On Creating Benchmark Dataset for Aerial Image Interpretation: Reviews, Guidances, and Million-AID

Yang Long, Gui-Song Xia, *Senior Member, IEEE*, Shengyang Li, Wen Yang, *Senior Member, IEEE*, Michael Ying Yang, *Senior Member, IEEE*, Xiao Xiang Zhu, *Fellow, IEEE*, Liangpei Zhang, *Fellow, IEEE*, and Deren Li

***Abstract*—The past years have witnessed 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. In this context, the benchmark datasets serve as an 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 interpretation. Specifically, we first analyze the current challenges of developing intelligent algorithms for RS image interpretation with bibliometric investigations. We then present the general guidances on creating benchmark datasets in efficient manners. Following the presented guidances, we also provide an example on building RS image dataset, i.e., Million Aerial Image Dataset (Online. Available: https://captain-whu.github.io/DiRS/), a new large-scale benchmark dataset containing a million instances for RS image 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 article will provide the RS community an overall perspective on constructing large-scale and practical image datasets for further research, especially data-driven ones.**

Manuscript received January 23, 2021; revised March 12, 2021; accepted March 22, 2021. Date of publication April 1, 2021; date of current version April 26, 2021. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61922065, Grant 61771350, Grant 41820104006, Grant 61871299, and Grant 92038301, in part by the Science and Technology Major Project of Hubei Province (Next-Generation AI Technologies) under Grant 2019AEA170, and in part 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." *(Corresponding author: Gui-Song Xia.)*

Yang Long, Liangpei Zhang, and Deren Li are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan 430079, China (e-mail: longyang@whu.edu.cn; zlp62@whu.edu.cn; drli@whu.edu.cn).

Gui-Song Xia is with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan 430079, China, and also with the School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: guisong.xia@whu.edu.cn).

Shengyang Li is with the Key Laboratory of Space Utilization, Technology, and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing 100094, China (e-mail: shyli@csu.ac.cn).

Wen Yang is with the School of Electronic Information, Wuhan University, Wuhan 430079, China, and also with State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan 430079, China (e-mail: yangwen@whu.edu.cn).

Michael Ying Yang is with the Faculty of Geo-Information Science and Earth Observation, University of Twente, 7522 NB Enschede, The Netherlands (e-mail: michael.yang@utwente.nl).

Xiao Xiang Zhu is with the Remote Sensing Technology, German Aerospace Center (DLR), 82234 Wessling, Germany, and also with the Technical University of Munich, 80333 Munchen, Germany (e-mail: xiaoxiang.zhu@dlr.de).

Digital Object Identifier 10.1109/JSTARS.2021.3070368

***Index Terms*—Annotation, benchmark datasets, Million Aerial Image Dataset (Million-AID), remote sensing (RS) image interpretation, scene classification.**

I. 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], urban planning [8], [9], etc. However, with the rapid development of the Earth observation technology, the volume of RS images increases dramatically, which raises high requirement of efficient image interpretation for real-world applications. Moreover, the rich details in RS images, such as the geometrical shapes, structural characteristics, and textural attributes, also pose great challenges to the interpretation of image content [10]–[13]. These motivate the increasing and stringent demands for automatic and intelligent interpretation of the blooming RS imagery.

To characterize RS image content, quite a few methods have been developed for various interpretation tasks, ranging from the scene-level content recognition [14]–[23], object-level image analysis [24]–[34] to the challenging pixelwise semantic understanding [35]–[46]. Benefiting from the increasing availability and various ontologies of RS images, the developed methods have reported promising performance on the interpretation of RS image content. However, many of the current methods are evaluated on small-scale image datasets which usually show domain bias for applications. Moreover, a dataset created toward specific algorithms rather than real application scenarios is hard to objectively validate the comprehensive performance of the algorithms. Recently, it is observed that data-driven approaches, particularly the deep learning ones [47]–[50], have become an important alternative to manual interpretation and provided a bright prospect for automatic interpretation, analysis, and content understanding for the massive RS images. However, the training and testing effectiveness could be curbed owing to the lack of adequate and accurately annotated ground-truth 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.

1) *The ever-growing volume of RS images is acquired while very few of them are annotated with valuable information:* With the rapid development and continuous improvement of sensor technology, it is 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 challenges as well as opportunities [51] for the interpretation of surface features [5], [28], [52]. However, in contrast to the huge amount of received RS images, those annotated with valuable information are relatively few, making them difficult to be productively utilized and also resulting in great waste.
2) *The generalization ability of algorithms for interpreting RS images is of great urgency to be enhanced:* Although a multitude of machine learning [53]–[55] and deep learning algorithms [47], [56], [57] have been developed for RS image interpretation, their interpretation capability could be constrained because of the complexity of RS image content. Besides, existing algorithms are usually trained on small-scale datasets, 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.
3) *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 [28], [58]–[60]. 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 feature distribution in the real world. Thus, it is natural to argue that the representative and large-scale RS image datasets are critical to push forward the development of practical interpretation algorithms, particularly deep learning based methods.
4) *There is a lack of public platforms for systematic evaluation and fair comparison among different interpretation algorithms:* A host of interpretation algorithms have been designed for RS image interpretation tasks and achieved excellent performances. However, many algorithms are designed toward 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.

With these points in mind, this article first provides a review of the available RS image datasets and discusses 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 discussion about challenges and perspectives in RS image annotation. To sum up, our main contributions are as follows.

1) Covering literature published over the past decade, we provide a comprehensive review on the existing RS image datasets concerning the current mainstream of RS image interpretation tasks, including scene classification, object detection, semantic segmentation, and change detection.
2) We present the general guidances, including the dataset property desirability, image acquisition via semantic coordinates collection, and annotation methodology, on creating benchmark datasets for RS image interpretation. The introduced guidances formulate an overall prototype, which we hope to provide a picture for RS image dataset creation with considerations in efficiency, quality assurance, and property assessment.
3) Following the general guidances of dataset creation, we establish the solution of building a scene classification dataset to further verify the practicability of the formed guidances. Consequently, we create a large-scale benchmark dataset for RS image scene classification, i.e., Million Aerial Image Dataset (**Million-AID**), which possesses a million 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 article is organized as follows. Section II reviews the existing datasets for RS image interpretation. Section III presents the guidances of constructing a meaningful annotated RS image dataset. Section IV gives an example of creating large-scale RS image dataset for scene classification. Section V discusses the challenges and perspectives concerning RS image annotation. Finally, in Section VI, we draw some conclusions.

II. Annotated Datasets for RS Image Interpretation: A Review

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

A. RS Image Interpretation Focus in the Past Decade

It is of great interest to check what the main research stream is in RS image interpretation. To do so, we analyzed the journal articles published in the past decade in RS community based on

TABLE I
INVESTIGATED JOURNALS AND NUMBER OF PAPERS

Name of journal	#Ref.
Remote Sensing	1,922
International Journal of Remote Sensing	587
IEEE Transactions on Geoscience and Remote Sensing	575
ISPRS Journal of Photogrammetry and Remote Sensing	536
Remote Sensing of Environment	528
IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing	493
Journal OF Applied Remote Sensing	329
International Journal of Applied Earth Observation and Geoinformation	304
Sensors	277
IEEE Geoscience and Remote Sensing Letters	276

Web of Science database. Specifically, we use "remote sensing" as the keyword to perform *topic* retrieval supported by *tile*, *abstract*, and *keywords*. Then, the retrieved references published in the last decade (i.e., 2011–2020) are gathered and those journals that published most articles ranked top 10 are selected to investigate the mainstream of RS image interpretation. Generally, RS image interpretation is closely related to the work of image/information/content extract/analysis/understanding. Relying on this idea, each term of "image interpretation," "image analysis," "image understanding," "content interpretation," "content analysis," "content understanding," "content extraction," "information extraction," "information analysis," "information interpretation," and "information understanding" was combined with the keyword of "remote sensing" to further screen those interpretation related works by topic retrieval. By excluding the irrelevant results (e.g., review articles), 5827 articles were obtained and then analyzed by CiteSpace [61]. Table I shows the final employed journals and number distribution of investigated references. It is shown that our investigated references are now well presented at the major international RS journals.

Fig. 1 shows the highest frequency terms appearing in the title, keyword, and abstract of the literature. The terms with higher frequency are presented with larger font size. As can be seen from this figure, RS image interpretation works mainly focus on *classification* tasks (e.g., land-cover classification and scene classification). Obviously, *change detection*, *(image) segmentation*, and *object detection* occupy prominent positions in the interpretation tasks. Specially, the terms around the center, e.g., *landsat*, *UAV* (unmanned aerial vehicle), *modis*, *synthetic aperture radar*, and *sentinel**, indicate the commonly used image sources for interpretation tasks. It is worth noting that *feature extraction* plays a significant role in the interpretation of RS images. This makes sense as the feature extraction performed by interpretation models and algorithms, reflected by the terms of *deep learning*, *machine learning*, *convolutional neural network* (CNN), *random forest*, and *support vector machine*, is indispensable to RS image interpretation tasks. Notably, *deep learning* represented by *convolutional neural network* also occupies the center of the tag cloud, where the currently most popular method for RS image interpretation is revealed. And this has heavily promoted dataset construction to advance the development of RS image interpretation. We subsequently filtered the meta articles by "deep learning" and "convolutional neural network."

Fig. 1. Tag cloud of RS image interpretation.

The highest frequency terms match well with Fig. 1, where scene classification, object detection, segmentation, and change detection possess the centrality of interpretation tasks, verified by [57]. Thus, the review given below focuses mainly on datasets concerning these topics.

B. Annotated Datasets for RS Image Interpretation

During the past decade, a number of datasets for RS image interpretation have been released publicly. The available datasets are arranged in chronological order as shown in Tables II– V, in which the corresponding references can be referred for more detailed information about these datasets. Instead of simply delivering descriptions of the datasets, we focus on analyzing the properties of the public RS image datasets from the perspective of annotation.[1]

1) Categories Involved in Interpretation: The interpretation of RS images aims to extract content of interest at pixel-, region-, and scene-level. Usually, the category information of image content is extracted through elaborately designed interpretation algorithms. Hence, some datasets are constructed to recognize common RS scenes [10], [14], [62]–[66], [72] in the earlier years. To extract specific information of objects, there are datasets focusing on one or several main categories [80], [82], [84]–[90], [93], [109], such as vehicle [80], [84], [85], [87], [90], building [82], [97], [98], [100], [110], airplane [85], [93], [109], and ship [91], [93], [99], [101], [106], [108]. The determination of semantic categories plays a significant role in real applications like land classification, urban planning, and environmental monitoring. Hence, a number of datasets are annotated for the purpose of land use and land cover (LULC) or agriculture application [5], [14], [111]–[115]. There are many

[1]We pay our attention mainly to the publicly released and popular RS image datasets, while those for special domains, e.g., contest and private applications, may not be fully covered due to their unstable accessibility or incomplete dataset information.

TABLE II
COMPARISON AMONG DIFFERENT RS IMAGE SCENE CLASSIFICATION DATASETS

Dataset	#Cat.	#Images per cat.	#Instances	Resolution (m)	Image size	GL/IT/SP	Year
UC-Merced [14]	21	100	2,100	0.3	256×256	✗✗✗	2010
WHU-RS19 [10]	19	50 to 61	1,013	up to 0.5	600×600	✗✗✗	2012
RSSCN7 [62]	7	400	2,800	–	400×400	✗✗✗	2015
SAT-4 [63]	4	89,963 to 178,034	500,000	1 to 6	28×28	✗✗✗	2015
SAT-6 [63]	6	10,262 to 150,400	405,000	1 to 6	28×28	✗✗✗	2015
BCS [64]	2	1,438	2,876	–	600×600	✗✗✓	2015
RSC11 [65]	11	∼100	1,232	∼0.2	512×512	✗✗✗	2016
SIRI-WHU [66]	12	200	2,400	2	200×200	✗✗✗	2016
NWPU-RESISC45 [67]	45	700	31,500	0.2 to 30	256×256	✗✗✗	2016
AID [52]	30	220 to 420	10,000	0.5 to 8	600×600	✗✗✗	2017
RSI-CB256 [68]	35	198 to 1,331	24,000	0.3 to 3	256×256	✗✗✗	2017
RSI-CB128 [68]	45	173 to 1,550	36,000	0.3 to 3	128×128	✗✗✗	2017
Planet-UAS [69]	17	–	40,480	3 to 5	256×256	✓✓✓	2017
RSD46-WHU [70]	46	500 to 3,000	117,000	0.5 to 2	256×256	✗✗✗	2017
MASATI [71]	7	304 to 1,789	7,389	–	512×512	✗✗✗	2018
EuroSAT [72]	10	2,000 to 3,000	27,000	10	64×64	✓✓✓	2018
PatternNet [73]	38	800	30,400	0.06 to 4.7	256×256	✗✗✗	2018
fMoW [74]	62	–	132,716	0.5	74×58 to 16,184×16,288	✓✓✓	2018
WiDS Datathon 2019 [75]	2	–	20,000	3	256×256	✗✗✗	2019
Optimal-31 [76]	31	60	1,860	–	256×256	✗✗✗	2019
BigEarthNet [77]	43	328 to 217,119	590,326	10,20,60	20×20;60×60;120×120	✓✓✓	2019
CLRS [78]	25	600	15,000	0.26 to 8.85	256×256	✗✗✗	2020
MLRSN [79]	46	1,500 to 3,000	109,161	0.1 to 10	256×256	✗✗✗	2020

*As *fMoW* is constructed with multiple temporal views for each scene, we ignore the *#Images per Cat.* and count the total number of unique scene instances, *i.e.*, #Instances. Note that *MLRSN* is a multilabel scene classification dataset. The *Cat.*, *GL*, *IT*, and *SP* are short for *Category*, *Geographic Location*, *Imaging Time*, and *Sensor parameter*, respectively. We present the *GL/IT/SP* column to indicate whether the datasets provide those complete and accurate meta information.

TABLE III
COMPARISON AMONG DIFFERENT RS IMAGE OBJECT DETECTION DATASETS

Datasets	Annot.	#Cat.	#Instances	#Images	Resolution (m)	Image width	GL/IT/SP	Year
TAS [80]	HBB	1	1,319	30	–	792	✗✗✗	2008
OIRDS [81]	OBB	5	1,800	900	up to 0.08	256 to 640	✓✓✓	2009
SZTAKI-INRIA [82]	OBB	1	665	9	–	∼800	✗✗✗	2012
NWPU-VHR10 [83]	HBB	10	3,651	800	0.08 to 2	∼1,000	✗✗✗	2014
DLR-MVDA [84]	OBB	2	14,235	20	0.13	5,616	✗✗✓	2015
UCAS-AOD [85]	OBB	2	14,596	1,510	–	∼1,000	✗✗✗	2015
VEDAI [86]	OBB	9	3,640	1,210	0.125	512;1,024	✓✗✗	2016
COWC [87]	CP	1	32,716	53	0.15	2,000 to 19,000	✓✗✗	2016
HRSC2016 [88]	OBB	26	2,976	1,061	–	∼1,100	✗✗✗	2016
RSOD [89]	HBB	4	6,950	976	0.3 to 3	∼1,000	✗✗✗	2017
CARPK [90]	HBB	1	89,777	1,448	–	1,280	✗✗✓	2017
SSDD/SSDD+ [91]	HBB/OBB	1	2,456	1,160	1 to 15	∼500	✗✗✓	2017
SpaceNet1-6* [92]	Polygon	1	859,982	–	up to 0.3	–	✓✓✓	2018
LEVIR [93]	HBB	3	11,028	22,000	0.2 to 1	800	✗✗✗	2018
VisDrone [94]	HBB	10	54,200	10,209	–	2,000	✗✗✗	2018
xView [95]	HBB	60	1,000,000	1,413	0.3	∼3,000	✓✗✓	2018
DOTA-v1.0 [28]	OBB	15	188,282	2,806	up to 0.3	800 to 13,000	✗✗✗	2018
ITCVD [96]	HBB	1	29,088	173	0.1	3,744;5,616	✗✗✗	2018
WHU building dataset [97]	Polygon	1	221,107	25,420	0.075 to 2.7	512	✗✗✗	2018
DeepGlobe Building [98]	Polygon	2	302,701	24,586	0.3	650	✗✗✓	2018
OpenSARShip [99]	Chip	1	11,346	41	∼10	–	✓✓✓	2018
CrowdAI Mapping Challenge [100]	Polygon	1	2,910,917	341,058	–	300	✗✗✗	2018
Airbus Ship Detection Challenge [101]	Polygon	1	∼131,000	208,162	–	768	✗✗✗	2018
iSAID [28], [102]	Polygon	15	655,451	2,806	up to 0.3	800 to 4,000	✗✗✗	2019
HRRSD [103]	HBB	13	55,740	21,761	0.15 to 1.2	152 to 10,569	✗✗✗	2019
DIOR [104]	HBB	20	192,472	23,463	0.5 to 30	800	✗✗✗	2019
DOTA-v1.5 [105]	OBB	16	402,089	2,806	up to 0.3	800 to 13,000	✗✗✗	2019
SAR-Ship-Dataset [106]	HBB	1	5,9535	43,819	up to 3	256	✗✗✓	2019
AIR-SARShip [107]	HBB	1	2,040	300	1;3	1,000	✓✓✓	2020
HRSID [108]	HBB	1	16,951	5,604	0.5;1;3	800	✗✗✓	2020
RarePlanes [109]	Polygon	1	644,258	50,253	0.3	–	✓✗✓	2020
DOTA-v2.0 [105]	OBB	18	1,793,658	11,268	up to 0.3	800 to 20,000	✗✗✗	2020

*For simplicity, we summarize the *SpaceNet1∼6* as a whole, considering their common functionality for building detection. Note that *SpaceNet3/5* are also associated with road network detection. *SpaceNet7* [92] with 11 080 000 and *xBD* [110] with 850 736 building footprints (referenced in Table V) can also be used for building object detection and instance segmentation. *CrowdAI Mapping Challenge* is presented with the train and validation sets for their accessibility. *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.

TABLE IV
COMPARISON OF DIFFERENT RS IMAGE SEMANTIC SEGMENTATION DATASETS

Datasets	#Cat.	#Images	Resolution (m)	#Channels	Image size	GL/IT/SP	Year
Kennedy Space Center [133]	13	1	18	224	512×614	✕✓✓	2005
Botswana [133]	14	1	30	242	1,476×256	✕✓✓	2005
Salinas [126]	16	1	3.7	224	512×217	✕✕✓	–
University of Pavia [126]	9	1	1.3	115	610×340	✕✕✓	–
Pavia Centre [126]	9	1	1.3	115 bands	1,096×492	✕✕✓	–
ISPRS Vaihingen [127]	6	33	0.09	IR,R,G,DSM,nDSM	∼2,500×2,500	✕✕✓	2012
ISPRS Potsdam [127]	6	38	0.05	IR,RGB,DSM,nDSM	6,000×6,000	✓✕✓	2012
Massachusetts Buildings [116]	2	151	1	RGB	1,500×1,500	✓✓✕	2013
Massachusetts Roads [116]	2	1,171	1	RGB	1,500×1,500	✓✓✕	2013
Indian Pines [134]	16	1	20	224	145×145	✓✓✓	2015
Zurich Summer [128]	8	20	0.62	NIR, RGB	1,000×1,150	✓✓✓	2015
SPARCS Validation [120]	7	80	30	11	1,000×1,000	✓✓✓	2016
Biome [122]	4	96	30	11	∼9,000×9,000	✓✓✓	2017
Inria [117]	2	360	0.3	RGB	5,000×5,000	✕✕✕	2017
EvLab-SS [135]	10	60	0.1 to 2	RGB	4,500×4,500	✕✕✓	2017
RIT-18 [136]	18	3	0.047	6	9,000×6,000	✓✓✓	2017
CITY-OSM [119]	3	1,671	0.1	RGB	2,500×2,500 to 3,300×3,300	✕✕✕	2017
Dstl-SIFD* [114]	10	57	up to 0.3	up to 16	∼3,350×3,400	✓✕✓	2017
IEEE GRSS Data Fusion Contest 2017	17	30	1,4	9	643×666;374×515	✓✓✓	2017
IEEE GRSS Data Fusion Contest 2018	20	1	1	48	4,172×1,202	✓✓✓	2018
Aeroscapes [137]	11	3,269	–	RGB	720×1,280	✕✕✕	2018
DLRSD [138]	17	2,100	0.3	RGB	256×256	✕✕✕	2018
DeepGlobe Land Cover [98]	7	1,146	0.5	RGB	2,448×2,448	✕✕✓	2018
So2Sat LCZ42 [139]	17	400,673	10	10	32×32	✓✕✓	2019
SEN12MS [130]	33	180,662 triplets	10 to 50	up to 13	256×256	✓✕✓	2019
95-Cloud [121]	1	43,902	30	NIR,RGB	384×384	✓✕✓	2019
Shakeel et al. [118]	1	2,682	0.3	RGB	300×300	✕✕✕	2019
ALCD Cloud Masks [123]	8	38	10	RGB	1,830×1,830	✓✓✓	2019
SkyScapes [132]	31	16	0.13	RGB	5,616×3,744	✕✕✕	2019
DroneDeploy [140]	7	55	0.1	RGB	up to 12,039×13,854	✕✕✕	2019
Slovenia LULC [141]	10	940	10	6	5,000×5,000	✓✓✓	2019
LandCoverNet [111]	7	1,980	10	NIR,RGB	256×256	✓✓✓	2020
UAVid [142]	8	420	–	RGB	∼4,000×2,160	✕✕✓	2020
GID [5]	15	150	0.8 to 10	4	6,800×7,200	✓✓✓	2020
LandCover.ai [112]	3	41	0.25,0.5	RGB	9,000×9,500;4,200×4,700	✓✕✕	2020
Agriculture-Vision [113]	9	94,986	0.1;0.15;0.2	NIR,RGB	512×512	✕✕✓	2020
S2CMC* [124]	18	513	20	13	1,024×1,024	✓✓✓	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. The *S2CMC* is short for *Sentinel-2 Cloud Mask Catalogue*. The *DSTL-SIFD* is short for the challenge of *Dstl Satellite Imagery Feature Detection*.

TABLE V
COMPARISON OF DIFFERENT RS IMAGE CHANGE DETECTION DATASETS

Datasets	#Cat.	#Image pairs	Resolution (m)	#Channels	Image size	GL/IT/SP	Year
SZTAKI AirChange [143]	2	13	1.5	RGB	952×640	✕✓✕	2009
AICD [144]	2	1,000	0.5	115	800×600	✕✕✕	2011
Taizhou Data [145]	4	1	30	6	400×400	✓✓✓	2014
Kunshan Data [145]	3	1	30	6	800×800	✓✓✓	2014
Cross-sensor Bastrop [146]	2	4	30,120	7,9	444×300; 1,534×808	✓✓✓	2015
MtS-WH [147]	9	1	1	NIR, RGB	7,200×6,000	✓✓✓	2017
Yancheng [148]	4	2	30	242	400×145	✓✓✓	2018
GETNET dataset [149]	2	1	30	198	463×241	✕✓✓	2018
Urban-rural boundary of Wuhan [150]	20	1	4/30	4, 9	960×960	✓✓✓	2018
Hermiston City, Oregon [151]	5	1	30	242	390×200	✓✓✓	2018
OSCD [152]	2	24	10	13	600×600	✓✓✓	2018
WHU building dataset [97]	2	1	0.2	RGB	32,507×15,354	✓✓✓	2018
Season-varing dataset [153]	2	16,000	0.03 to 0.1	RGB	256×256	✕✕✕	2018
ABCD [154]	2	16,950	0.4	RGB	128×128;160×160	✕✓✕	2018
California flood dataset [155]	2	1	5,30	RGB,11	1534×808	✓✓✓	2019
López-Fandiño et al. [156]	5	2	20	224	984×740; 600×500	✕✓✓	2019
xBD [110]	6	11,034	up to 0.8	RGB	1,024×1,024	✓✓✓	2019
HRSCD [157]	6	291	0.5	RGB	10,000×10,000	✓✓✓	2019
LEVIR-CD [158]	2	637	0.5	RGB	1,024×1,024	✕✕✕	2020
SECOND [131]	30	4,214	0.5 to 3	RGB	512×512	✕✕✕	2020
Google Dataset [159]	2	1,067	0.55	RGB	256×256	✓✓✕	2020
Zhang et al. [160]	2	4	2;2.4;5.8	NIR, RGB	1,431×1,431; 458×559; 1,154×740	✓✓✓	2020
Hi-UCD [115]	9	1,293	0.1	RGB	1,024×1,024	–/–/Y	2020
SpaceNet7 [92]	–	24	4	RGB	–	✓✓✓	2020
S2MTCP [129]	2	1,520	up to 10	13	600×600	✓✓✓	2021

semantic segmentation datasets that concern specific categories like building and road [97], [116]–[119], cloud [120]–[124], etc. Some datasets aim to interpret multiple land-cover categories within specific areas, e.g., city areas [115], [125]–[129], relating to intensive human activities. Even with accurate annotation of category information, these datasets are with relatively small numbers of interpretation categories, which can 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. However, there are more semantic categories in practical applications of RS image interpretation. As compensation for this situation, a lot of RS image datasets have been paid efforts to annotate dozens of semantic categories of interest, such as NWPU-RESISC45 [67], AID [52], RSI-CB [68], RSD46-WHU [70], Patternet [73], Optimal-31 [76], fWoM [74], CLRS [78], MLRSNet [79], xVew [95], SEN12MS [130], SECOND [131], and SkyScapes [132], emphasizing broadly on scene-, object-, and pixel-level information. Even with enriched semantic categories, to fully interpret the content of interest in RS images still remains difficult. Taking the LULC application as an example, there are a number of semantic categories and even hundreds of fine-grained classes. As a result, datasets with the limited number of scene categories are not able to extract the various and complex semantic content reflected in RS images. Moreover, categories in these datasets are set equal, while the relationship between different categories, e.g., the including, included, or cross-relationship, is ignored. This inevitably results in the chaotic category organization and management for semantic information. Particularly, the intraclass and interclass relationships are simply neglected in many datasets. Not only that, the context which can reveal the relationship between content of interest and their surrounding environment is rarely considered. Encouragingly, the significant exploration of relation modeling methods for RS image interpretation has been developed to address these issues [45]. Nevertheless, how to annotate datasets with rich semantic categories and reasonable relationship organization strives to be a key problem for practical dataset construction.

2) Dataset Annotation: To our knowledge, most of the datasets listed in Tables II–V are manually annotated by experts. Generally, the work of dataset annotation is to assign semantic tags to scenes, objects, or pixels of interest in RS images. For the task of scene classification, a category label is typically assigned to the scene components by visual interpretation of experts [52], [67]. In order to recognize specific objects, entities in images are usually labeled with closed areas. Thus, many existing datasets manually annotate objects in the form of bounding boxes, e.g., NWPU-VHR10 [83], RSOD [89], HRRSD [103], and DIOR [104], or enclosed polygons, e.g., iSAID [102] and xBD [110]. Before annotating content of interest, a fundamental issue is the acquisition of target RS images in which the intriguing content is contained. Usually, the target images are manually searched, distinguished, and screened in the image database by trained annotators. Along with the subsequent label assignment, the whole annotation process in the construction of RS image datasets is time-consuming and labor-intensive, especially for the pixelwise annotations as shown in Tables IV and V. As a result, dataset construction, from source image collection, semantic information annotation, and quality review, relies heavily on manual operations, making it an expensive project. This raises an urgent demand for developing more efficient and assistant strategies to lighten the burden of artificial annotation.

When it comes to the annotation tools, there is a lack of visualization methods for the annotation of large-scale and hyper-spectral RS images. Currently, annotation tools designed for natural images, e.g., LabelMe [161] and LabelImg [162], are introduced to annotate RS images. These annotation tools typically visualize an image with a 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 conduct the labeling operations within a local region of the RS image. In this situation, inaccurate annotation could be produced since it is difficult for the annotator to grasp the global content of the RS image. Meanwhile, the image roam process will inevitably constrain annotation efficiency. This problem is particularly serious when conducting annotation for semantic segmentation and change detection tasks where labels are typically assigned pixel-by-pixel [127], [134], [142], [143]. On the other hand, hyper-spectral RS images [125], [127], [133], [134], [148], [150]–[152], which characterize objects with rich spectral signatures, are usually employed for elaborate interpretation of semantic content. However, it is hard to label the hyper-spectral RS images since annotation tools developed for natural images are not able to visualize hyper-spectral images of hundreds of spectral bands. Therefore, universal annotation tools are desperately desired to be developed for efficient and convenient semantic annotation, especially for the large-scale and hyper-spectral RS images.

3) Image Source: A wide group of RS images has been employed as the source of interpretation datasets, including the optical, multi-/hyper-spectral, SAR images. Typically, the optical images from Google Earth are widely employed as the data standard, such as those for scene classification [14], [52], [62], [65]–[67], [70], [73], object detection [28], [29], [80], [85], [88], [89], [93], [104], and pixel-level analysis [116], [117], [128]. In these scenarios, RS images are typically interpreted by the visual content, of which the spatial pattern, texture structure, information distribution, as well as organization mode are more concerned. Although the Google Earth images are postprocessed with RGB formats using the original optical aerial images, they possess the potential for pixel-based LULC interpretation as there is no general statistical difference between the Google Earth images and optical aerial images [163]. Thus, Google Earth images can also be used as RS images for evaluating interpretation algorithms [52].

Different from the optical RS image datasets, the construction of hyper-spectral and SAR image datasets should adopt the original data formats. Compared to optical images, multi-/hyper-spectral images can capture the essential characteristics of ground features as the rich spectral and spatial information are simultaneously involved. Therefore, the content interpretation of hyper-spectral RS images is mainly based on the spectral properties of ground features. Naturally, this kind of images is typically employed to construct the dataset for subtle semantic

information extraction, such as semantic segmentation [125], [127], [130], [133], [134], [139] and change detection [41], [129], [148], [150]–[152], where more attention is paid to the knowledge of the fine-grained compositions. For SAR images acquired by microwave imaging, 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. And it is not encouraged to employ the modified data of SAR images for visual interpretation of interested content. It is worth noting that the advantages of different RS images can be integrated. This is why the multimodal learning framework has drawn much attention and been employed to greatly improve the performance of RS image interpretation [50], which provide significant reference for making the most of different RS image datasets, especially those from different imaging sensors.

4) Dataset Scale: A large number of RS image datasets have been constructed for various interpretation tasks. 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 an RS image dataset. RS images that are typically taken from the bird-view perspective have a 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 II–V, to fit specific models that are designed to extract features within the limited scale of images. In fact, the preservation of the original image size is more close to real-world applications [5], [28]. Some datasets with larger image sizes, say, width/height of a few thousand pixels, are limited with the number of annotated images or categories [87], [136], [147], [153], [157], [158]. Furthermore, quite a few datasets contain one or several images, especially those for semantic segmentation [125], [133], [134] and change detection [97], [145]–[151], [155], [156], [160], which are limited by the high cost of pixelwise annotation. 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, existing datasets often show deficiencies in image variation and sample diversity. Typically, content in RS images always shows 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 [64], [70], [84], [133]. In addition, content reflected in RS images are with complex texture, structure, and spectral features owing to the high complexity of the Earth's surface. Thus, datasets with limited images and samples [14], [65], [82], [87], [133], [136] are usually not able to completely characterize the properties of objects of interest. As a result, there is a lack of representativeness of real-world scenarios for datasets with small scales. This can lead to weak interpretation ability of algorithms with the change of application scenarios. Furthermore, constrained by the scale of datasets, the currently popular deep learning approaches are usually pretrained using the large-scale natural image datasets, e.g., ImageNet [58], and then used for RS image interpretation [164], [165]. Nevertheless, features learned by this strategy are hard to completely adapt to RS data because of the essential difference between RS images and natural images. For instance, the change of object orientation is common to be observed in RS images. All of these raise an urgent demand for annotating large-scale RS datasets with rich images to advance RS image interpretation.

III. Guidances of Building RS Image Benchmarks

The availability of a good RS image dataset has been shown critical for effective feature learning, algorithm development, and high-level semantic understanding [58]–[60], [166], [167]. More than that, the performance of almost all data-driven methods relies heavily on the training dataset. However, constructing a large-scale and meaningful image dataset for RS image interpretation is not an easy job, at least from the points of technology and cost factors. The challenge lies largely in the aspect of efficiency and quality control. The absence of systematic work involving these problems has limited the construction of practical datasets and continuous advancement of interpretation algorithms in RS community. Therefore, it is valuable to explore the feasible scheme for creating a practical RS image dataset. We believe that the following introduced aspects can be taken into account when creating a desirable dataset for RS image interpretation.

A. Desirable Properties of Benchmark Datasets

In order to enhance the practicality, the dataset for RS image interpretation should be created toward practical application requirements rather than the characteristics of interpretation algorithms. Essentially, the creation of RS image dataset aims at model training, testing, and screening for practical applications. It is of great significance to get the 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 designed interpretation model. To this end, the created dataset should consist of sufficient and accurately annotated samples that cover the challenges in practical application scenarios.

In this point of view, the annotation of RS image dataset is better to be conducted by the application sides rather than the algorithm developers. Annotations by algorithm developers will inevitably possess bias as they may be more familiar with the algorithm properties and lack of understanding of challenges lying in practical applications. As a result, the annotated dataset from developers could be at risk of being algorithm-oriented. On the contrary, the application sides have more opportunities to access the real application scenarios and, thus, are more familiar with the issues and challenges lying in the interpretation tasks. Therefore, the dataset annotation from application sides is more reliable and, thus, conducive to enhance the practicability of the interpretation algorithm.

In general, the RS image dataset should be constructed toward the real-world scenarios instead of the specific algorithms. Thus, it is possible to feed the interpretation system with high-quality data, which boost the interpretation algorithms to effectively learn and even extend knowledge that people desired. With

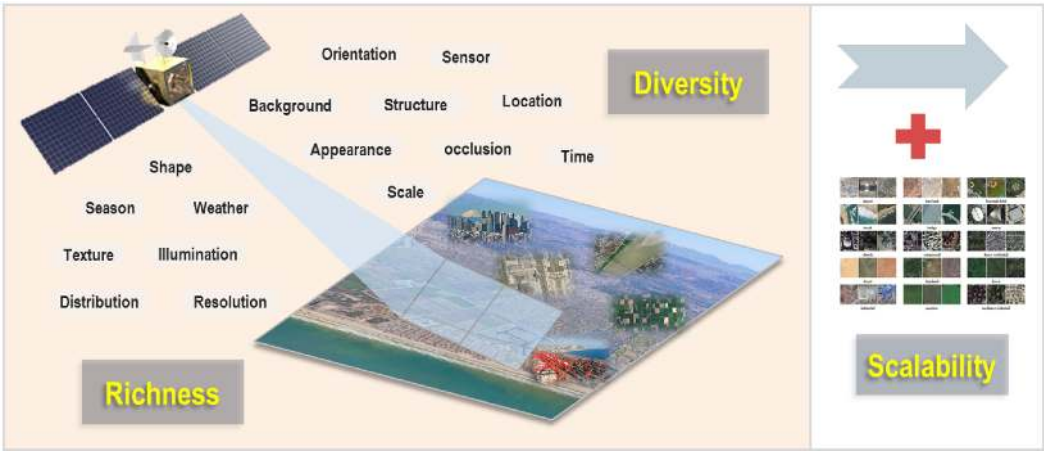


Fig. 2. **DiRS** properties: **di***versity*, **r***ichness*, and **s***calability*. DiRS formulates the basic properties which can be considered as basic desirability in the construction of datasets for RS image interpretation. We believe that these properties are complementary to each other. That is,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

these points in mind, we believe that the **di***versity*, **r***ichness*, and **s***calability* (called **DiRS**), as illustrated in Fig. 2, could be considered as the desirable properties when creating benchmark datasets for RS image interpretation.

1) Diversity: A dataset is considered to be diverse if its annotated objects depict various visual characteristics of relevant semantic content with a certain degree of complementarity. From the perspective of within-class diversity, annotated objects with large diversity are able to comprehensively represent content distribution in real world. To this end, it is better that each annotated object could reflect different attributes rather than the repeated characteristics. For example, the annotated objects in the same category, e.g., vehicle, can be distinguished from each other in properties like appearance, scale, and orientation that diversify the instances. Thus, the within-class objects of large diversity are conducive for an algorithm to learn the essential characteristics. In addition, it should be emphasized that the imaging and geographic properties are also desperately desired for dataset diversity improvement. In the real world, the properties of objects of interest could vary with its geographic location and imaging time. A fact is that the object of the same class could show differences in state, surroundings, and position with the spatio-temporal property variation. Thus, the imaging and geographic properties become nonnegligible when building an interpretation dataset of high diversity. Especially, this is very important for the large-scale geographic application using the method learned from a given dataset. Therefore, annotated objects with these distinct characteristics are able to provide insurance of training interpretation model with a powerful ability of feature representation and application generalization. In this regard, the within-class diversity actually emphasizes on the individual differences between objects of interest in the same class.

On the other hand, in order to learn an interpretation algorithm for effective discrimination of different classes, the between-class diversity should also be taken into consideration when constructing the RS image dataset. For this requirement, the fine-grained classes, particularly those with high semantic overlapping, should be contained as many as possible. Objects of different semantic classes usually take specific feature pattern and distribution. Thus, annotating objects with diverse semantic classes can enable an interpretation model to learn more powerful feature representation. Besides, high semantic overlapping in different categorical objects means large between-class similarity. It is easy to understand that the notable intervals of content features can make an interpretation model learning to distinguish different classes effortlessly. In contrast, objects with high semantic overlapping, denoting the small distance of different classes, will put forward higher requirements for interpretation models to discriminate similar semantic content. From this point of view, the between-class diversity pays more attention to the common characters among objects of different classes. Generally, the within-class and between-class diversities simultaneously offer the guarantee for feature complementarity and peculiarity for annotated objects, which is crucial for constructing datasets of large diversity.

2) Richness: In addition to the diversity that emphasizes on the otherness of objects, the richness of a dataset is another significant property, which attaches importance to the variation of images. Specifically, the rich image variation regards various content characteristics and large-scale samples as important when constructing an RS image dataset. In order to enrich the content characteristics, images can be collected under various circumstances, such as the weather, season, illumination, imaging condition, and sensor, which allow the dataset to possess rich variations in translation, viewpoint, object pose, spatial resolution, illumination, background, occlusion, etc. Not only that, images collected from different periods and geographic regions can also endow the dataset with rich spatio-temporal distribution.

Moreover, different from natural images that are usually taken from horizontal perspective with narrow extent, RS images are taken with bird-views, endowing the images with large geographic coverage, abundant ground features, and complex background information. Thus, an interpretation dataset is desired to contain images that reflect the rich characteristics, e.g., variation in geometrical shape, structure characteristic, textural attribute, etc. From this point of view, the constructed dataset

should consist of large-scale images to contain sufficient annotated samples, which is able to further ensure its comprehensive representativeness for real-world scenarios. The reality is that insufficient images and samples are more likely to lead to the overfitting problem in model training, particularly for data-driven interpretation methods (e.g., CNN). In this regard, the scale of an RS image dataset should be large enough to ensure the richness property. Thus, 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) 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 LULC. 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 application requirements. That is, a constructed benchmark dataset for RS image interpretation is better to be flexible and extendable, considering the change of application scenarios.

Notably, there is a large number of RS images received every day, which need to be efficiently labeled with valuable information to maximize their application value. To this end, the organization, preservation, and maintenance of annotations and images are of great significance to be controlled for the scalability of a dataset. Besides, it would be preferable if the newly annotated images could be involved in the constructed dataset effortlessly. Thus, the full operations of adding, updating, removing, and retrieving data and information in the constructed dataset become a significant property for scalability. With these considerations, a constructed RS image dataset with excellent scalability can be conveniently adapted to the changing requirements for real-world applications without impacting its inherent accessibility, thereby assuring sustainable utilization even as modifications are made.

B. Semantic Coordinates to Facilitate Image Acquisition

The acquisition of RS images that contain content of interest formulates the foundation of creating an interpretation dataset. Benefiting from the spatial property possessed by RS images, the RS images in the database can be accessed by utilizing their inherent information of geographic coordinates [74], [168]. And, further, a geographic feature is commonly presented with a series of geographic coordinates. Meanwhile, the feature is usually attached with specific tag attributes that present its semantic meaning. From this perspective, the geographic coordinates related to a specific feature element can be regarded as the semantic coordinates, by referencing the feature's tag attributes. Thus, we are able to collect the geographic coordinates and then access the corresponding tag attributes to efficiently identify the locations of RS images that contain content of interest.

Typically, this strategy can be performed to prepare a public optical RS image dataset, by utilizing the public map application interface, open source data, and public geodatabases. The coordinates collection may not be an optimal strategy but can also be employed as a reference when creating a private dataset of which images are from other sensors and databases.

1) Map Search Engines: A convenient way to collect RS images with content of interest is to utilize public map search engines, such as Google Map,[2] Bing Map,[3] and World Map.[4] As common digital map service solutions, they provide satellite images covering the whole world in different spatial resolutions. Many existing RS datasets, such as UCM [14] and NWPU-RESISC45 [67] for scene classification, LEVIR [93] and DOTA [28] for object detection, and Google Dataset [159] and LEVIR-CD [158] for change detection, have been built based on Google Map. When collecting RS images on such map search engines, the developed map application programming interface (API) 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 locations. The corresponding satellite images can be accessed using the coordinates of search results. Then, the acquired satellite images can be used to annotate airport scene and aircraft object samples.

2) Open Source Data: Open source geographic data is established on the global positioning system information, aerial photography images, other free content, and even local knowledge (such as social media data) from users. Open source geographic data, such as the Open Street Map (OSM) and WikiMapia, are 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, has low cost, and has a large amount in quantity compared with the manual collection strategy for RS images [68], [119]. 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. Then, we can match the collected geographic elements with their corresponding RS images. Moreover, the extracted geographic elements of interest 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) Geodatabase Integration: Different from the collection of natural images, which can be conveniently accessed through web crawling, search engines (e.g., Google image search), and sharing databases (e.g., Instagram, Flickr), the acquisition of RS images that contain content of interest is difficult because of the

[2]Online. [Available]: https://ditu.google.com
[3]Online. [Available]: https://cn.bing.com/maps
[4]Online. [Available]: http://map.tianditu.gov.cn

high searching cost. Nevertheless, the public geodatabases and geographic information products released by state institutions and communities usually provide accurate and rich geographic data. With this facility, the geographic coordinates attached with specific semantic information can be obtained through these databases. For example, the National Bridge Inventory[5] 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 geographic coordinates of bridges for the collection of bridge images. By integrating these kinds of public geodatabases, we are able to obtain the geographic locations of RS images with specific semantic information and, thus, efficiently collect a large number of RS images that contain content of interest at a relatively low cost.

C. Annotation Methodology

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

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.

a) *Manual annotation:* The common way to create an image dataset is to employ the manual annotation strategy. The great advantage of 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 [14], [52], [65], [66], object detection [28], [82], [86], and semantic segmentation [5], [142]. Regardless of the source from which the natural or RS images are acquired, the way to annotate content in RS images is similar. And many tools have been built to relieve the monotonous annotation work. Hence, image annotation tools developed for natural images can be further introduced for RS images (typically, the optical RS images) to pave the way for cost-effective construction of large-scale datasets. The resource concerning to image annotation tools will be introduced in Section V.
   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 [58]. To relieve this problem, crowd-sourcing annotation becomes an alternative solution that can be employed to create a large-scale image dataset [60], [74], [95] while paying efforts to its challenge with quality control. Besides, benefiting from excellent ability of image interpretation algorithms, annotators can also resort to machine learning schemes [169], [170], which can be integrated as the preliminary annotation, to speed up the efficiency of manual annotation.

b) *Automatic annotation:* In contrast to natural images, RS images are often characterized with complex structures and textures because of the spectral and spatial variation. It is difficult to annotate semantic content 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. In this situation, automatic image annotation methods are naturally employed to alleviate annotation difficulties and further reduce the cost of manual annotation [171].
   Automatic annotation methods reduce the cost of annotation by leveraging learning schemes [172]–[177]. In this strategy, a certain number of images are initialized to train an interpretation model, including the fully supervised [178] and weakly supervised methods [179]–[181]. The candidate images are then poured into the established model for content interpretation and the interpretation results finally serve as annotation information. And iterative and incremental learning [182] can be employed to filter noisy annotation and enhance the generalization ability of annotation model [180], [183]–[185]. Nevertheless, one disadvantage of automatic annotation is that the generalization ability of the annotation model can be affected by both the quality of the initial candidate images. In addition, to decompose the difficulty of annotation and enhance the connectivity between annotation and real applications, the existing semantic information, e.g., thematic products as a unique presentation for RS image content, can serve as the source for automatic RS image annotation and content update [186]. With the inherent semantic information contained in thematic products, reliable training samples are able to be extracted [187]. And this idea has also been successfully employed in dataset construction, e.g., BigEarth [77], which shows promising prospect in the automatic annotation of large-scale dataset for RS image interpretation.

c) *Interactive annotation:* In the era of big RS data, annotation with human–computer interaction, which falls in semiautomatic annotation, could be a more practical strategy considering the demand for RS image annotation with high quality and efficiency. In this strategy, an initial framework can be constructed using the existing archives with available annotation and then employed to annotate the unlabeled RS images. On this basis, the performance of an annotation model can be improved greatly with the intervention from annotators [188]. The intervention from annotators can be in the form of relevance feedback or identification of the relevant content in the images to be annotated. In this scheme, the overall performance of the annotation models mostly depends on the time that annotators spend on creating annotations [189].
   By employing active learning strategy [190], [191] and setting restrict constraints, those images that are difficult to be interpreted can be screened out and then manually annotated by experts. The received feedback can then be used to purify the annotation model through a loop learning way. Consequently, a large number of annotated images can be acquired to optimize the interpretation model and further boost the annotation task in an iterative way. With

[5]Online. [Available]: https://www.fhwa.dot.gov/bridge/nbi.cfm

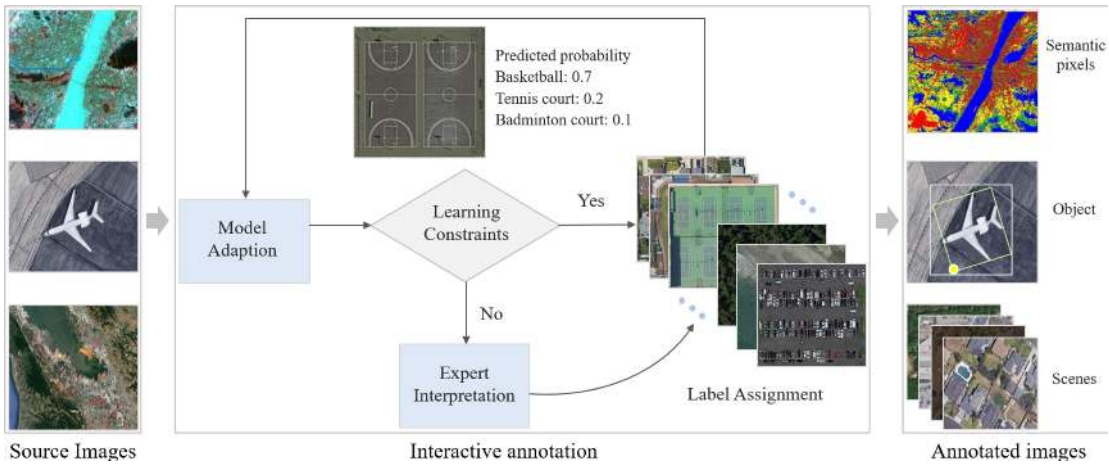


Fig. 3. General workflow of semiautomatic annotation in RS images.

the iteration process, the number of images to be annotated will be greatly reduced to relieve annotation labor. The general workflow of semiautomatic image annotation is shown in Fig. 3. Benefiting from the excellent feature learning ability, deep learning based methods can be developed for image annotation with significant improvement of quality and efficiency [170]. Instead of annotating the full image, human intervention by simple operations, e.g., point-clicks [192], boxes [193], and scribbles [194], can significantly improve the efficiency of interactive annotation. By utilizing the semiautomatic annotation strategy, a large-scale annotated RS image dataset can be constructed efficiently and also with quality assurance owing to the involvement of human labor.

2) Quality assurance: The dataset with high annotation quality is important for developing and evaluating interpretation algorithms. The following introduced strategies can be employed for the quality control when creating a dataset for RS image interpretation.

a) *Rules and samples:* The annotation rules without ambiguity are the guarantee of creating a high-quality dataset. Specifically, annotation rules like category definition, annotation format, viewpoint, occlusion, image quality, and others should be explained clearly. For example, whether to exclude the objects in occlusion and whether to annotate the objects of small sizes. If there are no clear rule descriptions, different annotators will annotate the image with their individual preferences [59]. For annotation in RS images, it is difficult for annotators to recognize the categories of ground features if they have no professional backgrounds. Therefore, samples are better to be provided by experts in the field of RS image interpretation and then presented to annotators as references.

b) *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 to meet the articulated requirements. Those annotators that failed to pass the test cannot be invited to participate in the later annotation project. With such a design, dataset builders are able to build an excellent annotation team. Taking xView [95] as an example, the annotation accuracy of objects is vastly improved with trained annotators. Therefore, the training of annotators can be a reliable guarantee for high-quality image dataset annotation.

c) *Multistage pipeline:* A series of different annotation operations are easy to cause fatigue and result in annotation errors. To avoid this problem, the pipeline of multistage 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 [60]. By this method, each annotator only needs to focus on one simple stage during the whole annotation project and the error rate can be effectively decreased.

d) *Grading and reward:* A comprehensive evaluation of annotators can be performed with the annotation result. For example, the analysis of an annotators' behavior, e.g., the required time per annotation stage and the amount of annotation result over a period, can be conducted to assess the potentially weak annotations. Thus, different types of annotators can be identified, e.g., spammers, sloppy, incompetent, competent, and diligent annotators [195]. Then, incentive mechanism (e.g., financial payment) can be employed to reward the excellent annotators and eliminate the inferior labels from unreliable annotators.

e) *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 [58]. To acquire high-quality annotations, majority voting can be utilized to merge multiple accurate annotations [196]. 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.

f) *Annotation review:* Another effective method to ensure the annotation quality is to introduce the review strategy, which is usually integrated among other annotation pipelines when creating a large-scale image dataset [161].

Specifically, some annotators can be invited to conduct peer review and rate the quality of the created annotations. Besides, further review work can be conducted by experts with professional knowledge. Based on the reviews of supervisors in each annotation step, the overall annotation quality can be strictly controlled in the whole annotation process.

g) *Spot check and assessment:* To check the annotation quality, a test set can be sampled from the annotated images. Also, gold data can be created by sampling and labeling a proper proportion of images annotated by experts. Then, one or several interpretation models can be trained based on these datasets, and the interpretation performance (e.g., $Recall$ and $Precision$ for object detection [28], [29]) can be evaluated to compare annotation from annotators and gold data from experts. If the evaluation result is lower than the preset expectation, annotations from the corresponding annotator would be rejected and required to be resubmitted for repetitive annotation.

IV. Example: Million-AID

Following the aforementioned prototype for building benchmark datasets for RS image interpretation, in this section, we present an example to construct a large-scale benchmark dataset for RS scene classification, i.e., the Million-AID. Limited by the scale of scene images and 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 reliable dataset for scene classification in RS community. In this section, we build Million-AID in the spirit of **DiRS**. And the introduced coordinates collection strategy is employed for efficient scene image acquisition. The dataset quality is guaranteed with a handful of human labor, which finally formulates a semiautomatic and reproducible framework for the construction of RS image scene dataset. The constructed Million-AID will be released for public accessibility.

A. Scene Category Organization

1) Main Challenges in Application: Benefiting from the advancement of RS technologies, the accessibility of RS images has been greatly improved. However, the construction of a large-scale scene classification dataset still faces challenges in aspects like scene taxonomy and image diversity. Obviously, a complete taxonomy of RS image scenes is better to have wide coverage of categorical space since there are a large number of semantic categories in practical applications, e.g., LULC. With various scene images in different categories, the completeness of a scene taxonomy is also significant to enhance the diversity of the dataset. Thus, the determination of scene categories is of great significance to construct a high-quality and practical RS image dataset for scene classification. Some existing datasets, such as the UCM [14], RSSCN7 [62], and RSC11 [65], contain limited scene categories, which make them not sufficiently represent the diverse content reflected by RS images. Consequently, the scene classification models learned from datasets of limited categories usually show weak generalization ability.

When facing practical applications, the excellent organization of scene categories is an important feature for scalability and continuous availability of a large-scale RS image dataset. Typically, the semantic categories that are closely related to human activities and land utilization are selected for the construction of scene categories. Because of the complexity of RS image content, 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 dataset builders. Therefore, effective organization of scene categories should be of great significance to construct an RS image dataset of high quality and scalability.

2) Scene Category Network: Faced with the above challenges, we build a hierarchical network to manage the categories of RS image scenes, as shown in Fig. 4. To satisfy the requirements of practical application rather than the classification algorithms, we construct the scene category system by referencing to the land-use classification standards of China (GB/T 21 010-2017). Considering the inclusion relationships and content discrepancies of different scene categories, 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 semantic category branch. Thus, the scene classification dataset can be constructed with category independence and semantic completeness.

As shown in Fig. 4, the proposed category network is established upon a multilayered 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 content. 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 land, commercial land, industrial land, public service land, residential land, transportation land, unutilized land, and water area. Benefiting from the hierarchical structure of category network, the scene labels from the parent nodes can be directly assigned to the images belonging to the corresponding leaf nodes. Therefore, each image will possess semantic labels with different category levels. This mechanic also provides potentiality for scene classification at flexible category levels.

As can be seen, the category definition and organization can be achieved by the proposed hierarchical category network. The synonyms of the category network are relevant to the practical application of LULC and hardly need to be purified. One of

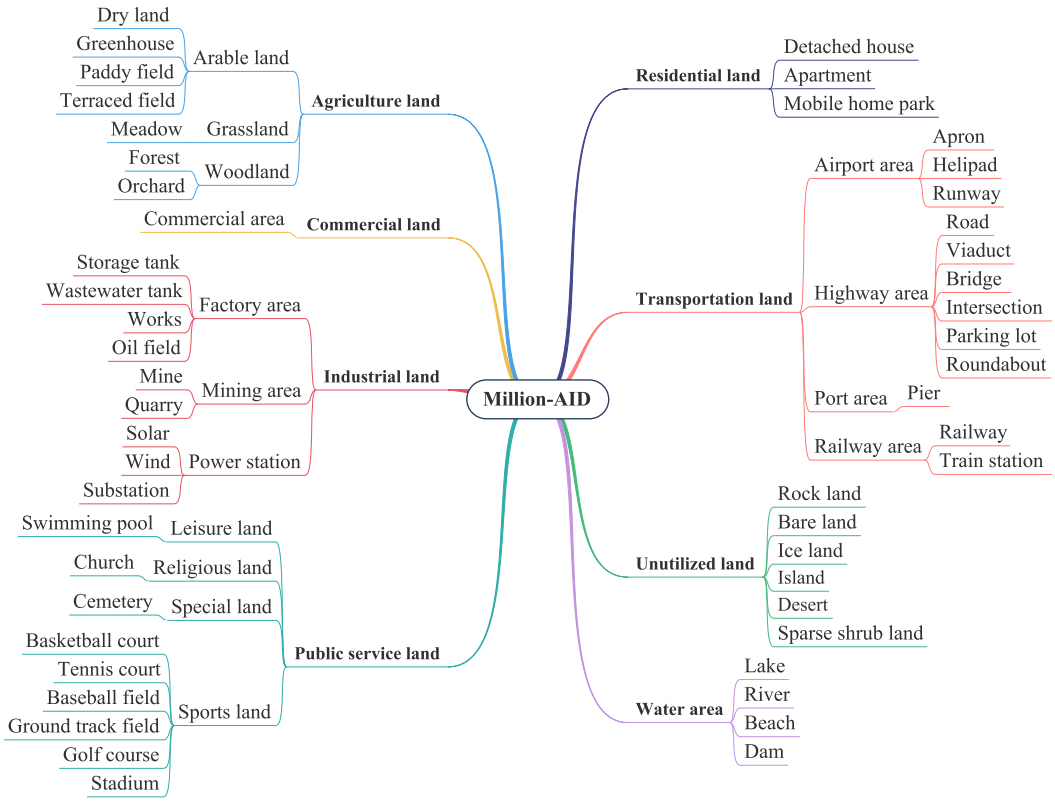


Fig. 4. Hierarchical scene category network of Million-AID. All categories are hierarchically organized in a three-level tree: 51 leaf nodes fall into 28 parent nodes at the second level which are grouped into eight nodes at the first level, representing the 8 underlying scene categories of agriculture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

the most prominent advantages of the category network lies in its semantic structure, i.e., its ontology of concepts. Hence, a new scene category can be easily embedded into the constructed category network as a new branch of synonym. The established category hierarchy can not only serve as the category standard for Million-AID dataset, but it also provides a valuable reference for dataset construction toward other interpretation tasks. Thus, these properties endow our proposed dataset with high practicability when facing real applications.

B. Semantic Coordinates Collection

In the conventional pipeline of constructing a scene classification dataset, one needs to manually search the target region that contains specific scenes. Then the scene images are collected from the image database. However, finding the target region with given semantic 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 alleviate this problem, we employ the introduced coordinates collection strategy and interactive annotation methodology to build the scene classification dataset. Specifically, we employ public map search engines, open sourced data, and public geodatabase resources to collect and label RS scene images. With the rapid development of geographic information and RS technologies, there are rich and publicly available geographic data like online map, open source data, and archives published by agencies as introduced before. Typically, these public geographic data present the surface features in forms like point, line, and plane, which describe the semantic information of ground objects and carry corresponding geographic location information. Based on the public 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. In the case of our practice, different methods are presented to acquire the labeling data. Google Map API and publicly available geographic data are mainly employed to obtain the coordinates of point features, while OSM API is mainly utilized to acquire the coordinates of line and plane features. In application, these methods can be combined to obtain different coordinate data of different forms. The acquired coordinates are then integrated into block data which presents the scene extent. Finally, the block data are further processed to obtain scene images which are automatically assigned with scene labels.

1) Point Coordinates: The point 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, e.g., over 100 million places. This provides a powerful solution to search the ground objects with specific semantic tags. Therefore, we develop a semantic tag search tool based on the Google Map API. With the customization search tool, we input semantic tags to retrieve corresponding point objects using the online map search engine and obtain

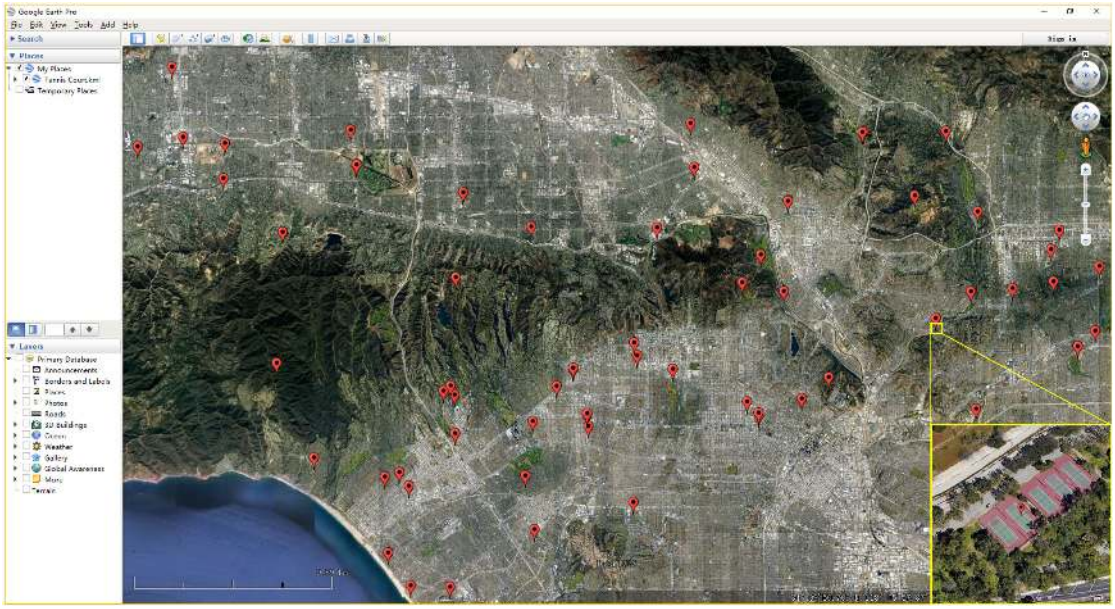


Fig. 5. Points of searched tennis courts shown in Google Earth Pro (©*2020 Google LLC.*), where the top-left and bottom-right coordinates are (34.1071° N, 118.3605° W) and (33.9823° N, 118.3605° W), respectively. We consider the tennis courts as point ground features. The red marks show the searched locations of tennis courts. The eagle window shows the detail of a tennis court scene, which confirms the validity of collecting semantic coordinates by our proposed method.

the geographic coordinates that match the semantic information within a certain range. The retrieved point results with location information, i.e., geographic coordinates, are naturally attached with scene tags. Fig. 5 shows the search result returned by the semantic tag "baseball field" based on the tool. To enhance the diversity of the dataset, we search points of interested objects through a wide range of geographic areas. This strategy makes it possible to cover individual point objects in distinct positions, which is able to greatly enhance the within-class diversity and quickly obtain a large number of points with semantic tags.

The map search engines have provided a powerful interface for accessing point data. However, many of them are associated with categories of common scenes, which will limit the diversity of dataset. For those scenes related to specific scene categories, it is reasonable to employ the publicly available geographic information and obtain the point data. Using the online platforms that publish geographic dataset, we collect the coordinate data of storage tanks, bridges, and wind turbines. For example, the U.S. Wind Turbine Database (USWTDB)[6] provides a large number of locations of land-based and offshore wind turbines in the United States. Fig. 6 shows 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 wind turbine scene is obtained. Consequently, with the publicly released geographic information, we are able to employ the strategy of geodatabase integration to collect coordinates of specific scenes.

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 planes. In order to obtain the scene images of line and plane features, we employ the open source data for scene coordinates collection as introduced before. Specifically, the OSM is utilized 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 by pair values of latitude and longitude. The way feature is composed of two or more connected nodes. 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 other relations. Every node or way has tags and geographic information that describe the corresponding ground object. Therefore, a line or plane feature that belongs to certain semantic classes can be obtained by searching its 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 line and plane features directly from the free, community-maintained data, e.g., Geofabrik,[7] produced by OSM. Fig. 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 scenes in the constructed network, we also search features by utilizing the powerful query language. The query criteria are associated with location information, type of objects, proximity of tag properties, and their combinations.

[6]Online. [Available]: https://eerscmap.usgs.gov/uswtdb

[7]Online. [Available]: http://www.geofabrik.de/geofabrik

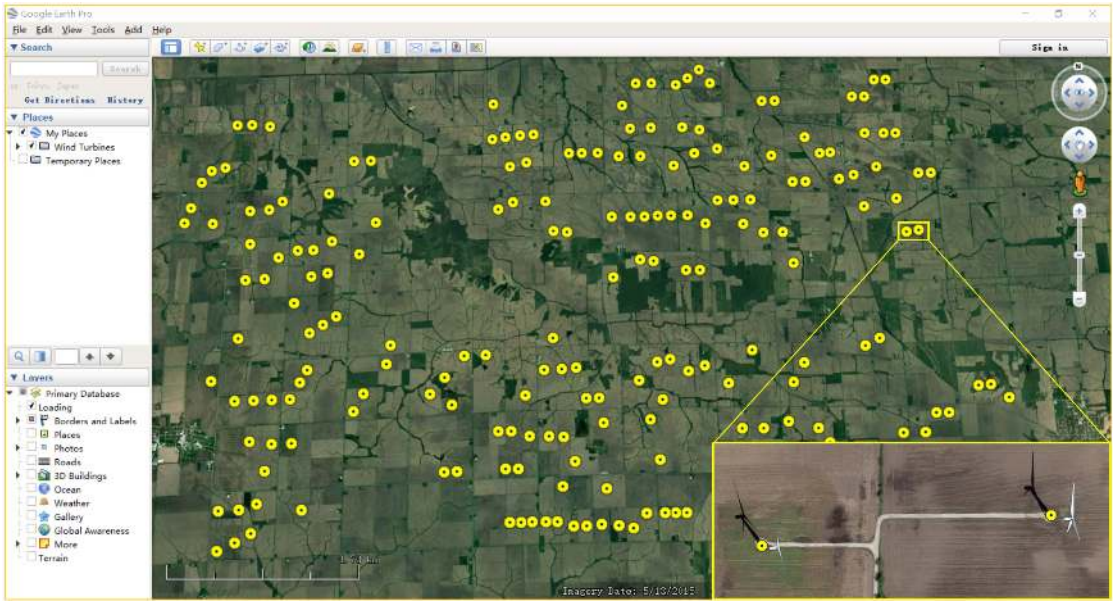


Fig. 6. Points of wind turbines extracted from USWTDB and integrated in Google Earth Pro (©*2020 Google LLC.*), where the geographic range is indicated with the top-left coordinates (41.2695° N, 90.3315° W) and bottom-right coordinates (41.1421° N, 90.0424° W). The eagle window shows the details of two wind turbines.

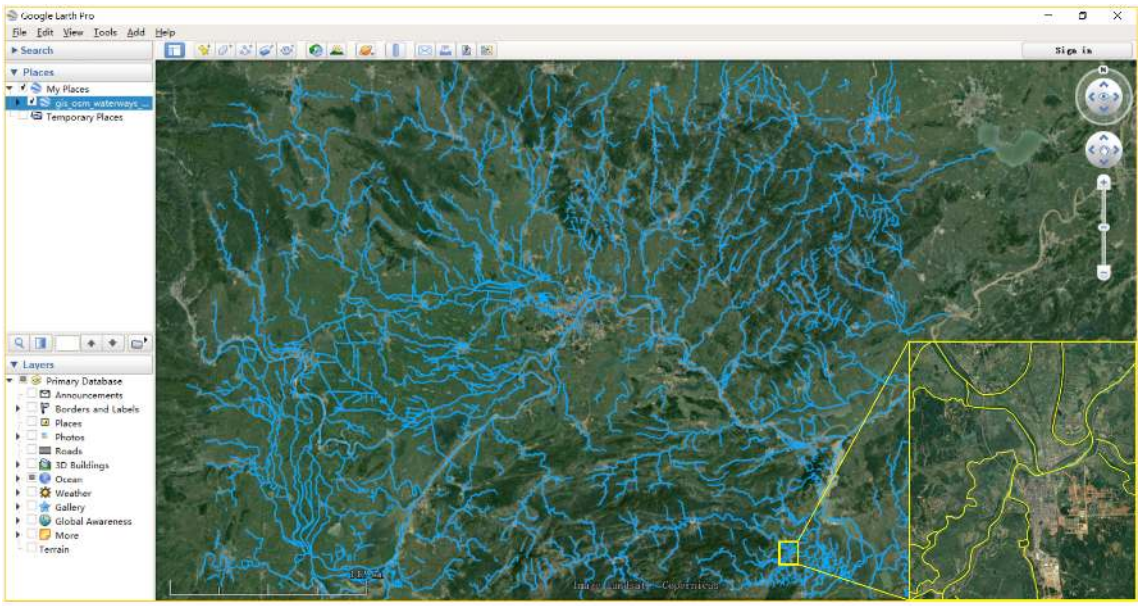


Fig. 7. River lines within a local area of China are extracted from OSM and displayed in Google Earth Pro (©*2020 Google LLC.*), where the upper-left and bottom-right coordinates of geographic range are (32.2826° N, 111.1027° E) and (28.7477° N, 118.5372° E), respectively. The zoomed image shows the details of river lines.

Fig. 8 shows the illustration of searching scenes of airport areas around the world. And the searched airports within a local area of the United States are integrated into Google Earth as shown in Fig. 9. It can be seen from Figs. 8 and 9 that the extracted plane data is consistent with the real-world airport scenes, and, thus, the semantic label is reliable. These results indicate that the former introduced method of employing the open source data is a practical, efficient, and reliable strategy for scene image acquisition via the collection of semantic scene coordinates.

C. Scene Image Acquisition

The geographic point, line, and plane coordinates collected through the above processes are employed to extract scene images from Google Earth. Fig. 10 illustrates the overall framework of collecting RS scene images. For the searched point data, the coordinates are attached with specific tags of semantic categories and we take the geographic coordinates as the center of a square box. For line data, we sample the points along the line by

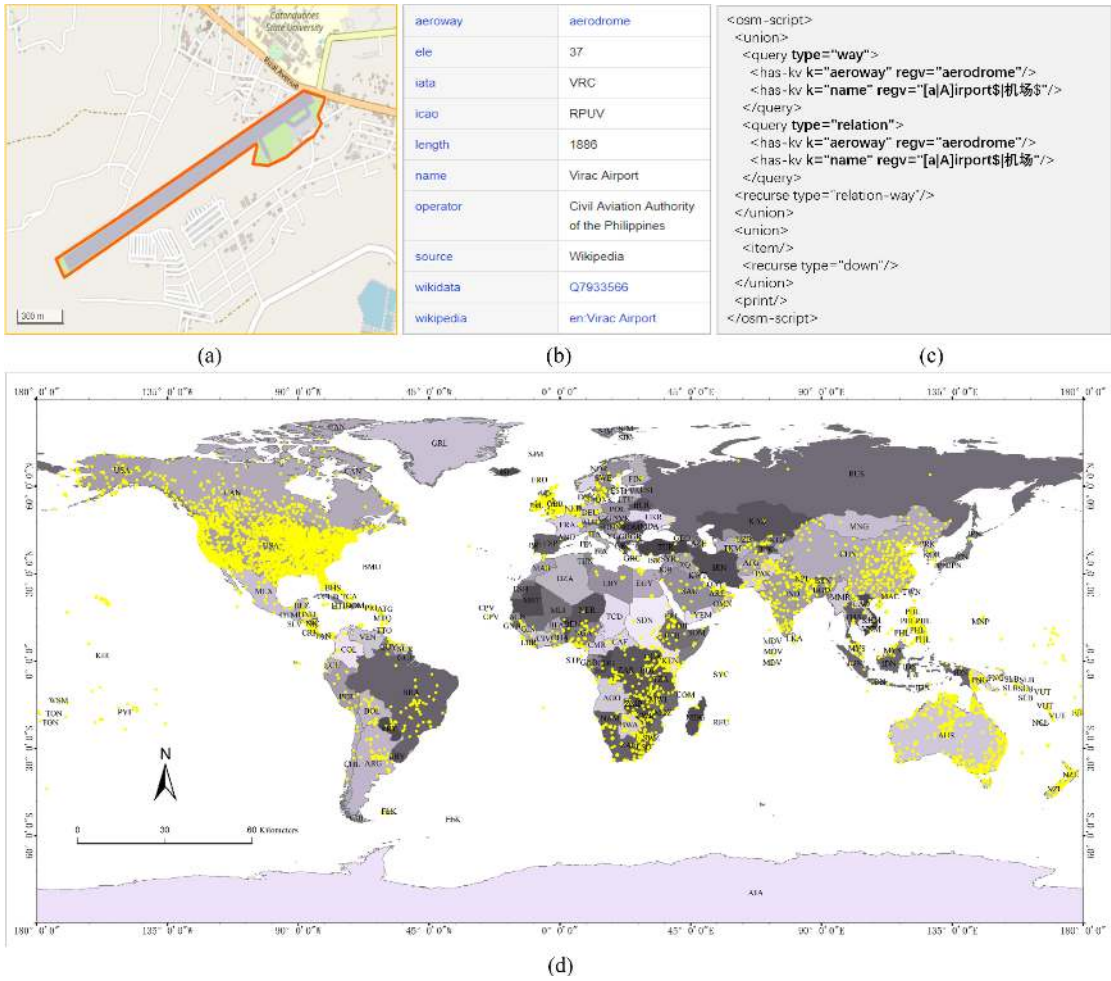


Fig. 8. Illustration of searching scenes of airports around the world. An airport in OSM contains a large amount of tags, which can be employed to search airports with specific semantic key-value labels, e.g., *aeroway* and *name*. More than 5000 world airports in forms like way and relation can be obtained with accurate geographic coordinates, using English and Chinese semantic tags.

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 box is consistent with its corresponding 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 and 512 × 512, according to the scene scales and resolutions of Google Earth images. There may be inaccurate semantic label assignments caused by noisy coordinates and wrong scene boxes. To ensure the correctness of the category labels, all of the scene images are checked by specialists in the field of RS image interpretation. Specifically, those downloaded images in a specific category are deleted if they are assigned with wrong scene labels. 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. With these operations, we are able to improve the accuracy of the scene images that are automatically annotated and, therefore, guarantee the quality of the constructed datasets for scene classification.

D. Discussion

By following the presented guidances, the above procedure formulates a framework for the collection of RS scene images. As shown in Table II, Million-AID consists of the most scene categories compared to the existing scene classification datasets except for fMoW [74]. Different from the existing datasets of which categories are organized with parallel or uncertain relationships, scene categories in Million-AID are organized with systematic relationship architecture, giving it superiority in management and scalability. More importantly, the scene categories are customized to match the land-use classification standards. All of these have greatly enhanced the practicability of the constructed dataset.

The property of diversity is important for a dataset to train interpretation algorithms of strong generalization ability. In the construction of Million-AID, the diversity of scenes in each

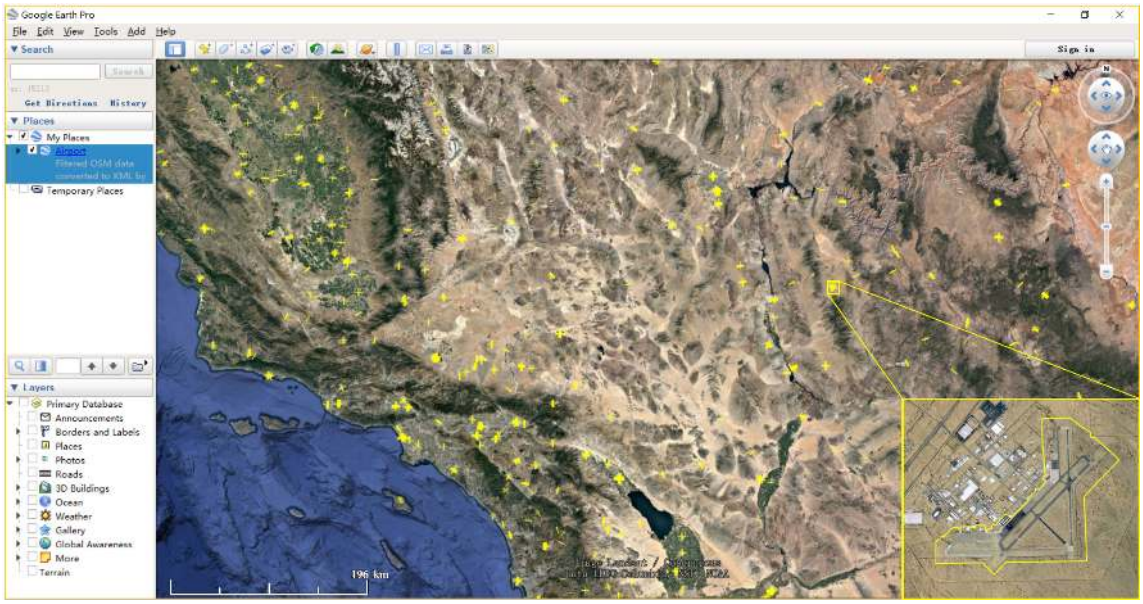


Fig. 9. Airport planes within a local area of the United States are extracted from OSM and shown in Google Earth Pro (©*2020 Google LLC.*), where the upper-left and bottom-right coordinates of geographic range are (37.2262° N, 115.8819° W) and (32.6005° N, 110.6497° W), respectively. The zoomed image shows the accurate extent of an airport in a closed OSM way feature.

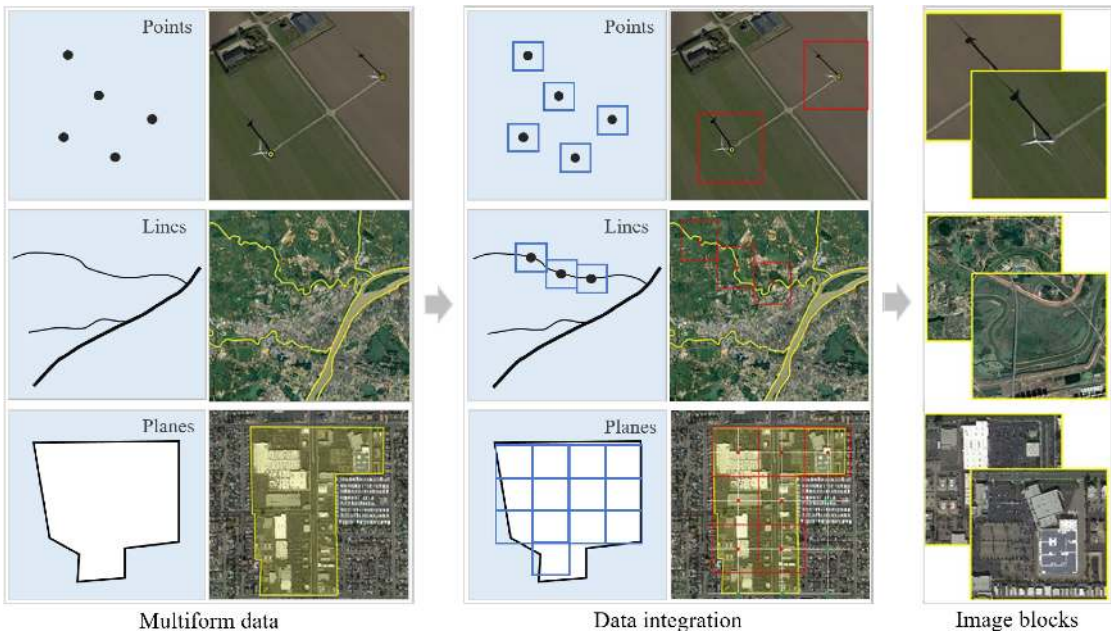


Fig. 10. Illustration of the acquisition of RS scene images based on the collected geographic point, line, and area data. The points are set as the centers of scene square boxes. For line data, the center points are sampled by intervals. For plane data, scene square boxes are generated by mesh grids.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

category is greatly enhanced by the wide geographic distribution of scene locations. Specifically, images in each scene category are extracted from different areas around the world. When some acquired scene coordinates are intensively located in a local area, we try to collect more scene coordinates from other different areas to increase the scene diversity in spatial distribution. The advantage of this strategy is obvious as the wide distribution of scene coordinates makes our collected scenes of interest be individually and spatially independent. Each scene object is able to reflect the unique characteristic from different perspectives. Thus, the within-class diversity of scene images can be greatly improved. In addition, the large-scale semantic categories also improve the between-class diversity of scene images. Some scenes belonging to different categories may share similar characteristics, e.g., stadium and ground track field. Especially, the fine-grained scene categories also possess the same semantic information as they all belong to the same parent scene class. In general, the independent scene objects within the same category

and the unique characteristics among different scene categories make the constructed dataset characterized with big diversity.

As introduced before, the collected images in Million-AID are mainly from Google Earth. It is well known that images in Google Earth are from different satellites, including but not limited to the GeoEye, WorldView, QuickBird, IKONOS, SPOT, and Landsat serial satellites. The multiple data sources can naturally improve the richness of the scene images. Besides, scene images in Million-AID are with broad resolutions, ranging from 0.5 to 153 m per pixel. Note that there are also images of different resolutions in the same scene category, in which scene images are acquired according to the scene scale. At the same time, in the process of scene image inspection, we pay more attention to choose scene images under different imaging conditions, e.g., viewpoint and illumination, to increase the richness of scene images. In the scene image acquisition stage, those scene images with regional overlap are eliminated and only one of them is retained. Thus, the collected content information of each scene image is not repeated and different scene images have different scene background information. In this way, the variety of dataset is greatly guaranteed by the large-scale and individual scene images. These characteristics allow us to greatly enhance the richness of the dataset at the image level.

The construction process of Million-AID also largely follows the idea of semiautomatic annotation. At the stage of scene location acquisition, map search engines, open source data, and public geographic information database are used to obtain the point, line, and plane features that indicate the scene objects. The scene labels are then acquired with the corresponding semantic tags of coordinates. In practice, each kind of scene objects can be acquired by combining several scene coordinate collection methods as introduced. Compared with the manual search method, our method can greatly reduce the difficulty of scene information acquisition and improve the efficiency of dataset construction. Thus, it is easy to collect large-scale scene images. Consequently, the number of images in each scene category goes beyond 2000 and reaches over 20 000 in average. All of these provide guarantees of diversity and richness for the constructed dataset. Not only that, owing to our strategy that automatically obtains scene coordinates and semantic annotations, only image check and deletion are performed manually, which is a really easy work. Therefore, the interactive annotation strategy makes the dataset construction fall into a semiautomatic annotation mode which can greatly reduce the manpower cost and ensure the label quality simultaneously. Thus, it is feasible to build the large-scale scene image dataset with high quality. Consequently, following the introduced guidances and annotation methodology, the Million-AID dataset is achieved with more than 1 000 000 annotated semantic images of 51 scene categories.

V. Challenges and Perspectives

Driven by the wide applications 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 are still a giant gap between the requirement of large-scale dataset and interpretation algorithm development, especially for 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. After investigating the current annotation strategies for RS image datasets, this section discusses the current challenges and potential perspectives for efficient dataset construction.

A. Visualization Technology for RS Image Annotation

In the process of RS image annotation, semantic content in the image is first recognized by visual interpretation of experts. Then, the semantic labels are assigned to the corresponding objects in pixel, region, or image levels. 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: A hyper-spectral image usually contains hundreds of spectral bands, which can provide rich spatial–spectral information of features of the Earth's surface. However, the high dimensionality of 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 that consists of hundreds of spectral bands using conventional display strategies. In order to alleviate this problem, the strategy of band selection can be explored to choose three representative bands of the original image as RGB channels [197]. The fundamental idea of this strategy is to select bands with as much information as possible from the original hyper-spectral image or directly refer the characteristics of the annotation objects. Alternatively, band transformation can also be considered by making the best use of the rich bands. 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. Then, the three representative channels can be selected for visualization [198]. These strategies should rely on effective algorithms developed for band selection and transformation. Besides, hyper-spectral RS images collected from different sensors usually suffer from spectral variability, making it difficult for information extraction and content annotation. Thus, effective hyper-spectral unmixing method [199] can be developed to accurately estimate the content to be annotated. 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 endows it with great superiority over optical RS images in some practical applications, e.g., disaster rescue. 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 with gray-scale mode and there is no color information except for full-polarimetric SAR images [200]. 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/smoothness and changes in the back-scattering signal intensity of surface conditions [201]. On the other hand, visualization technology should also be explored to enhance the legibility of SAR image content. One direction is to colorize the nonfull-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 [200]. With these considerations, more efforts should be poured into SAR image visualization to relieve the difficulties of annotation.

Large-size image annotation with high interaction efficiency: Annotation for large-size RS images is another important challenge. Currently, RS images are usually annotated by tools designed for natural image labeling [161], [162], where only images of limited sizes, e.g., image width/height of several hundred pixels, can be fully visualized for interactive annotation. However, with the improvement of image resolution, RS images taken from the bird-view have large geographic coverage and thus possess large sizes, e.g., width/height of tens of thousands pixels. Thus, the conventional annotation solution can only visualize the local region of an RS image for annotation operations. Besides, current machine monitor devices are also with limited sizes and resolution. It requires constant image roaming and zooming operations when annotating large-size RS images, which heavily hinders the interaction efficiency of annotation and loses the possibility of catching the features with spatial continuity from a global perspective of the image content. On the other hand, the RS images with spatial information need large storage space because of its large amount of data. Thus, the visualization of RS images also requires large-scale computing capability when conducting annotating operations. Considering these points, the visualization technology for displaying, roaming, zooming, and annotating large-size RS images needs to be stressed for efficient annotation.

B. 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 also play an important role in efficiency improvement. Besides, noise data is a common problem in RS image annotation, which makes the handling of noise annotation a valuable topic for dataset quality control as well as the 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 in the application department is produced, they naturally carry semantic information which can be utilized as the source of RS image annotations. For example, thematic map as the typical application product is able to be involved in creating training dataset and generating new annotation product [186], [187], [208]. Usually, the map data of land survey from the land-use institution is obtained through field investigation and thus possesses accurate land classification information, which can be easily combined with RS images to create reliable annotated datasets for model adaption. This scheme is reasonable because the product data from the application department is oriented to the real application scenarios. At this point, the created dataset for RS image interpretation can most truly reflect the key challenges in the real application scenarios. Thus, the interpretation algorithms built upon this kind of dataset would be more practical. Besides, the product data will change with the alternation of application department's business. Thus, the product data can be employed to update the created dataset promptly. In this way, it ensures the established dataset to be always oriented to real applications and, therefore, promote the design and training of practical interpretation algorithms. In general, the efficiency and practicality of the dataset for RS image interpretation can be greatly improved by cooperating with application departments.

Tools for RS image annotation: Another point worth noting is the necessity of developing professional and open-sourced tools for RS image annotation. A number of popular tools for image annotation have been published, as listed in Table VI. These include excellent tools for specific image content interpretation tasks, e.g., object recognition [162], [169], [203]. Some annotation tools strive to provide diverse annotation modalities, such as polygon, rectangle, circle, ellipse, line, and point [161], [204]–[207], serving as universal annotation platforms that are 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 VI 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 also algorithms tolerant to noise annotations are the common requirements in real application scenarios. In the construction of an RS image dataset, images can be annotated by multiple experts while different annotators possess varying levels of expertise. Besides, the opinions of annotators may conflict with each other because of the personal bias [209]. Not only that, but RS images with high complexity is hard to be correctly interpreted even for the experts due to the high demand for specialized background and knowledge of RS image interpretation. 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, cleansing annotations by the manual way usually results in

TABLE VI
ANNOTATION TOOLS FOR IMAGE DATASET CONSTRUCTION

No.	Name	Ref.	Year	Description
1	LabelMe	[161]	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)	[202]	2012	An online tool that efficiently scaling up video annotation with crowdsourced marketplaces (*e.g.*, AMT).
3	LabelImg	[162]	2015	A popular graphical image annotation application that labels objects in images with bounding boxes.
4	Visual Object Tagging Tool (VOTT)	[203]	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)	[204]	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	[205]	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++	[169]	2018	A deep learning-based annotation strategy, producing polygonal annotation of objects segmentation interactively using humans-in-the-loop.
8	Makesence.AI	[206]	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)	[207]	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 nonexhaustively presents the popular and representative image annotation tools.

high costs of time and labor. Thus, how to quickly find out the possible noise annotations in 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. Thus, the effect of noise annotation on the performance of interpretation algorithms is necessary to be explored for better utilization of the annotated dataset [210], [211]. Furthermore, it is crucial to consider the existence of annotation noise and develop noise-robust algorithms [212], [213] to efficiently fade away its negative effects on RS image interpretation.

VI. CONCLUSION

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

Through a bibliometric analysis, this article first presents a systematic review of the existing datasets related to the mainstream of RS image interpretation tasks. It reveals that many of 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 guidances for building the useful dataset for RS image interpretation. 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 presented guidances formulate a prototype for RS image dataset construction with consideration in efficiency and quality assurance. With the introduced guidances, we created a large-scale RS image dataset for scene classification, i.e., Million-AID, through a semiautomatic annotation strategy. It provides a new idea and approach for the construction of RS image datasets. And the discussion about challenges and perspectives in RS image dataset annotation delivers a new sight for the future work where efforts need to be dedicated for 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] C. Toth and G. Jóźków, "Remote sensing platforms and sensors: A survey," *ISPRS J. Photogrammetry Remote Sens.*, vol. 115, pp. 22–36, 2016.

[2] 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, Sep. 2019.

[3] L. Zheng *et al.*, "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, 2020, Art. no. 111522.

[5] X.-Y. Tong *et al.*, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, 2020, Art. no. 111322.

[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, 2020, Art. no. 111458.

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

[11] 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, Oct. 2017.

[12] 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.

[13] 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, Dec. 2018.

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

[15] G. Sheng, W. Yang, T. Xu, and H. Sun, "High-resolution satellite scene classification using a sparse coding based multiple feature combination," *Int. J. Remote Sens.*, vol. 33, no. 8, pp. 2395–2412, 2012.

[16] 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, Aug. 2015.

[17] 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, May 2018.

[18] 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*, vol. 6, no. 3, pp. 507–521, Sep. 2020.

[19] 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, Sep. 2019.

[20] 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, Jul. 2019.

[21] J. Liang, Y. Deng, and D. Zeng, "A deep neural network combined CNN and GCN for remote sensing scene classification," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 4325–4338, Jul. 2020.

[22] 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, Mar. 2020.

[23] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 3735–3756, Jun. 2020.

[24] J. Porway, Q. Wang, and S. C. Zhu, "A hierarchical and contextual model for aerial image parsing," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 254–283, 2010.

[25] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.

[26] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 117, pp. 11–28, 2016.

[27] F. Zhang, B. Du, L. Zhang, and M. Xu, "Weakly supervised learning based on coupled convolutional neural networks for aircraft detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 9, pp. 5553–5563, Sep. 2016.

[28] G.-S. Xia *et al.*, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.

[29] 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. Comput. Vis. Pattern Recognit.*, 2019, pp. 2849–2858.

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

[31] 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, Jul. 2019.

[32] 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.

[33] 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.

[34] Y. Xu *et al.*, "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1452–1459, Apr. 2021.

[35] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery, "Active learning methods for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 2218–2232, Jul. 2009.

[36] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari, "A survey of active learning algorithms for supervised remote sensing image classification," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 5, no. 3, pp. 606–617, Jun. 2011.

[37] A. Romero, C. Gatta, and G. Camps-Valls, "Unsupervised deep feature extraction for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1349–1362, Mar. 2016.

[38] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 645–657, Feb. 2017.

[39] 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, Nov. 2017.

[40] 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.

[41] 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, Jun. 2018.

[42] 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, Jun. 2019.

[43] 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.

[44] 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, Nov. 2019.

[45] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.3015157.

[46] C. Zhang *et al.*, "Deep feature aggregation network for hyperspectral remote sensing image classification," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 5314–5325, Sep. 2020.

[47] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[48] 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, Mar. 2017.

[49] M. Reichstein *et al.*, "Deep learning and process understanding for data-driven earth system science," *Nature*, vol. 566, no. 7743, pp. 195–204, 2019.

[50] D. Hong *et al.*, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.3016820.

[51] M. Chi, A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, and Y. Zhu, "Big data for remote sensing: Challenges and opportunities," *Proc. IEEE*, vol. 104, no. 11, pp. 2207–2219, Nov. 2016.

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

[53] 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.

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

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

[56] 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. 14 680–14707, 2015.

[57] 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.

[58] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

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

[60] T. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[61] C. Chen, "Citespace II: Detecting and visualizing emerging trends and transient patterns in scientific literature," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 57, no. 3, pp. 359–377, 2006.

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

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

[64] 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. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 44–51.

[65] L. Zhao, P. Tang, and L. Huo, "Feature significance-based multibag-of-visual-words model for remote sensing image scene classification," *J. Appl. Remote. Sens.*, vol. 10, no. 3, 2016, Art no. 0 35004.

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

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

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

[69] Planet, "Planet: Understanding the amazon from apace," 2017. Accessed: Dec. 16, 2020. [Online]. Available: https://www.kaggle.com/c/planet-understanding-the-amazon-from-space/overview

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

[71] A.-J. Gallego, A. Pertusa, and P. Gil, "Automatic ship detection from optical aerial images with convolutional neural networks," *Remote Sens.*, vol. 10, no. 4, 2018, Art. no. 511.

[72] 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, Jul. 2019.

[73] 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.

[74] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee, "Functional map of the world," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6172–6180.

[75] WiDS, "Women in data science (WiDS) datathon 2019," 2019. Accessed: Dec. 16, 2020. [Online]. Available: https://www.kaggle.com/c/widsdatathon2019/data

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

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

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

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

[80] G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 30–43.

[81] F. Tanner *et al.*, "Overhead imagery research data set: An annotated data library & tools to aid in the development of computer vision algorithms," in *Proc. Appl. Imagery Pattern Recognit. Workshops*, 2009, pp. 1–8.

[82] 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, Jan. 2012.

[83] 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.

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

[85] 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.

[86] 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.

[87] 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. Eur. Conf. Comput. Vis.*, 2016, pp. 785–800.

[88] 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, Aug. 2016.

[89] 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, May 2017.

[90] 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. Comput. Vis.*, 2017, pp. 4145–4153.

[91] J. Li, C. Qu, and J. Shao, "Ship detection in SAR images based on an improved faster R-CNN," in *Proc. SAR Big Data Era: Mod, Meth., Appl.*, 2017, pp. 1–6.

[92] SpaceNet, "Datasets: The spacenet catalog," Oct. 2018. Accessed: Jan. 15, 2021. [Online]. Available: https://spacenet.ai/datasets/

[93] 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, Mar. 2018.

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

[95] D. Lam *et al.*, "xview: Objects in context in overhead imagery," 2018, *arXiv:1802.07856*.

[96] M. Y. Yang, W. Liao, X. Li, and B. Rosenhahn, "Deep learning for vehicle detection in aerial images," in *Proc. IEEE Int. Conf. Image Process.*, 2018, pp. 3079–3083.

[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, Jan. 2019.

[98] I. Demir *et al.*, "Deepglobe 2018: A challenge to parse the earth through satellite images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 172–17209.

[99] L. Huang *et al.*, "OpenSARShip: A dataset dedicated to sentinel-1 ship interpretation," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 11, no. 1, pp. 195–208, Jan. 2018.

[100] CrowdAI, "Crowdai mapping challenge," 2018. Accessed: Feb. 26, 2021. [Online]. Available: https://www.crowdai.org/challenges/mapping-challenge

[101] Airbus, "Airbus ship detection challenge," 2018. Accessed: Feb. 26, 2021. [Online]. Available: https://www.kaggle.com/c/airbus-ship-detection/overview

[102] S. Waqas Zamir *et al.*, "iSAID: A large-scale dataset for instance segmentation in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 28–37.

[103] 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, Aug. 2019.

[104] 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.

[105] J. Ding *et al.*, "Object detection in aerial images: A large-scale benchmark and challenges," 2021, *arXiv:2102.12219*.

[106] Y. Wang, C. Wang, H. Zhang, Y. Dong, and S. Wei, "A SAR dataset of ship detection for deep learning under complex backgrounds," *Remote Sens.*, vol. 11, no. 7, 2019, Art. no. 765.

[107] X. Sun, Z. Wang, Y. Sun, W. Diao, Y. Zhang, and K. Fu, "AIR-SARShip-1.0: High-resolution SAR ship detection dataset," *J. Radars*, vol. 8, no. R19097, pp. 852–862, 2019.

[108] S. Wei, X. Zeng, Q. Qu, M. Wang, H. Su, and J. Shi, "HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation," *IEEE Access*, vol. 8, pp. 120 234–120254, 2020.

[109] J. Shermeyer *et al.*, "Rareplanes: Synthetic data takes flight," 2020, *arXiv:2006.02963*.

[110] R. Gupta *et al.*, "Creating xBD: A dataset for assessing building damage from satellite imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2019, pp. 10–17.

[111] S. Alemohammad, A. Ballantyne, G. Y. Bromberg, K. Booth, L. Nakanuku-Diggs, and A. Miglarese, "Landcovernet: A global land cover classification training dataset," Accessed: Jan. 10, 2021. [Online]. Available: http://registry.mlhub.earth/10.34911/rdnt.d2ce8i

[112] A. Boguszewski, D. Batorski, N. Ziemba-Jankowska, A. Zambrzycka, and T. Dziedzic, "Landcover. AI: Dataset for automatic mapping of buildings, woodlands and water from aerial imagery," 2020, *arXiv:2005.02264*.

[113] M. T. Chiu *et al.*, "Agriculture-vision: A large aerial image database for agricultural pattern analysis," 2020, *arXiv:2001.01306*.

[114] D. S. T. Laboratory, "DSTL satellite imagery feature detection," 2017. Accessed: Dec. 16, 2020. [Online]. Available: https://www.kaggle.com/c/dstl-satellite-imagery-feature-detection/overview

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

[116] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, Univ. Toronto, Toronto, Canada, 2013.

[117] 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 Int. Geosci. Remote Sens. Symp.*, 2017, pp. 3226–3229.

[118] A. Shakeel, W. Sultani, and M. Ali, "Deep built-structure counting in satellite imagery using attention based re-weighting," *ISPRS J. Photogrammetry Remote Sens.*, vol. 151, pp. 313–321, 2019.

[119] P. Kaiser, J. D. Wegner, A. Lucchi, M. Jaggi, T. Hofmann, and K. Schindler, "Learning aerial image segmentation from online maps," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6054–6068, Nov. 2017.

[120] M. J. Hughes and D. J. Hayes, "Automated detection of cloud and cloud shadow in single-date Landsat imagery using neural networks and spatial post-processing," *Remote Sens.*, vol. 6, no. 6, pp. 4907–4926, 2014.

[121] S. Mohajerani and P. Saeedi, "Cloud-net: A cloud segmentation CNN for Landsat 8 remote sensing imagery optimized with filtered Jaccard loss function," 2020, *arXiv:2001.08768*.

[122] S. Foga *et al.*, "Cloud detection algorithm comparison and validation for operational landsat data products," *Remote Sens. Environ.*, vol. 194, pp. 379–390, 2017.

[123] L. Baetens, C. Desjardins, and O. Hagolle, "Validation of copernicus Sentinel-2 cloud masks obtained from MAJA, Sen2Cor, and FMask processors using reference cloud masks generated with a supervised active learning procedure," *Remote Sens.*, vol. 11, no. 4, 2019, Art. no. 433.

[124] A. Francis, J. Mrziglod, P. Sidiropoulos, and J.-P. Muller, "Sentinel-2 cloud mask catalogue," 2020. Accessed: Dec. 16, 2020. [Online]. Available: https://zenodo.org/record/4172871#.YAPb2dAzZPY

[125] 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, Aug. 2009.

[126] B. A. M. Graña and M. A. Veganzons, "Pavia dataset," 2020. Accessed: Jan. 15, 2021. [Online]. Available: http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes

[127] ISPRS-Contest, "ISPRS 2D semantic labeling contest," Jun. 2018. Accessed: Jan. 15, 2021. [Online]. Available: http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html

[128] M. Volpi and V. Ferrari, "Semantic segmentation of urban scenes by learning local class interactions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2015, pp. 1–9.

[129] M. Leenstra, D. Marcos, F. Bovolo, and D. Tuia, "Self-supervised pre-training enhances change detection in Sentinel-2 imagery," 2021, *arXiv:2101.08122*.

[130] 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," 2019, *arXiv:1906.07789*.

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

[132] S. M. Azimi, C. Henry, L. Sommer, A. Schumann, and E. Vig, "Skyscapes fine-grained semantic understanding of aerial scenes," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 7393–7403.

[133] 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, Mar. 2005.

[134] M. F. Baumgardner, L. L. Biehl, and D. A. Landgrebe, "220 band aviris hyperspectral image data set: Jun. 12, 1992 Indian pine test site 3," Sep. 2015. Accessed: Jan. 10, 2021. [Online]. Available: https://purr.purdue.edu/publications/1947/1

[135] 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, 2017, Art. no. 500.

[136] 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.

[137] I. Nigam, C. Huang, and D. Ramanan, "Ensemble knowledge transfer for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 1499–1508.

[138] Z. Shao, K. Yang, and W. Zhou, "Performance evaluation of single-label and multi-label remote sensing image retrieval using a dense labeling dataset," *Remote Sens.*, vol. 10, no. 6, 2018, Art. no. 964.

[139] X. X. Zhu *et al.*, "So2Sat LCZ42: A benchmark dataset for global local climate zones classification," *IEEE Geosci. Remote Sens. Mag.*, 2020, *arXiv:1912.12171*.

[140] DroneDeploy, "Dronedeploy machine learning segmentation benchmark," 2019. Accessed: Dec. 16, 2020. [Online]. Available: https://github.com/dronedeploy/dd-ml-segmentation-benchmark

[141] Sentinelhub, "Example dataset of eopatches for Slovenia 2019," 2019. Accessed: Dec. 16, 2020. [Online]. Available: http://eo-learn.sentinel-hub.com/?prefix=

[142] 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.*, vol. 165, pp. 108–119, 2020.

[143] 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, Oct. 2009.

[144] N. Bourdis, D. Marraud, and H. Sahbi, "Constrained optical flow for aerial image change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2011, pp. 4176–4179.

[145] 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, May 2014.

[146] M. Volpi, G. Camps-Valls, and D. Tuia, "Spectral alignment of multitemporal cross-sensor images with automated kernel canonical correlation analysis," *ISPRS J. Photogrammetry Remote Sens.*, vol. 107, pp. 50–63, 2015.

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

[148] 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, 2018, Art. no. 1827.

[149] Q. Wang, Z. Yuan, Q. Du, and X. Li, "GETNET: A general end-to-end 2-D CNN framework for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 3–13, Jan. 2019.

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

[151] J. López-Fandiño, A. S. Garea, D. B. Heras, and F. Argüello, "Stacked autoencoders for multiclass change detection in hyperspectral images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 1906–1909.

[152] 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 Int. Geosci. Remote Sens. Symp.*, 2018, pp. 2115–2118.

[153] 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. Photogrammetry Remote Sens. Spatial Inf. Sci.*, vol. 42, no. 2, pp. 565–571, 2018.

[154] 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. Mach. Vis. Appl.*, 2017, pp. 5–8.

[155] L. T. Luppino, F. M. Bianchi, G. Moser, and S. N. Anfinsen, "Unsupervised image regression for heterogeneous change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9960–9975, Aug. 2019.

[156] J. López-Fandiño, D. B. Heras, F. Argüello, and M. Dalla Mura, "GPU framework for change detection in multitemporal hyperspectral images," *Int. J. Parallel Prog.*, vol. 47, no. 2, pp. 272–292, 2019.

[157] C. D. Rodrigo, L. S. Bertrand, B. Alexandre, and G. Yann, "Multitask learning for large-scale semantic change detection," *Comput. Vis. Image Underst.*, vol. 187, 2019, Art. no. 102783.

[158] 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, 2020, Art. no. 1662.

[159] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, H. Ding, and X. Huang, "SemiCDNet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, pp. 1–16, Aug. 2020.

[160] M. Zhang and W. Shi, "A feature difference convolutional neural network-based change detection method," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7232–7246, Oct. 2020.

[161] 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. Vis.*, vol. 77, no. 1–3, pp. 157–173, 2008.

[162] Tzutalin, "Labelimg," 2015. Accessed: Jan. 15, 2021. [Online]. Available: https://github.com/tzutalin/labelImg

[163] Q. Hu *et al.*, "Exploring the use of Google Earth imagery and object-based methods in land use/cover mapping," *Remote Sens.*, vol. 5, no. 11, pp. 6026–6042, 2013.

[164] 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, Jan. 2016.

[165] 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.

[166] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.

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

[168] N. Yang and H. Tang, "GeoBoost: An incremental deep learning approach toward global mapping of buildings from VHR remote sensing images," *Remote Sens.*, vol. 12, no. 11, 2020, Art. no. 1794.

[169] D. Acuna, H. Ling, A. Kar, and S. Fidler, "Efficient interactive annotation of segmentation datasets with Polygon-RNN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 859–868.

[170] 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.

[171] Y. Yao *et al.*, "Towards automatic construction of diverse, high-quality image datasets," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 6, pp. 1199–1211, Jun. 2020.

[172] P. Zhu *et al.*, "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, Jun. 2020.

[173] 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.

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

[175] 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.

[176] 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, 2019, Art. no. 145.

[177] 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.

[178] 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, Jul. 2013.

[179] 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, Sep. 2012.

[180] 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, Jun. 2016.

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

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

[183] 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.

[184] S. Dang, Z. Cao, Z. Cui, Y. Pi, and N. Liu, "Open set incremental learning for automatic target recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4445–4456, Jul. 2019.

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

[186] C. Paris, L. Bruzzone, and D. Fernández-Prieto, "A novel approach to the unsupervised update of land-cover maps by classification of time series of multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4259–4277, Jul. 2019.

[187] C. Paris and L. Bruzzone, "A novel approach to the unsupervised extraction of reliable training samples from thematic products," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 1930–1948, Mar. 2021.

[188] E. Agustsson, J. R. Uijlings, and V. Ferrari, "Interactive full image segmentation by considering all regions jointly," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11 622–11631.

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

[190] B. Demir, L. Minello, and L. Bruzzone, "Definition of effective training sets for supervised classification of remote sensing images by a novel cost-sensitive active learning method," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 2, pp. 1272–1284, Feb. 2014.
[191] 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. 15 014–15045, 2015.
[192] D.-J. Chen, J.-T. Chien, H.-T. Chen, and L.-W. Chang, "Tap and shoot segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 2119–2126.
[193] D. P. Papadopoulos, J. R. Uijlings, F. Keller, and V. Ferrari, "Extreme clicking for efficient object annotation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4930–4939.
[194] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3159–3167.
[195] 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.
[196] S. Vijayanarasimhan and K. Grauman, "Large-scale live active learning: Training object detectors with crawled data and crowds," *Int. J. Comput. Vis.*, vol. 108, no. 1-2, pp. 97–114, 2014.
[197] 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.
[198] 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, Apr. 2020.
[199] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.
[200] 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.*, vol. 18, no. 2, pp. 296–300, Feb. 2021.
[201] 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, Jul. 2012.
[202] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently scaling up crowdsourced video annotation," *Int. J. Comput. Vis.*, vol. 101, no. 1, pp. 184–204, 2013.
[203] Microsoft, "Visual object tagging tool," 2018. Accessed: Jan. 15, 2021. [Online]. Available: https://github.com/microsoft/VoTT
[204] Intel, "Computer vision annotation tool: A universal approach to data annotation," 2019. Accessed: Jan. 15, 2021. [Online]. Available: https://github.com/opencv/cvat
[205] N. Fiedler, M. Bestmann, and N. Hendrich, "Imagetagger: An open source online platform for collaborative image labeling," *RoboCup 2018: Robot World Cup XXII*. Berlin, Germany: Springer, 2018.
[206] P. Skalski, "Make sense," 2019. Accessed: Jan. 15, 2021. [Online]. Available: https://github.com/SkalskiP/make-sense
[207] A. Dutta and A. Zisserman, "The VIA annotation software for images, audio and video," in *Proc. ACM Int. Conf. Multimedia*, 2019, pp. 2276–2279.
[208] J. Inglada, A. Vincent, M. Arias, B. Tardy, D. Morin, and I. Rodes, "Operational high resolution land cover map production at the country scale using satellite image time series," *Remote Sens.*, vol. 9, no. 1, 2017, Art. no. 95.
[209] A. Khetan, Z. C. Lipton, and A. Anandkumar, "Learning from noisy singly-labeled data," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–15.
[210] C. Pelletier, S. Valero, J. Inglada, N. Champion, C. Marais Sicre, and G. Dedieu, "Effect of training class label noise on classification performances for land cover mapping with satellite image time series," *Remote Sens.*, vol. 9, no. 2, 2017, Art. no. 173.
[211] B. Swan, M. Laverdiere, and H. L. Yang, "How good is good enough? Quantifying the effects of training set quality," in *Proc. 2nd ACM SIGSPATIAL Int. Workshop AI Geogr. Knowl. Discov.*, 2018, pp. 47–51.
[212] B. Frénay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 5, pp. 845–869, May 2014.
[213] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Learning to learn from noisy labeled data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5051–5059.

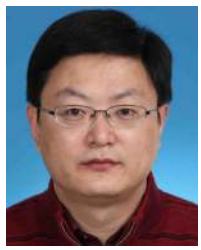
Yang Long received the B.S. degree in resources environment and the management of urban and rural planning from Hubei Normal University, Huangshi, China, in 2014 and the M.S. degree in cartography and geography information system from Wuhan University, Wuhan, China, in 2018. He is currently working toward the Ph.D. degree in photogrammetry and remote sensing at the State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing, Wuhan University, Wuhan, China.

His research interests include scene classification and object detection in remote sensing images.

Gui-Song Xia (Senior Member, IEEE) received the Ph.D. degree in image processing and computer vision from CNRS LTCI, Télécom ParisTech, Paris, France, in 2011.

From 2011 to 2012, he was a Postdoctoral Researcher with the Centre de Recherche en Mathématiques de la Decision, CNRS, Paris-Dauphine University, Paris, France, for one and a half years. He is currently working as a Full Professor of Computer Vision and Photogrammetry with Wuhan University, Wuhan, USA. He has also been working as a Visiting Scholar with DMA, École Normale Supérieure (ENS-Paris), Paris, France, for two months in 2018. He is also a Guest Professor with the Future Lab AI4EO, Technical University of Munich (TUM), Munich, Germany. His current research interests include mathematical modeling of images and videos, structure from motion, perceptual grouping, and remote sensing image understanding.

Dr. Xia serves on the Editorial Boards of the journals *Pattern Recognition*, *Signal Processing: Image Communications*, *EURASIP Journal on Image & Video Processing*, *Journal of Remote Sensing*, and *Frontiers in Computer Science: Computer Vision*.



Shengyang Li received the B.Eng. degree in computer science and technology from the Shandong University of Science and Technology, Qingdao, China, in 2003, and the M.Eng. degree in remote sensing image processing and analysis from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China, in 2006.

He is currently a Professor with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences. His research interests include computer vision, target detection and tracking in video satellite, and remote sensing image analysis and understanding.

Wen Yang (Senior Member, IEEE) received the B.S. degree in electronic apparatus and surveying technology, the M.S. degree in computer application technology, and the Ph.D. degree in communication and information system from Wuhan University, Wuhan, China, in 1998, 2001, and 2004, respectively.

He worked as a Visiting Scholar with the Apprentissage et Interfaces (AI) Team, Laboratoire Jean Kuntzmann (LJK), Grenoble, France, from 2008 to 2009. From 2010 to 2013, he worked as a Postdoctoral Researcher with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University. Since then, he has been a Full Professor with the School of Electronic Information, Wuhan University. His research interests include object detection and recognition, image retrieval and semantic segmentation, change detection, and multisensor information fusion.

Michael Ying Yang (Senior Member, IEEE) received the Ph.D. degree (*summa cum laude*) in photogrammetry from University of Bonn, Bonn, Germany, in 2011, and the venia legendi in computer science from Leibniz University Hannover, Hanover, Germany, in 2016.

He is currently an Assistant Professor with the Department of Earth Observation Science, ITC - Faculty of Geo-Information Science and Earth Observation, University of Twente, Enschede, The Netherlands, heading a group working on scene understanding. From 2008 to 2012, he worked as a Researcher with the Department of Photogrammetry, University of Bonn. From 2012 to 2015, he was a Postdoctoral Researcher with the Institute for Information Processing, Leibniz University Hannover. From 2015 to 2016, he was a Senior Researcher with Computer Vision Lab Dresden, TU Dresden, Dresden, Germany. His research interests are in the fields of computer vision and photogrammetry with specialization on scene understanding and semantic interpretation from imagery and videos. He authored or coauthored more than 100 articles in international journals and conference proceedings and cosupervised six Ph.D. students.

Dr. Yang serves as an Associate Editor for *Photogrammetric Engineering and Remote Sensing*, Cochair of ISPRS working group II/5 Dynamic Scene Analysis, and is the recipient of the ISPRS President's Honorary Citation (2016) and Best Science Paper Award at BMVC 2016. He co-organized 12 workshops with CVPR/ICCV/ECCV and is a Guest Editor of four journal special issues. He is regularly serving as program committee member of conferences and reviewer for international journals.

Xiao Xiang Zhu (Fellow, IEEE) received the M.Sc., Dr.Ing., and Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She is the Professor in data science in earth observation (former: signal processing in earth observation) with TUM and the Head of the Department "EO Data Science," Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany. Since 2019, she has been a Co-Coordinator of the Munich Data Science Research School. Since 2019, she has been the Head of the Helmholtz Artificial Intelligence-Research Field "Aeronautics, Space and Transport." Since May 2020, she has been the Director of the International Future AI Laboratory "AI4EO-Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond," Munich, Germany. Since October 2020, she has also been serving on the Board of Directors of the Munich Data Science Institute (MDSI), TUM. She was a Guest Scientist or a Visiting Professor with the Italian National Research Council (CNR-IREA), Naples, Italy; Fudan University, Shanghai, China; The University of Tokyo, Tokyo, Japan; and the University of California at Los Angeles, Los Angeles, CA, USA, in 2009, 2014, 2015, and 2016, respectively. Her main research interests are remote sensing and Earth observation, signal processing, machine learning, and data science, with a special application focus on global urban mapping.

Dr. Zhu is a member of Young Academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She is an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Liangpei Zhang (Fellow, IEEE) received the B.S. degree in physics from Hunan Normal University, Changsha, China, in 1982, the M.S. degree in optics from the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China, in 1988, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1998.

He was a Principal Scientist for the China State Key Basic Research Project from 2011 to 2016 appointed by the Ministry of National Science and Technology of China to lead the remote sensing program in China. He is a Chair Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing (LIESMARS), Wuhan University. He has authored more than 700 research articles and five books. He is the Institute for Scientific Information (ISI) highly cited author. He holds 30 patents. His research interests include hyperspectral remote sensing, high-resolution remote sensing, image processing, and artificial intelligence.

Dr. Zhang is a Fellow of the Institution of Engineering and Technology (IET). He was a recipient of the 2010 Best Paper Boeing Award, the 2013 Best Paper ERDAS Award from the American Society of Photogrammetry and Remote Sensing (ASPRS), and 2016 Best Paper Theoretical Innovation Award from the International Society for Optics and Photonics (SPIE). His research teams won the top three prizes of the IEEE GRSS 2014 Data Fusion Contest, and his students have been selected as the winners or finalists of the *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* student paper contest in recent years. He also serves as an Associate Editor or an Editor for more than ten international journals. He is serving as an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENING. He is the Founding Chair of the IEEE Geoscience and Remote Sensing Society (GRSS) Wuhan Chapter.

Deren Li received the Ph.D. degree in photogrammetry from the University of Stuttgart, Stuttgart, Germany, in 1986.

He was the President of the former Wuhan Technical University of Surveying and Mapping, Wuhan, China, from 1996 to 2000. He is currently a Professor and the Chair of the Academic Committee of the State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China. He has authored or coauthored eight books and more than 400 papers. His research interests include photogrammetry and remote sensing, global navigation satellite systems, and geographic information systems (GISs), and their innovation integrations and applications in China.

Dr. Li was elected as an Academician of the Chinese Academy of Sciences in 1991, the Chinese Academy of Engineering, and the Euro-Asia Academy of Sciences in 1995. He was the founding President of the Asia GIS Association from 2003 to 2006. He was the President of the Chinese Society of Geodesy, Photogrammetry and Cartography and the International Society for Photogrammetry and Remote Sensing Commissions III and VI. He is also the Vice President of the China Society of Image and Graphics, a Chief Scientist of the Optics Valley of China, and the Cochair of the Committee on Earth Observation and Satellites and the Integrated Global Observing Strategy Partnership. In the 1980s, his research findings on a posterior variance estimation-based iteration weighted method for bundler location was recognized internationally and named the Deren Li Method. His research on the separability theory of model errors scientifically solved a hundred-year baffling problem in geodetic science and earned him the 1988 Best Paper Award of the German Society for Photogrammetry and Remote Sensing and the Hansa Luftbild Award. He was a recipient of more than ten national- and provincial-level awards and prizes, such as the Sci-tech Progress Award, the National Excellent Textbook Award, and the excellent educational achievement awards.