I.J. Information Technology and Computer Science, **2020, 5, 63-72**
Published Online October 2020 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2020.05.05

Mask R-CNN for Geospatial Object Detection

Dalal AL-Alimi
Faculty of Engineering, Sana'a University, Sana'a, Yemen;
School of Computer Science, China University of Geosciences, Wuhan, China
E-mail: dalalm.ali@hotmail.com

Yuxiang Shao
School of Computer Science, China University of Geosciences, Wuhan, China
E-mail: shaoyx@cug.edu.cn

Ahamed Alalimi[1] and Ahmed Abdu[2]
[1]Faculty of Oil and Natural Gas from China University of Geosciences, Wuhan, China;
[2]Faculty of Information Engineering, China University of Geoscience, Wuhan, China
E-mail: [1]ahmedchina65@yahoo.com, [2]ahmedabd39@hotmail.com

Received: 02 December 2019; Accepted: 25 December 2019; Published: 08 October 2020

Abstract: Geospatial imaging technique has opened a door for researchers to implement multiple beneficial applications in many fields, including military investigation, disaster relief, and urban traffic control. As the resolution of geospatial images has increased in recent years, the detection of geospatial objects has attracted a lot of researchers. Mask R-CNN had been designed to identify an object outlines at the pixel level (instance segmentation), and for object detection in natural images. This study describes the Mask R-CNN model and uses it to detect objects in geospatial images. This experiment was prepared an existing dataset to be suitable with object segmentation, and it shows that Mask R-CNN also has the ability to be used in geospatial object detection and it introduces good results to extract the ten classes dataset of Seg-VHR-10.

Index Terms: Mask R-CNN, Faster R-CNN, RoIAlign, object detection, instance segmentation.

1. Introduction

The purpose of object detection (OD) is to provide the position and class of each object in the image if that image includes any target, and the aim of instance segmentation is to identify all pixels belonging to each object of interest. Remote sensing images (RSIs) are supplied from satellites and have been used in many applications and studies, including environmental monitoring, geological hazard identification, precision planting, military investigation, disaster relief, urban traffic control, and many more. OD in geospatial images is used to identify man-made objects, such as structures, ships, cars, airports, and bridges. However, object detection of RSIs, or geospatial images are more complicated and have many obstacles than natural images, which are taken by any camera, for several reasons. 1) Satellites provide us with large and incredibly high-resolution images that take a long time to execute. 2) Remote sensing images involve various objects that are very small, like cars and ships. The size of these objects is usually smaller than the size of objects in natural images. So, most method processing of natural images is not suitable for geospatial images. 3) Most objects found in dense groups like storage tanks, ships that make the manual annotation costly and there are lack of data acquisition. 4) Moreover, there are several other difficulties that make detection more complex in geospatial images, such as occlusion, background noise, illumination, and shadow, as evident in Fig. 1.

The success of image processing and the Region-based CNNs (RCNN) is due to the convolutional neural network (CNN) methods [1]. In several studies, these methods have been widely used. Furthermore, deep CNN is more effective than conventional methods which rely on the manual feature extraction [2, 3, 4]. After the effective success of the Spatial Pyramid Pooling Network (SPP-Net)[5, 4] in reducing the number of feature maps (FM), which first generates the FM of the entire input image, only once, then generates regions from those FMs. This operation speeds performance time up. Fast R-CNN had been proposed to improve and accelerate R-CNN and SPP-Net [6, 4, 2]. Many other methods are used with geospatial images, such as Faster R-CNN[7], Single Shot MultiBox Detector (SSD) [8, 2, 9, 10]. Deep CNN is used to design and optimize the feature extraction network like Visual Geometry Group (VGG) [11], Network of Residual (Res-Net), Feature Pyramid Network (FP-Net), network of squeeze and excitation (SE-Net) [7, 4, 3, 12, 2]. The early layers detect low-level features (edges and corners), and in the later layers, the higher-level features

are detected successfully, almost the whole object features (animal, plane, vehicle). Many methods use the performance of the last top layer of the extraction network of features. These kinds of methods are useful for the recognition of categories but are not helpful for localization. Moving deeper into the extraction of features across several layers of CNN, the resolution is reduced further and the output for small objects is much worse. By combining the performance of the upper layers with bottom layers of FP-Net, FP-Net structure leads to enhance the semantic and resolution value in the overall FMs. Using FP-Net with Faster R-CNN, the COCO dataset achieved state of the art [12].

Fig.1. Some images from NWPU VHR-10 dataset, showing the complexity of RSIs. Objects mostly appear in dense groups and others in very complex situations.

The components of Mask R-CNN was designed and chosen carefully to increase accuracy and reduce loss. In the part of feature extraction, Feature Pyramid Network (FP-Net) was used to obtain as much as possible of object features [13]. To optimize the accuracy of instance segmentation mask and to avoid losing small objects because of the successive pooling operation to minimize the size of the generated FMs, Mask R-CNN was applied RoIAlign in the second part. In addition, Fully Conventional Neural Networks (FC-Net) [14] were used to get classification and localization in the Fast R-CNN part. FC-Nets were also added to the branch of mask to generate the mask for each input region. Mask R-CNN aims to obtain the mask of each target separately, to identify object outlines at the pixel level. This study is the first study that has introduced different class segmentation in geospatial images. As known, OD in geospatial images has a scarcity of data more than in natural images, and in segmentation study, it is more. So, this study prepared an existing dataset (NWPU VHR-10) with ten classes to be suitable for object segmentation in the field study of geospatial images and called it Seg-VHR-10.

The structure of this study as follows: the next section is to review the work related to the detection and segmentation of objects. Section three provides details of this study framework, and information of this study experiment can be found in section four. The results and conclusions are section five and six.

2. Related Work

Determine the object's position in the provided image and its class in the provided image is The object detection objective. the pipeline could be divided into three stages in the traditional OD models: selection of informative regions, extraction of features and classification. A multiscale sliding window is used in the first stage to search the entire input image and figure out the position of all potential objects, but this way is costly and generates too many excrescent windows. During the second level, other methods can be used to extract the features, such as histograms of directed gradients (HOG) [15], Haar-like [16]. Due to the fact that input images usually have a lot of noise, it is difficult to manually construct the extraction function to identify various types of objects. The final stage is used to identify detected objects, and there are several methods to do this task. For example, support vector machine (SVM) [17], Deformable Part-based Model (DPM) [18], and AdaBoost algorithm [19].

RSIs are more complex than natural images, and they have been studied extensively for years. A number of handcrafted features were proposed in [20, 21, 22, 23], based on OD methods. In [22] each octave with five levels in the 15-level HOG feature pyramid was used to extract the feature, then SVM was used to train and detect the geospatial multi-class object. And in [24] a weakly supervised CNN-based aircraft detection combines a Network of Candidate Region Proposals (CRP-Net) and Localization Network (LOC-Net) to extract the proposals and locate the final location of the object [25]. Junwei et al.[20] suggested defining multi-class geospatial objects associated with visual saliency modeling and discrimina-tive learning with sparse coding.

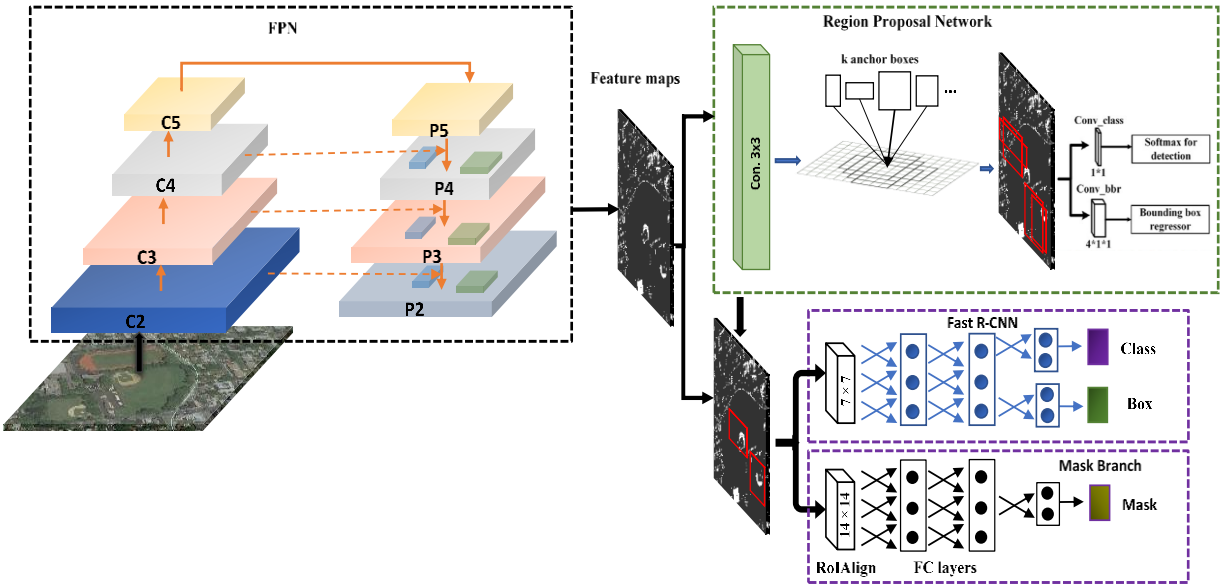

Fig.2. Mask R-CNN architecture which part one is feature extraction to get feature maps, part two is region proposal network and Faster R-CNN to obtain each target and its category.

With the advancement of deep learning and OD studies in geospatial images, [23, 24, 26, 37] used visual saliency to produce a small number of bounding boxes (BBXs), and then got features using Deep Belief Networks. These kinds of methods suit easy environments. In [23] Gong et al. suggested an efficient way of detecting objects by introducing a rotation-invariant layer on CNN to detect multiple-class objects. To improve accuracy and time of implementation, algorithms like R-CNN [1], Single-Shot Multibox Detector (SSD) [27], and You Only Look Once (YOLO) [28] have been developed to improve accuracy and time for implementation.

R-CNN uses Selective Search [29] to control quantities of region proposals (RPs) of the input image. After CNN extracts from each proposal a fixed-length function vector, it classifies them by SVM. For the sake of boosting the accuracy and speed of RPs generating in R-CNN, Fast R-CNN [6] replaced the SVM classifier with a pooling layer of the Region of Interest (RoI pooling) and supplied it with fully connected layers for classification and localization purposes. Faster R-CNN[7] has enhanced this stream by merging the Region Proposal Network (RP-Net) with Fast R-CNN to form a single network; sharing its convolutional features. In the first stage, Faster R-CNN uses VGG16 model to extract the FMs of each input image. These FMs are then inserted into RP-Net to predict RPs. Then the RoI pooling reshapes those expected RPs to a fixed scale. Eventually, the RPs were classified in the second stage, and the offset values of each BBX were predicted. However, SSD [27] and YOLO [28] are faster than two-stage systems, like RCNN, Fast R-CNN, and Faster R-CNN. YOLO is a region-free process, dividing the input image into S ×S grid, with each grid taking *m* BBXs with a confidence score. Although YOLO is faster than the previous methods (45 frames per second), it is not good for the small object. SSD and YOLOv2 [30] enhanced the YOLO detection methods by deleting fully connected layers and predicting BBXs using anchor boxes. Furthermore, the SSD network merged predictions from several FMs with different resolutions to manage objects of various sizes, but the detection of small objects is still not accurate.

Min et al. and Anurag et al. introduced another solution to get the Instance Segmentation by cutting the pixels of Semantic Segmentation of the same category [31, 32]. However, Mask R-CNN has a simple and effective way of generating each object's mask.

3. Framework

This part explains the components of the Mask R-CNN in each stage which based on Faster R-CNN.

3.1. Feature Pyramid Network

After Res-Net has given the solution to the problem of the vanishing gradient [33], Fig. 3, most deep feature extractions such as the Shallow-Deep Feature Extraction Network (SDFE) [2], FP-Net, have been optimized. The Res-Net Building Block equation is defined as:

$$Y = F(X, \{W_i\}) + X \tag{1}$$

Where Y is the output,
X is the input,
W_i is the parameters of the ith convolutional layers to be learned,
$F(X,\{W_i\})$ is the residual mapping, which is already learned.

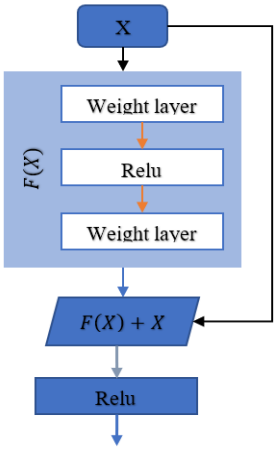

Fig.3. Structure of the residual block of Res-Net, this structure uses to improve the accuracy of feature extraction in deep networks to avert the vanishing gradient issue.

Table 1. The number of each object for each class in the Seg-VHR-10 dataset; objects with polygon points.

Class	Number	Class	Number
A	761	BC	184
S	354	GTF	169
ST	813	H	274
BD	418	B	304
TC	641	V	679
Total	4597		

Res-Net-50 has been used as an FP-Net bottom-up network in this experiment. Because the framework deals with many different scales and resolution images, it needs a suitable feature extraction that can do the balance and extracts more features. FP-Net works to extract FMs that represent the input image at different scales and features.

Simply, going deeper in multi-scale feature extraction to get out multi-scale feature maps, bring you to get more semantic information which in many cases is good for classification method but, it is not effective with detection (localization). The lower levels of FMs have strong resolution values that are very important to locate the location of objects. Therefore, FP-Net added two pathways to reinforcement FMs with resolution values (adding new Top-down pathway layers then connecting these layers by lateral connection with the previous bottom-up pathway layers). To generate feature maps {P5, P4, P3, P2}, the input image passes through the FP-Net layers. These FMs are then fed into the RP-NET to extract RPs, and into the second stage (Fast R-CNN + Mask branch) to get the final prediction.

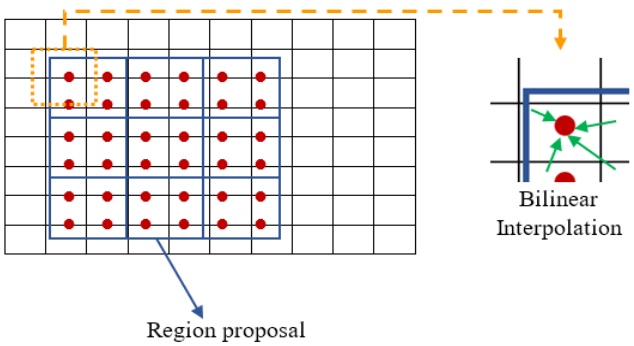

Fig.4. Shows the operation of Bilinear Interpolation and how to obtain values of the pixels in each region box.

3.2. Region Proposal Network

When the FM inserts through the RP-NET, first, RP-NET generates anchors over each FM. Number and size of ratios and scales control the number and shape of anchors in each point over each FM. Three ratios (0.5, 1, 2) and five scales (32, 64, 128, 256, 512) were used to generate the various size of anchors. Then, each anchor is categorized by Intersection-over-Union (IoU), in an attempt to pick the best precise box. If the value of IoU ≥ 0.7, that anchor is labeled as a positive label (1) which also means there is an object inside the anchor and matches well the ground-truth BBX; otherwise is labeled as a negative label (0). Hence, the output of RP-NET is two things: BBXs (RPs) and their class (1 or 0). The equation of IoU as the following:

$$IoU = \frac{area(B_{pb} \cap B_{gt})}{area(B_{pb} \cup B_{gt})} \tag{2}$$

Where $area(B_{pb} \cap B_{gt})$ is the area between the predicted BBX and the real BBX, and $area(B_{pb} \cup B_{gt})$ is their union.

In order to measure the loss of each detection layer at this stage, the classification loss and BBX regression values should be combined [6], from the above equation as follows:

$$L(X, Y, B_{gt}, B_{rb}) = L_{cts}(p(X), Y) + \lambda[Y \geq 1] L_{bbr}(B_{gt}, B_{rb}) \tag{3}$$

Where $L_{cts}(p(X), Y) = -\log p_y(X)$ is the cross-entropy loss,

X is the predicted probability anchor,

$\lambda = 1$,

$[Y \geq 1]$ is 1 when Y ≥ 1 and 0 otherwise,

B_{rb} is the regression area of BBX.

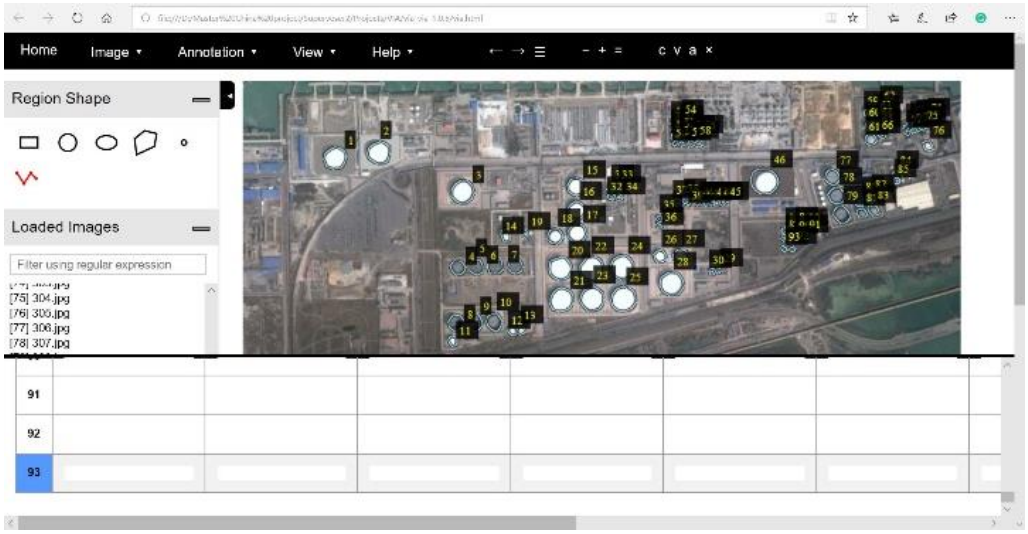

Fig.5. One example showing the operation of annotating each target in VGG image annotator application.

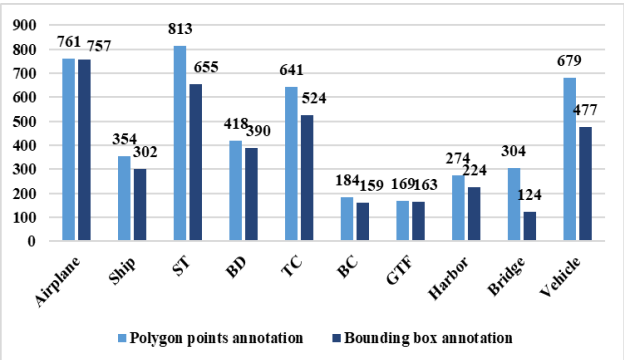

Fig.6. The number of different annotation ways. Light blue is the objects annotated by polygon points and the other by bounding boxes [23], for each class in both dataset.

The loss of regression of the bounding box will be measured as:

$$L_{bbr}(B_{gt}, B_{rb}) = Smooth_{L1}(xB_{gt} - B_{rb}) \tag{4}$$

In which

$$Smooth_{L1}(x) = \begin{cases} 0.5x^2, & |x| < 1 \\ |x| - 0.5, & otherwise \end{cases} \tag{5}$$

3.3. Fast R-CNN and Mask Branch

The detection stage of Mask R-CNN contains two parts: Fast R-CNN and mask branch. Mask R-CNN is replacing RoI-pooling of Faster R-CNN with RoIAlign. Because RoI pooling failed in case of instance segmentation, it effects on mask generation (pixel to pixel correspondence matters) and also causes the misalig-nment. To solve the above disadvantages of RoI pooling in this field, RoIAlign was used. Because region proposal boxes of RP-NET have various sizes, RoIAlign uses the "Bilinear Interpolation" method to obtain the image values on the pixels with the coordinates of floating-point values and to produce a fixed dimension for each region box. The size of objects reduces smoothly without losing during resize generated FMs. Simply researchers need to use Bilinear Interpolation in two situations. The first, when it is necessary to resample the data from one cell size to another (change the cell size). The second, if anybody needs to show up the data in another coordinate system (resampling data). Bilinear Interpolation determines the output value from the four nearest centers of input values. The new cell, which created from this output, generates a smoother-looking from that weighted average of those four nearest values, Fig. 4.

The output of RoIAlign was fed into FC-Net for classification and localization operations, and was also fed into the mask branch at the same time to produce the mask for each input region. In the present study, the dimension of RoIAlign is 14 × 14 in the mask branch and 7×7 in Fast R-CNN stage, Fig. 2.

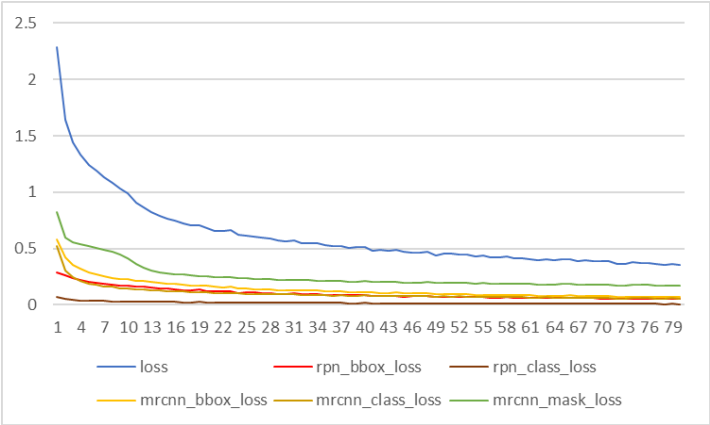

Fig.7. Loss values of each task in Mask R-CNN. Showing the RoI loss values in RP-Net, Fast R-CNN, and Mask part. Mask branch has the highest loss and classification loss of RP-Net has lower values.

The difference in the RP-NET loss function is that the RP-NET classification layer calculates only two classes, the positive and negative, and Fast R-CNN (second stage) classification layer deals with all classes of objects [37, 26, 10, 6].

Mask R-CNN not only detects the position of each object but it also provides an individual outline of the object (without its background). Mask R-CNN uses multi-task loss defined on each sampled RoI:

$$L = L_{cls} + L_{box} + L_{mask} \tag{6}$$

Classification loss (L_{cls}) and BBX loss (L_{box}) both of them are defined in [6], mask loss (L_{mask}) is known as the mean binary cross-entropy loss [13]. Integrating those three losses improves performance and speeds up[6].

4. Experiments

The implementation and evaluation were on Core i7-4790 CPU with 8 GB RAM.

4.1. Dataset

Gong Cheng and his team, from Northwestern Polytechnical University (NWPU), have collected and provided the NWPU VHR-10 dataset with 10 classes for research purposes only in OD area [37, 23, 22, 2]. All the images of the

NWPU VHR-10 dataset are very-high-resolution (VHR) Geospatial images. It contains 800 images collected and cropped from two different satellites with different resolutions, Google Earth and Vaihingen data. Then it was manually annotated by experts to create the four coordinates of ground truth BBXs (x1, y1), (x2, y2) of desired classes, for 650 images, and the rest of them used for another proposes. The desired classes are airplane (A), ship (S), storage tank (ST), baseball diamond (BD), tennis court (TC), basketball court (BC), ground track field (GTF), harbor (H), bridge (B), and vehicle (V). As shown in Fig. 6, the dark blue color represents the number of objects annotated by BBXs.

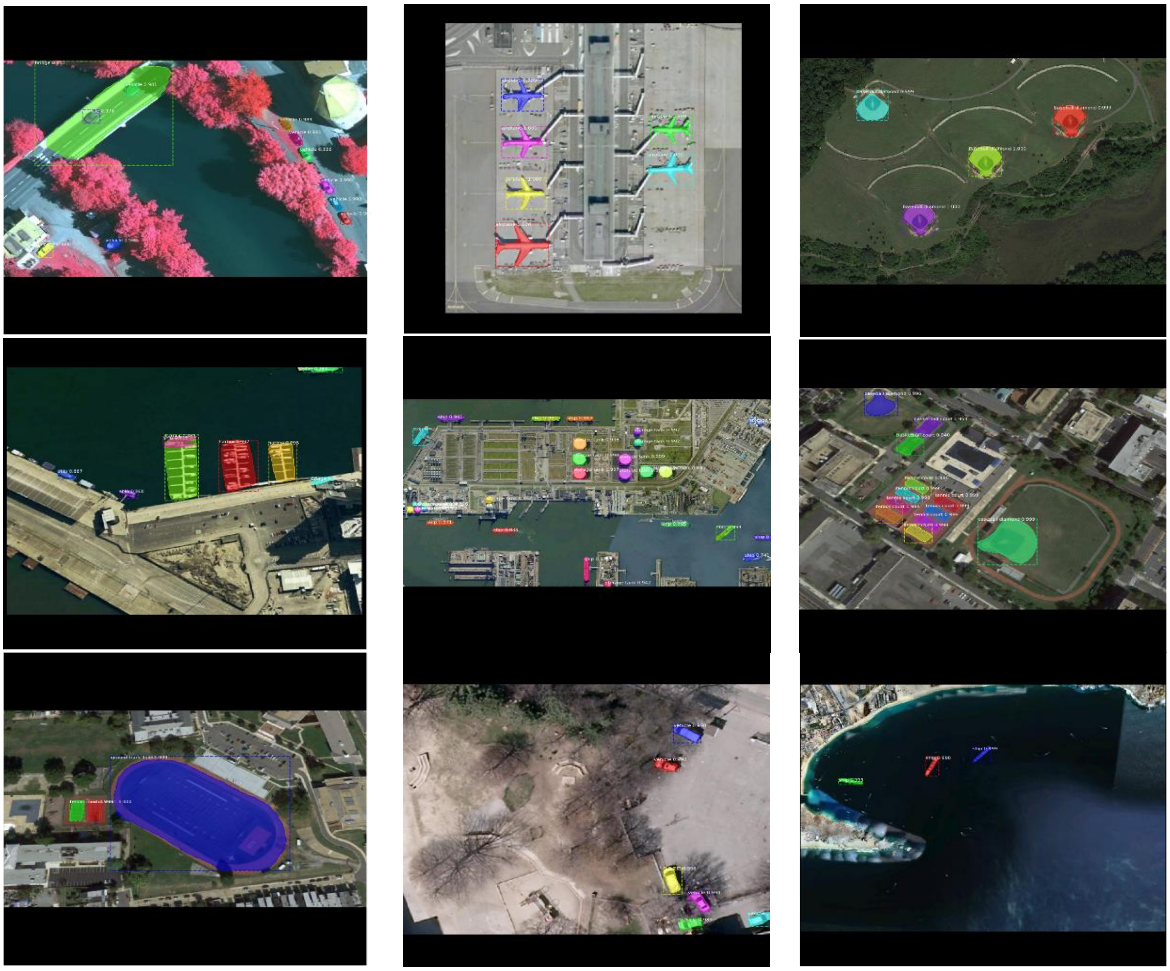

Fig.8. The detection results of the Mask R-CNN model for 10 classes of Seg-VHR-10 dataset.

In this study, the NWPU VHR-10 dataset was re-prepared to be suitable for the segmentation research area. So, this experiment used the VGG Image Annotator (VIA) application. VIA is an open-source application or website based on HTML, it opens in any web browser and does not need any setup [34, 35]. 650 images manually annotated which have targets of the same 10 classes. Drawing the mask of polygon points-set for 4597 objects in 650 images. The VIA program automatically saves all annotation points in one JSON file. Fig. 5 Displays one example of the VIA manual annotation. Fig. 6 shows the contrast between the quantities of the rectangular BBXs manually annotated and the polygon point-set mask. As seen, this experiment has annotated more objects (822 objects). In this experiment, the dataset was split to 80% for the training part and 20% for the validation part. This new dataset called Seg-VHR-10 dataset; "Seg" is the short name of Segmentation.

4.2. Implementation Details

In the implementation time, each input image was resized to 800 × 1024, also used zero-padding to modify the size of the input image. The zero-padding helps to control the size of any input and give uniform size. The mini-batch includes 2 images per GPU, each image has 200 RoIs, the ratio of the positive number of proposals to the negative is 1:3. This study trained the dataset on 1 GPU; the iteration was 325 and the epoch was 80. Weight decay of 1e-4, momentum and the learning rate were 9e-1 and 1e-3, respectively. As can be seen in Fig. 2, RPs are fed into two networks, one to generate class detection and the BBX regressor of each detected object, and the other one is to get the detected mask of the same object. The part of Fast R-CNN uses 7 × 7 dimension RoIAlign and the mask branch uses 14 × 14 dimension RoIAlign to get the unified size of each part.

5. Results

Fig. 7 displays the Mask R-CNN loss values of training operation in each RoI operation. The loss value of rpn_bbox was calculated in the region proposal network. Fig. 8 showing the accuracy of generated coordinates of proposal boxes, which also resized by RoIAlign. The rpn_class is the result of the classified operations of each proposal that was also created in the RP-Net. The mrcnn_bbox and mrcnn_class are in the Fast R-CNN stage which is the loss values of getting the BBX regression of each detected object and its classification. Finally, in the mask branch, the mrcnn_mask loss comes from the mask generation of the same detected object. From Fig. 7 it can be seen that the loss of mrcnn_mask has the highest value. In general, the loss value of Mask R-CNN is 0.35.

Fig. 8 shows some results of Mask R-CNN of the Seg-VHR-10 dataset. It detected and drew successfully the mask of each object, the model is very effective to obtain individually pixels of the objects of remote sensing images, and to detect each object separately. This means that Mask R-CNN can beat many complicates of geospatial OD, and the new dataset successfully generated. In this experiment, a very small memory 8 BG device was used, which affected badly on training time and limited the test operation.

6. Conclusions

Object detection in geospatial images is widely used for many purposes. This paper presented Mask R- CNN to train the new dataset. It clarified how prepared the existing dataset (NWPU VHR-10) to detect object segmentation by using the VIA application and called it Seg-VHR-10. Mask R-CNN combines between OD and semantic segmentation. It is the effective method to detect and extract the outlines of each object individually. We used Mask R-CNN to extract the ten classes of Seg-VHR-10 dataset and we got good results to extract each object effectively. In the future study, we will work to redesign the feature extraction to improve the localization, which is strong in the shallow CNN, and add more different scales over different feature maps to optimize the extraction of various small objects.

Acknowledgment

We would like to express our appreciation to all those who have supported us during our research and study in the School of Computer Science at China University of Geoscience (Wuhan) and in the Faculty of Engineering at Sana'a University.

References

[1] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," IEEE Conference on Computer Vision and Pattern Recognition, pp. 580-587, 2014.
[2] D. AL-Alimi , Y. Shao, F. Ruyi , M. A. A. Al-qaness, M. Abd Elaziz and S. Kim, "Multi-Scale Geospatial Object Detection Based on Shallow-Deep Feature Extraction," Remote Sensing, vol. 11, no. 21, pp. 1-19, 29 10 2019.
[3] Z.-Q. Zhao, P. Zheng, S.-T. Xu and X. Wu, "Object Detection With Deep Learning: A Review," IEEE Transactions on Neural Networks and Learning Systems, pp. 1-21, 2019.
[4] Z. Zou, Z. Shi, Y. Guo and J. Ye, "Object Detection in 20 Years: A Survey," arXiv:1905.05055v1, pp. 1-40, 2019.
[5] K. He, X. Zhang, S. Ren and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904-1916, 2015.
[6] R. Girshick, "Fast R-CNN," 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1440-1448, 2015.
[7] K. H. R. G. a. J. S. Shaoqing Ren, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 2017.
[8] D. A. Wei Liu, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu and A. C. Berg, "SSD: Single Shot MultiBox Detector," Proc. European Conf. Computer Vision, 2016,, p. 21–37, 2016.
[9] K. F. H. S. J. Y. Z. G. Xue Yang, "R2CNN++: Multi-Dimensional Attention Based Rotation Invariant Detector with Robust Anchor Strategy," arXiv, pp. 1-10, 2018.
[10] W. Guo, W. Yang, H. Zhang and G. Hua, "Geospatial Object Detection in High Resolution Satellite Images Based on Multi-Scale Convolutional Neural Network," Remote Sens., vol. 10, no. 1, pp. 1-21, 2018.
[11] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv, 2014.
[12] T.-Y. Lin, P. Doll ár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature Pyramid Networks for Object Detection," arxiv, pp. 1-9, 2017.
[13] K. He, G. Gkioxari, P. Doll ár and R. Girshick, "Mask R-CNN," arXiv, pp. 1-12, 2018.
[14] J. Long, S. Evan and T. Darrell , "Fully Convolutional Networksfor Semantic Segmentation," CVPR, 2015.
[15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, pp. 886-893, 2005.
[16] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 511-518, 2001.
[17] C. Cortes and V. Vapnik, "Support vector machine," Machine learning, vol. 20, no. 3, p. 273–297, 1995.
[18] F. F. Pedro, B. G. Ross, M. David and R. Deva, "Object Detection with Discriminatively Trained Part-Based Models," IEEE

Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, 2010.
[19] R. E. S. Yoav Freund, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119-139, 1997.
[20] J. Han, P. Zhou, D. Zhang, G. Cheng, L. Guo, Z. Liu, S. Bu and J. Wu, "Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 89, pp. 37-48, 2014.
[21] J. Han, D. Zhang, G. Cheng, L. Guo and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," IEEE Transactions on Geoscience and Remote Sensing, vol. 53, no. 6, pp. 3325-3337, 2015.
[22] 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, pp. 119-132, 2014.
[23] G. Cheng, P. Zhou and J. Han, "Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images," IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 12, pp. 7405-7415, 2016.
[24] F. Zhang, B. Du, L. Zhang and M. Xu, "Weakly Supervised Learning Based on Coupled Convolutional Neural Networks for Aircraft Detection," IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 9, pp. 5553-5563, 2016.
[25] F. Zhang, B. Du, L. Zhang and M. Xu, "Weakly Supervised Learning Based on Coupled Convolutional Neural Networks for Aircraft Detection," IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 9, 2016.
[26] Z. Deng, H. Sun, S. Zhou, J. Zhao, L. Lei and H. Zou, "Multi-scale object detection in remote sensing imagery with convolutional neural networks," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 145, pp. 3-22, 2018.
[27] D. A. Wei Liu, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu and A. C. Berg, "SSD:SingleShotMultiBoxDetector," Proc. European Conf. Computer Vision, 2016,, p. 21–37, 2016.
[28] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779-788, 2016.
[29] J. Uijlings, K. v. d. Sande, T. Gevers and A. Smeulders, "Selective Search for Object Recognition," International Journal of Computer Vision, vol. 104, pp. 154-171, 2013.
[30] J. Redmon and A. Farhadi, "YOLO9000: Better,Faster,Stronger," arXiv, pp. 1-9, 2016.
[31] M. Bai and R. Urtasun , "Deep Watershed Transform for Instance Segmentation," CVPR, vol. 3, 2017.
[32] A. Arnab and P. H. S. Torr, "Pixelwise Instance Segmentation with a Dynamically Instantiated Network," CVPR, vol. 3, 2017.
[33] K. Z. X. R. S. He, "Deep residual learning for image recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p. 770–778, 2016.
[34] A. Dutta and A. Zisserman, "The VIA Annotation Software for Images, Audio and Video," in Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 2019.
[35] R. Anantharaman, M. Velazquez and Y. Lee, "Utilizing Mask R-CNN for Detection and Segmentation of Oral Diseases," in 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, 2018.
[36] J. H. Gong Cheng, "A survey on object detection in optical remote sensing images," ISPRS Journal of Photogrammetry and Remote Sensing, no. 177, pp. 11-28, 2016.
[37] D. AL-Alimi , Y. Shao, F. Ruyi , M. A. A. Al-qaness, M. Abd Elaziz and S. Kim, "Multi-Scale Geospatial Object Detection Based on Shallow-Deep Feature Extraction," Remote Sensing, vol. 11, no. 21, pp. 1-19, 2019.

Authors' Profiles

Dalal AL-Alimi has received her Master of Computer Science and Technology from School of Computer Science and Technology at China University of Geosciences (CUG), Wuhan, China. She received her B.Sc. Electrical Computer and IT from Engineering Faculty at Sana'a University, Sana'a, Yemen. Her research interests include Object Detection, Object Classification, Machine Learning, Deep Learning, AI, Image Processing and Analysis, Time Series Methods and IoT.

Yuxiang Shao has received the Ph.D. degree in Mineral prospecting and Exploration from CUG. He is currently an associate professor in School of Computer Science and Technology at CUG, Wuhan, China. His current research interests include Data Mining, Data Warehouse, High-Performance Computing and Web Information Exploration for Train Transportation Fields.

Ahmed Alalimi has gotten his Bachelor's and Master of Oil And Gas Engineering from the Faculty of Oil And Natural Gas from CUG, Wuhan, China. Currently his Ph.D. in Reservoir Engineering at CUG. His research interesting in Oil Production Forecasting and Reservoir Modeling.

Ahmed Abdu has received his Master of Information Engineering from Software Engineering at China University of Geoscience (Wuhan). Brain-Inspired Navigation, Mobile Robotics Navigation, Intelligent Systems Analysis, Machine Learning, and Image Processing are his research interests.

How to cite this paper: Dalal AL-Alimi, Yuxiang Shao, Ahamed Alalimi, Ahmed Abdu, "Mask R-CNN for Geospatial Object Detection", International Journal of Information Technology and Computer Science(IJITCS), Vol.12, No.5, pp.63-72, 2020. DOI: 10.5815/ijitcs.2020.05.05

