remote sensing

MDPI

Article

Geospatial Object Detection in Remote Sensing Imagery Based on Multiscale Single-Shot Detector with Activated Semantics

Shiqi Chen, Ronghui Zhan * and Jun Zhang

Science and Technology on Automatic Target Recognition Laboratory, National University of Defense Technology, Changsha 410073, China; chenshiqi12@nudt.edu.cn (S.C.); Zhj64068@sina.com (J.Z.)

* Correspondence: zhanrh@nudt.edu.cn

Received: 2 May 2018; Accepted: 22 May 2018; Published: 24 May 2018

Abstract: Geospatial object detection from high spatial resolution (HSR) remote sensing imagery is a heated and challenging problem in the field of automatic image interpretation. Despite convolutional neural networks (CNNs) having facilitated the development in this domain, the computation efficiency under real-time application and the accurate positioning on relatively small objects in HSR images are two noticeable obstacles which have largely restricted the performance of detection methods. To tackle the above issues, we first introduce semantic segmentation-aware CNN features to activate the detection feature maps from the lowest level layer. In conjunction with this segmentation branch, another module which consists of several global activation blocks is proposed to enrich the semantic information of feature maps from higher level layers. Then, these two parts are integrated and deployed into the original single shot detection framework. Finally, we use the modified multi-scale feature maps with enriched semantics and multi-task training strategy to achieve end-to-end detection with high efficiency. Extensive experiments and comprehensive evaluations on a publicly available 10-class object detection dataset have demonstrated the superiority of the presented method.

Keywords: high spatial resolution (HSR) remote sensing images; geospatial object detection; segmentation; semantic information

1. Introduction

Geospatial object detection is one of the concerned fields in remote sensing. The development of high spatial resolution (HSR) remote sensing image sensors accelerates the acquisition of various aerial and satellite images with adequate detailed spatial structural information. These remote sensing imagery can facilitate a wide range of military and civil applications, such as marine monitoring [1], urban area detection [2,3], cargo transportation, and port management, etc. Different from natural imagery obtained on the ground from a horizontal view, HSR remote sensing imagery is obtained from a top-down view, which is an approach that can be easily affected by weather and illumination conditions. Apart from this, the small-size and scale-variable properties of multi-class geospatial objects as well as the dearth of manually annotated training samples make the detection tasks more challenging.

Many investigations related to object detection in remote sensing imagery have been carried out. The existing methods can be generally divided into four main categories: template matching-based methods, knowledge-based methods, object based image analysis (OBIA)-based methods, and machine learning-based methods [4]. Template matching-based methods [5–8] are widely applied in remote sensing field and can be further divided into two classes—rigid template matching and deformable template matching, which involve two main steps, namely, template generation and similarity measurement [9,10]. For knowledge-based methods, the prior knowledge including geometric and contextual information is

used to address a hypothesis-testing problem [11–13]. OBIA-based methods contain two main procedures: image segmentation and object classification, where the appropriate segmentation scale is the key factor influencing the performance [14]. For the machine learning methods, the following processing steps are feature extraction, feature fusion, dimension reduction, and classifier training, respectively [3,15]. Owing to the powerful techniques in machine learning area, detection tasks can be formulated as feature extraction and classification stages. The feature extraction stage relying on the proposals chosen by selective search (SS) [16] usually involves extracting handcrafted features such as scale-invariant feature transform (SIFT), histograms of oriented gradients (HOG) [17], which are widely applied in computer vision and other image related fields. Bag of words (BoW) feature represents the image of a scene by the collection of local regions by unsupervised learning and it has been widely used in the geospatial object detection with excellent performance. Sparse coding based features are learned by sets of over-completed bases to represent data with high efficiency. The subsequent classification stage mainly deals with training a classifier such as a support vector machine (SVM) [18], conditional random fields [19], k-nearest neighbors (KNN), and so on [20,21]. In addition to the uncertainty of human feature design and complex time-consuming procedures, these methods divide the object detection tasks into region proposal generation and object localization stages, which greatly influences the efficiency of the algorithm.

The popularity of deep neural networks, which are capable of hierarchical feature representation, has enabled a highly promising method for end-to-end object detection. Owing to the rapid development of large-scale public natural image datasets such as PASCAL VOC [22] and well-performing graphics processing units (GPUs), the algorithms based on convolutional neural networks (CNN) have achieved prominent success in many visual recognition tasks [23,24].

The deep learning methods for object detection can be grouped into two main streams, including the region-based methods and the region-free (regression-based) methods. Since the region-based convolutional neural network (R-CNN) [25] has made breakthroughs on the PASCAL VOC dataset, the procedure consisting of the region proposal-based extractor with a detection network has become a classical paradigm in recent years. However, these approaches such as RCNN, Spatial Pyramid Pooling (SPP-Net) [26], and Fast-RCNN [27] are still hindered by the time consumption of the proposal of generation procedures and detection procedure. Ren et al. [28], put forward a region proposal network (RPN) to substitute the typical region proposal methods, which shares convolutional features with a RPN and Fast R-CNN, and achieves end-to-end object detection. In these region-based methods, the proposed object boxes are generated and then passed to the deep convolutional neural networks for classification and location regression in the second stage. The other regression-based methods, including You Only Look Once (YOLO) [29] and Single Shot Multi-Box Detector (SSD) [30], treat object detection as a single shot problem, and directly predict bounding boxes and classification results simultaneously. In SSD, small convolutional filters are applied to each feature map to predict box offsets and category scores rather than fully connected layers in region-based methods. Additionally, SSD uses multi-representation that detect objects with different scales and aspect ratios. Multi-scale Convolutional Neural Networks (MS-CNN) [31] and Feature Pyramid Networks (FPN) [32] adopts the multi-scale feature pyramid form and fuse the output detection in the end.

CNN's amazing benchmark breaking records and presented innovative structure have continuously motivated us to explore the better solutions for small, multi-scale object detection. As for remote sensing imagery, Han et al. [33] proposed R-P-Faster R-CNN which added RPN to the original Faster R-CNN architecture and this method has achieved higher precision than other CNN-based models in the Northwestern Polytechnical University very high spatial resolution-10 (NWPU VHR-10) dataset [24,34]. Another workflow tried to address the geometric modeling problem in object detection and introduced CNN with deformable convolution layers embedded on Region-based Fully Convolutional Networks (R-FCN) [35]. To reduce the number of false positive bounding boxes in distorted aspect ratio, the authors also proposed aspect ratio constrained non-maximum suppression (NMS) to eradicate false results and improve precision. The method proposed by Kang in [36] also concatenate multi-layer features of each region proposal using region of interest (ROI) pooling before

predicting final results in ship detection tasks. Although these region proposal based methods have achieved promising performance and fit the geometric properties of VHR targets well, they have poor adaptability in complex practical scenarios and are somewhat time consuming both in training and inference stage.

To remedy the limit of the two-stage frameworks, we mainly focus on the improvement of regression-based detection. In order to cover the shortage of small-sized object detection, experiments have been conducted by making combination of different layers in CNN. Deconvolution single shot detector (DSSD) [37] introduces additional context into SSD via deconvolution to improve the accuracy. Deeply supervised object detector (DSOD) [38] designs an efficient architecture and a set of principles to learn object detectors from scratch, also following the framework of SSD. Furthermore, GRP-DSOD [39], a network which concatenates high-level semantic features and low-level spatial features in a single pyramid is proposed by Shen et al. Some researchers [40] even aim to address the extreme class imbalance problem by re-designing the loss function or classification strategies. Although the progress made by one-stage detectors are inspiring, their ability for accurate positioning on multi-scale objects and small objects still has a large room for improvement. Some investigations such as [41,42] have leveraged scene context information to help object detection. Zhang et al. [43] use semantic segmentation-aware CNN features to strengthen detection features by activation function for natural image challenge object detection. Nevertheless, this approach only considers learning relationship between channels and object classes in a channel-level attention while the scale information is too restricted to enrich the semantics at higher level layers.

Based on a backbone network that generates a low level detection feature map, semantic information from features of different level are learned in a hierarchical manner. Smaller objects are detected by lower layers, while larger ones are detected by higher layers. However, the feature of small objects generated by shallow layers only capture elementary visual patterns without enough semantic information. This may affect the performance on small object detection, and the quality of coarse features from top layers would also be impaired by the inadequate fine features from shallow layers.

To tackle the problem of restricted detection speed as well as inaccurate positioning especially for small objects, a novel end-to-end single shot detection framework, namely, recurrent detection with activated semantics (RDAS), is presented here for multi-class geospatial object detection from HSR remote sensing imagery. Semantic augmentation is divided into two stages and implemented during the whole network structure. In the first stage, a low-level detection feature map is enriched with strong semantic meaningful information supervised by bounding-box level segmentation ground-truth. This can be interpreted as an attention mechanism, where each channel of original low-level features is activated by a semantic attention map. For higher level detection feature maps, a refined location module which consists of several identical blocks is employed to enrich semantic information again and prune out unnecessary locating information in the second stage. By seamlessly incorporated into the SSD framework, which elegantly customizes it towards fast, accurate, single shot object detection, RDAS results in a forceful object detector that works reliably on multi-scale and multi-orientation objects with single input resolution.

The main contributions of this paper can be summarized as follows:

(a) Different from previous two-stage detectors for geospatial object detection in HSR images, we build a new regression-based detection framework which can play a crucial part in complex scenarios, detect small sized objects or objects with extremely different scales, and reduce repetitive detection regions of densely arranged targets.

(b) We improve the traditional single shot detectors by augmented semantic information. Our thought is achieved by developing a semantic segmentation branch to strength low-level spatial features with more meaningful high-level semantic features, which ensures the effectiveness of segmentation information and can be interpreted as an attention mechanism. This object regional attention greatly suppresses background interference in convolutional features, which turns out to reduce false detections as well as highlight challenging object patterns.

(c) We adopt gating functions to control message transmission and enhance the representation power of modules throughout the networks, which integrates attention from local and global view.

The proposed method with enriched semantics is evaluated on a publicly available remote sensing object detection dataset and then compared with current state-of-the-art approaches. The experimental results confirm our elaborate design and demonstrate the effectiveness and efficiency of our method. With this method, objects in VHR remote sensing images will be better understood by RDAS with less time consumption and more condensed computation resources.

The rest of this paper is organized as follows. Section 2 describes the details of the proposed method. Section 3 presents the experiments conducted on NWPU VHR-10 dataset to validate the effectiveness of the modified framework. Section 4 presents the analysis of the experimental results and the discussion of the results, respectively. Finally, Section 5 concludes this paper.

2. Proposed Method

Figure 1 shows the overall architecture of RDAS. The framework is mainly composed of two components, namely segmentation branch and the main detection branch for prediction upon multi-scale response maps. The backbone VGG16 used in typical single shot detector acts as a sub-network for detection in RDAS and it is represented in the upper part of Figure 1. The convolutional architecture of the SSD is extended from the 16-layer VGGNet, by replacing the fully-connected (FC) layers fc6, fc7 with several convolutional layers and changing pool5 from 2 × 2-*s*2 to 3 × 3-*s*1. As represented in Figure 1, conv4_3 denotes the third convolutional operation in the fourth convolutional block and other identifiers mean similarly, while fc7 is the last layer of VGG16 because we remove all the dropout layers and the fc8 layer. Since the early network layers are based on VGG16 used for image classification (truncated before classification layers), the convolutional feature maps from conv6_2 to conv9_2 are added as auxiliary structure on top of the base network. The conv4_3 up to conv9_2 indicates the detection source layers. To augment the low level detection feature map with strengthened semantic information, a segmentation module is introduced as shown in the lower left part of Figure 1. For the high level feature maps, we explore an alternate direction of recalibrating the feature maps adaptively, to boost meaningful features with activated semantics, while suppressing weak ones. In the lower right part of Figure 1, a refined location module which contains several identical blocks called concurrent global attention blocks (CABs), is presented to strengthen the useful feature maps from the main detection branch. Both segmentation module and refined location module play an indispensable role of enriching semantic information and multi-scale information incorporation is achieved in a recurrent way. The rest of this section describes the details of the proposed method and analyzes the motivation behind our design.

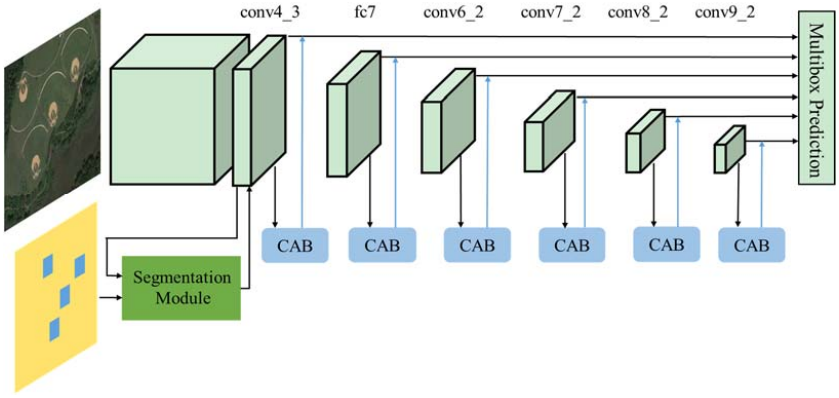


Figure 1. Architecture of the newly presented Recurrent Detection with Activated Semantics. CAB blocks in the lower part of Figure denote concurrent global attention blocks.

2.1. Semantic Enrichment at Low Level Layer

In a typical CNN model, convolutional features in a lower layer often focus on local image details, while the features from deeper layer generally capture more abstracted information. The segmentation branch was introduced to enrich sematic information at low-level detection feature layer, which was achieved in a weakly supervised way. The low-level detection layer conv4_3 from original SSD and segmentation ground truth from bounding-box level were taken as inputs, and a semantic oriented feature map with the same dimension as the ground truth was generated.

Our segmentation module, identified as a green rectangular in Figure 1, was able to automatically learn rough spatial regions of objects from the convolutional features. This semantic segmentation to object was then directly encoded back into the convolutional features, where target-related features are strongly strengthened by suppressing background interference in the convolutional maps, as shown in Figure 2.

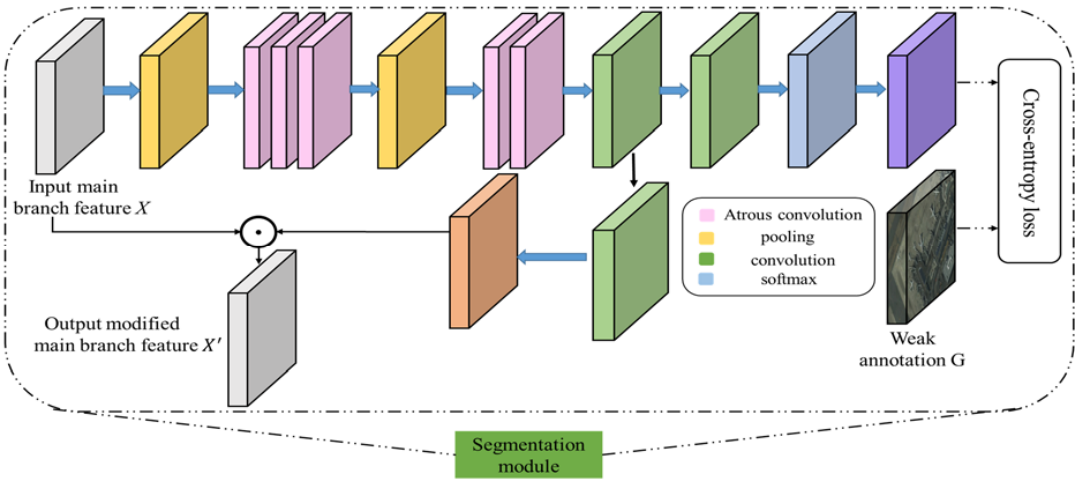


Figure 2. The specific framework of segmentation branch attached to low level feature map. The intermediate feature map from object detection branch (e.g., conv4_3 for SSD300) acts as the input, which generates a semantically meaningful feature map to activate the original feature map and the output is then used in the main branch.

Since atrous convolution [44] allows us to explicitly control the resolution at which features responses are computed within CNNs, the dense prediction tasks such as semantic image segmentation usually highlight this atrous convolution. We deployed this powerful operator in the segmentation branch. It allowed us to effectively enlarge the field of visualization of filters to incorporate large context without increasing the number of parameters or the amount of computation. Specifically, we added five atrous convolutional layers with 3×3 kernel size after the input feature map X from main branch. The first three atrous convolutional layers had a dilation rate of 2 and the last two atrous convolutional layers had a dilation of 4. After that, another 1×1 convolutional layer was deployed to generate the intermediate feature map denoted as $G(X)$. This feature map was used to generate segmentation prediction and provide high semantic information to strengthen the input feature map from main branch. To this end, there were two paths linked to $G(X)$. The first path included a 1×1 convolution layer with N+1 output channels followed with a softmax layer to generate the segmentation prediction Y. The second path took another 1×1 convolutional layer to generate a semantic meaningful feature map Z, which enjoys the same output channel number with that of X. After that, we activated the feature map from main branch by element-wise multiplication. All these layers were elaborately designed to ensure the size of feature maps was invariable.

As shown in Figure 2, we gave a mathematically definition of the segmentation branch. Let $X \in \mathbb{R}^{C \times H \times W}$ be the low level detection feature map from the main detection branch, C,H,W represents the channel number,

the height and width of the feature map respectively. The weak annotation $G \in \{0,1,2,\ldots N\}^{H\times W}$ means the segmentation ground truth where N is the number of classes in the dataset. The segmentation branch predicts the per-pixel class assignment where:

$$Y = F(G(X)) \tag{1}$$

satisfying:

$$\sum_{c=0}^{N} Y_{c,h,w} = 1, Y \in [0,1]^{(N+1)\times H\times W} \tag{2}$$

$G(X) \in \mathbb{R}^{C'\times H\times W}$ is the intermediate result which will be further used to generate the output local activation feature map Z:

$$Z = H(G(X)) \in \mathbb{R}^{C\times H\times W} \tag{3}$$

This sematic meaningful feature map Z is then used to strengthen the original low level detection feature map X by element-wise multiplication: $X' = X \odot Z = X \odot H(G(X))$.

In this way, the semantically strengthened low level detection feature map will contain both basic visual patterns and high level semantic information. The proposed segmentation branch is formulated in an unified framework which is trained end to end by allowing for computation back propagations through all layers.

As for the problem of generating segmentation ground-truth given the object bounding boxes, we denoted this segmentation ground-truth as SG and ensure it has the same resolution as the input of segmentation branch. This strategy guarantees that there is only one class to be distributed to each pixel in SG. For a pixel SG_{hw} locates within a bounding-box on an image patch I, we assigned the label of the bounding box to SG_{hw}. If this pixel was located in more than one bounding boxes, the label of the bounding box with the smallest size was chosen as the class of the pixel. Otherwise, if it did not locate in any bounding box, we assigned it to the background class.

2.2. Semantic Activation at Low Level Layer

Motivated by the goal of guaranteeing that our detection network could adaptively filtrate the representative scales for objects with different sizes, we proposed to introduce a two-level attention mechanism and broadcasted identity mapping principle to enrich the semantic information of each prediction layer. This ingenious design was partially inspired by SENets [45], which showed the best performance in the ILSVRC 2017 classification challenge. In this way, the useful features in a suitable resolution would be intensified and less useful ones would be weakened correspondingly. The block in the refined location module called CAB, consisted of two main stages that could be seen as an effective gating mechanism which combined attention both locally and globally. Figure 3 shows a diagram of our structure.

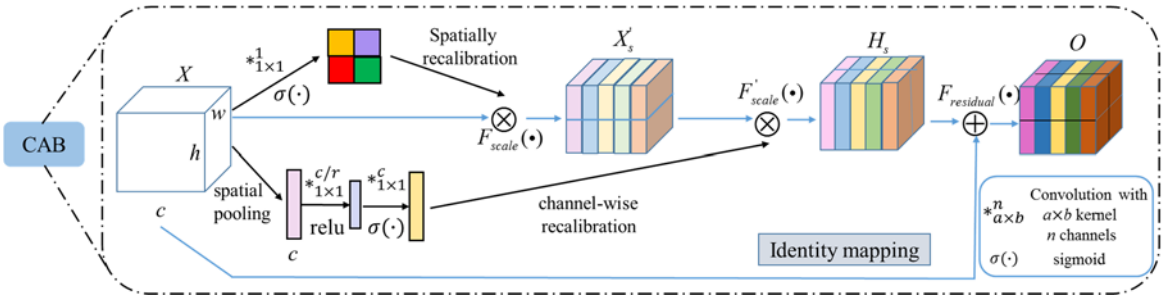


Figure 3. The structure of concurrent attention block used to enrich the semantic information at low level layer.

As the previously introduced SE block only excited channel-wise, which proved to be effective for classification, we introduced variants of SE blocks for object detection. Considering the pixel-wise

spatial information is more representative for object detection, we introduced another SE block, which squeezed along the channels and excited spatially. Finally, we proposed the introduction of concurrent attention blocks that recalibrated the feature maps to be informative both channel-wise and spatially. The two SE blocks were integrated into CAB and we explained it in two parts: concurrent attention with channel-level, and global-level and identity mapping.

For the upper branch part in Figure 3, the Squeeze and Excitation block [45] was applied in our structure as channel level attention, which consisted of: (i) a spatial pooling stage F_{sp} for global information embedding; and (ii) an excitation stage F_{ex} for channel-wise recalibration. In this way the channel-level will output feature map as:

$$X' = F_{ex}(F_{sp}(X)) \tag{4}$$

Formally, given the input feature map $X \in \mathbb{R}^{C\times H\times W}$, the spatial pooling stage will produce a vector $Z \in \mathbb{R}^{C}$ calculated by:

$$z_c = F_{sp}(x_c) = \frac{1}{w \times h}\sum_{i=1}^{w}\sum_{j=1}^{h} x_c(i,j) \tag{5}$$

where z_c is the c-th element of z.

The excitation stage is composed of two fully-connected layers with different dimensions of output plus a sigmoid activation.

This can be seen as a gating mechanism to limit the model complexity and aid generalization, so a dimensionality-reduction layer with parameters W_1 with reduction ratio r and a dimensionality-increasing layer with parameters W_2 are produced. Then, we can express the excitation stage as:

$$e = F_{ex}(z) = sigmoid(W_2\delta(W_1 z)) \tag{6}$$

where δ refers to the ReLU function, $W_1 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $W_2 \in \mathbb{R}^{C\times\frac{C}{r}}$. Here, we choose $r = 16$.

Then, we can express:

$$X'_s = F_{scale}(X') = e_c \cdot x_c = e \otimes X \tag{7}$$

where $\otimes$ denotes channel-wise multiplication.

While our global attention in the bottom branch in Figure 3 takes the feature map X as the input, and different from the upper branch, we squeeze the feature map along the channels and excites spatially. The modified excitation stage can be denoted as:

Here, we considered an arbitrary slicing of the input tensor $X_c = [x^{1,1}, x^{1,2}, \ldots, x^{i,j}, \ldots, x^{H,W}]$, where $x^{i,j} \in \mathbb{R}^{1\times1\times C}$ corresponding to the spatial location (i,j) with $i \in \{1,2,\ldots,H\}$ and $j \in \{1,2,\ldots,W\}$. The spatial squeeze operation is achieved by a convolution with weight $W_3 \in \mathbb{R}^{1\times1\times C\times1}$, generating a projected tensor $q \in \mathbb{R}^{H\times W}$. Each element of the projection represents the linearly combined representation for all channesl C for a global location (i,j) in the channel-wise attentioned feature map X'_s. This projection can then be passed through a sigmoid layer to rescale activations to [0,1]. Finally, the global-level attention used to excite X'_s can be calculated by:

$$H_s = F'_{\text{scale}}(X'_s) = sigmoid(W_3 X'_s) \tag{8}$$

In the stage of broadcasted identity mapping, addition operation [46] $F_{residual}$ is used to obtain the final outputs:

$$X_O = F_{residual}(X, H_s) = X \oplus H_s \tag{9}$$

where $\oplus$ denotes the element-wise addition.

The final combined feature map X_O contains both fundamental visual patterns and high level semantic information. All these layer are well-designed to guarantee the size of feature maps unchanged.

Since the semantic information from higher level detection layers was already inherited from previous layers, it was not necessary to adopt the segmentation branch for higher layers. Moreover, the small resolution of coarse features made it harder to do the segmentation task on them. Due to the above reasons, we employed an unambiguous refined location module, to strengthen the semantic information of feature maps from fc7 to conv9_2 in a recurrent fashion.

The refined location module was shown in the right lower part of Figure 1, which included several identical global attention blocks. Those blocks were attached at each object detection source layers in the detection branch. As illustrated in Figure 3, concurrent attention blocks can learn the relationship between channels and object classes, by recalibrating the input feature map both spatial and channel-wise information. In this way, the refined location module encouraged the network to learn more representaive feature maps, thus making the final location of bounding boxes more accurate.

2.3. Multi-Task Training

The modified feature maps obtained by semantic activation can be fused to train our newly presented detector. Similar to the principle in SSD, two 3×3 convolutional layers are added on top of the feature maps of detection branch to obtain the classification scores and bounding box results. The original object detection loss function L_{det} can be represented as:

$$L_{\text{det}}(\{c_i\}, \{t_i\}) = \sum_i L_m(c_i, l_i^*) + \sum_i [l_i^* \geq 1] L_r(t_i, g_i^*) \tag{10}$$

where i is the index of default box in a mini-batch, and c_i and t_i are the predicted object category and coordinates of the bounding box being processed by CAB. l_i^* means the ground truth class label of default box i, g_i^* is the ground truth location and size of the default box. The multi-class classification loss L_m means the softmax loss over multiple classes confidences and the smooth L_1 loss L_r represents the regression loss similar to that in Fast R-CNN [27]. The Iverson bracket indicator function $[l_i^* \geq 1]$ outputs 1 when the condition is true, and 0 otherwise.

For segmentation branch, we introduce another cross-entropy loss function L_{seg} which is denoted as:

$$L_{seg}(I, G^s) = -\frac{1}{HW} \sum_{h,w} \log(Y_{G^s{}_{h,w},h,w}) \tag{11}$$

where $Y \in [0,1]^{(N+1)\times H\times W}$ is the segmentation prediction.

By adding the new segmentation loss in conjunction with the original detection loss function, the final object function we are optimizing is defined as:

$$L(I, G^s, c_i, t_i) = L_{\text{det}}(\{c_i\}, \{t_i\}) + \alpha L_{\text{seg}}(I, G^s) \tag{12}$$

where α is a parameter to tradeoff between the task of segmentation and detection.

Current one-stage methods rely on one-step regression and use various feature maps with different scales to predict the locations and sizes of objects, which is rather inaccurate in some challenging scenarios especially for small scaled objects. To this end, we supposed to use semantic information and then modify the original feature layers in the detection branch. That is, CAB provided better initialization for the final regression. Specifically, we associated n anchor boxes with each regularly divided cell on the feature map. The initial position of each anchor box relatively to its corresponding cell was fixed. Processed by CAB, the corresponding feature maps were passed to generate specific object categories and accurate object locations and sizes. This procedure was similar to the default boxes used in SSD. Instead of directly using the regularly tiled default boxes for detection, CAB takes the initial anchor boxes as input for further detection, leading to more precise detection results.

3. Dataset and Experimental Settings

3.1. Dataset Description

Nowadays, many datasets are published for researchers to compare the performance of various deep learning based methods and conduct further investigations in remote sensing imagery. However, obvious drawbacks do exist. On the one hand, the annotations of the objects in these datasets are difficult to obtain, which restricts the potential of CNN owing to the requirements of large-scale training samples. Furthermore, datasets are constrained as certain types of objects such as airplanes, ships or vehicles, but the diversified benchmark for remote sensing imagery has not been established.

In our work, the performance of the proposed approach was tested on a multi-class object detection benchmark: NWPU VHR-10 dataset proposed in prior studies [33,34]. The NWPU VHR-10 dataset [47] contains multi-source and multi-resolution object detection imagery, which not only includes optical remote sensing images, but also includes pan-shaped color infrared images. In addition, there are 800 optical remote sensing images, where 715 images were obtained from Google Earth Pro with spatial resolutions from 0.5 m to 2.0 m, and 85 pan-sharpened color infrared images were acquired from the Vaihingen data with a 0.08 m spatial resolution. Meanwhile, the NWPU VHR-10 dataset contains 650 annotated images, with each image containing at least one target to be recognized. The comprehensive object types covered airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle. For the positive image set in VOC 2007 formula, 757 airplanes, 302 ships, 655 storage tanks, 390 baseball diamonds, 524 tennis courts, 150 basketball courts, 163 ground track fields, 224 harbors, 124 bridges, and 477 vehicles were manually annotated with rectangular bounding boxes, and were utilized as the ground truth. For the fair comparison with baseline method, the split ratios of the training, validation and testing dataset were set to 20%, 20%, and 60%, respectively. Then, 130, 130, and 390 images from the positive section in the NWPU VHR-10 dataset were randomly selected to fill the three subsets.

3.2. Evaluation Indicators

To quantitatively evaluate the performance of various object detectors, we applied the widely utilized evaluation indicators of average precision (AP) and precision-recall curve (PRC), which are prevalent and extensively adopted for the object detection framework.

3.2.1. Average Precision

As usually defined, the AP computes the average value of the precision over the interval ranging from recall = 0 to 1, also known as the area under the PRC. Mean AP (mAP) computes the average value of AP over all object categories. Since most papers recognize higher AP as the sign of benchmark breakthrough, the higher the AP, the better the performance.

3.2.2. Precision-Recall Curve

PRC is calculated from four frequently used evaluation components in information retrieval, true positive (TP), false positive (FP), true negative (TN) and false negative (FN). The precision indicator measures the proportion of detections that are TP, while the recall indicator generates the the percentage of positives that are correctly identified. On the basis of these four components, the precision and recall indicators are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

If the area overlap ratio between the predicted bounding box and the ground-truth bounding box is larger than 0.5, the proposed detection map will be considered to be a TP; otherwise, it will be

determined as a *FP*. Additionally, if several proposals overlap with the same ground-truth bounding box, only the one with maximum overlap is considered as a *TP*, and the others are considered as *FN*. Generally speaking, the value of precision and recall are inversely related.

3.3. Baseline Method

The performance of the proposed algorithm was compared with both the handcrafted feature based methods of the BoW feature [48], spatial sparse coding BoW (SSCBoW) [49], Fisher discrimination dictionary learning (FDDL) [50], the collection of part detectors (COPD) [51] method, and the deep learning feature based methods such as transferred CNN [34], newly trained CNN [34], rotation-invariant CNN (RICNN) with or without fine-tuning [34], R-P-Faster R-CNN proposed in [33] and Multi-Scale CNN proposed in [52].

3.4. Implementation Details

Single-Shot detectors integrated with semantics and comparison models were implemented using the open source Caffe framework [53] and executed on a 64-bit Ubuntu 16.04 computer with CPU Intel(R) Core(TM) i7-6770K @4.00 GHz × 8 and GeForce GTX1080 GPU with 8 GB memory CUDA8.0 cuDNN5.0.

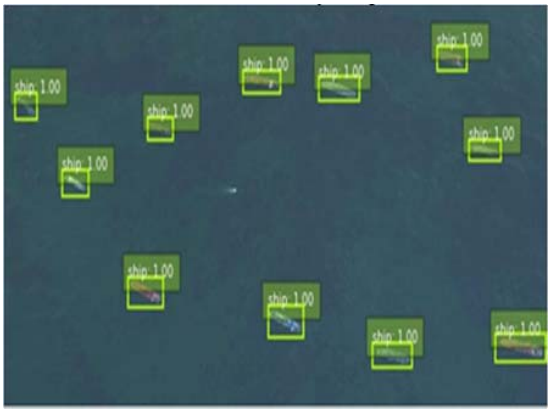
Our detection network was trained end-to-end by using the mini-batch stochastic gradient descent algorithm, where the momentum was fixed to be 0.9 and the weight decay was set to be 0.0005.

To validate the effectiveness of our proposed network, we adopted the backbone network VGG16 to ensure a reasonable comparison as applied in [33,34] for NWPU object detection. The pre-trained SSD model is utilized to initialize the parameters of conv5_1, conv5_2, conv5_3, fc6 and fc7 in the main detection branch while the rest layers of the segmentation branch are initialized by Xavier initialization [54]. For the SSD model, the parameters were set the same as the proposed method. The learning rate (*lr*) is initialized to 0.0005, with a step strategy of $\gamma = 0.2$ and the step size N = 25,000. The total iteration number of the proposed method was set 75000. For both RDAS with or without segmentation branches, we used aspect ratios of the default boxes as: [2], [2,3], [2,3], [2,3], [2], [2] for multi-scales of predictions.

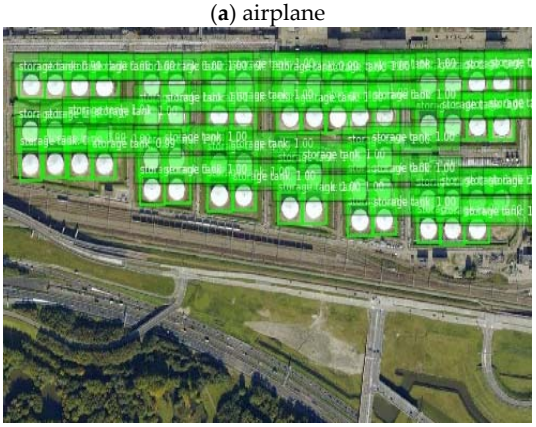
4. Results

Experiments on enriched semantics strategies verified the effectiveness of the proposed method in remote sensing object detection, especially in the detection of small-sized targets and densely packed objects. Visualization of the objects detected by the proposed approach in NWPU VHR-10 dataset is shown in Figure 4. The predicted bounding boxes which match the ground truth bounding boxes with Intersection-over-Union (IoU) > 0.5 are plotted in different colors according to the categories. From Figure 4, the newly presented detector could generate satisfactory bounding boxes that covered most of the targets even when recognizing orientation variant targets or densely arranged objects such as storage tanks, tennis courts, and especially harbors. Besides, the detection performance for extremely small scales objects, like ships in Figure 4b and vehicles in Figure 4j was also promising. Unlike general SSD, which produces a plethora of bounding boxes with low IOU and confidence scores by multi-scale prediction, our method could exclude most of the false bounding boxes and suppress some false alarms by refined location module. Although our detectors could cover most objects, there still existed a small number of overlapped bounding boxes which had a large intersection with others, as shown in Figure 4h where harbors were closely aligned.

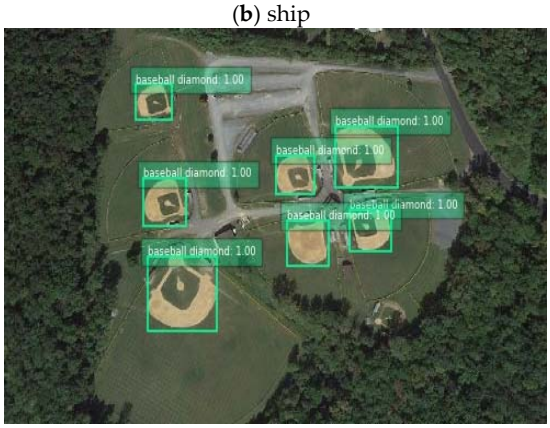
(**a**) airplane



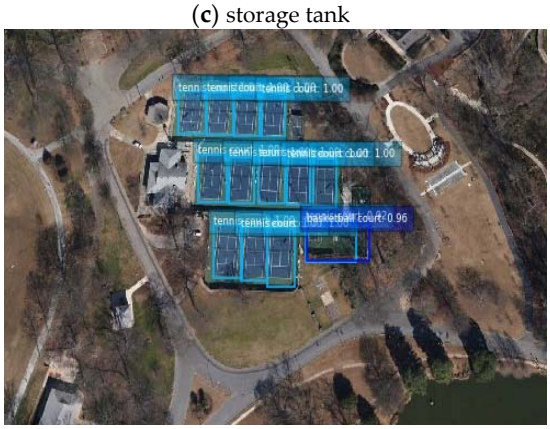
(**b**) ship



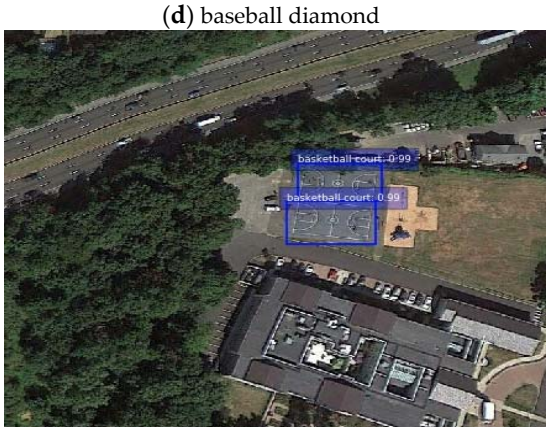
(**c**) storage tank



(**d**) baseball diamond



(**e**) tennis court



(**f**) basketball court

(**g**) ground track field

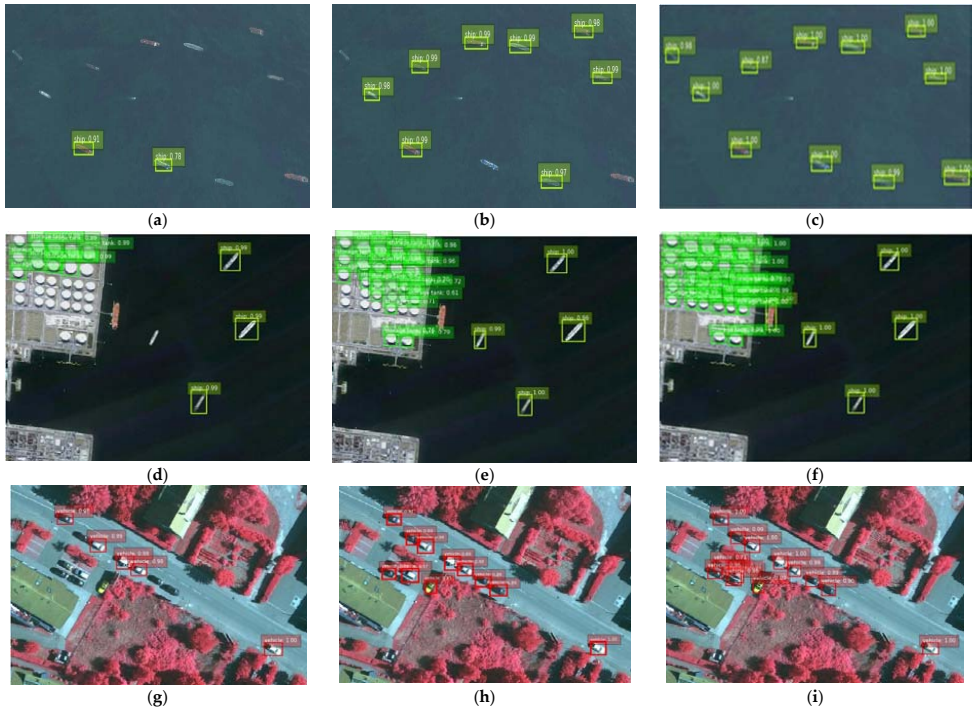
(**h**) harbor

Figure 4. *Cont.*

(**i**) bridge (**j**) vehicle

Figure 4. Visualization of some object detection examples by recurrent detection with activated semantics (RDAS) in the NWPU VHR-10 dataset. Detection results such as (**a**) airplanes; (**b**) ships; (**c**) storage tanks; (**d**) baseball diamonds; (**e**) tennis courts; (**f**) basketball courts; (**g**) ground track fields; (**h**) harbors; (**i**) bridges; (**j**) vehicles are displayed. The numbers on the bounding boxes denote the confidence score for each object category.

To explicitly analyze the superiority of our detector, we displayed ships, storage tanks, vehicles, basketball courts, and baseball diamonds detected by applying SSD, improved methods based on SSD, and our method in Figure 5. As shown in Figure 5a,d,g,j, only a small fraction of objects with extremely small sizes were detected by SSD. Nevertheless, the improved single-shot stage based methods indeed detected more small scaled targets to some extent, while the remaining parts were still omitted. The newly presented method resulted in a better performance in the given scenes. Not only small-sized targets such as ships and closely aligned targets such as storage tanks and tennis courts are successfully detected, but also a basketball court arranged around the tennis court was well detected in Figure 5l. All the amazing results benefited from the segmentation branch and attention mechanism used to enrich the semantic information.



(**a**) (**b**) (**c**)

(**d**) (**e**) (**f**)

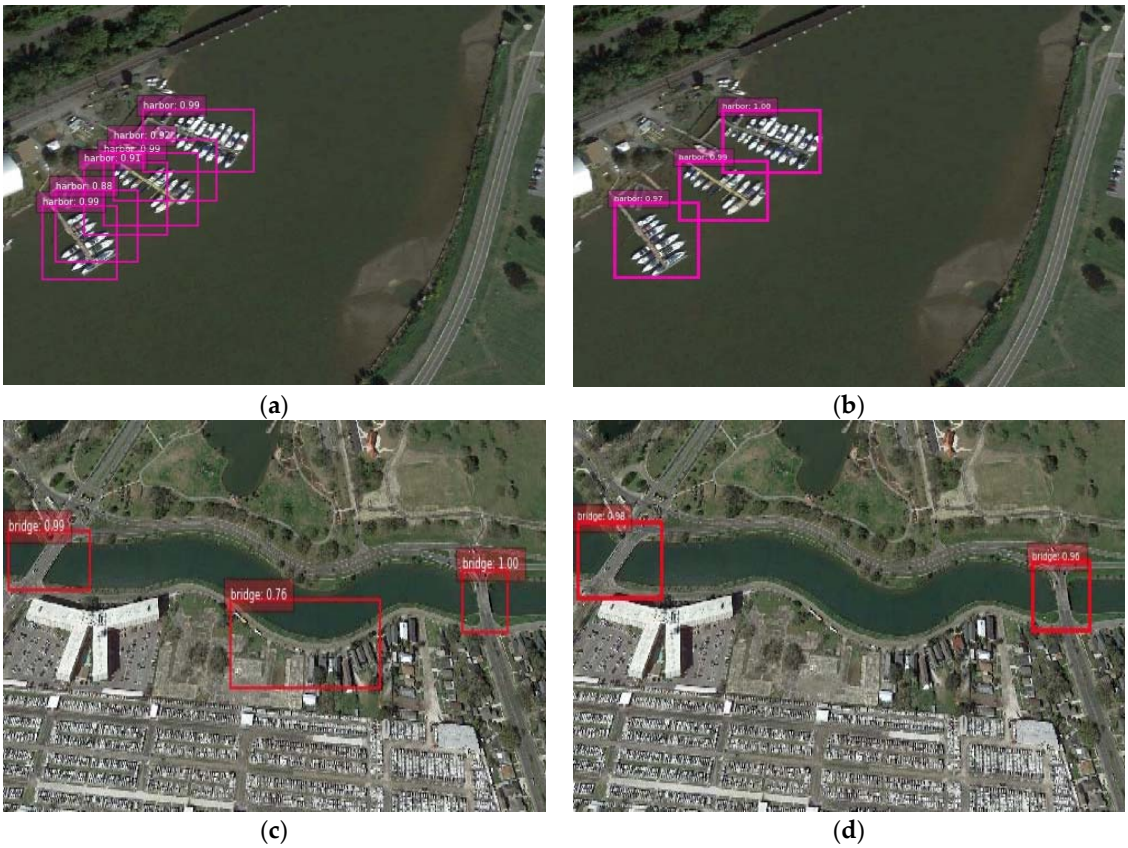
(**g**) (**h**) (**i**)

Figure 5. *Cont.*

(j) (k) (l)

Figure 5. Detection results using Single Shot Multi-Box Detector (SSD), other single-shot stage based methods and our method. The first column denotes the results of SSD; the second column represents the results of improved SSD; the third column means the results of our method. (**a**–**l**) are the results of ships, storage tanks, vehicles and tennis courts respectively.

The method of enriching semantics at higher level layer proposed in [43] only introduces channel-level attention while spatial information also counts in the task of object detection. For the purpose of validating the effectiveness of two-level attention mechanism introduced in the CAB module, we demonstrate the experimental results using the segmentation branch combined with channel-level attention and two-level attention, respectively. In Figure 6a, for one object with irregular shape, more than one bounding box was predicted to locate the target if we applied channel-level attention; nevertheless, the redundant bounding boxes were eliminated by two-level attention. As for the false alarm displayed in Figure 6c, the margin of the river was mistaken as a bridge while it was suppressed in our method.



(**a**) (**b**)

(**c**) (**d**)

Figure 6. Detection results using a method with channel-level attention and a two-level attention mechanism. The first column denotes the results of segmentation branch combined with channel-level attention; the second column represents our method with two-level attention mechanism. (**a**–**d**) denote the results of harbors and bridges respectively.

In the following subsection, quantitative and qualitative analysis will be presented to evaluate the performance of the SSD with enriched semantics.

4.1. Quantitative Evaluation of NWPU VHR-10 Dataset

To validate the effectiveness of the proposed method, object detection experiments are divided into methods based on traditional or region-based approaches and regression-based approaches. In Tables 1 and 2, we display the quantitative comparisons measured by AP values from two main stream methods, respectively. The best AP value of each category is bold in Table 2. Table 3 shows the average running time per image of some methods. The recall rate value is displayed in Figure 7. In review of some region-based methods, the RICNN [34] with fine-tuning method uses AlexNet pre-trained on ImageNet. Apart from the Zeiler and Fergus (ZF) model, the R-P-Faster R-CNN used the VGG16 training mechanism, which contains single fine-tuning and double fine-tuning. The deformable R-FCN with arcNMS was fine-tuned on the ResNet-101 ImageNet pre-trained model. The Multi-Scale CNN initializes the parameters with VGG16 by the model pre-trained on ImageNet. As shown in Table 1, the application of multi-scale strategy in region-based method has shown a superior high average precision value. From Table 2, among all the single-shot stage based methods, the proposed RDAS, once fine-tuned on the VGG16 ImageNet pre-trained model, obtained the amazing performance and pushes the benchmark into 89.5%, which is slightly lower than the best result in the region-based method while greatly cuts down the processing time. Although some innovative methods such as DSSD, DSOD, GRP-DSOD have achieved surprising improvement on the whole, there still exists some weakness in performances when compared to the newly presented method as indicated in Table 2. After segmentation branch was added to the recurrent detection with activated semantics, the APs among all the objects all increased, including basketball court (0.844 to 0.948), harbor (0.759 to 0.826), bridge (0.738 to 0.772), and vehicle (0.808 to 0.865) are all increased. Compared with the region-based method, the APs of ship, ground track field as well as harbor still can be boosted by further strategies.

Table 1. The AP (Average Precision) values of traditional and region-based methods.

	BoW	SSC BoW	FDDL	CPOD	Transferred AlexNet	Newly Trained AlexNet	RICNN without Fine-Tuning	RICNN with Fine-Tuning	R-P-Faster R-CNN (ZF)	R-P-Faster R-CNN (Single) (VGG16)	Deformable R-FCN (ResNet-101) with arcNMS	Multi-Scale CNN
Airplane	0.025	0.506	0.292	0.623	0.661	0.701	0.860	0.884	0.803	0.906	0.873	0.993
Ship	0.585	0.508	0.376	0.689	0.569	0.637	0.760	0.773	0.681	0.762	0.814	0.920
Storage tank	0.632	0.334	0.770	0.637	0.843	0.843	0.850	0.853	0.359	0.403	0.636	0.832
Baseball diamond	0.090	0.435	0.258	0.833	0.816	0.836	0.873	0.881	0.906	0.908	0.904	0.972
Tennis court	0.047	0.003	0.028	0.321	0.350	0.355	0.396	0.408	0.715	0.797	0.816	0.908
Basketball court	0.032	0.150	0.036	0.363	0.459	0.468	0.579	0.585	0.677	0.774	0.741	0.926
Ground track field	0.078	0.101	0.201	0.853	0.800	0.812	0.855	0.867	0.892	0.880	0.903	0.981
Harbor	0.530	0.583	0.254	0.553	0.620	0.623	0.665	0.686	0.769	0.762	0.753	0.851
Bridge	0.122	0.125	0.215	0.148	0.423	0.454	0.585	0.615	0.572	0.575	0.714	0.719
Vehicle	0.091	0.336	0.045	0.440	0.429	0.448	0.680	0.711	0.646	0.666	0.755	0.859
Mean AP	0.246	0.308	0.245	0.546	0.597	0.618	0.710	0.726	0.702	0.743	0.791	0.896

Table 2. The AP (Average Precision) values of single-shot stage based methods.

	SSD300 (VGG16)	SSD512 (VGG16)	DSSD321 (ResNet101)	DSOD300	GRP-DSOD300	RDES300 without Segmentation Branch	RDES300 with Segmentation Branch (VGG16)	RDAS512 without Segmentation Branch	RDAS512 with Segmentation Branch (VGG16)
Airplane	0.958	0.904	0.865	0.816	0.827	0.998	0.996	0.989	**0.996**
Ship	0.784	0.609	0.654	0.580	0.628	0.700	0.801	0.827	**0.855**
Storage tank	0.855	0.798	**0.903**	0.853	0.892	0.897	0.883	0.902	0.890
Baseball diamond	0.899	0.899	0.896	0.866	0.901	0.908	0.948	0.924	**0.950**
Tennis court	0.894	0.826	0.851	0.847	0.878	0.895	0.896	0.875	**0.896**
Basketball court	0.829	0.806	0.804	0.816	0.809	0.818	0.884	0.844	**0.948**
Ground track field	0.012	**0.983**	0.782	0.833	0.798	0.791	0.807	0.957	0.953
Harbor	0.688	0.734	0.705	0.785	0.821	0.705	0.716	0.759	**0.826**
Bridge	0.702	0.767	0.682	0.689	**0.812**	0.732	0.741	0.738	0.772
Vehicle	0.819	0.521	0.742	0.589	0.613	0.851	0.858	0.808	**0.865**
Mean AP	0.744	0.784	0.788	0.790	0.798	0.830	0.853	0.862	**0.895**

Table 3. Computation time comparisons for different detection methods.

	BoW	SSC BoW	FDDL	CPOD	Transferred AlexNet	Newly Trained AlexNet	RICNN with Fine-Tuning	R-P-Faster R-CNN (ZF)	R-P-Faster R-CNN (Single) (VGG16)	Deformable R-FCN (ResNet-101) with arcNMS	Multi-Scale CNN	SSD512 (VGG16)	RDAS512 without Segmentation Branch	RDAS512 with Segmentaion Branch (VGG16)
Average running time per image (second)	5.32	40.32	7.17	1.06	5.24	8.77	8.77	0.005	0.155	0.201	0.11	0.061	0.054	0.057

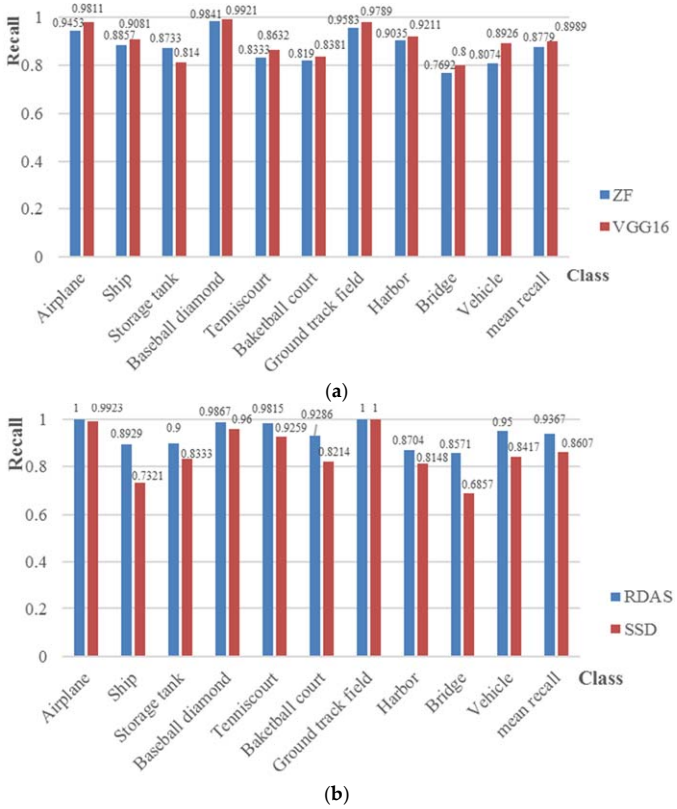


Figure 7. Quantitative evaluation results measured by recall rate for all 10 object categories. (1. airplane; 2. ship; 3. Storage tank; 4. Baseball diamonds; 5. Tennis court; 6. Basketball court; 7. Ground track field; 8. Harbor; 9. Bridge; 10. Vehicle; and 11. Average.) The number on the bars denote the recall rates for each object category. (**a**) The blue bar means the recall value obtained with the ZF model for fine-tuning. The red one means the recall value obtained with the VGG16 model for fine-tuning. (**b**) The red bar denotes the recall value obtained with SSD512 with the VGG16 model for fine-tuning. The blue one means the recall value obtained with our proposed method.

Except the above evaluation indexes, we also take the computational efficiency as an important factor for evaluating the performance of the proposed algorithm. To quantitatively evaluate the inference speed, we run RDAS, SSD, as well as some other competitors on our machine to compare the speed fairly and Table 2 shows the average running time of all previous approaches. From the results in Table 2, it can be confirmed that our method can achieve a relatively high average precision without large time consumption.

For object detection in remote sensing, the running time per image of RDAS was slightly slower than the VGG16-based SSD given its additional modules while it is faster than the two-stage detectors. Figure 7 demonstrates the recall values of Faster R-CNN fine-tuned with ZF and VGG16 model, the VGG16-based SSD and the proposed RDAS, which achieves an overall recall rate of 0.9367. From the overall view, it can be concluded that the recall value of the proposed one-stage detector with enriched semantics based on VGG16 is higher than the original SSD under the same training conditions. Compared with region-based methods such as Faster R-CNN, the recall values of tennis court, basketball court, ground track field, and vehicle were escalated by at least 0.1 on average. In Figure 7, it can be seen that the classes of airplane, baseball diamond, tennis court, ground track field and vehicle can obtain high recall values of greater than 0.95; however, the classes of harbor and bridge presents worse recall values. Furthermore, to balance the trade-off existing between precision and recall, the following PRC explains the limitation of evaluating the performance of various CNN architectures.

4.2. PRC Evaluation of NWPU VHR-10 Dataset

For object detection approaches, PRC acts as one of the elementary indicators of robustness and effectiveness. The precision vector generated in experiments is measured on the *y*-axis and the recall rate on the *x*-axis. The curve at the top of the PRCs indicates a better performance. In this paper, we focused on the single-shot based detector and the improved version, i.e., SSD512, RDAS without segmentation branch and RDAS trained on VGG16 with an optimized hyper-parameter. Figure 8 shows the PRCs of these three methods as well as the traditional comparison methods. From Figure 8, we can conclude that our RDAS achieves the best recall for all classes since this detector could produce more precise anchor boxes which cover most objects. In particular, the recall rates of small objects like airplanes, storage tanks and vehicles increase more than other objects, which further validate the effectiveness of our method for small object detection. Considering another indicator, it is seen that most of the classes in RDAS achieve higher precision than SSD and traditional methods especially in detecting baseball diamonds, tennis courts and ground track fields. Nevertheless, for ships, bridges, harbors and basketball courts, RDAS requires improvement. Moreover, the segmentation branch was proved effective in elevating the AP value by preventing PRC from decreasing.

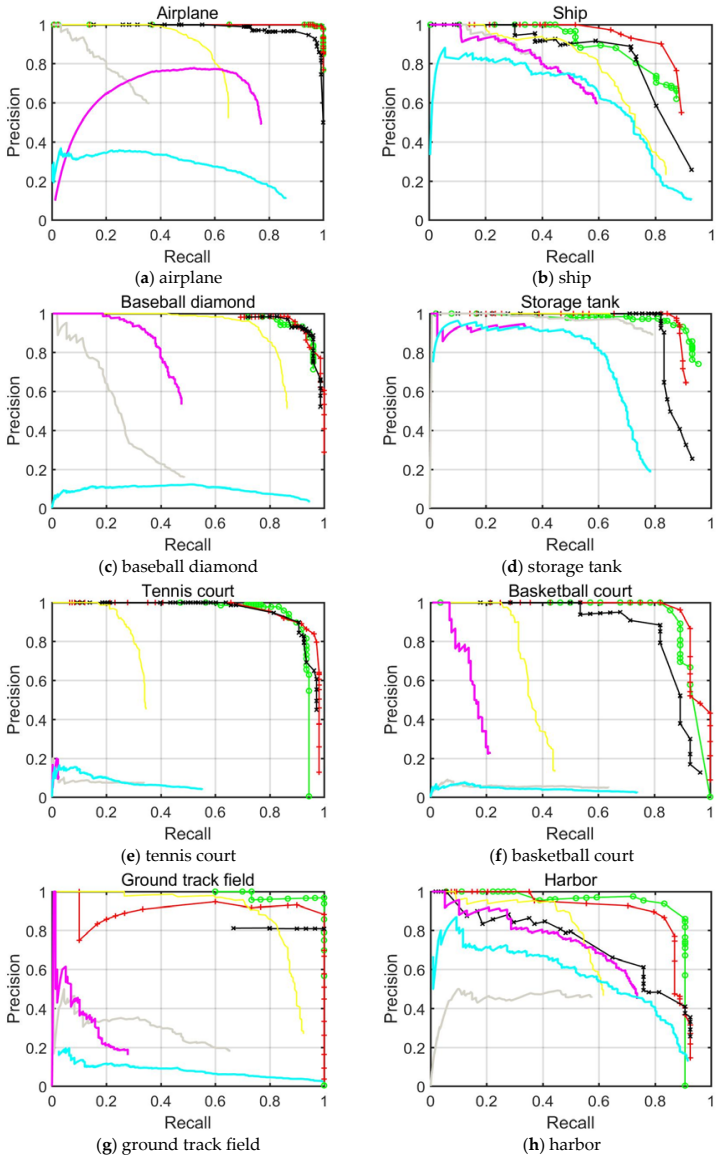


Figure 8. *Cont.*

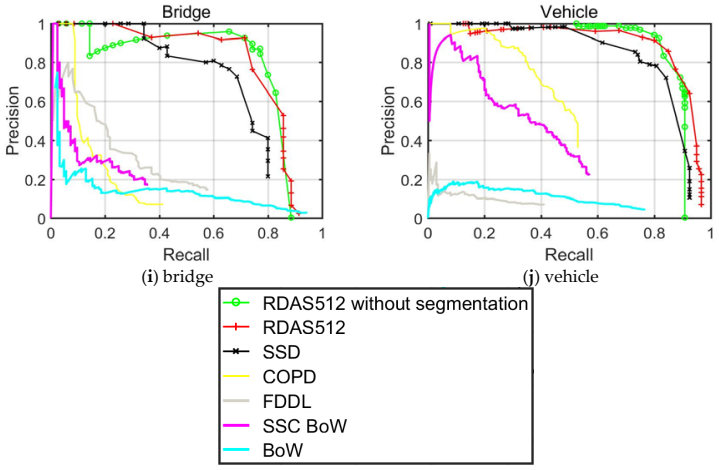


Figure 8. PRCs (Precision Recall Curves) of three single-shot based object detection approaches, as well as the comparison methods. PRCs of the ten class objects in NWPU VHR-10 are displayed in (**a**–**j**), respectively.

By jointly analyzing the AP values, the recall rate and the PRCs, it can be drawn that the proposed RDAS algorithm displays a superior detection performance especially for small scaled objects and geometrically variant objects such as ships, bridges and baseball diamonds. We also confirmed our intuition that the performance can be further boosted by the addition of segmentation branch, and introducing high level semantic knowledge to the early stage of the detection network can contribute to a stronger object detector.

5. Conclusions

In this paper, a novel single-shot detector named Recurrent Detection with Activated Semantics (RDAS) structure is presented for addressing the small-scaled object fast detection problem in VHR remote sensing imagery. Since low level detection feature maps usually lack high level semantic information, a segmentation branch is introduced to provide semantic-related and class-aware features to activate and adjust the original feature map utilized in the detection branch. We also deploy the concurrent attention block to combine context information both locally and globally and enhance the representation power throughout the network. Considering the complex distribution of geospatial objects and the low efficiency of the current two-stage based detection methods for HSR remote sensing imagery, the robust properties of segmentation branch and concurrent attention blocks are considered and combined in the original SSD object detection As the original feature map from standard SSD is substituted by recalibrated ones, this modified version is capable of detecting small objects under a more complicated visual appearance.

Our workflow was experimented using the NWPU VHR-10 dataset. The quantitative evaluation results indicate that the proposed SSD with enriched semantic information approach excels state-of-the-art benchmarks for object detection in both accuracy and speed. A detailed investigation confirmed that the added segmentation module attached at low-level feature map shows better performance for geometrically diverse objects such as baseball diamond, bridge and vehicle, and it can be easily applied in existing two-stage frameworks or single shot object detectors with stronger backbone network. In our future work, a more effective and accurate object detection architecture will be considered for HSR imagery geospatial object detection.

Author Contributions: All of the authors made significant contributions to the work. S.C. designed the research and analyzed the results. R.Z. provided suggestion for the preparation and revision of the paper.

Acknowledgments: This work was supported by National Natural Science Foundation of China under Grant No. 61471370. The authors would also like to thank the anonymous reviewers for their very competent comments and helpful suggestions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pelich, R.; Longépé, N.; Mercier, G.; Hajduch, G.; Garello, R. AIS-Based Evaluation of Target Detectors and SAR Sensors Characteristics for Maritime Surveillance. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 3892–3901. [CrossRef]
2. Zhong, P.; Wang, R. A multiple conditional random field's ensemble framework for urban area detection in remote sensing optical images. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 3978–3988. [CrossRef]
3. Li, X.; Cheng, X.; Chen, W.; Chen, G.; Liu, S. Identification of Forested Landslides Using LiDAR Data, Object-based Image Analysis, and Machine Learning Algorithms. *Remote Sens.* **2015**, *7*, 9705–9726. [CrossRef]
4. Cheng, G.; Han, J. A Survey on Object Detection in Optical Remote Sensing Images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
5. Li, Y.; Zhang, Y.; Huang, X.; Zhu, H.; Ma, J. Large-scale remote sensing image retrieval by deep hashing neural networks. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 950–965. [CrossRef]
6. Ma, J.; Jiang, J.; Liu, C.; Li, Y. Feature guided Gaussian mixture model with semi-supervised EM and local geometric constraint for retinal image registration. *Inf. Sci.* **2017**, *471*, 128–142. [CrossRef]
7. Li, Y.; Tao, C.; Tan, Y.; Shang, K.; Tian, J. Unsupervised Multilayer Feature Learning for Satellite Image Scene Classification. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 157–161. [CrossRef]
8. Ma, J.; Zhou, H.; Zhao, J.; Gao, Y.; Jiang, J.; Tian, J. Robust Feature Matching for Remote Sensing Image Registration via Locally Linear Transforming. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 6469–6481. [CrossRef]
9. Stankov, K.; He, D.C. Detection of buildings in multispectral very high spatial resolution images using the percentage occupancy hit-or-miss transform. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 4069–4080. [CrossRef]
10. Jain, A.K.; Ratha, N.K.; Lakshmanan, S. Object detection using Gabor filters. *Pattern Recognit.* **1997**, *30*, 295–309. [CrossRef]
11. Leninisha, S.; Vani, K. Water flow based geometric active deformable model for road network. *ISPRS J. Photogramm. Remote Sens.* **2015**, *102*, 140–147. [CrossRef]
12. Ok, A.O. Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts. *ISPRS J. Photogramm. Remote Sens.* **2013**, *86*, 21–40. [CrossRef]
13. Ok, A.O.; Senaras, C.; Yuksel, B. Automated detection of arbitrarily shaped buildings in complex environments from monocular VHR optical satellite imagery. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 1701–1717. [CrossRef]
14. Blaschke, T.; Hay, G.J.; Kelly, M.; Lang, S.; Hofmann, P.; Addink, E.; Feitosa, R.Q.; Vander Meer, F.; Vander Werff, H.; Van Coillie, F. Geographic object-based image analysis-towards a new paradigm. *ISPRS J. Photogramm. Remote Sens.* **2014**, *87*, 180–191. [CrossRef] [PubMed]
15. Li, Y.; Wang, S.; Tian, Q.; Ding, X. Feature representation for statistical-learning-based object detection: A review. *Pattern Recognit.* **2015**, *48*, 3542–3559. [CrossRef]
16. Siva, P.; Russell, C.; Xiang, T. In defense of negative mining for annotating weakly labeled data. In Proceedings of the European Conference on Computer Vision, Firenze, Italy, 7–13 October 2012.
17. Cheng, G.; Han, J.; Guo, L. Object detection in remote sensing imagery using a discriminatively trained mixture model. *ISPRS J. Photogramm. Remote Sens.* **2013**, *85*, 32–43. [CrossRef]
18. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
19. Zhang, Y.; Du, B.; Zhang, L. A sparse representation-based binary hypothesis model for target detection in hyperspectral images. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 1346–1354. [CrossRef]
20. Geva, S.; Sitte, J. Adaptive nearest neighbor pattern classification. *IEEE Trans. Neural Netw.* **2002**, *2*, 318–322. [CrossRef] [PubMed]
21. Tim, K. Random decision forests. In Proceedings of the International Conference on Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013; Volume 1, p. 278.
22. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
23. Guo, Y.; Oerlemans, A.; Lao, S.; Wu, S.; Lew, M.S. Deep learning for visual understanding: A review. *Neurocomputing* **2016**, *187*, 27–48. [CrossRef]
24. Druzhkov, P.N.; Kustikova, V.D. A survey of deep learning methods and software tools for image classification and object detection. *Pattern Recognit. Image Anal.* **2016**, *26*, 9–15. [CrossRef]

25. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 580–587.
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 1904–1916. [CrossRef] [PubMed]
27. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015; pp. 1440–1448.
28. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
29. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
30. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S. SSD: Single Shot MultiBox Detector. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 21–37.
31. 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*; Springer: Cham, Switzerland, 2016; pp. 354–370.
32. Shelhamer, E.; Long, J.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef] [PubMed]
33. Han, X.; Zhong, Y.; Zhang, L. An Efficient and Robust Integrated Geospatial Object Detection Framework for High Spatial Resolution Remote Sensing Imagery. *Remote Sens.* **2017**, *9*, 666. [CrossRef]
34. 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.* **2016**, *54*, 7405–7415. [CrossRef]
35. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object Detection via Region-based Fully Convolutional Networks. *arXiv*, 2016.
36. Kang, M.; Ji, K.; Leng, X.; Lin, Z. Contextual Region-based Convolutional Neural Network with Multilayer Fusion for SAR Ship Detection. *Remote Sens.* **2017**, *9*, 860. [CrossRef]
37. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. DSSD: Deconvolutional single shot detector. *arXiv*, 2017.
38. Shen, Z.; Liu, Z.; Li, J.; Jiang, Y.G.; Chen, Y.; Xue, X. Dsod: Learning deeply supervised object detectors from scratch. *arXiv*, 2017.
39. Shen, Z.; Shi, H.; Feris, R.; Cao, L.; Yan, S.; Liu, D.; Wang, X.; Xue, X.; Huang, T.S. Learning Object Detectors from Scratch with Gated Recurrent Feature Pyramids. *arXiv*, 2017.
40. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. *arXiv*, 2017.
41. Gidaris, S.; Komodakis, N. Object detection via a multi-region and semantic segmentation-aware cnn model. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1134–1142.
42. Shrivastava, A.; Gupta, A. Contextual priming and feedback for faster r-cnn. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 330–348.
43. Zhang, Z.; Qian, S.; Xie, C.; Shen, W.; Wang, B.; Yuille, A.L. Single-Shot Object Detection with Enriched Semantics. *arXiv*, 2017.
44. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv*, 2016.
45. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. *arXiv*, 2017.
46. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *arXiv*, 2015.
47. NWPU VHR-10 Dataset. Available online: http://www.escience.cn/people/gongcheng/NWPU-VHR-10.html (accessed on 20 April 2018).
48. Xu, S.; Fang, T.; Li, D.; Wang, S. Object classification of aerial images with bag-of-visual words. *IEEE Geosci. Remote Sens. Lett.* **2010**, *7*, 366–370.
49. Sun, H.; Sun, X.; Wang, H.; Li, Y.; Li, X. Automatic target detection in high-resolution remote sensing images using spatial sparse coding bag-of-words model. *IEEE Geosci. Remote Sens. Lett.* **2012**, *9*, 109–113. [CrossRef]
50. Han, J.; Zhou, P.; Zhang, D. Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding. *ISPRS J. Photogramm. Remote Sens.* **2014**, *89*, 37–48. [CrossRef]
51. Cheng, G.; Han, J.; Zhou, P. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS J. Photogramm. Remote Sens.* **2014**, *98*, 119–132. [CrossRef]

52. 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.* **2018**, *10*, 131. [CrossRef]
53. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv*, 2014.
54. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. *J. Mach. Learn. Res.* **2010**, *9*, 249–256.