MultiScene: A Large-Scale Dataset and Benchmark for Multiscene Recognition in Single Aerial Images

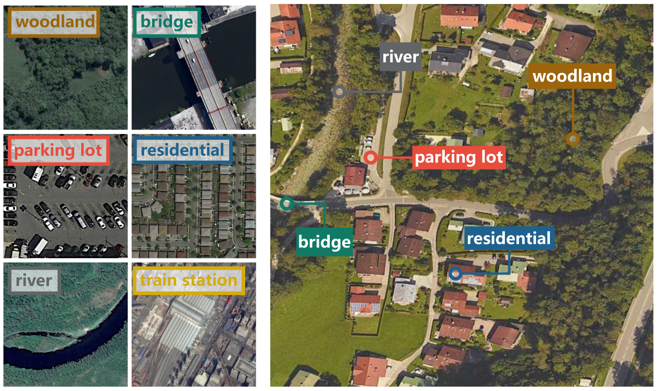
Yuansheng Hua, *Graduate Student Member, IEEE*, Lichao Mou, Pu Jin, *Member, IEEE*, and Xiao Xiang Zhu, *Fellow, IEEE*

***Abstract*— Aerial scene recognition is a fundamental research problem in interpreting high-resolution aerial imagery. Over the past few years, most studies focus on classifying an image into one scene category, while in real-world scenarios, it is more often that a single image contains multiple scenes. Therefore, in this article, we investigate a more practical yet underexplored task—multiscene recognition in single images. To this end, we create a large-scale dataset, called MultiScene, composed of 100 000 unconstrained high-resolution aerial images. Considering that manually labeling such images is extremely arduous, we resort to low-cost annotations from crowdsourcing platforms, e.g., OpenStreetMap (OSM). However, OSM data might suffer from incompleteness and incorrectness, which introduce noise into image labels. To address this issue, we visually inspect 14 000 images and correct their scene labels, yielding a subset of cleanly annotated images, named MultiScene-Clean. With it, we can develop and evaluate deep networks for multiscene recognition using clean data. Moreover, we provide crowdsourced annotations of all images for the purpose of studying network learning with noisy labels. We conduct experiments with extensive baseline models on both MultiScene-Clean and MultiScene to offer benchmarks for multiscene recognition in single images and learning from noisy labels for this task, respectively. To facilitate progress, we make our dataset and trained models available on https://gitlab.lrz.de/ai4eo/reasoning/multiscene.**

***Index Terms*— Convolutional neural network (CNN), crowdsourced annotations, large-scale aerial image dataset, learning from noisy labels, multiscene recognition in single images.**

I. Introduction

With the recent development of Earth observation techniques, massive aerial imagery is now accessible for a variety of applications, such as environmental monitoring [1]–[6], urban planning [7]–[12], land cover and land use mapping [13]–[16], and disaster assessment [17], [18]. As one of the crucial steps toward these applications, aerial scene recognition has been extensively studied in the remote sensing community. During the last few years, the emergence of deep convolutional neural networks (CNNs) pushed ahead research in this field, and enormous achievements [19]–[26] have been obtained. Albeit successful, most existing scene classification studies only focus on a specific scenario, where an aerial image is assumed to include a single scene [27]–[34]. Basically, these studies regard aerial scene recognition as a single-label classification problem and learn models on well-cropped single-scene aerial images [see Fig. 1(a)]. However, in practical applications, an aerial image often contains multiple scenes, as it is collected overhead and usually has a large coverage [cf. Fig. 1(b)]. We also note that even in public single-scene aerial image datasets, the coexistence of multiple scenes in a single image is inevitable, especially in images covering large areas. For example, as shown in the bottom two images in Fig. 1(a), although they are assigned single-scene labels according to their central/dominant scenes (i.e., river and train station), there actually exists more than one scene in each of them.



(a) (b)

Fig. 1. Examples of images utilized in (a) single-scene and (b) multiscene recognition tasks. In (a), each aerial image is assigned one scene label, while in (b), labels of all present scenes are inferred. In comparison with (b), (a) might suffer from partial scene understanding, as only one label is predicted even if there indeed exist multiple scenes in an image. For a clear visualization, locations of scenes are marked in (b).

Hence, in this article, we aim to tackle a more realistic yet challenging problem, namely multiscene recognition in

Manuscript received April 6, 2021; revised July 30, 2021; accepted August 18, 2021. Date of publication September 15, 2021; date of current version January 31, 2022. This work was supported in part by the European Research Council (ERC) through the European Union's Horizon 2020 Research and Innovation Programme (acronym: So2Sat) under Grant ERC-2016-StG-714087, in part by the Helmholtz Association through the Framework of Helmholtz AI under Grant ZT-I-PF-5-01—Local Unit "Munich Unit @Aeronautics, Space and Transport (MASTr)," in part by the Helmholtz Excellent Professorship "Data Science in Earth Observation—Big Data Fusion for Urban Research" under Grant W2-W3-100, 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) under Grant 01DD20001. *(Corresponding authors: Lichao Mou; Xiao Xiang Zhu.)*

Yuansheng Hua, Lichao Mou, and Xiao Xiang Zhu are with the Remote Sensing Technology Institute, German Aerospace Center, 82234 Weßling, Germany, and also with the Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany (e-mail: yuansheng.hua@dlr.de; lichao.mou@dlr.de; xiaoxiang.zhu@dlr.de).

Pu Jin is with the Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany (e-mail: pu.jin@tum.de).

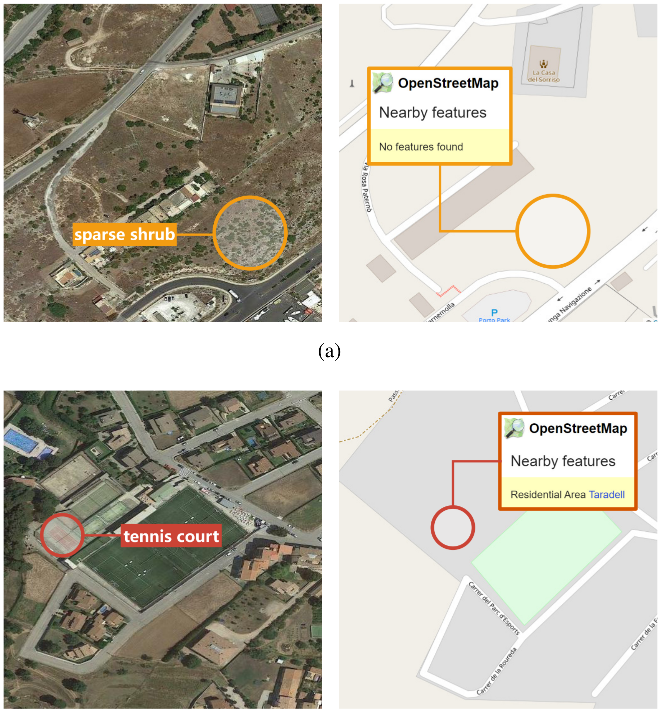
Digital Object Identifier 10.1109/TGRS.2021.3110314

single aerial images. This task refers to assigning an aerial image multiple scene labels, and there are no constraints on image preparations, such as centering dominant scenes and eliminating clutter scenes. Compared to the conventional scene recognition task, multiscene recognition is more arduous because images are large-scale and unconstrained and all present scenes in an aerial image need to be exhaustively recognized. Fig. 1(b) shows an example of multiscene aerial image and corresponding multiple scene-level labels. We can see that not only dominant scenes (e.g., residential and woodland) but also trivial scenes (e.g., bridge and parking lot) are annotated, which draws a more comprehensive picture for the unconstrained image.

However, very few efforts have been deployed to this problem in the remote sensing community. In order to advance the progress of multiscene recognition in single images, we propose a large-scale multiscene recognition (MultiScene) dataset, where 100 000 aerial images are collected around the world. In the phase of data preparation, we note that although massive high-resolution aerial images can be effortlessly obtained from remote sensing data platforms, such as Google Earth,[1] it is extremely time- and labor-consuming to yield their corresponding multiple scene labels. To alleviate such annotation burden, in this article, we resort to crowdsourced data, e.g., OpenStreetMap[2] (OSM) annotations, which has been proven to be successful in generating image-level labels [27], [28], [35] and pixel-wise footprints [12], [36] for training deep networks. However, we observe that OSM data might suffer from two common defects, incompleteness and incorrectness, which could introduce severe noise into image labels. Fig. 2 shows two examples of incorrect OSM annotations, where sparse shrubs are neglected and the tennis court is mislabeled as residential. With this in mind, here, we do not directly use crowdsourced labels as ground truth data. Instead, we visually inspect 14 000 images and correct their labels, producing a subset of cleanly labeled images, named MultiScene-Clean. It allows developing and evaluating deep networks for unconstrained multiscene recognition using clean data. Moreover, we note that the noisy crowdsourced data are not completely useless, for example, they can be used to study network learning with noisy labels for this task. Therefore, we also provide crowdsourced annotations of all images.

The contributions of this article are fourfold.

1) Unlike conventional aerial scene recognition where all images are well-cropped and each of them contains only one scene-level label, in this article, we explore a more practical task—multiscene recognition in single images.
2) We propose a large-scale dataset, namely MultiScene, consisting of 100 000 unconstrained multiscene aerial images, and each is assigned OSM labels. We visually inspect 14 000 images and correct their labels, yielding a subset of cleanly labeled images.
3) The proposed dataset provides not only ground truth data but also crowdsourced labels, which enables studies in learning from enormous noisy labels for our task.
4) We extensively evaluate commonly used classification networks on both MultiScene-Clean and MultiScene and provide benchmarks for recognizing multiple scenes in single images and learning from noisy labels for this task, respectively.



(a)

(b)

Fig. 2. Examples of (a) incomplete and (b) incorrect OSM annotations. In (a), sparse shrubs are not annotated in OSM data, while in (b), the tennis court is mislabeled as residential.

The remaining sections of this article are organized as follows. Section II reviews studies in aerial single-scene classification and multilabel object classification. Section III briefly recalls existing scene datasets and delineates the proposed dataset. Experimental configurations and results are exhibited in Section IV, and finally, Section V draws a conclusion.

II. Related Work

This section briefly reviews related works in two fields: aerial single-scene classification and multilabel object recognition.

A. Aerial Single-Scene Classification

Aerial single-scene classification refers to categorize an aerial image into a single-scene class. Early studies propose to construct scene representations with variant low-level features, e.g., local structures [40], [41], color attributes [42], [43], and texture information [44], [45]. Concerning that low-level features fail to comprehensively depict complex scenes, midlevel algorithms, such as Bag-of-Visual-Words (BoVW) [37], [46] and topic models [47], [48], are devised to encode local features (so-called "visual words") into more holistic midlevel scene representations for the classification task. However,

[1] https://earth.google.com/web/

[2] https://www.openstreetmap.org/

these methods show limited performance in recognizing scenes of high diversity due to their dependence on handcrafted features.

Recently, the emergence of deep CNNs brings immense advancements to the community, and many achievements [19]–[34] have been obtained in the field of aerial single-scene classification. These deep networks have hierarchical architectures, where convolutional and max-pooling layers are periodically interleaved for learning high-level features of intricate scenes. With layers going deeper, the learned features are more abstract and supposed to contain richer semantic information, which is crucial for judicious decisions. A popular trend of deep learning algorithms in single-scene classification is to take a CNN as the backbone and introduce well-designed modules for further enhancing the feature efficiency. For instance, Bi *et al.* [31] proposed to learn multiple instances from feature maps extracted by a densely connected CNN and integrated them into bag-level features for single-scene classification. Li *et al.* [49] proposed a key region capturing method to learn class-specific features and retain global information for inferring scene labels. To leverage features of variant levels, feature aggregation plays a key role in the single-scene classification. Lu *et al.* [50] fused features learned by the last three blocks and the second fully connected layer of VGG-16, and Cao *et al.* [51] designed a nonparametric self-attention layer to enhance spatial and channel responses of fused features for the final prediction. Sun *et al.* [20] developed a gated bidirectional network for aggregating features extracted by different convolutional layers with a gated function in both top-down and bottom-up directions. Besides, exploiting supplementary data, such as geotagged audios and multitemporal images, has been a new research direction. Hu *et al.* [19] proposed to predict scene categories by transferring sound event knowledge learned from sound–image pairs. Ru *et al.* [25] proposed a two-branch network to learn deep features of bitemporal images and fused them through a CorrFusion module for aerial scene classification. Our literature review demonstrates that most of the existing studies assume that an aerial image includes only one scene and focus on well-cropped single-scene aerial images. Hence, these studies tend to regard entities present in an image as compositions of a scene, while in multiscene recognition, this would trigger networks to learn erroneous feature representations. However, very few efforts have been deployed to explore multiscene recognition in the remote sensing community.

B. Multilabel Object Classification

Multilabel object classification refers to assigning an aerial image multiple object-level labels, such as car, tree, and building. Similar to our work, these studies aim to provide a holistic understanding of aerial images, but from the perspective of object. Early attempts [52], [53] follow the idea of simply combining a deep CNN with a postprocessing approach for identifying multiple objects in an aerial image. Zeggada *et al.* [52] fed outputs of a CNN into a customized thresholding operation for inferring multiple object labels, while in [53], a conditional random field (CRF) is utilized as the postprocessing model. In recent literature, more efforts are deployed to endow deep neural networks with the capacity of reasoning about relations among various objects for more accurate predictions. Hua *et al.* [54] proposed an end-to-end network comprising a CNN and a long short-term memory (LSTM) network that is responsible for modeling label dependencies through its recurrent units for multilabel object classification. Sumbul and Demir [55] exploited a bidirectional LSTM network to learn spatial relations among all patches in an image for the final prediction. Hua *et al.* [56] proposed a relational reasoning network module to model label dependencies and gained better classification results. Instead of encoding label relations, Koda *et al.* [57] divided an aerial image into several patches with the same size and modeled spatial relationships among them for multilabel object interpretation. Compared to these studies, our task is more challenging, because compared to object, the concept of scene is more abstract and intricate.

III. Multiscene Dataset for Multiscene Recognition in Single Aerial Images

This section first reviews existing single-scene aerial image datasets and then delineates the proposed dataset.

A. Existing Single-Scene Aerial Image Dataset

During the last decades, various aerial image datasets are published for single-scene classification, and here, we briefly review several commonly used ones.

1) *UC-Merced [37]:* The UC-Merced dataset is composed of 2100 images collected from the United States Geological Survey (USGS) National Map, and each of them is categorized into one of 21 scene classes: overpass, golf course, river, harbor, beach, building, airplane, freeway, intersection, medium residential, runway, agricultural, storage tank, parking lot, forest, sparse residential, chaparral, tennis courts, dense residential, baseball diamond, and mobile home park. The number of images per scene is evenly defined as 100, and only cities in the United States are covered in data acquisition. The size of each image is 256 × 256 pixels, and the spatial resolution is 1 ft. Chaudhuri *et al.* [58] focused on the task of recognizing multiple objects in an image and relabeled the UC-Merced dataset, yielding a multilabel dataset. In this dataset, 2100 images are relabeled, and each is assigned one or several labels from 17 newly defined object classes: airplane, sand, pavement, building, car, chaparral, court, tree, dock, tank, water, grass, mobile home, ship, bare soil, sea, and field.
2) *WHU20 [38]:* The WHU20 dataset is an extended version of the WHU-RS dataset that was originally proposed in [59]. This dataset expands numbers of aerial images and scene classes from 950 to 5000 and from 12 to 20, respectively. For each scene category, more than 200 images with a size of 600 × 600 pixels are collected, and their spatial resolutions range from 0.26 to 7.44 m/pixel.

TABLE I
COMPARISON WITH EXISTING AERIAL SCENE DATASETS FROM VARIOUS PERSPECTIVES

Dataset	# images	spatial resolutions	# scenes	# labels per image	crowdsourced label	Year
UC-Merced [37]	2,100	0.3 m/pixel	21	1	✗	2010
WHU20 [38]	5,000	0.3-7.4 m/pixel	20	1	✗	2015
RSSCN7 [39]	2,800	0.2-1.4 m/pixel	7	1	✗	2015
AID [27]	10,000	0.5-8 m/pixel	30	1	✗	2017
NWPU-RESISC45 [40]	31,500	0.2-30 m/pixel	45	1	✗	2017
MultiScene (Ours)	100,000	0.3-0.6 m/pixel	36	1-13	✓	2021

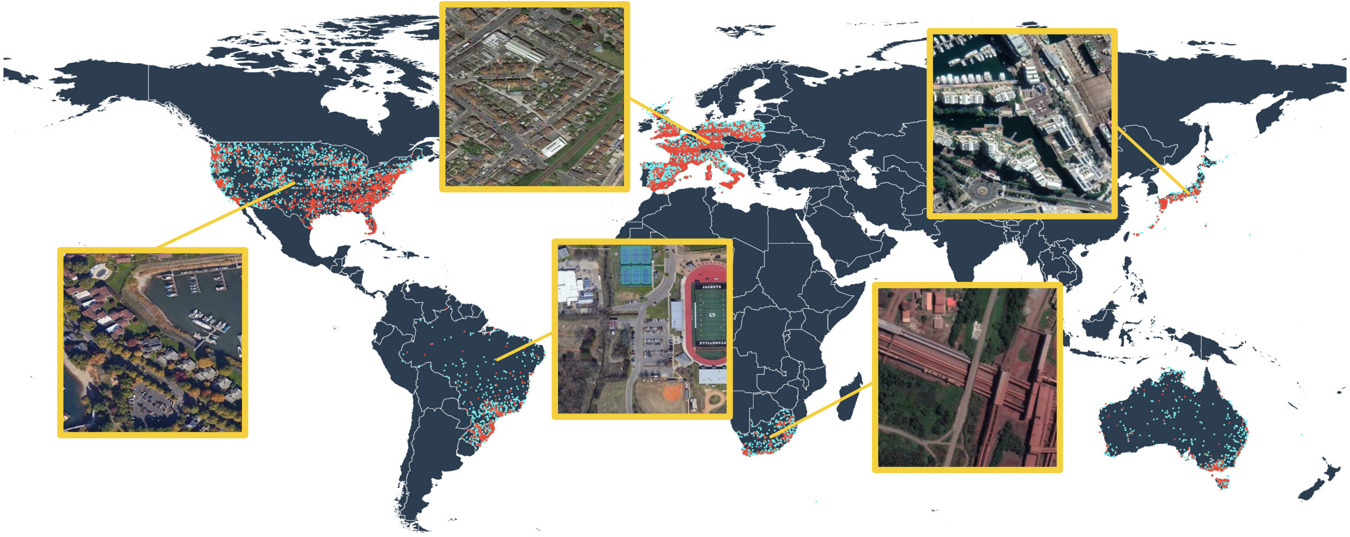


Fig. 3. Coordinate distributions and examples of multiscene aerial images in our dataset. Red dots denote images with both crowdsourced and clean labels, and cyan dots represent images with only crowdsourced scene labels.

3) *RSSCN7 [39]:* The RSSCN7 dataset is a collection of 2800 high-resolution images each belonging to one of seven scene categories: grassland, forest, farmland, parking lot, river/lake, industrial region, and residential region; 400 images with different spatial resolutions are cropped from Google Earth imagery for each scene, and the image size is 400 × 400 pixels.

4) *AID [27]:* The AID dataset is a large-scale benchmark consisting of 10 000 aerial images and 30 scene types: airport, pond, forest, baseball field, resort, bare land, center, beach, bridge, commercial, desert, storage tanks, farmland, industrial, mountain, park, parking, playground, viaduct, church, railway station, river, school, meadow, sparse residential, dense residential, medium residential, square, stadium, and port. Google Earth is exploited to acquire image samples, and the spatial resolution of each sample varies from 0.5 to 8 m/pixel. The size of images is 600 × 600 pixels, and the number of images for each class ranges from 220 to 420.

5) *NWPU-RESISC45 [2]:* The NWPU-RESISC45 dataset contains 31 500 high-resolution images and each is assigned with one of 45 scene labels. For each scene, 700 images with a size of 256 × 256 pixels are acquired from Google Earth imagery, and their spatial resolutions vary from 0.2 to 30 m/pixel.

In addition, we note that BigEarthNet [60] is a large-scale dataset for multilabel learning, where 590 326 Sentinel-2 images are captured over the European Union, and their spatial resolutions range from 10 to 60 m/pixel. Since BigEarthNet focuses on land covers instead of scenes, we do not specify it here. Table I presents an overview of public high-resolution aerial image datasets from the perspectives of dataset scales, image resolutions, scene categories, and annotations.

B. Multiscene for Multiscene Recognition

Although there are already variant datasets for aerial scene recognition, most of them can only be used for single-scene classification. In this article, we aim to take a step toward a more general scenario, multiscene recognition in single images, and produce the MultiScene dataset.

To be more specific, we collect 100 000 high-resolution aerial images from Google Earth imagery, which cover six continents, Europe, Asia, North America, South America, Africa, and Oceania, and 11 countries including Germany, France, Italy, England, Spain, Poland, Japan, USA, Brazil, South Africa, and Australia (cf. Fig. 3). This can ensure high intraclass diversity, as different scene appearances resulted from different cultural regions are covered. The spatial resolution of each image ranges from 0.3 to 0.6 m/pixel, and the spatial size of images is 512 × 512 pixels. In contrast to single-scene image datasets [27], [37]–[39], we put no constraints on the location and area of the dominant/trivial scene in an image during the data collection process. Some example multiscene images are shown in Fig. 4. In total, 36 scene categories are defined: apron, baseball field, basketball field, beach, bridge, cemetery, commercial, farmland,

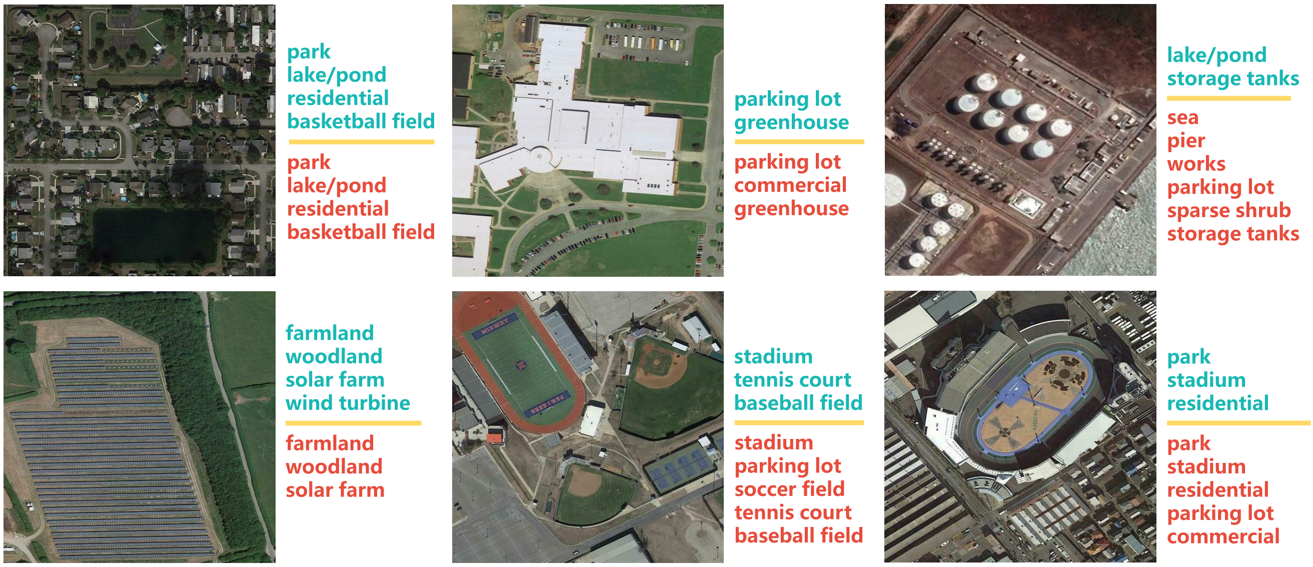


Fig. 4. Example multiscene aerial images with their crowdsourced and clean annotations in the MultiScene dataset.

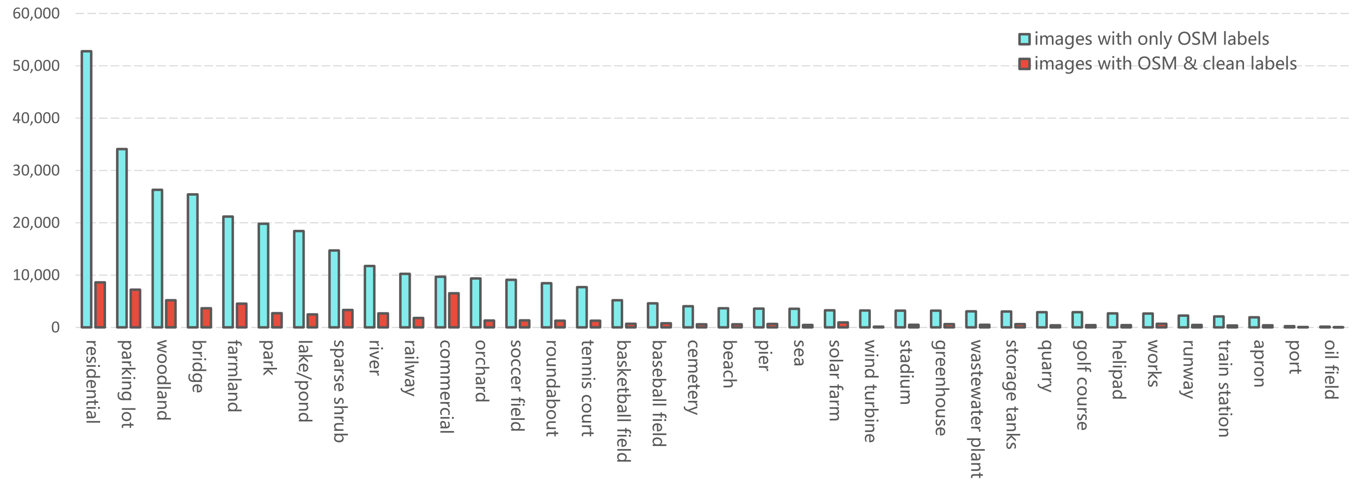


Fig. 5. Sample distributions of all scene categories in our dataset. Each cyan bar indicates the number of images assigned only OSM labels with respect to each scene category, and red bars represent the numbers of images with both OSM and clean labels.

woodland, golf course, greenhouse, helipad, lake/pond, oil field, orchard, parking lot, park, pier, port, quarry, railway, residential, river, roundabout, runway, soccer field, solar farm, sparse shrub, stadium, storage tanks, tennis court, train station, wastewater, plant, wind turbine, works, and sea.

To obtain crowdsourced annotations, we first localize each image in OSM with coordinates of its four corners. Afterward, we parse properties of scenes present in the corresponding region and label images accordingly. In this way, crowdsourced annotations of all aerial images can be automatically yielded at a very low cost compared to conventional manual labeling. However, these almost free annotations might suffer from noise as aforementioned in Section I, and the performance of networks directly trained on them could be degraded. Therefore, we visually inspect 14 000 images from all six continents and correct their labels, yielding a subset, MultiScene-Clean. Fig. 3 shows the coordinate distribution of all images, and the number of samples associated with each scene is present in Fig. 5. Compared to other scene recognition datasets (cf. Table I), our dataset is featured by its manifold labels per image and the available crowdsourced annotations. Fig. 6 further shows the number of images associated with different numbers of scenes.

C. Challenges

Compared to existing aerial scene datasets, our dataset brings more challenges to the field of scene interpretation from the following three perspectives.

1) Images are unconstrained and large-scale, and thus, scenes are likely to be incomplete and trivial, which makes recognition more difficult.
2) The long-tail sample distribution (see Fig. 5) poses a challenge of learning unbiased models on an imbalanced dataset.

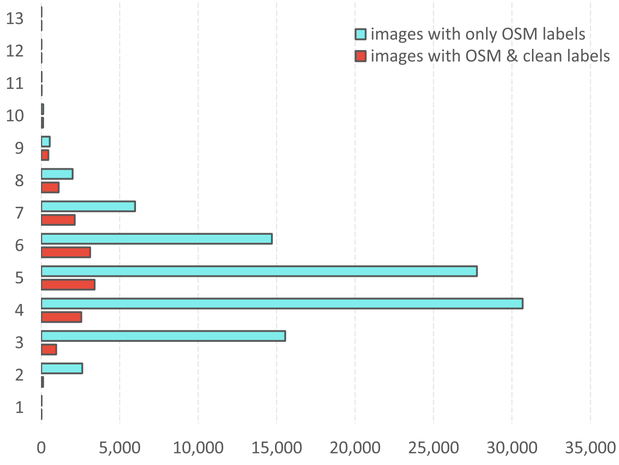


Fig. 6. Number of images associated with different numbers of scenes. Y-axis indicates the number of scenes and X-axis represents the number of images. The legend is the same as that in Fig. 5.

3) We gather images from different cultural regions, which results in a high intraclass variation.

IV. Experiments

A. Experimental Setup

1) Data Configuration: We evaluate the performance of existing models on both MultiScene-Clean and MultiScene datasets. As to the former, we use 7000 cleanly labeled images to train and validate networks, and the remaining images are utilized to test networks. For the latter, we leverage the same test set but train deep neural networks on the other 93 000 images with only crowdsourced annotations.

2) Evaluation: For a comprehensive evaluation, we measure the performance of baseline models with class-based, example-based, and overall metrics. Let L and N be numbers of classes and examples,[3] and these metrics are calculated as follows.

1) *Class-Based Metrics:* Mean class-based precision (mCP), mean class-based recall (mCR), F_1 (mCF$_1$) score, and per-class average precision (AP) are calculated for measuring the performance of networks from the perspective of class. Specifically, mCP, mCR, and mCF$_1$ score are computed as

$$\mathrm{mCP} = \frac{1}{L}\sum_{c=1}^{L}\frac{\mathrm{TP}_c}{\mathrm{TP}_c+\mathrm{FP}_c},\ \mathrm{mCR} = \frac{1}{L}\sum_{c=1}^{L}\frac{\mathrm{TP}_c}{\mathrm{TP}_c+\mathrm{FN}_c}$$
$$\mathrm{mCF}_1 = \frac{1}{L}\sum_{c=1}^{L}\frac{\mathrm{TP}_c}{\mathrm{TP}_c+\frac{1}{2}(\mathrm{FP}_c+\mathrm{FN}_c)} \tag{1}$$

where TP_c, FN_c, and FP_c represent the numbers of true positives, false negatives, and false positives with respect to the cth class, respectively. As to the per-class AP, we first rank all examples according to the predicted probability of the cth class in each of them. Then, we calculate the corresponding AP with the following formula:

$$\mathrm{AP} = \frac{1}{N_c}\sum_{k=1}^{N}\frac{\mathrm{TP}_c@k}{\mathrm{TP}_c@k+\mathrm{FP}_c@k}\times \mathrm{rel}@k \tag{2}$$

where N_c denotes the number of examples including the cth class and $\mathrm{TP}_c@k$ and $\mathrm{FP}_c@k$ represent the numbers of true and false positives in top-k examples, respectively. Notably, $\mathrm{TP}_c@k$ and $\mathrm{FP}_c@k$ are equivalent to TP_c and FP_c when k equals to N. rel@k denotes the relevance between the kth example and the cth class, and it is set to 0/1 when the cth class is included/excluded. Besides, the mean average precision (mAP) can be computed by averaging APs for all categories.

2) *Example-Based Metrics:* Mean example-based precision (mEP), mean example-based recall (mER), and mean example-based F_1 (mEF$_1$) score are computed to validate networks from the perspective of example with the following equations:

$$\mathrm{mEP} = \frac{1}{N}\sum_{k=1}^{N}\frac{\mathrm{TP}_k}{\mathrm{TP}_k+\mathrm{FP}_k},\ \mathrm{mER} = \frac{1}{N}\sum_{k=1}^{N}\frac{\mathrm{TP}_k}{\mathrm{TP}_k+\mathrm{FN}_k}$$
$$\mathrm{mEF}_1 = \frac{1}{N}\sum_{k=1}^{N}\frac{\mathrm{TP}_k}{\mathrm{TP}_k+\frac{1}{2}(\mathrm{FP}_k+\mathrm{FN}_k)} \tag{3}$$

where TP_k, FP_k, and FN_k denote the numbers of true positives, false positives, and false negatives in the kth example.

3) *Overall Metrics:* Overall precision (OP), overall recall (OR), and overall F_1 (OF$_1$) score can be used to measure the performance of models from a more holistic perspective, and they are calculated as

$$\mathrm{OP} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},\quad \mathrm{OR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$
$$\mathrm{OF}_1 = \frac{\mathrm{TP}}{\mathrm{TP}+\frac{1}{2}(\mathrm{FP}+\mathrm{FN})} \tag{4}$$

where TP, FP, and FN are counted based on predictions of all scenes and examples.

B. Baselines

To provide comprehensive benchmarks, we evaluate the performance of extensive popular deep neural networks. Since they were originally designed for single-label classification, we substitute sigmoid functions for their softmax activations to predict multiple scene labels that are encoded into multihot binary sequences. Besides, several classical machine learning algorithms are also evaluated. In total, 22 models are tested on both MultiScene-Clean and MultiScene datasets, and a brief review is as follows.

1) *Support Vector Machine (SVM) [61]:* It aims to learn one or several hyperplanes for separating samples of different classes with the largest margin. Usually, the hyperplanes are constructed in a high-dimensional space and can be learned directly (linear SVM) or through kernel functions (nonlinear SVM). In our experiments, we select the latter and use a radial basis function (RBF) kernel [62] to learn SVM.

[3] An example indicates an image that has multiple labels.

TABLE II

NUMERICAL RESULTS OF BASELINE MODELS ON THE MULTISCENE-CLEAN DATASET (%). MODELS ARE TRAINED AND TESTED ON CLEANLY LABELED IMAGES, AND THE BEST SCORES ARE SHOWN IN BOLD

Model	mAP	mCP	mCR	mCF$_1$	mEP	mER	mEF$_1$	OP	OR	OF$_1$
SVM	14.9	19.6	8.4	8.6	62.2	32.8	41.1	66.9	32.2	43.5
RF	15.6	25.4	8.7	9.5	64.6	32.5	41.4	70.9	32.1	44.2
XGBOOST	16.9	34.1	11.2	12.8	67.0	37.4	45.8	69.6	36.5	47.9
VGG-16	56.5	63.3	47.9	53.6	74.9	64.3	67.0	73.6	63.1	67.9
VGG-19	56.4	62.9	47.7	53.3	74.8	64.1	66.8	73.5	62.7	67.7
Inception-V3	53.5	65.0	40.8	48.5	74.2	59.9	63.9	73.0	58.6	65.0
ResNet-50	62.0	74.8	45.9	55.1	79.7	62.7	67.9	79.0	61.4	69.1
ResNet-101	63.0	75.9	46.6	55.8	79.9	64.3	69.1	79.2	63.1	70.3
ResNet-152	63.8	74.9	49.1	57.7	**80.8**	64.0	69.2	**80.1**	62.8	70.4
SqueezeNet	46.3	58.1	36.8	43.5	71.3	58.0	61.3	70.0	56.9	62.7
MobileNet-V2	58.8	70.9	44.8	53.1	77.6	62.7	67.0	76.6	61.6	68.3
ShuffleNet-V2	50.7	61.8	38.1	45.7	73.8	58.2	62.5	73.0	57.0	64.0
DenseNet-121	62.2	74.6	45.1	54.4	79.5	61.8	67.3	79.1	60.6	68.6
DenseNet-169	63.2	76.7	45.8	55.3	80.4	63.4	68.6	79.6	62.3	69.9
ResNeXt-50	63.4	**77.3**	45.0	54.2	78.5	64.3	68.6	77.8	63.2	69.8
ResNeXt-101	**64.8**	76.5	48.6	57.3	79.3	66.6	**70.2**	78.5	65.4	**71.3**
MnasNet	53.8	61.8	42.9	49.9	73.0	59.4	63.0	72.1	58.1	64.3
KFBNet	58.8	68.8	45.2	53.3	77.9	64.2	68.1	77.3	63.0	69.4
FACNN	56.5	60.3	48.7	52.6	73.1	65.3	66.8	71.6	64.1	67.7
SAFF	61.8	72.5	48.1	56.7	79.4	63.9	68.6	78.7	62.8	69.9
LR-VGG-16	58.1	67.7	46.7	54.2	77.3	64.6	68.0	76.2	63.5	69.2
LR-ResNet-50	63.1	68.1	**53.1**	**59.0**	76.7	**67.6**	69.7	75.3	**66.5**	70.6

2) *Random Forest (RF) [63]:* It is an ensemble of decision trees, which are trained with random subspaces of image features and make final predictions through the majority voting. The number of decision trees is set to 200 in our experiments.
3) *XGBOOST:* XGBOOST[4] is a computationally efficient implementation of gradient-boosted trees [64] that optimizes tree ensembles (e.g., an ensemble of decision trees) through successive learning steps [65]. In each step, the existing trees are fixed, and a new tree is added and optimized with objective functions. Considering the difficulty of our task, we set the number of trees to 200 for training XGBOOST on both datasets.
4) *VGGNet [66]:* VGGNet utilizes five convolutional blocks and three fully connected layers to extract high-level features for image classification. Each block has multiple stacked convolutional layers and ends with one max-pooling layer. The size of convolutional filters is 3 × 3, and the stride of max-pooling layers is 2. In our experiments, a 16-layer VGGNet (VGG-16) and a 19-layer VGGNet (VGG-19) are trained on our dataset.
5) *Inception Networks [67]–[70]:* They are characterized by their wide modules, where convolutional filters of variant sizes and max-pooling operators are jointly employed to learn diverse features. Besides, a bottleneck architecture made of 1 × 1 convolutions is introduced to mitigate the boosted computational cost resulting from heavy inception modules. In Tables II and III, we report the performance of Inception-v3 [69] in multiscene recognition.
6) *ResNet [71]:* ResNet aims to address the degradation problem by learning residual mappings with shortcut connections. By doing so, ResNet can go much deeper than plain CNNs and achieve outstanding performance in not only image classification but also semantic segmentation and object detection tasks. In our experiments, we evaluate a 50-layer ResNet (ResNet-50), a 101-layer ResNet (ResNet-101), and a 152-layer ResNet (ResNet-152) on the proposed dataset. Notably, residual blocks in these deep ResNets are modified into bottleneck architectures for reducing the computational burden.
7) *SqueezeNet [72]:* It focuses on preserving network performance with fewer parameters. To achieve this, most of 3 × 3 convolutional filters are replaced with 1 × 1 filters, and features are squeezed in the channel dimension before being fed into the remaining 3 × 3 filters. In addition, bypass connections are introduced to features of the same size for improving the classification performance. Experimental results of SqueezeNet on our dataset are reported in Sections IV-D and IV-D2.
8) *MobileNet [73]:* It is a lightweight deep neural network, which is applicable on mobile devices with restricted computational sources. The network is designed in a streamlined architecture, and depth-wise separable convolutions play a significant role in increasing computational efficiency. Specifically, such convolutions are implemented by factorizing standard convolutions into depth-wise and point-wise convolutions. The former is conducted on each channel, and the latter aggregates channel-wise outputs via 1 × 1 convolutions. To further reduce the computational cost, two hyperparameters, width multiplier α and resolution multiplier β, are designed to shrink feature channels and input resolutions, respectively. In the advanced variation of MobileNet, i.e., MobileNet-V2 [74], inverted residual connections and linear bottlenecks are developed to

[4] https://xgboost.readthedocs.io/en/latest/tutorials/model.html

TABLE III

COMPARISONS OF APs ON THE MULTISCENE-CLEAN DATASET (%). THE BEST APs ARE SHOWN IN BOLD

Model	apron	baseball field	basketball field	beach	bridge	cemetery	commercial	farmland	woodland	golf course	greenhouse	helipad	lake/pond	oil field	orchard	parking lot	park	pier	port	quarry	railway	residential	river	roundabout	runway	soccer field	solar farm	sparse shrub	stadium	storage tanks	tennis court	train station	wastewater plant	wind turbine	works	sea
SVM	3.1	6.3	5.4	4.0	29.9	4.5	60.5	44.1	55.2	3.1	5.2	3.0	18.0	0.1	9.9	65.5	20.2	5.5	0.6	3.0	13.4	68.3	19.3	8.7	3.7	9.5	8.0	25.7	3.7	5.2	9.1	2.7	3.6	0.8	5.2	3.4
RF	3.1	6.3	5.4	10.9	28.1	4.5	62.7	53.2	54.6	3.1	5.4	3.0	17.4	0.1	10.1	65.6	20.2	7.1	0.6	3.0	13.6	73.3	19.5	8.7	3.7	9.5	10.6	26.2	3.7	4.9	9.1	2.7	3.6	0.8	5.2	3.4
XGBOOST	3.1	9.4	5.4	15.8	33.4	4.5	62.8	56.2	57.6	3.1	5.2	3.0	21.5	0.1	11.5	68.8	23.3	9.3	0.6	3.0	13.8	74.7	20.3	8.7	4.2	9.8	10.9	32.9	3.7	5.2	9.1	2.7	3.6	0.8	5.2	4.3
VGG-16	72.2	81.7	24.2	70.0	72.1	28.9	81.6	87.8	85.7	65.1	42.8	34.9	63.2	1.9	72.0	86.2	50.1	72.7	19.0	50.9	55.6	93.7	52.5	65.8	68.7	61.6	32.4	59.7	53.7	49.4	63.6	35.5	45.9	51.6	27.2	54.9
VGG-19	70.1	80.7	21.3	67.1	71.7	28.0	80.6	87.4.	85.5	64.5	44.4	33.5	64.1	2.6	73.4	86.6	50.7	72.4	17.2	50.7	56.4	93.8	52.3	68.7	68.6	62.3	32.3	58.2	53.0	49.1	66.8	37.6	47.0	50.3	28.3	54.5
Inception-V3	67.8	83.2	20.0	68.8	69.9	19.8	81.4	85.0	83.1	51.4	36.4	33.0	55.2	0.5	68.3	86.0	50.0	70.5	13.5	49.0	47.9	92.3	49.8	63.9	65.4	59.7	31.1	56.6	57.9	44.6	55.5	34.5	46.9	44.5	22.7	59.6
ResNet-50	77.3	86.7	26.4	79.4	74.6	39.7	83.3	88.4	86.7	76.7	49.3	43.1	66.0	0.5	76.8	88.2	55.3	77.2	24.7	55.7	62.4	94.3	59.6	71.2	74.8	67.7	40.3	61.9	63.0	55.0	68.8	46.7	54.5	50.3	36.0	68.9
ResNet-101	79.7	88.0	27.6	80.2	75.9	44.5	84.2	88.4	87.3	75.6	49.7	45.3	68.7	0.9	77.8	88.6	58.3	77.6	20.9	61.2	62.7	94.5	61.5	73.3	77.4	70.0	40.0	62.1	64.6	54.2	70.9	46.6	55.8	47.2	**37.8**	68.1
ResNet-152	79.1	**88.6**	27.4	**84.0**	77.1	42.4	83.9	88.7	87.6	77.4	51.8	46.9	68.8	0.4	78.3	89.2	**59.4**	**79.3**	20.3	59.4	65.1	94.5	61.9	74.4	77.7	70.7	41.0	62.8	65.1	**57.2**	72.5	51.3	57.3	48.8	35.5	**71.1**
SqueezeNet	51.4	73.8	18.2	57.6	59.4	15.6	77.4	83.8	81.6	50.1	34.0	12.3	51.3	0.7	65.9	83.7	44.2	56.8	20.5	34.7	46.2	93.2	44.9	57.7	58.0	45.7	27.4	53.2	38.9	35.9	56.1	22.8	26.8	18.0	18.9	49.1
MobileNet-V2	74.3	84.4	24.8	78.5	73.4	32.5	81.9	87.2	85.9	72.2	46.3	38.9	64.3	1.6	73.8	88.3	53.7	72.2	18.1	51.8	60.0	93.6	55.8	71.7	67.8	64.1	34.3	60.4	58.7	46.3	69.8	40.7	47.1	48.5	31.5	64.0
ShuffleNet-V2	62.0	78.4	23.5	67.5	66.8	14.4	80.5	84.1	82.5	61.0	36.6	18.0	56.5	0.4	65.0	85.5	48.6	61.9	11.0	40.4	50.6	92.5	48.9	57.0	66.5	59.4	31.0	57.3	51.7	37.5	56.3	30.2	36.4	20.4	27.0	58.8
DenseNet-121	79.0	87.5	28.3	80.9	75.1	37.9	83.5	87.7	85.5	**78.0**	48.0	47.3	66.3	2.9	75.7	88.5	58.3	77.7	25.3	58.4	62.3	94.1	58.5	73.6	73.9	67.1	39.1	61.6	61.4	54.3	69.9	47.4	56.4	50.2	31.0	66.5
DenseNet-169	81.8	88.1	26.3	81.4	76.9	42.6	**84.5**	88.1	86.0	77.2	49.0	44.1	67.4	4.3	78.7	88.8	56.6	78.1	21.4	58.2	62.9	94.4	60.5	72.4	75.4	69.4	41.2	61.2	**65.9**	54.6	72.4	51.4	58.6	53.2	35.1	67.6
ResNeXt-50	81.5	87.1	27.6	81.6	75.5	41.3	83.4	88.2	86.3	76.1	50.8	50.3	66.3	9.6	75.3	89.0	57.6	77.7	23.8	58.6	64.5	94.1	61.4	73.3	76.7	68.7	40.9	62.2	63.5	53.8	71.8	50.2	56.9	**54.7**	34.9	66.6
ResNeXt-101	**82.3**	87.7	**30.2**	82.9	**77.2**	**45.6**	84.1	**88.8**	87.3	77.1	**52.6**	**54.8**	**71.0**	1.3	**79.4**	**89.7**	58.6	76.1	20.3	**61.8**	**66.5**	94.5	**64.3**	75.0	**79.2**	**72.0**	**42.3**	63.3	64.5	55.4	75.1	**54.4**	**58.7**	52.5	37.8	67.2
MnasNet	69.4	84.0	21.6	70.9	67.4	20.7	78.9	84.7	82.4	63.0	42.0	27.5	58.3	0.7	70.5	85.7	49.1	69.5	18.1	45.2	51.6	91.4	48.0	66.1	66.4	59.0	33.3	55.9	53.2	42.4	61.5	32.8	42.6	33.4	25.6	62.6
KFBNet	68.4	83.0	27.2	75.1	75.4	37.6	82.2	88.7	86.7	68.4	47.8	47.0	67.4	6.3	75.7	89.1	55.3	73.1	10.1	57.2	58.5	**94.6**	55.1	72.5	68.7	64.7	35.4	60.3	46.5	48.5	74.8	31.6	47.3	49.1	26.0	61.2
FACNN	69.4	83.7	21.5	70.5	72.9	31.7	81.8	88.4	85.4	66.4	36.9	36.2	65.9	4.8	72.8	87.5	50.4	71.1	12.6	57.8	54.5	93.6	55.2	70.1	68.4	64.5	32.4	56.4	50.0	50.8	68.4	32.3	46.1	46.4	23.9	54.2
SAFF	74.5	86.2	29.8	76.8	75.1	41.4	83.0	88.7	86.6	76.9	50.0	49.6	67.7	1.9	76.8	88.9	58.7	74.8	17.1	51.6	62.1	94.2	58.8	**76.5**	72.4	68.3	39.8	**63.3**	57.6	54.5	**77.8**	41.9	54.6	46.6	33.7	65.9
LR-VGG-16	76.3	82.5	19.9	74.7	71.0	26.6	82.5	86.8	86.4	70.7	41.4	41.0	65.1	**11.0**	72.0	87.5	52.5	73.1	10.4	57.4	58.5	93.5	56.3	70.7	71.2	62.7	27.3	58.1	51.6	50.5	69.0	43.9	52.0	45.9	30.5	62.5
LR-ResNet-50	78.5	88.2	24.6	80.8	75.9	44.2	83.8	88.7	**87.8**	76.2	50.9	48.1	67.4	0.9	77.1	88.8	57.3	76.8	**29.3**	57.5	63.3	94.2	61.0	72.4	73.0	70.5	42.1	62.3	64.8	55.1	72.0	48.7	55.9	48.9	36.4	67.3

improve the network performance. In our experiments, we train MobileNet-V2 and set both α and β as the default value, 1.

9) *ShuffleNet [75]:* It improves the computational efficiency by utilizing point-wise group convolutions and channel shuffle. Specifically, the former divides feature maps into several groups and conducts 1 × 1 convolutions on each group independently. The latter rearranges feature channels for enabling information to flow across channels belonging to different groups. Besides, element-wise addition, which is often used in a residual block, is replaced with concatenation for enlarging channel dimension at a low computational cost. In ShuffleNet-V2 [76], features are grouped by channel split, and point-wise group convolutions are discarded. As a consequence, two feature groups are yielded and fed into two branches, of which one is an identity mapping and the other is a set of convolutions. Afterward, outputs are concatenated and shuffled along the channel dimension. In our experiments, we evaluate the performance of ShuffleNet-V2 on our dataset.

10) *DenseNet [77]:* It proposes to enhance information flow by directly connecting each layer to all subsequent layers with equivalent feature-map sizes. To preserve information learned by proceeding layers, concatenation is employed to combine features from various layers. By reusing feature maps throughout entire networks, DenseNet can learn compact internal representations for visual recognition tasks. Two variations, a 121-layer DenseNet (DenseNet-121) and a 169-layer DenseNet (DenseNet-169), are tested.

11) *ResNeXt [78]:* It learns residuals with aggregated residual transformations but not a stack of convolutional layers (e.g., ResNet). The aggregated residual transformation is implemented by first slicing features into low-dimensional embeddings and then conducting convolutions on them. Afterward, outputs are aggregated with element-wise addition. With this design, ResNeXt outperforms its ResNet counterpart on ImageNet-5K [78] and COCO [79] datasets. We test a 50-layer ResNext (ResNeXt-50) and a 101-layer (ResNeXt-101) in our experiments.

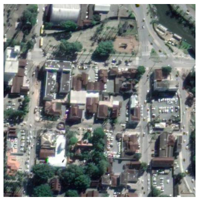
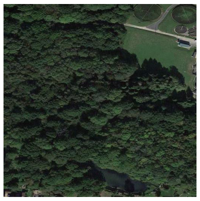
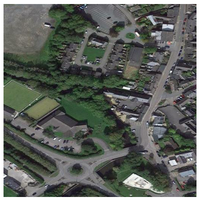
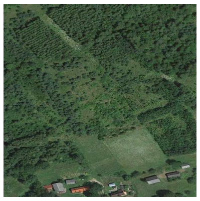
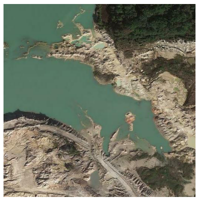
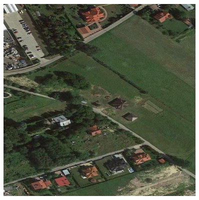
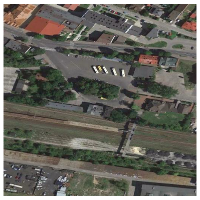
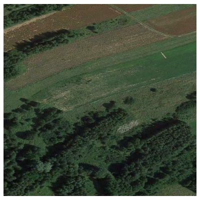
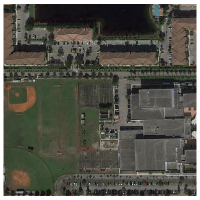
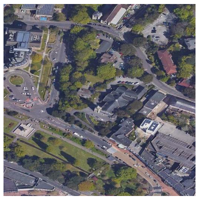
12) *MnasNet [80]:* MnasNet architectures are automatically learned on target datasets through a mobile neural architecture search (MNAS) algorithm [80]. Compared to conventional NAS algorithms [81], MNAS takes not only classification accuracy but also model latency into consideration and is executed on mobile phones for measuring real-world inference latency. As a consequence, MnasNet searched on target datasets is expected to achieve a good tradeoff between accuracy and latency. To control the model size, a depth multiplier is designed for scaling the number of channels in each layer. In our experiments, the depth multiplier is set to 1, and the best-performing MnasNet searched on the ImageNet dataset [82] is chosen to perform multiscene recognition in the wild.

13) *KFBNet [49]:* It exploits a key region capturing method namely key filter bank (KFB) for aerial image scene classification. The proposed KFB is composed of two streams: a global stream (G-Stream) and a key stream (K-Stream). The former predicts labels using features learned by the last block of a CNN, while the latter highlights key features in both spatial and channel dimensions for inferring scene categories. Finally, predictions made by the two streams are merged via an element-wise addition as the final decision. We take VGG-16 as the backbone and report numerical results in Tables II, III, and V.

14) *FACNN [50]:* It is a scene classification network composed of a CNN backbone and a feature aggregation module. In the latter, features extracted by the last

TABLE IV

EXAMPLE PREDICTIONS OF RESNEXT-101 ON THE MULTISCENE-CLEAN DATASET

Multi-scene Aerial Images in the MultiScene-Clean dataset					
Ground Truths	bridge, parking lot, river, roundabout, and residential	woodland, lake/pond, and wastewater plant	bridge, parking lot, river, roundabout, and residential	farmland, woodland, orchard, residential, and sparse shrub	lake/pond and quarry
Predictions	bridge, parking lot, river, roundabout, and residential	woodland, lake/pond, and wastewater plant	bridge, parking lot, river, roundabout, and residential	farmland, woodland, orchard, residential, and sparse shrub	lake/pond and quarry
Multi-scene Aerial Images in the MultiScene-Clean dataset					
Ground Truths	commercial, farmland, parking lot, and residential	commercial, parking lot, park, railway, residential, train station, and works	farmland, woodland, sparse shrub	baseball field, basketball field, lake/pond, parking lot, residential, soccer field, and tennis court	bridge, commercial, parking lot, park, residential, river, roundabout, and solar farm
Predictions	commercial, farmland, woodland, parking lot, and residential	commercial, parking lot, park, railway, residential, train station, and works	farmland, woodland, lake/pond, and sparse shrub	baseball field, baseketball field, lake/pond, parking lot, residential, soccer field, and tennis court	bridge, commercial, lake/pond, parking lot, park, river, solar farm, and residential

Purple predictions indicate false negatives, while blue predictions are false positives.

TABLE V

NUMERICAL RESULTS OF BASELINE MODELS ON THE MULTISCENE DATASET (%). MODELS ARE TRAINED ON IMAGES WITH NOISY CROWDSOURCED ANNOTATIONS AND TESTED ON CLEANLY LABELED IMAGES. THE BEST SCORES ARE SHOWN IN BOLD

Model	**mAP**	**mCP**	**mCR**	**mCF$_1$**	**mEP**	**mER**	**mEF$_1$**	**OP**	**OR**	**OF$_1$**
SVM	14.7	24.7	4.1	5.4	51.4	15.7	23.1	77.7	15.5	25.8
RF	15.1	49.7	4.4	6.1	55.4	16.4	34.3	78.7	15.8	26.3
XGBOOST	18.4	54.6	10.6	14.9	62.0	26.7	35.1	70.4	25.5	37.4
VGG-16	63.4	71.0	46.9	54.1	78.4	51.6	59.3	79.3	49.6	61.0
VGG-19	59.8	68.9	47.2	54.1	75.5	52.2	58.9	75.1	50.2	60.2
Inception-V3	65.8	74.1	50.8	58.5	79.1	53.8	61.2	79.5	51.9	62.8
ResNet-50	63.9	73.7	47.7	55.9	78.3	52.5	60.0	78.5	50.7	61.6
ResNet-101	63.4	73.0	47.5	55.5	77.1	52.5	59.7	77.2	50.6	61.2
ResNet-152	62.8	73.2	47.6	55.7	76.2	53.1	59.9	76.5	51.3	61.4
SqueezeNet	61.4	74.4	41.1	50.5	78.9	47.7	56.4	80.7	45.9	58.5
MobileNet-V2	65.5	72.3	48.4	56.0	79.6	54.6	62.0	80.1	52.8	63.6
ShuffleNet-V2	65.1	74.6	46.7	55.1	81.7	51.0	59.9	**82.9**	49.0	61.6
DenseNet-121	67.5	77.0	49.4	58.2	**82.2**	54.4	**62.6**	82.8	52.3	**64.1**
DenseNet-169	64.2	71.3	**53.3**	**59.3**	77.2	**55.7**	62.0	77.1	**53.9**	63.4
ResNeXt-50	63.9	73.6	49.0	56.9	77.5	52.6	59.8	77.6	50.7	61.3
ResNeXt-101	60.8	68.5	47.4	53.7	73.8	51.2	57.7	74.0	49.5	59.3
MnasNet	58.1	74.1	31.0	40.4	75.0	38.0	47.6	80.4	36.0	49.7
KFBNet	67.1	**77.7**	46.2	54.3	80.2	54.0	61.7	81.1	52.3	63.6
FACNN	65.2	73.9	47.6	55.6	78.9	53.9	61.1	80.0	52.1	63.1
SAFF	64.8	74.0	47.4	55.3	80.9	51.2	59.8	81.8	49.4	61.6
LR-VGG-16	**67.8**	76.0	48.4	56.1	80.5	52.1	60.5	81.3	50.3	62.2
LR-ResNet-50	65.5	71.2	51.6	57.9	79.2	53.1	60.7	79.4	51.2	62.3

three blocks of VGG-16 are aggregated through pooling operations and 1 × 1 convolutions. Afterward, they are concatenated with outputs of the second fully connected layer of VGG-16 to form discriminative scene representations for the final prediction.

15) *SAFF [51]:* It proposes a nonparametric self-attention layer for enhancing spatial and channel responses of feature maps. Specifically, features extracted by the last three blocks of a pretrained CNN (e.g., VGG-16) are fused and fed into the proposed self-attention layer.

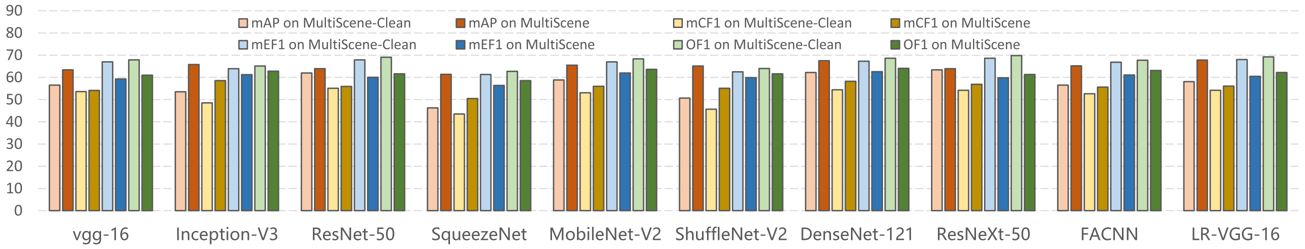


Fig. 7. Comparisons of the performance of networks trained on images with clean (light-color bars) and crowdsourced (dark-color bars) annotations. For each network, the left four bars represent class-based scores, mAPs and CF_1, while the right four bars indicate EF_1 and OF_1 scores.

In this layer, spatial- and channel-wise weightings are conducted to emphasize the importance of locations of salient objects and channels with infrequently occurring features, respectively. Principal component analysis (PCA) whitening is also introduced to reduce the information redundancy and squash channels. However, since this operation frequently fails the network training, we replace it with a learnable fully connected layer. Besides, VGG-16 is selected as the backbone in our experiments.

16) *LR-CNN [56]:* It is a multilabel classification network, which consists of three elements: a class-wise feature extraction module, an attentional region extraction module, and a relational reasoning module. Specifically, the first module learns deep features with respect to each category from the input images. Afterward, the second module extracts attentional regions of class-wise features, which are eventually leveraged to reason about relations between different objects for inferring their existences through the third module. In our experiments, we validate LR-VGG-16 and LR-ResNet-50, where VGG-16 and ResNet-50 are taken as backbones, respectively.

C. Training Details

Before training SVM, RF, and XGBOOST, we use histogram of oriented gradient (HOG) [83] and local binary pattern (LBP) [84] as visual features as recommended in [85]. The size of each cell is set to 32 × 32 pixels for HOG, and the radius is defined as 16 pixels for LBP. We use Scipy to implement these machine learning classifiers and apply them to multiscene recognition using the function *MultiOutputClassifier*[5]. As to baseline classification neural networks, we initialize them with weights pretrained on the ImageNet dataset and fine-tune them on the proposed multiscene image dataset. The loss is defined as binary cross entropy, and stochastic gradient descent (SGD) with momentum [86] is selected as the optimizer. To accelerate the network convergence, the momentum is set to a large value, 0.9. Besides, the initial learning rate and the weight decay are set to 0.02 and $1e^{-}4$, respectively. All deep networks are implemented on Pytorch and validated on one NVIDIA Tesla V100-SXM2 32GB GPU. For experiments on both MultiScene-Clean and MultiScene, we train networks for 87k and 581k iterations, respectively, and the size of each training batch is set to 16 for both versions.

D. Experimental Results Across Different Tasks

1) Multiscene Recognition With Cleanly Labeled Data: To evaluate baselines for our task, we conduct experiments on the MultiScene-Clean dataset and report quantitative results in Table II. It can be seen that ResNeXt-101 achieves the best mAP (64.8%), mEF_1 (70.2%), and OF_1 score (71.3%), which demonstrates its high performance and robustness in this task from almost all perspectives. LR-ResNet-50 gains the highest value in mCF_1 (59.0%) due to its capability of reasoning about relations among various scenes. Moreover, such a reasoning capability also enables LR-ResNet-50 to surpass the other baselines in all recall metrics, as scenes tend to be predicted as positive once its related scenes are recognized. Another observation is that MnasNet, SqueezeNet, and ShuffleNet-V2 show relatively poor performance due to their lightweight designs. Compared to deep neural networks, traditional machine learning algorithms achieve lower scores in all metrics.

For an insight into the performance of networks in identifying different scenes, we also report per-class APs in Table III. As we can see, ResNeXt-101 achieves the highest APs in most scenes, which is in line with the previous observations. Furthermore, we note that most networks fail to accurately recognize scenes having scarce training samples, e.g., oil field and port. This suggests that learning unbiased models on an imbalanced dataset is a big challenge. Besides numerical results, we exhibit several predictions in Table IV.

2) Learning From Noisy Crowdsourced Labels: We investigate networks learned from noisy crowdsourced labels for our task on the MultiScene dataset. To ensure a fair comparison, we utilize the same test set as in Section IV-D and report numerical results in Table V. It can be observed that OF_1 scores of all models are decreased by an average of 8.2% compared to the values in Table II, which demonstrates that noise in crowdsourced annotations significantly affects the learning of deep neural networks. Moreover, it is interesting to note that the values of class-based metrics, mAP and mCF_1 score, are increased by 4.6% and 1.2%, respectively, in comparison with those in Table II. This can be attributed to the fact that numbers of training samples, especially for scenes seldomly appearing, are effortlessly increased by crawling OSM data with keyword

[5] https://scikit-learn.org/stable/modules/generated/sklearn.multioutput.MultiOutputClassifier.html

searching. Compared to models showing high performance on the MultiScene-Clean dataset, we find that DenseNet gains the highest scores in mCF$_1$ (59.3%), mEF$_1$ (62.6%), and mOF$_1$ (64.1%), as it can sufficiently reuse features and has relatively few parameters. Besides, LR-VGG-16 achieves the highest mAP (67.8%), which demonstrates that taking advantage of underlying relations among various scenes can suppress the influence of noise introduced by OSM data. Furthermore, we compare the performance of several networks trained on the MultiScene-Clean and MultiScene datasets in Fig. 7, and it can be again observed that higher class-based scores (see orange and brown bars in Fig. 7) are obtained when using massive crowdsourced labels. All in all, although crowdsourced labels influence the overall performance of networks, comparisons in class-based scores also suggest their great potential.

V. Conclusion

In this article, we propose a large-scale dataset, MultiScene, for multiscene recognition in single images, which is featured by unconstrained multiscene aerial images and the available both crowdsourced and clean labels. The proposed dataset allows studies in not only recognizing aerial scenes in the wild but also learning from noisy crowdsourced labels. We comprehensively evaluate popular baseline models on both MultiScene-Clean (a subset consisting of only cleanly labeled images) and MultiScene datasets. Experimental results on the former demonstrate that unconstrained multiscene recognition is still a challenging task, and those on the latter showcase the great potential of exploiting a large number of crowdsourced annotations. Looking into the future, the dataset can be applied to develop more efficient networks and learning strategies for exploiting noisy labels for aerial scene understanding in the wild.

References

[1] Q. Weng, Z. Mao, J. Lin, and X. Liao, "Land-use scene classification based on a CNN using a constrained extreme learning machine," *Int. J. Remote Sens.*, vol. 39, no. 19, pp. 1–19, Oct. 2018.

[2] 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.

[3] D. Wen, X. Huang, H. Liu, W. Liao, and L. Zhang, "Semantic classification of urban trees using very high resolution satellite imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 4, pp. 1413–1424, Apr. 2017.

[4] S. Manfreda *et al.*, "On the use of unmanned aerial systems for environmental monitoring," *Remote Sens.*, vol. 10, no. 4, p. 641, 2018.

[5] L. Mou and X. X. Zhu, "IM2HEIGHT: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network," 2018, *arXiv:1802.10249*. [Online]. Available: https://arxiv.org/abs/1802.10249

[6] C. Qiu, L. Mou, M. Schmitt, and X. X. Zhu, "Local climate zone-based urban land cover classification from multi-seasonal Sentinel-2 images with a recurrent residual network," *ISPRS J. Photogramm. Remote Sens.*, vol. 154, pp. 151–162, Aug. 2019.

[7] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 135, pp. 158–172, Jan. 2018.

[8] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 20–32, Jun. 2018.

[9] L. Mou and X. X. Zhu, "RiFCN: Recurrent network in fully convolutional network for semantic segmentation of high resolution remote sensing images," 2018, *arXiv:1805.02091*. [Online]. Available: https://arxiv.org/abs/1805.02091

[10] 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.

[11] C. Qiu, M. Schmitt, C. Geiß, T.-H.-K. Chen, and X. X. Zhu, "A framework for large-scale mapping of human settlement extent from Sentinel-2 images via fully convolutional neural networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 163, pp. 152–170, May 2020.

[12] Q. Li, Y. Shi, X. Huang, and X. X. Zhu, "Building footprint generation by integrating convolution neural network with feature pairwise conditional random field (FPCRF)," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7502–7519, Nov. 2020.

[13] G. Cheng, L. Guo, T. Zhao, J. Han, H. Li, and J. Fang, "Automatic landslide detection from remote-sensing imagery using a scene classification method based on BoVW and pLSA," *Int. J. Remote Sens.*, vol. 34, no. 1, pp. 45–59, Jan. 2013.

[14] 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.

[15] 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.

[16] D. Marcos, M. Volpi, B. Kellenberger, and D. Tuia, "Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 96–107, Nov. 2018.

[17] A. Vetrivel, M. Gerke, N. Kerle, and F. Nex, "Disaster damage detection through synergistic use of deep learning and 3D point cloud features derived from very high resolution oblique aerial images, and multiple-kernel-learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 45–59, Jun. 2018.

[18] W. Lee, S. Kim, Y.-T. Lee, H.-W. Lee, and M. Choi, "Deep neural networks for wild fire detection with unmanned aerial vehicle," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Aug. 2017, pp. 252–253.

[19] D. Hu *et al.*, "Cross-task transfer for geotagged audiovisual aerial scene recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 68–84.

[20] H. Sun, S. Li, X. Zheng, and X. Lu, "Remote sensing scene classification by gated bidirectional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 82–96, Jan. 2020.

[21] J. Murray, D. Marcos, and D. Tuia, "Zoom in, zoom out: Injecting scale invariance into landuse classification CNNs," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 5240–5243.

[22] 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. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, no. 99, pp. 3735–3756, Jun. 2020.

[23] A. P. Byju, G. Sumbul, B. Demir, and L. Bruzzone, "Remote sensing image scene classification with deep neural networks in JPEG 2000 compressed domain," 2020, *arXiv:2006.11529*. [Online]. Available: https://arxiv.org/abs/2006.11529

[24] Y. Xu, B. Du, and L. Zhang, "Assessing the threat of adversarial examples on deep neural networks for remote sensing scene classification: Attacks and defenses," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1604–1617, Feb. 2021.

[25] L. Ru, B. Du, and C. Wu, "Multi-temporal scene classification and scene change detection with correlation based fusion," 2020, *arXiv:2006.02176*. [Online]. Available: https://arxiv.org/abs/2006.02176

[26] Q. Li, C. Qiu, L. Ma, M. Schmitt, and X. Zhu, "Mapping the land cover of Africa at 10 m resolution from multi-source remote sensing data with Google Earth engine," *Remote Sens.*, vol. 12, no. 4, p. 602, Feb. 2020.

[27] 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.

[28] P. Jin, G.-S. Xia, F. Hu, Q. Lu, and L. Zhang, "AID++: An updated version of AID on scene classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 4721–4724.

[29] D. Tuia, D. Marcos, and G. Camps-Valls, "Multi-temporal and multi-source remote sensing image classification by nonlinear relative normalization," *ISPRS J. Photogramm. Remote Sens.*, vol. 120, pp. 1–12, Oct. 2016.

[30] S. Niazmardi, B. Demir, L. Bruzzone, A. Safari, and S. Homayouni, "Multiple kernel learning for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1425–1443, Mar. 2018.

[31] 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. [Online]. Available: https://ieeexplore.ieee.org/document/9023551

[32] J. Lin, L. Mou, T. Yu, X. X. Zhu, and Z. J. Wang, "Dual adversarial network for unsupervised ground/satellite-to-aerial scene adaptation," in *Proc. ACM Int. Conf. Multimedia (ACMMM)*, 2020, pp. 10–18.

[33] X. Wang, X. Xiong, and C. Ning, "Multi-label remote sensing scene classification using multi-bag integration," *IEEE Access*, vol. 7, pp. 120399–120410, 2019.

[34] Q. Zhu, X. Sun, Y. Zhong, and L. Zhang, "High-resolution remote sensing image scene understanding: A review," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 3061–3064.

[35] Y. Long *et al.*, "On creating benchmark dataset for aerial image interpretation: Reviews, guidances and million-AID," 2020, *arXiv:2006.12485*. [Online]. Available: https://arxiv.org/abs/2006.12485

[36] S. Zorzi and F. Fraundorfer, "Regularization of building boundaries in satellite images using adversarial and regularized losses," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 5140–5143.

[37] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst. (GIS)*, 2010, pp. 270–279.

[38] J. Hu, T. Jiang, X. Tong, G. Xia, and L. Zhang, "A benchmark for scene classification of high spatial resolution remote sensing imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 5003–5006.

[39] 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.

[40] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[41] V. Risojević, S. Momić, and Z. Babić, "Gabor descriptors for aerial image classification," in *Proc. Int. Conf. Adapt. Natural Comput. Algorithms*, 2011, pp. 51–60.

[42] M. J. Swain and D. H. Ballard, "Color indexing," *Int. J. Comput. Vis.*, vol. 7, no. 1, pp. 11–32, 1991.

[43] J. dos Santos, O. Penatti, and R. da Silva Torres, "Evaluating the potential of texture and color descriptors for remote sensing image retrieval and classification," in *Proc. Int. Conf. Comput. Vis. Theory Appl. (VISAPP)*, 2010, pp. 203–208.

[44] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 837–842, Aug. 1996.

[45] V. Risojevic and Z. Babic, "Aerial image classification using structural texture similarity," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol. (ISSPIT)*, Dec. 2011, pp. 190–195.

[46] L.-J. Zhao, P. Tang, and L.-Z. Huo, "Land-use scene classification using a concentric circle-structured multiscale bag-of-visual-words model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 12, pp. 4620–4631, Dec. 2014.

[47] M. Lienou, H. Maitre, and M. Datcu, "Semantic annotation of satellite images using latent Dirichlet allocation," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 28–32, Jan. 2010.

[48] Y. Zhong, Q. Zhu, and L. Zhang, "Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 11, pp. 6207–6222, Nov. 2015.

[49] F. Li, R. Feng, W. Han, and L. Wang, "High-resolution remote sensing image scene classification via key filter bank based on convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 8077–8092, Nov. 2020.

[50] X. Lu, H. Sun, and X. Zheng, "A feature aggregation convolutional neural network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7894–7906, Oct. 2019.

[51] R. Cao, L. Fang, T. Lu, and N. He, "Self-attention-based deep feature fusion for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 1, pp. 43–47, Jan. 2020.

[52] A. Zeggada, F. Melgani, and Y. Bazi, "A deep learning approach to UAV image multilabeling," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 694–698, May 2017.

[53] A. Zeggada, S. Benbraika, F. Melgani, and Z. Mokhtari, "Multilabel conditional random field classification for UAV images," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 3, pp. 399–403, Mar. 2018.

[54] Y. Hua, L. Mou, and X. X. Zhu, "Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 149, pp. 188–199, Mar. 2019.

[55] G. Sumbul and B. Demir, "A deep multi-attention driven approach for multi-label remote sensing image classification," *IEEE Access*, vol. 8, pp. 95934–95946, 2020.

[56] Y. Hua, L. Mou, and X. X. Zhu, "Relation network for multilabel aerial image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4558–4572, Jul. 2020.

[57] S. Koda, A. Zeggada, F. Melgani, and R. Nishii, "Spatial and structured SVM for multilabel image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5948–5960, May 2018.

[58] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1144–1158, Feb. 2017.

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

[60] 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. (IGARSS)*, Jul. 2019, pp. 5901–5904.

[61] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[62] J.-P. Vert, K. Tsuda, and B. Schölkopf, "A primer on kernel methods," *Kernel Methods Comput. Biol.*, vol. 47, pp. 35–70, Dec. 2004.

[63] T. K. Ho, "Random decision forests," in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, Aug. 1995, pp. 278–282.

[64] T. Hastie, R. Tibshirani, and J. Friedman, "Boosting and additive trees," in *The Elements of Statistical Learning*. New York, NY, USA: Springer, 2009.

[65] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[66] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[67] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[68] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 448–456.

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

[70] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2017, pp. 4278–4284.

[71] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[72] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5 MB model size," 2016, *arXiv:1602.07360*. [Online]. Available: https://arxiv.org/abs/1602.07360

[73] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: https://arxiv.org/abs/1704.04861

[74] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "MobileNets: Efficient convolutional neural networks for mobile vision applications," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Apr. 2018, pp. 1–9.

[75] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 6848–6856.

[76] N. Ma, X. Zhang, H. Zheng, and J. Sun, "ShuffleNetV2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 116–131.

[77] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[78] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.

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

[80] M. Tan *et al.*, "MnasNet: Platform-aware neural architecture search for mobile," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2820–2828.

[81] B. Zoph and Q. Le, "Neural architecture search with reinforcement learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–16.

[82] 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. (CVPR)*, Jun. 2009, pp. 248–255.

[83] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.

[84] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[85] Y. Lin *et al.*, "Large-scale image classification: Fast feature extraction and SVM training," in *Proc. CVPR*, Jun. 2011, pp. 1689–1696.

[86] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2013, pp. 1139–1147.

Yuansheng Hua (Graduate Student Member, IEEE) received the bachelor's degree in remote sensing science and technology from Wuhan University, Wuhan, China, in 2014, and the double master's degrees in Earth oriented space science and technology (ESPACE) and photogrammetry and remote sensing from the Technical University of Munich (TUM), Munich, Germany, and Wuhan University in 2018 and 2019, respectively. He is currently pursuing the Ph.D. degree with the German Aerospace Center (DLR), Weßling, Germany, and TUM.

In 2019, he was a Visiting Researcher with Wageningen University and Research, Wageningen, The Netherlands. His research interests include remote sensing, computer vision, and deep learning, especially their applications in remote sensing.

Lichao Mou received the bachelor's degree in automation from Xi'an University of Posts and Telecommunications, Xi'an, China, in 2012, the master's degree in signal and information processing from the University of Chinese Academy of Sciences (UCAS), Beijing, China, in 2015, and the Dr.Ing. degree from the Technical University of Munich (TUM), Munich, Germany, in 2020.

He is currently a Guest Professor at the Munich AI Future Lab AI4EO, TUM, and the Head of the Visual Learning and Reasoning Team, Department "EO Data Science," Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Weßling, Germany. Since 2019, he has been a Research Scientist at DLR-IMF and an AI Consultant for the Helmholtz Artificial Intelligence Cooperation Unit (HAICU). In 2015, he spent six months at the Computer Vision Group, University of Freiburg, Freiburg im Breisgau, Germany. In 2019, he was a Visiting Researcher with Cambridge Image Analysis Group (CIA), University of Cambridge, Cambridge, U.K.

Dr. Mou was a recipient of the First Place in the 2016 IEEE GRSS Data Fusion Contest and finalists for the Best Student Paper Award at the 2017 Joint Urban Remote Sensing Event and the 2019 Joint Urban Remote Sensing Event.

Pu Jin (Member, IEEE) received the bachelor's degree in electronic information science and technology from Wuhan University, Wuhan, China, in 2017, and the double master's degrees in Earth oriented space science and technology (ESPACE) and photogrammetry and remote sensing from the Technical University of Munich (TUM), Munich, Germany, and Wuhan University in 2020 and 2021, respectively. He is currently pursuing the Ph.D. degree with the German Aerospace Center (DLR), Weßling, Germany, and TUM.

His research interests include remote sensing, computer vision, and deep learning, especially their applications in remote sensing.

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

Since 2019, she has been a Co-Coordinator of the Munich Data Science Research School. Since 2019, she has been heading the Helmholtz Artificial Intelligence—Research Field "Aeronautics, Space and Transport." Since May 2020, she has been the Director of the International Future AI Lab (AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond), Munich. Since October 2020, she has been the Co-Director of the Munich Data Science Institute (MDSI), TUM. She was a Guest Scientist or a Visiting Professor at 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. She is currently a Professor with the Data Science in Earth Observation (former: Signal Processing in Earth Observation), TUM, and the Head of the Department "EO Data Science," Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany. She is also a Visiting AI Professor at ESA's Phi-lab. 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 the young academy (Junge Akademie/Junges Kolleg) at Berlin-Brandenburg Academy of Sciences and Humanities, the German National Academy of Sciences Leopoldina, and the Bavarian Academy of Sciences and Humanities. She serves on the Scientific Advisory Board in several research organizations, including the German Research Center for Geosciences (GFZ) and Potsdam Institute for Climate Impact Research (PIK). She is also an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. She serves as an Area Editor for Special Issues of *IEEE Signal Processing Magazine*.