Proceedings of The 2019 IEEE International
Conference on Real-time Computing and Robotics
August 4-9, 2019, Irkutsk, Russia

A New Remote Sensing Image Dataset for Large-Scale Remote Sensing Detection *

Dongyang Xie, Jun Cheng, and Dapeng Tao

***Abstract*— With recent developments in sensor technology, remote sensing is playing an increasingly important role in military and civilian life such as town planning, mapping, resource and crop monitoring, and disaster prevention. These applications are based on remote sensing images, which ideally comprehensively reflect the information needed. However, remote sensing images are unusual, in the sense that they tend to be larger scale, contain more irrelevant information, and contain more complex scenes. These factors greatly increase the difficulty of the object detection for remote sensing task and, moreover, there is currently a lack of large datasets. Here we build a new image dataset for object detection in remote sensing. The dataset includes 108,989 images in two categories, with the images cropped from large-scale remote sensing images. To increase the relevance of our large-scale remote sensing images, we include the same area imaged at different time points. We test the dataset using faster R-CNN detection models trained on different CNN networks. In addition, we use fixed faster R-CNN for detection in large-scale remote sensing images using a model trained with VGG16. Our experimental results show that the proposed dataset provides a new benchmark for remote sensing detection, and the fixed faster R-CNN detected correctly in large-scale remote sensing images.**

***Keywords*— remote sensing, object detection, dataset**

I. INTRODUCTION

The object detection in remote sensing image method aims to detect objects in remote sensing images. Remote sensing images tend to be of larger scale and contain complex scenes and irrelevant information. Current datasets for remote sensing detection cropped from remote sensing images have poor generalization performance, and many remote sensing object detection methods are based on traditional object detection methods. Although deep convolutional neural networks (CNNs) [1] have significantly improved object detection [2, 3], the large scale of remote sensing images poses challenges to object detection in these images.

Many remote sensing image datasets contain images cropped from public remote sensing images, which are then subsequently labeled. The WHU-RS19 dataset [26] contains 50 images in each category; the UC Merced Land Use dataset [25] contains 2100 small size remote sensing images with only 100 images in each category; the SIRI-WHU dataset [27] contains 200 images in each category, overcoming the shortcomings of traditional datasets for object detection; while DOTA [22] published a large number of high-resolution remote sensing images with detailed labels. Although remote sensing datasets are gradually improving, some problems still remain:

1) The quantity in each category is insufficient;

2) Most remote sensing dataset images are too small to reflect the scenes;

3) The images are sampled from irrelevant large-scale remote sensing images.

Deep CNN networks [4] make it possible to fully extract image features, and a trained model should describe the image features more appropriately. Over recent years, several excellent object detection methods have been proposed. RCNN [5] improved the mean average precision (mAP) to 66% from 34.3% with DPM [2] on VOC2007 [6]. In 2015, fast R-CNN [7] regarded the BBox as a regression problem, which was solved with a CNN model. SSD [8] introduced a multi-scale design to make the features more effective. Faster R-CNN [9] used the RPN network to accelerate training and testing. RFCN [10] optimized the network structure, while YOLO [11], YOLO-v2 [12], and YOLO-v3 all used one-step detection as an alternative way to solve the problem. RetinaNet [14] used a one-step detection method, thereby making the original object detection method more flexible. While these methods have proven excellent for small size image detection, when they are used to detect large-scale remote sensing images, two problems must be solved:

1) These methods cannot detect the entire large-scale remote sensing image; and

2) When detecting local areas in the large-scale remote sensing image, the labeled object may not correspond to that area, which will influence the detection result.

Many studies have explored the problem of remote sensing object detection. Cheng et al. [15] tried to make the model learn rotation-invariant feature representations to overcome the absence of a regular point of view in remote sensing images.

*Research supported by National Key R&D Program of China (2018YFB1308000), National Natural Science Foundation of China (61772455, U1713213, 61772508, 61572486), Guangdong Technology Project (2017B010110007, 2016B010108010), Shenzhen Technology Project (JCYJ20180507182610734, JCYJ20170413152535587), CAS Key Technology Talent Program Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences (2014DP173025), Shenzhen Engineering Laboratory for 3D Content Generating Technologies ([2017]476), the Program for Excellent Young Talents of National Natural Science Foundation of Yunnan University (2018YDJQ004), Yunnan Natural Science Foundation (2016FB105, 2018FY001(-013)), the Program for Excellent Young Talents of Yunnan University under Grant (WX069051).

D. Xie and D. Tao are with the FIST LAB, School of Information Science and Engineering, Yunnan University, Kunming 650091, Yunnan, P.R. China and also with the Guangdong Provincial Key Laboratory of Robotics and Intelligent System, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, 518055, China.
(e-mail: shining_young@outlook.com, dapeng.tao@gmail.com).

J. Cheng is with the Guangdong Provincial Key Laboratory of Robotics and Intelligent System, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, 518055, China, and also with the Chinese University of Hong Kong, Hong Kong, 999077, China (corresponding author to provide phone: 0755-86392121; fax: 0755-86392121; e-mail: jun.cheng@siat.ac.cn).

978-1-7281-3726-1/19/$31.00 © 2018 IEEE

Authorized licensed use limited to: Zhejiang University. Downloaded on November 20,2022 at 18:13:28 UTC from IEEE Xplore. Restrictions apply.

Zhang et al. [16] proposed a classifier for fine resolution remote sensing image classification to enhance the mAP, but some detail could still be improved upon. In view of these problems, here we build the remote sensing object detection dataset (RSODD), which contains two classes, airplanes and ships. Images were selected from Google Earth, and most contained an airport and seaport sampled from all over the world. The dataset provides a large number of images in each category, 232 areas, and the same area (airport or seaport) imaged at different time periods and cropped into many small size images at a resolution of 600*600 pixels, thereby reflecting the scene well and increasing the relevance of images cropped from the same area. We also make several adjustments to faster R-CNN for improved remote sensing object detection, allowing it to complete detection in large-scale images with a model trained using unbroken object labels.

The structure of this paper is as follows. Section II provides a brief overview of related work. Section III presents our approach including the new dataset's properties and our adjustment to faster R-CNN. In section IV, we present our experiments and results, and we conclude in section V.

II. RELATED WORK

As object detection research has become more common, several object detection datasets have appeared, including the famous VOC [6] and COCO [17] datasets for object detection and segmentation research. High-quality datasets promote the development of object detection technology. However, the original object detection datasets lacked remote sensing object detection images, and the labels lacked detail. Therefore, several researchers have developed remote sensing image datasets. NWPU VHR-10 [27,28] collected the relatively small number of 800 images in 10 categories. UCAS-AOD [24] added negative labeled images, the main target objects being airplanes and motor vehicles. Han et al. [23] published a remote sensing image dataset called NWPU-RESISC45 containing 45 classes. UC Merced Land Use [25] contains 21 categories with 2100 images, while WHU-RS19 [26] includes 19 classes but only 50 images in each class. DOTA [22] contains 2806 aerial images from different aerial sensors in 15 classes, supplementing the lake in ImageNet [13] and COCO [17] for remote sensing.

The detection processes in remote sensing object detection methods and the original object detection methods are similar. The early methods included deformable part model (DPM) [2] and SIFT [3], both of which extracted the feature to describe the image using hand-crafted descriptors, although the results using these methods were relatively poor in contrast to current methods [4-10]. In a groundbreaking development in object detection, Girshick et al. [4-6] proposed conception region convolutional neural networks R-CNN [5], in which features are extracted through a CNN. R-CNN has four steps: 1) using the selective search algorithm (SS) [18] to generate 2000 regions; 2) resizing the regions and making them a regular size; 3) extracting the features from the regions and using support vector machine (SVM) [19] for classification; 4) if the region belongs to one of the target classes, using regression to box the selected regions. For improved results, before training the CNN model, the authors used the pre-trained model on ImageNet data and fine-tuned on the VOC dataset [6].

To improve performance, fast R-CNN had been modified several times, thereby not only improving the mAP but also significantly increasing the speed. The main idea of fast R-CNN is: 1) using SS to obtain 2000 regions, i.e., the same as in R-CNN; 2) using the whole image as input to extract the features by the CNN; 3) mapping the regions selected by SS [19] from the original image to the feature map; and 4) introducing the concept of SPPnet [20], where regions on the feature map are pooled to a regular size (ROI pooling) and then using the softmax classifier to classify and the bounding box regressor to ensure the location. Finally, faster R-CNN [9] has taken object detection to a new level, delivering performance gains over fast R-CNN. The RPN network pre-divides the foreground and background to improve the recall rate.

Based on the R-CNN framework, some new methods have also been proposed. For example, Zhu et al. [21] improved accuracy by introducing Markov random fields. RFCN [10] presented the concept of position-sensitive score maps to fuse the information of targets into ROI pooling, thereby simplifying and speeding up detection. For remote sensing object detection, in order to improve the robustness of detectors when processing remote sensing images, Cheng et al. [15] proposed the RIFD-CNN network to learn rotation-invariant feature representations and, at the end of the process, adding the rotation-invariant regularizer and Fisher discrimination regularizer to overcome undeterminedness caused by irregular points of view and to classify the object more accurately.

Based on these successful previous works, here we report a remote sensing dataset called RSODD (remote-sensing object detection dataset) containing 100k images, each resized to 600*600 pixels, and with two classes (airplane and ship). We also make several adjustments to faster R-CNN to improve the remote sensing object detection task and conduct detection on large-scale images.

III. OUR WORK

A. *Our Previous Work*

Before proposing this dataset, we previously collected 10k images for single object detection in remote sensing images, containing only one class (ship). However, we found that the generalization performance of the model trained with faster R-CNN was poor. Images were collected from Google Earth, 20 seaports imaged over one period (date) were selected, and these large-scale images containing the seaport were cropped into 600*600 pixel images to obtain 10k 600*600 pixel images. Trained with faster R-CNN [9] and testing on images selected from Google Earth containing lake and river images with ships (Figure 1), we found that the model worked well when the area contained seaports but not in other areas (lakes, rivers). We concluded that: 1) the test images contained areas with lakes and rivers that were not present in the training set; 2) the training images were all imaged over the same period, so were subjected to only one condition (weather, light), while the test images contained more conditions; and 3) the training set was not sufficiently large.

Authorized licensed use limited to: Zhejiang University. Downloaded on November 20,2022 at 18:13:28 UTC from IEEE Xplore. Restrictions apply.

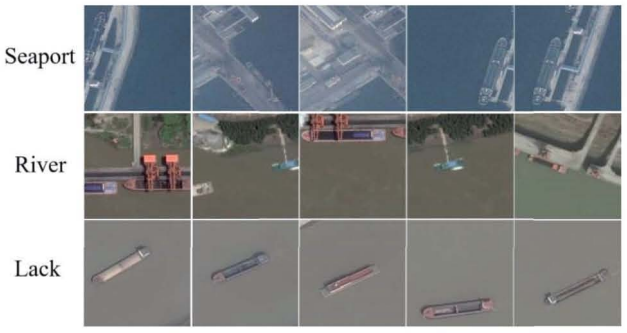

Figure 1. Line 1 are seaport images, line 2 are river images and line 3 are lake images. In wich, line 1 for training, line 2 and line 3 for testing. All of them are in size of 600*600.

B. Remote-sensing Object Detection Dataset

Having identified these problems, we expanded the dataset and added a new class. Therefore, the differences between the new and previous dataset are: 1) for the ship object, more scenes were introduced into the dataset, i.e., lake and river scenes; 2) for the same location, images from several periods were included; 3) a new class (airplane) was added. The added classes are shown in Figure.2.

Figure 2. The added class images, all from airports.

All images were collected from Google Earth. For the object ship, we obtained 29 areas containing different seaports, lakes, and rivers. To enhance the generalization performance, we obtained several images from each area at different periods to obtain 393 large-scale images in total. Owing to the use of different time periods, the weather and illumination are different for the same area in different images, which might enhance model robustness. For the category airplane, we selected 203 airports from all over the world, and again obtained images for every airport at different time periods to obtain 3361 images of different airports at different periods. The large-scale images are illustrated in Figure 3.

Since the large-scale images are too large to be used directly for detection, all large-scale images were cropped into small size images of 600*600 pixels. However, cropping large-scale images directly may divide objects into two or more parts, and it is difficult to retain the whole object in small-size images cropped from the large-scale images. To overcome this problem, large-scale images were cropped with overlap, i.e., an image of size of 900*900 pixels was cropped into four images of size 600*600 pixels, the smaller image occupying [0-600,0-600] pixels from the original large-scale image, and the second smaller image occupying [300-900,0-600] pixels from original large-scale image. The third and fourth images are [0-600,300-900] and [300-900,300-900] pixels in the original large-scale images, respectively. This overlapping strategy not only retained all whole objects in large-scale images, but also produced a new sample with part of the whole object. This may be useful for generalization performance. The final number of small images in the dataset is 108,989.

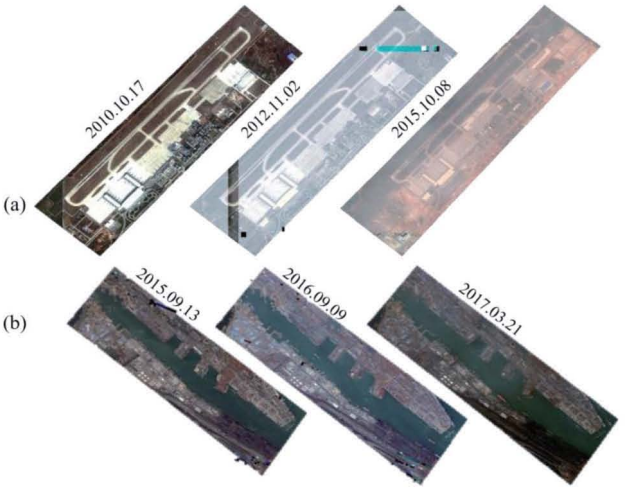

Figure 3. The large-scale images imaged at different periods. (a) WUHAN TIANHE international airport. (b) TIANJIN seaport.

The small image labels are the same as in the VOC dataset [6], and one annotation file (XML) corresponds to one image. The annotation contains the name of image, the categories, and the objects' BBox.

C. Fixed Faster R-CNN for Remote Sensing Image Detection

Based on faster R-CNN [9], we propose a version of fixed faster R-CNN [9], which enables fixed faster R-CNN to detect original large-scale images imaged by real remote sensors. First, input a large-scale image and divide it into 600*600 pixels using overlapping. Second, in parallel, detect small images and store the results. For example, if there are 4 detectors and the large-scale image is divided into 200 small images, every detector will detect 50 images, which are stored. Third, gather the results and map the results into the original images. Finally, use the NMS [28] algorithm to eliminate reduplicated results. The process is shown in Figure 4.

The cropped image's order numbers of rows and columns:

$$R_i = \frac{X_i - 600}{300} + 1 \quad (1)$$

$$C_j = \frac{Y_j - 600}{300} + 1 \quad (2)$$

R_i is the ith row, the X_i is ith row image's coordinate on the x axis, C_j is jth column, and Y_j is jth column image's coordinate on the y axis.

The final coordinate of the detected object:

$$(R_i - 1) * 300 + x_i \quad (3)$$

$$(C_j - 1) * 300 + y_j \quad (4)$$

Authorized licensed use limited to: Zhejiang University. Downloaded on November 20,2022 at 18:13:28 UTC from IEEE Xplore. Restrictions apply.

IV. EXPERIMENTS

In this section, we test the dataset using the faster R-CNN model trained with several networks. Then, using the proposed fixed faster R-CNN, we detect objects in large-scale remote sensing images to show that the dataset can train the model with faster R-CNN and the model can be used for real remote sensing image detection. For the evaluation covering the two classes (airplane and ship), the performance is measured according to PASCAL VOC criteria [6], with average precision (AP) and mAP as the metrics. We also adopt the standard IoU (intersection over union) criterion of 0.5 for all experiments.

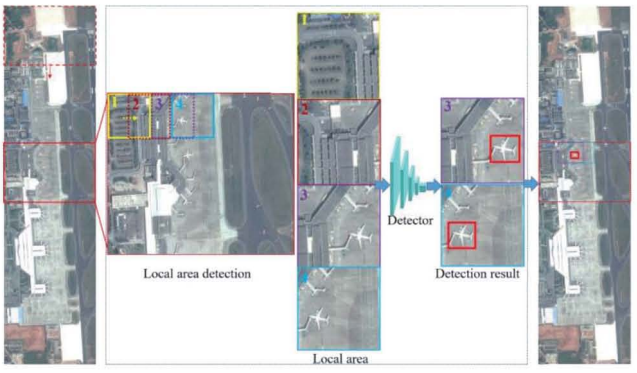

Figure 4. The proposed fixed faster R-CNN adjustment. In the local area of the large-scale image, overlap crop the small image to 600*600 pixels; in each small image, detect the object and store the results; finally, using NMS, eliminate the reduplicated results.

A. Datasets

The dataset contains 108,989 images of size 600*600 pixels, with the object airplane in 50% of images and the remaining training images being ship images. Therefore, there are 98,090 images for training and 10,899 images for testing. There are two classes (airplane, ship) and, in every image, while most contain one object, some contain more than one object. Owing to image cropping, some labels in the training or testing datasets are part of the objects. The training and testing datasets are in the same form as VOC [6].

B. Parameter Setting and Experiment Implementation

During training, 2 images are used per minibatch, the input data (ROIs) of RPN are 128, the overlap threshold for ROI is 0.5, and, when the ROI's class score is greater or equal to 0.5, the ROI will be considered as foreground. The NMS threshold for RPN is 0.7, and the batch size for RPN is 256. During testing, the NMS threshold for RPN is 0.7, and the threshold for the final class score is 0.5. The hardware used was: 1) CPU: 2 Intel Xeon E5-2620 2.1GHZ, 8 cores in each CPU. 2) Memory: 128 G .3) GPU: 8 NVIDIA GTX1080TI.

C. Object Detection on Our Dataset

We implement three different networks to extract the image features. The total number of iterations was 700,000, and in the testing process we selected the best performing model as the final result. In ZF, we selected the 35,000th iteration model. In VGG_CNN_1024, we selected the 36,000th iteration model, while in VGG16 we selected 33,000th model. The results are listed in Table 1. The model trained with VGG16 obtained the best mAP but was also the most time consuming. The model trained with ZF was the fastest, but performance was poor. Finally, the model trained with VGG_CNN_1024 obtained a balance between performance and speed.

TABLE I. THE RESULTS OF THE MODEL TRAINED ON OUR DATASET

networks	AP for airplane	AP for ship	mAP	Time consuming per image
ZF	0.888	0.739	0.814	0.042s
VGG_CNN_1024	0.902	0.847	0.874	0.045s
VGG16	0.906	0.880	0.893	0.077s

D. Testing on Satellite Images

Finally, we tested the model trained with VGG16 and tested .

The test images were selected from the location at a seaside airport, so contained airplanes and ships in the same images. The results are shown in Figure 5.

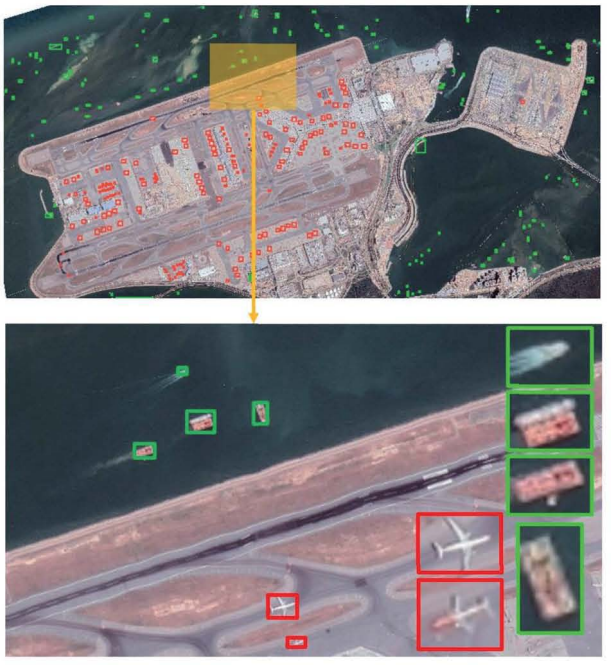

Figure 5. Using fixed faster R-CNN to detect whole large-sale remote sensing images, the red BBoxs detected airplanes and the green BBoxs detected ships. (b) Part of the local area of large-scale remote sensing images. The area is Hong Kong airport.

V. CONCLUSION

In this paper, we construct a remote sensing image dataset containing more scenes of ships in harbors, lakes, and rivers. The images collected of the same scene at different time periods are more relevant, making the proposed dataset more usable. Moreover, a new category, aircraft, has been added to the dataset. We evaluate the dataset on the models trained with faster R-CNN. By analyzing different networks, we improve faster R-CNN so that it can detect directly on original remote sensing images.

Authorized licensed use limited to: Zhejiang University. Downloaded on November 20,2022 at 18:13:28 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229,* 2013.

[2] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence,* vol. 32, no. 9, pp. 1627-1645, 2010.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580-587.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580-587.

[6] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision,* vol. 88, no. 2, pp. 303-338, 2010.

[7] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440-1448.

[8] W. Liu *et al.*, "Ssd: Single shot multibox detector," in *European conference on computer vision*, 2016, pp. 21-37: Springer.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91-99.

[10] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in neural information processing systems*, 2016, pp. 379-387.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779-788.

[12] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," *arXiv preprint,* 2017.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 248-255: Ieee.

[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE transactions on pattern analysis and machine intelligence,* 2018.

[15] G. Cheng, J. Han, P. Zhou, and D. Xu, "Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection," *IEEE Transactions on Image Processing,* vol. 28, no. 1, pp. 265-278, 2019.

[16] C. Zhang *et al.*, "A hybrid MLP-CNN classifier for very fine resolution remotely sensed image classification," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 140, pp. 133-144, 2018.

[17] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *European conference on computer vision*, 2014, pp. 740-755: Springer.

[18] J. Theeuwes, "Stimulus-driven capture and attentional set: selective search for color and visual abrupt onsets," *Journal of Experimental Psychology: Human perception and performance,* vol. 20, no. 4, p. 799, 1994.

[19] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery,* vol. 2, no. 2, pp. 121-167, 1998.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European conference on computer vision*, 2014, pp. 346-361: Springer.

[21] Y. Zhu, R. Urtasun, R. Salakhutdinov, and S. Fidler, "segdeepm: Exploiting segmentation and context in deep neural networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4703-4711.

[22] G.-S. Xia *et al.*, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. CVPR*, 2018.

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

[24] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao, "Orientation robust object detection in aerial images using deep convolutional neural network," in *Image Processing (ICIP), 2015 IEEE International Conference on*, 2015, pp. 3735-3739: IEEE.

[25] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, 2010, pp. 270-279: ACM.

[26] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sensing,* vol. 7, no. 11, pp. 14680-14707, 2015.

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

[28] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, 2006, vol. 3, pp. 850-855.

Authorized licensed use limited to: Zhejiang University. Downloaded on November 20,2022 at 18:13:28 UTC from IEEE Xplore. Restrictions apply.