Received: August 3, 2019. Revised: October 22, 2019. 65

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Object Detection Using Adaptive Mask RCNN in Optical Remote Sensing Images

Amira S. Mahmoud[1]* Sayed A. Mohamed[1] Reda A. El-Khoribi[2]
Hisham M. AbdelSalam[2]

[1] *National Authority for Remote Sensing and Space Science, Cairo, Egypt*
[2] *Faculty of Computers and Information, Cairo university, Giza, Egypt*
* Corresponding author's Email: Amira.sobhy@narss.sci.eg

Abstract: Fast and automatic object detection in remote sensing images is a critical and challenging task for civilian and military applications. Recently, deep learning approaches were introduced to overcome the limitation of traditional object detection methods. In this paper, adaptive mask Region-based Convolutional Network (mask-RCNN) is utilized for multi-class object detection in remote sensing images. Transfer learning, data augmentation, and fine-tuning were adopted to overcome objects scale variability, small size, the density of objects, and the scarcity of annotated remote sensing image. Also, five optimization methods were investigated namely: Adaptive Moment Estimation (Adam), stochastic gradient decent (SGD), adaptive learning rate method (Adelta), Root Mean Square Propagation (RMSprop) and hybrid optimization. In hybrid optimization, the training process begins Adam then switches to SGD when appropriate and vice versa. Also, the behaviour of adaptive mask RCNN was compared to baseline deep object detection methods. Several experiments were conducted on the challenging NWPU-VHR-10 dataset. The hybrid method Adam_SGD acheived the highest Accuracy precision, with 95%. Experimental results showed detection performance in terms of accuracy and intersection over union (IOU) boost of performance up to 6%.

Keywords: Object detection, Deep learning, Mask RCNN, Adam, SGD, RmsProp.

1. Introduction

Object detection is a multi-objectives complex problem considering classification and localization single or multi-object in an image [1], In remote sensing domain, object detection becomes even more complicated due to the complex nature of remote sensing images. The term object may include both sharp boundaries (man-made) and vague boundaries fused with the background (landscape) [2]. Object detection has a comprehensive range of applications such as robot vision, face recognition, content-based image retrieval, military applications, and pedestrian detection [1]. Very high-resolution satellite images capture detailed information about the size, shape, texture, and topology of objects on earth in addition to a complicated background, variety of illumination intensities, the influence of weather, noise obscured. Recently, object detection is a hot research topic in remote sensing domain [3]. Object detection methods can be divided into four main groups [3]. 1) template matching based methods which can be subdivided into the rigid template and deformable template-based methods [3]; 2) Knowledge-based method [4], which divided into geometric knowledge and context information [5]; 3) Object Based Image analysis (OBIA) based method which required two main steps object segmentation and object classification [6] and 4) Machine learning based method which consist of two main stages. The first stage is feature extraction using feature engineer methods [7] such as histogram of oriented gradients (HOG) [8], Bag of Words (BOW) [9], sparse representation and Human activity recognition (Har) was adopted in [10]. A group of these features may be used, and different feature reduction methods can also be utilized to improve feature selection stage. In the second stage, a classifier is trained using these features. The widely used classifiers include support vector machine [10], Ada-boost [11], artificial neural network [12].

Recently, deep learning algorithms show their superiority in feature representation tasks in different computer vision and remote sensing domain. The recent evolution of deep learning (DL) in detecting complicated patterns in big remote sensing imagery exposes its high potential to address various challenges such as complexity of satellite images, lack of training datasets, multi-sensor data, complex background, atmospheric conditions. These are the primary challenges to achieve a robust automatic object detection using deep learning.

Region-based Convolutional Network (R-CNN) [18] achieved an excellent object detection accuracy using very deep CNN to classify object proposals. R-CNN has notable drawbacks such as multi-stage pipeline training, extensive training time and space, and slow detection. An enhancement was introduced by Spatial Pyramid Pooling Networks (SPPnets)[19] by sharing convolutions across proposals to limit time cost in training. Fast RCNN [20] operates on a single stage with a multi-task loss during the training phase. This enhancement limits the used storage space and improves accuracy, but region proposal computation still considered the main bottleneck. To overcome this problem, Ren et al. introduced an additional region proposal network (RPN) [13] that replaced the selective search for region proposal generation, thereby combining region proposal, classification, and localization regression improve speed and accuracy but still too slow to achieve real-time detection. Another approach to overcome the time-consumed in region selection step was to directly predict confidences for both classification and localization bounding boxes. YOLO [14] introduced real-time performance by computing a single loss. YOLOv2 [15] is an enhancement that provided a smooth trade-off Between speed and accuracy. SSD method [16], achieved significantly accurate performance compared with YOLO by adding feature map at each scale YOLO versions and SSD methods struggle with small objects within the image, due to the spatial constraints of the algorithm. R-FCN [17, 32] is considered as two-stage object detector which applies the position-sensitive ROI-pooling to tackle the dilemma between translation-invariance in classification and translation-variance in localization however it less accurate than faster R-CNN.

In [12], a deep neural network was utilized for ship detection task in optical images. Various augmentation methods, such as rotation, scaling, and illuminations conditions, were adopted to enhance the learning procedure. In [22], Pan et al. utilized a cascade convolutional neural network (CCNN) framework based on transfer-learning and geometric feature constraints (GFC) to improve the accuracy of aircraft detection. The detection accuracy increased by an average of 3.66%. In [23], an enhancement of Faster R-CNN was introduced to detect densely packed objects in satellite images. Enormous experiments were conducted to evaluate the effectiveness of the proposed method in terms of accuracy and IOU. Results showed the effectiveness of the proposed method.

In [24], AlexNet was adopted to extract generic feature for ship detection task in very high-resolution images. The proposed method outperforms You Only Look Once (YOLO) and SSD in terms of accuracy and IOU. Moreover, Nie et al. [25] proposed a novel framework based on Mask R-CNN for the inshore ship detection task. They adopted Soft-Non-Maximum Suppression (Soft-NMS) to improve the proposed method of performance robustness and efficiency. In [26], Yang et al. proposed a three stages framework for object detection. In the first stage, a sliding window technique utilized to generate the candidate region proposal. Next, AlexNet and GoogleNet were chosen to extract generic image features from each region proposal. Finally, unsupervised score-based bounding box regression (USB-BBR) algorithm was proposed to optimize the bounding box of the detected object. Results of the proposed framework surpass other methods in terms of accuracy and IOU quality with complex backgrounds. Inspired by Faster-RCNN, Li et al. [27] used region proposal network to generate translation-invariant and multi-scale candidate region. Next, local-contextual feature fusion network was used to form a discriminative joint representation (local-and contextual feature) for each candidate region. Finally, accurate classification and accurate object localization were implemented. In [28], Cheng et al. presented a two stages approach based on Faster R-CNN, namely deep adaptive proposal network (DAPNet). The input image is feed to the backbone network to generate the high-level features representation of the image, then the category prior network (CPN) sub-network and fine-region proposal network (F-RPN) used the aforementioned high-level features to obtain the category prior information and candidate regions for each image respectively. Both results were combined to achieve an adaptive region proposal. Finally, the accuracy detection network sub-network was used to classification and regression for each adaptive candidate boxes. Several experiments were carried out on a public NWPUVHR dataset to evaluate the proposed approach performance and results show its superiority. Ammour et al. [29] proposed a car detection method in unmanned aerial vehicle images

(UAV). A mean-shift algorithm was used to segment the UAV input image into small homogeneous regions. Then, a pre-trained Vgg-16 was adopted to extract a generic feature for each segment. Finally, linear support vector (SVM) classifier was adopted to binary map each segment as into "car" and "no car." The proposed method outperformed state-of-the-art methods, both in terms of accuracy and computational time. To overcome the limited accuracy of the traditional ship detection methods, Yang et al. [30]proposed an approach called Rotation Dense Feature Pyramid Network (R-DFPN) method. The proposed method has two stages: dense feature Pyramid Network (DFPN) for feature fusion and Rotation Region Detection Network (RDN) for prediction. Comprehensive evaluations on remote sensing images extracted from Google Earth for ship detection demonstrated the superiority of the proposed method. In [31], Cheng et al. proposed an effective approach to learn a rotation-invariant CNN mode. First, the new rotation-invariant layer was trained by optimizing a new objective function via imposing a regularization constraint then fine-tune the whole CNN network to boost the performance further. The proposed method was evaluated on a public NWPUVHR dataset, and the results denoted the effectiveness of the proposed method.

The problem investigated in this paper, we utilized mask-RCNN to boost the object detection accuracy in the RS domain. The main contribution of this paper is utilizing adaptive Mask RCNN framework to detect multi-scale object in optical remote sensing images. The proposed adaptive mask RCNN efficiently reduce the redundancy of detectors boxes and allow multi-scale targets under complex background images. Transfer learning and fine-tune were adopted to overcome the scarcity and complexity of remote sensing images. The paper also studies the behaviour of adaptive mask RCNN towards baseline optimization methods namely: Adam, SGD, Ada-delta, RMSprop, hybrid SGD_Adam, hybrid Adam_SGD. The paper also studies compare adaptive mask RCNN towards baseline object detection methods Faster RCNN (FRCN) method [13], You only look once (YOLO) method [14], (YOLO2) method [15], Single Shot Multibox Detector (SSD) method [16], Region-based Fully Convolutional Network (R FCN) [17]. All experiments were conducted on a publicly available10-class geospatial object NWPU VHR-10 dataset[33].

The remainder of this paper is organized as follows, proposed adaptive Mask R-CNN is proposed in section 2. Experimental results and discussion were introduced in section 3. Finally, section 4 draws the conclusion.

2. Proposed method

In recent years, deep learning techniques have achieved state-of-the-art results for object detection on standard benchmarks. Mask R-CNN outperformed other deep learning object detection model and won a COCO object detection challenge in 2016. However, the performance of Mask R-CNN in remote sensing domain hardly achieved comparable results due to the complex nature of satellite images, the lake of annotated sampled, and varied object scales. This work study the behavior of different optimization methods and a hybrid training strategy that starts with an adaptive method (Adam) then switches to SGD (SWATS), and vice versa.

Mask-RCNN[33] was introduced by He et al. in 2018 as an extension to Faster RCNN [13] to allow an accurate pixel-based segmentation. It consists of two main stages namely: Feature Pyramid Network (FPN) and Region Proposal Network (RPN). In feature pyramid network, a different number of proposals was generated about the regions where there might be an object based on the input image. First, we utilized a standard convolutional neural network to serve as a feature extractor. The state of art architectures AlexNet, VGG Net and GoogleNet had (5, 19, 22) layers respectively. By getting deeper, the network suffers from vanishing gradient problem, which results in performance saturation or even degrading rapidly. Several attempts [34] had been introduced to overcome the vanishing gradient problem. Based on the residual block, [35] was firstly introduced ResNet50 architecture. Skip connection or shortcut which allow to take activation from one layer and feed it to another layer s that about 2–3 hops away. ResNet50 becomes seminal architecture to different computer vision applications. In this paper, we used a pre-trained architecture on ImageNet (1000 class) dataset. Generally, the size of the recent model is substantially smaller due to the usage of global average pooling rather than fully-connected layers. We choose ResNet50 as a feature extractor network which encodes input image into 32x32x2048 feature map. The FPN extracts regions of interest from features of different levels according to the size of the feature which feeds as input to Next stage (RPN). In Region Proposal Network (RPN), the regions scanned individually and predicted whether or not an object is present. The actual input image is never scanned by RPN instead RPN network scans the feature map, making it much faster. Next, each of

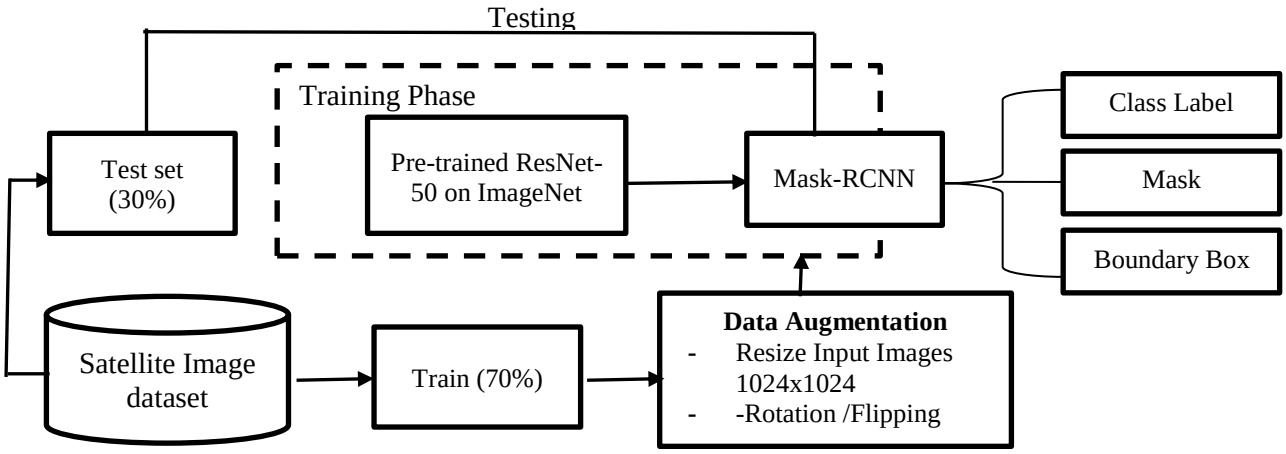

Figure.1 The proposed object detection method for optical remote sensing image

regions of interest proposed by the RPN as inputs and outputs a classification (SoftMax) and a bounding box (regressor). Finally, Mask- RCNN adds a new branch to output a binary mask that indicates whether the given pixel is or not part of an object. This added branch is a Fully Convolutional Network on top of the backbone architecture. The proposed method consists of two main phases: Training and testing phase as illustrated in Fig. 1.

2.1 Loss function

Mask R-CNN utilized a multi-task loss function that combined the loss of classification, localization and segmentation mask as illustrated in Eq. (1).

$$L = L_{cls} + L_{bbox} + L_{mask} \quad (1)$$

Where L_{cls}, L_{bbox} are same as in Faster R-CNN [13]. The added mask L_{mask} is illustrated in *Eq. (2)*. as the average binary cross-entropy that only includes k^{th} mask if the region is associated with the ground truth class k .

$$L_{mask} = -\frac{1}{m^2} \sum_{1 \le i,j \le m} y_{ij} \ log \hat{y}_{ij}^{k} + (1 - y_{ij}) \log(1 - \hat{y}_{ij}^{k}) \quad (2)$$

Where the mask branch generates a mask of dimension m x m for each *RoI* and each classy_{ij} and $\hat{y}_{ij}^{k}$ are cell *(i, j)* label of the true mask and the predicted value respectively.

2.2 Training phase

Mask-RCNN requires a large amount of annotated data for training to avoid overfitting. To overcome the problem of limited annotated dataset in remote sensing domain, we adopted transfer learning by selected the pre-trained network weights of the resnet50 model, which was successfully trained with the image net dataset [36]. We utilized the pre-trained resnet50 and fine-tuned the network weights to the NWPUVHR dataset. Due to limited memory, we consider three different strategies in fine-tuning. First strategy, we train the head layer for 30 epochs while freezing other layers with learning rate 0.1. Second, the convolution layer (+5) and convolution layer (+4) were trained for 30 epochs each using a learning rate 0.01and 0.001, respectively. Finally, the convolution layer (+3) were trained for 400 epochs with learning rate 0.001. We used different argumentation methods such as horizontal flip, vertical flip, image rotation, and image translation to enlarge the training data. One can observe that this domain-specific fine-tuning allows learning good network weights for a high-capacity CNN for NWPUVHR dataset.

2.3 Testing phase

The learned model used directly to predict class label, boundary box, and masked segment for each image in testing data. To evaluate the learned model performance, the predicted labels and boundary box is matched with those in the dataset.

2.4 Optimization techniques

Neural network optimization played an essential role in training deep neural networks. Generally, there are two metrics to evaluate the efficiency of optimizer: speed of convergence and generalization. Stochastic gradient descent (SGD) [37] is commonly used for training deep neural networks. Compared

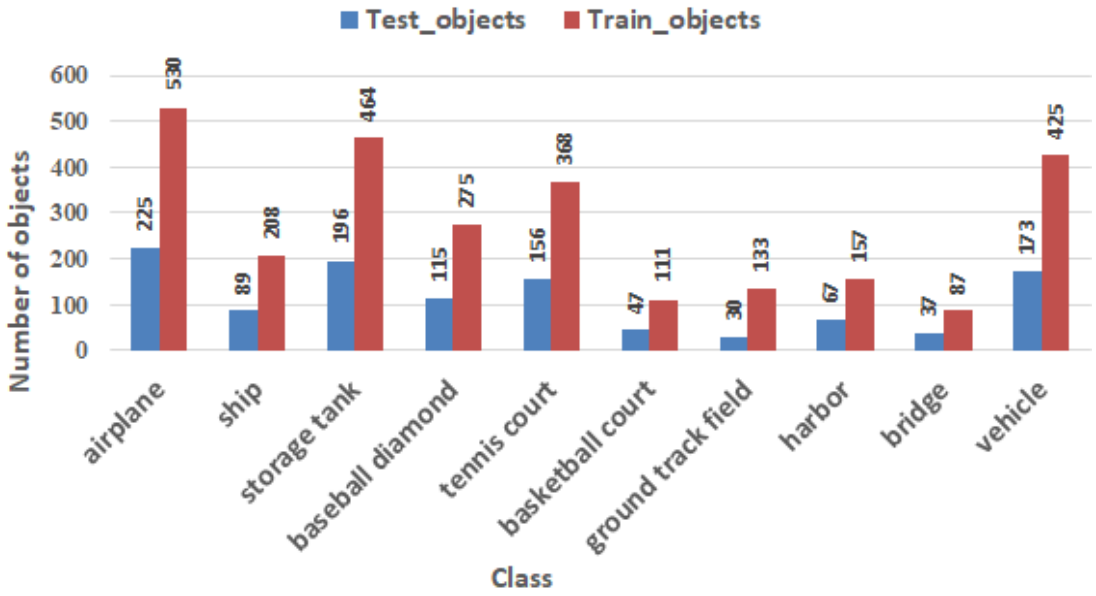

Figure. 2 Statistics of total number of objects of each category used in training and testing in the NWPU VHR-10 data set

with SGD, Adaptive optimization methods such as Adam [38] , Adelta [39], RMSprop [40] perform well in the initial stages of training but tend to generalize poorly. Inspired by their work, Keskar, and Soche [41]. We introduced two-hybrid training strategy that starts with an adaptive method (Adam) then switches to SGD (SWATS), and vice versa. An evaluation of their performance of the hybrid approach in object detection in remote sensing domain. We conducted several experiments to investigate the triggering condition to switch between Adam and SGD. The triggering condition includes the number of epochs and value of learning rate. The optimal triggering condition in object detection was to set the learning rate to 0.001 or epochs achieved 400.

3. Results

In this paper, NWPU-VHR-10geospatial object detection dataset is used to evaluate the performance of our proposed method. We describe the data set and the evaluation metrics in section 3.1 and 3.2, respectively. The implementation details of the proposed method are presented in section 3.3. Finally, the proposed adaptive Mask RCNN method is compared with other state-of-art deep object detection methods, including the results presentation and numerical analysis were depicted in section 3.4.

3.1 Dataset

The NWPU VHR-10 [2-4] is one of the pioneering works in remote sensing object detection filed, which is designed to provide a standard dataset for multi-class in remote sensing images. This data set was cropped from Google Earth then manually annotated by experts; it contains ten classes of objects, namely “airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle” samples as shown in Fig. 2.

In our work, the total number of objects in the NWPU VHR-10 data set is divided into 70% and 30% for training and testing in class level. Fig. 2 presents the statistics of the total number of objects in each class used in both training and testing. Overall, it can be seen that the 10- classes included in NWPU dataset are not equally distributed in terms of the number of images or objects.

3.2 Evaluation matrices

Two evaluation metrics were used to evaluate the proposed object detection method: The Average Precision (AP) [42] and the Precision-recall curves (PRC). The Precision measures the fraction of detections that are true positives as illustrated in *Eq. (3)*and the Recall measures the fraction of positives that are correctly identified as illustrated in *Eq. (4)* The area under the PRC measures the AP metric. The higher the AP value, the better the performance, and vice versa. The precision indicator measures the percentage of your positive predictions are truly positive, and the recall indicator represents the fraction of positives that are correctly identified. The precision and recall indicators are formulated as follows.

Table 1. Performance for YOLO, Faster RCNN, SSD, R-FCN, and proposed method on NWPU dataset in terms of AP percentage values and average running time in seconds per image

Class	**FRCN [13]**	**YOLO1 [14]**	**YOLO2 [15]**	**SSD [16]**	**R-FCN [17]**	**Proposed Method**
Airplane	82.8	60.8	87.3	95.7	96.1	**99.9**
Ship	77.5	62.7	84.7	93.6	**98.3**	92.7
Storage tank	52.5	28.7	42.7	60.9	72.5	**94.5**
Baseball diamond	96.3	85.7	93.1	99.4	99.4	**99.5**
Tennis court	62.9	58.4	65.7	87.7	90.7	**97.3**
Basketball	68.8	82.2	85.5	**92**	97.8	88.9
Ground track field	98.4	88.7	97.1	**98.6**	99.3	93.8
Harbor	82.5	75	80.5	94.6	92.5	**95.9**
Bridge	78.8	72.5	90	**97.0**	93.4	95.8
Vehicle	63.8	52.3	70.8	74.5	88.4	**91.6**
Mean AP	76.4	66.7	79.7	89.4	92.8	**95**
Time (s)	6.21	3.36	4.24	5.72	4.32	7.1

$$percision = \frac{tp}{tp+fp} \quad (3)$$

$$recall = \frac{tp}{tp+fn} \quad (4)$$

Where tp = True Positives, fp = False Positives and fn = False Negatives.

3.3 Implementation details

We randomly selected 500 images from the positive images as training images. The rest 150 images were used to evaluate the performance of the proposed object detection method. Owing to the limited size of the training set, different data augmentation was adopted such as rotation, flipped it horizontally and vertically to expand the number of samples. The augmented images were considered as a representation for rotation of target, lighting changes, and the variety of sensors. We conducted our experiments using NVIDIA GEFORCE® GTX 1080 Ti, 11 GB of memory, to considerably speed up deep learning training computations. Tensor Flow [43] was selected as the implementing framework.

3.4 Results

In this section, we conducted several experiments to evaluate the performance of the proposed method in term of average precision, computation time, Intersection over Union (IOU), and Precision-Recall Curves (PRC).

First, we compare the performance of the proposed method against the deep learning baseline object detection techniques namely Faster- Region-based Convolutional Network (FRCN) method [13],You only look once (YOLO1) [14] You only look once (YOLO2) method [15], Single Shot Multibox Detector (SSD) method [16], and Region-based Fully Convolutional Network (R-FCN) [17] in term of average precision and computation time.

Table 1 shows the obtained results achieved measured by AP values for each class of NWPU dataset. One can observe that YOLO2 and SSD achieved a comparable performance 79.7%, and 89.4% respectively. However, YOLO is slightly faster compared with other techniques. R-FCN has the highest mean AP value (92.8%). Our proposed method outperforms other techniques and boosts performance by 6%. The proposed method achieved a better trade-off between detection accuracy and speed. However, although our method has achieved the best performance in terms of AP, the detection accuracy for the object categories of basketball court and vehicle is still relatively low.

Table 1 shows the quantitative comparison results of six different methods measured by AP values, and average running time per image. The best performances are highlighted in bold. The proposed method achieved the best performance in terms of mean AP. However, it reached the lowest performance in terms of speed. However, the proposed method still needs to be improved in terms of speed to achieve real-time performance. The balance between computational complexity and performance remains a big challenge. In the future, we would like to investigate other approaches to meet lower computational burden and system complexity.

Second, we conduct several experiments to evaluate six optimization method in the remote sensing object detection task. Four optimization techniques: Adam, Adelta, RMSprop and SGD, and two hybrid techniques Adam- SGD, and SGD_Adam were tested method in terms of IOU. The recall rates of these optimization techniques under different IOU thresholds are plotted in Fig. 3. It can be observed that

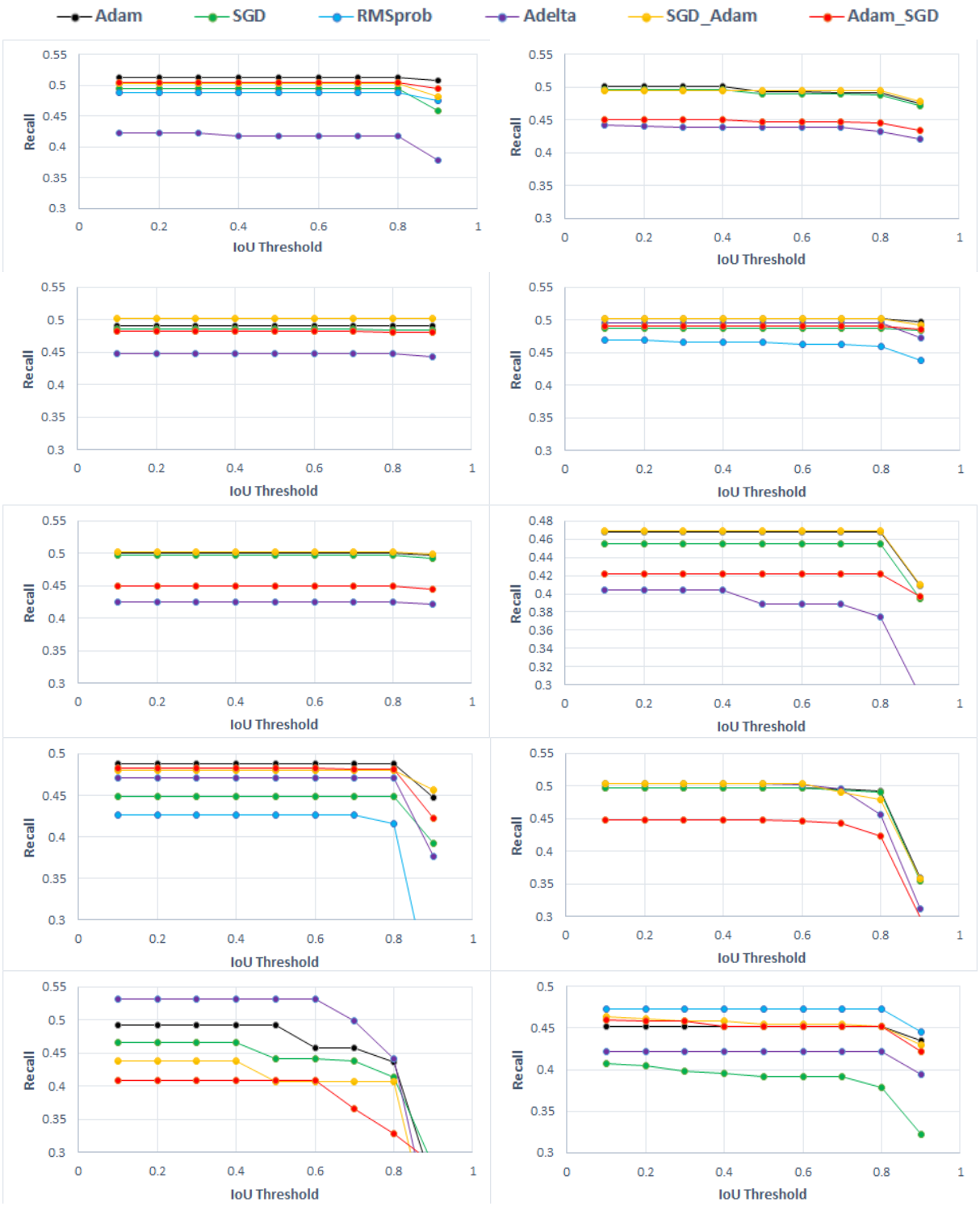

Figure. 3 Recall vs. IOU overlap ratio on the NWPU VHR-10 data set for airplane, ship, storage tank, baseball diamond, tennis court, basketball, ground track field, and harbour, bridge, and vehicle classes, respectively

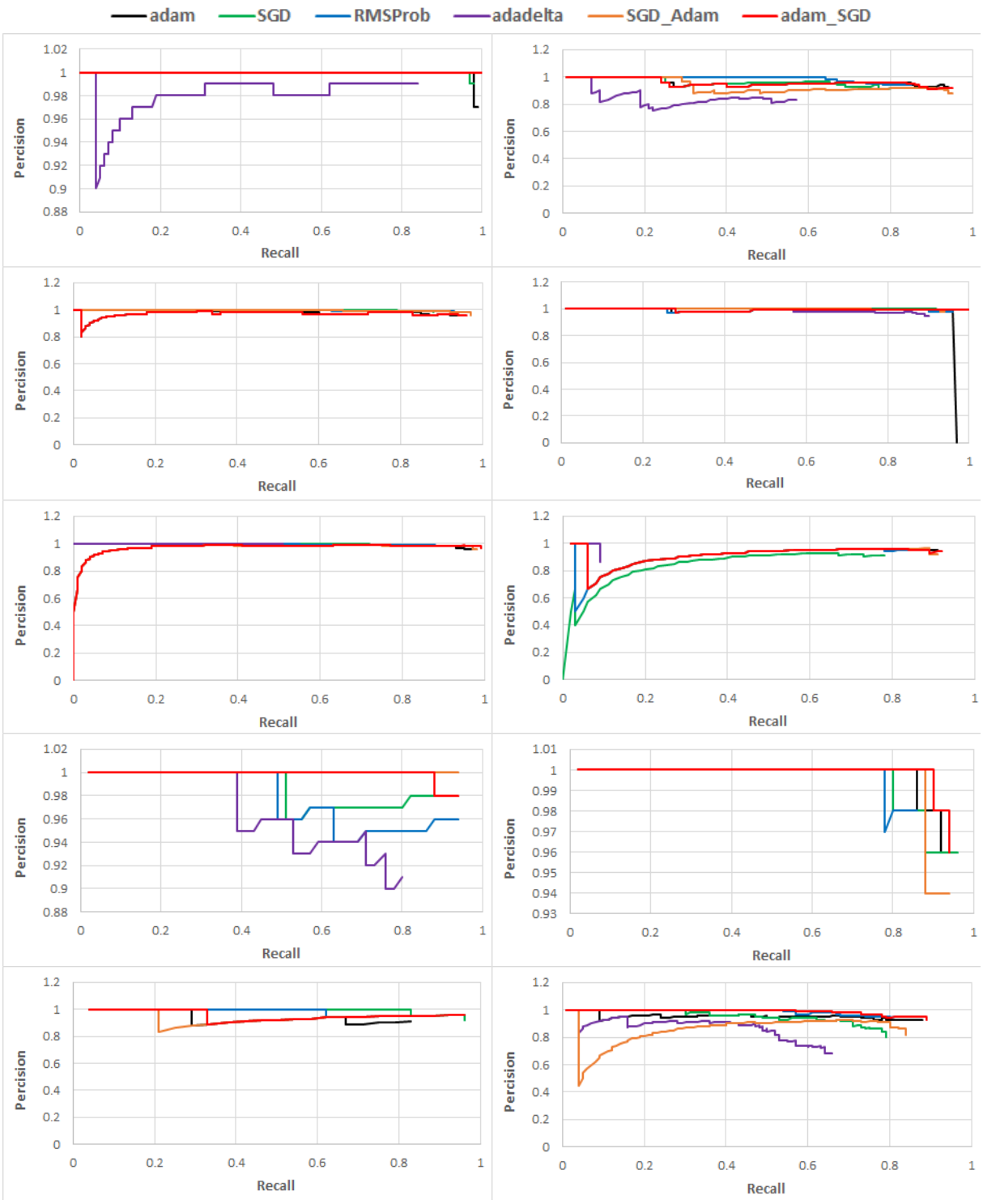

Figure.4 precision and recall on the NWPU VHR-10 dataset for airplane, ship, storage tank, baseball diamond, tennis court, basketball, ground track field, harbour, bridge, and vehicle classes, respectively

(1) The recall curves declined with the increasing of IoU thresholds. In detail, the recall of Adelta and SGD optimization decreased more quickly compared with other techniques, which demonstrates their limited performance in object detection task in remote sensing domain. (2) For object classes such as basketball, ground-track, and harbor, the recall of different optimization techniques is higher compared with other object classes. This is due to small size objects with a complex background are harder to detect. (3) Hybrid based optimization Adam-SGD

Figure. 5 samples of object detection result with the proposed approach

achieves the highest recall rate compared with other optimization techniques. The remaining optimization techniques have comparable performance. Overall, our proposed method achieved higher recall for a small object, which is vital in object detection in remote sensing domain due to the different resolution of satellite data.

Third, to evaluate the quality of the proposed method in each class detection, Fig. 4 displays the precision-recall curves (PRCs) of six optimizations aforementioned techniques. For better comparison and visualization, we plot (1-recall) for X-axes and (1- precision) for Y-axes. As can be seen from them, (1) all optimization techniques achieved superb performance for the object categories of airplane and baseball diamond. However, for other eight object categories, the PRC of different optimization techniques are varied. This is due to that both classes have relatively larger in training samples count and size. (2) Hybrid Adam-SGD and Hybrid SGD_Adam have higher precision than Adam and SGD, respectively. This demonstrates that hybrid optimization method can boost detection performance. 3) Adelta optimization method is not favored in object detection (3) other optimization techniques achieved comparable performance except for Adelta. Overall, the hybrid optimization (SWATS) is very effective for object detection in remote sensing images.

Finally, Table 2 shows the quantitative comparison results of the six, as mentioned earlier optimization method in terms of AP values, and average running time per image. The best performances are highlighted in bold. The performance of Ada-delta optimization shows struggles with small-size objects, whereas the hybrid-optimization method (ADAM-SGD) is more effective for detection small size objects in remote sensing images. (4) Hybrid-optimization (ADAM-SGD) method attained the highest AP values for the most class of objects. This show that our method is effective for detecting objects with various size. Adam and Hybrid SGD_Adam obtained a comparable detection performance compared with Hybrid Adam-SGD. An obtained gain in performance up to 6 % in terms of mean AP, which illustrates that the switching between Adam to SGD can effectively improve the generalization of object detection in remote sensing domain. Compared with SGD and Ada-delta, Hybrid Adam-SGD achieved up to 6% and 40% performance gains in terms of mean AP respectively. Fig. 5 shows object detection results of samples with the proposed approach. Some objects such as storage tank, tennis court are densely peaked, vehicles and ships are small in size with a complex background, and ground track field has a large size. The proposed method has successfully detected most of these objects, demonstrating the effectiveness of our method. As can be seen from Table 2, Adam and Hybrid SGD_Adam achieved near 100% AP for airplane and baseball diamond, but the AP value has degenerated. This is mainly because the small size of the object leads to limited feature representation for accurate object detection.

4. Conclusion

This paper proposed an adaptive Mask RCNN approach for detecting multi-class objects in remote sensing images. We utilized transfer learning, fine-tuning, and augmentation techniques such as rotation, scaling, and illuminations conditions to overcome the insufficient labeled remote sensing imagery. The paper also draws a comparison between the proposed method and the baseline deep object detection techniques in term of average precision, computation time, Intersection over Union (IOU), and Precision-Recall Curves (PRC). Numerous experiments were conducted using challenging multi-class NWPU-VHR-10 dataset. The dataset was split into 70% and 30% for training and testing respectively. Also, several experiments were performed to evaluate the effectiveness of optimization techniques, namely: Adam, SGD, Ada-delta, RMSprop, hybrid SGD_Adam, hybrid Adam_SGD.

Analyze the results, the proposed method outperforms other baseline object detection methods and boot the performance by 6% in terms of AP. In terms of IOU and PRCS, the results obtained from all

Table 2. performance of six optimization techniques in terms of AP percentage values and average running time per image

class	Adam	SGD	RMSprop	A-delta	SGD_Adam	Adam_SGD
Airplane	99.0	97.6	97.2	82.8	**100**	99.9
Ship	91	81.7	84.3	50	89.8	**92.7**
Storage tank	93.3	87.6	94.3	43.5	**96.9**	94.5
Baseball diamond	96.6	97.1	95.8	89.2	98.4	**99.5**
Tennis court	95.3	83.2	87.8	51.2	96.8	**97.3**
Basketball	87.6	72.8	81.2	9.4	87.7	**88.9**
Ground track field	87.7	90.9	92.2	77.4	**93.9**	93.8
Harbor	93.8	95.5	85.8	16	93.6	**95.9**
Bridge	79.6	95.3	74.3	8.3	76.3	**95.8**
Vehicle	83.5	75.6	79.8	58.5	79.2	**91.6**
Mean AP	90.8	87.7	87.3	48.6	91.2	**95.0**

optimization techniques clarify superb performance for the object categories of airplane and baseball diamond. However, for other eight object categories, are varied. This is due to that both classes have relatively larger in training samples count and size.

The AP metric measures the area under the PRC. The higher the AP value, the better the performance, and vice versa. the results of the average precision (AP) in optimization techniques Adam, SGD, RMSprop, Adelta, hybrid SGD_Adam, and hybrid Adam_SGD were 90.8%, 87.7%, 87.3%, 48.6%, 91.2, and 95% respectively. The proposed adaptive Mask RCNN firstly, outperformed other deep learning methods and achieved the highest accuracy in terms of IOU and PRC by utilizing the switch between optimizers SWATS (switch from Adam to SGD) in training phase compared with utilizing default optimizer (SGD) in other methods. Secondly, SWATS achieved a verified high accuracy with reducing the computation time and cost. Hence, in our future work, we intend to implement an ensemble of heterogeneous object detection approaches. In addition to incorporate a multi-GPU configuration to further reduce the computation time.

References

[1] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images", *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol.117, pp.11-28, 2016.

[2] T.R. Martha, N. Kerle, C.J. Westen, V. Jetten, and K.V. Kumar, "Segment optimization and data-driven thresholding for knowledge-based landslide detection by object-based image analysis", *IEEE Transactions on Geoscience and Remote Sensing*, Vol.49, No.12, pp. 4928-4943, 2011.

[3] D. Chaudhuri, N. K. Kushwaha, and A.Samal, "Semi-automated road detection from high resolution satellite images by directional morphological enhancement and segmentation techniques", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol.5, No.5, pp. 1538-1544, 2012.

[4] A.O. Ok, "Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts", *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol.86, pp. 21-40, 2013.

[5] T. Blaschke, G.J. Hay, M. Kelly, S. Lang, P. Hofmann, E. Addink, R.Q. Feitosa, F.V. Meer, H.V. Werff, F.V. Coillie, and D. tiede, "Geographic object-based image analysis–towards a new paradigm", *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol.87, pp. 180-191, 2014.

[6] Y. Li, S. Wang, Q. Tian, and X. Ding, "Feature representation for statistical-learning-based object detection: A review", *Pattern Recognition*, Vol.48, No.11, pp. 3542-3559, 2015.

[7] G. Cheng, J. Han, L. Guo, Z. Liu, S. Bu, and J. Ren, "Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images", *IEEE Transactions on Geoscience and Remote Sensing*, Vol.53, No.8, pp. 4238-4249, 2015.

[8] D. Zhang, J. Han, G. Cheng, Z. Liu, S. Bu, and L. Guo, "Weakly supervised learning for target detection in remote sensing images", *IEEE Geoscience and Remote Sensing Letters*, Vol.12, No.4, pp. 701-705, 2015.

[9] N. Yokoya and A. Iwasaki, "Object detection based on sparse representation and Hough voting for optical remote sensing imagery", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol.8, No.5, pp. 2053-2062, 2015.

[10] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review", *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol.66, No.3, pp. 247-259, 2011.

[11] Z. Shi, X. Yu, Z. Jiang, and B. Li, "Ship detection in high-resolution optical imagery based on anomaly detector and local shape feature", *IEEE Transactions on Geoscience and Remote Sensing*, Vol.52, No.8, pp.4511-4523, 2014.

[12] J. Tang, C. Deng, G.B. Huang, and B. Zhao, "Compressed-domain ship detection on spaceborne optical image using deep neural network and extreme learning machine", *IEEE Transactions on Geoscience and Remote Sensing*, Vol.53, No.3, pp.1174-1185, 2015.

[13] 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.

[14] 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.

[15] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger", In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7263-7271, 2017.

[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, and A.C. Berg, "Ssd: Single shot multibox detector", In: *Proc. of European Conference on Computer Vision*, pp. 21-37, 2016.

[17] 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.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation", In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580-587, 2014.

[19] 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.

[20] R. Girshick, "Fast r-cnn", In: *Proc. of the IEEE International Conference on Computer Vision*, pp. 1440-1448, 2015.

[21] M.M.U. Rathore, A. Paul, A. Ahmad, B.W. Chen, B. Huang, and W. Ji, "Real-time big data analytical architecture for remote sensing application", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol.8, No.10, pp. 4610-4621, 2015.

[22] B. Pan, J. Tai, Q. Zheng, and S. Zhao, "Cascade Convolutional Neural Network Based on Transfer-Learning for Aircraft Detection on High-Resolution Remote Sensing Images", *Journal of Sensors*, Vol.2017, 2017.

[23] Z. Deng, L. Lei, H. Sun, H. Zou, S. Zhou, and J. Zhao, "An enhanced deep convolutional neural network for densely packed objects detection in remote sensing images", *International Workshop on Remote Sensing with Intelligent Processing*, pp. 1-4, 2017.

[24] T. Wang and Y. Gu, "Cnn Based Renormalization Method for Ship Detection in Vhr Remote Sensing Images", In: *Proc. of IGARSS IEEE International Geoscience and Remote Sensing Symposium*, pp.1252-1255, 2018.

[25] S. Nie, Z. Jiang, H. Zhang, B. Cai, and Y.Yao, "Inshore Ship Detection Based on Mask R-CNN", In: *Proc. of IGARSS IEEE International Geoscience and Remote Sensing Symposium*, pp.693-696, 2018.

[26] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate Object Localization in Remote Sensing Images Based on Convolutional Neural Networks", *IEEE Transactions on Geoscience and Remote Sensing*, Vol.55, No.5, pp. 2486-2498, 2017.

[27] K. Li, G. Cheng, S. Bu, and X. You, "Rotation-insensitive and context-augmented object detection in remote sensing images", *IEEE Transactions on Geoscience and Remote Sensing*, Vol.56, No.4, pp. 2337-2348, 2017.

[28] L. Cheng, X. Liu, L. Li, L. Jiao, and X. Tang, "Deep Adaptive Proposal Network for Object Detection in Optical Remote Sensing Images", *arXiv preprint arXiv:1807.07327*, 2018.

[29] N. Ammour, H. Alhichri, Y. Bazi ,B. Benjdira, N. Alajlan, and M. Zuair, "Deep learning approach for car detection in UAV imagery", *Remote Sensing*, Vol.9, No.4, pp.31, 2017.

[30] X. Yang, H. Sun, k. Fu, J. Yang, X. Sun, M. Yan, and Z. Guo, "Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense

feature pyramid networks", *Remote Sensing*, Vol.10, No.1, pp.132, 2018.

[31] 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.

[32] 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.

[33] K. Zhao, J. Kang, J. Jung, and G.Soh, "Building Extraction from Satellite Images Using Mask R-CNN with Building Boundary Regularization", In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp.247-251, 2018.

[34] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks", *Pattern Recognition*, Vol.77, pp.354-377, 2018.

[35] K. He, X. Zhang, S. Ren, and J. Sun "Deep residual learning for image recognition", In: *Proc.of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.770-778, 2016.

[36] J. Wang, C. Luo, H. Huang, H. Zhao, and S. Wang, "Transferring Pre-Trained Deep CNNs for Remote Scene Classification with General Features Learned from Linear PCA Network", *Remote Sensing*, Vol.9, No.3, pp.225, 2017.

[37] H. Robbins and N. Carolina, "A stochastic approximation method", *The Annals of Mathematical Statistics*, pp. 400-40, 1951.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", *arXiv preprint arXiv:1412.6980*, 2014.

[39] M.D. Zeiler, "ADADELTA an adaptive learning rate method", *arXiv preprint arXiv:1212.5701*, 2012.

[40] T. Tieleman and G. Hinton ,"Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning", *Tech. rep., Technical Report*. Available online: https://zh.coursera.org/learn/neuralnetworks/lecture/YQHki/rmsprop-divide-the-gradient-by-a-running-average-of-its-recent-magnitude (Accessed on 21 April 2017)

[41] N.S. Keskar and R. Socher, "Improving generalization performance by switching from adam to sgd", *arXiv preprint arXiv:1712.07628*, 2017.

[42] K. Oksuz, B.C. Cam, E. Akbas, S. Kalkan, "Localization recall precision (lrp): A new performance metric for object detection", In: *Proc.of the European Conference on Computer Vision (ECCV)*, pp.504-519, 2018.

[43] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning", In: *Proc. of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.