DiRS: On Creating Benchmark Datasets for Remote Sensing Image Interpretation

Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang,
Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, Deren Li

Abstract—The past decade has witnessed the great progress on remote sensing (RS) image interpretation and its wide applications. With RS images becoming more accessible than ever before, there is an increasing demand for the automatic interpretation of these images, where benchmark datasets are essential prerequisites for developing and testing intelligent interpretation algorithms. After reviewing existing benchmark datasets in the research community of RS image interpretation, this article discusses the problem of how to efficiently prepare a suitable benchmark dataset for RS image analysis. Specifically, we first analyze the current challenges of developing intelligent algorithms for RS image interpretation with bibliometric investigations. We then present some principles, *i.e.*, ***di****versity*, ***r****ichness*, and ***s****calability* (called **DiRS**), on constructing benchmark datasets in efficient manners. Following the DiRS principles, we also provide an example on building datasets for RS image classification, *i.e.*, **Million-AID**, a new large-scale benchmark dataset containing million instances for RS scene classification. Several challenges and perspectives in RS image annotation are finally discussed to facilitate the research in benchmark dataset construction. We do hope this paper will provide RS community an overall perspective on constructing large-scale and practical image datasets for further research, especially data-driven ones.

Index Terms—Remote sensing image interpretation, annotation, benchmark datasets, scene classification, Million-AID

1 Introduction

The advancement of remote sensing (RS) technology has significantly improved the ability of human beings to characterize features of the earth surface [1], [2]. With more and more RS images being available, the interpretation of RS images has been playing an important role in many applications, such as environmental monitoring [3], [4], resource investigation [5]–[7], and urban planning [8], [9], *etc*. However, with the rapid development of the earth observation technology, the volume of RS images is confronted with dramatic growth. And the rich details in RS images, such as the geometrical shapes, structural characteristics, and textural attributes, pose challenges to the interpretation of image contents [10]–[13]. Moreover, the information extraction, analysis and application of RS images in practice rely heavily on visual interpretation and manual processing by experts [14]. These present the increasing and stringent demands for automatic and intelligent interpretation on the blooming RS imagery.

To characterize contents in RS images, quite a few methods have been developed for a wide range of interpretation tasks, from complex scene recognition [15]–[20], object-level image analysis [21]–[27] to the challenging pixel-wise semantic understanding [28]–[34]. Benefiting from the increasing availability and various ontologies of RS images, the developed methods have reported promising performance on the interpretation of RS image contents. However, most of the current methods are evaluated upon small-scale image datasets which usually show domain bias for applications. Moreover, the existing datasets created toward specific algorithms rather than practical scenarios cannot objectively and fairly validate the effectiveness of the others [35], [36]. Recently, it is observed that data-driven approaches, particularly the deep learning ones [37]–[39], have become an important alternative to manual interpretation and provide a bright prospect for automatic interpretation, analysis and content understanding from the ocean of RS images. However, the training and testing effectiveness could be curbed owing to the lack of adequate and accurately annotated groundtruth datasets. As a result, it usually turns out to be difficult to apply the interpretation models in real-world applications. Thus, it is natural to argue that a great amount of efforts need to be paid for datasets construction considering the following points:

- *The ever-growing volume of RS images is acquired while very few of them are annotated with valuable information.*

- *The study of this paper is funded by the National Natural Science Foundation of China (NSFC) under grant contracts No.61922065, No.61771350 and No.41820104006 and 61871299. It is also partially funded by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond".*
- *Y. Long, L. Zhang, D. Li are with the State Key Lab. LIESMARS, Wuhan University, Wuhan, China. e-mail:* {longyang, zlp62, drli}*@whu.edu.cn*
- *G.-S. Xia is with the School of Computer Science and also the State Key Lab. LIESMARS, Wuhan University, Wuhan, China. e-mail: guisong.xia@whu.edu.cn*
- *S. Li is with the Key Laboratory of Space Utilization, Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing, China. e-mail: shyli@csu.ac.cn*
- *W. Yang is with the School of Electronic Information and the State Key Lab. LIESMARS, Wuhan University, Wuhan China. e-mail: yangwen@whu.edu.cn*
- *M. Y. Yang is with the Faculty of Geo-Information Science and Earth Observation, University of Twente, Hengelosestraat 99, Enschede, Netherlands. e-mail: michael.yang@utwente.nl*
- *X. Zhu is with the German Aerospace Center (DLR) and also Technical University of Munich, Germany. e-mail: xiaoxiang.zhu@dlr.de*
- *The corresponding author is Gui-Song Xia (guisong.xia@whu.edu.cn).*

With the rapid development and continuous improvement of sensor technology, it is incredibly convenient to receive RS data with various modalities, *e.g.*, optical, hyper-spectral and synthetic aperture radar (SAR) images. Consequently, a huge amount of RS images with different spatial, spectral, and temporal resolutions is received every day than ever before, providing abundant information of the ground features [5], [22], [35]. However, in contrast to the huge amount of received RS images, images annotated with valuable information are relatively few, making it impossible to be productively utilized and also resulting in great waste.

- *The generalization ability of algorithms for interpreting RS images is of great urgency to be enhanced.* Although a multitude of machine learning [40]–[42] and deep learning algorithms [37], [43], [44] have been developed for RS image interpretation, the interpretation capability of the algorithms could be constrained since RS images are usually characterized with complex and diverse content. Besides, existing algorithms are usually trained on a small-scale dataset, which shows weak representation ability for the real-world feature distribution. Consequently, the constructed algorithms inevitably show limitations, *e.g.*, weak generalization ability, in practical applications. Therefore, more robust and intelligent algorithms need to be further explored accounting for the essential characteristics of RS images.
- *Representative and large-scale RS image datasets with accurate annotations are demanded to narrow the gap between algorithm development and real applications.* An annotated dataset with large volume and variety has proven to be crucial for feature learning [22], [45]–[47]. Although various datasets have been built for different RS image interpretation tasks, there are inadequacies, *e.g.*, the small scale of images, the limited semantic categories, and deficiencies in image diversity, which severely limit the development of new approaches. From another point of view, large-scale datasets are more conducive to characterize the pattern of information distribution in the real-world applications. Thus, it is natural to argue that the representative and large-scale RS image datasets become critical to push forward the development of practical image interpretation algorithms, particularly deep learning-based methods.
- *There is a lack of public platforms for systematic evaluation and fair comparison among different algorithms.* A host of interpretation algorithms have been designed for RS image interpretation tasks and achieved excellent performances. However, many algorithms are designed for specific datasets, rather than practical applications. Without the persuasive evaluation and comparison platforms, it is an arduous task to fairly compare and optimize different algorithms. Moreover, the established image datasets may show deficiencies in scale, diversity and other properties as mentioned before. This makes the learned algorithms inherently deficient. As a result, it is difficult to effectively and systematically measure the validity and practicability of different algorithms for real interpretation applications of RS images.

With these points in mind, this paper first provides an overview to the available RS image datasets and discuss the creation of benchmark datasets for RS image interpretation. Then, we present an example of constructing a large-scale dataset for scene classification as well as the challenges and perspectives in RS image annotation. To sum up, our main contributions are as follows:

- Covering literature published over the past decade, we perform a systematic review of the existing RS image datasets concerning the current mainstream of RS image interpretation tasks, including scene classification, object detection, semantic segmentation and change detection.
- We present some principles, *i.e.*, ***di**versity*, ***r**ichness*, and ***s**calability* (called **DiRS**), on creating benchmark datasets for RS image interpretation. The introduced principles formulate a brief prototype, which we hope to provide a picture for RS image dataset construction with considerations in efficiency, quality assurance, and property assessment.
- Following the principles of dataset construction, we create a large-scale benchmark dataset for RS image scene classification, *i.e.*, **Million-AID**, which possesses millions of RS images. Besides, we conduct a discussion about the challenges and perspectives in RS image dataset annotation to which efforts need to be dedicated in the future work.

The remainder of this paper is organized as follows. Section 2 reviews the existing datasets for RS image interpretation. Section 3 presents the principles of constructing a useful annotated RS image dataset. Section 4 gives an example of constructiong large-scale RS image dataset for scene classification. Section 5 discusses the challenges and perspectives concerning RS image annotation. Finally, in Section 6, we draw some conclusions.

2 Annotated Datasets for RS Image Interpretation: A Review

The interpretation of RS images has been playing an increasingly important role in a large diversity of applications, and thus, has attracted remarkable research attentions. Consequently, various datasets have been built to advance the development of interpretation algorithms for RS images. In this section, we firstly investigate the mainstream of RS image interpretation. And then, a comprehensive review is conducted from the perspective of dataset annotation.

2.1 RS Image Interpretation Focus in the Past Decade

It is of great interest to check what is the main research stream in RS image interpretation. To do so, we analyzed the journal articles published in the past decade in RS community based on Web of Science (WoS) database. Specifically, four journals (*i.e.*, *ISPRS Journal of Photogrammetry and Remote Sensing*, *Remote Sensing of Environment*, *IEEE Transactions on Geoscience and Remote Sensing*, and *IEEE Journal of Selected Topics in Applied Earth Observations and*

TABLE 1: Top 5 Keywords with the Strongest Citation Bursts.

Year	Keywords	Strength	Duration	Year	Keywords	Strength	Duration	Year	Keywords	Strength	Duration
2011	atmospheric correction	2.27		2014	surface	2.08		2017	convolutional neural network	3.19	
	endmember extraction	1.86			scattering model	1.95			deep learning	2.90	
	segmentation	1.63			landslide	1.62			sparse representation	2.31	
	object detection	1.24			extraction	1.59			field	2.27	
	reflectance spectroscopy	1.24			reflectance	1.23			radiative transfer	1.81	
2012	hyperspectral data	2.24		2015	time series	2.17		2018	regression	2.70	
	signal	1.87			hyperspectral imaging	2.10			random forest	2.45	
	backscatter	1.41			remote sensing	2.00			dimentionality reduction	2.20	
	model	1.37			subsidence	1.74			reconstruction	2.20	
	optical property	1.28			water quality	1.74			machine learning	1.91	
2013	object	2.45		2016	inversion	2.96		2019	ground penetrating radar	3.72	
	urban area	2.45			simulation	2.84			water	3.19	
	monitoring	2.04			information	2.44			emission	2.12	
	hyperspectral image	1.85			point cloud	2.37			temperature	1.79	
	algorithm	1.68			restoration	2.28			propagation	1.59	

* The *minimum duration* is 1 and $\gamma = 0.20$ for burstness detection in CiteSpace. A rectangular bar in the *Duration* column denotes the years from 2011 to 2019. The red rectangles indicate the duration time of the keywords with citations bursts.

Fig. 1: Tag cloud of RS image interpretation.

Remote Sensing) were mainly investigated considering their typically high level of academic rigor, peer review process and forefront of research. In total, there are 11, 337 articles were employed as the meta-data for analysis.

To identify the interpretation work of RS images, a *tile/topic/keyword* search was performed using the item, *e.g.*, "interpretation". By excluding the irrelevant results (*e.g.*, review articles), 516 articles were obtained and then analyzed by CiteSpace [48]. Figure 1 shows the highest-frequency terms appearing in the title, keyword, and abstract of the literature. The higher-frequency terms are presented with larger font size. As can be seen from this figure, interpretation works mainly focus on *classification* tasks (*i.e.*, land-cover classification, object detection, and scene classification). *Segmentation* and *change detection* occupy prominent positions in the interpretation tasks. An interesting discovery is that *synthetic aperture radar* is an import part in the field of RS image interpretation. It is worth noting *model* and *algorithm* construction plays an significant role in the interpretation of RS images. These have heavily promoted dataset construction to advance the interpretation performance of RS images.

Table 1 shows the top 5 keywords with the strongest citation bursts during the past decade. The higher the value of *strength*, the more cutting-edge the corresponding research is in the year. From this table it appears that the keywords of frontier topics change year over year. Most notably, *deep learning* and *convolutional neural network* (CNN) have come to be the most prominent features since 2017. This makes sense as deep learning with the represented CNN method have been widely introduced into the field of RS for visual content interpretation and applications [37], [44]. And the subsequent citation burst, *i.e.*, *sparse representation*, confirms its significant role in data-driven interpretation schemes. With the strongest citation bursts, we subsequently filtered the meta articles by "deep learning" and "convolutional neural network". The highest-frequency terms match well with Figure 1, where scene classification, object detection, segmentation as well as change detection possess the centrality of interpretation tasks, and, as a result, the review given below focuses mainly on datasets concerning these topics.

2.2 Annotated Datasets for RS Image Interpretation

During the past years, a number of RS image datasets for RS image interpretation have been released publicly, which can be arranged in chronological order as shown in Tables 2-5. For detailed information about these datasets, the corresponding references listed in the tables can be referred. Different from several previously published surveys that deliver introductions about the datasets, we focus on analyzing the properties of the current RS image datasets from the perspective of annotation.

2.2.1 Categories Involved in Interpretation

The interpretation of RS images aims to extract contents of interest at pixel-, region- and scene-levels. Usually, category information of interested image contents is extracted through elaborately designed interpretation algorithms. Therefore, some datasets were constructed to recognize common RS scenes [36], [49]–[54] in the earlier years. Concentrating on specific object information, there are datasets focusing on one or several main categories [59], [60], [62]–[69], such as vehicle [59], [62], [65], [68], building [60], airplane [63], [69], and ship [69]. The determination of land use and land cover (LULC) categories plays a significant role in urban planning or land-use survey. Hence, several RS image

TABLE 2: Comparison among different RS image scene classification datasets.

Dataset	#Cat.	#Images per cat.	#Images	Resolution (m)	Image size	Year
UC-Merced [49]	21	100	2,100	0.3	256×256	2010
WHU-RS19 [13]	19	50 to 61	1,013	up to 0.5	600×600	2012
RSSCN7 [50]	7	400	2,800	–	400×400	2015
SAT-4 [51]	4	89,963 to 178,034	500,000	1 to 6	28×28	2015
SAT-6 [51]	6	10,262 to 150,400	405,000	1 to 6	28×28	2015
BCS [52]	2	1,438	2,876	–	600×600	2015
RSC11 [53]	11	∼100	1,232	∼0.2	512×512	2016
SIRI-WHU [54]	12	200	2,400	2	200×200	2016
NWPU-RESISC45 [55]	45	700	31,500	0.2 to 30	256×256	2016
AID [35]	30	220 to 420	10,000	0.5 to 8	600×600	2017
RSI-CB256 [56]	35	198 to 1,331	24,000	0.3 to 3	256×256	2017
RSI-CB128 [56]	45	173 to 1,550	36,000	0.3 to 3	128×128	2017
RSD46-WHU [57]	46	500 to 3,000	117,000	0.5 to 2	256×256	2017
EuroSAT [36]	10	2,000 to 3,000	27,000	10	64×64	2018
PatternNet [58]	38	800	30,400	0.06 to 4.7	256×256	2018

* *Cat.* is short for *Category* and the same in the following text.

TABLE 3: Comparison among different RS image object detection datasets.

Datasets	Annot.	#Cat.	#Instances	#Images	Image width	Year
TAS [59]	HBB	1	1,319	30	792	2008
SZTAKI-INRIA [60]	OBB	1	665	9	∼800	2012
NWPU-VHR10 [61]	HBB	10	3,651	800	∼1,000	2014
DLR 3k [62]	OBB	2	14,235	20	5,616	2015
UCAS-AOD [63]	OBB	2	14,596	1,510	∼1,000	2015
VEDAI [64]	OBB	9	3,640	1,210	512/1,024	2016
COWC [65]	CP	1	32,716	53	2,000–19,000	2016
HRSC2016 [66]	OBB	26	2,976	1,061	∼1,100	2016
RSOD [67]	HBB	4	6,950	976	∼1,000	2017
CARPPK [68]	HBB	1	89,777	1,448	1280	2017
LEVIR [69]	HBB	3	11,028	22,000	800	2018
VisDrone [70]	HBB	10	54,200	10,209	2,000	2018
xView [71]	HBB	60	1,000,000	1,413	∼3,000	2018
DOTA-v1.0 [22]	OBB	15	188,282	2,806	800–4,000	2018
HRRSD [72]	HBB	13	55,740	21,761	152–10,569	2019
DIOR [73]	HBB	20	192,472	23,463	800	2019
DOTA-v1.5 [22]	OBB	16	402,089	2,806	800–13,000	2019
DOTA-v2.0 [22]	OBB	18	1,488,666	11,067	800–20,000	2020

* *Annot.* refers to the *Annotation* style of instances, *i.e.*, *HBB* (*horizontal bounding box*) and *OBB* (*oriented bounding box*). *CP* refers to the annotation with only the *center point* of an instance.

semantic segmentation datasets concern specific land-cover categories like building and road [74]–[76]. Also, there are datasets that concern multiple land-cover categories within specific areas, e.g., city areas [77]–[80]. Even with accurate annotation of category information, these datasets are with relatively small numbers of interpretation categories, which can only be used for content interpretation when certain specific objects are concerned.

It is obvious that the above mentioned datasets prefer to advance interpretation algorithms with limited semantic categories. To avoid this situation, some datasets were annotated with dozens of semantic categories of interest, such as NWPU-RESISC45 [55], AID [35], RSI-CB [56], RSD46-WHU [57], Patternet [58], SEN12MS [81] and SCDN [82]. However, more categories of semantic contents are encountered in practical RS applications. For example, there are many categories and hundreds of fine-grained classes in real LULC applications. As a result, datasets with a limited number of scene categories are not able to extract the various and complex semantic contents reflected in RS images. Moreover, categories in these datasets are set equal while the relationship between different interpretation categories is ignored. Particularly, the intra-class and inter-class relationships are simply neglected by dataset creators. This inevitably results in the chaotic category organization and management for semantic information mining. Therefore, how to annotate datasets with rich semantic categories and reasonable relationship organization strives to be a key problem for practical dataset construction.

2.2.2 Dataset Annotation

our knowledge, nearly all the created datasets as listed in Tables 2-5 are manually annotated by experts. Generally, dataset annotation attempts to assign scenes, objects or pixels of interest with semantic tags in images. For the task of scene classification, a category label is typically assigned to the scene components by visual interpretation of experts [35], [55]. In order to recognize specific objects, entities in images are usually labeled with closed areas. Thus, many existing datasets, such as NWPU-VHR10 [61], RSOD [67], xView [71], HRRSD [72], and DIOR [73], manually annotate objects in bounding boxes. Another fundamental issue is the acquisition of target RS images in which the intriguing contents are contained. Usually, target images are manually searched, distinguished and screened in the image database (*e.g.*, Google Earth) by trained annotators. Along with the subsequent label assignment, the whole annotation process in the construction of RS image datatsets is time-consuming and labor-intensive, especially those for specialized applications [87]. As a result, dataset construction, form source image collection and filtering to semantic information annotation and quality review, rely heavily on manual operations, which make it an expensive project. This raises an urgent demand for developing more efficient and assistant strategies of dataset annotation to lighten the burden of artificial annotation.

TABLE 4: Comparison of different RS image semantic segmentation datasets.

Datasets	#Cat.	#Images	Resolution (m)	#Bands	Image size	Year
Kennedy Space Center [83]	13	1	18	224 bands	512×614	2005
Botswana [83]	14	1	30	242 bands	1476×256	2005
Salinas [78]	16	1	3.7	224 bands	512×217	–
University of Pavia [77], [78]	9	1	1.3	115 bands	610×340	–
Pavia Centre [78]	9	1	1.3	115 bands	1096×492	–
ISPRS Vaihingen [79]	6	33	0.09	IR,R,G,DSM,nDSM	∼2,500×2,500	2012
ISPRS Potsdam [79]	6	38	0.05	IR,RGB,DSM,nDSM	6,000×6,000	2012
Massachusetts Buildings [74]	2	151	1	RGB	1,500×1,500	2013
Massachusetts Roads [74]	2	1,171	1	RGB	1,500×1,500	2013
Indian Pines [84]	16	1	20	224 bands	145×145	2015
Zurich Summer [80]	8	20	0.62	NIR, RGB	1,000×1,150	2015
Inria Dataset [75]	2	360	0.3	RGB	1,500×1,500	2017
EVlab-SS [85]	10	60	0.1 to 2	RGB	4,500×4,500	2017
RIT-18 [86]	18	3	0.047	6 bands	9,000×6,000	2017
WHU Building-Aerial Imagery [76]	2	8,189	0.3	RGB	512×512	2019
WHU Building-Satellite Imagery I [76]	2	204	0.3 to 2.5	RGB	512×512	2019
WHU Building-Satellite Imagery II [76]	2	17,388	2.7	RGB	512×512	2019
So2Sat LCZ42 [87]	17	400,673	10	10 bands	32×32	2019
SEN12MS [81]	33	180,662 triplets	10 to 50	up to 13 bands	256×256	2019
UAVid [88]	8	420	–	RGB	∼4,000×2,160	2020
GID [5]	15	150	0.8 to 10	4 bands	6,800×7,200	2020

* The UAVid consists of 30 video sequences captured by unmanned aerial vehicle and each sequence is annotated by every 10 frames, resulting in 420 densely annotated images.

TABLE 5: Comparison of different RS image change detection datasets.

Datasets	#Cat.	#Image pairs	Resolution (m)	#Bands	Image size	Year
SZTAKI AirChange [89]	2	13	1.5	RGB	952×640	2009
AICD [90]	2	1000	0.5	115 bands	800×600	2011
Taizhou Data [91]	4	1	30	6 bands	400×400	2014
Kunshan Data [91]	3	1	30	6 bands	800×800	2014
Yancheng [92]	4	2	30	242 bands	400×145	2018
Urban-rural boundary of Wuhan [93]	20	1	4/30	4/9 bands	960×960	2018
Hermiston City area, Oregon [94]	5	1	30	242 bands	390×200	2018
OSCD [95]	2	24	10	13 bands	600×600	2018
Quasi-urban areas [96]	3	1	0.5	8 bands	1,200×1,200	2018
WHU Building-Change Detection [97]	2	1	0.2	RGB	32,207×15,354	2018
Season-varing Dataset [98]	2	16,000	0.03 to 0.1	RGB	256×256	2018
ABCD [99]	2	4,253	0.4	RGB	160×160	2018
HRSCD [100]	6	291	0.5	RGB	10,000×10,000	2019
MtS-WH [101]	9	1	1	NIR, RGB	7,200×6,000	2019
LEVIR-CD [102]	2	637	0.5	RGB	1,024×1,024	2020
SCDN [82]	30	4,214	0.5 to 3	RGB	512×512	2020

When it comes to the annotation tools, there is a lack of visualization method for the annotation of large sale and hyper-spectral RS images. Currently, tagging instruments designed for natural images, *e.g.*, LabelMe [103] and LabelImg [104], are introduced to annotate RS images. Those annotation tools typically visualize a image with limited scale. However, different from natural images, RS images taken from the bird-view are with large scale and wide geographic coverage. Thus, the annotator can only perform the labeling operations within a local region of the RS image, which is easy to make influence on the annotator to grasp the global content of the RS image and cause inaccurate annotation. Meanwhile, the image roam process will inevitably constrain the annotation efficiency. This problem is particularly serious when facing the annotation for semantic segmentation and change detection tasks where labels are usually assigned semantically pixel-by-pixel [79], [84], [88], [89]. On the other hand, hyper-spectral RS images [77], [79], [83], [84], [92]–[95] which characterize objects with rich spectral signatures, are usually employed for semantic content interpretation. However, it is hard to label the hyper-spectral RS images since annotation tools developed for natural images are not able to present hyper-spectral information for visualization and accurate semantic annotation. Therefore, universal annotation tools are pressing needed to be developed for fast and convenient semantic annotation especially for the large scale and hyper-spectral RS images.

2.2.3 *Image Source*

A wide group of RS images has been employed as the source of various interpretation datasets, including the optical, hyper-spectral, SAR images. Typically, the optical images from Google Earth are widely employed as the data standard, such as those for scene classification [35], [49], [50], [53]–[55], [57], [58], object detection [21], [22], [59], [63], [66], [67], [69], [73], and pixel-based understanding [74], [75], [80]. In these scenarios, RS images are typically interpreted by the visual contents, of which the spatial pattern, texture structure, information distribution as well as organization mode are more concerned. Although the Google Earth images are post-processed with RGB formats using the original optical aerial images, they possess potential for pixel-based LULC interpretation as there is no general statistical difference between the Google Earth images and the optical aerial images [105]. Thus, the Google Earth images can also be used as aerial images for evaluating interpretation algorithms [35].

Different from the optical RS image dataset, the construction of hyper-spectral and SAR image datasets should adopt the data with original forms. Compared to optical images, multi-/hyper-spectral images can capture the essential characteristics of ground features in detail which simultaneously

involves spectral and spatial information. Therefore, the content interpretation of hyper-spectral RS images is mainly based on the spectral properties of ground features. Naturally, this kind of images are typically employed to construct dataset for subtle semantic information extraction, such as semantic segmentation [77], [79], [81], [83], [84], [87] and change detection [34], [90], [92]–[96], in which more attention are paid to the knowledge of the fine-gained compositions. For SAR images acquired by microwave imaging, the content interpretation is usually performed by the radiation, transmission, and scattering properties. Hence, SAR images are employed for abnormal object detection by utilizing the physical properties of ground features. To this end, it is not encouraged to employ the modified data of SAR images for the interpretation of interested content. But most importantly, images in a constructed dataset should be selected from the real-world application scenarios. Otherwise, the employed image is not able to reflect the difficulties among the real interpretation tasks, which will inevitably cause weak generalization ability of the learned interpretation algorithms.

2.2.4 Dataset Scale

A large number of RS image datasets have been constructed for various interpretation tasks at different semantic levels. However, many of them are with small scales, reflected in aspects like the limited number, small size, and lacked diversity of annotated images. On the one hand, the size and number of images are important properties concerning the scale of a RS image dataset. RS images that typically taken from the bird-view perspective have the large geographic coverage and thus possess large image size. For example, an image from GF-2 satellite usually exceeds $30,000 \times 30,000$ pixels. However, many of the current datasets employ the chipped images, usually with the width/height of a few hundred pixels as shown in Tables 2-5, to fit specific models which are generally designed to extract features within limited visual space of images. In fact, preservation of the original image size is more close to the real-world applications [5], [22]. Some datasets with larger image sizes, say, width/height of a few thousand pixels, are limited with the number of annotated images or categories [65], [86], [98], [100]–[102]. Furthermore, quite a few datasets only contain one or several images, especially those for semantic segmentation [77], [83] and change detection [34], [92]–[94], [96], [97], [101]. As a result, the scale limitations in size and number of images could easily lead to performance saturation for interpretation algorithms.

On the other hand, due to the constraint of data scale, the existing datasets often show deficiencies in image variation and sample diversity. Typically, content in RS images always show differences with the change of spatio-temporal attributes while images in some of the datasets are selected from local areas or with limited imaging conditions [52], [57], [62], [83]. In addition, contents reflected in RS images are with complex textural, structural and spectral features owing to the high complexity of the earth's surface. Thus, datasets with limited images and samples [49], [53], [60], [65], [83], [86] are usually not able to completely characterize the properties of interested objects. As a result, there is a lack of content representation of real-world scenarios for datasets with small scales, causing weak generalization ability of interpretation algorithms. Furthermore, constrained by the scale of dataset, the currently popular deep learning approaches usually pre-train the models with the large-scale natural image datasets, *e.g.*ImageNet [45], for RS image interpretation [106], [107]. Nevertheless, features learned by this strategy are hard to completely adaptive to RS data because of the essential difference between RS images and natural images. For instance, Dramatic change of object orientation is common to be observed in RS images. All of these raise an urgent demand for annotating large-scale RS dataset with abundant and diverse images for the advancement of RS image interpretation methods.

3 Principles to Build RS Image Benchmarks

The availability of a shiny RS image dataset has been shown critical for effective feature learning, algorithm development, and high-level semantic understanding [45]–[47], [108], [109]. More than that, the performance of almost all data-driven methods rely heavily on the training dataset. However, constructing a large-scale and meaningful image dataset for RS interpretation issues is not an easy job, at least from the points of technology and cost factors. The challenge lies largely in the aspects of efficiency and quality control, which make it difficult to manage each practical step in the dataset construction process. The absence of systematic work involving these problems has largely limited the construction of useful datasets and continuous advancement of interpretation algorithms in the RS community. Therefore, it is valuable to explore the feasible scheme for creating a practical RS dataset. We believe that the following introduced aspects can be taken into account when creating a desirable dataset for RS image interpretation.

3.1 Principles for Dataset Construction

The primary principle to construct a meaningful RS image dataset is that the dataset should be created on the basis of the requirements of practical applications rather than the characteristics of algorithms to be employed. Essentially, the creation of RS image dataset should be aimed at model training, testing, and screening for practical applications. It is of great significance to get a whole picture of a designed interpretation model before it is poured into practical applications. Thus, the reliable benchmark dataset becomes critical to comprehensively verify the validity of the designed interpretation models and eventually eliminate those with weak interpretation capability. To this end, it requires the created dataset to consist of rich annotated samples, *e.g.*, variation in background, scale, and imaging conditions, that cover the practical application scenarios.

From these points of view, the annotation of RS image dataset is better to be conducted by the application sides rather than the algorithm developers. Basically, annotations by algorithm developers will inevitably possess bias as they are more familiar with the algorithm properties and lack of understanding of the challenges in practical applications. As a result, the annotated dataset by developers will be algorithm-oriented. On the contrary, the application sides have more opportunities to access the real application scenarios, and thus, are more familiar with the challenges lying

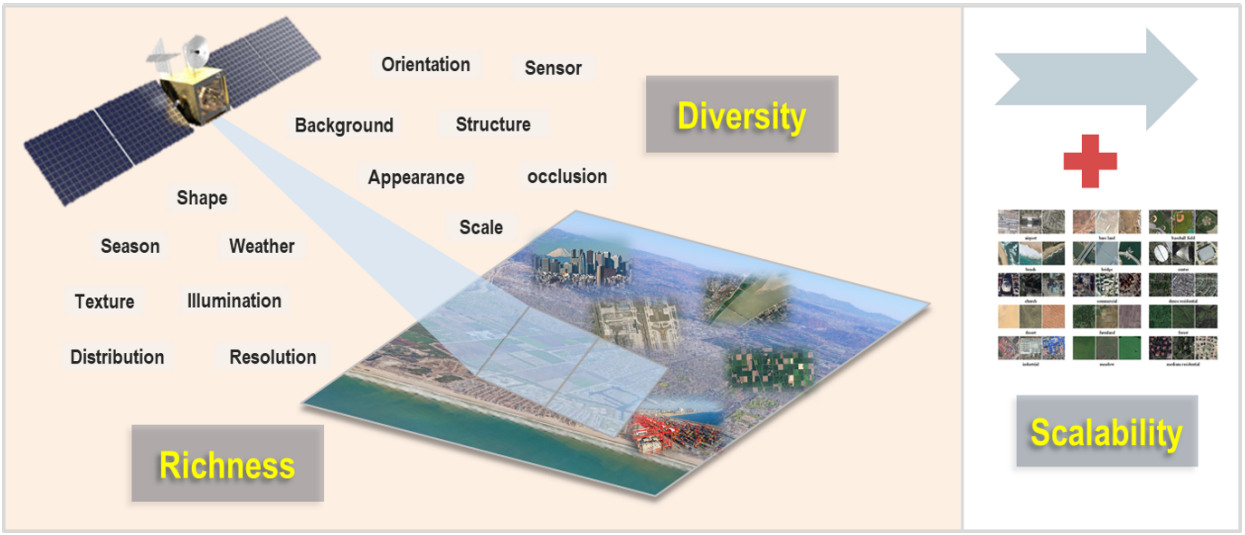


Fig. 2: The ***DiRS*** principles: ***di****versity*, ***r****ichness*, and ***s****calability*. DiRS formulates the basic principles which can be employed as the guidance to the construction of RS image interpretation dataset. We believe that these principles are complementary to each other. That is, the improvement of dataset in one principle can simultaneously promote the dataset quality reflected in other principles.

in the interpretation tasks. Consequently, dataset annotation by application sides is naturally application-oriented, which is more conducive to enhance the practicability of the interpretation algorithm.

Generally, the RS image dataset should be constructed toward the real-world scenarios, not the specific algorithms. On this basis, it is possible to feed the interpretation system with high-quality data and make the interpretation algorithms effectively learn and even extend knowledge that people desired. With these points in mind, we believe that the ***di****versity*, ***r****ichness*, and ***s****calability* (called ***DiRS***), as illustrated in Figure 2, could be considered as the basic principles when creating benchmark datasets for RS image interpretation.

3.1.1 Diversity

A dataset is considered to be diverse if its images depict various visual characteristics of the relevant semantic contents with a certain degree of complementarity. From the perspective of within-class diversity, annotated samples with large diversity are able to represent the real-world content distribution more comprehensively. To this end, it is better that each annotated sample could reflect different attributes for the content of the same class rather than the repeated characteristics. Thus, the within-class samples of large diversity are more conducive for an algorithm to learn the essential characteristics for specific class recognition. For example, objects of the same category (*e.g.*, vehicle) usually differences in appearance, scale, and orientation that diversify the instances. These diverse characteristics could provide insurance to train detectors with the more powerful ability of feature representation and model generalization.

On the other hand, in order to learn an interpretation algorithm for the effective discrimination of similar classes, the between-class similarity should also be taken into consideration when constructing the RS image dataset. For this requirement, more fine-grained classes with high semantic overlapping should be contained in the created dataset. It is easy to understand that the notable intervals of content features can enable an interpretation model to learn to distinguish the image content of different categories effortlessly. In contrast, between-class similarity means the small distance of different classes, which will put forward higher requirements for interpretation models to discriminate semantically similar content. Generally, the within-class diversity and between-class similarity offer a guarantee for feature complementarity, which is crucial for constructing datasets with a large diversity. On this basis, it is able to train an interpretation model to adapt to the real-world scenarios.

3.1.2 Richness

In addition to the diversity, the richness of a dataset is another significant determinant for learning the interpretation model with strong stability. Specifically, the rich image variation, various content characteristics, and large-scale samples should be considered as the desired properties when constructing an interpretation dataset. However, most of the current datasets as summarized in Tables 2-5, to some extent, show deficiencies in these aspects. To ensure the rich image variation of a dataset, images can be collected under various circumstances, such as the weather, season, illumination, imaging condition, and sensor, which will allow the dataset to possess rich variations in translation, viewpoint, object pose and appearance, spatial resolution, illumination, background, occlusion, *etc*. Not only that, images collected from different periods and geographic regions will also endow the dataset with various attributes concerning spatio-temporal distribution.

Moreover, different from natural images that are usually taken from horizontal perspectives with narrow extents, RS images are taken with bird-views, endowing the images with large geographic coverage, abundant ground features, and complex background information. Faced with situation, the images in an interpretation dataset is better to contain contents with variety of characteristics, varying in the geometrical shape, structure characteristic, textural attribute,

etc. From this point of view, the constructed dataset should consist of large-scale images and sufficient annotated samples to ensure strong representation power for the real-world scenarios. The reality is that insufficient images and samples are more likely to lead to the over-fitting problem, particularly for data-driven interpretation algorithms (*e.g.*, CNN frameworks). In general, the interpretation models built upon the dataset in accordance with the above lines are able to possess more powerful representation and generalization ability for practical applications.

3.1.3 Scalability

Scalability can be a measure of the ability to extend a constructed dataset. With the increasingly wide applications of RS images, the requirements for a dataset usually change along with the specific application scenarios. For example, a new category of scene may need to be differentiated from the collected categories with the change of land-cover and land-use. Thus, the constructed dataset must be organized with sufficient category space to involve the new category scenes while keeping the existing category system extensible. Not only that, but the relationship among the annotated features is also better to be well managed according to the real-world situation. That is, an interpretation dataset is better to be flexible and extendable, considering the change of application scenarios.

Notably, there is a large number and diversity of remote sensing images received every day, which need to be efficiently labeled with valuable information to play the value of applications. To this end, the organization, preservation, and maintenance of annotation and images are of great significance to be performed for the scalability of a dataset. Besides, it would be preferable if the newly annotated images could be involved in the constructed dataset effortlessly. With these considerations, a constructed RS image dataset with excellent scalability can be conveniently adapted to the changing requirements of real-world applications without impacting its inherent accessibility, thereby assuring continuing availability even as modifications are made.

3.2 Coordinates Collection

The acquisition of RS images formulates the foundation of creating an interpretation dataset. Benefiting from the spatial property possessed by the RS images, the candidate RS images can be accessed by utilizing their inherent information of geographic coordinates. Typically, this operation is performed to prepare a public optical RS image dataset, such as by utilizing the map application interface, open source data, and public geodatabases. The coordinates collection may not be the optimal strategy but can also be employed as a reference when creating a private dataset or dataset in which images are from other sensors.

3.2.1 Map Search Engines

A convenient way to collect RS images is to utilize public map search engines, such as Google Map[1], Bing Map[2], Baidu Map[3], and World Map[4]. As common digital map service solutions, they provide satellite images covering the whole world with different spatial resolutions. Many existing RS datasets, such as the AID [35] for scene classification, DOTA [22] for object detection, have been built based on Google Map. While collecting RS images on such map search engines, the developed map API (application programming interface) can be utilized to extract images and acquire the corresponding semantic tags. Based on the rich positional data composed of millions of point, line and region vectors that contain specific semantic information, the large amount of candidate RS images can be collected through these map engines. For example, by searching "airport" on Google Earth, all searched airports in a specific area will be indicated with specific geographic positions. The corresponding satellite images can then be accessed using the coordinates of search results and the acquired satellite images can be used to build airport scene and aircraft object samples.

3.2.2 Open Source Data

Open source geographic data is usually established upon the global positioning system (GPS) information, aerial photography images, other free contents and even local knowledge (such as social media data) from users. Open source geographic data, such as Open Street Map (OSM) and WikiMapia, are generally created upon the collaboration plan which allows users to label and edit the ground feature information. Therefore, the open source geographic data can provide rich semantic information that is timely updated, low cost and has a large amount in quantity compared with the manual collection strategy for RS images [56]. With the abundant geographic information provided by various open source data, we are able to collect elements of interest like points, lines and regions with specific geographic coordinates and then match the collected geographic elements with corresponding RS images. Moreover, the extracted geographic elements of interests can be aligned with temporal RS images which can be downloaded from different map engines as described above. With these advantages and operations, it is possible to collect large-scale RS images of great diversity for dataset construction.

3.2.3 Geodatabase Integration

Different from the collection of natural images, which can be conveniently accessed through web crawling, search engines (*e.g.*, Google Images search) and sharing databases (*e.g.*, Instagram, Flickr), the acquisition of RS images is difficult because of the high cost. Nevertheless, the public geodatabases released by the state institutions and communities usually provide accurate and complete geographic data. With this facility, the geographic coordinates of scenes that belong to the specific labels can also be obtained based on these databases. For example, the National Bridge Inventory (NBI) Bridges database presents detailed information of the bridges, including the geographic locations, length, material, and so on. Benefiting from this advantage, we can extract a large number of coordinates of bridges for

1. https://ditu.google.com
2. https://cn.bing.com/maps
3. https://map.baidu.com
4. http://map.tianditu.gov.cn

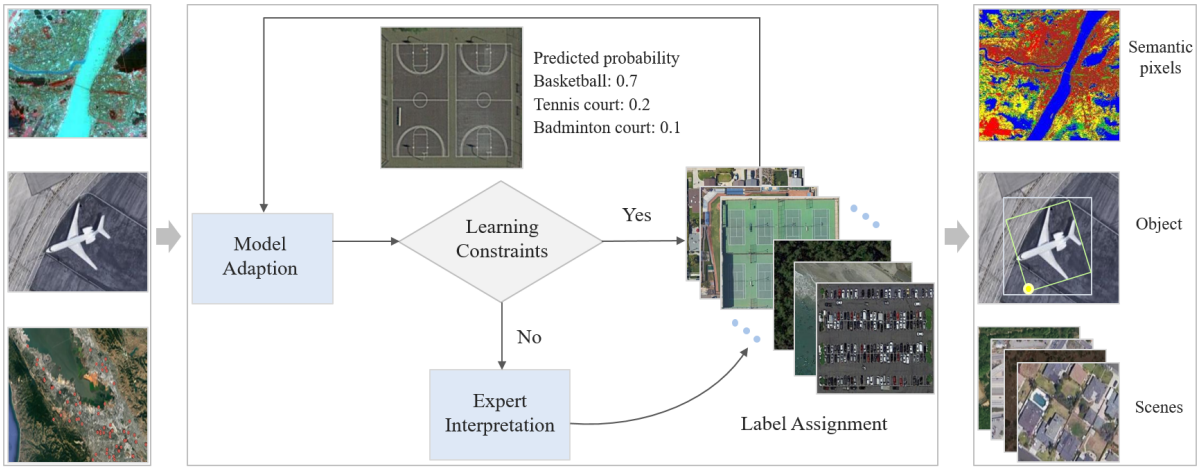


Fig. 3: General workflow of Semi-automatic annotation for RS images.

the collection of corresponding scene images. Generally, by integrating these kinds of public geodatabases, we are able to obtain the geographic coordinate information of scene images with categories, and thus, efficiently collect a large number of scene images with relatively low cost.

3.3 Annotation Methodology

With the collected images for a specific interpretation task, annotation is performed to assign specific semantic labels to the elements of interest contained in the images. Next, the common image annotation strategies will be introduced.

3.3.1 Annotation Strategies

Depending on whether human intervention is involved, the solutions to RS image annotation can be classified into three types: manual, automatic, and interactive annotation.

- **Manual Annotation** The common way to create an image dataset is the manual based annotation strategy. The great advantage of the manual annotation is its high accuracy because of the fully supervised annotation process. Based on this consideration, many RS image datasets have been manually annotated for various interpretation tasks, such as those for scene classification [35], [49], [53], [54], object detection [22], [60], [64] and semantic segmentation [5], [88]. Regardless of the source from which the natural or RS images are acquired, the way to annotate contents in images is similar and many task-oriented annotations tools have been built to relieve the monotonous annotation process. Thus, image annotation tools developed for natural images can be further introduced for RS images to pave the way for the cost-effective construction of large-scale annotated datasets. A resource concerning to image annotation tools will be introduced in Section 5.
  In practice, constructing a large-scale image dataset by manual scheme is laborious and time-consuming as introduced before. For example, a number of people spent several years to construct the ImageNet [45]. To relieve this problem, crowd-sourcing annotation becomes an alternative strategy that can be employed to create a large-scale image dataset [46], [71] while paying efforts to its challenge with quality control. Besides, benefiting from excellent ability of image content interpretation, annotators also resort to machine learning schemes [110], [111], which are integrated for preliminary annotation, to speed up the efficiency of manual annotation.
- **Automatic Annotation** In contrast to natural images, RS images are often characterized with diverse structures and textures because of spectral and spatial variation. It is difficult to annotate semantic contents for annotators without domain knowledge. As a result, the manually annotated dataset is prone to have bias problem because of annotators' difference in domain knowledge, educational background, labelling skill, life experience, etc. To this situation, automatic image annotation methods are naturally employed to alleviate annotation difficulties and further reduce the cost of manual annotation [112].
  Automatic methods reduce the cost of annotation by leveraging learning schemes [113]–[118]. In this strategy, a certain number of images are initialized to train an interpretation model, including the supervised methods [119] and weakly supervised methods [120]–[122]. The candidate images are then poured into the established model for content interpretation and the interpreted results finally serve as annotation information. As RS images are characterized with complex contents of large geographic extent [121], iterative and incremental learning [123] can be employed to filter noisy annotation and enhance the generalization ability of annotation, including scene classification [124], object detection [125], semantic segmentation [121], [126]. Generally, the key of the automatic image annotation is to narrow the semantic gap between low-level visual features and high-level semantic concepts, *i.e.*, to learn high-level semantic labels from visual features by exploring strong and stable image content descriptors [118]. Nevertheless, one disadvantage of automatic annotation is that the generalization ability can be affected by both the quality of the initial

candidate images and the capability of models to retain images from different distributions, resulting in a weak domain adaption ability.

- **Interactive Annotation** In the era of big RS data, annotation with human-computer interaction, which falls in semi-automatic annotation, could be a more practical scheme considering the demand for high-quality and efficient annotation for real-world applications. In this scheme, an initial interpretation model can be trained with the existing datasets and then employed to annotate the unlabeled RS images. Besides, the performance of an annotation model can be improved greatly with intervention of the annotator [127]. The annotator intervention can be in the form of relevance feedback or identification of the relevant contents in the images to be annotated. Generally, the overall performance of the interpretation algorithms mostly depends on the time spent on creating annotations [128].
  Typically, by employing active learning strategy [129] and setting restrict constraints, those images that are difficult to be interpreted can be screened out and then annotated by manual method. The feedback received can be used to purify the annotation model through a learning loop way. Consequently, a large volume of annotated images can be acquired to retrain the interpretation model and further boost the annotation task in an iterative way. With the iteration process, the number of images to be annotated will be greatly reduced to relieve annotation labor. The general workflow of semi-automatic image annotation is shown in Figure 3. Benefiting from the excellent feature learning ability, deep learning based methods can be developed for image semantic annotation with significant quality and efficiency improvement [111]. Instead of full image annotation, human intervention by simple operations, *e.g.*, point-clicks [130], boxes [131], and scribbles [132], can significantly improve the efficiency of interactive annotation. Based on the semi-automatic annotation strategy, a large-scale annotated RS image dataset can be constructed efficiently and also with quality assurance owing to the involvement of human intervention.

3.3.2 Quality Assurance

The dataset with high annotation quality is important for the development of effective algorithms and the evaluation of interpretation performance. The following introduced strategies can be employed for the quality control when creating an RS image dataset.

- **Rules and Samples** The annotation rules without ambiguity are the foundation to create a high-quality dataset. The rules involve the category definition, annotation format, viewpoint, occlusion, image quality, etc. For example, whether to consider the fine-grained classes, whether to exclude the objects in occlusion, whether to annotate the extremely small objects. If there are no specific rule instructions, different annotators will annotate the data with their individual preferences [47]. For annotation in RS images, it is difficult for annotators to recognize the categories of ground elements if they have no professional backgrounds. Therefore, samples are better to be annotated by experts in the field RS image interpretation and then presented to annotators.
- **Training of Annotators** Each annotator is required to pass the test of annotation training. Specifically, each annotator is given a small part of the data and asked to annotate the data meeting the articulated requirements. Those annotators that failed to pass the test are not invited to participate in the later annotation project. With such design, dataset builders are able to build an excellent annotation team. Take xView [71] as an example, the annotation accuracy of objects is vastly improved on annotation accuracy with trained participants. Therefore, the training of annotators could be a reliable guarantee for high-quality image dataset annotation.
- **Multi-stage Pipeline** A serial of different annotation operations is easier to cause fatigue and result in errors. To dismiss this effect, the multi-stage pipeline of image annotation can be designed to decouple the difficulties of the annotation task. For example, the annotation of object detection can be decoupled to be spotting, super-category and sub-category recognition [46]. Through this kind of disposal, each annotator only needs to focus on one simple stage during the whole annotation project and the error rate can be effectively decreased.
- **Grading and Reward** A comprehensive study of annotators based on a survey and annotated images can be performed with prepared referencing images. Besides, an analysis of annotators' behavior, *e.g.*, the required time per annotation step and the amount of annotation over a period of time, can be conducted to assess the potentially weak annotations. Thus, different types of annotators can be identified, *i.e.*, spammers, sloppy, incompetent, competent and diligent annotators [133]. Then, incentive mechanism (*e.g.*, financial payment) can be employed to reward the excellent annotators and eliminate the inferior labels from unreliable annotators.
- **Multiple Annotations** A feasible measurement to guarantee high-quality image annotation is to obtain multiple annotations from different annotators, merge the annotations and then utilize the response contained in the majority of annotations [45]. To acquire high-quality annotations, majority voting can be utilized to merge multiple accurate annotations [134]. One disadvantage of this approach is that multiple annotations require more annotators and it is not reliable if the majority of annotators produce low-quality annotations.
- **Annotation Review** Another effective method to ensure the annotation quality is to introduce a review strategy, which is usually integrated among other annotation pipelines when creating a large-scale image dataset [103]. Specifically, additional annotators can be invited to conduct peer review and rate the quality of the create annotations. Besides, further review

work can be conducted by experts with professional knowledge. Based on the reviews of different levels of supervisors in each annotation step, the overall annotation quality can be strictly controlled through the whole annotation process.

- **Spot Check and Assessment** To check the annotation quality for application, a test dataset can be extracted from the annotated images. Also, gold data is created by sampling and labeling a proper proportion of images from the dataset by experts who are professional imagery analysts. Then, one or several interpretation models can be trained based on these datasets and the evaluation metric (*e.g.*$Recall$ and $Precision$ in object detection [21], [22]) is calculated by comparing the annotations produced by annotator and the gold data. If the evaluation metric is lower than a preset threshold. The corresponding annotations of that annotator would be rejected and required to be resubmitted for annotation.

4 An Example: Million-AID

Limited by the scale of scene images and the number of scene categories, current datasets for scene classification are far from meeting the requirements of the real-world feature representation and the scale for interpretation model development. It is desperately expected that there is a much larger image dataset for scene classification in the RS community. Based on the proposed principles to build RS image benchmarks, we present the following approaches to construct the *Million Aerial Image Dataset for Scene Classification* (named *Million-AID*), which will be released publicly.

4.1 Scene Category Organization

4.1.1 Main challenges

Benefiting from the advancement of RS technologies, the accessibility of various RS images has been greatly improved. However, the construction of a large-scale scene classification dataset for RS images still faces challenges in the aspects of category determination and organization of scene taxonomy. The determination of scene categories is of great significance to construct a high-quality and large-scale RS image dataset for scene classification. Undeniably, a complete taxonomy of RS image scenes should have wide coverage of categorical space since there are a large number of semantic categories in practical applications, *e.g.*, LULC. Existing datasets, such as UCM [49], RSSCN7 [50] and RSC11 [53], contain limited scene categories, which make the datasets not sufficiently representative for the complex contents reflected by RS scenes. Consequently, the datasets with category inadequacy are of weak generalization and inevitably curbed when facing real-world applications.

On the other hand, the excellent organization of scene categories has become an important feature for scalability and continuous availability of a large-scale RS image dataset. Typically, the semantic categories which are closely related to human activities and land utilization are selected manually for the construction of RS scene categories. Because of the complexity of RS image contents, there is a large number of semantic categories and also a hierarchical relationship among different scene categories. Usually, it is difficult to completely cover all the semantic categories and the relationship information between different scene categories can be easily neglected owing to the subjectivity of category selection from dataset builders. Therefore, effective organization of scene categories should be of great significance to construct a RS image dataset of high quality and scalability.

4.1.2 Category network

Faced with the above challenges, we build a hierarchical network to manage the categories of RS image scenes, as shown in Figure 4. To satisfy the requirements of practical application, we construct the scene category system by referencing to the land-use classification standards of China (GB/T 21010-2017). Taking the inclusion relationships and content discrepancies of different semantic scene categories into consideration, the hierarchical category network is finally built with three semantic layers. In accordance with the semantic similarity, those categories with overlapping relationships are merged into a unique category branch. Thus, the scene classification dataset can be constructed with category independence and semantic coverage completeness.

As can be seen from Figure 4, the proposed category network is established upon a multi-layered structure, which provides scene category organization with different semantic levels. When it comes to the specific categories, we extract aerial images on Google Earth and determine whether the images can be assigned with the semantic scene labels in the category network. For those images that cannot be recognized with specific categories within the existing nodes, new category nodes will be embedded into the original category network by experts according to the image scene contents. In view of the fact that there are inclusion relationship among different scene categories, all classes are hierarchically arranged in a three-level tree: 51 leaf nodes fall into 28 parent nodes at the second level, and the 28 parent nodes are grouped into 8 nodes at the first level, representing the 8 underlying scene categories of agriculture, commercial industrial, public service, residential, transportation, unutilized land, and water area. Benefiting from the hierarchical structure of category network, the category labels from the corresponding parent nodes can be directly assigned to the images, so that each image will possess semantic labels with different category levels. This mechanic also provides potentiality for dataset construction and scene classification at flexible category levels.

The category definition and organization can be conveniently achieved by the proposed hierarchical category network. The synonyms of the category network are usually relevant to practical application particularly for LULC and hardly need to be purified. One of the most prominent assets of the category network lies in its semantic structure, *i.e.*its ontology of concepts. Hence, the category network is usually comprehensive enough for class expansion. That is, a new scene category can be easily embedded into the built category network as a new branch of synonym. The established category hierarchy can not only serve as the category standard for the Million-AID dataset but also provides a reliable unified system for future dataset construction.

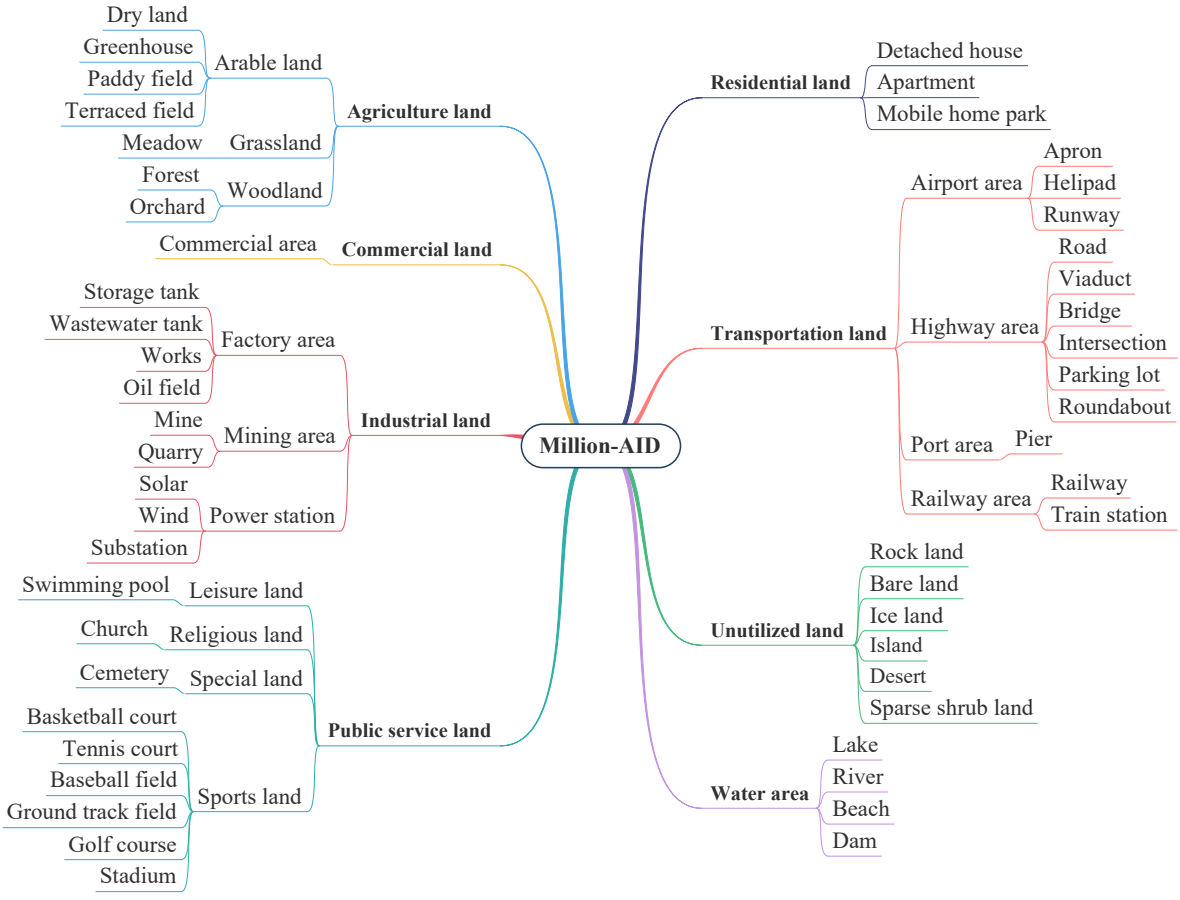


Fig. 4: The hierarchical scene category network of Milion-AID. All categories are hierarchically organized in a three-level tree: 51 leaf nodes fall into 28 parent nodes at the second level which are grouped into 8 nodes at the first level, representing the 8 underlying scene categories of agriculture land, commercial land, industrial land, public service land, residential land, transportation land, unutilized land, and water area.

4.2 Semantic Coordinates Collection

The traditional pipeline of constructing a scene classification dataset is to manually search RS images according to specific semantic categories in the target area and then collect the image blocks attached with specific scene categories. However, finding the target image area with given semantic category scenes is a time-consuming procedure and usually requires high-level technical expertise. Besides, in order to ensure the reliability of scene information, images need to be labeled by specialists with domain knowledge of RS image interpretation. To tackle this problem, we consider making use of public maps and geo-information resources to collect and label RS scene images. With the rapid development of geographic information and RS technologies, there are rich and public geographic data available, such as online map, open source and agency published data. Typically, the publicly geographic data present the surface features in forms like point, line, and plane, which describe the semantic information of ground objects and carries corresponding geographic location information. Based on the publicly geographic data, we search for coordinates of specific semantic tags, and then utilize the semantic coordinates to collect the corresponding scene images.

In RS images, scenes are presented with different geometric appearances. Therefore, different methods are employed to acquire the labeling data which indicates different forms of the ground features. Google Map API and publicly available geographic data are employed to obtain the coordinates of point features while OSM API is mainly utilized to acquire the coordinates of line and plane features. The acquired feature coordinates are then integrated into block data which presents the scene extent. Finally, the block data are further processed to obtain scene images which will be labeled with the attached semantic information.

4.2.1 *Point coordinates*

The point ground features, such as tennis courts, baseball fields, basketball courts, and wind turbines, take relatively small ground space in the real-world. The online Google map makes it possible to discover the world with rich location data for over 100 million places. This provides a powerful function of searching semantic tags of the ground objects. Therefore, we develop a semantic tag search tool based on the Google Map API to retrieve the corresponding coordinates of point data. With the customization search tool, we input semantic tags to retrieval corresponding point objects as using the online map search engine and obtain the geographic coordinates that match the semantic information within a certain range. The retrieved point results attached with location information, *i.e.*, geographic coordinates, are

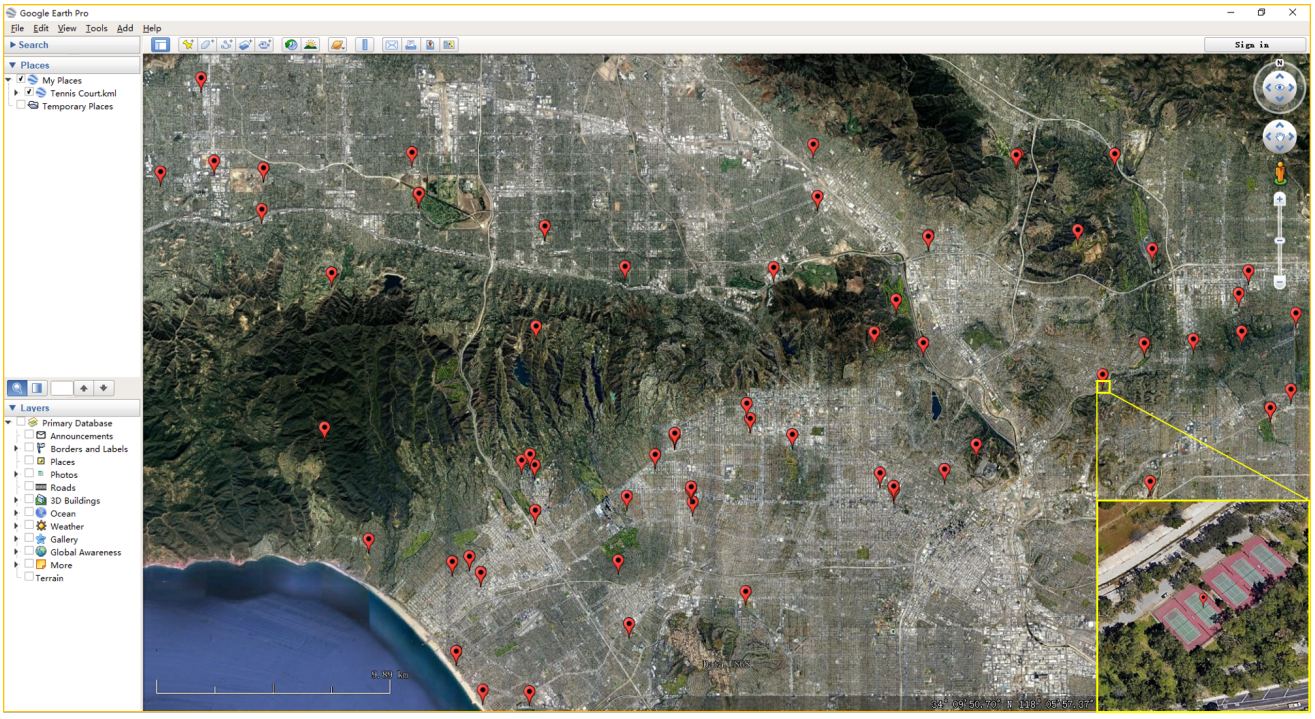


Fig. 5: The points of searched tennis courts shown in Google Earth, where the top-left and bottom-right coordinates are (34.1071° N, 118.3605° W) and (33.9823° N, 118.3605° W), respectively. We consider the tennis courts as point ground features and the red marks show the searched positions of tennis courts. The eagle display shows the detail of a tennis court scene, which confirms the rationality and validity of collecting semantic point coordinates by our proposed method.

naturally assigned with semantic annotation of scene tags. Figure 5 shows the search result returned by the semantic tag "baseball field" based on the tool. By searching through a wide range of areas, we can quickly obtain a large number of matching points with semantic tags.

The Google Map API has provided a powerful interface for accessing point data. However, most of them are associated with common life scene categories. For those scene categories related to production activities, it is reasonable to employ the publicly available geographic datasets and obtain the point data. With some online geographic dataset publishing websites, we collect the coordinate data of storage tanks, bridges, and wind turbines. For example, the U.S. Wind Turbine Database (USWTDB)[5] provides a large number of locations of land-based and offshore wind turbines in the United States. Figure 6 shows the detail of the tagged data of wind turbines, which can indicate the accurate positions of wind turbines in a local area. By processing these data, a single point coordinate data corresponding to a specific scene category is obtained.

4.2.2 *Line and plane coordinates*

The ground features, such as river and railway, are usually presented in the form of lines; other features like grassland and residential land are typically presented by plane areas in RS images. In order to obtain the scene images of line and plane features, we employ OSM to extract the location information of line and plane features. OSM is a collaborative project to create a free editable map of the world. The elements in OSM consist of node, way, and relation, which are also the basic components of OSM conceptual data model that depicts the physical world. A node represents a point feature on the ground surface. It can be defined as the latitude and longitude. The way feature is composed of two or more connected nodes that define a polyline. An open way describes a linear feature, such as roads, streams and railway lines. A plane or area feature can be described in a closed way. A relation element in OSM is utilized to describe one or several complex objects with a data structure that records a relationship between nodes, ways, and even other relations. Every node or way has tags and geographic information that describe the ground object. Therefore, a line or plane feature that belongs to certain semantic classes can be captured by searching the corresponding tags.

Many methods can be employed to obtain the geographic coordinates of ground features from OSM. As the most convenient way, we collect the semantic categories of line and plane features directly from the free, community-maintained data, *e.g.*, Geofabrik[6], produced by OSM. Figure 7 shows the river line features collected through Geofabrik, which provides maps and geographic data extracted from OSM. Besides, we also employ the OSM interface, *i.e.*, Overpass API, to extract the features of interest. In order to obtain the semantic coordinates of categories in the constructed network, we search by the powerful query language with criteria like *e.g.*location, type of objects, tag properties proximity, or combinations of them. Figure 8 shows the illustration of searching scenes of airport areas

5. https://eerscmap.usgs.gov/uswtdb

6. http://www.geofabrik.de/geofabrik

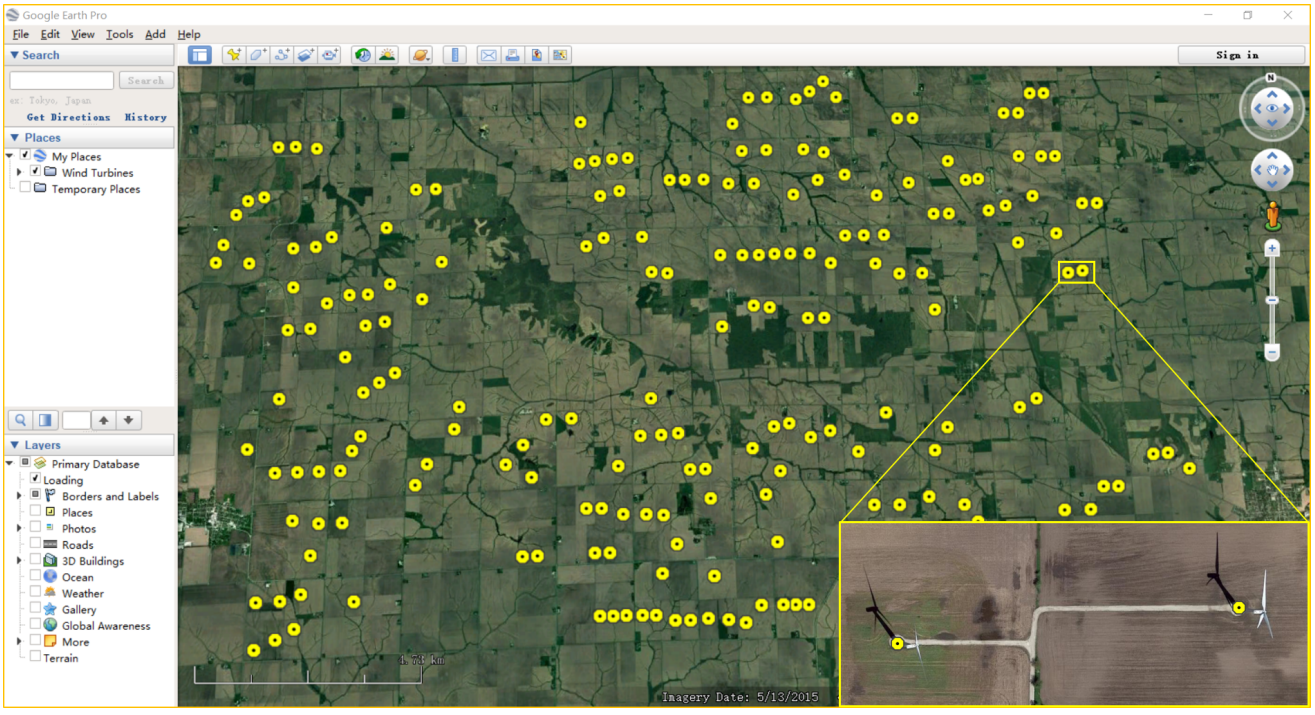


Fig. 6: The points of wind turbines extracted from USWTDB and integrated in Google Earth, where the geographic range is indicated with the top-left coordinates (41.2695° N, 90.3315° W) and bottom-right coordinates (41.1421° N, 90.0424° W). The eagle display shows the details of two wind turbines.

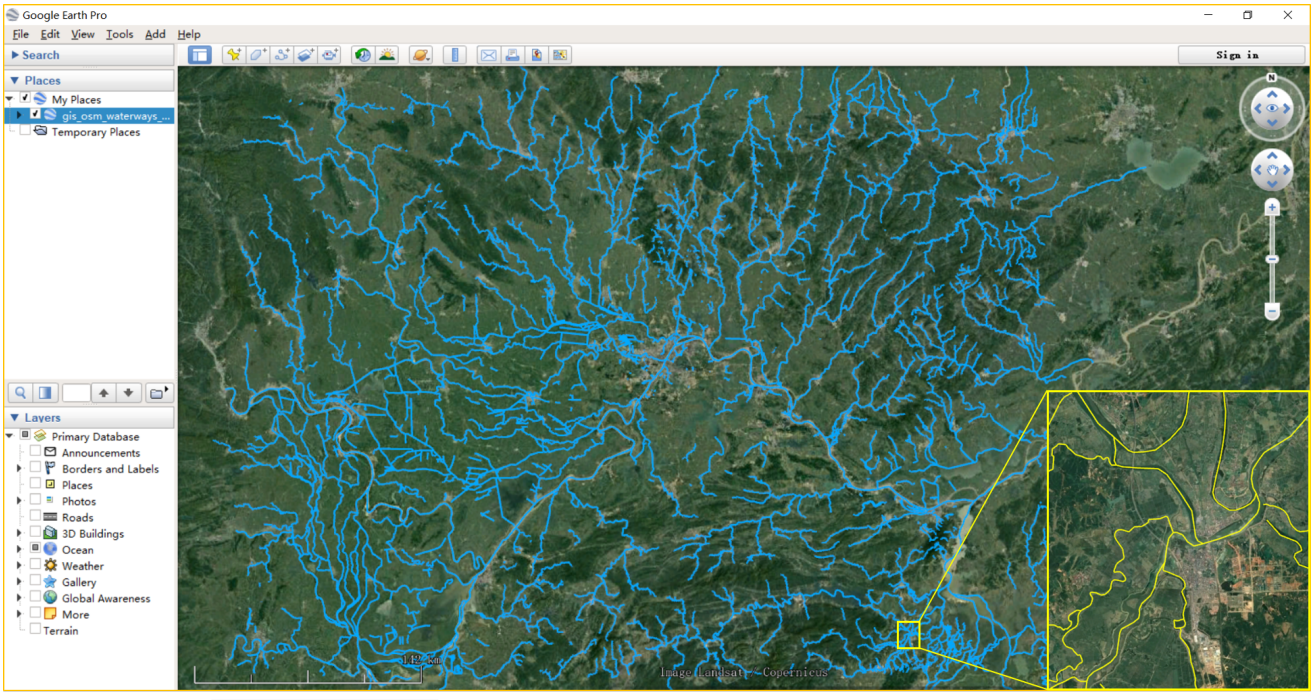
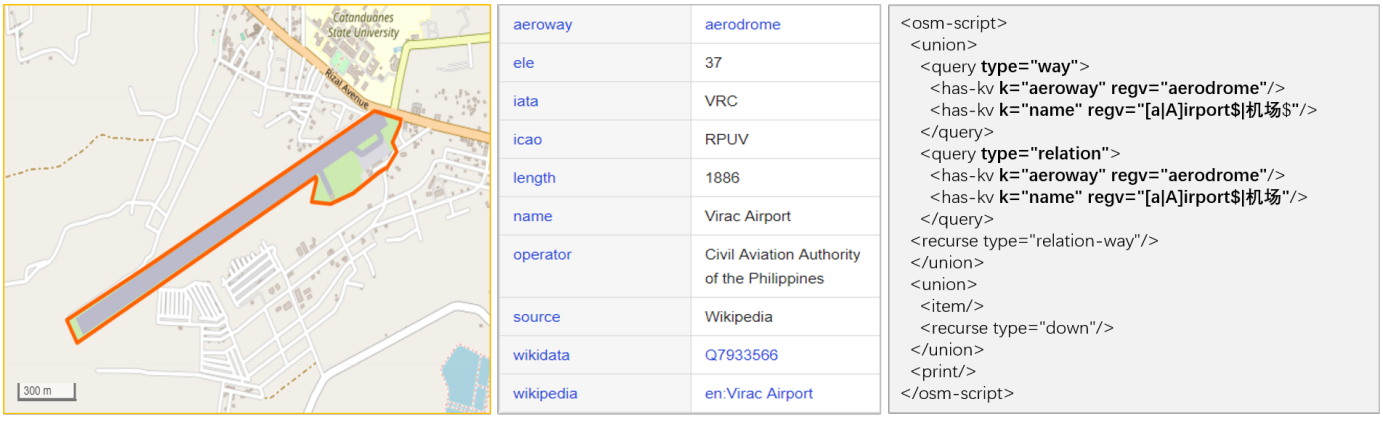


Fig. 7: The river lines within a local area of China are extracted from OSM and displayed in Google Earth, where the upper left and bottom right coordinates of geographic range are (32.2826° N, 111.1027° E) and (28.7477° N, 118.5372° E), respectively. The zoomed image shows the details of river lines.

around the world. And the searched airports within a local area of the United States are integrated into Google Earth and shown in figure 9. It can be seen from Figure 8-9 that the extracted plane data is consistent with the real-world airport scenes, and thus, the semantic label is of great reliability.

4.3 Scene Image Acquisition

The geographic point, line and plane coordinates collected through the above processes are employed to extract scene images from Google Earth. For the searched point data, the coordinates are attached with specific semantic category



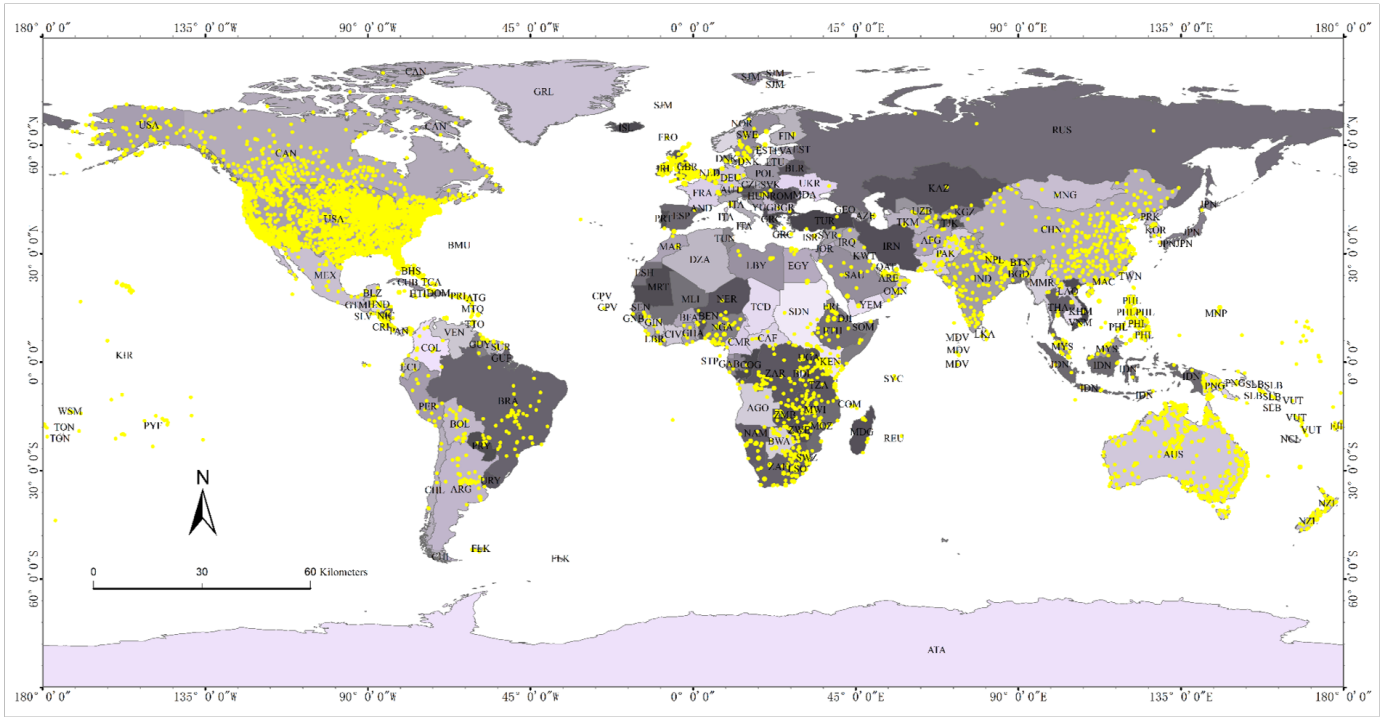
(a)

aeroway	aerodrome
ele	37
iata	VRC
icao	RPUV
length	1886
name	Virac Airport
operator	Civil Aviation Authority of the Philippines
source	Wikipedia
wikidata	Q7933566
wikipedia	en:Virac Airport

(b)

```
<osm-script>
 <union>
  <query type="way">
   <has-kv k="aeroway" regv="aerodrome"/>
   <has-kv k="name" regv="[a|A]irport$|机场$"/>
  </query>
  <query type="relation">
   <has-kv k="aeroway" regv="aerodrome"/>
   <has-kv k="name" regv="[a|A]irport$|机场"/>
  </query>
 <recurse type="relation-way"/>
 </union>
 <union>
  <item/>
  <recurse type="down"/>
 </union>
 <print/>
</osm-script>
```

(c)



(d)

Fig. 8: The illustration of searching scenes of airports around the world. An airport in OSM contains a large amount of tags, which can be employed to search airports with specific semantic key-value labels, *e.g.aeroway* and *name*. More than 5,000 world airports in forms like way and relation can be obtained with accurate geographic coordinates, using English and Chinese semantic tags.

tags and we take the geographic coordinates as the center of a square box. For line data, we sample the points along the line by intervals and a sampled point is selected as the center of a square box. Based on the center point, a square box of customized size is generated to serve as a scene box characterized by four geographic coordinates. For the plane data, *e.g.*, commercial area, a mesh grid is generated to divide the plane area into individual scene boxes. Some scenes like airport and train station are usually presented with individual blocks. Therefore, the envelop rectangles of these blocks are extracted as the scene boxes directly. Thus, the content of a scene image box is consistent with its corresponding semantic scene category.

All the scene boxes are utilized to outline and download scene images from Google Earth. The scene images are extracted with different sizes, such as 256×256, 512×512 and 1024×1024, according to the scene scale and resolution of Google Earth image. Therefore, there are images of different resolutions ranging from 0.5m to 10m per pixel. We collect the scene images that belong to the same categories and organize them to be a dataset according to the established hierarchical scene category network. Figure 10 illustrates the overall framework of collecting RS scene images. Note that there are images with different resolutions even within the same scene category. Images for each scene category are also extracted from different areas around the world, ensuring the wide distribution of the geographic space. The multiple data sources, *e.g.*, QuickBird, IKONOS, SPOT, and Landsat serial satellites, of Google Earth also greatly improved the richness of the scene images. Consequently, the number of images in each scene category goes beyond 5,000. All of these provide guarantees of scale, variation, and diversity

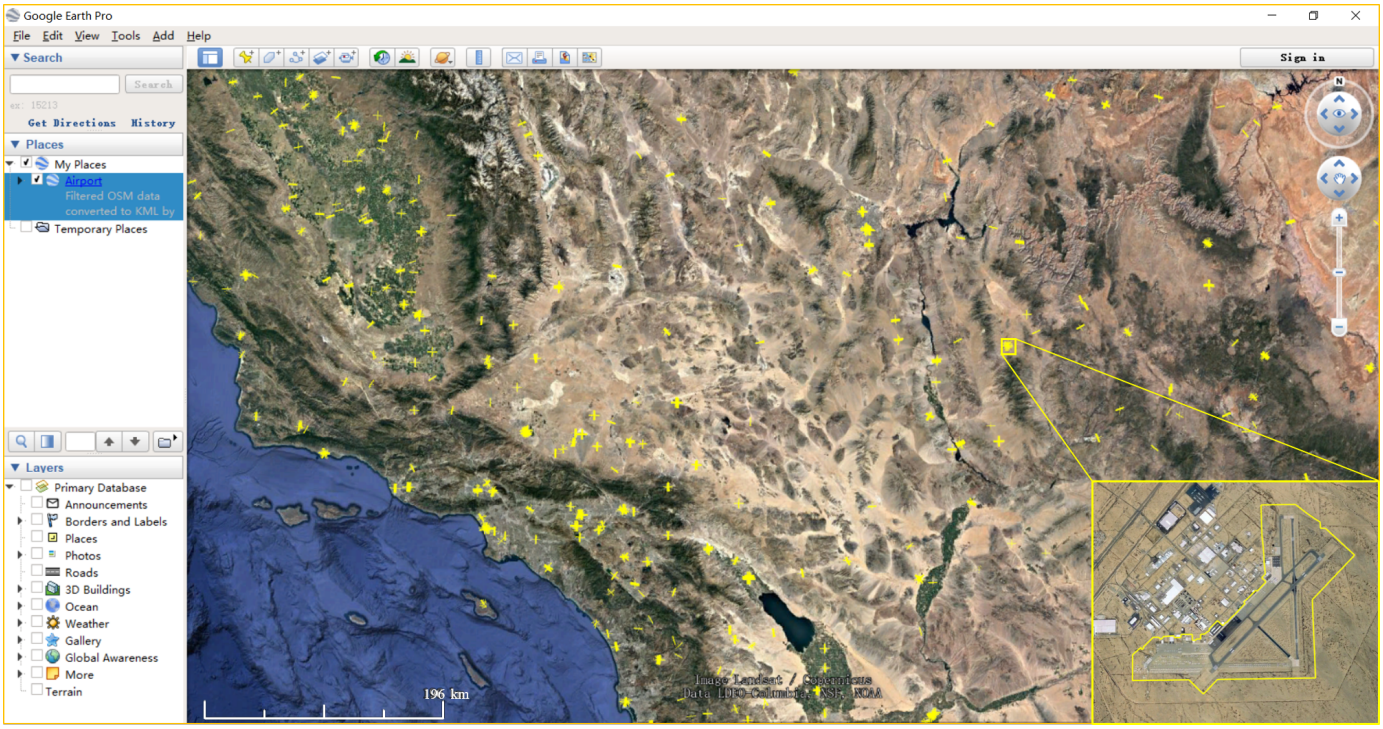


Fig. 9: The airport planes within a local area of the United States are extracted from OSM and shown in Google Earth, where the upper left and bottom right coordinates of geographic range are (37.2262° N, 115.8819° W) and (32.6005° N, 110.6497° W), respectively. The zoomed image shows the accurate extent of an airport in a closed OSM way feature.

for the constructed scene dataset.

There may be inaccurate semantic label assignments in the dataset caused by coordinate searching and image acquiring. For those scene boxes that are overlapped with each other, only one of the scene boxes will be chosen to extract the corresponding scene image. To ensure the correctness of the category labels, all of the scene images are checked by specialists in the field of RS interpretation. Specifically, those downloaded images for a specific category are deleted if they are assigned with wrong scene labels. This strategy is mainly conducted automatically while only image check and deletion are performed manually. Therefore, the strategy for scene dataset construction falls in a semi-automatic annotation method which can greatly reduce the manpower cost and ensure the label quality. Thus, it is feasible to build a large-scale RS scene image dataset of high reliability. With the introduced method, the Million-AID dataset is constructed with more than 1,000,000 annotated semantic images of 51 scene categories.

5 Challenges and Perspectives

Driven by the wide application of RS image interpretation, various datasets have been constructed for the development of interpretation algorithms. In spite of the great success in RS image datasets construction over the past years, there still exists a giant gap between the requirement of large-scale dataset and interpretation algorithm evolution, especially for those data-driven methods. Thus, how to speed up the annotation process of RS images remains to be a key issue for the construction of interpretation datasets. By investigating the current annotation strategies for RS image datasets, this paper catches a glimpse of the current challenges and potential perspectives for efficient dataset construction.

5.1 Visualization Technology for RS image Annotation

In the process of annotation for a RS image, semantic content in the image is firstly recognized by the visual interpretation of experts. Then, the semantic labels are assigned to the corresponding objects, typically the pixel-wise, regional, or image level. Thus, the visualization technology for RS image plays a significant role in the process of accurate semantic annotation, especially for the hyper-spectral, SAR, and large size RS images.

hyper-spectral image annotation with visualization technology. hyper-spectral image contains continuous hundreds of spectral bands, which can provide rich spatial-spectral information of the earth's surface features. However, the high dimensionality of the hyper-spectral image brings the challenge for semantic information annotation. The reality is that the existing display devices are designed for gray or color images with typical RGB channels. Thus, it is impossible to directly display a hyper-spectral RS image which consists of hundreds of spectral bands using traditional display strategies. In response to this challenge for hyper-spectral RS image annotation, the strategy of band selection can be explored to choose three representative bands of the original image as RGB channels [135]. The fundamental idea of this strategy is to select bands with as much information as possible from the original hyper-spectral image or directly according to the characteristics of the annotation objects. Alternatively, band transformation can also be considered by making the best use of the rich

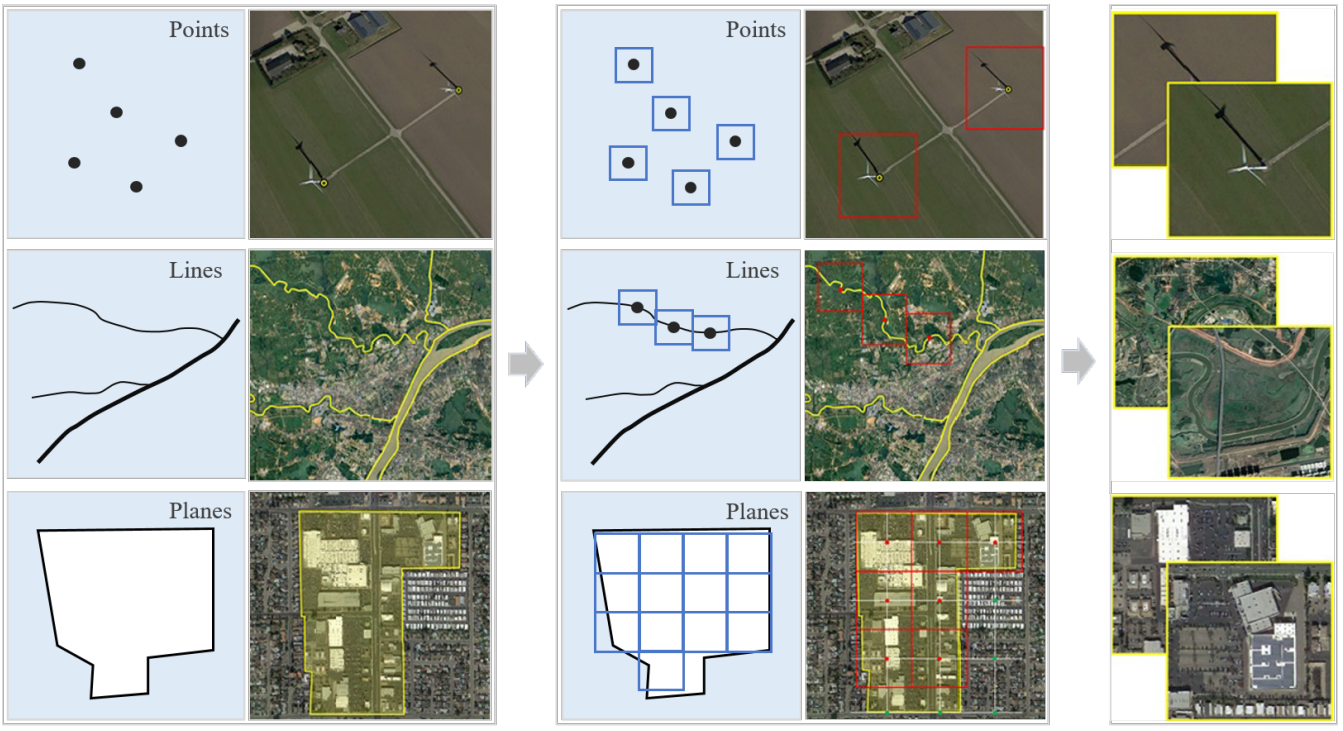


Fig. 10: The illustration of the acquisition of RS scene images based on the collected geographic point, line and area data. The points are set as the centers of scene square boxes. For line data, the center points are sampled by intervals. For plane data, scene square boxes are generated by mesh grids. The red frames indicate the generated scene square boxes which is consistent with the final scene image blocks.

bands in hyper-spectral images. The basic principle is to transform the original image into a new feature space by spectral transformation, *e.g.*, dimensionality reduction, band fusion and clustering, and select the first three features as RGB channels for image visualization [136]. Besides, these strategies should rely on effective algorithms developed for band selection or transformation. Thus, hyper-spectral RS images can be well visualized, providing a guarantee for annotating reliable information.

SAR image annotation via physical signal expression. Compared with optical RS images, the challenge of SAR image annotation mainly comes from the weak legibility in visual appearance. SAR's all-weather, all-day, and penetrating imaging ability show its superiority over optical RS images in some practical applications, *e.g.*disaster monitoring. However, due to the interference of coherent returns scattered by small reflectors within each resolution cell, SAR images are contaminated by multiplicative noise called speckle. Also, SAR images are usually grayscale without any color information except for full-polarimetric SAR images [137]. All of these pose great challenges for SAR image annotation. An essential point is that the SAR image is represented with signal information, where different objects show different polarimetric features. Thus, the utilization of physical information of objects can be a promising solution for SAR image annotation, relying upon the basic principles related to surface roughness or smoothness and changes in the back-scattering signal intensity of the surface conditions [138]. On the other hand, visualization technology should also be explored to enhance the legibility of SAR images. One direction is to colorize the non-full-polarimetric to full-polarimetric SAR images based on the radar polarimetry theories. Inspired by the success of transfer learning in computer vision, it is also valuable to color the SAR images through simulating RGB images using DCNNs [137]. With these considerations, there is much work to do with SAR image visualization to relieve the difficulties of annotation.

Large-size image annotation with high interaction efficiency. The rapid annotation for large size images is another important challenge in the field of RS image annotation. Currently, RS image annotation is usually conducted by resorting to natural image annotation tools [103], [104], where only images with limited sizes, *e.g.*, image width/height of several hundred pixels, can be fully visualized for annotation. However, with the improvement of image resolution, RS images taken from the bird-view have large geographic coverage and thus possess large size, usually with width/height of tens of thousands pixels. Thus, the traditional annotation solution can only visualize the local region of a RS image for annotation operations. Besides, current machine monitor devices are also with limited sizes and resolution. This requires constant image roaming and zooming operations when annotating large size RS images, which heavily hinders the interaction efficiency of annotation and loss the possibility of catching the spatially continuous features from a global perspective of image content. On the other hand, the RS image with spatial information needs large storage space because of its large amount of data. Thus, the visualization will also involve in large-scale computing when annotating RS images of large sizes. Considering these points, the

TABLE 6: Annotation tools for image dataset construction

No.	Name	Ref.	Year	Description
1	LabelMe	[103]	2008	An online image annotation tool that supports various annotation primitives, including polygon, rectangle, circle, line and point.
2	Video Annotation Tool from Irvine, California (VATIC)	[139]	2012	An online tool that efficiently scaling up video annotation with crowdsourced marketplaces (*e.g.*, AMT).
3	LabelImg	[104]	2015	A popular graphical image annotation application that labels objects in images with bounding boxes.
4	Visual Object Tagging Tool (VOTT)	[140]	2017	An open source annotation and labeling tool for image and video assets, extensible for importing/exporting data to local or cloud storage providers, including Azure Blob Storage and Bing Image Search.
5	omputer Vision Annotation Tool (CVAT)	[141]	2018	A universal data annotation approach for both individuals and teams, supporting large-scale semantic annotation for scene classification, object detection and image segmentation.
6	Image Tagger	[142]	2018	An open source online platform to create and manage image data and diverse labels (*e.g.*, bounding box, polygon, line and point), with friendly support for collaborative image labeling.
7	Polygon RNN++	[110]	2018	A deep learning-based annotation strategy, producing polygonal annotation of objects segmentation interactively using humans-in-the-loop.
8	Makesence.AI	[143]	2019	An open source and online image annotation platform, using different artificial model to give recommendations as well as automate repetitive and tedious labeling activities.
9	VGG Image Annotator (VIA)	[144]	2019	A simple and standalone manual annotation software for image and video, providing rich labels like point, line, polygon as well as circle and ellipse without project management.

* This table non-exhaustively presents the popular and representative image annotation tools.

visualization technology for displaying, roaming, zooming, and annotating large size RS images needs to be addressed for efficient annotation.

5.2 Annotation Efficiency and Quality Improvement

There is no doubt that the constructed RS image dataset is ultimately utilized for various interpretation applications. Thus, the application products can be employed in turn to facilitate the annotation of RS images. And in the annotation process, the reliable tools developed for RS image annotation tasks play an important role in efficiency improvement. Besides, noise data is a common problem in RS image semantic annotation, which makes the handling of noise annotation a valuable topic for the control of dataset quality and development of interpretation algorithms.

Cooperation with application departments. A feasible way to improve the efficiency of RS image annotation is to cooperate with application departments and convert the application products to annotated datasets. Once the product data of the application department is produced, they naturally carry semantic information which can be utilized as the source of RS image annotations. For example, the map data of land survey from the land-use institution is usually obtained through field investigation and thus possesses accurate land classification information, which can be easily combined with RS images to create high-quality annotated datasets. This scheme is reasonable because the product data from the application department is oriented to the real application scenarios. At this point the created RS image interpretation dataset can most truly reflect the key challenges and problems in the real application scenarios. Thus, the interpretation algorithms built upon this kind of dataset would have more practicability. Besides, the product data will change with the alternation of the application department's business. Thus, the product data can be employed to update the created RS image interpretation dataset promptly. In this way, it can ensure that the established dataset is always oriented to the real applications, and therefore, promote the design and training of practical interpretation algorithms. In general, the efficiency and practicality of the RS image interpretation dataset can be greatly improved by the cooperation with application departments.

Tools for RS image annotation. Another point worth noting is the necessity of developing open-sourced and professional tools for RS image annotation. A number of popular tools concerning image annotation have been published, as listed in Table 6. These include excellent tools for specific image content interpretation tasks, *e.g.*, object recognition [104], [110], [140]. Some annotation tools strive to provide diverse annotation modalities, such as polygon, rectangle, circle, ellipse, line, and point [103], [141]–[144], serving as universal annotation platforms applicable to build image-level labels, the local extent of objects and semantic information of pixels. Due to the differences in interpretation tasks and application requirements, the most common concerns among annotators are the features and instructions of these tools. The properties of these annotations tools are summarized in Table 6 and more details can be found in the corresponding reference materials. In general, when building an annotated dataset for RS image content interpretation, the choice of a flexible annotation tool is of great significance for efficiency and quality assurance.

Processing for noisy annotations. The processing of noise annotations and the development of algorithms tolerant to

noise annotations are the common issues in real application scenarios. In the construction of a RS image dataset, images can be annotated by multiple experts while each annotator possesses the varying level of expertise and their opinions may commonly conflict with each other because of personal bias [145]. Not only that, but RS images can also be too complex to be correctly interpreted even for the experts due to the high demand for specialized background and knowledge for RS image content recognition. All of these will inevitably lead to noisy annotations. An intuitive approach to overcome this problem is to remove the noisy annotations by manual cleaning and correction. However, manual annotation cleansing usually results in high costs in terms of time and labor. Thus, how to quickly find out the possible noise annotations for a constructed dataset becomes a challenging problem. Faced with this situation, it is valuable to build effective algorithms to model and predict noise annotations for data cleansing and quality improvement. On the other hand, in order to obtain a high-performance algorithm for RS image interpretation, most data-driven methods require a fair amount of data with precise annotations for proper training, particularly the deep learning algorithms. With this in mind, the effect of noise annotation on the performance of interpretation algorithms is necessary to be explored for the better utilization of the annotated dataset. Furthermore, it is crucial to consider the existence of annotation noise and develop noise-robust algorithms [146], [147] to efficiently fade away its negative effects for RS image interpretation.

6 Conclusions

Remote sensing technology over the past years has made tremendous progress and been providing us an ocean of RS images for systematic observation of the earth surface. However, the lack of publicly available large-scale RS image datasets of accurate annotation has become a bottleneck problem to the development of new and intelligent approaches for image interpretation.

Through a bibliometric analysis, this paper first presents a systematic review of the existing datasets related to the mainstream of RS image interpretation tasks. It reveals that the annotated RS image datasets, to some extent, show deficiencies in one or several different aspects, *e.g.*, diversity and scale, that hamper the development of practical interpretation models. Hence, the creation of RS image datasets needs to be paid with more attention, from the annotation process to property control for real applications. Subsequently, we paid efforts to explore the principles for building the useful annotated RS image dataset. It is suggested that the construction of the RS image datasets should be created toward the requirements of practical applications, rather than the interpretation algorithms. The introduced principles formulates a prototype for RS image dataset construction with consideration in efficiency, quality assurance, and property assessment. With the established principles, we created a large-scale RS image dataset for scene classification, *i.e.*, Million-AID, through a semi-automatic annotation strategy. It will provide new ideas and approaches for the construction and improvement of RS image datasets. And the discussion about challenges and perspectives in RS image dataset annotation delivers a new sight for the future work that efforts needed to be dedicated to RS image dataset annotation.

In the future, we will devote our endeavor to develop a publicly online evaluation platform for various interpretation datasets and algorithms. We believe that the trend of intelligent interpretation for RS images is unstoppable, and more practical datasets and algorithms oriented to real RS applications will be created in the coming years. It should be encouraged that more datasets and interpretation frameworks be shared within the RS community to advance the prosperity of intelligent interpretation and applications of RS images.

References

[1] T.-Z. Xiang, G.-S. Xia, and L. Zhang, "Mini-unmanned aerial vehicle-based remote sensing: Techniques, applications, and prospects," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 3, pp. 29–63, 2019.

[2] C. Toth and G. Jóźków, "Remote sensing platforms and sensors: A survey," *ISPRS J. Photogrammetry Remote Sens.*, vol. 115, pp. 22–36, 2016.

[3] L. Zheng, G. Zhao, J. Dong, Q. Ge, J. Tao, X. Zhang, Y. Qi, R. B. Doughty, and X. Xiao, "Spatial, temporal, and spectral variations in albedo due to vegetation changes in china's grasslands," *ISPRS J. Photogrammetry Remote Sens.*, vol. 152, pp. 1–12, 2019.

[4] M. E. Bauer, "Remote sensing of environment: History, philosophy, approach and contributions, 1969–2019," *Remote Sens. Environ.*, vol. 237, p. 111522, 2020.

[5] X.-Y. Tong, G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, p. 111322, 2020.

[6] A. Shaker, W. Y. Yan, and P. E. LaRocque, "Automatic land-water classification using multispectral airborne lidar data for near-shore and river environments," *ISPRS J. Photogrammetry Remote Sens.*, vol. 152, pp. 94–108, 2019.

[7] Y. Shendryk, Y. Rist, C. Ticehurst, and P. Thorburn, "Deep learning for multi-modal classification of cloud, shadow and land cover scenes in planetscope and sentinel-2 imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 157, pp. 124–136, 2019.

[8] A. M. Coutts, R. J. Harris, T. Phan, S. J. Livesley, N. S. Williams, and N. J. Tapper, "Thermal infrared remote sensing of urban heat: Hotspots, vegetation, and an assessment of techniques for use in urban planning," *Remote Sens. Environ.*, vol. 186, pp. 637–651, 2016.

[9] W. Zhou, D. Ming, X. Lv, K. Zhou, H. Bao, and Z. Hong, "So–cnn based urban functional zone fine division with vhr remote sensing image," *Remote Sens. Environ.*, vol. 236, p. 111458, 2020.

[10] G.-S. Xia, G. Liu, X. Bai, and L. Zhang, "Texture characterization using shape co-occurrence patterns," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 5005–5018, 2017.

[11] R. M. Anwer, F. S. Khan, J. Vandeweijer, M. Molinier, and J. Laaksonen, "Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 138, pp. 74–85, 2018.

[12] Q. Li, L. Mou, Q. Liu, Y. Wang, and X. X. Zhu, "Hsf-net: Multiscale deep feature embedding for ship detection in optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7147–7161, 2018.

[13] 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 Symposium - 100 Years ISPRS*, 2010, pp. 298–303.

[14] A. R. White, "Human expertise in the interpretation of remote sensing data: A cognitive task analysis of forest disturbance attribution," *Int. J. Appl. Earth Obs. Geoinform.*, vol. 74, pp. 37–44, 2019.

[15] Q. Bi, K. Qin, Z. Li, H. Zhang, K. Xu, and G.-S. Xia, "A multiple-instance densely-connected convnet for aerial scene classification," *IEEE Trans. Image Process.*, vol. 29, pp. 4911–4926, 2020.

[16] X.-Y. Tong, G.-S. Xia, F. Hu, Y. Zhong, M. Datcu, and L. Zhang, "Exploiting deep features for remote sensing image retrieval: A systematic investigation," *IEEE Trans. Big Data*, pp. 1–1, 2019.

[17] W. Yang, X. Yin, and G. Xia, "Learning high-level features for satellite image classification with limited labeled samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4472–4482, 2015.

[18] J. Xie, N. He, L. Fang, and A. Plaza, "Scale-free convolutional neural network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6916–6928, 2019.

[19] X. Zheng, Y. Yuan, and X. Lu, "A deep scene representation for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4799–4809, 2019.

[20] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, 2018.

[21] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning roi transformer for oriented object detection in aerial images," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2019, pp. 2849–2858.

[22] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2018, pp. 3974–3983.

[23] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *arXiv preprint arXiv:1905.05055*, 2019.

[24] X. Wu, D. Hong, J. Tian, J. Chanussot, W. Li, and R. Tao, "Orsim detector: A novel object detection framework in optical remote sensing imagery using spatial-frequency channel features," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 5146–5158, 2019.

[25] M. D. Hossain and D. Chen, "Segmentation for object-based image analysis (obia): A review of algorithms and challenges from remote sensing perspective," *ISPRS J. Photogrammetry Remote Sens.*, vol. 150, pp. 115–134, 2019.

[26] L. Ma, M. Li, X. Ma, L. Cheng, P. Du, and Y. Liu, "A review of supervised object-based land-cover image classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 130, pp. 277–293, 2017.

[27] Y. Xu, M. Fu, Q. Wang, Y. Wang, K. Chen, G.-S. Xia, and X. Bai, "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–8, 2020.

[28] S. Liu, D. Marinelli, L. Bruzzone, and F. Bovolo, "A review of change detection in multitemporal hyperspectral images: Current techniques, applications, and challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 140–158, 2019.

[29] Y. Gu, J. Chanussot, X. Jia, and J. A. Benediktsson, "Multiple kernel learning for hyperspectral image classification: A review," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6547–6565, 2017.

[30] C. Geiß, P. A. Pelizari, L. Blickensdörfer, and H. Taubenböck, "Virtual support vector machines with self-learning strategy for classification of multispectral remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 151, pp. 42–58, 2019.

[31] Z. Zhu, "Change detection using landsat time series: A review of frequencies, preprocessing, algorithms, and applications," *ISPRS J. Photogrammetry Remote Sens.*, vol. 130, pp. 370–384, 2017.

[32] F. Ghazouani, I. R. Farah, and B. Solaiman, "A multi-level semantic scene interpretation strategy for change interpretation in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8775–8795, 2019.

[33] G.-S. Xia, G. Liu, W. Yang, and L. Zhang, "Meaningful object segmentation from sar images via a multiscale nonlocal active contour model," *IEEE Trans. geosci. remote sens.*, vol. 54, no. 3, pp. 1860–1873, 2015.

[34] W. Zhang, X. Lu, and X. Li, "A coarse-to-fine semi-supervised change detection for multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3587–3599, 2018.

[35] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, 2017.

[36] 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. Top. Appl. Earth Obs. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, 2019.

[37] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, 2017.

[38] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais *et al.*, "Deep learning and process understanding for data-driven earth system science," *Nature*, vol. 566, no. 7743, pp. 195–204, 2019.

[39] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, "Advanced spectral classifiers for hyperspectral images: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 1, pp. 8–32, 2017.

[40] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review," *ISPRS J. Photogrammetry Remote Sens.*, vol. 66, no. 3, pp. 247–259, 2011.

[41] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS J. of Photogrammetry Remote Sens.*, vol. 114, pp. 24–31, 2016.

[42] A. Lagrange, M. Fauvel, S. May, and N. Dobigeon, "Hierarchical bayesian image analysis: From low-level modeling to robust supervised learning," *Pattern Recognition*, vol. 85, pp. 26–36, 2019.

[43] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sens.*, vol. 7, no. 11, pp. 14 680–14 707, 2015.

[44] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogrammetry Remote Sens.*, vol. 152, pp. 166–177, 2019.

[45] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[46] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," in *Proc. European Conf. Computer Vision*, 2014, pp. 740–755.

[47] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *Int. J. Comput. Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[48] C. Chen, "Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature," *Journal of the American Society for information Science and Technology*, vol. 57, no. 3, pp. 359–377, 2006.

[49] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. Int. Conf. Advances in Geographic Information Systems*. ACM, 2010, pp. 270–279.

[50] 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, 2015.

[51] S. Basu, S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, and R. Nemani, "Deepsat: A learning framework for satellite imagery," in *Proc. Int. Conf. Advances in Geographic Information Systems*. ACM, 2015, pp. 1–10.

[52] O. A. B. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshop*, 2015, pp. 44–51.

[53] 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, p. 035004, 2016.

[54] 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, 2016.

[55] 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, 2017.

[56] H. Li, X. Dou, C. Tao, Z. Wu, J. Chen, J. Peng, M. Deng, and L. Zhao, "Rsi-cb: A large-scale remote sensing image classification benchmark using crowdsourced data," *Sensors*, vol. 20, no. 6, 2020.

[57] 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, p. 725, 2017. [Online]. Available: https://github.com/RSIA-LIESMARS-WHU/RSD46-WHU

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

[59] G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in *Proc. European Conf. Computer Vision*, 2008, pp. 30–43.

[60] C. Benedek, X. Descombes, and J. Zerubia, "Building development monitoring in multitemporal remotely sensed image pairs with stochastic birth-death dynamics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 33–50, 2012.

[61] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS J. Photogrammetry Remote Sens.*, vol. 98, pp. 119–132, 2014.

[62] K. Liu and G. Máttyus, "Fast multiclass vehicle detection on aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 9, pp. 1938–1942, 2015.

[63] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao, "Orientation robust object detection in aerial images using deep convolutional neural network," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 3735–3739.

[64] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery: A small target detection benchmark," *J. Vis. Commun. Image Represent.*, vol. 34, pp. 187–203, 2016.

[65] T. N. Mundhenk, G. Konjevod, W. A. Sakla, and K. Boakye, "A large contextual dataset for classification, detection and counting of cars with deep learning," in *Proc. European Conf. Computer Vision*, 2016, pp. 785–800.

[66] Z. Liu, H. Wang, L. Weng, and Y. Yang, "Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 8, pp. 1074–1078, 2016.

[67] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2486–2498, 2017.

[68] M.-R. Hsieh, Y.-L. Lin, and W. H. Hsu, "Drone-based object counting by spatially regularized regional proposal networks," in *Proc. IEEE Int. Conf. Computer Vision*, 2017, pp. 4145–4153.

[69] Z. Zou and Z. Shi, "Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1100–1111, 2017.

[70] P. Zhu, L. Wen, X. Bian, L. Haibin, and Q. Hu, "Vision meets drones: A challenge," *arXiv preprint arXiv:1804.07437*, 2018.

[71] D. Lam, R. Kuzma, K. McGee, S. Dooley, M. Laielli, M. Klaric, Y. Bulatov, and B. McCord, "xview: Objects in context in overhead imagery," *arXiv preprint arXiv:1802.07856*, 2018.

[72] Y. Zhang, Y. Yuan, Y. Feng, and X. Lu, "Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5535–5548, 2019.

[73] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 296 – 307, 2020.

[74] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, University of Toronto, 2013.

[75] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark," in *Proc. IEEE International Geoscience and Remote Sensing Symposium*, 2017, pp. 3226–3229.

[76] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, 2018.

[77] Y. Tarabalka, J. A. Benediktsson, and J. Chanussot, "Spectral-spatial classification of hyperspectral imagery based on partitional clustering techniques," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 8, pp. 2973–2987, 2009.

[78] "Pavia dataset." [Online]. Available: http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes

[79] "Isprs 2d semantic labeling contest," June 2018. [Online]. Available: http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html

[80] M. Volpi and V. Ferrari, "Semantic segmentation of urban scenes by learning local class interactions," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshop*, June 2015, pp. 1–9.

[81] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, "Sen12ms–a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion," *arXiv preprint arXiv:1906.07789*, 2019.

[82] K. Yang, Z. Liu, G.-S. Xia, and L. Zhang, "CSDN: A cross spatial difference network for semantic change detection in remote sensing images," in *Proc. IEEE International Geoscience and Remote Sensing Symposium*, 2020.

[83] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 492–501, 2005.

[84] M. F. Baumgardner, L. L. Biehl, and D. A. Landgrebe, "220 band aviris hyperspectral image data set: June 12, 1992 indian pine test site 3," Sep 2015. [Online]. Available: https://purr.purdue.edu/publications/1947/1

[85] M. Zhang, X. Hu, L. Zhao, Y. Lv, M. Luo, and S. Pang, "Learning dual multi-scale manifold ranking for semantic segmentation of high-resolution images," *Remote Sens.*, vol. 9, no. 5, p. 500, 2017.

[86] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 60–77, 2018.

[87] X. X. Zhu, J. Hu, C. Qiu, Y. Shi, J. Kang, L. Mou, H. Bagheri, M. Häberle, Y. Hua, R. Huang *et al.*, "So2sat lcz42: A benchmark dataset for global local climate zones classification," *IEEE Geosci. Remote Sens. Mag.*, 2020.

[88] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, "Uavid: A semantic segmentation dataset for uav imagery," *ISPRS J. Photogrammetry Remote Sens.*, 2020.

[89] C. Benedek and T. Szirányi, "Change detection in optical aerial images by a multilayer conditional mixed markov model," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 10, pp. 3416–3430, 2009.

[90] N. Bourdis, D. Marraud, and H. Sahbi, "Constrained optical flow for aerial image change detection," in *Proc. IEEE International Geoscience and Remote Sensing Symposium*, 2011, pp. 4176–4179.

[91] C. Wu, B. Du, and L. Zhang, "Slow feature analysis for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2858–2874, 2013.

[92] A. Song, J. Choi, Y. Han, and Y. Kim, "Change detection in hyperspectral images using recurrent 3d fully convolutional networks," *Remote Sens.*, vol. 10, no. 11, p. 1827, 2018.

[93] 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 International Geoscience and Remote Sensing Symposium*, 2018, pp. 1970–1973.

[94] J. López-Fandiño, A. S. Garea, D. B. Heras, and F. Argüello, "Stacked autoencoders for multiclass change detection in hyperspectral images," in *Proc. IEEE International Geoscience and Remote Sensing Symposium*, 2018, pp. 1906–1909.

[95] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral earth observation using convolutional neural networks," in *Proc. IEEE International Geoscience and Remote Sensing Symposium*, 2018, pp. 2115–2118.

[96] S. Saha, F. Bovolo, and L. Brurzone, "Unsupervised multiple-change detection in vhr optical images using deep features," in *Proc. IEEE International Geoscience and Remote Sensing Symposium*, 2018, pp. 1902–1905.

[97] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, 2018.

[98] M. Lebedev, Y. V. Vizilter, O. Vygolov, V. Knyaz, and A. Y. Rubis, "Change detection in remote sensing images using conditional adversarial networks." *Int. arch. photogramm. remote sens. spat. inf. sci.*, vol. 42, no. 2, 2018.

[99] A. Fujita, K. Sakurada, T. Imaizumi, R. Ito, S. Hikosaka, and R. Nakamura, "Damage detection from aerial images via convolutional neural networks," in *Proc. IAPR Int. Conf. Machine Vision Applications*. IEEE, 2017, pp. 5–8.

[100] C. D. Rodrigo, L. S. Bertrand, B. Alexandre, and G. Yann, "Multi-task learning for large-scale semantic change detection," *Comput. Vis. Image Underst.*, vol. 187, p. 102783, 2019.

[101] 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, 2017.

[102] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sens.*, vol. 12, no. 10, p. 1662, 2020.

[103] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," *Int. J. Comput. Vision*, vol. 77, no. 1-3, pp. 157–173, 2008.

[104] Tzutalin, "Labelimg," 2015. [Online]. Available: https://github.com/tzutalin/labelImg

[105] Q. Hu, W. Wu, T. Xia, Q. Yu, P. Yang, Z. Li, and Q. Song, "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, 2013.

[106] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning earth observation classification using imagenet pretrained networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 105–109, 2016.

[107] K. Nogueira, O. A. Penatti, and J. A. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognit.*, vol. 61, pp. 539–556, 2017.

[108] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.

[109] X. Zhu, C. Vondrick, C. C. Fowlkes, and D. Ramanan, "Do we need more training data?" *Int. J. Comput. Vision*, vol. 119, no. 1, pp. 76–92, Aug 2016.

[110] D. Acuna, H. Ling, A. Kar, and S. Fidler, "Efficient interactive annotation of segmentation datasets with polygon-rnn++," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2018, pp. 859–868.

[111] M. Andriluka, J. R. Uijlings, and V. Ferrari, "Fluid annotation: a human-machine collaboration interface for full image annotation," in *Proc. ACM Int. Conf. Multimedia*, 2018, pp. 1957–1966.

[112] Y. Yao, J. Zhang, F. Shen, L. Liu, F. Zhu, D. Zhang, and H. T. Shen, "Towards automatic construction of diverse, high-quality image datasets," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 6, pp. 1199–1211, 2020.

[113] P. Zhu, Y. Tan, L. Zhang, Y. Wang, J. Mei, H. Liu, and M. Wu, "Deep learning for multilabel remote sensing image annotation with dual-level semantic concepts," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 4047–4060, 2020.

[114] M. J. Afridi, A. Ross, and E. M. Shapiro, "On automated source selection for transfer learning in convolutional neural networks," *Pattern Recognit.*, vol. 73, pp. 65–75, 2018.

[115] V. Maihami and F. Yaghmaee, "Automatic image annotation using community detection in neighbor images," *Physica A*, vol. 507, pp. 123–132, 2018.

[116] D. Tian and Z. Shi, "Automatic image annotation based on gaussian mixture model considering cross-modal correlations," *J. Vis. Commun. Image Represent.*, vol. 44, pp. 50–60, 2017.

[117] X. Zhuo, F. Fraundorfer, F. Kurz, and P. Reinartz, "Automatic annotation of airborne images by label propagation based on a bayesian-crf model," *Remote Sens.*, vol. 11, no. 2, p. 145, 2019.

[118] Q. Cheng, Q. Zhang, P. Fu, C. Tu, and S. Li, "A survey and analysis on automatic image annotation," *Pattern Recognit.*, vol. 79, pp. 242–259, 2018.

[119] X. Zheng, X. Sun, K. Fu, and H. Wang, "Automatic annotation of satellite images via multifeature joint sparse coding with spatial relation constraint," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 4, pp. 652–656, 2013.

[120] W. Yang, D. Dai, B. Triggs, and G.-S. Xia, "Sar-based terrain classification using weakly supervised hierarchical markov aspect models," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 4232–4243, 2012.

[121] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo, "Semantic annotation of high-resolution satellite images via weakly supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3660–3671, 2016.

[122] H. Lin, P. Upchurch, and K. Bala, "Block annotation: Better image annotation with sub-image decomposition," in *Proc. IEEE Int. Conf. Computer Vision*, 2019, pp. 5290–5300.

[123] L. Jia and L. Fei-Fei, "Optimol: automatic online picture collection via incremental model learning," *Int. J. Comput. Vision*, vol. 88, no. 2, pp. 147–168, 2010.

[124] 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. Photogrammetry Remote Sens.*, vol. 145, pp. 23–43, 2018.

[125] S. Dang, Z. Cao, Z. Cui, Y. Pi, and N. Liu, "Open set incremental learning for automatic target recognition," *IEEE Trans. Geosci. Remote Sens.*, 2019.

[126] O. Tasar, Y. Tarabalka, and P. Alliez, "Incremental learning for semantic segmentation of large-scale remote sensing data," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 12, no. 9, pp. 3524–3537, 2019.

[127] E. Agustsson, J. R. Uijlings, and V. Ferrari, "Interactive full image segmentation by considering all regions jointly," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2019, pp. 11 622–11 631.

[128] A. Zlateski, R. Jaroensri, P. Sharma, and F. Durand, "On the importance of label quality for semantic segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2018, pp. 1479–1487.

[129] G.-S. Xia, Z. Wang, C. Xiong, and L. Zhang, "Accurate annotation of remote sensing images via active spectral clustering with little expert knowledge," *Remote Sens.*, vol. 7, no. 11, pp. 15 014–15 045, 2015.

[130] D.-J. Chen, J.-T. Chien, H.-T. Chen, and L.-W. Chang, "Tap and shoot segmentation," in *Proc. AAAI Conf. Artificial Intelligence*, 2018, pp. 2119–2126.

[131] D. P. Papadopoulos, J. R. Uijlings, F. Keller, and V. Ferrari, "Extreme clicking for efficient object annotation," in *Proc. IEEE Int. Conf. Computer Vision*, 2017, pp. 4930–4939.

[132] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 3159–3167.

[133] G. Kazai, J. Kamps, and N. Milic-Frayling, "Worker types and personality traits in crowdsourcing relevance labels," in *Proc. ACM int. conf. Inform. Knowl. Manag.*, 2011, pp. 1941–1944.

[134] S. Vijayanarasimhan and K. Grauman, "Large-scale live active learning: Training object detectors with crawled data and crowds," *Int. J. Comput. Vision*, vol. 108, no. 1-2, pp. 97–114, 2014.

[135] X. Kang, P. Duan, and S. Li, "Hyperspectral image visualization with edge-preserving filtering and principal component analysis," *Inf. Fusion*, vol. 57, pp. 130–143, 2020.

[136] P. Duan, X. Kang, S. Li, and P. Ghamisi, "Multichannel pulse-coupled neural network-based hyperspectral image visualization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2444–2456, 2020.

[137] G. Ji, Z. Wang, L. Zhou, Y. Xia, S. Zhong, and S. Gong, "Sar image colorization using multidomain cycle-consistency generative adversarial network," *IEEE Geosci. Remote Sens. Lett.*, pp. 1–5, 2020.

[138] S. G. Dellepiane and E. Angiati, "A new method for cross-normalization and multitemporal visualization of sar images for the detection of flooded areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 7, pp. 2765–2779, 2012.

[139] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently scaling up crowdsourced video annotation," *Int. J. Comput. Vision*, vol. 101, no. 1, pp. 184–204, 2013.

[140] Microsoft, "Visual object tagging tool," 2019. [Online]. Available: https://github.com/microsoft/VoTT

[141] Intel, "Computer vision annotation tool: A universal approach to data annotation," 2018. [Online]. Available: https://github.com/opencv/cvat

[142] N. Fiedler, M. Bestmann, and N. Hendrich, "Imagetagger: An open source online platform for collaborative image labeling," in *RoboCup 2018: Robot World Cup XXII*. Springer, 2018.

[143] P. Skalski, "Make sense," 2019. [Online]. Available: https://github.com/SkalskiP/make-sense

[144] A. Dutta and A. Zisserman, "The VIA annotation software for images, audio and video," in *Proc. ACM Int. Conf. Multimedia*, 2019, pp. 2276–2279.

[145] A. Khetan, Z. C. Lipton, and A. Anandkumar, "Learning from noisy singly-labeled data," in *Proc. Int. Conf. Learn. Repr.*, 2018, pp. 1–15.

[146] G. Algan and I. Ulusoy, "Image classification with deep learning in the presence of noisy labels: A survey," *arXiv preprint arXiv:1912.05170*, 2019.

[147] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Learning to learn from noisy labeled data," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2019, pp. 5051–5059.