Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 9, August 2021:7366 – 7376

Research Article

An Effective Multi-class Object Detection Model for Remotely Sensed Image using Mask R- DCNN

P.Deepan[1], Dr. L.R. Sudha[2], Dr. T. Poongothai[3], Dr. Rajalingam[4], Dr. R.Santhoshkumar[5]

Abstract

Object detection in remote sensing image has received increasing attention from the research community in recent days. Over the past few decades, variety of deep learning based detection model such as Region based Convolutional Neural Network (R-CNN), Fast R-CNN and Faster R-CNN has been applied for object detection. However, most of the existing detection methods localize each object using the bounding box, but cannot segment the object from the background. So in order to tackle the issue, we introduce the Mask R- Dilated CNN model, which incorporates both object detection and segmentation. In Mask R-DCNN, ResNet-50 and ResNet-101 act as backbone for feature extraction, Region Proposal Network (RPN) is utilized to generate RoIs and RoIAlign is to carefully hold the exact spatial location to generate mask through Fully Convolution Network (FCN). The aim of Mask R- DCNN model is to incorporate more relevant information by increasing the receptive field of convolutional layer for improving the robustness. Experimental results on the NWPU VHR 10-class benchmark dataset demonstrated the effectiveness of the proposed model by providing 95.7% accuracy for Dilated ResNet-50 & 96.2% accuracy for Dilated ResNet-101, which is better than traditional Mask R-CNN model.

Keywords: Object detection, Region Proposal Network, Deep learning, Mask R-DCNN, Remote sensing image, ResNet-50 and ResNet-101.

1. Introduction

With the rapid development of remote sensing technologies in the field of remote sensing many satellite sensor provides high resolution satellite images. These images are mainly used to detect the various objects such as airplane, building, ship and vehicles in the field of military civilians, intelligent monitoring, agricultural monitoring, disaster management, geographical information system (GIS) updating, urban planning, etc.,[1]. The general term ‘object’ consists of man-made (eg. building, ships and vehicles) and landscape objects (land use, land cover). Man-made objects have sharp boundary and independent on background environment, but landscape object have blur

[1]Assistant Professor, Department of CSE, St. Martin's Engineering College, Secunderabad, Telangana, India
[2] Associate Professor, Departemnt of CSE, Annamalai University, Chidambaram, Tamilnadu, India
[3]Professor, Department of CSE, St. Martin's Engineering College, Secunderabad, Telangana, India
[4]Associate Professor, Department of CSE, St. Martin's Engineering College, Secunderabad, Telangana, India
[5]Associate Professor, Department of CSE, St. Martin's Engineering College, Secunderabad, Telangana, India
[1]deepanp87@gmail.com,[2]sudhaselvin@ymail.com, [3]poongothait@gmail.com, [4]rajalingam35@gmail.com, [5]santhoshkumar.aucse@gmail.com

boundary and dependent on background environment [2]. Object detection in remote sensing images or geospatial images are more complicated and have big challenges than natural images, because large variation in the visual appearance of objects [3]. These phenomena may be affected by resolution of image, viewpoint variation, background clutter, occlusion, variation in illumination intensities and shadow effects.

Figure 1. Different visual appearance in remote sensing image

In addition, sizes of the remote sensing images (airplane, ship and vehicle) are too small that may difficult to detect the accurate object. An image from NWPU VHR 10-class dataset shown in Figure 1 depicts challenges like illumination, shadow effect and small objects with background clutter [4].

The aim of object detection is to recognize and localize object based on pre-processing, feature extraction, classification and localization steps. Many researches have been developed huge number of object detection algorithm and it is mainly categorized into two namely, Machine learning based detection and deep learning based detection. The early research on remote sensing image focused on the extraction of handcrafted or shallow or human engineering features, such as color[5], shape, texture, spatial and spectral information. The histogram of gradients (HOG), color histogram (CH), gray level co-occurrence matrix (GLCM), local binary pattern (LBP), scale in-variant feature transform (SIFT) are some of the familiar handcraft feature extraction methods used for image scene classification and object detection.

These low level features are producing better results, but they require domain expertise and consume more time. In addition, handcrafted features require an artificial dilation for extracting the features [6]. To overcome the limitation of handcraft features, automatically learning the features from images are considered as best way. In recent years, deep learning method has great success in the field of object detection and scene classification. It is composed of hierarchical layers that can automatically learn more multilevel features from the data. In addition, the deep layers of representations have great potential to characterise robust features with complex patterns and semantics, such as land use, land cover, functional sites etc [7]. Currently, many deep learning based object detection models are available such as Region based Convolutional Neural Network (R-CNN), Fast R-CNN, Faster R-CNN, Mask R-CNN, You Only Look Once (YOLO) and Single Shot Multi-box Detector (SSD).

This paper has been organised as follows: Section 2 covers literature research work carried out for various object detection models. Section 3 includes the architecture of proposed Mask R-DCNN model. Section 4 includes the experimental analysis of proposed approach on the NWPU VHR - 10 class benchmark dataset and evaluation metrics. Finally conclusions are given in Section 6.

2. Related Works

Recently, most of the object detection algorithms are based on deep learning techniques. These algorithms are broadly classified into two streams namely, one-stage detectors and two-stage detectors. The YOLO and SSD is most familiar one-stage detector, which can perform object detection without region proposal. W. Liu et al. developed [8] a Single-Shot Multi-box Detector (SSD) to predict all at once the bounding boxes and the class probabilities with an end-to-end CNN architecture. The model takes input image which passes through multiple convolutional layers with various filter size. Redmon et al.[9] developed aircraft detection based on YOLO which is region free model, dividing the input image into S×S grid, with each grid taking four bounding box with a confidence score. However, region free method extremely fast but still there are some lacking in detection of object in efficient manner.

The other object detection method is two-stage model that generates proposals and then makes predictions for these proposals, such as Region with Convolutional Neural Networks (R-CNN) [10], Fast R-CNN and Faster R-CNN. The Region-based CNN[11] method initially select the Region of Interest (RoIs) by using Selective Search (SS) [12], Region Proposal Networks (RPN), or edge box. Then, feature extraction is carried out for every RoI by CNNs. Finally, the detection of an object with a bounding box is obtained using the soft max classifiers. Due to the fact that there is a huge amount of overlaps between these RoIs, inefficient result can occur in redundant calculations. In order to handle this R-CNN issue, Girshick et al. [13] proposed a Fast R-CNN model. The model is a shared feature method that can extract features only one time for the entire image instead of using a CNN for each region. RoIs are detected with SS methods applied on the produced feature maps. The feature map size is reduced by using RoI pooling layer and each RoI pooling layer feeds the feature into a fully connected layer. The final vector is used to predict the object detection using softmax classifiers.

The main intention of Fast R-CNN is to reduce the time consumption of the high number of models necessary to analyze all region proposals. The region proposals detected with SS methods are computationally expensive. So in order to tackle the problem, Ren et al. [14] introduced RPN model. These models directly generate region proposals and predict the bounding box as well as detect the objects. The Faster R-CNN model [15] was proposed by combining the RPN and Fast R-CNN model. The model takes an input image and produces feature maps. All the feature maps and feature vectors are linked to two fully connected layers, one for box regression and another one for box classification.

The aim of this model was to detect the object and draw or locate the bounding box over the object, which may cause some difficulties due to background clutter and shadow effect during the feature extraction. Instance segmentation is a challenging task because it requires the correct prediction of objects in an image and also accurate segmentation of each instance with the help of masking. Taking the above disadvantages into consideration, our contribution is to propose a Mask Region based Dilated Convolutional Neural Network (Mask R-DCNN) for object detection of RSI images by using instance segmentation and object detection. The model efficiently increases the receptive fields of filer without loss of information in convolutional layer and also allows multi-class target detection in complex background situations [16]. The paper uses four backbone networks (ResNet-50, ResNet-101, Dilated ResNet-50 and Dilated ResNet-101) as feature extraction model.

3. Proposed Works

Mask R-CNN [17] is one of the familiar neural network architecture for pixel based instance segmentation which was introduced by He et al., in 2017. This section describes the Mask R-CNN

model for remote sensing image object detection using deep learning networks. The model consists of two parts namely, backbone network (Feature Extraction) and Region Proposal Network. A different number of proposals were generated in the backbone network regarding the region where an object could be based on the input image. First, we have used the standard deep convolutional neural network architecture for feature extraction. The architecture AlexNet, VGG-16 and Inception with 5,19 and 22 convolutional layers respectively. By getting the deeper, the model suffers from vanishing gradient issue, which may affect saturation of performance accuracy rapidly. In order to solve vanishing gradient issue, we have used ResNet-50 and ResNet-101 model as backbone network for feature extraction. To improve the computational time of ResNet model convolutional layer, we have use the dilated convolutional filter instead of traditional filter. The new dilated convolution filter expands the receptive field without increasing parameters. So, we can improve the performance of this model and also reduce the computational time. The architecture of proposed Mask R-Dilated CNN model was shown in Figure 2.

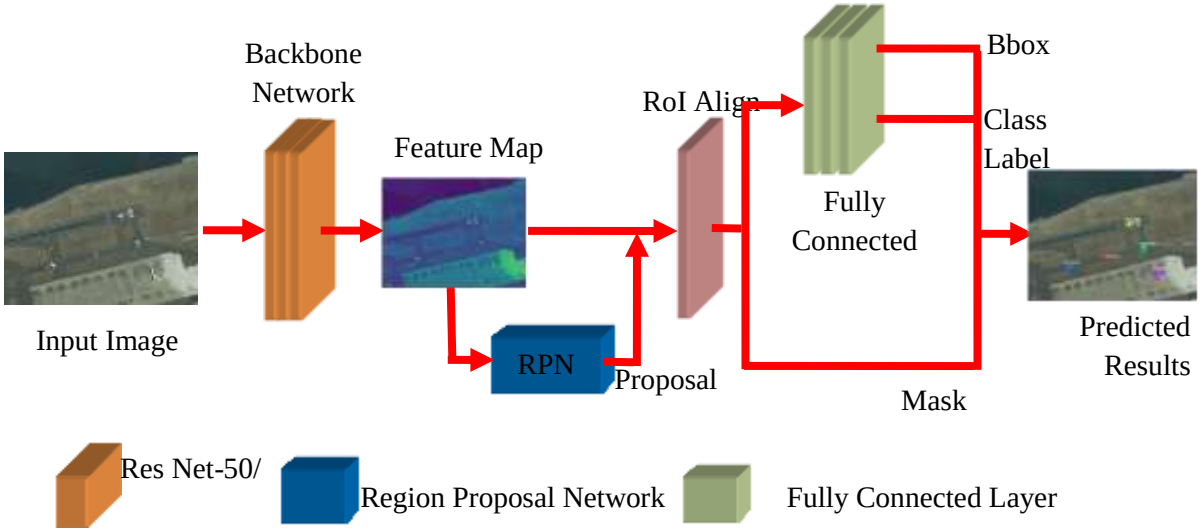


Figure 2. The architecture of Mask R-DCNN object detection

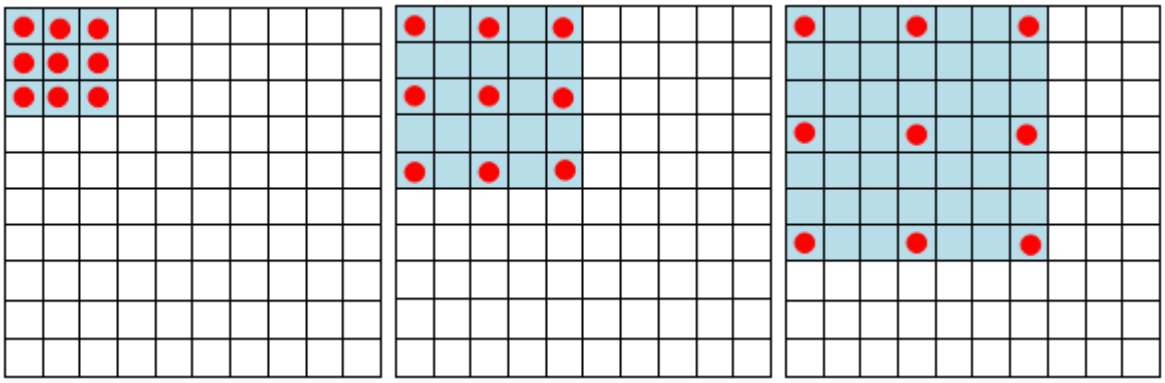
3.1. Dilated Convolutional Filter

The Convolutional layer is most important and essential layer in the CNN, which is used to extract or learn the features from the input images. The general form of convolution is defined as:

$$s(i,j) = \sum_{k=1}^{n_{in}} (X_k \times W_k)(i,j) + b \qquad (1)$$

where n_{in} represents the input matrices of the tensor. X_k is k^{th} input matrix. W_k is the k^{th} sub-convolution kernel matrix of the convolution kernel. s(i,j) represents the output values for matrix of corresponding elements to the kernel w and b represents an bias values for corresponding weight. As shown in Figure 3, the kernel's receptive field is 3×3, 5×5 and 7×7 respectively. The receptive field size is increased by adding the gap between the matrices; however, the number of parameters in all the dilated convolution kernels is same. Therefore, using such a dilated convolutional kernel to process images can get more information from the convolution kernel without increasing the computation. In dilated convolution, a small kernel size w×w is extended to w + (w-1) (dr-1) with dilate rate dr. In traditional convolutional kernel with size of 3×3, the receptive field is 3×3. While performing dilated convolutional kernel with size of 3×3, its receptive field is 5×5 when dilation rate dr = 2, and 7×7 when dr = 3. The receptive field size is generally defined as $[k + (k-1)(i-1)] \times [k + (k-1)(i-1)]$ when dr = i. The Dilated ResNet model extracts

Region of Interests (RoIs) from different levels of features and applied the features as input to the Region Proposal Network.



(a) (b) (c)

Figure 3. Conceptual illustration of traditional and dilated convolution; (a) traditional convolution (b) dilated convolution with dilation rate 2; (c) dilated convolution with dilation rate 3.

3.2. Region Proposal Network

The regions were scanned individually in the Region Proposal Network and predicted whether an object was present or not in the feature map. The RPN never scans the actual input image instead of that scan it the feature map by the RPN network, making it much faster. Next, the region proposals generated from RPN require RoIAlign to adjust their dimension to meet the multi-branch prediction models. The RoIAligned region of interest values are fed into fully connected layers. Fully connected layers classify the output as label with bounding-box and perform masking over the specified location of object detection.

4. Experimental Results and Analysis

In this section, we focus on analysing the performance and effectiveness of RSI object detection based on Mask R-Dilated CNN model and Mask R-CNN model under the same parameters and conditions. First, we introduced the benchmark datasets for RSI object detection, then analyzed the performance of Mask R-CNN for two backbone networks (ResNet-50 and ResNet-101), and finally presented the experimental results for proposed Mask R-DCNN with dilated convolutional filter using the above said backbone networks. The proposed model (Mask R-DCNN) was developed on Python and Anaconda IDE tools.

4.1. Dataset Descriptions

For experimental evaluation, we have used North-Western Polytechnical University Remote 10-class dataset which is a challenging Very High Resolution (VHR) dataset for object detection and it was developed by Gong Cheng et al. The dataset contains 650 aerial images from more than 100 countries extracted by different platforms and sensors. Each image has different resolution with different shapes, scales and orientations. These NWPU 10-class images are annotated by the VGG Image Annotator (VIA Tool). The total annotated images contain 3,775 instances with bounding boxes and instance masks used as ground truth. To demonstrate the effectiveness of our proposed approach, we have used NWPU VHR-10 classes namely bridge(BR), basketball court(BBC),

harbour(HR), ground track field(GTF), ship(SH), baseball diamond(BBD), vehicle(VE), tennis court(TC), storage tank(ST) and airplane(A) and the number of instances corresponding to the images is shown in Table 1. In our proposed work, the NWPU-10 data set was utilized 70% for training, and 30% for testing.

Table 1. The Number of object instances in each class

S. No.	Class	No. of Instance
1.	Bridge	124
2.	Basketball court	159
3.	Harbour	163
4.	Ground track field	224
5.	Ship	302
6.	Baseball diamond	390
7.	Vehicle	477
8.	Tennis court	524
9.	Storage tank	655
10.	Airplane	757

4.2. Evaluation Metrics

We have evaluated the performance of a proposed model by using three standard performance metrics namely Average Precision (AP), Precision-Recall Curve (PRC) and Intersection of Union (IoU). The precision can be measured by number of properly detected object divided by total number of all samples in a class. Precision value of the class c, Pc can be shown in equation (2) where, tc is a total number of properly detected objects in class c and nc is a total number of samples in the class c.

$$P_C = \frac{t_c}{n_c} \quad (2)$$

The recall can be measured by number of properly detected objects divided by the number of all relevant samples in the corresponding class. Recall value of the class, Rc is shown in equation (3) where, tc is total number of properly detected object samples in class c and kc is number of samples detected as relevant to class c.

$$R_C = \frac{t_c}{k_c} \quad (3)$$

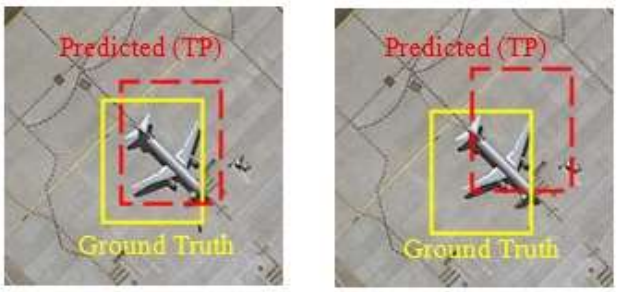


Figure 2. Intersection of Union Calculation Process

The AP computes the average value of the precision over the interval from Pc = 0 to Rc = 1, i.e., the area under the PRC; hence, higher the AP, better the performance. The IoU, which is a standard evaluation metric to measure the accuracy of object detection and the overlap rate of the predicted bounding box(Bp) and ground truth(Bgt) generated by the model. Figure 4, shows the IoU of actual ground truth and predicted object and it is calculated by using the equation 4.

$$IoU = \frac{B_p \cap B_{gt}}{B_p \cup B_{gt}} \geq a0 \quad (4)$$

4.3. Result Analysis

To evaluate the proposed approach Mask R-Dilated CNN model was trained on the NWPU 10-class very high resolution dataset. All the experiments were trained and performed using PC with Intel® i7 processors @ 3.40 GHz speed and 16GB of RAM. We have used four combination of backbone network such as ResNet-50, ResNet-101, Dilated ResNet-50 and Dilated ResNet-101 as the backbone of feature extraction network and trained the network for 100 epochs with learning rate of 0.0001 and Adam optimizer. Similarly, we have evaluated the object detection of models based on the three parameters namely AP, AP50 and AP75. Here, AP represents the average IoU threshold from 0.50 to 0.95. The AP50 represents the IoU threshold for 0.50 and The AP75 represents the IoU threshold for 0.75.

Compared to Mask R-CNN, our proposed model was more efficient and accurately detects the object in remote sensing image. Since the ResNet-50 and ResNet-101 model have more number of convolutional layers, the computational time will also be more. With the help of new dilated convolution filter, the receptive field is expanded. So, we can improve the performance of this model and also reduce the computational time. The experimental result of proposed model was compared with traditional ResNet-50 and ResNet-101 backbone network model and results are showed in Table 2 and Figure 6. From the Table 2, we observed that the proposed Dilated ResNet-101 model most reliable to detect object in remote sensing images. Figure 5, shows the result of some multi-class object detection in remote sensing images based on Mask R-DCNN model.

Table 2. Performance Analysis of Mask R-CNN and Mask R-DCNN

S.No.	Model	Backbone	AP	AP50	AP75
1.	YOLOv1	-	57.1	89.5	56.3
2.	YOLOv2	DarkNet-19	58.2	91.5	59.4
3.	Fast R-CNN	VGG-16 Net	61.4	89.3	61.7
4.	Faster R-CNN	ResNet	62.1	91.2	63.1
5.	Mask R-CNN	ResNet-50	57.7	94.1	67.7
6.	Mask R-CNN	ResNet-101	59.3	92.8	71.3
7.	Mask R-DCNN	Proposed ResNet-50	61.3	95.7	70.1
8.	Mask R-DCNN	Proposed ResNet-101	64.4	96.2	71.5

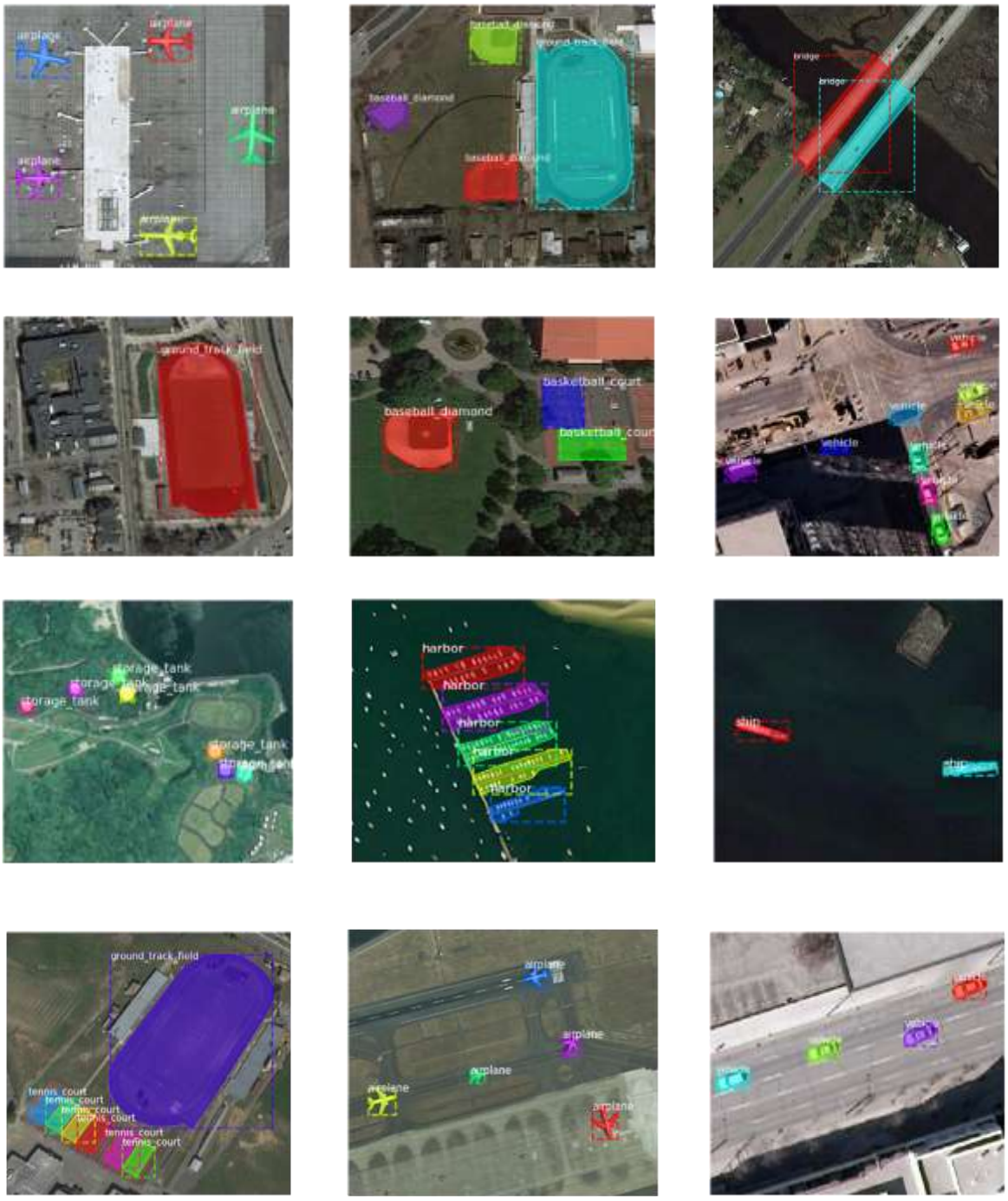


Figure 5. Object detection sample results of Mask R-CNN approach

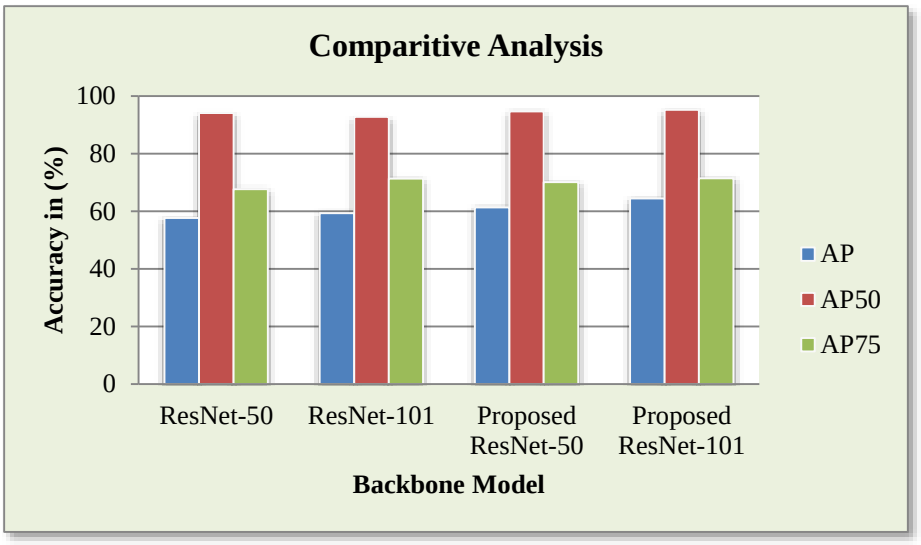


Figure 6. Performance Analysis of Proposed Model with other detection model

5. Conclusion

A Mask R- Dilated CNN model for remote sensing object detection and instance segmentation was proposed in this paper. The approach will extract feature from proposed backbone dilated ResNet-50 and dilated ResNet-101 and generate RoIs by RPN, by maintaining RoIAlign's spatial position to generate binary masks via a full convolutional network (FCN). The proposed model reduces the computational time in convolutional layer operations and also detects the objects in efficient manner. In future, we have planned to implement the model into large scale remote sensing image datasets.

References

1. Zhong-Qiu Zhao, Peng Zheng, Shou-Tao Xu, and Xindong Wu , Object Detection With Deep Learning: A Review, IEEE Transactions on Neural Networks and Learning Systems, pp.1-21, 2019.
2. G. Cheng, X. Xie, J. Han, L. Guo and G. Xia, "Remote Sensing Image Scene Classification Meets Deep Learning: Challenges, Methods, Benchmarks, and Opportunities," in IEEE Jo. of Selected Topics in Applied Earth Observations and Remote Sen., vol. 13, pp. 3735-3756, 2020.
3. P.Deepan and L.R. Sudha, “Object Classification of Remote Sensing Image Using Deep Convolutional Neural Network”, The Cognitive Approach in Cloud Computing and Internet of Things Technologies for Surveillance Tracking Systems, pp.107-120, 2020. https://doi.org/10.1016/B978-0-12-816385-6.00008-8
4. Lei, M., Yu, L., Xueliang, Z., Yuanxin, Y., Gaofei, Y and Brian Alan, J., (2019). Deep learning in remote sensing applications: A meta-analysis and review, ISPRS Journal of Photogrammetry and Remote Sensing: 166–177.
5. Yu, H., Yang, W., Xia, G.S., and Liu, G., (2016). A color-texture-structure descriptor for high resolution satellite image classification”, Journal of Remote Sensing: 259-269.
6. Qin Zou, Lihao Ni, Tong Zhang and Qian Wang, “Deep Learning Based Feature Selection for Remote Sensing Scene Classification”, IEEE, 2015.

7. P. Deepan, and L.R. Sudha., (2019). Fusion of Deep Learning Models for Improving Classification Accuracy of Remote Sensing Images. Journal of Mechanics of Continua and Mathematical Sciences. 14. doi: 10.26782/jmcms.2019.10.00015.
8. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, and A.C. Berg, "SSD: Single Shot multi-box Detector", In: Proc. of European Conference on Computer Vision, pp. 21-37, 2016.
9. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection", In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp.779-788, 2016.
10. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 580–587.
11. J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks", Advances in Neural Information Processing Systems, pp.379-387, 2016.
12. S. Ren, K. He, R. Girshick, X. Zhang, J. Sun. Object detection networks on convolutional feature maps. arXiv: 1504.06066, 2015.
13. Uijlings, J.; Van de Sande, K.; Gevers, T.; Smeulders, A., "Selective Search for Object Recognition", International Journal of Comput. Vis., 104, pp.154–171, 2013.
14. R. Girshick, "Fast R-CNN", In: Proc. of the IEEE International Conference on Computer Vision, pp. 1440-1448, 2015.
15. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks", Advances in Neural Information Processing Systems, pp.91-99, 2015.
16. P.Deepan and L.R. Sudha, "Remote Sensing Image Scene Classification using Dilated Convolutional Neural Networks", International Journal of Emerging Trends in Engineering Research, Vol. 8, No.7, pp.3622-3630, 2020.
17. P.Deepan and L.R. Sudha, "Comparative Analysis of Remote Sensing Images using Various Convolutional Neural Networks", EAI Endorsed Transaction on Cognitive Communications, Vol. 8, No.7, pp.3622-3630, 2021.
18. P. Deepan, L.R. Sudha, "Effective utilization of YOLOv3 model for aircraft detection in Remotely Sensed Images", Materials Today: Proceedings, 2021, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2021.02.831.
19. R. Santhoshkumar, M. Kalaiselvi Geetha, J. Arunnehru, 'SVM-KNN based Emotion Recognition of Human in Video using HOG feature and KLT Tracking Algorithm, *International Journal of Pure and Applied Mathematics*, vol. 117, No. 15, 2017, pp.621-624, ISSN: 1314-3395.
20. R. Santhoshkumar, M. Kalaiselvi Geetha, 'Deep Learning Approach: Emotion Recognition from Human Body Movements', *Journal of Mechanics of Continua and Mathematical Sciences (JMCMS),* Vol.14, No.3, June 2019, pp.182-195, ISSN: 2454-7190.
21. R. Santhoshkumar, M. Kalaiselvi Geetha, *'*Vision based Human Emotion Recognition using HOG-KLT feature*' Advances in Intelligent System and Computing, Lecture Notes in Networks and Systems*, Vol.121, pp.261-272, ISSN: 2194-5357, Springer https://doi.org/10.1007/978-981-15-3369-3_20

22. R. Santhoshkumar, M. Kalaiselvi Geetha, 'Human Emotion Prediction Using Body Expressive Feature', *Microservices in Big Data Analytics, IETE Springer Series,* ISSN 2524-5740, 2019, (Springer), https://doi.org/10.1007/978-981-15-0128-9_13
23. R. Santhoshkumar, M. Kalaiselvi Geetha, 'Emotion Recognition System for Autism Children Using Non-verbal Communication', *International Journal of Innovative Technology and Exploring Engineering (IJITEE),* Vol.8, No.8, June 2019, pp.159-165, ISSN: 2278-3075.
24. B.Rajalingam, R. Santhoshkumar (2020) "Intelligent Multimodal Medical Image Fusion with Deep Guided Filtering", Multimedia Systems, Springer-Verlag GmbH Germany, part of Springer Nature 2020
25. K.P. Sanal Kumar, S Anu H Nair, Deepsubhra Guha Roy, B. Rajalingam, R. Santhosh Kumar " Security and privacy-aware Artificial Intrusion Detection System using Federated Machine Learning" Computers & Electrical Engineering, Volume 96, Part A, December 2021, 107440
26. Dr. B. Rajalingam, Dr. R.Santhoshkumar, Dr. G. Govinda Rajulu, Dr. R. Vasanthselvakumar, Dr. G. JawaherlalNehru, Dr. P. Santosh Kumar Patra "Survey On Automatic Water Controlling System For Garden Using Internet Of Things (Iot)" The George Washington International Law Review, Vol.- 07 Issue -01 April-June 2021.
27. K. He., G. Gkioxari, P. Doll´ar, "Mask R-CNN," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2961–2969, Honolulu, HI, USA, July 2017.