planning, urban dynamic monitoring, and ecological assessment. This is due to their high resolution, wide coverage, and large number [7]–[10]. Scene classification (SC) based on the landscapes and their spatial pattern encoded in VHR RS image scenes has attracted increasing attention in the field of RS research. In addition, urban land use change detection is also paying more attention to scene changes. Scene change detection (SCD) is the process of identifying changes and differences between multitemporal image scenes. Although pixel-level and object-level change detection can identify changes in features, these changes do not always change the category of the scene [11], [12]. Therefore, SCD has significant potential for application at the semantic level of urban development and land management.

Owing to its powerful feature learning capability and model expansion ability, deep learning has come to play a broad and important role [13], [14] in object detection [15]–[20], SC [11], [21]–[24], semantic segmentation [25]–[30], change detection [31], [32], image denoising [33], and video tracking [34]–[36] in the context of computer vision and RS images [37]–[46]. Compared with the traditional methods, deep learning is better able to learn contextual information during interpretation and has thus been widely utilized in remote sensing scene classification (RSSC) and SCD studies. However, deep learning requires numerous high-quality data samples for algorithm optimization.

Thus, large numbers of RSSC datasets are essential for the implementation and development of both SC and SCD.

SC datasets have grown from just a few thousand samples (for the early UC Merced dataset [47]) to tens of thousands [48]–[58] or even millions of samples [59]. The number of categories has also grown from a mere handful [9], [13] to 40–50 [51], [54], [55], [58]–[60], making the data richer and more diverse. Moreover, the sources of data collection have become various; not only RGB data captured from Google Earth (GE) images [47], [49]–[54], [56], [59]–[63] but also multispectral data obtained from Sentinel-2 images [57], [58]. The annotation process has also evolved from manual annotation in the beginning to crowdsourced assistance [54], [56] and interactive annotation [59]. While the resolution of these datasets is mainly at the meter level and submeter level, there are also some 10-m-level datasets, such as EuroSAT [57] and BigEarthNet [58].

Benefiting from the currently available SC datasets, high-resolution SC algorithms have evolved from only being able to classify low-level features (such as grass and buildings) to understanding high-level semantic information, such as commercial areas and industrial zones. This is known as breaking the "semantic gap" between the low-level and high-level semantic information.

While a large number of scene-based classification datasets have been proposed, utilization of these existing datasets is still seriously limited when it comes to solving practical production problems such as urban landscape patterns and urban dynamic monitoring. The main issues related to urban land use applications are as follows.

A. Lack of More Practical RS Scene Datasets

Samples in each category are labeled to be as typical as possible, and typical samples are collected in many different cities. In practical LULC applications, however, the complexity of the LULC features can be much higher, and some scene images may not be as typical as the training samples in existing datasets. While numerous studies have used transfer learning or other methods to map urban land use through existing RS scene imagery datasets, the accuracy of high-level semantic mapping using these methods is unsatisfactory [49]. To obtain LULC maps with high confidence in practical applications, it is instead necessary to build a practical scene dataset containing complex and easily confused scene samples in real applications.

B. Lack of Georeferenced RS Scene Datasets for Multiple Applications

Most current datasets are built for algorithm study, meaning that they are extracted from various regions and are not georeferenced. When mapping LULC distributions in a megacity, points of interest (POI), OpenStreetMap (OSM), and different kinds of auxiliary data are also beneficial and additionally need to be linked with the RS image scene [54], [56]. Therefore, to explore the potential of RSSC images, a georeferenced scene dataset is necessary. This dataset is also available for multiple applications, focusing on city structure, shanty towns, real estate valuation, and so on.

C. Lack of Multitemporal Correlated Datasets

SCD is also useful for monitoring the changes in LULC from the semantic level, which indicates the development and expansion of urban areas. Recent RSSC datasets focus primarily on land use mapping at a single point in time; there is no wide-ranged multitemporal RS scene dataset designed specifically for studying SCD, SC with temporal correlation, or updating scene maps.

These problems mentioned above, indicating that it is difficult to directly apply existing datasets from the algorithmic level of RS scene interpretation to the level of real LULC-related applications, is known as the "application gap" of RS scene interpretation. In this article, we newly propose the **W**u**h**an **M**ulti**a**pplication **V**HR **S**cene classification dataset (WH-MAVS). It is a novel large range multiapplication SC dataset annotated around the central region of Wuhan, a megacity in China. There are 14 classes based on China's urban land use planning criteria, which involve diverse factual urban applications.

The dataset is a good opportunity for subsequent application-oriented algorithm research in real land use scenario. Moreover, it also provides the potential for national or global high-level urban land fine-grained mapping research, along with the study of urban land change and expansion, slum detection, social vulnerability assessment, housing value estimation, and so on.

Overall, the present work makes the following major contributions.

1) We construct the first novel multiapplication large range VHR SC dataset labeled, with 14 categories based on high-level semantics for urban functions, and accordingly bridge the application gap between the algorithms and real-life applications.
2) WH-MAVS is a unique SC dataset collected in the urban center of a megacity and marked with geographical coordinates. Therefore, our dataset is more realistic, more complex, and closer to the application requirements compared to the existing datasets.
3) Our proposed SC dataset with change information is capable of being utilized not only for SC algorithm studies but also for SCD algorithm studies. To the best of our knowledge, it is also the only ultrawide range urban SCD dataset suggesting that it may come to play a key role in promoting the study of urban SCD.
4) We evaluate existing SC methods using our new dataset to provide a baseline for future SC and SCD researches. WH-MAVS is also used to achieve application in the fields of urban planning, land mapping and assistance, landscape pattern analysis, and urban dynamic monitoring. The source code and our dataset will be made open source for future academic research.

II. Related Work

In this section, we review the existing open source urban LULC high-resolution RSSC datasets, along with the existing SCD datasets.

TABLE I
EXISTING OPEN SOURCE URBAN LULC RSSC DATASETS

Dataset	Year	Cla	CC	LM	Image number in each class	Total images	Image size (pixel)	Bands	GL	MA	Resolution (m)	Data Region	Download source
UC-Merced [47]	2010	21	N	M	100	2,100	256×256	RGB	N	N	0.3	U.S.regions	USGS
WHU-RS19 [61]	2012	19	N	M	50 to 61	1,013	600×600	RGB	N	N	Up to 0.5	Worldwide	GE
RSSCN7 [62]	2015	7	N	M	400	2,800	400×400	RGB	N	N	-	-	GE
SAT4/6 [48]	2015	4/6	N	M	125,000/ 675,000	500,000/ 405,000	28×28	RGB, NIR	N	N	1 to 6	U.S.regions	NAIP
RSC11 [49]	2016	11	N	M	~100	1,232	512×512	RGB	N	N	~0.2	U.S.regions	GE
SIRI-WHU [50]	2016	12	N	M	200	2,400	200×200	RGB	N	N	2	Urban areas in China	GE
NWPU-RESISC45 [51]	2016	45	N	M	700	31,500	256×256	RGB	N	N	0.2 to 30	Worldwide	GE
AID [52]	2017	30	N	M	220 to 420	10,000	600×600	RGB	N	N	0.5 to 8	Worldwide	GE
PatternNet [53]	2018	38	N	M	800	30,400	256×256	RGB	N	N	0.062 to 4.693	U.S.regions	GE
OPTIMAL-31 [63]	2019	31	N	M	60	1,860	256×256	RGB	N	N	-	-	GE
MLRSNet [60]	2020	46	N	M	1,500 to 3,000	109,161	256×256	RGB	N	N	0.1 to 10	Worldwide	GE
Million-AID [59]	2021	51	Y	IA	Over 20,000 in average	1,000,000	256×256; 512×512	RGB	Y	N	0.5 to 153	Worldwide	GE
RSI-CB128/ RSI-CB256 [54]	2017	45/ 35	Y	CA	About 800/690	36,000/ 24,000	128×128/ 256×256	RGB	N	N	0.22 to 3	Worldwide	GE & BM
RSD46-WHU [55]	2017	46	N	M	500 to 3,000	117,000	256×256	RGB	N	N	0.5 to 2	-	GE & Tianditu
CLRS [56]	2020	25	Y	CA	600	15,000	256×256	RGB	Y	N	0.26 to 8.85	Worldwide	GE, BM, GM, Tianditu
EuroSAT [57]	2018	10	Y	M	2,000 to 3,000	27,000	64×64	13	Y	N	10	Europe	Sentinel-2
BigEarthNet [58]	2019	43	Y	M	328 to 217,119	590,326	10,20,60	13	Y	N	10,20,60	Europe	Sentinel-2

BM = Bing maps, Cla = Class, CC = Classification criteria, M = Manually annotation, CA = Crowdsource auxiliary annotation, IA = Interactive annotation, LM = Label method, GL = Geographic location, GE = Google Earth, GM = Google Map, MA = Multiapplication, NAIP = National Agriculture Imagery Program, USGS = United States Geological Survey.

A. *Existing Datasets for SC*

As RSSC has attracted considerable and increasing focus, a large number of RSSC datasets have been developed to propel the algorithms forward. These have been summarized in Table I. The earliest RSSC dataset to emerge was UC Merced [47]. Although the data size was small, with only 100 samples per category, it greatly advanced research into RSSC algorithms and is still widely used in SC tasks. Only a decade after its emergence, nearly 20 RSSC datasets had sprung up. The total number of images in these datasets escalated from mere thousands at the beginning [47], [49], [50], [61]–[63] into the tens of thousands [51]–[53], [56], [57], hundreds of thousands [48], [55], [58], [60], or even millions [59], making the diversity of SC samples extremely broad. Likewise, the number of categories has climbed from single digits [48] to over 40 [51], [54], [58]–[60]. This indicates that these datasets are advancing toward increased abundance and diversity, which also leads to the increasing universality of SC algorithms. The resolutions of the existing datasets are at the meter level and submeter level; while these are mainly collected from GE images [49]–[56], [59]–[63], a small portion is also collected from sources such as the United States Geological Survey (USGS) [47], Bing Maps [54], [56], Tianditu [55], [56], and so forth. Moreover, other datasets such as EuroSAT [57] and BigEarthNet [58] utilize free and open 10-m-level medium-resolution satellite imagery (like Sentinel-2) for data collection. The vast majority of these datasets are in three bands [red (R), green (G), and blue (B)], but a rare few are four-band high-resolution data [such as SAT4/6 [48], in R, G, B, and near infrared (NIR)] or well as multispectral data [57], [58]. As a result, the download sources of SC datasets are varied. The classification criteria or the datasets have also evolved from the initial personal perception into a more confirmed standard based on certain LULC classification criteria. Simultaneously, the overwhelming amount of data has made it impractical to rely solely on manual sample selection and annotation. As a result, a series of crowdsourcing-assisted annotation efforts have been implemented, including the use of map search engines, OSM, and POI, such as RSI-CB [54] and CLRS [56]. Furthermore, the interactive annotation has recently evolved in the creation of Million-AID. The emergence of these annotation methods has significantly eased the pressure on manual annotation. In the below sections, we review a few representative existing RSSC datasets, namely UC Merced [47], SIRI-WHU [50], EuroSAT [57], and Million-AID [59].

1) UC Merced: This is one of the earliest datasets proposed for the study of RSSC algorithms. The source RGB images for the dataset were downloaded from the USGS. The dataset was divided into 21 categories, each with 100 manually annotated patches of 256 × 256 pixels in size and a spatial resolution of 0.3 m. This is one of the most classic SC datasets, as well as the first open source and available SC dataset, which has been widely used in SC algorithm research. Nevertheless, it is negatively impacted by the small number of samples in each class and insufficient intraclass variation.

2) SIRI-WHU: This dataset was obtained from 2 m spatial resolution images of major urban areas in China on GE. It comprises 12 land use categories: 1) agriculture; 2) commercial; 3) harbor; 4) idle land; 5) industrial; 6) meadow; 7) overpass; 8) park; 9) pond; 10) residential; 11) river; and 12) water. Each

class consists of 200 RGB color image patches of 200 × 200 pixels of size. This dataset is, to the best of our knowledge, the only publicly available dataset to have been created with only Chinese urban areas as the sample region, featuring the typical characteristics of Chinese cities. However, it still has a weak sample size and lack of sample diversity.

3) EuroSAT: Sentinel-2 satellite imagery is openly and freely accessible, and further allows for services such as continuous ground monitoring for the next 20 to 30 years. Moreover, the dataset categories were established on the basis of the European Urban Atlas 2012. Accordingly, this dataset offers possibilities for land use applications such as assisted mapping and change detection monitoring. The EuroSAT has ten classes of concern with 2000 to 3000 patches per class. All image patches have a dimension of 64 × 64 pixels with a 10 m/pixel resolution. The entire dataset is made up of 27 000 georeferenced image patches spread across Europe. It is worth noting that this is a multispectral dataset with 13 bands. However, despite its potential applications in multispectral images, it still has a problem with fewer types and difficulties in applying it to VHR images.

4) Million-AID: The dataset is archived in a semiautomatic and reproducible framework through a semantic coordinate collection strategy that relies on the GE API, as well as OSM data. It has three category levels established with reference to relevant Chinese land use standards (GB/T 21010-2017)—specifically, there are 8 first-level categories, 28 second-level categories, and 51 third-level categories. Each third-level scene category has over 2000 instances, with an average of over 20 000 per type. The quantity of total image tiles reaches 1 000 000, with a wide range of spatial resolutions (from 0.5 to 153 m/pixel) covering locations all over the world. There are two patch sizes: 1) 256 × 256 pixels and 2) 512 × 512 pixels. The Million-AID dataset possesses the desirable attributes of a real-world scenario benchmark dataset, namely diversity, richness, and scalability. However, this dataset is not yet openly available. In addition, there are some imperfections in the third-level types of Million-AID. For example, only the "swimming pool" class in the third level is included in the second-level "leisure" land class, while other categories that might be of interest (such as parks or playgrounds) are lacking.

To date, a large number of RSSC datasets have emerged. Although the diversity and quantity of data are constantly increasing, they are also largely algorithm oriented; the "application gap" has not been solved by the increase in the data volume, and the current datasets are still not directly applicable to the relevant real-world LULC.

B. Existing Datasets for SCD

The monitoring of urban dynamics has long been a fascinating topic [64], [65]. As a result, an abundance of change detection datasets has been created, including OSCD [66], urban–rural boundary of Wuhan [67], SECOND [68], and Hi-UCD [69]. The overwhelming majority of extant change detection datasets are at the pixel level, but relevant open source SCD datasets are very scarce. To the best of our knowledge, the only existing dataset of this kind is MtS-WH [11], [12]. However, the scene-level feature changes are also of great research significance [11], [12].

1) MtS-WH: The two raw images in this dataset were acquired from RS imagery taken by the IKONOS sensor covering the Hanyang district of Wuhan, China. These images contain four bands (blue, green, red, and NIR) with 1 m/pixel spatial resolution. The acquisition dates of the original images are February 11, 2002 and June 24, 2009. This SCD dataset is divided into two parts: 1) a training set consisting of 190 image pairs and 2) a test set consisting of 1050 image pairs. The size of each image patch is 150 × 150 pixels. It is worthy of note that, there were 413 changed scenario pairs and 637 unchanged scenario pairs in the test set. MtS-WH is annotated in the following eight classes: 1) parking; 2) water; 3) sparse house; 4) dense house; 5) residential; 6) idle; 7) farmland; and 8) industrial. The dataset has facilitated excellent advancement in conducting both scene-level LULC region change detection algorithm research and dynamic analysis [11], [12], [70], [71]. Moreover, it is the only freely and publicly available dataset for SCD. However, it still undeniably suffers from the data study area being too confined and the number of data samples being too small; this severely limits the further development of SCD algorithms and application analysis.

In addition, there is a strong correlation between SC and SCD. However, there are currently inconsistencies between the two kinds of datasets in terms of the category criteria; this makes it inevitable that new datasets must be reconstructed to meet the needs of SC and SCD research, resulting in a massive drain on manpower, labor, and time. Moreover, when the classification criteria are made uniform, the combination of multitemporal and single temporal SC datasets can be employed to enhance the classification accuracy of single temporal data through correlation among multitemporal images.

In light of the above, and considering the current problems with classification and change detection datasets, we are the first to unify the category criteria of SC and SCD datasets and create a large range application-oriented SC dataset—WH-MAVS—that is adequate for both SC and SCD.

III. Dataset Creation

Although there are a large number of existing datasets for related algorithm research, it is still extremely difficult to apply the obtained result directly to actual land use applications. According to our analysis, one of the main reasons for this is that the samples of the publicly available datasets collected across different cities are too fragmented. As a result, the obtained samples are not conducive to verification and analysis without ground references. Hence, to bridge the "application gap," we herein present WH-MAVS, an original multitemporal and application-oriented large range dataset, which is the first to comprise samples collected from GE imagery covering the entire central megacity area. Significantly, although the GE images were preprocessed into RGB color bands from real original optical satellite images, it had been confirmed that no obvious

difference between both data sources exists, even in the pixel-/object-based classification [72]. It is therefore also possible to use GE images to evaluate the scene-level classification.

A. GE Image Acquisition

We collected the Wuhan images obtained from the same area via GE and scaled them to 18 levels (a spatial resolution of 1.2 m) for screen capture. Since the central city area of Wuhan (Hubei Province, China, 2608 km^2) is large range, neither the 2014 nor the 2016 GE data could be obtained at the same time. These were filtered into clearer mosaic images, which we refer to as "large range mosaic images." Both temporal images are stored in the WGS-84 coordinate system. Notably, the temporal data collected from the same sensors may exhibit geographic mismatch or geographical bias.

First, the image-to-image registration was brought into effect for the multitemporal GE images to ensure the spatial alignment of the manually collected tie points and cubic convolution resampling. We then cropped the overlapping region of the multitemporal images into identical images with a spatial size of 47 537 $\times$ 38 100 pixels.

B. Classification Criteria

The classification standard was designed on the basis of two references: 1) "Code for classification of urban land use and planning standards of development land" (GB50137 – 2011),[1] edited by the Ministry of Housing and Urban–Rural Development of the People's Republic of China (MOHURD) and 2) the city planning data (CPD) established in 2014 by the Wuhan Land Resources and Planning Bureau from the School of Urban Design, Wuhan University. We additionally considered that the scene image understanding implemented by the computer should be more in line with human visual understanding. Therefore, when conducting category criteria development, if categories or subcategories are identified that can be distinguished only by land use attributes and are difficult to distinguish visually (such as sanitary and epidemic prevention land and scientific research land), we merge them into the same category. Furthermore, there are categories with both distinctive visual features and obvious analytical significance, which we also separate (for example, separating playgrounds from administration and public services to form a new type). The application-based classification criteria are detailed in Table II and comprise the following 14 categories: 1) administration; 2) commercial region; 3) water; 4) agricultural region; 5) green space; 6) transportation; 7) industrial region; 8) residential region 1; 9) residential region 2; 10) residential region 3; 11) road; 12) parking lot; 13) bare land; and 14) playground.

The administration class consists primarily of institutions and facilities such as administration, culture, education, sports, and health. It should be noted that the playground class is classified as a single category that is not included in the administration class. The commercial region class includes both land for commercial facilities (such as wholesale markets, retail stores, and catering) and land for comprehensive office such as finance, insurance, art, and media. The water class covers rivers, lakes, reservoirs, ponds, trench drains, mudflats, glaciers, and permanent snow, but excludes parkland and water within workplaces. The agricultural region class encompasses arable land, gardens, forest, pasture, and agricultural facility land. The green space class is defined as open space land in the form of parkland, protective green space, and so on, but excludes green space allocated within residential regions and institutions. The transportation class chiefly comprises railways, viaducts, ports, airports, etc. The industrial region class mainly refers to the land used for production workshops, warehouses, and ancillary facilities of industrial and mining enterprises, ranging from huge manufactories to small buildings within an area.

TABLE II
SUMMARY OF SEMANTIC SCENE CLASSES IN WH-MAVS

Type ID	Type Name	2014	2016	Total	Trans
1	Administration	520	521	1041	+1
2	Commercial Region	269	273	542	+4
3	Water	5,832	5,850	11682	+18
4	Agriculture Region	4,811	4,801	9612	-10
5	Greenspace	1,126	1,131	2257	+5
6	Transportation	507	509	1016	+2
7	Industrial Region	1,827	1,840	3667	+13
8	Residential Region 1	268	268	536	0
9	Residential Region 2	2,633	2,662	5295	+29
10	Residential Region 3	1,374	1,365	2739	-9
11	Road	2,312	2,422	4734	+110
12	Parking Lot	120	126	246	+6
13	Bare Land	1,781	1,610	3391	-171
14	Playground	187	189	376	+2

Residential regions can be divided into three categories depending on the predominant dwelling type and the surrounding environment are as follows.

1) The *residential region 1* class mainly refers to villas, detached houses, and quadrangle courtyards with complete facilities, well-landscaped environments, and low-rise housing.
2) The *residential region 2* class primarily includes multiple, medium, and high-rise flats with relatively complete public utilities, transport facilities, and a comparatively integrated layout, representing a better environment compared to *residential region 3*.
3) The *residential region 3* class indicates residential regions with poor environments, incomplete public facilities, and transport facilities that need to be renovated (including dilapidated houses, shantytowns, and temporary housing).

A focus on the residential region 3 class would be appropriate to facilitate the formulation of a suitable renewal policy for old districts.

The road class comprises mostly expressways, main roads, subsidiary roads, and bypass roads (which contains intersections, etc.). The parking plot class focuses on car parks of various sizes and shapes in addition to coach parks. Bare land is also an important element of urban supervision research. The bare land class covers open land, saline land, marshland, sandy land, etc. For its part, the playground class comprises a wide range of open-air stadiums and playgrounds of diverse sizes.

[1][Online]. Available: http://www.suifenhe.gov.cn/upload/2013/9/291005871.pdf

Fig. 1. Image cropping process. The red box indicates the cropping box, the single black arrow indicates the cropping direction, and the red box inside the green rectangular box is the image patches to be omitted.

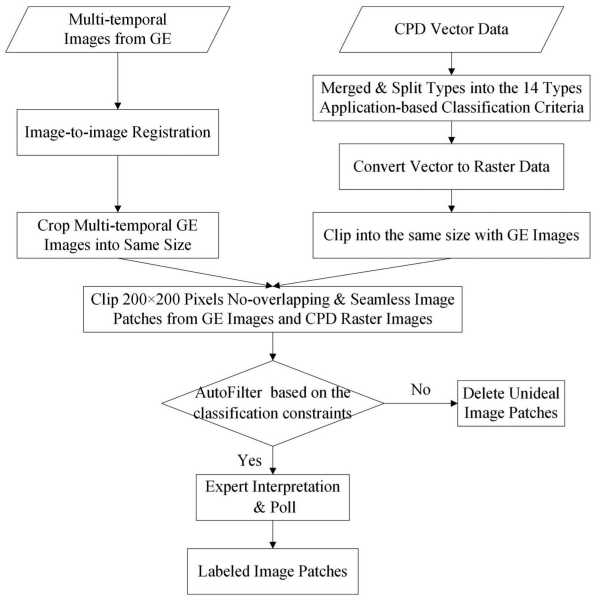

Fig. 2. Flowchart of WH-MAVS dataset creation.

Even though the aforementioned 2014 CPD can assist in creating an application-oriented dataset, it cannot be utilized directly as ground truth. One of the reasons for this is that the CPD is only partially updated in actual production (as opposed to updating the whole city on an annual basis). Therefore, we take the CPD as a ground truth reference only. Since the CPD is vector data, we split and merge the original classification standards published by MOHURD according to our application-based criteria, then convert the vector data to raster to obtain the ground truth reference. Concretely, the CPD is located in the same WGS-84 coordinate system as the GE images of both time phases. Subsequently, the raster data of CPD is clipped to the same extent as the treated GE images (spatial size of 47 537 × 38 100 pixels).

C. Clipping

We created separate nonoverlapping patches for the treated multitemporal GE images as well as for the CPD. The specific cropping process is illustrated in Fig. 1. We start from the top left corner of the image and perform seamless cropping with a step size of 200 pixels and without overlap; the size of each cropping box is 200 × 200 pixels. Note that we regard cropped patches smaller than 200 × 200 pixels at the right and bottom edge of the image as discarded samples and abandon them directly.

D. Inspection Standard

The WH-MAVS annotation process is semiautomatic. The classification constraints of patches can be described as follows.

1) The image patch is considered to fall within the type when more than 60% of its pixels are in the same pixel-based labels supplied by the new ground truth reference.
2) The damaged patches (including dead pixels and those heavily obscured by clouds or mist) are requested to be eliminated.
3) The image patches that do not satisfy the above condition and are too complex to assess are regarded as *0-undefined* type.

E. Labeled Image Patches

First, the patches of the collected 2014 GE images were automatically screened with reference to the above inspection constraints, based on the city planning reference data. Furthermore, several relevant specialists in the field of RS were engaged to manually check the initial classification labels through the original GE images and perform cross-manual interpretation based on a novel poll strategy to filter out mislabeled patches. Next, obsolete samples that do not meet the criteria are deleted. After single phase sample production was offered, the other temporal data (2016) were executed on the basis of the 2014 labels, followed by a poll strategy. Finally, we selected patches of the same area that were labeled in both time spots to create the WH-MAVS dataset. Fig. 2 illustrates the creation of WH-MAVS.

IV. Dataset Description

In this section, we present a detailed elaboration of the WH-MAVS dataset. It is a multiapplication and multitemporal dataset that includes two time phases, namely 2014 and 2016. The WH-MAVS facilitates relevant investigations of not only SC but also SCD. To the best of our knowledge, the WH-MAVS dataset is unique in being a real dataset currently annotated around the central urban areas of a megacity. Both time-phased data in WH-MAVS are based on large range mosaic RGB GE images with the spatial size of 47 537 × 38 100 pixels [see Fig. 3(a) and (b)], at a spatial resolution of 1.2 m, covering 2608 km^2 in the central megacity of Wuhan, Hubei Province, China. The WH-MAVS contains 23 567 labeled patch pairs that exhibit a one-to-one geographical correspondence between

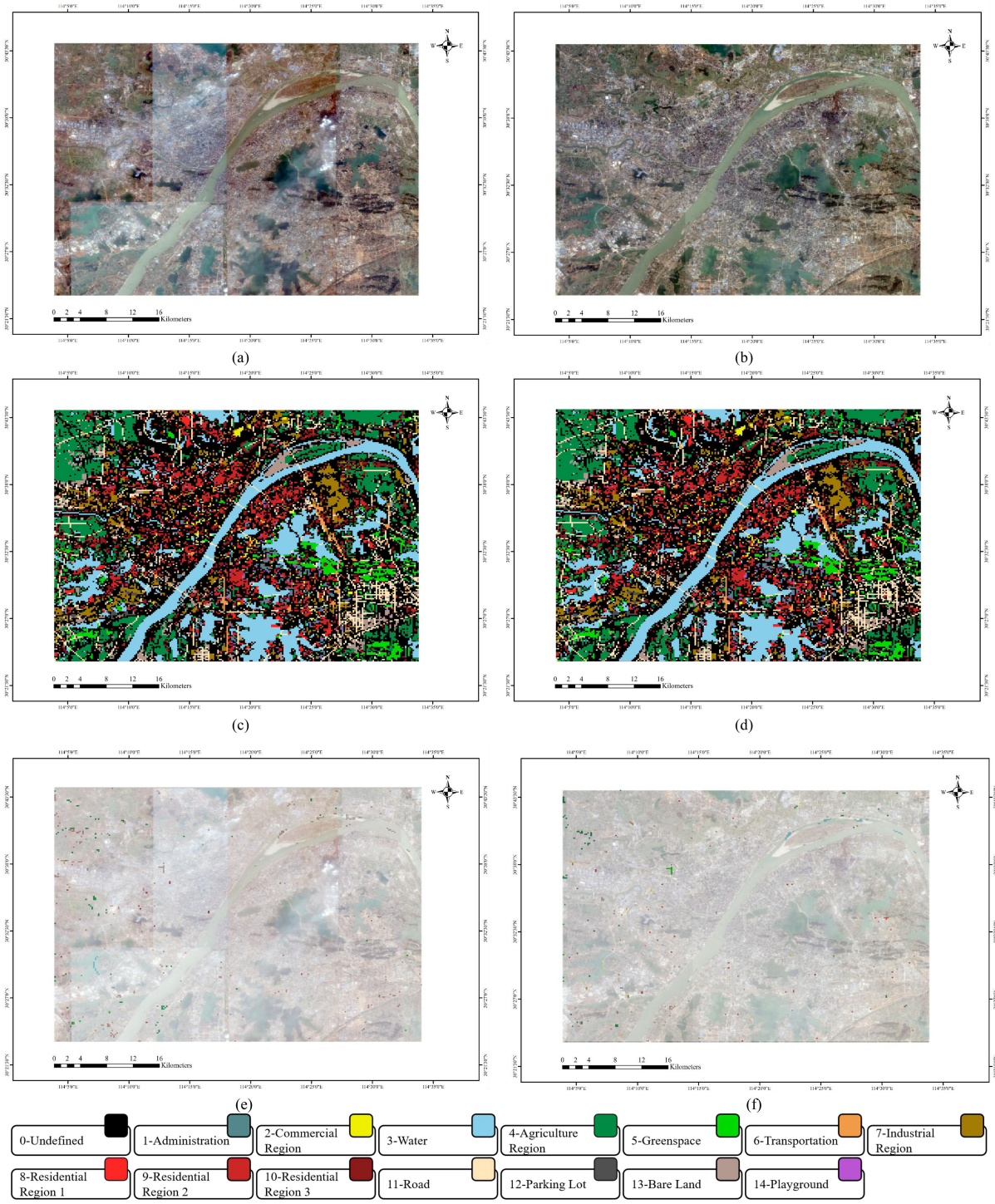

Fig. 3. Acquisition and distribution of a large range, high-resolution and multiapplication SC dataset, "WH-MAVS." A fusion image of the central area of Wuhan, obtained in (a) 2014 and (b) 2016. (c) and (d) Spatial layout of the WH-MAVS dataset in 2014 and 2016, respectively. (e) and (f) Spatial distribution maps of scene changes from 2014 to 2016 in the WH-MAVS dataset, respectively.

2014 and 2016, all of which are 200 × 200 pixels in size. On the basis of the new application-oriented classification system developed above, the WH-MAVS dataset comprises the following 14 categories: 1) administration; 2) commercial region; 3) water; 4) agricultural region; 5) green space; 6) transportation; 7) industrial region; 8) residential region 1; 9) residential region 2; 10) residential region 3; 11) road; 12) parking lot; 13) bare land; and 14) playground. The specific spatial layout of the data for the two time phases is outlined in Fig. 3(c) and (d). The SCD-labeled maps represent in Fig. 3(e) and (f). Moreover, Fig. 4 presents five samples for each class in the WH-MAVS.

In Table II, we coded each category and calculated the number of samples per category separately for 2014 and 2016, as well as the total and transformations for each category. Overall, the

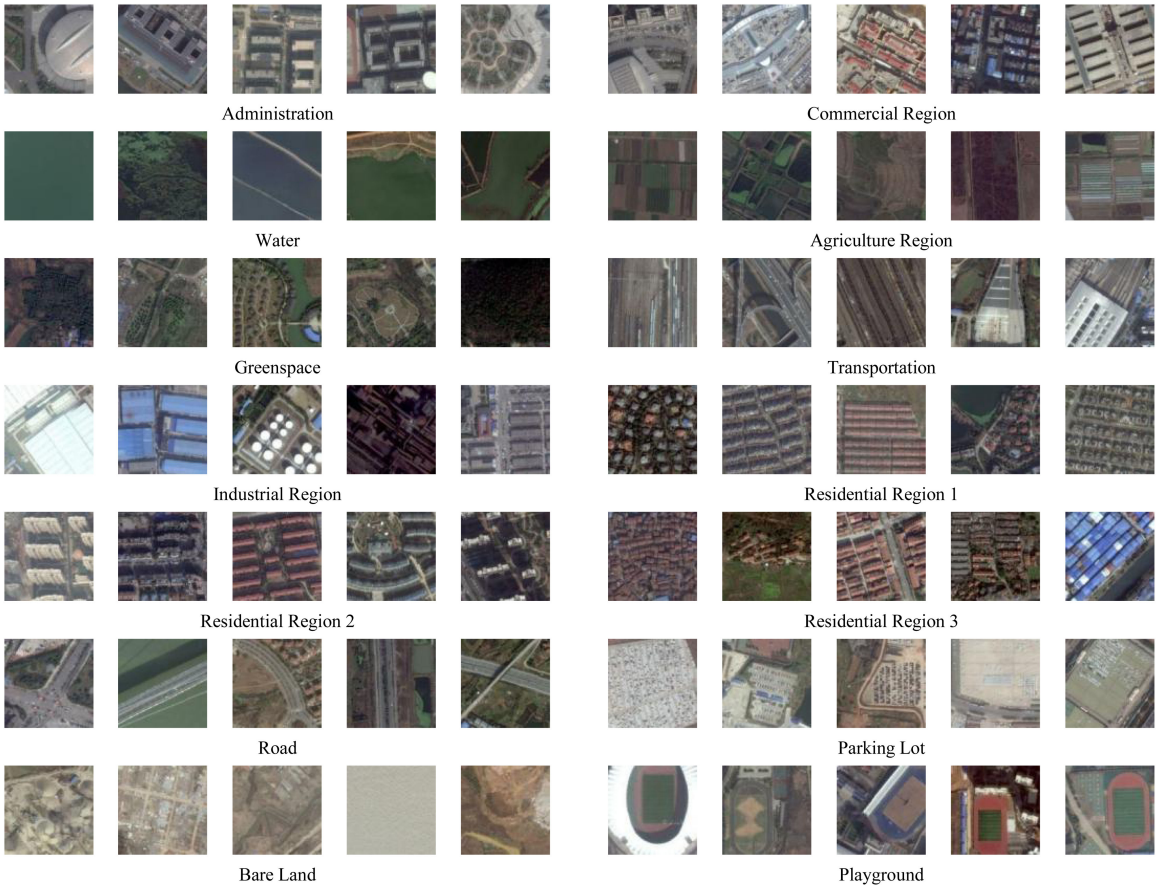

Fig. 4. Sample patches of WH-MAVS.

number of sample image patches in each category is highly unbalanced, ranging from 246 to 11 682. Specifically, 3-water and 4-agriculture region each contain a total of around 10 000 images; this stands in contrast to 12-parking lot, which has less than 250 images. This undoubtedly poses a significant challenge for the uses of the WH-MAVS dataset. Table II also presents the overall transformation (Trans) of each category, where "+" indicates an increase from 2014 to 2016, while "–" represents a "from–to" decrease. Thereinto, 11-road has the most significant rise with 110 patches, followed by 9-residential region 2, increased by 29 patches. The number of samples decreased in only three categories, namely 4-agriculture region, 10-residential region 3, and 13-bare land; among these, 13-bare land exhibited the largest reduction and the largest number of transformations of all categories, with 171 patches. The shrinkage of ten images in 4-agriculture region and nine images in 10-residential region 3 is much less than that in 13-bare land.

In light of the fact that the geographic coordinates of the two-time phase image patches correspond one-to-one, Table III elaborates the number of "from–to" changed and unchanged samples for each semantic scene class. In terms of changes among the samples in the semantic scene category from 2014 to 2016, the changes in 13-bare land were the most complex, followed by 11-road. The number of samples converted from 13-bare land to 11-road and 4-agriculture region ranked first and second, respectively. There are also a large number of transformations from other categories to 11-road and 13-bare land. As the statistics show, out of the total 23 567 samples taken per year, there were 23 202 unchanged areas and 365 changed areas in WH-MAVS. We refer to a transformation from one semantic scene category in 2014 to another category in 2016 as a "from–to" change type; for example, from 7-industrial region to 13-bare land or from 10-residential region 3 to 8-residential region 1. There are 31 "from–to" change types in WH-MAVS. Examples of change types are illustrated in Fig. 5.

In summary, our proposed WH-MAVS dataset has the following distinctive characteristics.

1) Authentic and Complex Dataset: In order to be closer to practical applications, our WH-MAVS dataset is annotated with as many samples as possible within a limited urban area and without overlap; as a result, the dataset is more atypical, but also more realistic. Therefore, a significant imbalance of categories inevitably exists in WH-MAVS, which represents a tremendous challenge when it comes to verifying the adaptation of the algorithm in the presence of highly unbalanced data volumes. Moreover, the application-oriented classification constraints and the atypical samples cause some classes of the dataset, such as the administration class and commercial region class in Fig. 4, to be characterized by little interclass differentiation and large intraclass variation, similar to industrial region, residential

TABLE III
NUMBER AND "FROM–TO" CHANGES OF EACH SCENE CATEGORY IN 2014 AND 2016 OF THE WH-MAVS DATASET

		2016														sum
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	
2014	**1**	520	0	0	0	0	0	0	0	0	0	0	0	0	0	520
	2	0	269	0	0	0	0	0	0	0	0	0	0	0	0	269
	3	0	0	5811	2	0	0	0	0	0	0	15	0	4	0	5832
	4	0	0	1	4735	4	0	0	0	0	0	36	0	35	0	4811
	5	0	0	0	2	1117	0	0	0	0	0	0	0	7	0	1126
	6	0	0	0	0	0	507	0	0	0	0	0	0	0	0	507
	7	0	0	0	0	0	0	1823	0	0	0	0	0	4	0	1827
	8	0	0	0	0	0	0	0	268	0	0	0	0	0	0	268
	9	0	0	0	0	0	0	0	0	2629	0	0	1	2	1	2633
	10	0	0	0	0	0	0	0	0	1	1365	2	0	6	0	1374
	11	0	0	0	0	0	2	2	0	2	0	2302	1	3	0	2312
	12	0	0	0	0	0	0	0	0	0	0	0	120	0	0	120
	13	1	4	38	62	10	0	15	0	30	0	67	4	1549	1	1781
	14	0	0	0	0	0	0	0	0	0	0	0	0	0	187	187
sum		521	273	5850	4801	1131	509	1840	268	2662	1365	2422	126	1610	189	23567

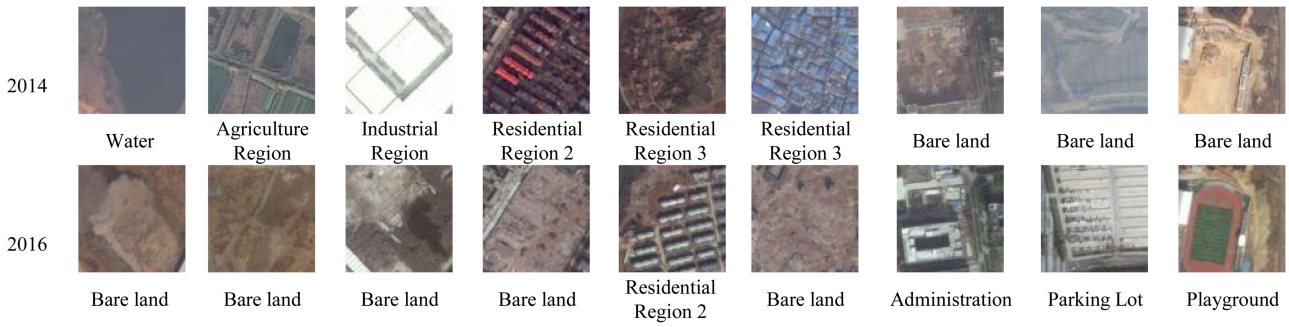

Fig. 5. Sample patches of "from–to" SCD in WH-MAVS.

region 2, etc. The realism and complexity of the WH-MAVS data can either contribute to measuring the robustness of SC and SCD algorithms or be more effectively applied to real applications.

2) Multitemporal Dataset: Considering that changes in pixel- or object-based features do not inherently lead to transformation in the high-level semantics of the surrounding scene, scholars are increasingly focused on detecting change at the scene level so as to meet the need for monitoring urban land use. However, few datasets are currently oriented toward SCD, and the size of the data is quite small. Accordingly, to advance the development of SCD algorithms, we present a multitemporal scene dataset with 23 567 samples per time phase; to the best of our knowledge, this is the largest volume of existing open source SCD datasets.

3) Real Dataset for a Large Range Metropolitan Area: Existing SC datasets are mainly developed for the purpose of promoting algorithm research. They accordingly tend to collect typical features scattered throughout the world as SC dataset samples. However, not all features are typical in the actual production process of land use, resulting in huge obstacles when the algorithm is applied to reality. Our proposed WH-MAVS dataset covers an urban central area of a megacity with nonoverlapping and seamless annotation that is as close as possible to human interpretation. Therefore, many scenario samples in this dataset are more complex and atypical, as well as realistic. Not only can the WH-MAVS facilitate the parallel advancement of algorithmic and application research but it can also build a bridge to close the "application gap" that presently exists in the field of SC.

4) Multiapplication Dataset: The WH-MAVS dataset is unique in being a publicly available, multiapplication-oriented RS dataset designed specifically for SC and SCD. Alongside its ability to promote algorithms that improve accuracy, our dataset can also be used to address practical applications; these encompass assisted mapping, landscape pattern analysis, urban land use dynamics monitoring, etc.

V. DATASET BENCHMARK

In this section, to verify the feasibility and effectiveness of our proposed WH-MAVS dataset, we investigate a variety of currently popular deep learning backbone networks designed for natural image classification. DenseNet [61] is generally the most accurate for classification tasks due to its dense fusion of multilayer characteristics. Within the same framework, a deeper model tends to outperform shallower models on larger range datasets. Nevertheless, it is possible that the network may do well on one dataset and not on another. Accordingly, in this article, we tested several deep learning methods—namely VGG16 [22], VGG19 [22], Inception V3 [23], ResNet50 [73], ResNet101 [73], ResNet152 [73], DenseNet121 [21], DenseNet169 [21], and DenseNet201 [21]—as benchmark algorithms to evaluate the performance of our proposed dataset.

TABLE IV
BENCHMARKED CLASSIFICATION ACCURACIES (%) OF DIFFERENT DEEP LEARNING METHODS ON VALIDATION SET OF WH-MAVS IN 2014 AND 2016

	OA_T1	OA_T2	OA_B	OA_T	KC_T1	KC_T2	Time (s)
VGG16	85.60%	90.14%	84.32%	80.91%	83.17%	88.49%	44.36
VGG19	85.68%	87.42%	82.39%	78.89%	83.29%	85.30%	53.40
Inception V3	88.06%	88.48%	85.60%	81.74%	86.05%	86.56%	44.72
ResNet50	85.41%	89.75%	84.73%	80.80%	82.96%	88.05%	43.43
ResNet101	84.26%	86.61%	81.06%	77.24%	81.59%	84.36%	76.68
ResNet152	83.09%	87.68%	80.63%	77.13%	80.22%	85.62%	100.88
DenseNet121	90.88%	91.75%	**89.90%**	86.61%	89.34%	90.36%	59.80
DenseNet169	**91.07%**	**92.09%**	89.56%	**86.70%**	**89.57%**	**90.76%**	71.86
DenseNet201	90.88%	91.51%	89.01%	86.02%	89.34%	90.09%	93.81

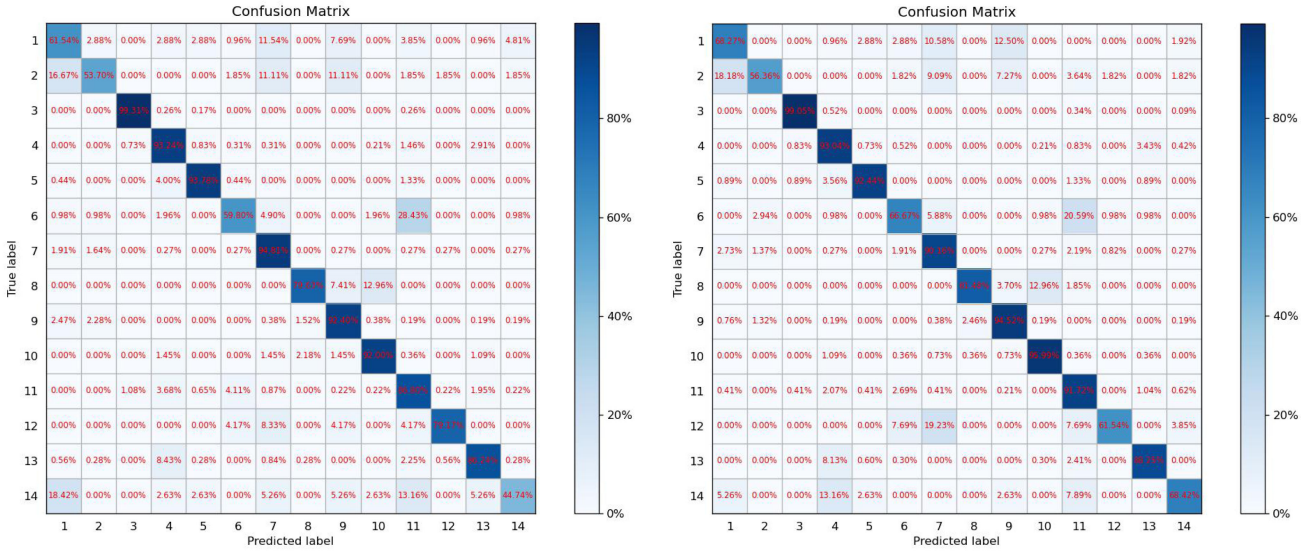

Fig. 6. Confusion matrix computed by DenseNet169 on the validation set of our proposed WH-MAVS dataset. (a) 2014 and (b) 2016.

The dataset was divided into a training set, a validation set, and a test set (70/20/10 ratio), respectively. It should be noted that all demonstrated methods were performed using TensorFlow on a single RTX 2080Ti GPU. We employed momentum optimization [74] with an initial learning rate of 0.001 and a momentum parameter of 0.9 in the training phase. The number of epochs was set to 100 and the batch size was 16. During the training process, we set the $\mathcal{L}_2$ regularization factor value to 0.0001.

A. Assessment Criteria

We herein focus primarily on the overall accuracy (OA), Kappa coefficient (KC), and confusion matrix as the accuracy evaluation metrics for the WH-MAVS. Notably, because the WH-MAVS is multitemporal, we assess the accuracy of each method in 2014 and 2016 individually. As illustrated in Table IV, OA_T1 and OA_T2 illuminate the OA of the validation set classification results in 2014 and 2016. Moreover, OA_B and OA_T are the rubrics for measuring the accuracy of SCD. Here, OA_B illuminates the OA metric of the binary change detection (i.e., it focuses only on whether each scene patch has changed) while OA_T denotes the from–to transition accuracy (which is specifically concerned with the accuracy of the classification results of the two temporal phases agreeing with the corresponding ground truth). Furthermore, we use the confusion matrix to clearly depict the specific accuracy of each category. A heat map was employed to indicate the level of accuracy; the deeper the color, the higher the accuracy, and *vice versa*. The KC also needs to be evaluated owing to the extremely imbalanced distribution of the category samples in our dataset. We separately evaluate the 2014 and 2016 single time-phase KC, denoted as KC_T1 and KC_T2.

B. Experimental Results

Our results are presented in Table IV. Here, the best result in each column has been highlighted in bold, while the second best result in each column has been underlined. As the table shows, DenseNet performs significantly better than other networks in SC and SCD. DenseNet121 obtains the best performance of OA_B in SCD, while DenseNet169 scores the highest not only for classifying the scene images in two temporal phases but also for the detection accuracy of the two-time phase changes in the WH-MAVS dataset, and even for KC. Subsequently, we select the best overall performing network, DenseNet169, and the SC confusion matrices are calculated for this network on

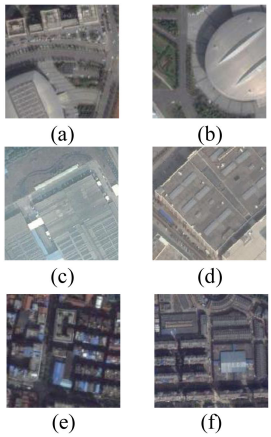

Fig. 7. Comparison of confusing samples in 2-commercial region. The left column is for category 2-commercial region, while the right column is the example of easily confusable categories. (a) 2-commercial region, (b) 1-administration, (c) 2-commercial region, (d) 7-industrial region, (e) 2-commercial region, and (f) 9-residential region.

the validation sets of the WH-MAVS dataset in 2014 and 2016, respectively.

The confusion matrices of the two temporal phases in Fig. 6(a) and (b) reveal that the following eight categories exhibit a superior classification accuracy that exceeds 80%: 3-water, 4-agricultural region, 5-green space, 7-industrial region, 9-residential region 2, 10-residential region 3, 11-road, and 13-bare land. This is mainly due to the fact that the number of samples in each of these categories exceeds 1000, and moreover that each category is relatively simple in terms of its texture and geometric structure.

The remaining six categories suffer from serious misclassification. It is interesting to note that the classification accuracy results for the 14-playground class were over 68% in 2014, but only 44% in 2016. Considering that the sample size of this category only just exceeds 180, we propose that the unstable classification results may be due to the highly unbalanced sample sizes. The existence of more incorrect classifications of category 1-administration into category 14-playground in the confusion matrix of 2014 can be attributed to the separation of category 14-playground from category 1-administration, as their contexts are more similar. Another class that exhibits misclassification in both time phases, 11-road class, has far more samples than and is similar to 14-playground class in terms of spatial distribution or texture configuration.

In addition, by comparing Fig. 6 with Table II, it can be determined that several other categories with very low OA—2-commercial region class, 6-transportation class, and 12-parking lot class—also have a small number of samples. This is due to the fact that the texture and color characteristics of the parking lot are very similar to a large portion of white factory roofs, as well as that the number of instances in the 7-industrial region class is drastically higher than that in the 12-parking lot class, causing poor classification results.

One of the reasons why 2-commercial region and 1-administration cannot be easily distinguished is that some

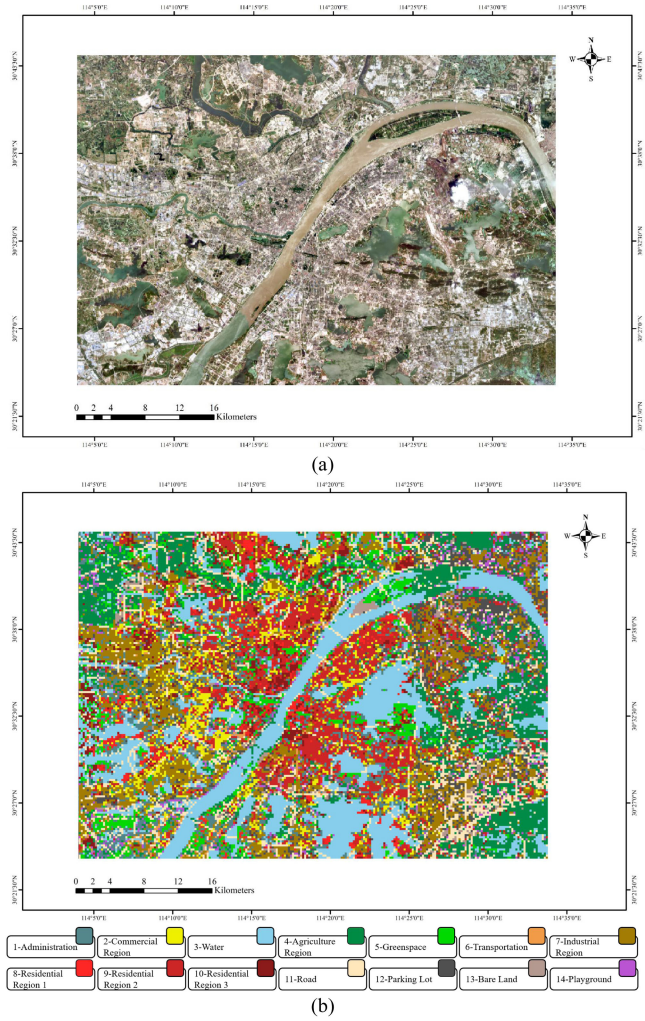

Fig. 8. (a) Free 2 m resolution GaoFen-1 satellite RS image of the Wuhan area in 2020. (b) Results of classifying the urban planning land use map in central Wuhan in 2020 after training on the WH-MAVS dataset using the DenseNet169 classification network.

large administrative buildings in category 1-administration (such as science and technology museums and art museums) are very similar in terms of geometric shapes to the large shopping malls and concert halls in category 2-commercial region, as shown in Fig. 7(a) and (b). The percentage of category 2-commercial region that is misclassified as 7-industrial region and 9-residential region 2 is also high. Among these, the sample size of 9-residential region 2 class is approximately seven times larger than that of 2-commercial region class, while the texture of some samples in 9-residential region 2 class is very similar to that of the 2-commercial region class, as shown in Fig. 7(c) and (d); this is also the main reason why they are more frequently scored incorrectly. Moreover, as shown in Fig. 7(e) and (f), some of the retail-based stores in 2-commercial region class are very similar to some of the residential areas in category 9-residential region 2, which leads to incorrect classification.

The type that causes 6-transportation class to be less accurate is mainly 11-road class. This is because their texture structures

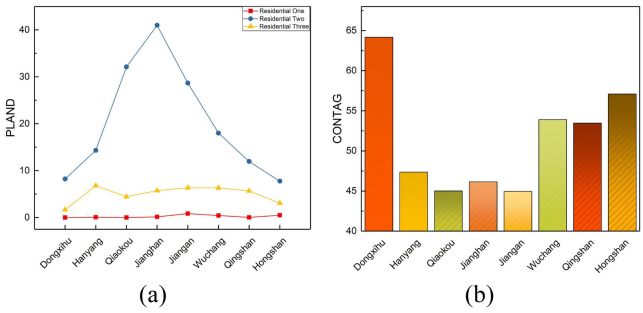

Fig. 9. Several indicators for landscape pattern analysis. (a) PLAND and (b) CONTAG.

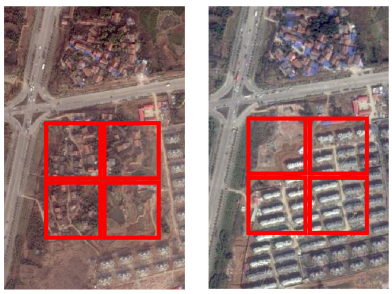

Fig. 10. Sample of urban SCD results. (a) Detected area before the change and (b) corresponding detection result after the change.

are very similar; notably, the difficulty of distinguishing between the 6-transportation and 11-road classes is caused by extreme sample imbalance. Although the number of samples in category 8-residential region 1 is very small, the results are quite good in terms of classification accuracy due to the distinctive characteristics of the scene structure.

VI. Application

In this section, we present the scope of the application and development prospects of the new application-oriented WH-MAVS dataset in the direction of SC and SCD, respectively. Our proposed novel dataset can not only advance the research development of SC and SCD algorithms, but also meet the requirements of various urban LULC applications. We present several examples, including updates to urban planning maps, the analysis of urban landscape patterns, the monitoring of urban land use changes, etc.

A. Scene Classification

Our WH-MAVS can be used as an SC dataset in a single temporal phase. It is also possible to integrate two temporal phases of data together for SC algorithm research. Given that our proposed WH-MAVS is characterized by large range, rich intraclass variation, indistinguishable interclass similarity, and extremely unbalanced sample data, it provides a priceless research dataset for use by SC algorithms in solving related problems. In addition, since our classification criteria have been developed for applications related to urban LULC, the more accurate classification results obtained from the improved algorithms can be applied directly to urban planning, decision formulation, and evaluation, thereby achieving a seamless bridge between algorithm and application.

1) Updating and Assistance in Urban Planning Mapping: In the actual production tasks faced by the relevant departments, such as urban planning and mapping bureaus, there is a large demand for map updates. For example, the urban planning department updates the urban LULC map annually according to the demand and makes further urban planning decisions and ministries on this basis. However, rather than updating the map in its entirety, a portion of areas are selected each year for fieldwork and updating. This is not only costly in terms of labor, material, and financial resources but also makes it impossible to obtain the latest LULC maps for all urban areas in the city each year. Insufficiently accurate land use maps can have a detrimental effect on decision-making authorities responsible for the development of urban planning maps. Worse yet, it may lead to more imprecise urban planning assessments and deviations in the development of new policies. For example, we train a benchmark classification network, DenseNet169, using the WH-MAVS dataset. We then test the latest images of Wuhan central in 2020 acquired by the 2 m/pixel high-resolution RS satellite, GaoFen-1, which is released for free by the High-Resolution Earth Observation System of Systems Hubei Data and Applications website (www.hbeos.org.cn). In Fig. 8(b), the fast, accurate, and efficient urban planning LULC reference can be achieved. Moreover, this approach enables the identification of undefined categories in the WH-MAVS dataset. Thus, it can be used as a land use reference for review and verification by urban planning departments.

2) Analysis of Urban Landscape Patterns: The ecology of urban landscapes has received a great deal of attention among those concerned with the construction of cities and their sustainable development. After training the network with our proposed dataset and obtaining an urban LULC map with high confidence, the landscape pattern within the city can be analyzed as an effective aid in analyzing the ecological rationality of the landscape. The distribution of the landscape in the city's various districts was analyzed using class metrics and landscape metrics, with the aid of relevant analysis software such as FRAGSTATS.

We used the SC results in the central city of Wuhan in 2020 as a base map for landscape pattern analysis. As shown in Fig. 9(a), the percentage of landscape (PLAND) index was statistically analyzed in terms of class metrics for the different types of residential areas in each district of the city. The overall PLAND index for residential region 2 is much higher than the other two classes, reflecting the higher standard of living per capita in Wuhan. The PLAND index of the Hanyang district in Wuhan for residential region 3 class is significantly higher than in other districts, indicating that the phenomenon of "urban villages" is more serious in Hanyang district.

Fig. 9(b) represents the degree of contagion (CONTAG) index at the landscape metric. It is inversely proportional to the level of urbanization; that is, poorer connectivity indicates greater fragmentation and higher levels of urbanization. Dongxihu district has the highest value, which represents the lowest level

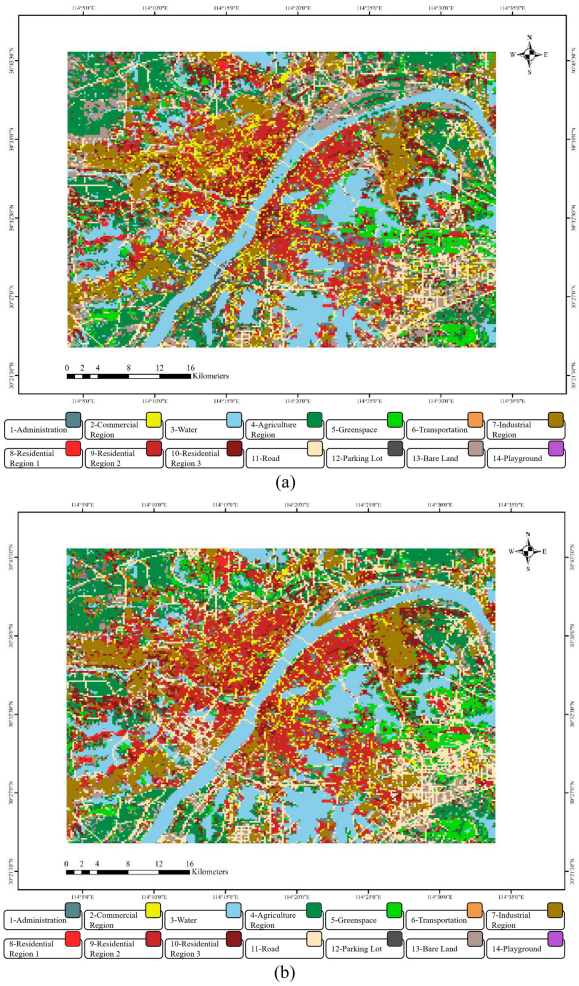

Fig. 11. Prediction LULC maps before and after the changes obtained on WH-MAVS through the use of a trained SCD network.

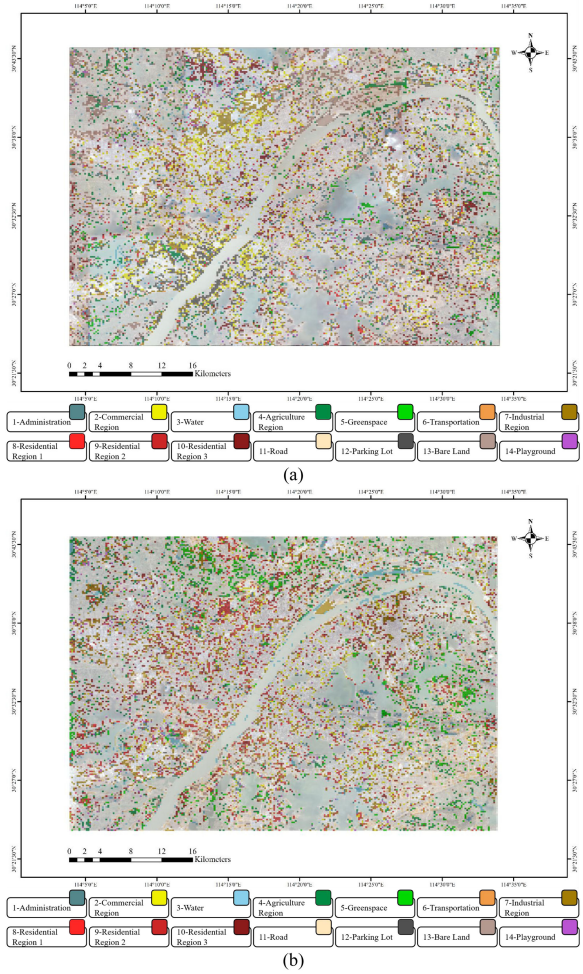

Fig. 12. Land use scene change maps obtained by using a trained SCD network on WH-MAVS. (a) Before the change and (b) after the change.

of urbanization. Since this region is a distant suburban area in Wuhan and still has large areas of interconnected farmland, it exhibits hysteresis compared to the main districts. Hanyang district, Qiaokou district, Jianghan district, and Jiang'an district are all the main districts with a very high degree of urban development. By contrast, although Hongshan district is also one of the main districts, it has a higher CONTAG value. One very important reason for this concerns the presence of a larger area of water within its district.

B. Change Detection

As urbanization continues to intensify, land use changes are continuously occurring and have attracted widespread scholarly attention. The task of scene-level change detection is more valuable for research, as it focuses on the situation of change in high-level semantic scenes rather than only pixel-level and object-level land use changes. Our open source, freely available high-resolution dataset provides the largest dataset to date for SCD research, while also being large range with massive sample size and multitemporal characteristics. The total number of samples in WH-MAVS is consistent across all temporal phases; each scene region exhibits one-to-one geographic correspondence, and a rich sample of changes is also present. The dataset will now facilitate progress in algorithmic research [70] and represents an important contribution to the advancement of algorithms for SCD. It may also come to play an important role in the application of urban dynamic monitoring.

In recent years, the dynamic monitoring of cities has become a hot topic. The analysis of land use changes within cities over a long time series is useful for monitoring illegal construction, assessing real estate prices, and studying economic development within cities.

One fact of note is that our multitemporal dataset, being trained on the SCD network, can precisely identify urban regions in which semantic changes occur. In the sample of detected results shown in Fig. 10, the red box represents the location of the detected scene image patches. Here, the transition from Fig. 10(a) to (b) indicates that there is a portion of residential region 3 land at the edge of the city that has been bulldozed and rebuilt as an upscale residential area over the course of a year or two. This enables the relevant authorities to quickly and precisely detect and locate change areas, as well as to explore the potential drivers of the existing change.

Moreover, the results of the detection facilitate analysis of changes occurring in the city. Although our dataset already

contains a large number of annotated samples pertaining to the city, many undefined regions remain. With the labeled samples taken as the training set, the network learning of change detection (such as that in CorrFusionNet [70]) is performed to predict the classes of features in the undefined areas in order to obtain the overall maps of changes occurring across the whole city (see Figs. 11 and 12). The predicted maps can provide more complete maps of the urban regions of change, which can in turn lead to more effective monitoring of urban dynamics.

VII. Conclusion

In this article, we present the first large range dataset capable of bridging the gap between algorithms and applications and pushing down the barriers of datasets between different tasks, such as SC and SCD. This new scene dataset, WH-MAVS, has multitemporal and multiapplication characteristics. It is the only application-oriented and open source dataset currently in existence that is large range and focuses on the central area of one megacity. In order to create the WH-MAVS, we collected two high-resolution satellite RS images of Wuhan's central urban area from 2014 and 2016, respectively, via GE, both with a resolution of 1.2 m/pixel. A set of application-oriented classification criteria was developed in accordance with the relevant classification reference for urban planning. Dataset creation was achieved through a semiautomatic and interactive approach. The dataset contains 14 categories with a total of two time phases, each with 23 567 image patches; moreover, the geographic location of each image patch in each time phase corresponds to that of the same patch in the other phase. We retain information on the geographic coordinates of each sample, which can be remapped back onto the map. We select the most popular deep learning classification networks of the current time for benchmark classification and change detection on multiple time phases separately. The dataset was found to perform best on DenseNet169. Tests were conducted on the 2014 and 2016 datasets with the accuracy of 91.07% and 92.09%, respectively. The binary change detection accuracy for SCD was found to be 89.56%, while the multiple (from–to) change accuracy was 86.70%. The KC of the two-time phases was found to reach 89.57% in 2014 and 90.76% in 2016. Overall, the proposed dataset is the first to unify SC and SCD in terms of categories and further provides the largest dataset for SCD that is currently public and freely available. Not only does our proposed dataset make a significant contribution to algorithmic research in the context of SC and SCD but it is already playing an important role in real-life production, such as urban planning mapping updates and assistance, landscape pattern analysis, and monitoring of urban LULC dynamics. In the future, our proposed dataset will also be applied to more and broader practical applications through methods such as transfer learning. This will lead to the rapid integration of industry, academia, and research.

References

[1] B. Huang, B. Zhao, and Y. Song, "Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery," *Remote Sens. Environ.*, vol. 214, pp. 73–86, Sep. 2018.

[2] J. Gong, C. Liu, and X. Huang, "Advances in urban information extraction from high-resolution remote sensing imagery," *Sci. China Earth Sci.*, vol. 63, no. 4, pp. 463–475, Dec. 2019.

[3] J. Peng, P. Xie, Y. Liu, and J. Ma, "Urban thermal environment dynamics and associated landscape pattern factors: A case study in the Beijing metropolitan region," *Remote Sens. Environ.*, vol. 173, pp. 145–155, Feb. 2016.

[4] J. Peng *et al.*, "Low-rank and sparse representation for hyperspectral image processing: A review," *IEEE Geosci. Remote Sens. Mag.*, to be published, doi: 10.1109/MGRS.2021.3075491.

[5] J. Peng, Y. Zhou, W. Sun, Q. Du, and L. Xia, "Self-paced nonnegative matrix factorization for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1501–1515, Feb. 2021.

[6] W. Sun *et al.*, "A multiscale spectral features graph fusion method for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2021.3102246.

[7] G. Grekousis, "Artificial neural networks and deep learning in urban geography: A systematic review and meta-analysis," *Comput. Environ. Urban Syst.*, vol. 74, pp. 244–256, Mar. 2019.

[8] D. Phiri and J. Morgenroth, "Developments in Landsat land cover classification methods: A review," *Remote Sens.*, vol. 9, no. 9, 2017, Art. no. 967.

[9] S. Liu and Q. Shi, "Local climate zone mapping as remote sensing scene classification using deep learning: A case study of metropolitan China," *ISPRS J. Photogramm. Remote Sens.*, vol. 164, pp. 229–242, Jun. 2020.

[10] W. Sun *et al.*, "A simple and effective spectral-spatial method for mapping large-scale coastal wetlands using China ZY1-02D satellite hyperspectral images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 104, Dec. 2021, Art. no. 102572.

[11] C. Wu, L. Zhang, and B. Du, "Kernel slow feature analysis for scene change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 4, pp. 2367–2384, Apr. 2017.

[12] C. Wu, L. Zhang, and L. Zhang, "A scene change detection framework for multi-temporal very high resolution remote sensing images," *Signal Process.*, vol. 124, pp. 184–197, Jul. 2016.

[13] G. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

[14] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[15] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[19] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, vol. 9905, pp. 21–37.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., Jun. 2016, pp. 779–788.

[21] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., Jul. 2017, pp. 2261–2269.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–14.

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit* ., Jun. 2016, pp. 2818–2826.

[24] W. Han, R. Feng, L. Wang, and Y. Cheng, "A semi-supervised generative framework with deep learning features for high-resolution remote sensing image scene classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 23–43, Nov. 2018.

[25] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[26] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[27] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[28] S. Liu, Q. Shi, and L. Zhang, "Few-shot hyperspectral image classification with unknown classes using multitask deep learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5085–5102, Jun. 2021.

[29] S. Šćepanović, O. Antropov, P. Laurila, Y. Rauste, V. Ignatenko, and J. Praks, "Wide-area land cover mapping with Sentinel-1 imagery using deep learning semantic segmentation models," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10357–10374, Sep. 2021. doi: 10.1109/JSTARS.2021.3116094.

[30] M. Schmitt, L. Hughes, C. Qiu, and X. Zhu, "SEN12MS – A curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion," *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. 2019, pp. 153–160, Sep. 2019.

[31] L. Mou, L. Bruzzone, and X. Zhu, "Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 924–935, Feb. 2019.

[32] H. Chen, C. Wu, B. Du, L. Zhang, and L. Wang, "Change detection in multisource VHR images via deep siamese convolutional multiple-layers recurrent neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2848–2864, Apr. 2020.

[33] Q. Shi, X. Tang, T. Yang, R. Liu, and L. Zhang, "Hyperspectral image denoising using a 3-D attention denoising network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10348–10363, Dec. 2021.

[34] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.

[35] J. Shao, B. Du, C. Wu, M. Gong, and T. Liu, "HRSiam: High-resolution siamese network, towards space-borne satellite video tracking," *IEEE Trans. Image Process.*, vol. 30, pp. 3056–3068, Feb. 2021. doi: 10.1109/TIP.2020.3045634.

[36] J. Shao, B. Du, C. Wu, and L. Zhang, "Can we track targets from space? A hybrid kernel correlation filter tracker for satellite video," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8719–8731, Nov. 2019.

[37] Y. Dong, T. Liang, Y. Zhang, and B. Du, "Spectral-spatial weighted kernel manifold embedded distribution alignment for remote sensing image classification," *IEEE Trans. Cybern.*, vol. 51, no. 6, pp. 3185–3197, Jun. 2021.

[38] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[39] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data a technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[40] E. Protopapadakis, A. Doulamis, N. Doulamis, and E. Maltezos, "Stacked autoencoders driven by semi-supervised learning for building extraction from near infrared remote sensing imagery," *Remote Sens.*, vol. 13, no. 3, 2021, Art. no. 371.

[41] D. Hong, N. Yokoya, G.-S. Xia, J. Chanussot, and X. Zhu, "X-ModalNet: A semi-supervised deep cross-modal network for classification of remote sensing data," *ISPRS J. Photogramm. Remote Sens.*, vol. 167, pp. 12–23, Sep. 2020.

[42] K. Chen, K. Fu, X. Gao, M. Yan, X. Sun, and H. Zhang, "Building extraction from remote sensing images with deep learning in a supervised manner," in *Proc. IEEE Int. Geosci. Remote Sens.*, Jul. 2017, pp. 1672–1675.

[43] E. Maltezos, A. Doulamis, N. Doulamis, and C. Ioannidis, "Building extraction from LiDAR data applying deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 1, pp. 155–159, Jan. 2019.

[44] A. Turlapaty, B. Gokaraju, Q. Du, N. Younan, and J. Aanstoos, "A hybrid approach for building extraction from spaceborne multi-angular optical imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 1, pp. 89–100, Feb. 2012.

[45] W. Sun, K. Ren, X. Meng, C. Xiao, G. Yang, and J. Peng, "A band divide-and-conquer multispectral and hyperspectral image fusion method," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5502113, doi: 10.1109/TGRS.2020.3046321.

[46] W. Sun, J. Peng, G. Yang, and Q. Du, "Fast and latent low-rank subspace clustering for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 3906–3915, Jun. 2020.

[47] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. ACM Int. Conf. Adv. Geograph. Inf. Syst.*, 2010, pp. 270–279.

[48] S. Basu, S. Ganguly, S. Mukhopadhyay, R. Dibiano, M. Karki, and R. Nemani, "DeepSat: A learning framework for satellite imagery," in *Proc. Int. Conf. Adv. Geographic Inf. Syst.*, 2015, pp. 1–10.

[49] L. Zhao, P. Tang, and L. Huo, "Feature significance-based multibag-of-visual-words model for remote sensing image scene classification," *J. Appl. Remote Sens.*, vol. 10, no. 3, 2016, Art no. 35004.

[50] Q. Zhu, Y. Zhong, B. Zhao, G.-S. Xia, and L. Zhang, "Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 6, pp. 747–751, Jun. 2016.

[51] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

[52] G.-S. Xia *et al.*, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.

[53] W. Zhou, S. Newsam, C. Li, and Z. Shao, "PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 197–209, Nov. 2018.

[54] H. Li *et al.*, "RSI-CB: A large-scale remote sensing image classification benchmark using crowdsourced data," *Sensors*, vol. 20, no. 6, Mar. 2020, Art. no. 1594.

[55] Z. Xiao, Y. Long, D. Li, C. Wei, G. Tang, and J. Liu, "High-resolution remote sensing image retrieval based on CNNs from a dimensional perspective," *Remote Sens.*, vol. 9, no. 7, 2017, Art. no. 725.

[56] H. Li *et al.*, "CLRS: Continual learning benchmark for remote sensing image scene classification," *Sensors*, vol. 20, no. 4, Feb. 2020, Art. no. 1226.

[57] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, Jul. 2019.

[58] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "BigEarthNet: A large-scale benchmark archive for remote sensing image understanding," in *Proc. IEEE Int. Geosci. Remote Sens.*, Jul. 2019, pp. 5901–5904.

[59] Y. Long *et al.*, "On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4205–4230, Apr. 2021, doi: 10.1109/JSTARS.2021.3070368.

[60] X. Qi *et al.*, "MLRSNet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding," *ISPRS J. Photogramm. Remote Sens.*, vol. 169, pp. 337–350, Nov. 2020.

[61] G.-S. Xia, W. Yang, J. Delon, Y. Gousseau, H. Sun, and H. Maître, "Structural high-resolution satellite image indexing," in *Proc. ISPRS TC VII Symp. - 100 Years ISPRS*, 2010, pp. 298–303.

[62] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, Nov. 2015.

[63] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019.

[64] M. Liu, Q. Shi, A. Marinoni, D. He, and L. Zhang, "Super-resolution-based change detection network with stacked attention module for images with different resolutions," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2021.3091758.

[65] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2021.3085870.

[66] R. Daudt, B. Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral earth observation using convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens.*, Jul. 2018, pp. 2115–2118.

[67] D. He, Y. Zhong, and L. Zhang, "Land cover change detection based on spatial-temporal sub-pixel evolution mapping: A case study for urban expansion," in *Proc. IEEE Int. Geosci. Remote Sens* ., Jul. 2018, pp. 1970–1973.

[68] K. Yang, G.-S. Xia, Z. Liu, B. Du, and M. Pelillo, "Asymmetric siamese networks for semantic change detection," 2020, *arXiv:2010.05687*.

[69] S. Tian, Y. Zhong, A. Ma, and Z. Zheng, "Hi-UCD: A large-scale dataset for urban semantic change detection in remote sensing imagery," 2020, *arXiv:2011.03247*.

[70] L. Ru, B. Du, and C. Wu, "Multi-temporal scene classification and scene change detection with correlation based fusion," *IEEE Trans. Image Process.*, vol. 30, pp. 1382–1394, Nov. 2021, doi: 10.1109/TIP.2020.3039328.

[71] C. Wu, B. Du, X. Cui, and L. Zhang, "A post-classification change detection method based on iterative slow feature analysis and Bayesian soft fusion," *Remote Sens. Environ.*, vol. 199, pp. 241–255, Sep. 2017.
[72] Q. Hu *et al.*, "Exploring the use of Google Earth imagery and object-based methods in land use/cover mapping," *Remote Sens.*, vol. 5, no. 11, pp. 6026–6042, Nov. 2013.
[73] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[74] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. 30th Int. Conf. Mach. Learn.*, Jun. 2013, pp. 1139–1147.

Jingwen Yuan (Student Member, IEEE) received the B.S. degree in photogrammetry and remote sensing in 2015 from Wuhan University, Wuhan, China, where she is currently working toward the Ph.D. degree in photogrammetry and remote sensing.

Her research interests include urban functional zones extraction and analysis in multispectral and hyperspectral remote sensing images.

Lixiang Ru (Graduate Student Member, IEEE) received the B.S. degree in information security in 2018 from the School of Cyber Science and Engineering, Wuhan University, Wuhan, China, where he is currently working toward the Ph.D. degree in artificial intelligence with the School of Computer Science.

His research interests include deep learning and computer vision.

Shugen Wang (Member, IEEE) received the B.S. degree in aerial photogrammetry from the Wuhan College of Surveying and Mapping, Wuhan, China, in 1984, the M.S. degree in photogrammetry and remote sensing from the Wuhan University of Surveying and Mapping Science and Technology, Wuhan, China, in 1994, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2003.

He is currently a Professor with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China. His research interests include digital photogrammetry, high-spatial-resolution remote sensing image processing, and computer vision.

Chen Wu (Member, IEEE) received the B.S. degree in surveying and mapping engineering from Southeast University, Nanjing, China, in 2010, and the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2015.

He is currently an Associate Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China. His research interests include multitemporal remote sensing image change detection and analysis in multispectral and hyperspectral images.