Arabian Journal for Science and Engineering
https://doi.org/10.1007/s13369-021-06288-x

RESEARCH ARTICLE-COMPUTER ENGINEERING AND COMPUTER SCIENCE

A Unified Deep Learning Framework of Multi-scale Detectors for Geo-spatial Object Detection in High-Resolution Satellite Images

Sultan Daud Khan[1] · Louai Alarabi[2] · Saleh Basalamah[3]

Received: 8 April 2021 / Accepted: 5 October 2021

Abstract
Geo-spatial object detection in high-resolution satellite images has many applications in urban planning, military applications, maritime surveillance, environment control and management. Despite the success of convolutional neural networks in object detection tasks in natural images, the current deep learning models face challenges in geo-spatial object detection in satellite images due to complex background, arbitrary views and large variations in object sizes. In this paper, we propose a framework that tackles these problems in efficient and effective way. The framework consists of two stages. The first stage generates multi-scale object proposals and the second stage classifies each proposal into different classes. The first stage utilizes feature pyramid network to obtain multi-scale feature maps and then convert each level of the pyramid into an independent multi-scale proposal generator by appending multiple region proposal networks (RPNs). We define scale range for each RPN to capture different scales of the target. The multi-scale object proposals are provided as input to the detection sub-network. We evaluate proposed framework on publicly available benchmark dataset, and from the experiment results, we demonstrate that proposed framework outperformed other reference methods

Keywords Geo-spatial object detection · Region proposal networks · Multi-scale object proposals

1 Introduction

Remote sensing technology allows the scientists and researcher to acquire information about the objects from far distance via remote sensors on satellite and unmanned aerial vehicle (UAV). With the advancement in remote sensing technology, high-resolution satellite images can be easily obtained [1,2]. Object detection in satellite imagery has become the focus of many researcher. The task of automated object detection has several applications in maritime surveillance [3], truck traffic monitoring [4], building detection [5], land cover classification [6].

The task of geo-spatial object detection is to identify the class and location of the object [7]. In this paper, our focus is on detecting man-made objects, for example, ships, aircraft, etc.). Object detection in normal images is relatively challenging than satellite images, since in later the images are captured from long distance and top-down view and quality of images are mainly affected by weather and other environmental conditions [8]. Complex and cluttered background, small object size and multiple object scales further make geo-spatial object detection a challenging problem.

Several methods have been reported in literature to detect multi-class objects in satellite images. Some of traditional methods [9–11] extract hand-crafted features, for example, SIFT [12], BoW [13], HoG [14], etc., from input image and employ machine learning algorithms to classify the objects. Although hand-crafted features work well in detecting some specific objects, however, these features demonstrate poor robustness and generalization capabilities in other object detection tasks [2], particularly in satellite images.

During recent years, deep learning models achieved tremendous performance in classification, object detection,

✉ Sultan Daud Khan
sultandaud@nutech.edu.pk

Louai Alarabi
lmarabi@uqu.edu.sa

Saleh Basalamah
smbasalamah@uqu.edu.sa

[1] Department of Computer Science, National University of Technology, Islamabad, Pakistan

[2] Department of Computer Science, Umm Al-Qura University, Makkah, Saudi Arabia

[3] Department of Computer Engineering, Umm Al-Qura University, Makkah, Saudi Arabia

Published online: 22 October 2021

semantic segmentation tasks. These deep learning models, unlike other machine learning methods, extract hierarchical features from raw images. Convolutional neural network (CNN), on the other hand, directly learns hierarchical features from 2-D raw images. The local receptive field of different layers of CNN learns different contexts and spatial relationship of the objects in the scene.

Although deep learning model has achieved tremendous success in object detection task for natural images, the performance of these methods degrades when employing to detect objects in satellite images. Small sizes of the objects and wide range of object scales are the two main causes leading to the poor performance of object detectors in satellite images. The promising solutions are efficient learning of multi-scale features and feature fusion strategy that fuses feature maps of different convolution layers, remarkably achieve performance boost. Significant amount of literature is devoted to address the scale problem by utilizing image pyramid [15,16]. In feature pyramid strategy, image is re-sized by using a scale-aware network [17] to bring all objects in a single scale and trained a single object detector. However, training a single detector to coup with all scales is hard [18]. Moreover, processing of image pyramid causes huge computational cost due to increase in memory requirement.

To detect multi-class, multi-scale object in satellite images, we propose a framework that utilizes feature pyramid network (FPN) as feature extractor. FPN extracts multi-scale feature map (feature pyramid) from an input image of arbitrary size. To detect multiple objects of different scales, instead of training a single-scale detector, we learn multiple detectors, each of which utilizes single level of feature pyramid as feature map and responsible of detection in a certain scale range. Generally, we utilize multiple region proposal networks (RPNs), each of which has its own scale range. We believe that proposed framework can be applied in other detection problems, for example, landslide detection [19] that will provide an aid to landslide risk prediction problem [20] and detection of rolling contact fatigue in rail tracks [21] for the safety and maintenance of rail tracks.

The key contributions of proposed framework are listed as follows:

1. A multi-scale and multi-class unified framework that detects objects in high-resolution satellite images.
2. The framework deals with the multi-scale problem by employing multiple RPNs, each of which has its own scale range and utilizes the independent level of the pyramid for generating scale-specific object proposals.
3. From quantitative and qualitative analysis, we exhibit that proposed framework outperforms other related methods on challenging benchmark datasets.

2 Related Work

We now review some methods related to object detection in satellite images. Traditional methods extract low-level features from input image by sliding window approach and utilizes machine learning models to classify object proposals. These low-level features includes, LBP [22], HOG [14], BoW [23] and sparse coding [24]. Support vector machine [25] successfully employed and demonstrated strong discriminative abilities in detecting different geo-spatial objects, for example, ship detection [26], air plane [27]. Similarly. AdaBoost algorithm that combines multiple weak classifiers to form a single robust classifier has been utilized to detect geo-spatial objects in [28]. k-nearest-neighbor (kNN) [29] has been used for different geo-spatial object detection and classification tasks [30,31]. Conditional random field (CRF) [32] has been used in task of urban area detection [33], building detection [34,35] and airport detection [36]. Similarly, different artificial neural networks (ANNs), for example, multilayer perceptron (MLP), extreme learning machine [37] have been utilized in several remote sensing applications, for example, ship detection [38], vehicle detection [39], tree detection [37], road detection [40], land-cover classification [41].

Aforementioned methods achieve impressive performance, however, these models rely on computation of complex hand-crafted features. These features do not have the discriminative power to classify and detect multi-class objects in satellite images with complex background. Recently, deep learning models achieve remarkable success in object detection, semantic segmentation and classification tasks in natural imagery. Moreover, deep learning models have achieved tremendous success in defect detection [42,43] and time-series classification tasks [44,45]. However, deep learning models have a long way to go to achieve high precision to detect objects in satellite images.

For object detection task, we classify deep learning models into two categories: (1) Region-based models and (2) Regression-based models [46]. Region-based methods consist of two stages, where object proposals are extracted during the first stage and obtained proposals are then classified during the second stage. Therefore, these models are also called as two-stage models. Region-based model proposed in [47] proposed a model(R-CNN) that utilizes statistical method, i.e., selective search (SS) [48] to extract multi-scale region proposal from the input image. CNN is then utilized to extract hierarchical features from each region proposal. Finally, non-maximum suppression (NMS) is employed to suppress low confidence bounding boxes and obtain the results. To automate region proposal generation process, Faster-RCNN [49] uses region proposal network (RPN) that generates object proposal through a learning process. Mask R-CNN [50] further extends and improves the performance of Faster R-CNN

by parallel branches to predict the mask and bounding box of the object simultaneously. To further enhance the speed and accuracy of Faster-RCNN, a model is proposed in [51] that accumulates the feature map of last convolutional layer by employing region of interest (RoI) layer that produces score for each ROI. Two-stage models achieve best results in detecting large objects in natural images. However, these models suffer from the following limitations: (1) Two-stage models rely on the complex process for generating object proposals. (2) The inference speed of these models is relatively slow compared to single-stage models. (3) Due to slow inference speed, these models may discard important frames and therefore, not suitable for real-time applications. (4) These models face challenges while detecting small objects and therefore not convenient for object detection in satellite images.

Single-stage deep learning models treat the object detection problem as regression problem. Popular single-stage models include You only look once (YOLO) [52], and its variants, YOLOV2 [53], YOLOv3 [54], single-shot multibox detector (SSD) [55], over-feat [56], etc. These models do not require to generate object proposals and predict the class of the object directly. These models improve the run-time speed, however, at the cost of accuracy compared to two-stage models. Apart from traditional single-stage and two-stage models, different CNN architectures have been proposed for objects detection task in satellite images [57–59]. Most recent and comprehensive survey of different models and datasets for object detection in satellite imagery can be found in [46]. A new variant of YOLO series, YOLOv4, is proposed in [60] that uses novel CSPDarknet53 classifier and works twice as fast as EfficientDet [61] with comparable detection accuracy. A spatial pyramid pooling block is added to CSPDarknet53 classifier that increases the receptive field of the network and enhances its learning capability. In contrast to YOLOv3 that uses feature pyramid network [62], YOLOv4 uses PANet [63] for object detection.

The problem with the above-mentioned deep learning models is that these models cannot be transferred to high-resolution satellite images due to difference between natural and satellite images. Therefore, to date, several attempts have been made to detect objects in high-resolution satellite images. For example, Cheng et al. [64] trained a discriminative CNNs by optimizing novel discriminative objective function to enhance the performance of scene classification in remote sensing images. A rotation-invariant CNN is proposed in [65] that enhances the performance by introducing a rotation-invariant layer in the traditional CNN architecture. A two-stage deep adaptive proposal network (DAPNet) is proposed in [66] to detect objects in satellite imagery. The network adopts new learning strategy based on category prior network (CPN) to generate suitable number of object proposal for each input image. Similarly, selective search and EdgeBoxes methods are used in [67] to generate object proposals followed by a classifier. From experiment results, the authors demonstrated that EdgeBoxes method generates high-quality object proposal compared to selective search method and achieved high recall rate. A multi-scale deep model is proposed in [68] that detects objects in satellite imagery. The network consists of two networks. The first network generates multi-scale object proposals that are provided as input to classification network that classifies proposals into different categories. Deep Boltzmann machine is employed in [69] to learning low- and mid-level features that can best describe small objects in satellite images. Then, weekly supervised learning is used to detect objects in the images. A robust pre-trained Faster-RCNN is proposed in [70] for object detection in high-resolution satellite imagery. A novel deep learning network is proposed in [71] that exploits contextual information. The network then learns local and contextual features in independent ways. A novel two-stage network is proposed in [72] for image classification and object detection in satellite images. Three stages deep network is proposed in [73] to precisely detect objects in satellite imagery. First stage of the network utilizes selective search algorithm [48] for object proposal generation. In the second stage, a pre-trained 2-D convolutional network is utilized to extract features from each candidate region, followed by the classification and object detection stage. Non-maximum suppression algorithm combine with score-based bounding box regression algorithm is used to refine the bounding boxes. The model follows traditional pipeline of RCNN [47] and suffers from the following limitations. (1) The model uses selective search that generates object proposal without using a learning process, therefore leads to the generation of in-appropriate object proposals [74]. (2) Selective search cannot efficiently handle multi-scale problem [75], therefore, the model cannot detect objects of different sizes. (3) The second stage of the model utilizes pre-trained network (i.e., AlexNet, GoogleNet) and utilizes last convolutional layers for object detection. Due to this strategy, the model is not able to detect small objects, since information of small objects in the last convolutional layers is lost. We overcome these limitations, by employing RPN that generates appropriate object proposals through a learning process. Furthermore, multiple RPNs enhance the multi-scale capability of the network to detect objects of different sizes. To date, a comprehensive survey of object detection in satellite images can be found in [46]. The survey includes the review of 110 existing state-of-the-art methods, datasets for object detection optical remote sensing images.

3 Proposed Methodology

The overall architecture of proposed framework is shown in Fig. 1. Proposed framework consists of two main compo-

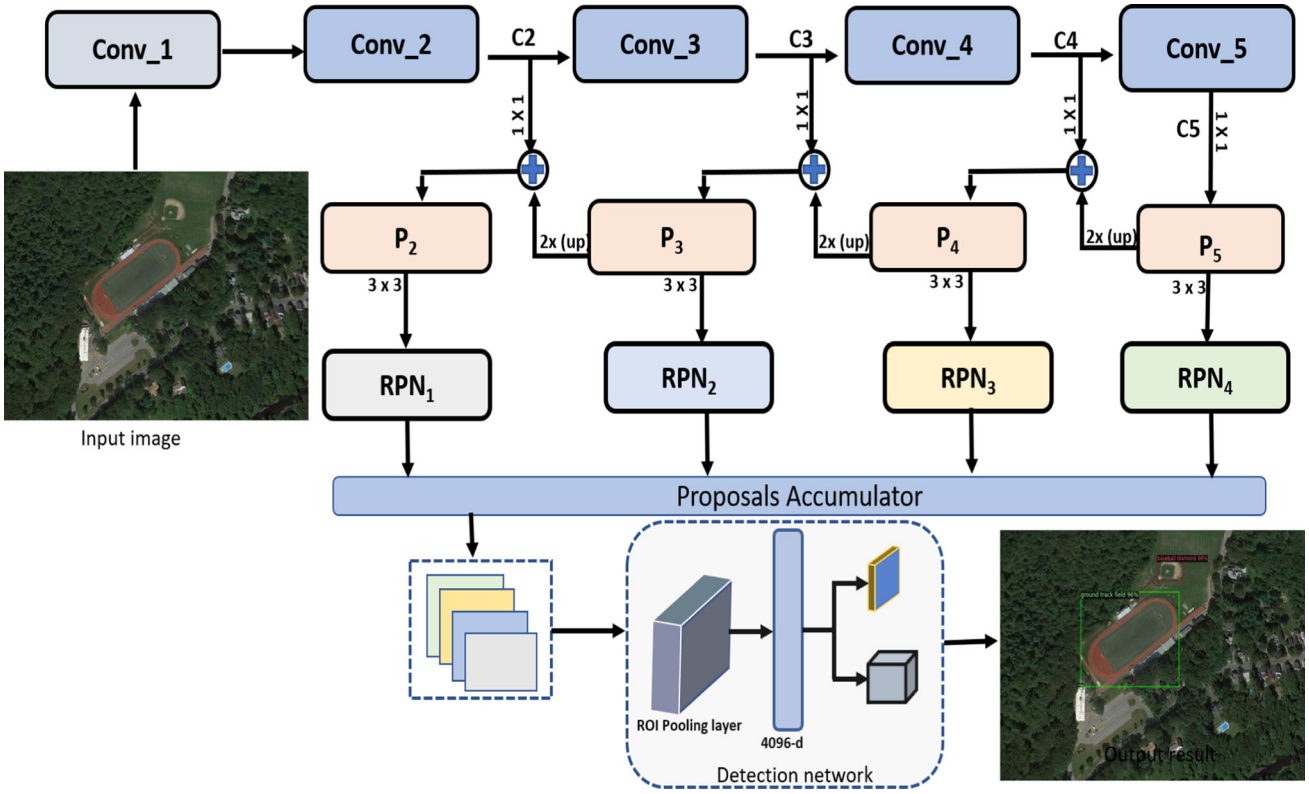

Fig. 1 Pipeline of proposed framework for object detection in satellite images

nents, namely, feature extraction and object detection part that detects multiple objects of different scales. For feature extraction, we use feature pyramid network (FPN) [62]. Region-based object detection frameworks, for example, Faster R-CNN [49] that uses VGG-16 as backbone network extracts features from the last convolutional layer. The resolution of last convolutional layer is reduced by 1/32 times of original image after passing through the network. The last convolutional layer retains the contextual information which is helpful in detecting large objects, however, it lost details of small objects. Therefore, with this configuration, it is hard for likewise region-based object detectors to detect small objects in high-resolution satellite images, where the minimum size of the object is 33×33 or even less. To tackle small object detection problem, some methods adopt fusion strategy, where feature maps from multiple convolutional layers are merged [76,77], however, ignore low-level feature maps. Generally, CNN consists of a stack of convolutional and pooling layers. The size of the input image is reduced after passing through these subsequent layers. The resolution of feature map of low-level convolutional layers is high and contains much details about location of the objects, however, lack contextual information. On the other hand, the resolution of last convolutional layer is small and contains rich contextual information due to large receptive field. Due to the small resolution, these higher layers skip important details about the small objects [78–80]. For a good detector, it is important to utilize high-resolution feature maps that capture strong contextual information.

In remote sensing images, most of objects, for example, cars, ships, airplane, etc. appear smaller than they appear in natural images while other objects, such as ground track field, tennis court and basketball court appear large in remote sensing images. This poses a challenge for a generic object detector to detect multiple objects with such large-scale variations. To detect multiple objects with different scales, we utilize Feature Pyramid Network (FPN) [62] that captures contextual information from both high and low levels of the network in the form of multi-scale feature pyramid without additional memory requirement. FPN follows the pipeline of fully convolutional neural network that takes an image of arbitrary size and outputs multi-scale feature maps. We use FPN with the backbone of ResNets [81], however, any backbone architecture can be adopted such as VGG-16 [82], AlexNet [83], DenseNet [84], etc. Generally, ResNets consists of four convolutional blocks [81]. Each convolutional block captures different contexts of the object in the input image. For example, Conv2_x contains much information about the small objects due to its small receptive field, Conv3_x captures details of the medium objects and the last two convolutional blocks, i.e., Conv4_x and Conv5_x contain details of the large objects and can capture rich contextual

information. To detect multiple objects of different sizes, we utilize all four levels of the pyramid and append multiple RPNs with different scale sets. The first convolutional block, Conv1 consists of one convolutional and a pooling layer. Conv1 applies the convolutional kernel with size of 7×7 and stride 2 superseded by a max pooling layer with filter size of 3×3 and stride 2. Each subsequent block consists of three convolutional layers with the first layer applies the convolution with kernel size of 1×1, the kernel size of second layer is 3×3 and kernel size of third convolution layer is 1×1. The size of feature map is reduced by half after passing through each convolution block. FPN utilizes the output (feature map) of each block of ResNet to generate feature pyramid. To build bottom-up pathway, FPN utilizes backbone network (ResNet) to compute feedforward computation and generates feature maps at each block of the backbone network [49]. Generally, the last convolution layer of each block of the backbone network is considered as reference feature map and defined as one level of the pyramid. The reference feature maps for 2nd, 3rd, 4th and 5th blocks are labeled as $\{C_2, C_3, C_4, C_5\}$, respectively. The resolution of each reference map C_i is half the size of previous map C_{i-1}. The feature maps of the higher/top layers have low resolution (spatially coarser) but have rich contextual information. This means that each pixel in the feature map of the top convolutional layer looks to a large window (image patch) in the input image, therefore looking into a larger context. In other words, the receptive fields of the top layers are large. Since the receptive field of top layers is large, therefore, these top layers capture much contextual information. This fact is mathematically proved in [85]. To build top-down pathway, FPN upsamples the top layers and enriches the feature maps of the top layers by using lateral connection. Lateral connection then fuses the feature maps of the top-down pathway (after up-sampled by a factor of 2) and bottom-up pathway by element-wise addition. Before merging, the channel dimension of feature map of bottom-up layer is reduced by subjecting feature map to 1×1 convolutional layer. To obtain the final map of the top-down layers, each merged feature map is subjected to 3×3 convolutional layer to remove aliasing of up-sampling operation. The final map at each pyramid level is represent by $\{P_2, P_3, P_4, P_5\}$. Now, these feature maps are spatially enhanced and semantically strong enough to detect small and multiple scale objects in high-resolution satellite images.

After building pyramid of feature maps, the next step is to generate object proposals for object detection task. We employ Region Proposal Network (RPN) for object proposals generation which is a class agnostic object detector [49], operates in sliding window fashion. Originally, RPN was used in Faster-RCNN for object detection in natural images. Although, Faster-RCNN achieved commendable performance in object detection task of large objects, however, the performance degrades when applied to satellite images. This is due to fact that RPN uses three scales [49] (128 pixels, 256 pixels and 512 pixels) and three aspect ratios (1:1, 1:2 and 2:1) which are specially designed for detecting large objects in natural images. While the scale and size of objects in satellite images are smaller than objects in natural images. Therefore, RPN in the original settings cannot be directly employed for object detection in satellite images. Some researchers adopted a straightforward way of improving the accuracy of detector by including smaller scales to the original set of RPN [49]. They have improved the result to some extent but still the accuracy is not enough.

To cover multi-scales of multiple objects in satellite images, we employ multiple RPNs instead of single RPN as shown in figure, since each level of feature pyramid captures different contextual information of the image. For example, feature map at level P_5 contains more contextual information since it is obtained from the top convolutional layer (5_{th}) of the backbone network. This feature map contains the details of large objects but important information about small object is lost. On contrary, feature map at level P_2 contains more details about the small objects with no context. From our empirical studies, we observe that a single RPN cannot provide enough region proposal [86] that covers multi-scale objects in satellite images. For generating multi-scale object proposals, we provide feature map of each pyramid level to an independent RPN. Since in our case, we have four levels of pyramid, therefore, we use four RPNs with different scale settings and aspect ratios.

After appending RPNs, each level of the pyramid then becomes an object detector that detects objects in satellite images with its own scale range. The use of multiple detectors (with different scale ranges) in a single framework effectively used for small face detection problem in [86]. We define the scale range of each detector similar to [86], however, we modified the scale ranges for each detector according to the requirement of problem. Let D_1 is first detector defined at the bottom level of pyramid P_2. We use a scale set of four different sizes, i.e., $\{8, 16, 24, 32\}$ with three aspect ratios (1:1, 1:2, 2:1).

Generally, D_1 uses feature map P_2 and generate 12 anchors at each sliding position to obtain region proposals. Similarly, detector D_2 uses P_3 feature map and scale set $\{40, 64, 90, 112\}$ with same aspect ratios (1:1, 1:2, 2:1) also yields 12 anchors. Detector D_3 and D_4 use scale set $\{140, 165, 190, 215\}$ and $\{240, 265, 290, 315\}$, respectively, with same aspect ratios. We define these scale sets by using an assumption adopted in [86]. We assume that objects with different sizes can be effectively modeled by using multiple networks with different scale sets. Using this assumption, we divide all objects (of both datasets) into four groups, i.e., small, medium, large and very large. We assume that size of small objects ranges from 8×8 pixels to 35×35 pixels. Sim-

ilarly, the size of medium objects varies from 40×40 pixels to 120×120 pixels, 125×125 pixels to 220×220 pixels for large objects and 225×225 pixels to 320×320 pixels for very large objects. We use multiple RPNs with different scale sets to effectively capture scale variations.

Each detector uses its respective feature map and generates 12 object proposal at each location of the sliding window. To detect small objects in satellite images, we generate object proposals at each location of sliding window with a stride of 1. From the empirical studies, we observe that the performance of the framework is decreased with increase in the stride [85] . Our empirical studies also validate the theoretical investigation of sliding window in [87]. Each detector will feed the object proposals generated by its respective RPN to two sibling layers, i.e., regression layer (*Reg*) and classification layer (*Cls*). The regression layer (*Reg*) predicts four outputs (x, y, w, h) for each input proposal, where x and y are the spatial coordinates, w and h are the width and height of the bounding box. The classification layer (*Cls*) assigns the score for being an object or background to each input object proposal. Proposal accumulator simply accumulates all object proposals generated by these four RPNs, and resizes all object proposal to a common size to make it fit to the input of detection network. The detection network classifies each object proposal into a desired category.

To train the RPNs, we assign labels to anchors based on Intersection-over-Union (IoU) criteria [49]. A positive label is assigned to an anchor if IoU between the anchor and ground truth bounding box is greater than 0.7. Similarly, negative label is assigned to an anchor if the IoU among the anchor and ground truth bounding box is less than 0.3. The anchors whose IoU is greater than 0.3 and less than 0.7 are discarded and are not used in training. Since, we have different set of scales at different pyramid levels, therefore, we train each RPN with its own set of training samples [86]. Let $S_1 = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$ is scale range of the detector D_1. To generate training samples D_1, we define a scale range $R = [\alpha_{\min}, \alpha_{\max}]$ and greedily samples anchors that falls within the scale range R. In the same way, we define scale range for each detector and samples anchors according to corresponding scale range. We follow the above-mentioned criteria for selecting positive and negative anchors. We then define multi-task loss as in [49] to optimize the following objective function.

$$L(c_i, b_i) = \frac{1}{M_{\text{cls}}} \sum_{i=1}^{N} L_{\text{cls}}(c_i, \hat{c_i}) + \lambda \frac{1}{M_{\text{reg}}} \sum_{j=1}^{N} L_{\text{reg}}(b_i, \hat{b_i}) \tag{1}$$

where i is the index of the anchor, N is the amount of anchors per mini-batch, $\hat{c_i}$ is the predicted score of the anchor and $\hat{c_i}$ is the ground truth label. The ground truth label $\hat{c_i}$ is 0 if the anchor is negative and 1 if the anchor is positive. b_i is the four parametrized coordinates of the predicted anchor i represented as $[x_i, y_i, w_i, h_i]^T$. $\hat{b_i}$ is the four parametrized coordinates of the ground truth bounding box i and represented as $[\hat{x_i}, \hat{y_i}, \hat{w_i}, \hat{h_i}]^T$. L_{cls} is log loss over two classes, i.e., object/background and formulated as $L_{\text{cls}} = -c_i(\log \hat{c_i})$. L_{reg} is the regression loss formulated as $L_{\text{reg}}(b_i, \hat{b_i}) \sum_{i \in \{x,y,w,h\}} L_1(\hat{b_i}, b_i)$, where L_1 is the robust loss function defined in [88]. The regression loss L_{reg} is activated only if anchor is positive and remain disable for negative anchors. The two terms in Eq. 1 is normalized by M_{cls} and M_{reg}, where λ is the balancing parameter and we keep its value to 10 to approximately balance the two terms of the equation.

We initialize the weights of all layers by Xavier initialization [89]. We set the learning rate to 0.001 and decrease the learning rate by a factor of 10 after every 10k iterations.

4 Experiment Results

To quantitatively and qualitatively evaluate the performance of the proposed framework, we conduct series of experiments. Furthermore, we perform comparison of the proposed framework with other reference methods on challenging benchmark datasets.

4.1 Benchmark Datasets

To evaluate the effectiveness of proposed framework, we use two publicly available benchmark datasets, namely, NWPU VHR-10 and DOTA.

NWPU VHR-10 dataset is proposed by Cheng et al. [90,91]. NWPU VHR-10 is challenging dataset widely used to evaluate geo-spatial object detection models. The dataset consists of multi-source and multi-resolution images that contain objects of multiple scales. The dataset contains 800 images obtained from the Google Earth Pro and Vaihingen dataset [92]. Among these 800 images, 715 obtained from the Google Earth Pro with spatial resolution ranges from 0.2 to 0.5 m. The remaining 85 images are acquired from Vaihingen dataset [92] with spatial resolution of 0.08 m. These 85 images are color infrared images. The dataset consists of 10 classes of objects with the following labels: These ten classes of objects are labeled as, airplane as labeled 1, ship is 2, storage tank is 3, baseball diamond is 4, tennis court is 5, basketball court is 6, ground track field is 7, harbor is 8, bridge is 9, and vehicle is labeled as 10. The data is divided into two sets, i.e., positive and negative set. The positive set contains 650 images with each image containing at least one object, while the negative set contains 150 images containing most of the background and does not contain any object.

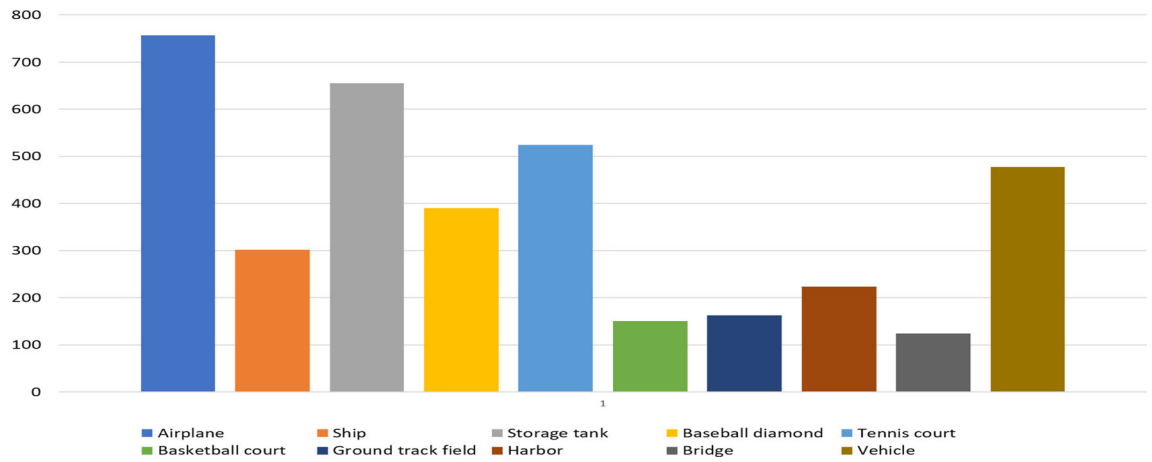

Fig. 2 Distribution of number of instances per object class. The objects are manually annotated with bounding boxes that can be utilized for training the network

Figure 2 summarizes the distribution of number of instances per each class for positive set. The negative set is used for object detection based on semi-supervised [90] and weakly supervised [93] learning which is not the focus of this paper.

For training, we followed the same convention as adopted in other methods and divide the positive image set into three splits. We kept 2% positive images for training, 20% for validation and remaining 60% for testing.

DOTA dataset recently proposed by [94] is the most challenging dataset for object detection in satellite imagery. The dataset consists of 15 categories and surpasses NWPU VHR-10 not only in number of categories but also number of samples per each category. The categories include 10 classes from NWPU VHR-10 dataset in addition to four new categories, soccer ball field, helicopter, swimming pool and roundabout. Objects are labeled as, 1: Airplane, 2: Ship, 3: Storage tank, 4: Baseball diamond, 5: Tennis court, 6: Basketball court, 7: Ground track field, 8: Harbor, 9: Bridge, 10: Vehicle, 11: Soccer ball field, 12: Helicopter, 13: Swimming pool, 14: Roundabout. The dataset contains 2806 satellite images of different resolutions, perspectives and object scales. The resolution of images ranges from 800 $\times$ 800 to 4000 $\times$ 4000 pixels. The dataset contains 188,282 of both horizontal and oriented annotations, which make it suitable for both horizontal and orientated object detection in satellite images

For training and testing, we follow the same convention as adopted in [94] and use 1403 images selected randomly for training. The remaining 1/6th and 1/3rd images are used for validation and testing, respectively.

4.2 Evaluation Metrics

To evaluate the detection accuracy of proposed framework and other reference methods, we use average precision (AP) and Precision–Recall curve (P–R curve), widely adopted evaluation metrics for object detection tasks.

Average Precision computes the average percentage of accurate detection and its value varies from 0 to 1 and formulated as $\frac{TP}{TP+FP}$, where TP is the true positive and FP is the false positive. Recall, computes the percentage of finding all the positive detection and formulated as $\frac{TP}{TP+FN}$, where FN represents the false negative. To find TP, FP and FN, we use Intersection over union (IoU). IoU is the overlap of area of predicted bounding box and ground truth bounding box. IoU measures how accurately bounding box is predicted. Generally, the location of the object in image is represented by a bounding box. IoU measures how much predicted location overlaps the ground truth location. A threshold value is defined to decide the fate of each predicted bounding box. Generally, for object detection tasks, threshold value of 0.5 is used. The predicted bounding box is regarded as TP if IoU $\geq$ 0.5, otherwise, it will be regarded as FP. After computing TP, FP and FN for all predicted bounding boxes, we then compute average precision (AP). Generally, the higher value of AP is considered as the sign of breakthrough.

Average precision uses a fixed threshold value of 0.5, therefore, it cannot measure the performance of a detector using wide range of threshold values. Generally, average precision can divide the given data into positive and negative classes and it may be useful for some applications, however, it cannot generalize the performance of a detector. Therefore, for comprehensive evaluation, we use precision–recall curve.

5 Baseline Methods for Comparison

In this section, we briefly discuss the state-of-the-art methods used for object detection in satellite images. Generally,

we divide the methods into group, i.e., hand-crafted feature-based methods and deep learning methods.

Hand-crafted feature base methods include Bag-of-Visual Words (BoW) [95], collection of part detectors (COPD) [90], SSCBoW [11] and (FDDL) [96]. Xu et al. use BoW in [95] to classify aerial images into two classes, i.e., land-use/land-cover. The model generates visual words and learns the occurrences of visual words from the training data. The authors also combine both spectral and texture features for the classification that further boosts the performance. Cheng et al. [90] proposed a multi-class object detector and classifier based on collection of part detectors, where each part detector is support vector machine that detects spatial object or recurring patterns within range of specific orientation. Sun et al. [11] proposed spatial sparse coding method based on BoW to detect complex shapes in satellite images. The model adopts sliding window approach and extracts features from each window. A novel mapping strategy is proposed to detect relative parts of the target object. Moreover, the model encodes geometric information to handle rotation variations. Han et al. [96] proposed an efficient multi-class geo-spatial object detector based on discriminative sparse coding. The authors incorporated Fisher discrimination criterion to learn the dictionary.

In addition to aforementioned methods, we also compare our results with deep learning models. These models include R-P-Faster [97], RICNN [65], D-R-FCN [98], Large-RAM [99], SSD [55], Sig-NMS [100] and YOLOv3 [54].

We evaluate and compare the performance of different methods using average precision evaluation metric and the results are reported in Table 1. We first categorize the reference methods in two categories, i.e., hand-craft feature-based models and deep learning models. Similarly, we report comparison results of different deep models on DOTA dataset in Table 3. From Tables 1 and 3, it is obvious that proposed model achieves superior performance compared to other reference methods. We also report quantitative results of different methods in terms of average precision and recall in Table 2 on NWPU VHR-10 dataset. Each average precision and recall value is computed by taking the average of precision and recall values computed over threshold range of 0–1. From Table 2, it is obvious that proposed framework achieves higher precision and recall values (Table 3).

5.1 Ablation Study

We perform ablation study to verify and analyze the effect of combining different feature maps from different convolutional layers. We evaluate the performance of the network using four different configurations. In all configurations, we used VGG-16 as backbone network. In the first configuration, *config-1* we used single feature map of C_5 and apply single RPN for generating object proposals. In the second

Table 1 Performance comparison of statistical feature-based and deep learning-based models using average precision (AP) on NWPU VHR-10 dataset

Labels	Hand-crafted feature models				Deep learning models							
	BoW	COPD	SSCBoW	FDDL	RICNN	R-P-Faster	D-R-FCN	SSD	Large-RAM	Sig-NMS	YOLO v3	Proposed
1	0.25	0.62	0.50	0.29	0.88	0.80	0.87	0.95	0.94	0.90	0.80	0.96
2	0.58	0.68	0.50	0.37	0.77	0.68	0.81	0.78	0.85	0.80	0.82	0.80
3	0.63	0.63	0.33	0.77	0.85	0.35	0.63	0.85	0.86	0.59	0.78	0.85
4	0.09	0.83	0.43	0.25	0.88	0.90	0.90	0.90	0.90	0.90	0.87	0.91
5	0.04	0.32	0.003	0.02	0.40	0.71	0.81	0.89	0.67	0.80	0.78	0.89
6	0.03	0.36	0.15	0.03	0.58	0.67	0.74	0.01	0.64	0.90	0.75	0.76
7	0.07	0.85	0.10	0.20	0.86	0.89	0.90	0.68	0.50	0.99	0.63	0.91
8	0.53	0.55	0.58	0.25	0.68	0.76	0.75	0.70	0.63	0.90	0.60	0.77
9	0.12	0.14	0.12	0.21	0.61	0.57	0.71	0.81	0.58	0.67	0.54	0.81
10	0.09	0.44	0.33	0.04	0.71	0.64	0.75	0.74	0.81	0.78	0.79	0.76
Average	0.24	0.54	0.30	0.24	0.72	0.70	0.79	0.74	0.73	0.82	0.73	0.84

Table 2 Performance comparison of statistical feature-based and deep learning-based models using average precision and recall on NWPU VHR-10 dataset

Labels		Hand-crafted feature models				Deep learning models							
		BoW	COPD	SSCBoW	FDDL	RICNN	R-P-Faster	D-R-FCN	SSD	Large-RAM	Sig-NMS	YOLO v3	Proposed
1	Precision	0.01	0.62	0.45	0.03	0.78	0.75	0.77	0.91	0.87	0.89	0.83	0.97
	Recall	0.20	0.79	0.52	0.23	0.89	0.75	0.86	0.93	0.92	0.78	0.69	0.95
2	Precision	0.37	0.69	0.51	0.34	0.65	0.62	0.71	0.79	0.79	0.79	0.78	0.87
	Recall	0.60	0.65	0.47	0.32	0.76	0.55	0.78	0.67	0.83	0.71	0.75	0.84
3	Precision	0.35	0.64	0.34	0.75	0.78	0.29	0.64	0.78	0.78	0.43	0.82	0.84
	Recall	0.42	0.79	0.29	0.72	0.82	0.32	0.45	0.84	0.84	0.57	0.65	0.85
4	Precision	0.02	0.82	0.42	0.02	0.74	0.85	0.87	0.86	0.91	0.87	0.74	0.97
	Recall	0.003	0.78	0.39	0.23	0.91	0.81	0.83	0.84	0.89	0.85	0.87	0.94
5	Precision	0.01	0.32	0.00	0.02	0.37	0.75	0.70	0.87	0.47	0.82	0.76	0.84
	Recall	0.002	0.43	0.001	0.001	0.31	0.61	0.83	0.84	0.59	0.67	0.72	0.89
6	Precision	0.00	0.35	0.20	0.002	0.47	0.72	0.69	0.00	0.65	0.92	0.75	0.91
	Recall	0.19	0.45	0.15	0.001	0.53	0.58	0.72	0.00	0.52	0.72	0.61	0.84
7	Precision	0.02	0.85	0.24	0.37	0.83	0.92	0.89	0.54	0.49	0.97	0.68	0.94
	Recall	0.19	0.58	0.01	0.19	0.87	0.81	0.92	0.69	0.54	0.98	0.40	0.95
8	Precision	0.65	0.67	0.75	0.29	0.71	0.79	0.71	0.68	0.62	0.93	0.52	0.87
	Recall	0.40	0.49	0.37	0.19	0.58	0.67	0.69	0.64	0.49	0.74	0.57	0.86
9	Precision	0.01	0.19	0.19	0.25	0.61	0.45	0.78	0.73	0.41	0.71	0.56	0.91
	Recall	0.20	0.07	0.14	0.05	0.47	0.52	0.54	0.75	0.57	0.57	0.51	0.82
10	Precision	0.07	0.04	0.45	0.001	0.69	0.67	0.61	0.62	0.77	0.81	0.79	0.92
	Recall	0.19	0.80	0.27	0.02	0.82	0.49	0.73	0.69	0.72	0.67	0.73	0.82

Average precision and recall values are obtained by changing the threshold value from 0 to 1

Table 3 Performance comparison of different methods on DOTA dataset using average precision (AP)

Labels	Faster-RCNN	R-FCN	SSD	YOLO v2	Proposed
1	0.82	0.81	0.58	0.77	0.85
2	0.50	0.49	0.24	0.52	0.56
3	0.60	0.67	0.47	0.34	0.71
4	0.77	0.59	0.33	0.34	0.79
5	0.90	0.69	0.81	0.61	0.92
6	0.75	0.52	0.25	0.48	0.76
7	0.68	0.59	0.19	0.35	0.68
8	0.62	0.45	0.14	0.36	0.65
9	0.33	0.32	0.16	0.23	0.45
10	0.54	0.50	0.05	0.39	0.55
11	0.57	0.42	0.11	0.29	0.61
12	0.42	0.34	0.00	0.11	0.44
13	0.56	0.53	0.09	0.38	0.57
14	0.50	0.51	0.31	0.36	0.52
Average	0.61	0.53	0.26	0.39	0.65

Table 4 Results of the ablation study using different configurations using average precision (AP) on NWPU VHR-10 dataset

Labels	Config-1	Config-2	Config-3	Config-4
1	0.75	0.84	0.89	0.96
2	0.62	0.69	0.75	0.80
3	0.56	0.79	0.82	0.85
4	0.81	0.85	0.89	0.91
5	0.78	0.81	0.87	0.89
6	0.53	0.64	0.69	0.76
7	0.72	0.84	0.87	0.91
8	0.62	0.70	0.76	0.77
9	0.52	0.67	0.79	0.81
10	0.58	0.69	0.74	0.76
Average	0.64	0.75	0.80	0.84

Objects are labeled as, 1: Airplane, 2: Ship, 3: Storage tank, 4: Baseball diamond, 5: Tennis court, 6: Basketball court, 7: Ground track field, 8: Harbor, 9: Bridge, 10: Vehicle

configuration, *config-2*, we utilized feature maps of C_5 and C_4 and append two RPNs for object proposals generation. In the third configuration, *config-3*, we append three RPNs to C_3, C_4 and C_5. In the fourth configuration, *config-4* (proposed framework), we append four RPNs to C_2, C_3, C_4 and C_5, respectively.

We compare the performance of all configurations on NWPU VHR-10 dataset. Since the dataset contains wide variety of classes with objects of different scales, orientations and appearances. We report the results of all configurations in Table 4. From Table 4, it can be seen that the performance improves by using multiple RPNs. We achieved best results when four RPNs are appended to feature maps of four convolutional blocks. The superior performance may be that using four RPNs can capture complementary information that can better represent the scale of the object in satellite images. On the other hand, *config-1* does not yield good results, since the model uses last convolutional block C_5 for generating object proposals. The receptive field of C_5 is large that contains rich contextual information about the large objects, however, contains no information about small objects. Therefore, *config-1* receives performance set back in detecting small objects.

Table 5 Inference time comparisons of different models

Methods	Average test time per image (in seconds)
BoW	5.32
COPD	1.06
SSBoW	40.32
FDDL	7.17
RICNN	8.77
R-P-Faster	0.005
D-R-FCN	0.20
SSD	0.06
Proposed	0.18

We compare time complexity of different methods on NWPU VHR-10 dataset and results are reported in Table 5. From Table 5, it is obvious that most of hand-crafted feature models take large inference time compared to deep learning models. Furthermore, the inference time of proposed framework is relatively slower than other deep learning models. This is due to the reason that multiple scale-specific RPNs are used for generation of multi-scale proposals. However, proposed model is faster than D-R-FCN, RICNN and achieves reasonable runtime speed with superior detection performance.

5.1.1 Discussion

From experiments results, we report the following conclusions. BoW [95] model achieves relatively low AP values compare to state-of-the-art methods. This is due to reason that BoW computes the histogram from each block of the image and then generates visual vocabulary by employing K-means clustering. Since histogram cannot retain the spatial relationship among the local features, therefore, the method cannot detect complex objects. This is obvious from the table, BoW produces lower AP values for detecting baseball diamond, tennis court, ground track filed, bridge and vehicle. Similarly, SSCBoW [11] also generates histogram from each local region of the image, however, k-means clustering is replaced by sparse coding scheme for visual encoding that results in the performance boost compared to BoW [95] model. FDDL

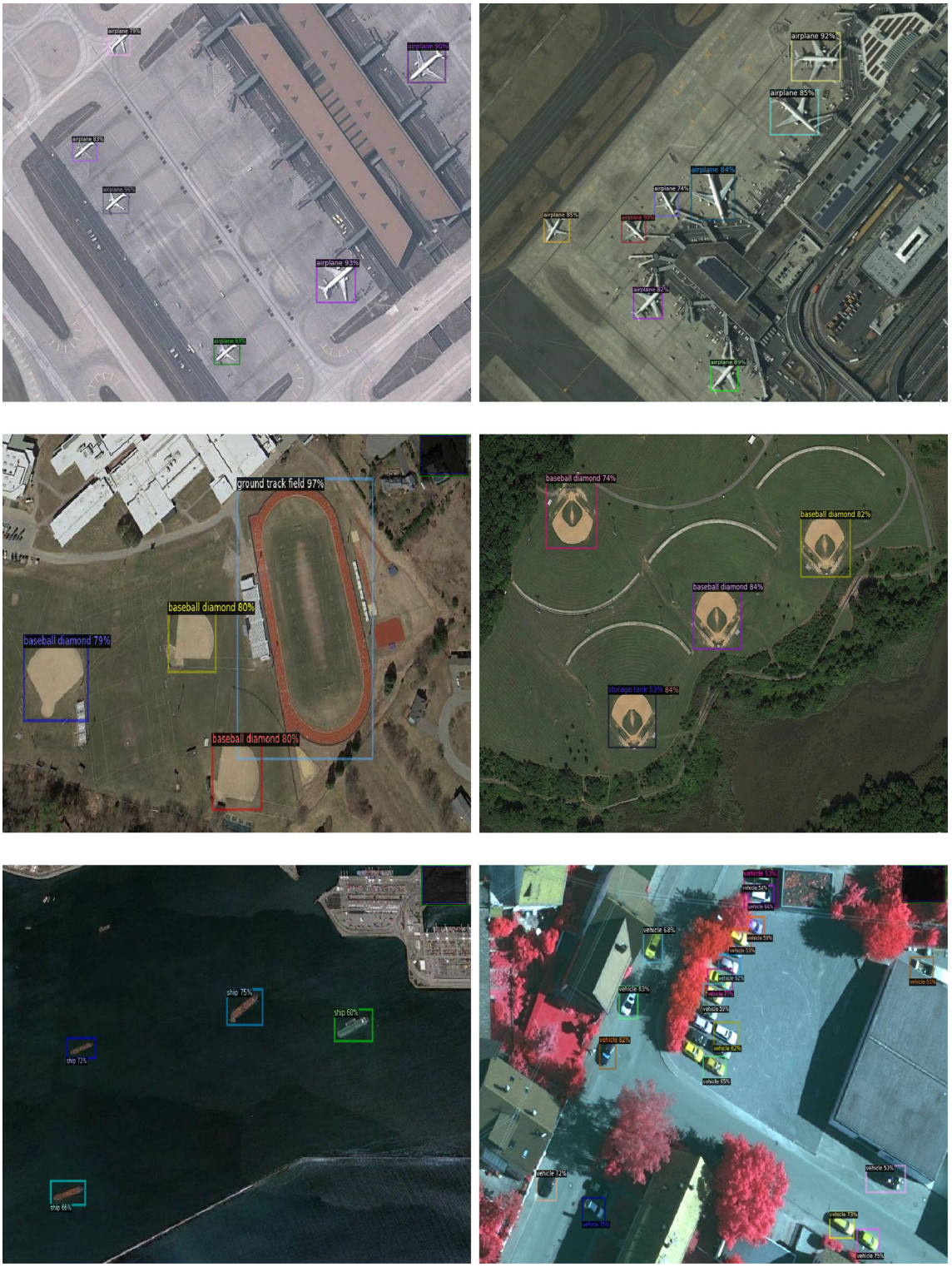

Fig. 3 Visualization of the objects detected by proposed framework using NWPU VHR-10 dataset. The figure also shows the confidence value and bounding box of each predicted object in the given image (best view zoom in)

Fig. 4 Visualization of the objects detected by proposed framework using DOTA dataset. The figure also shows the confidence value and bounding box of each predicted object in the given image (best view zoom in)

[96] adopts sparse representation for multi-class object detection. The method extracts few representative atoms from image patch. The method first reduces the size of image patch to fit the size of atoms. Important and critical details are lost due to this re-sizing step that significantly affects the detection accuracy of FDDL model. COPD [90], on the other hand, achieves superior performance compared to other hand-crafted feature base models. COPD uses collection of part detectors, where each detector classifies the class of the object with different viewpoints and then adopts iterative strategy to further refine the part detectors during training. Since COPD uses multiple part detectors to capture different viewpoints of the objects, therefore, the collection of part detectors boosts the detection accuracy by effectively detecting oriented objects.

We further discuss the performance of different deep learning models reported in Table 1. From Table 1, it is observed that R-P-Faster achieves relatively low values of AP for all ten categories. This is due to the fact that the method builds the model based on Faster-RCNN that uses last convolutional layer for object detection. The last convolutional layers have rich contextual information, however, due to large receptive field, the information about small objects is lost. SSD also produces inferior results compared to other methods. This is due to reason that On the other hand, SSD uses shallow layers to detect small objects, that SSD uses feature map of shallow layer for object detection. The feature maps of shallow layers are enriched with information about the small objects, however, cannot capture the context and information of large objects. This is also evident from Table 1, where SSD achieves good AP values for small objects, like airplane, ship, tennis court, etc., but produces lower AP values while detecting large objects, like Basketball court. RICNN [65] uses classification network, i.e., AlexNet and follows pipeline of R-CNN [47] for object detection in satellite images. Similar to R-CNN, the authors use selective search (SS) [48] method to generate fix number object proposals. However, the selective search algorithm generates object proposals based on hand-crafted feature and involves no learning process. The algorithm generates in-appropriate object proposals and cannot handle multi-scales properly. Deformable region-based fully convolutional network [98] produces comparable results.

To visual the detection performance of proposed framework, we report qualitative results in Figs. 3 and 4 on NWPU VHR-10 and DOTA datasets, respectively. From Figs. 3 and 4, it is obvious that proposed framework demonstrates good detection performance not only in detecting small objects, such as ship, vehicles and airplanes but also achieve great performance in detecting large objects like Ground track, Baseball diamond.

6 Conclusion

In this paper, we developed a framework that detects multi-class geo-spatial objects in high-resolution satellite images. Proposed framework deals multi-scale problem by employing multiple RPNs instead of a single RPN, with each RPN has its own scale range. We conduct comprehensive evaluation of the framework on challenging datasets that contain multiple objects classes. We demonstrate through experiment results that proposed framework achieves state-of-the-art performance. However, proposed framework suffers from the computational cost during training, since we are training multiple detectors in a single framework. This will be one of important consideration in future to improve the performance of framework in terms of speed and accuracy. Also we will extend the framework to incorporate oriented bounding boxes for the object with different orientations.

Acknowledgements This research is supported by National University of Technology, Islamabad, Pakistan, and Umm Al-Qura University, Makkah, Saudi Arabia.

References

1. Dong, Z.; Wang, M.; Wang, Y.; Zhu, Y.; Zhang, Z.: Object detection in high resolution remote sensing imagery based on convolutional neural networks with suitable object scale features. IEEE Trans. Geosci. Remote Sens. **58**(3), 2104–2114 (2019)
2. Aksoy, S.; Akçay, H.G.; Wassenaar, T.: Automatic mapping of linear woody vegetation features in agricultural landscapes using very high resolution imagery. IEEE Trans. Geosci. Remote Sens. **48**(1), 511–522 (2009)
3. Holsten, S.: Global maritime surveillance with satellite-based ais. In: OCEANS 2009-EUROPE, pp. 1–4. IEEE (2009)
4. Kaack, L.H.; Chen, G.H.; Morgan, M.G.: Truck traffic monitoring with satellite images. In: Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies, pp. 155–164 (2019)
5. Sirmacek, B.; Unsalan, C.: A probabilistic framework to detect buildings in aerial and satellite images. IEEE Trans. Geosci. Remote Sens. **49**(1), 211–221 (2010)
6. Kwan, C.; Ayhan, B.; Budavari, B.; Yan, L.; Perez, D.; Li, J.; Bernabe, S.; Plaza, A.: Deep learning for land cover classification using only a few bands. Remote Sens. **12**(12), 2000 (2020)
7. Zhang, X.; Zhu, K.; Chen, G.; Tan, X.; Zhang, L.; Dai, F.; Liao, P.; Gong, Y.: Geospatial object detection on high resolution remote sensing imagery based on double multi-scale feature pyramid network. Remote Sens. **11**(7), 755 (2019)
8. Chen, S.; Zhan, R.; Zhang, J.: Geospatial object detection in remote sensing imagery based on multiscale single-shot detector with activated semantics. Remote Sens. **10**(6), 820 (2018)
9. Tao, C.; Tan, Y.; Cai, H.; Tian, J.: Airport detection from large ikonos images using clustered sift keypoints and region information. IEEE Geosci. Remote Sens. Lett. **8**(1), 128–132 (2010)
10. Xiao, Z.; Liu, Q.; Tang, G.; Zhai, X.: Elliptic Fourier transformation-based histograms of oriented gradients for rotationally invariant object detection in remote-sensing images. Int. J. Remote Sens. **36**(2), 618–644 (2015)

11. Sun, H.; Sun, X.; Wang, H.; Yu, L.; Li, X.: Automatic target detection in high-resolution remote sensing images using spatial sparse coding bag-of-words model. IEEE Geosci. Remote Sens. Lett. **9**(1), 109–113 (2011)
12. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
13. Wallach, H.M.: Topic modeling: beyond bag-of-words. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 977–984 (2006)
14. Dalal, N.; Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)
15. Kim, S.-W.; Kook, H.-K.; Sun, J.-Y.; Kang, M.-C.; Ko, S.-J .: Parallel feature pyramid network for object detection. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 234–250 (2018)
16. Kong, T.; Sun, F.; Tan, C.; Liu, H.; Huang, W.: Deep feature pyramid reconfiguration for object detection. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 169–185 (2018)
17. Liu, Y.; Li, H.; Yan, J.; Wei, F.; Wang, X; Tang, X.: Recurrent scale approximation for object detection in CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 571–579 (2017)
18. Singh, B.; Davis, L.S.: An analysis of scale invariance in object detection snip. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3578–3587 (2018)
19. Piralilou, S.T.; Shahabi, H.; Jarihani, B.; Ghorbanzadeh, O.; Blaschke, T.; Gholamnia, K.; Meena, S.R.; Aryal, J.: Landslide detection using multi-scale image segmentation and different machine learning models in the higher Himalayas. Remote Sens. **11**(21), 2575 (2019)
20. Tengtrairat, N.; Woo, W.L.; Parathai, P.; Aryupong, C.; Jitsangiam, P.; Rinchumphu, D.: Automated landslide-risk prediction using web gis and machine learning models. Sensors **21**(13), 4620 (2021)
21. Chen, X.; Tian, G.Y.; Ding, S.; Ahmed, J.; Woo, W.L.: Tomographic reconstruction of rolling contact fatigues in rails using 3d eddy current pulsed thermography. IEEE Sens. J. **6**, 66 (2021)
22. Ahonen, T.; Hadid, A.; Pietikäinen, M.: Face recognition with local binary patterns. In: European Conference on Computer Vision, pp. 469–481. Springer (2004)
23. Dang, E.K.F.; Luk, R.W.P.; Allan, J.: Beyond bag-of-words: bigram-enhanced context-dependent term weights. J. Assoc. Inf. Sci. Technol. **65**(6), 1134–1148 (2014)
24. Lee, H.; Battle, A.; Raina, R.; Ng, A.Y.: Efficient sparse coding algorithms. In: Advances in Neural Information Processing Systems, pp. 801–808 (2007)
25. Inglada, J.: Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features. ISPRS J. Photogramm. Remote Sens. **62**(3), 236–248 (2007)
26. Hwang, J.; Jung, H.-S.: Automatic ship detection using the artificial neural network and support vector machine from x-band sar satellite images. Remote Sens. **10**(11), 1799 (2018)
27. Li, W.; Xiang, S.; Wang, H.; Pan, C.: Robust airplane detection in satellite images. In: 2011 18th IEEE International Conference on Image Processing, pp. 2821–2824. IEEE (2011)
28. Shi, Z.; Yu, X.; Jiang, Z.; Li, B.: Ship detection in high-resolution optical imagery based on anomaly detector and local shape feature. IEEE Trans. Geosci. Remote Sens. **52**(8), 4511–4523 (2013)
29. Cover, T.; Hart, P.: Nearest neighbor pattern classification. IEEE Trans. Inf. Theory **13**(1), 21–27 (1967)
30. Ma, L.; Crawford, M.M.; Tian, J.: Local manifold learning-based *k*-nearest-neighbor for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. **48**(11), 4099–4109 (2010)
31. Yang, J.-M.; Yu, P.-T.; Kuo, B.-C.: A nonparametric feature extraction and its application to nearest neighbor classification for hyperspectral image data. IEEE Trans. Geosci. Remote Sens. **48**(3), 1279–1293 (2009)
32. Lafferty, J.; McCallum, A.; Pereira, F.C.N.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data (2001)
33. Zhong, P.; Wang, R.: A multiple conditional random fields ensemble model for urban area detection in remote sensing optical images. IEEE Trans. Geosci. Remote Sens **45**(12), 3978–3988 (2007)
34. Li, E.; Femiani, J.; Shibiao, X.; Zhang, X.; Wonka, P.: Robust rooftop extraction from visible band images using higher order crf. IEEE Trans. Geosci. Remote Sens. **53**(8), 4483–4495 (2015)
35. Wegne, J.D.; Soergel, U.; Rosenhahn, B.: Segment-based building detection with conditional random fields. In: 2011 Joint Urban Remote Sensing Event, pp. 205–208. IEEE (2011)
36. Yao, X.; Han, J.; Guo, L.; Shuhui, B.; Liu, Z.: A coarse-to-fine model for airport detection from remote sensing images using target-oriented visual saliency and crf. Neurocomputing **164**, 162–172 (2015)
37. Malek, S.; Bazi, Y.; Alajlan, N.; AlHichri, H.; Melgani, F.: Efficient framework for palm tree detection in UAV images. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. **7**(12), 4692–4703 (2014)
38. Tang, J.; Deng, C.; Huang, G.-B.; Zhao, B.: Compressed-domain ship detection on spaceborne optical image using deep neural network and extreme learning machine. IEEE Trans. Geosci. Remote Sens. **53**(3), 1174–1185 (2014)
39. Jin, X.; Davis, C.H.: Vehicle detection from high-resolution satellite imagery using morphological shared-weight neural networks. Image Vis. Comput. **25**(9), 1422–1431 (2007)
40. Mokhtarzade, M.; Valadan, M.J.; Zoej, A.: Road detection from high-resolution satellite images using artificial neural networks. Int. J. Appl. Earth Observ. Geoinform. **9**(1), 32–40 (2007)
41. Pacifici, F.; Chini, M.; Emery, W.J.: A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification. Remote Sens. Environ. **113**(6), 1276–1292 (2009)
42. Hu, B.; Gao, B.; Woo, W.L.; Ruan, L.; Jin, J.; Yang, Y.; Yu, Y.: A lightweight spatial and temporal multi-feature fusion network for defect detection. IEEE Trans. Image Process. **30**, 472–486 (2020)
43. Ruan, L.; Gao, B.; Wu, S.; Woo, W.L.: Deftectnet: joint loss structured deep adversarial network for thermography defect detecting system. Neurocomputing **417**, 441–457 (2020)
44. David Koh, B.H.; Lim, C.L.P.; Rahimi, H.; Woo, W.L.; Gao, B.: Deep temporal convolution network for time series classification. Sensors **21**(2), 603 (2021)
45. Ircio, J.; Lojo, A.; Mori, U.; Lozano, J.A.: Mutual information based feature subset selection in multivariate time series classification. Pattern Recognit. **108**, 107525 (2020)
46. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J.: Object detection in optical remote sensing images: a survey and a new benchmark. ISPRS J. Photogramm. Remote Sens. **159**, 296–307 (2020)
47. 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, pp. 580–587 (2014)
48. Uijlings, J.R.R.; Van De Sande, K.E.A.; Gevers, T.; Smeulders, A.W.M.: Selective search for object recognition. Int. J. Comput. Vis. **104**(2), 154–171 (2013)
49. Ren, S.; He, K.; Girshick, R.; Sun, J.: Faster r-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)

50. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R.: Mask r-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
51. Dai, J.; Li, Y.; He, K.; Sun, J.: R-FCN: object detection via region-based fully convolutional networks. In: Advances in Neural Information Processing Systems, pp. 379–387 (2016)
52. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
53. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271 (2017)
54. Redmon, J.; Farhadi, A.: Yolov3: an incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
55. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S., Fu, C.-Y., Berg, A.C.: Ssd: single shot multibox detector. In: European Conference on Computer Vision, pp. 21–37. Springer (2016)
56. Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; LeCun, Y.: Overfeat: integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229 (2013)
57. Ren, Y.; Zhu, C.; Xiao, S.: Small object detection in optical remote sensing images via modified faster r-CNN. Appl. Sci. **8**(5), 813 (2018)
58. Pang, J.; Li, C.; Shi, J.; Zhihai, X.; Feng, H.: Fast tiny object detection in large-scale remote sensing images. IEEE Trans. Geosci. Remote Sens. **57**(8), 5512–5524 (2019)
59. Cheng, G.; Si, Y.; Hong, H.; Yao, X.; Guo, L.: Cross-scale feature fusion for object detection in optical remote sensing images. IEEE Geosci. Remote Sens. Lett. **6**, 66 (2020)
60. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M.: Yolov4: optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
61. Tan, M.; Pang, R.; Le, Q.V.: Efficientdet: scalable and efficient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10781–10790 (2020)
62. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
63. Wang, K.; Liew, J.H.; Zou, Y.; Zhou, D.; Feng, J.: Panet: few-shot image semantic segmentation with prototype alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9197–9206 (2019)
64. Cheng, G.; Yang, C.; Yao, X.; Guo, L.; Han, J.: When deep learning meets metric learning: remote sensing image scene classification via learning discriminative CNNs. IEEE Trans. Geosci. Remote Sens. **56**(5), 2811–2821 (2018)
65. Cheng, G.; Zhou, P.; Han, J.: Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. IEEE Trans. Geosci. Remote Sens. **54**(12), 7405–7415 (2016)
66. Cheng, L.; Liu, X.; Li, L.; Jiao, L.; Tang, X.: Deep adaptive proposal network for object detection in optical remote sensing images. arXiv preprint arXiv:1807.07327 (2018)
67. Farooq, A.; Hu, J.; Jia, X.: Efficient object proposals extraction for target detection in vhr remote sensing images. In: 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 3337–3340. IEEE (2017)
68. Guo, W.; Yang, W.; Zhang, H.; Hua, G.: Geospatial object detection in high resolution satellite images based on multi-scale convolutional neural network. Remote Sens. **10**(1), 131 (2018)
69. Han, J.; Zhang, D.; Cheng, G.; Guo, L.; Ren, J.: Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. IEEE Trans. Geosci. Remote Sens. **53**(6), 3325–3337 (2014)
70. Han, X.; Zhong, Y.; Feng, R.; Zhang, L.: Robust geospatial object detection based on pre-trained faster r-CNN framework for high spatial resolution imagery. In: 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 3353–3356. IEEE (2017)
71. Li, K.; Cheng, G.; Shuhui, B.; You, X.: Rotation-insensitive and context-augmented object detection in remote sensing images. IEEE Trans. Geosci. Remote Sens. **56**(4), 2337–2348 (2017)
72. Ševo, I.; Avramović, A.: Convolutional neural network based automatic object detection on aerial images. IEEE Geosci. Remote Sens.Lett. **13**(5), 740–744 (2016)
73. Long, Y.; Gong, Y.; Xiao, Z.; Liu, Q.: Accurate object localization in remote sensing images based on convolutional neural networks. IEEE Trans. Geosci. Remote Sens. **55**(5), 2486–2498 (2017)
74. Turner, J.T.; Gupta, K.; Morris, B.; Aha, D.W.: Keypoint density-based region proposal for fine-grained object detection and classification using regions with convolutional neural network features. arXiv preprint arXiv:1603.00502 (2016)
75. Arbeláez, P.; Pont-Tuset, J.; Barron, J.T.; Marques, F.; Malik, J.: Multiscale combinatorial grouping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 328–335 (2014)
76. Cai, Z.; Fan, Q.; Feris, R.S.; Vasconcelos, N.: A unified multi-scale deep convolutional neural network for fast object detection. In: European Conference on Computer Vision, pp. 354–370. Springer (2016)
77. Zhuang, S.; Wang, P.; Jiang, B.; Wang, G.; Wang, C.: A single shot framework with multi-scale feature fusion for geospatial object detection. Remote Sens. **11**(5), 594 (2019)
78. Sultan Daud Khan and Saleh Basalamah: Multi-scale person localization with multi-stage deep sequential framework. Int. J. Comput. Intell. Syst. **14**(1), 1217–1228 (2021)
79. Khan, S.D.; Basalamah, S.: Scale and density invariant head detection deep model for crowd counting in pedestrian crowds. Vis. Comput. **66**, 1–11 (2020)
80. Tan, X.; Xiao, Z.; Wan, Q.; Shao, W.: Scale sensitive neural network for road segmentation in high-resolution remote sensing images. IEEE Geosci. Remote Sens. Lett. **18**(3), 533–537 (2020)
81. He, K.; Zhang, X.; Ren, S.; Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
82. Simonyan, K.; Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
83. Krizhevsky, A.; Sutskever, I.; Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
84. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
85. Jie, Z.; Lu, W.F.; Sakhavi, S.; Wei, Y.; Tay, E.H.F.; Yan, S.: Object proposal generation with fully convolutional networks. IEEE Trans. Circuits Syst. Video Technol. **28**(1), 62–75 (2016)
86. Yang, S.; Xiong, Y.; Loy, C.C.; Tang, X.: Face detection through scale-friendly deep convolutional networks. arXiv preprint arXiv:1706.02863 (2017)
87. Müller, J.; Fregin, A.; Dietmayer, K.: Disparity sliding window: object proposals from disparity images. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 5777–5784. IEEE (2018)
88. Girshick, R.: Fast r-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)

89. Glorot, X.; Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256 (2010)
90. Cheng, G.; Han, J.; Zhou, P.; Guo, L.: Multi-class geospatial object detection and geographic image classification based on collection of part detectors. ISPRS J. Photogramm. Remote Sens. **98**, 119–132 (2014)
91. Cheng, G.; Han, J.: A survey on object detection in optical remote sensing images. ISPRS J. Photogramm. Remote Sens. **117**, 11–28 (2016)
92. Cramer, M.: The dgpf-test on digital airborne camera evaluation—overview and test design. Photogrammetrie Fernerkundung Geoinform. **66**(2), 73–82 (2010)
93. Zhang, D.; Han, J.; Cheng, G.; Liu, Z.; Shuhui, B.; Guo, L.: Weakly supervised learning for target detection in remote sensing images. IEEE Geosci. Remote Sens. Lett. **12**(4), 701–705 (2014)
94. Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L.: Dota: a large-scale dataset for object detection in aerial images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3974–3983 (2018)
95. Sheng, X.; Fang, T.; Li, D.; Wang, S.: Object classification of aerial images with bag-of-visual words. IEEE Geosci. Remote Sens. Lett. **7**(2), 366–370 (2009)
96. Han, J.; Zhou, P.; Zhang, D.; Cheng, G.; Guo, L.; Liu, Z.; Shuhui, B.; Jun, W.: Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding. ISPRS J. Photogramm. Remote Sens. **89**, 37–48 (2014)
97. Han, X.; Zhong, Y.; Zhang, L.: An efficient and robust integrated geospatial object detection framework for high spatial resolution remote sensing imagery. Remote Sens. **9**(7), 666 (2017)
98. Xu, Z.; Xu, X.; Lei, W.; Rui, Y.; Pu, F.: Deformable convnet with aspect ratio constrained NMS for object detection in remote sensing imagery. Remote Sens. **9**(12), 1312 (2017)
99. Zou, Z.; Shi, Z.: Random access memories: a new paradigm for target detection in high resolution aerial remote sensing images. IEEE Trans. Image Process. **27**(3), 1100–1111 (2017)
100. Dong, R.; Xu, D.; Zhao, J.; Jiao, L.; An, J.: Sig-nms-based faster r-CNN combining transfer learning for small target detection in vhr optical remote sensing imagery. IEEE Trans. Geosci. Remote Sens. **57**(11), 8534–8545 (2019)