Journal of Theoretical and Applied Information Technology
15th February 2020. Vol.98. No 03

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

OBJECT DETECTION IN HIGH-RESOLUTION AERIAL IMAGES BASED ON IMAGE PYRAMID AND PATCH DETECTION

HOANH NGUYEN

Faculty of Electrical Engineering Technology, Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam

E-mail: nguyenhoanh@iuh.edu.vn

ABSTRACT

Recently, many approaches based on deep convolutional neural networks (CNNs) for object detection have showed better performance compared with traditional approaches. Since objects in high-resolution aerial images are usually very small with complex backgrounds, the performance of deep CNN-based approaches on object detection in high-resolution aerial images is still limited. In addition, with the constraint of the amount of memory on graphics processing units (GPUs), current state-of-the-art CNN architectures cannot directly process high-resolution aerial images. This paper proposes an improved deep CNN-based framework for object detection in high-resolution aerial images. To detect objects at different scales, image pyramid with different sizes is generated from single input image by down-sampling the original high-resolution input image. Each image level in the image pyramid can be used to detect objects at a different scale. For each image in the image pyramid, small patches with small fixed-size are generated. Each patch is then fed to the detection network, which is based on SSD framework with VGG-16 architecture as the base network, to generate patch detection results. Patch detection results are then projected to the image pyramid at the original scale to form image detection results. Finally, Non-Maximum Suppression (NMS) algorithm is adopted after image detection results to create final detection results. Experimental results on public datasets with high-resolution aerial images show that the proposed approach is very simple and efficient while achieving nearly as detection accuracy as recent state-of-the-art methods.

Keywords: *Object Detection, Deep Learning, Image Pyramid, Aerial Image, Convolutional Neural Network*

1. INTRODUCTION

With the development of earth observation technology and the diversity of remote sensing platforms, vision-based object detection in aerial images has attracted more and more attention. However, due to difficult conditions such as the complex backgrounds, high-resolution images, small objects, the uneven distributions of training samples in terms of size and quantity, illumination, and occlusion, object detection in aerial images are still challenging. Existing vision-based object detection in aerial images can be divided into two groups: traditional methods and deep learning-based methods. Traditional methods first use the traditional filters to extract features and then perform feature fusion and dimension reduction to concisely extract features. Finally, the features are fed into a classifier like Support Vector Machine or AdaBoost, which rely on hand-crafted features. However, these classifiers have difficulty to efficiently processing aerial images in the context of big data. In addition, hand-crafted features can detect only specific targets, when applying them to other objects, the detection results are unsatisfactory.

Recently, deep learning-based methods for object detection in aerial images have achieved good performance. However, these innovations usually fail to detect very small objects because small object features are lost during the downsampling processes of convolution layers. In addition, objects in aerial images usually have small size, and the objects are usually blurry, which has created considerable challenges in normal object detection with no good solutions to date. To alleviate the issues of small object detection, many methods such as feature pyramid network, deeply supervised object detectors, and scale normalization for image

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

pyramids have been proposed. To a certain extent, these methods strengthen the feature extraction of small objects. However, they do not perform well when detecting aerial objects because many objects in aerial images have complex backgrounds due to terrain or illumination factors, and the above methods cannot easily distinguish them.

To address previous issues, this paper proposes a framework to improve the performance of object detection in high-resolution aerial images. In the proposed framework, image pyramid with different image sizes is first generated from high-resolution input image. The image pyramid can help to detect objects at a different scale. For each image in the image pyramid, small patches with small fixed-size are generated. Each patch is then fed to the detection network, which is based on SSD framework with VGG-16 architecture as the base network, to create patch detection results. Patch detection on small patches can solve the problem of the memory constraint on GPUs. Patch detection results are then projected to the image pyramid at the original scale to form image detection results. Finally, Non-Maximum Suppression algorithm is adopted after image detection results to create final detection results. Experimental results on public datasets show that the proposed approach achieves nearly as detection accuracy as recent state-of-the-art methods while being simpler and more efficient.

This paper is organized as follows: an overview of previous methods on vehicle detection is presented in Section 2. Section 3 describes detail the proposed method. Section 4 demonstrates experimental results. Finally, the conclusion is made in Section 5.

2. RELATED WORK

Existing object detection approaches can be divided into two main categories: traditional methods and deep learning methods. Traditional methods include the scale-invariant feature transform [13] [37] and histogram of oriented gradients [14]. These methods first use the traditional filters to extract features and then perform feature fusion and dimension reduction to concisely extract features. Finally, the features are fed into a classifier like Support Vector Machine [15], AdaBoost [16], which rely on hand-crafted features.

Recently, deep learning-based object detection approaches, including two-stage networks such as Faster R-CNN [7], and one-stage networks such as You Only Look Once (YOLO) [12] and single shot multibox detector (SSD) [1], have achieved good performance in object detection tasks. Faster R-CNN introduces a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals. SSD framework skips the region proposal stage and directly uses multiple feature maps with different resolutions to perform object localization and classification. YOLO solves object detection as a regression problem to spatially separated bounding boxes and associated class probabilities. A single neural network predicts bounding boxes and class probabilities directly from full images in one evaluation. Since the performance of deep learning-based object detector is significantly affected by the base network, many theoretical studies concerning the architecture of the base network have been conducted. AlexNet, developed by Krizhevsky et al. [20], was a groundbreaking CNN architecture. The main feature of the GoogleNet [22] is its improved utilization of the computing resources inside the network. This improvement was achieved through a carefully crafted design that allowed the depth and width of the network to increase while keeping the computational demands constant. The VGG models proposed by Simonyan and Zisserman [8] were used to investigate the relationship between the depth of a convolutional network and its accuracy in large-scale image recognition regardless of the size or scale of the image, thus eliminating the requirement for a fixed-size input image. ResNet [22] was reformulated to learn residual functions with reference to the layer inputs instead of learning unreferenced functions to ease the training of networks that are substantially deeper than those used previously. DenseNet [23] based on the ResNet uses dense connections to enhance the feature propagation, and greatly reduce the numbers of parameters.

To better handle the issues of small object detection, many methods such as feature pyramid network (FPN) [17], deeply supervised object detectors (DSOD) [18], and scale normalization for image pyramids [19] have been proposed. To a certain extent, these methods strengthen the feature extraction of small objects. FPN exploited the inherent multi-scale, pyramidal hierarchy of deep convolutional networks to construct feature pyramids with marginal extra cost. Furthermore, a top-down architecture with lateral connections is developed for building high-level semantic feature maps at all scales. DSOD designed a set of design principles for training object detectors from scratch. Scale normalization for image pyramids applied high-capacity convolutional neural networks to

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

bottom-up region proposals in order to localize and segment objects.

In the field of object detection in aerial images, many deep learning-based methods have been proposed to improve the detection accuracy. Yang et al. [24] proposed a framework called Rotation Dense Feature Pyramid Networks which can effectively detect ship in different scenes including ocean and port. Furthermore, a rotation anchor strategy is designed to predict the minimum circumscribed rectangle of the object so as to reduce the redundant detection region and improve the recall. Zhang et al. [25] developed a CNN-based method to extract the high-level features and the hierarchical feature representations of the objects. An iterative weakly supervised learning framework is then employed to automatically mine and augment the training data set from the original image. In [26], the authors first adopted a region proposal method to generate candidate regions with the aim of detecting all objects of interest within these images. Then, generic image features from a local image corresponding to each region proposal are extracted by a combination model of 2-D reduction convolutional neural networks. Finally, to improve the location accuracy, an unsupervised score-based bounding box regression algorithm combined with a non-maximum suppression algorithm was proposed to optimize the bounding boxes of regions that detected as objects. The deep learning-based method [27] used CNN features from combined layers to perform orientation-robust aerial object detection. A position-sensitive balancing framework [28] based on the ResNet and a novel end-to-end adaptively aspect-ratio multi scale network [29] can significantly improve detection accuracy. Wu et al. [30] proposed an efficient way to automatically learn the presentations from the passive image data and increase the computational efficiency of aircraft detection. Ding et al. [31] investigated the capabilities of a CNN model combined with data augmentation operations in SAR target recognition. Zhang et al. [32] designed a network with a deconvolution layer after the last convolution layer of base network for small object detection on high resolution remote sensing images. In [33], the authors presented an automatic content-based analysis of aerial imagery in order to detect and mark arbitrary objects or regions in high resolution images. Zhang et al. [34] presented a hierarchical oil tank detector with deep surrounding features combined with local features to describe oil tanks and then applied gradient orientation to select candidate regions from satellite images. Salberg et al. [35] investigated an algorithm for automatic detection of seals in aerial remote sensing images using features extracted from a pre-trained deep convolutional neural network. Jiang et al. [36] proposed a vehicle detection method in satellite images using DCNNs based on super-pixel segmentation. Zhu et al. [38] used CNN features from combined layers to perform orientation-robust aerial object detection.

3. METHODOLOGY

In this section, this paper presents the details of the proposed model in Section 3.1. The details of training and testing process are presented in Section 3.2.

3.1 Model

Figure 1 illustrates the overall architecture of the proposed framework. As shown in Figure 1, to reduce the memory usage on high-resolution aerial images, the input image is cropped into small fixed-sized patches. These patches are then fed into the detection network. Moreover, since large objects may not be entirely covered in a single image patch, the original image is down-sampled to form an image pyramid. Image pyramid allows the proposed framework to achieve scale-invariance and to process input images with a variety of resolutions. In the detection network, the proposed method makes predictions on only one feature map of small scale. Details of each module of the proposed method will be explain in the following sections.

3.1.1 ResNet

Since the detection network is designed to be sensitive to small objects in high-resolution aerial images, large objects will not be detected in the original image. Thus, this paper constructs an image pyramid based on input image as shown in Figure 2. With the image pyramid, the larger objects that cannot be detected in the image with original resolution become detectable on images with smaller scales. Moreover, since the memory available on GPUs is limited, the VGG-16 [8] network cannot process large images. To solve this problem, small patches with fixed size $W \times H$ will be cropped from each image level in the image pyramid as the input to the detection network. Each patch is created by using a sliding window with a stride of s in both horizontal and vertical direction on image.

3.1.2 Enhanced Feature Map Generation

The detection network is based on SSD framework [1] with VGG-16 [8] as the base convolution layers. SSD framework is built on top of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

a base network that ends with some convolutional layers. SSD adds a series of progressively smaller convolutional layers. Each of the added layers, and some of the earlier base network layers are used to predict scores and offsets for some pre-defined default bounding boxes. These predictions are performed by two 3x3 filters, one filter for each category score and one for each dimension of the bounding box that is regressed. In this paper, single scale feature map is used for detection. More specific, this paper produces the object detection on the feature map generated by the highest-level feature map, i.e. *Conv4-3* as shown in Figure 3. The receptive field of this layer is 97 × 97, which is adequate for small object detection [9], [10].

For the default anchor boxes, similar to the approach in Faster-RCNN [7] and SSD [1], a set of pre-defined default boxes with different sizes and aspect ratios are introduced at each location of the highest-level feature map of the base convolution layers to assist producing the predictions for bounding boxes. Instead of directly predicting the location of the bounding boxes for each object in an image, for each position of the feature map, the detection network predicts the offsets relative to each of the default boxes and the corresponding confidence scores over the target classes simultaneously. Specifically, given n default boxes associated with each location on the highest-level feature map with a size of $w \times h$, there are $n \times w \times h$ default boxes in total. For each of the default boxes, c classes and 4 offsets relative to the default box location should be computed. As a result, $(c + 4) \times n \times w \times h$ predictions are generated for the feature map.

3.2 Training and Testing

3.2.1 Training Samples

For the training samples, 200×200 patches centered at target objects are cropped from the original images as input of the detection network. For each training samples, the target objects may be larger than the patch at the current pyramid level. Moreover, multiple objects might be included in one patch. Thus, to create the training samples, this paper considers an object as positive only if over 1/2 area of the object is covered in the patch. In addition, to include more background information, a set of patches containing only background are randomly cropped from the original training images for learning the model. The ratio between the number of background patches and that of the positive patches is roughly 2:1.

3.2.2 Choosing the Default Boxes

In SSD framework, there are six default boxes per feature map location. By combining predictions for all default boxes with different scales and aspect ratios from all locations of many feature maps, a diverse set of predictions covering various input object sizes and shapes are obtained. This paper chooses the default boxes with small size to ensure the performance of the proposed method for small object detection in high-resolution aerial images. More specific, the size of the square default boxes is $S_1 = 0.1 \times 200$ and $S_2 = \sqrt{(0.1 \times 200) \times (0.2 \times 200)}$. With the resolution of input image of the detection network is 200×200, default boxes occupy around 10% of area of the input image. To make the network fit better to objects with a shape other than square, the aspect ratios are chosen as $a_R \in \left\{2,3,\frac{1}{2},\frac{1}{3}\right\}$. Thus, there are total 6 default boxes with size of 25×25 at each location of the highest-level feature map. The width w_R and the height h_R of the corresponding default box can be calculated as follows:

$$w_R = S_1\sqrt{a_R} \quad (1)$$

$$h_R = \frac{S_1}{\sqrt{a_R}} \quad (2)$$

3.2.3 Matching Default Boxes

In training phase, this paper first finds the correspondence between the default boxes and the ground-truth bounding boxes by calculating the Jaccard overlap between each default box and the ground truth boxes as in MultiBox [11]. The default boxes are labeled as "matched" when the Jaccard overlap is over 0.5. Analogous to regressing multiple boxes at each location in YOLO [12], different default boxes can be matched to one ground truth box. For each of the matched boxes, offsets relative to the box shape and the corresponding confidence scores used to calculate the loss and update the parameters of the detection network are produced.

3.2.4 Loss Function

The overall objective loss function is to minimize the localization loss (L_{conf}) and the confidence loss (L_{loc}) [1]. The overall objective loss function is defined as follows:

$$L = \frac{L_{conf}(x,c) + \lambda L_{loc}(x,\hat{b},b)}{N} \quad (3)$$

where x represents a matched default box; N represents the number of matched default boxes; L_{loc} is the Smooth L1 loss based on the predicted box $\hat{b}$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

and the ground truth bounding box b; L_{conf} is the softmax loss over target classes; and λ represents the weight to balance between the two losses. In this paper, λ is set to 1 by cross validation.

3.2.5 Data Augmentation

To make the network more robust to various input object sizes and shapes, this paper adopts similar data augmentation approach as in [1]. Training samples will be produced by cropping patches from the input images. The overlapped part of the ground-truth box will be kept if over 70 percent of its area falls in the sampled patch, and the sampled patch is resized to a fixed size.

3.2.6 Hard Negative Sampling

During the training process, hard negative samples are selected for training according to the confidence scores after each iteration. More specific, at the end of each training iteration, the miss-classified negative samples will be sorted based on the confidence scores and the ones with the highest confidence scores will be chosen as hard negative samples. Hard negative samples are then used to update the weights of the network. Following the implementation in SSD [1], the number of hard negatives used for training the model is at most 3 times larger than the number of positives.

3.2.7 Testing

In the testing phase, since the limited amount of memory available on current GPUs, it is impossible for deep networks to process large image size. Thus, 200×200 patches will be cropped from the input image, which will be fed into the trained detection network for testing. Since the detection network is designed to focus on small objects in high-resolution aerial images, some large objects in the original image will be missed at the original resolution. To solve this problem, an image pyramid based on the input image is created. More specific, given an input image, a smaller image is obtained by sub-sampling the input image by a factor of r along each coordinate direction. The sample procedure is repeated several times until a stop criterion is met. Patches with size of 200×200 are cropped from each of the images in the pyramid, which are employed as input to the detection to produce patch-level detection. Then, image-level detection can be obtained by applying Non-Maximum Suppression (NMS) algorithm. Furthermore, it is impossible to put all the patches from a single image into one testing batch because of the limitation of memory on current GPUs. Thus, this paper divides the patches from the same image into several batches. All the patch-level predictions will be projected back onto the image at the original scale after all the patches from the same image are processed. Then, NMS is employed to generate the final image-level predictions.

4. EXPERIMENTAL RESULTS

4.1 Implementation Details

The initial learning rate for training the detection network is set at 0.001. The learning rate is then decreased to 0.0001 after 40,000 iterations and continues training for another 30,000 iterations. A momentum is set at 0.9, and a weight decay is set at 0.0005. During testing phase, an image pyramid will be constructed with a down sampling ratio r = 0.5, until the area of the down-sampled image falls below 0.4x200. Patches are cropped from each of the images in the pyramid with a stride of s = 150 in both horizontal and vertical directions. The last part in the horizontal direction will be padded by zeros if it does not fit the patch completely. The last part in the vertical direction gets discarded if it does not make a whole patch. When evaluating the results, this paper uses a threshold at 0.5 for the confidence score and an intersection over union (IoU) at 0.5 between the predicted bounding box and ground-truth. The proposed method is implemented on a Window system machine with Intel Core i7 8700 CPU, NVIDIA GTX 1080 GPU and 8 GB of RAM. TensorFlow is adopted for implementing deep CNN frameworks.

4.2 Dataset

To evaluate and compare the performance of the proposed approach with that of other state-of-the-art approaches, this paper conducts experiments on DOTA dataset [3] and RSOD dataset [26]. DOTA is so far the largest and most diverse dataset for multi-object detection in aerial and satellite images. There are 15 object categories: plane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, large vehicle, small vehicle, helicopter, roundabout, soccer ball field and swimming pool. The DOTA dataset contains 2806 aerial images. The resolution of each image is from 800×800 to 4000×4000 pixels. The image contains objects of different scales, orientations, and shapes. This paper utilizes the horizontal annotation of DOTA to evaluate the proposed network. Furthermore, this paper divides the DOTA dataset into three types based on the size of object instance: small instance categories for helicopter, swimming pool, small vehicle, ship, and plane; medium categories for large vehicle, bridge,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

harbor, storage tank, and roundabout; large categories for soccer ball field, ground track field, basketball court, tennis court, and baseball diamond. RSOD dataset is an open dataset for object detection in remote sensing images. This dataset contains 2326 images captured by Google Earth and was divided into four classes: aircraft, overpass, oil tank, and playground. Table 1 shows the numbers of objects and images in each class.

4.3 Evaluation Metrics

The mean average precision, which has been used in many deep learning-based methods [1], [4], is adopted to evaluate the performance of the proposed object detection networks. The Precision (P), Average Precision (AP), Mean Average Precision (mAP) are defined as follows:

$$P = \frac{TP}{TP+FP} \quad (4)$$

$$AP = \frac{1}{N}\sum_{I=1}^{N} P \quad (5)$$

$$mAP = \frac{1}{M}\sum_{I=1}^{M} P \quad (6)$$

where TP represents the number of true positive samples which means the positive samples be predicted positively; FP represents the number of false positive samples which means the negative samples be predicted positively; N is the number of the samples in one class and M is the number of class. A sample is considered as correct detection if the Jaccard overlap between this sample and ground truth sample is at least 0.5.

4.4 Detection Results on DOTA Dataset

To demonstrate the effectiveness of the proposed method on small object detection in high-resolution aerial images while maintaining the effectiveness for detecting larger objects , the DOTA dataset is divided into three different groups based on the size of object instance in image: small group with helicopter, swimming pool, small vehicle, ship, and plane; medium group with large vehicle, bridge, harbor, storage tank, and roundabout and large group with soccer ball field, ground track field, basketball court, tennis court, and baseball diamond. Notably, even objects falling in the large group has relatively small size compared to the size of the original image. This paper conducts experiments on all three groups and then compares the detection results with the results of recent state-of-the-art methods, including Li et al. [5], Wang et al. [6], and Faster R-CNN [7]. Li et al. [5] proposed to detect coarse candidate regions that may contain objects at the first stage. At the second stage, fine candidate regions are cropped from coarse candidate regions, and are classified as objects or backgrounds. Wang et al. [6] used skip-connected encoder-decoder model to extract multiscale features from a full-size image. For feature maps in each scale, a visual attention network is learned, which is followed by a classification branch and a regression branch, to highlight the features from object region and suppress the cluttered background. Faster R-CNN [7] introduced a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals. The comparison of detection results is shown in Table 2. As shown in Table 2, the proposed approach achieves the best results in both small and medium group. More specific, with small group, the performance of the proposed method is improved by 19%, 3.8%, and 3.9% compared with Li et al. [5], Wang et al. [6], and Faster R-CNN [7] respectively. With medium group, the performance of the proposed method is improved by 24.6%, 3.7%, and 3.9% compared with Li et al. [5], Wang et al. [6], and Faster R-CNN [7] respectively. For large group, the proposed model achieves nearly as performance as other state-of-the-art models. It can be seen from Table 2 that the proposed method shows obvious advantages for multi-scale object detection in high-resolution aerial images. Figure 4 shows some detection results of the proposed approach on DOTA dataset. As shown in Figure 4, the proposed approach can locate exactly multi-scale objects in high-resolution aerial images. Figure 5 shows some failed detection results. When the size of objects in image is too small with heavy occlusion between objects, the proposed method cannot accurately detect them.

4.5 Detection Results on RSOD Dataset

To further evaluate the performance of proposed approach on object detection in high-resolution aerial image, this paper conducts experiments on RSOD dataset. Table 3 shows the comparison results of the proposed approach with Faster R-CNN and R-FCN [39]. R-FCN proposed position-sensitive score maps to address a dilemma between translation-invariance in image classification and translation-variance in object detection. As shown in Table 3, the proposed approach outperforms both Faster R-CNN and R-FCN. More specific, in terms of the mAP, the performance of the proposed method is improved by 9.1% and 7.9% compared with Faster R-CNN and R-FCN respectively. The results show the effectiveness

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

of the proposed method on object detection in high-resolution aerial images.

5. CONCLUSIONS

In this paper, a deep learning-based framework for addressing the object detection problem in high-resolution aerial images is presented. In particular, due to the limited memory available on current GPUs, it is hard for CNNs to process large input images. Furthermore, detect small objects from large images is still a challenging in recent years. To address the above challenges, the large input image is broken into small patches with fixed size, which are employed as input to a detection network. Moreover, since objects with large sizes may not be detected in the original resolution, an image pyramid is constructed by down-sampling the original image to make the large objects detectable by the detection network. The detection network is derived from an SSD model with a VGG-16 network as the base network, where only the first 4 convolutional stages of VGG-16 network are kept. A group of default boxes are associated with each location on the feature map to assist the detection network to produce object detection. A set of convolutional layers with a kernel size of 3×3 is employed to produce the confidence scores and coordinates of the corresponding bounding box for each of the default boxes. Experimental results on DOTA dataset and RSOD dataset, which include images containing small objects occupying only a small proportion of an image, have demonstrated the effectiveness of the proposed method in terms of alleviating the memory usage while maintaining a good object detection performance, especially for objects with small sizes. Since the proposed framework employed a sliding window strategy, it is time consuming. In the future, this paper plans to make the system more efficient.

REFERENCES:

[1] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed, "SSD: Single shot multibox detector," in *ECCV*, 2016.

[2] R. Girshick, "Fast R-CNN," in *ICCV*, 2015, pp. 1440–1448.

[3] G.-S. Xia et al., "Dota: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 1–17.

[4] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," Int. J. Comput. Vis., vol. 104, no. 2, pp. 154–171, 2013.

[5] X. Li and S. Wang, "Object detection using convolutional neural networks in a coarse-to-fine manner," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 11, pp. 2037–2041, Nov. 2017.

[6] C. Wang, X. Bai, S. Wang, J. Zhou and P. Ren, "Multiscale Visual Attention Networks for Object Detection in VHR Remote Sensing Images," in *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 2, pp. 310-314, Feb. 2019.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *NIPS*, 2015.

[9] S. Bell, C. Lawrence Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *CVPR*, 2016, pp. 2874–2883.

[10] S. Gidaris and N. Komodakis, "Object detection via a multi-region and semantic segmentation-aware cnn model," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1134–1142.

[11] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. "Scalable object detection using deep neural networks", In *CVPR*, 2014.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *ICCV*, 2016.

[13] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection snip," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3578-3587.

[14] Z. Xiao, Q. Liu, G. Tang, and X. Zhai, "Elliptic Fourier transformation-based histograms of oriented gradients for rotationally invariant object detection in remote-sensing images," *Int. J. Remote Sens.*, vol. 36, no. 2, pp. 618-644, 2014.

[15] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273-297, 1995.

[16] Y. Freund, "Boosting a weak learning algorithm by majority," *Inf. Com- put.*, vol. 121, no. 2, pp. 256-285, 1995.

[17] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," 2016,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

arXiv:1612.03144. [Online]. Available: https://arxiv.org/abs/1612.03144.

[18] Z.-Y. Shen, Z. Liu, J. G. Li, Y. Jiang, Y. Chen, and X. Xue, "DSOD: Learning deeply supervised object detectors from scratch," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1919-1927.

[19] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 580-587.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097-1105.

[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 Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325-3337, Jun. 2015.

[22] Y. Ren, C. Zhu, and S. Xiao, "Small object detection in optical remote sensing images via modified faster R-CNN," *Appl. Sci.*, vol. 8, no. 5, p. 813, 2018.

[23] G. Huang, Z. Liu, L. van der Maaten, and K.-Q. Weinberger, "Densely connected convolutional networks," 2016, arXiv:1608.06993. [Online]. Available: https://arxiv.org/abs/1608.06993.

[24] 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 Sens.*, vol. 10, no. 1, p. 132, 2018.

[25] F. Zhang, B. Du, L. Zhang, and M. Xu, "Weakly supervised learning based on coupled convolutional neural networks for aircraft detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 9, pp. 5553-5563, Sep. 2016.

[26] Y. Long, Y. Gong, Z. F. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2486-2498, May 2017.

[27] K. Li, G. Cheng, S. Bu, and X. You, "Rotation-insensitive and contextaugmented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2337-2348, Apr. 2018.

[28] Y. Zhong, X. Han, and L. Zhang, "Multi-class geospatial object detection based on a position-sensitive balancing framework for high spatial resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 138, pp. 281-294, Apr. 2018.

[29] H. Q. Qiu, H. L. Li, Q. B. Wu, and F. M. Meng, "A2RMNet: Adaptively aspect ratio multi-scale network for object detection in remote sensing images," *Remote Sens.*, vol. 11, p. 1594, Jan. 2019.

[30] H.Wu, H. Zhang, J. Zhang, and F. Xu, "Typical target detection in satellite images based on convolutional neural networks," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Oct. 2015, pp. 2956-2961.

[31] J. Ding, B. Chen, H. Liu, and M. Huang, "Convolutional neural network with data augmentation for SAR target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 3, pp. 364-368, Mar. 2016.

[32] W. Zhang, S. H. Wang, S. Thachan, J. Z. Chen, and Y. T. Qian, "Deconv R-CNN for small object detection on remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 2483-2486.

[33] I. Ševo and A. Avramović, "Convolutional Neural Network Based Automatic Object Detection on Aerial Images," in *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 5, pp. 740-744, May 2016.

[34] L. Zhang, Z. Shi, and J. Wu, "A hierarchical oil tank detector with deep surrounding features for high-resolution optical satellite imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 10, pp. 4895-4909, Oct. 2015.

[35] A. B. Salberg, "Detection of seals in remote sensing images using features extracted from deep convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2015, pp. 1893-1896.

[36] Q. Jiang, L. Cao, M. Cheng, C. Wang, and J. Li, "Deep neural networksbased vehicle detection in satellite images," in *Proc. Int. Symp. Bioelec-tron. Bioinf.*, Oct. 2015, pp. 184-187.

[37] Ibtissam, Zaaj, Brahim El Khalil Chaouki, and Lhoussaine Masmoudi, "Road extraction in a very high resolution image based on Hough transformation and local binary patterns," *Journal of Theoretical and Applied Information Technology* 91.1 (2016): 94.

[38] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao, "Orientation robust object detection in

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

aerial images using deep convolutional neural network," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 3735-3739.

[39] Dai, Jifeng, Yi Li, Kaiming He, and Jian Sun, "R-fcn: Object detection via region-based fully convolutional networks," In *Advances in neural information processing systems*, pp. 379-387. 2016.

Journal of Theoretical and Applied Information Technology
15th February 2020. Vol.98. No 03

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

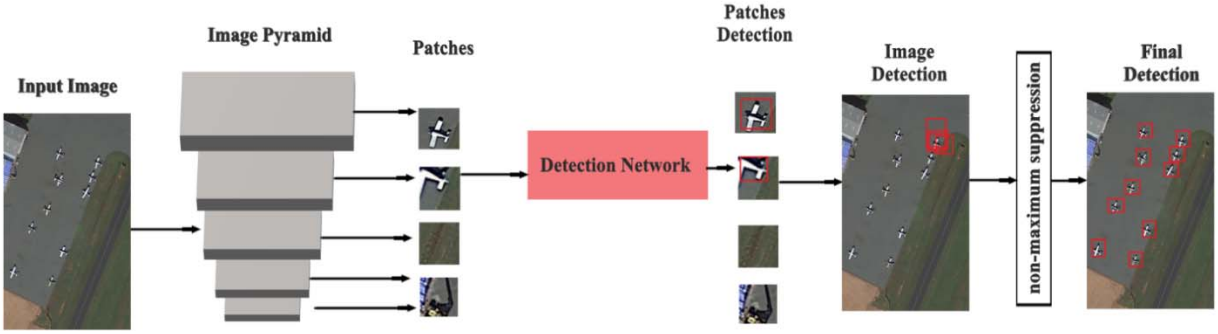

Figure 1: The Overall Architecture of The Proposed Approach.

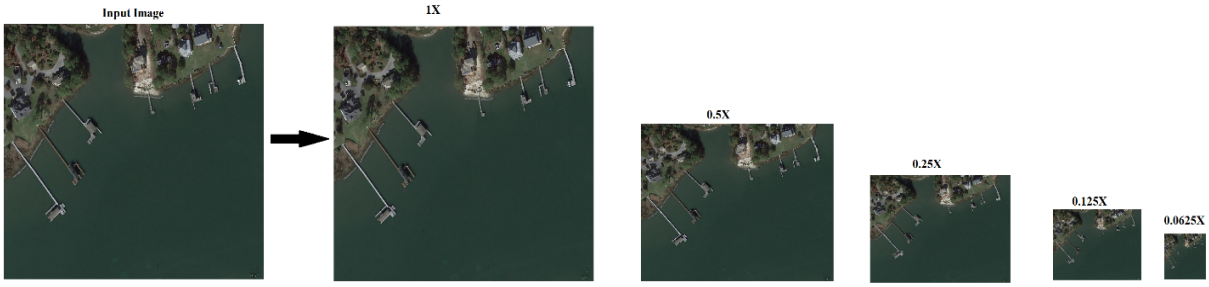

Figure 2: Image Pyramid Construction from Input Image.

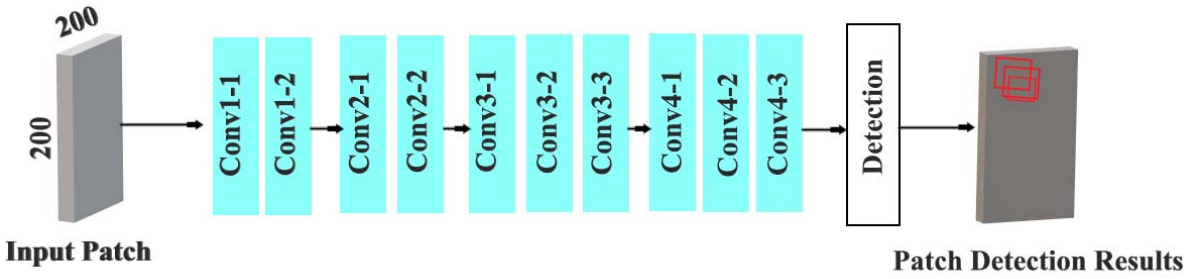

Figure 3: The Architecture of The Proposed Detection Network.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: The Numbers of Objects and Images in Each Class of RSOD Dataset.

Class	Number of Image	Number of Object
Aircraft	446	4993
Playground	189	191
Overpass	176	180
Oil tank	165	1586

Table 2: Performance Comparisons on DOTA Dataset.

Methods	Average Precision			mAP (%)
	Small (%)	Medium (%)	Large (%)	
Li et al. [5]	41.4	30.6	40.6	37.6
Wang et al. [6]	56.6	51.5	73.4	60.5
Faster R-CNN [7]	56.5	51.3	74.2	60.6
Proposed method	60.4	55.2	72.8	62.8

Table 3: Performance Comparison on RSOD Dataset.

Method	Aircraft (%)	Oil Tank (%)	Overpass (%)	Playground (%)	mAP (%)
Faster R-CNN	70.8	90.2	78.7	98.1	84.5
R-FCN	71.5	90.2	81.5	99.5	85.7
Proposed Method	86.4	95.1	93.1	99.8	93.6

Journal of Theoretical and Applied Information Technology
15th February 2020. Vol.98. No 03

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 4: Detection Results on DOTA Dataset.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

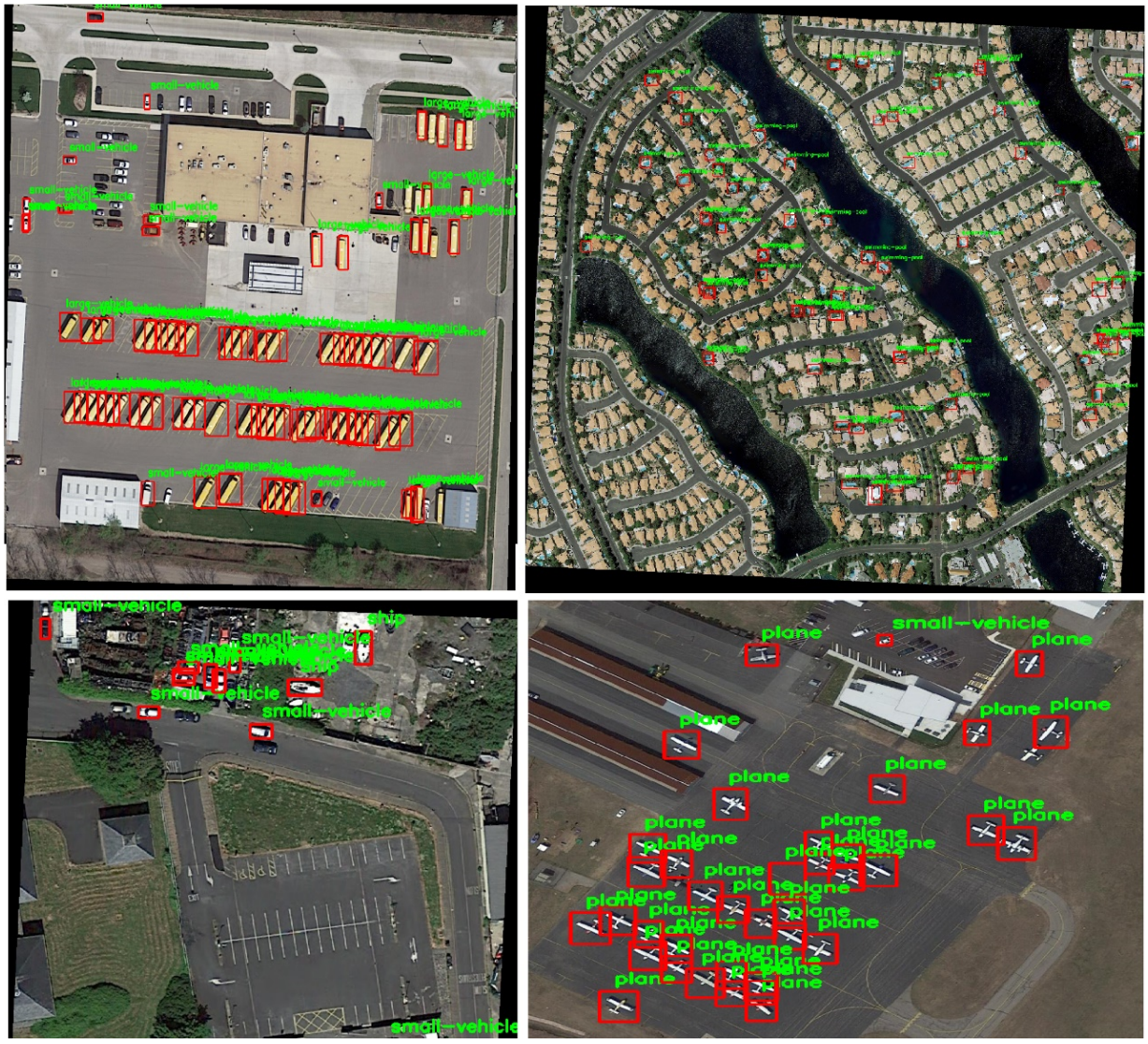

Figure 5: Examples of Failed Detection Result.