UNCONSTRAINED AERIAL SCENE RECOGNITION WITH DEEP NEURAL NETWORKS AND A NEW DATASET

Yuansheng Hua[1,2], *Lichao Mou*[2], *Pu Jin*[1], *Xiao Xiang Zhu*[1,2]

[1]Data Science in Earth Observation, Technical University of Munich (TUM), Munich, Germany
[2]Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany

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 paper, we investigate a more practical yet underexplored task—multi-scene recognition in single images. To this end, we create a large-scale dataset, called MultiScene dataset, composed of 100,000 unconstrained images each with multiple labels from 36 different scenes. Among these images, 14,000 of them are manually interpreted and assigned ground-truth labels, while the remaining images are provided with crowdsourced labels, which are generated from low-cost but noisy OpenStreetMap (OSM) data. By doing so, our dataset allows two branches of studies: 1) developing novel CNNs for multi-scene recognition and 2) learning with noisy labels. We experiment with extensive baseline models on our dataset to offer a benchmark for multi-scene recognition in single images. Aiming to expedite further researches, we will make our dataset and pre-trained models available[1].

Index Terms— Convolutional neural network (CNN), multi-scene recognition in single images, crowdsourced annotations, large-scale aerial image dataset

1. INTRODUCTION

With the recent development of earth observation techniques, massive aerial imagery is now accessible for a variety of applications, such as urban planning [1, 2] and land-used/land-cover mapping [3, 4, 5, 6]. As one of crucial steps towards 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) has drastically pushed ahead with researches in this field, and enormous achievements [7, 8, 9] have been obtained. Albeit successful, most existing scene classification researches only focus on a specific scenario, where an aerial image is assumed to include a single scene.

[1]https://github.com/Hua-YS/Multi-Scene-Recognition

(a) Single-scene recognition (b) Multi-scene recognition

Fig. 1. Examples of images utilized in (a) single-scene and (b) multi-scene 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 there indeed exist multiple scenes. For a clear visualization, locations of scenes are marked in (b).

As a consequence, 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 has a large coverage (cf. Fig. 1(b)).

Hence, in this paper, we aim to tackle a more realistic yet challenging problem, namely multi-scene recognition in 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, multi-scene recognition is more arduous because 1) images are large-scale and unconstrained, and 2) all present scenes in an aerial image need to be exhaustively recognized. However, very few efforts have been deployed to this problem in the remote sensing community.

In order to advance the progress of multi-scene recognition in single images, we propose a large-scale Multi-Scene

Table 1. Comparison with existing scene recognition datasets from various perspectives.

Dataset	# images	# scenes	# labels per image	crowdsourced label	Year
UC-Merced [10]	2,100	21	1	Not available	2010
WHU20 [11]	5,000	20	1	Not available	2015
RSSCN7 [12]	2,800	7	1	Not available	2015
AID [7]	10,000	30	1	Not available	2017
NWPU-RESISC45 [13]	31,500	45	1	Not available	2017
MultiScene (Ours)	100,000	36	1-13	Available	2020

Recognition (MultiScene) dataset, where 100,000 aerial images are collected around the world and assigned with multiple scene labels. In the phase of data acquisition, we note that although massive high-resolution aerial images can be effortlessly obtained from remote sensing data platforms, such as Google Earth [2], it is extremely time- and labor-consuming to yield their corresponding multiple scene labels. To tackle such annotation burden, in this paper, we resort to crowdsourced data, i.e., OpenStreetMap [3] (OSM), and it has been proven to be successful in generating image-level annotations [7, 8] and pixel-wise footprints [14] for training deep networks. However, we note that OSM data might suffer from two common defects, incompleteness and incorrectness, which could introduce severe noise into image labels. With this in mind, here we did not treat crowdsourced labels as ground truths like previous works [7, 8]. Instead, we manually inspect a portion of collected images and rectify their crowdsourced annotations to yield corresponding ground-truth labels. As a consequence, the MultiScene dataset provides two types of labels, *ground-truth* and *crowdsourced* annotations, and enables two branches of studies: 1) developing novel deep networks and training them on clear data (i.e., ground-truth data) for multi-scene recognition and 2) learning networks from massive noisy crowdsourced labels.

2. THE MULTISCENE DATASET FOR UNCONSTRAINED MULTI-SCENE RECOGNITION

We collect 100,000 high-resolution aerial images from Google Earth imagery, which cover six continents, i.e., Europe, Asia, North America, South America, Africa, and Oceania, and eleven countries, including Germany, France, Italy, England, Spain, Poland, Japan, the United States, Brazil, South Africa, and Australia. This can ensure large intra-class diversities of scenes, as their variant appearances resulted from cultural differences are covered. The spatial resolution of each image ranges from 0.3 m/pixel to 0.6 m/pixel, and the size is 512×512 pixels. Since we will annotate all scenes in each image, there are no specific constraints on locations and areas of dominant/trival scenes in an image. In total, 36 scene categories are defined: 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, and sea.

To yield crowdsourced annotations, we first localize each image in OSM with coordinates of its four corners. Afterwards, we parse properties of scenes present in the corresponding region and distill multiple scene labels accordingly. In this way, crowdsourced annotations of all aerial images can be automatically yielded at a very low cost compared to conventional manual annotation. However, these cheap annotations might suffer from noise as aforementioned in Section 1, and the performance of networks directly trained on them could be degraded. Therefore, we manually inspect 14,000 images from all six continents and rectify their multiple scene labels, yielding a subset of cleanly labeled images, MultiScene-Clean. In the MultiScene-Clean dataset, multiple ground-truth scene-level labels are available for each high-resolution aerial image. The number of samples associated with each scene is present in Fig. 2. Moreover, we compare our dataset with commonly used scene recognition datasets in Table 1. It can be seen that our dataset is featured by manifold labels per image and available crowdsourced annotations.

- Unlike conventional aerial scene recognition where all images are well-cropped and each of them contains only one scene-level label, in this paper, we explore a more practical task—multi-scene recognition in single images.

- We propose a large-scale dataset, namely MultiScene, consisting of 100,000 unconstrained multi-scene 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.

- The proposed dataset provides not only ground truth data but also crowdsourced labels, which enables researches in learning from enormous noisy labels for our task.

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

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

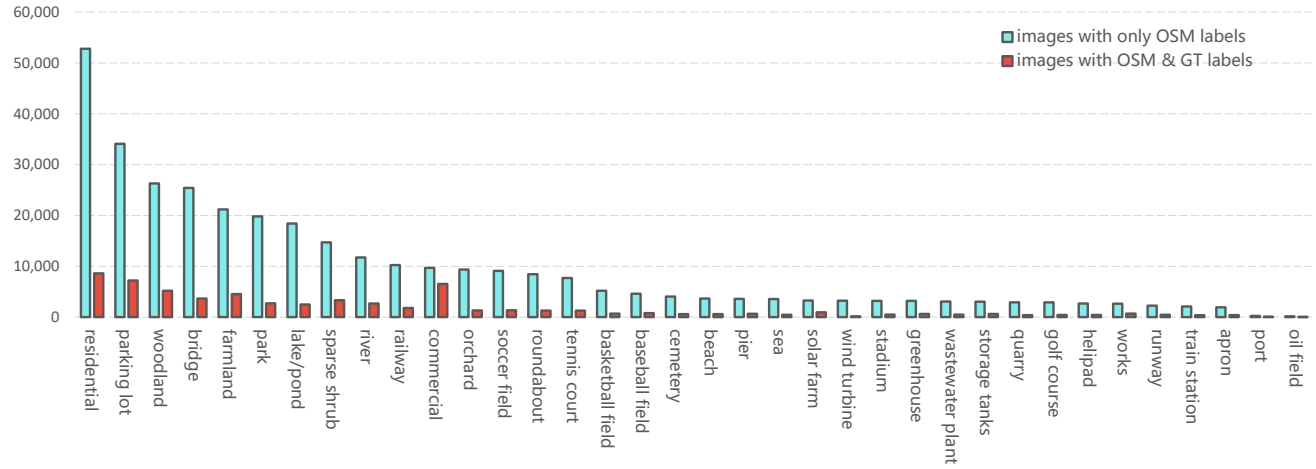

Fig. 2. Sample distributions of all scene categories in the MultiScene dataset. Each blue bar indicates the number of images assigned only OSM labels with respect to each scene category, and red bars represent numbers of images with both OSM and ground-truth (GT) labels.

3. EXPERIMENTS

3.1. Experimental Setup

Data. In this work, we intend to answer two questions: 1) How current classification models perform in unconstrained multi-scene recognition? 2) Can crowdsourced annotations help with this task? To address the first question, we validate the performance of existing models on the MultiScene-Clean dataset and leverage only clean labels. In the training phase, 7,000 images are utilized for training and validation, while others are chosen to build a test set. As to the second question, we keep the test set as the same as that in the first configuration and train deep neural networks on the remaining 93,000 images with only crowdsourced annotations.

Evaluation metric. For a comprehensive evaluation, the performance of baseline models are measured from the perspectives of the class and example[4]. As to class-based metrics, we calculate the class-based precision (CP), recall (CR), and F_1 score of each scene class is computed as follows:

$$\begin{aligned} \mathrm{CP} = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c}, \ \mathrm{CR} = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c}, \\ \mathrm{CF}_1 = 2\frac{\mathrm{CP} \cdot \mathrm{CR}}{\mathrm{CP} + \mathrm{CR}}, \end{aligned} \tag{1}$$

where TP_c, FP_c, and FN_c represent numbers of true positives, false positives, and false negatives with respect to each scene class c, respectively. By averaging CP, CR and CF_1 of all scene categories, the mean CP, CR, and CF_1 can be obtained and abbreviated as mCP, mCR, mCF_1. Similarly, the mean example-based precision (mEP), recall (mER), and F_1 score (mEF_1) are computed by taking average of EP, ER, and EF_1, which are calculated with the following equations:

$$\begin{aligned} \mathrm{EP} = \frac{\mathrm{TP}_e}{\mathrm{TP}_e + \mathrm{FP}_e}, \ \mathrm{ER} = \frac{\mathrm{TP}_e}{\mathrm{TP}_e + \mathrm{FN}_e}, \\ \mathrm{EF}_1 = 2\frac{\mathrm{EP} \cdot \mathrm{ER}}{\mathrm{EP} + \mathrm{ER}}, \end{aligned} \tag{2}$$

where TP_e, FP_e, and FN_e denote numbers of true positives, false positives and false negatives in an example. Moreover, we measure the performance of all baselines from an overall perspective with the overall precision (OP), recall (OR), and F_1 score (OF_1). These metrics are calculated as:

$$\begin{aligned} \mathrm{OP} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \ \mathrm{OR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \\ \mathrm{OF}_1 = 2\frac{\mathrm{OP} \cdot \mathrm{OR}}{\mathrm{OP} + \mathrm{OR}}, \end{aligned} \tag{3}$$

where TP, FP, and FN are counted based on the prediction of each scene in each example. Therefore, here we mainly consider OF_1 as it can holistically evaluate the performance of each model under the circumstance of imbalanced distribution across different classes.

3.2. Baseline results

To validate the performance of existing classification networks in unconstrained multi-scene recognition, we conduct experiments on the MultiScene-Clean dataset and report quantitative results in Tabel 2. It can be seen that LR-ResNet-50 achieves the best mCF_1 (59.7%), mEF_1 (69.7%), and OF_1 (70.6%), which demonstrate its high performance in multi-scene recognition from class-based, example-based, and overall perspectives. Besides, we also report numerical results of baselines trained on crowdsourced annotations in Table 3. It can be seen that DesNet-121 gains the highest

[4]An example indicates an image with multiple scene labels.

Table 2. Numerical results of baseline models on the MultiScene-CLean dataset (%). Models are trained and tested on clean annotations, and the best scores are shown in bold.

Model	mCP	mCR	mCF_1	mEP	mER	mEF_1	OP	OR	OF_1
ResNet-50 [15]	74.8	45.9	56.9	**79.7**	62.7	67.9	79.0	61.4	69.1
DenseNet-121 [16]	74.6	45.1	56.2	79.5	61.8	67.3	**79.1**	60.6	68.6
ResNeXt-50 [17]	**77.3**	45.0	56.9	78.5	64.3	68.6	77.8	63.2	69.8
SqueezeNet [18]	58.1	36.8	45.0	71.3	58.0	61.3	70.0	56.9	62.7
LR-ResNet-50 [19]	68.1	**53.1**	**59.7**	76.7	**67.6**	**69.7**	75.3	**66.5**	**70.6**

Table 3. Numerical results of baseline models on the MultiScene dataset (%). Models are trained on crowdsourced annotations and tested on clean annotations. The best scores are shown in bold.

Model	mCP	mCR	mCF_1	mEP	mER	mEF_1	OP	OR	OF_1
ResNet-50 [15]	73.7	47.7	57.9	78.3	52.5	60.0	78.5	50.7	61.6
DenseNet-121 [16]	**75.0**	**54.0**	**62.8**	**80.9**	**55.3**	**63.0**	**81.1**	**53.4**	**64.4**
ResNeXt-50 [17]	73.6	49.0	58.8	77.5	52.6	59.8	77.6	50.7	61.3
SqueezeNet [18]	74.4	41.1	53.0	78.9	47.7	56.4	80.7	45.9	58.5
LR-ResNet-50 [19]	71.2	51.6	59.8	79.2	53.1	60.7	79.4	51.2	62.3

values of all metrics, which suggests its superior capability in learning from noisy crowdsourced labels. By comparing Table 3 and Table 2, we can see that the performance of baselines learned from crowdsourced labels is lower than those trained on ground-truth labels due to introduced noise. This imposes a great challenge on developing networks and learning strategies.

4. CONCLUSION AND OUTLOOK

In this paper, we propose a large-scale dataset for multi-scene recognition in single images, MultiScene, which is featured by unconstrained aerial images and available crowdsourced and ground-truth labels. The proposed dataset allows researches in not only unconstrained multi-scene recognition but also using crowdsourced data for network training. Looking into the future, the dataset can also be applied to the field of learning from noisy labels for multi-scene recognition.

5. ACKNOWLEDGEMENTS

This work is jointly supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. [ERC-2016-StG-714087], Acronym: *So2Sat*), by the Helmholtz Association through the Framework of Helmholtz Artificial Intelligence Cooperation Unit (HAICU) - Local Unit "Munich Unit @Aeronautics, Space and Transport (MASTr)" and Helmholtz Excellent Professorship "Data Science in Earth Observation - Big Data Fusion for Urban Research" and 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".

6. REFERENCES

[1] L. Mou and X. X. Zhu, "RiFCN: Recurrent network in fully convolutional network for semantic segmentation of high resolution remote sensing images," *arXiv:1805.02091*, 2018.

[2] 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)," *TGRS*, 2020.

[3] 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 P& RS*, vol. 145, pp. 96–107, 2018.

[4] L. Mou, Y. Hua, and X. X. Zhu, "Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images," *TGRS*, vol. 58, no. 11, pp. 7557–7569, 2020.

[5] Y. Hua, D. Marcos, L. Mou, X. X. Zhu, and D. Tuia, "Semantic segmentation of remote sensing images with sparse annotations," *GRSL*, in press.

[6] L. Mou, . Saha, . Hua, F. Bovolo, L. Bruzzone, and X. X. Zhu, "Deep reinforcement learning for band selection in hyperspectral image classification," *TGRS*, 2021.

[7] G. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "AID: A benchmark data set for performance evaluation of aerial scene classification," *TGRS*, 2017.

[8] P. Jin, G. Xia, F. Hu, Q. Lu, and L. Zhang, "AID++: An updated version of aid on scene classification," in *IGARSS*, 2018.

[9] Y. Hua, L. Mou, J. Lin, K. Heidler, and X. X. Zhu, "Aerial scene understanding in the wild: Multi-scene recognition via prototype-based memory networks," *ISPRS P & RS*, in press.

[10] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2010.

[11] J. Hu, T. Jiang, X. Tong, G. Xia, and L. Zhang, "A benchmark for scene classification of high spatial resolution remote sensing imagery," in *IGARSS*, 2015.

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

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

[14] S. Zorzi and F. Fraundorfer, "Regularization of building boundaries in satellite images using adversarial and regularized losses," in *IGARSS*, 2019.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[16] G. Huang, Z. Liu, L. Van Der Maaten, and K. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017.

[17] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *CVPR*, 2017.

[18] F. Iandola, S. Han, M. Moskewicz, K. Ashraf, W. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and¡ 0.5 MB model size," in *ICLR*, 2017.

[19] Y. Hua, L. Mou, and X. X. Zhu, "Relation network for multilabel aerial image classification," *TGRS*, vol. 58, no. 7, pp. 4558–4572, 2020.