remote sensing

MDPI

Article

A Novel Multi-Model Decision Fusion Network for Object Detection in Remote Sensing Images

Wenping Ma [1], Qiongqiong Guo [1], Yue Wu [2,*], Wei Zhao [1], Xiangrong Zhang [1] and Licheng Jiao [1]

[1] Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education,
International Research Center for Intelligent Perception and Computation,
Joint International Research Laboratory of Intelligent Perception and Computation,
School of Artificial Intelligence, Xidian University, Xi'an 710071, China;
wpma@mail.xidian.edu.cn (W.M.); qqiongguo@126.com (Q.G.); weizhao_90@163.com (W.Z.);
xrzhang@mail.xidian.edu.cn (X.Z.); lchjiao@mail.xidian.edu.cn (L.J.)

[2] School of Computer Science and Technology, Xidian University, Xi'an 710071, China

* Correspondence: ywu@xidian.edu.cn

Received: 28 January 2019; Accepted: 19 March 2019; Published: 27 March 2019

Abstract: Object detection in optical remote sensing images is still a challenging task because of the complexity of the images. The diversity and complexity of geospatial object appearance and the insufficient understanding of geospatial object spatial structure information are still the existing problems. In this paper, we propose a novel multi-model decision fusion framework which takes contextual information and multi-region features into account for addressing those problems. First, a contextual information fusion sub-network is designed to fuse both local contextual features and object-object relationship contextual features so as to deal with the problem of the diversity and complexity of geospatial object appearance. Second, a part-based multi-region fusion sub-network is constructed to merge multiple parts of an object for obtaining more spatial structure information about the object, which helps to handle the problem of the insufficient understanding of geospatial object spatial structure information. Finally, a decision fusion is made on all sub-networks to improve the stability and robustness of the model and achieve better detection performance. The experimental results on a publicly available ten class data set show that the proposed method is effective for geospatial object detection.

Keywords: convolutional neural networks (CNNs); object detection; remote sensing images; contextual information; part-based; multi-model

1. Introduction

Nowadays, optical remote sensing images with high spatial resolution are obtained conveniently due to the significant progress in remote sensing technology, which leads to a wide range of applications such as land planning, disaster control, urban monitoring, and traffic planning [1–4]. As one of the most fundamental and challenging tasks required for understanding remote sensing images, object detection has gained increasing attention in recent years. To deal with a variety of problems faced in optical remote sensing image object detection, numerous approaches have been proposed [5,6]. A deep review on object detection in optical remote sensing images can be found in [7].

As is known to all, a common method for object detection is to extract features. The quality of the extracted features is critical as it will directly affect the final result of object detection. Powerful feature representation can make an object more discriminative and its location more explicit, which makes the object easier to detect. On the contrary, insufficient ability to represent objects will result in inaccurate detection. Therefore, it is important for us to choose a method to extract features for object detection in remote sensing images. Currently, because of the advantage of directly generating more powerful feature representations from raw image pixels through neural networks, deep learning methods, especially CNN-based [4,8–25], are recognized as predominate techniques for extracting features in object detection. Therefore, we select a CNN-based approach to extract features for object detection in optical remote sensing images.

Object detection in remote sensing images becomes more complicated because of the diversity of illumination intensities, noise interference, and the influence of weather. At present, there are still a lot of problems to be solved, such as the diversity and complexity of geospatial object appearance, and the insufficient understanding of geospatial object spatial structure information.

In the field of optical remote sensing images, lots of object detection algorithms only pay attention to the features of objects themselves [16,17,26]. However, due to the diversity and complexity of geospatial object appearance, in many cases, relying solely on the characteristics of an object itself cannot effectively identify the object, and sometimes may even cause mis-detection between two objects which belong to two different classes but look very similar in appearance factor. For instance, recognizing a storage tank only through exploiting its features may be difficult as its appearance is just circular, and a bridge is often mistaken for part of the road (as shown in Figure 1). In this case, the application of auxiliary information can effectively help detect objects. Therefore, contextual information is a choice. Some existing works [18,20,27] take local contextual information into account and obtain good performance. For example, the work in [20] used features surrounding the regions of interest, thus alleviating false detection caused by object appearance ambiguity. Although those methods yield good results, there are still deficiencies. Also, relationships among objects play an important role in improving the performance of detection. Therefore, in addition to the use of local contextual information, the proposed method takes object-object relationship contextual information into consideration.

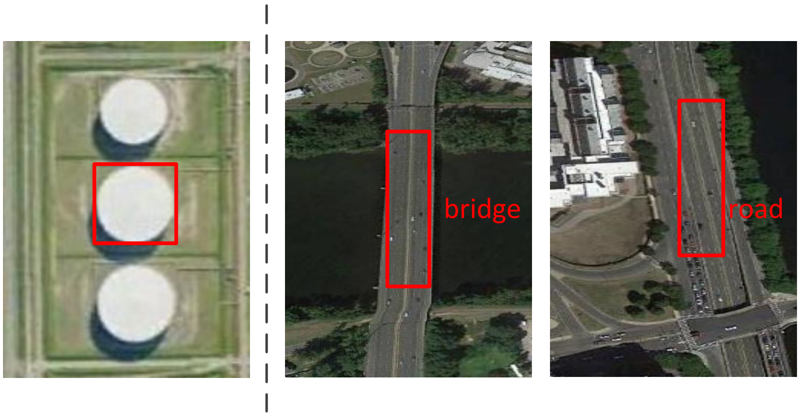

Figure 1. Examples difficult to detect. (**Left**) Only using the sample appearance features in the red rectangle, just a circle, is hard to identify the storage tank. (**Right**) The bridge and the road are easily confused.

The spatial structure of geospatial objects plays an important role in recognizing the objects. Optical remote sensing images with high spatial resolutions always contain abundant spatial structure information about objects. Therefore, investigating deeply the structural information about objects can result in good detection results. It is necessary to design an object detector to effectively alleviate the insufficient understanding of geospatial object spatial structure information. Each part of a geospatial object provides many local visual properties and much geometric information about the object. Paying attention to the various parts of an object can help us to understand more details about its spatial structure. There are lots of part-based models [28–32] concentrating on using the various parts of objects to improve detection performance. For example, Zhang et al. [28] proposed a generic discriminative part-based model (GDPBM), which divides a geospatial object with arbitrary orientation into several parts to achieve good performance for object detection in optical remote sensing images. Unlike the previous part-based approaches [28–32], which use traditional features such as histogram of oriented gradients (HOG) [33], the proposed method applies the CNN-based technique to extract high-level features for better feature representation. In addition, it is easier to obtain and process parts of objects in the proposed approach.

In this paper, we propose a novel multi-model decision fusion framework for object detection in remote sensing images. Aiming at the diversity and complexity of geospatial object appearance, we build a local contextual information and object-object relationship contextual information fusion sub-network. Focusing on the insufficient understanding of geospatial object spatial structure information, we construct a part-based multi-region feature fusion sub-network. Furthermore, unlike many methods just using single model, we make a decision fusion on several models for better stability and robustness. For the implementation of the multi-model decision fusion strategy, in addition to the above two sub-networks, we also fuse a baseline sub-network based on Faster R-CNN model.

In summary, the major contributions of this paper are presented as follows.

(1) We propose a local contextual information and object-object relationship contextual information fusion sub-network based on gated recurrent unit (GRU) to form discriminative feature representation, which can effectively recognize objects and reduce false detection between different types of objects with similar appearance. The object-object relationship contextual information is introduced for the first time in the field of remote sensing image object detection as far as we know.

(2) We propose a new part-based multi-region feature fusion sub-network to investigate more details of objects, which can diversify object features and enrich semantic information.

(3) We propose a multi-model decision fusion strategy to fuse the detection results of the three sub-networks, which can improve the stability and robustness of the model and obtain better algorithm performance.

The remainder of this paper is organized as follows. The second section gives a brief review of the related work on geospatial object detection, contextual information fusion, and the RoIAlign layer. In the third section, we introduce the proposed method in detail. The details of our experiments and results are presented in the fourth section. The last section concludes this paper with a discussion of the results.

2. Related Work

2.1. Geospatial Object Detection

In the past decades, the research on the field of remote sensing image object detection has made a breakthrough development. Many object detection algorithms have been proposed to address various problems [17,20,34]. For example, Cheng et al. [17] proposed a novel and effective approach to learn a rotation-invariant CNN (RICNN) model for addressing the problem of object rotation variations, which is achieved by introducing and learning a new rotation-invariant layer on the basis of the existing CNN frameworks. Han et al. [34] combined the weakly supervised learning (WSL) and high-level feature learning to tackle the problems of manual annotation and insufficiently powerful descriptors. Li et al. [20] put forward a novel region proposal network (RPN) including multiangle, multiscale, and multiaspect-ratio anchors to address the problem of geospatial object rotation variations, and also proposed a double-channel feature fusion network which can learn local and contextual properties to deal with the geospatial object appearance ambiguity issue.

Low-level features are often used for image analysis [35]. Employing the extracted low-level features of objects for object detection has been a very common method used by many scholars. Those low-level features contain scale-invariant feature transform (SIFT) [3,34,36], histogram of oriented gradients (HOG) [5,6,33], the bag-of-words (BoW) model [37–39], Saliency [40,41], etc. For example, Tuermer et al. [5] used the HOG feature and disparity maps to detect airborne vehicles in dense urban areas. Shi et al. [6] developed a circle frequency-HOG feature for ship detection by combining circle frequency features with HOG features. Han et al. [40] proposed to detect multiple-class geospatial objects through integrating visual saliency modeling and the discriminative learning of sparse coding. Although those low-level features show impressive success in some specific object detection tasks, they have certain limitations because they do not represent the high-level semantic information required for identifying objects, especially when visual recognition tasks become more challenging.

Currently, deep convolutional neural network (CNN) models are widely used in the field of visual recognition [42–44], such as object detection, owing to the powerful ability of CNN to capture both low-level and high-level features. The region-based convolutional neural network (R-CNN) [8] is considered as a milestone among CNN-based object detection approaches, and achieves superior performance. Subsequently, many advanced object detection algorithms in natural images, such as Fast R-CNN [9], Faster R-CNN [10], YOLO [11], SSD [12], Mask R-CNN [13], are proposed successively and yield unusually brilliant results. However, the aforementioned models can not be directly utilized for geospatial object detection, because the properties of remote sensing images and natural images are different and the direct application of those models to remote sensing images is not optimal. Researchers have done a lot of work in applying CNN-based models to detect geospatial objects in remote sensing images and achieved remarkable consequences [4,15–25,45]. For example, the work in [4] utilized a hyperregion proposal network (HRPN) and a cascade of boosted classifiers to detect vehicles in remote sensing images. Long et al. [16] proposed a new object localization framework based on convolutional neural networks to efficiently achieve the generalizability of the features used to describe geospatial objects, and obtained accurate object locations. Yang et al. [21] constructed a Markov random field (MRF)-fully convolutional network to detect airplanes.

2.2. Contextual Information Fusion

Contextual information is advantageous to various visual recognition tasks [18,20,27,46–53], such as object detection. For example, in order to promote object detection performance, the work in [48] developed a novel object detection model, attention to context convolution neural network (AC-CNN), through incorporating global and local contextual information into the region-based CNN detection framework. Bell et al. [49] presented the Inside-Outside Net (ION) to exploit information both inside and outside the regions of interest, which integrates the contextual information outside the regions of interest by using spatial recurrent neural networks. Furthermore, some recent works [50–52] proposed new architectures to investigate the contextual information about object-object relationships for better object detection performance. In the field of remote sensing images, the work in [20] fused local and contextual features to address the problem of object appearance ambiguity in object detection. Considering that the appearance is not enough to distinguish oil tanks from the complex background, Zhang et al. [27] applied trained CNN models to extract contextual features, which makes oil tanks easier to recognize. Xiao et al. [18] fused auxiliary features both within and surrounding the regions of interest to represent the complementary information of each region proposal for airport detection, effectively alleviating detection problems caused by the diversity of illumination intensities in remote sensing images. Motivated by those models, we believe that the local contextual information and the object-object relationship context are very useful for object detection in optical remote sensing images. It is necessary to remember features of the object itself before incorporating contextual information. The process of merging messages follows the memory characteristics of Gated Recurrent Units (GRU) [54]. Therefore, we use GRU to fuse the two types of features.

Next we introduce how the j-th hidden unit in a GRU cell works. First, the *reset* gate r_j is obtained by:

$$r_j = \sigma([\mathbf{W}_r\mathbf{x}]_j + [\mathbf{U}_r\mathbf{h}_{t-1}]_j) \tag{1}$$

where σ is the logistic sigmoid function, and $[.]_j$ indicates the j-th element of a vector. $\mathbf{x}$ is the input, while $\mathbf{h}_{t-1}$ denotes the previous hidden state. Both $\mathbf{W}_r$ and $\mathbf{U}_r$ are learnable weight matrices.

Similarly, the *update* gate z_j is calculated by:

$$z_j = \sigma([\mathbf{W}_z\mathbf{x}]_j + [\mathbf{U}_z\mathbf{h}_{t-1}]_j) \tag{2}$$

The actual activation of the proposed unit h_j is then calculated by:

$$h_j^t = z_j h_j^{t-1} + (1 - z_j)\widetilde{h}_j^t \tag{3}$$

where

$$\widetilde{h}_j^t = \phi([\mathbf{W}\mathbf{x}]_j + [\mathbf{U}(\mathbf{r} \odot \mathbf{h}_{t-1})]_j) \tag{4}$$

ϕ denotes *tanh* activate function, and $\odot$ indicates element-wise multiplication. **W** and **U** are weight matrices which are learned. As described in [54], the reset gate **r** effectively allows the hidden state to drop any information that is found to be irrelevant later in the future, which provides a more compact information representation. On the other side, the update gate **z** dominates how much information from the previous hidden state will carry over to the current hidden state. More details about GRU can be seen in Figure 2.

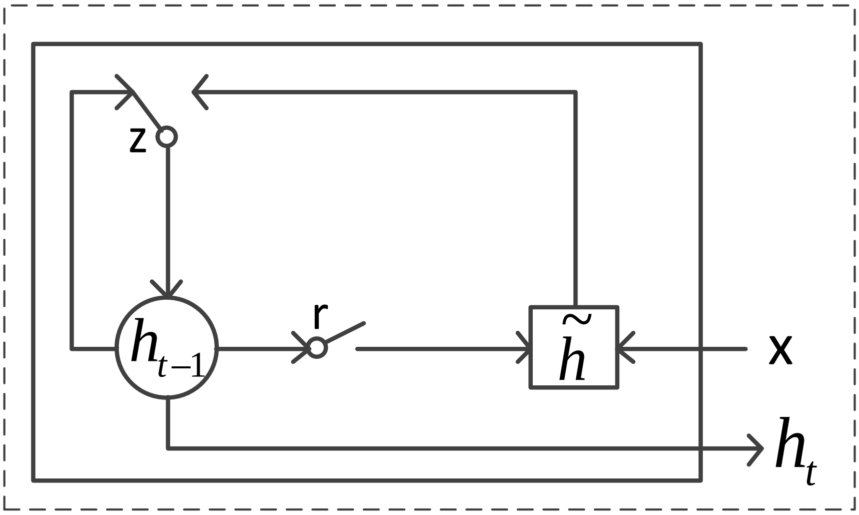

Figure 2. An illustration of a gated recurrent unit (GRU) [54]. The update gate z selects whether the hidden state h_t is to be updated with a new hidden state $\widetilde{h}$. The reset gate r decides whether the previous hidden state h_{t-1} is ignored.

2.3. The RoIAlign Layer

RoIAlign [13] is based on RoIPooling [10]. As we know, RoIPooling performs two quantizations, first quantizing a floating-number RoI to the discrete granularity of the feature map and then subdividing the quantized RoI into spatial bins which are themselves quantized. Unlike RoIPooling, RoIAlign avoids any quantization of the RoI boundaries or bins. In the execution of RoIAlign, bilinear interpolation [55] is exploited to calculate the exact values of the input features at four regularly sampled locations in each RoI bin. The result after bilinear interpolation is aggregated by average pooling.

3. Proposed Framework

The flowchart of the proposed object detection method is shown in Figure 3. The framework is based on the VGG16 model [56] and the popular detection frame Faster R-CNN [10]. First, given a remote sensing image, we employ the parts of VGG16 to extract object features and use the region proposal network (RPN) to generate region proposals. Unlike the work of Faster R-CNN using a RoI pooling layer to convert the features inside any valid region of interest into a small feature map with a fixed spatial extent, we apply the RoIAlign layer proposed in Mask R-CNN. There are misalignments between the RoIs and the extracted features in RoI pooling. RoIAlign can address the problem of misalignments introduced by quantizations, thus, enhancing the ability to detect small and intensive objects. Second, motivated by the work in [51] and for adapting to remote sensing images which contain complex backgrounds, we extract both local contextual information and object-object relationship contextual information, and fuse them by GRU. The fused feature is employed subsequently to obtain the classification and regression results of the contextual information fusion sub-network. Then, we divide the object in candidate regions generated by RPN into several parts and utilize the RoIAlign layer to pool each part. All parts are merged to gain better feature representations for detecting objects. After that, we perform classification and regression to obtain the consequences of the part-based multi-region sub-network. Finally, in the case of separately gaining results of the contextual information fusion sub-network, the part-based multi-region fusion sub-network, and the baseline sub-network, we execute a decision fusion on those results to acquire the bottom detection result, which we call multi-model decision fusion. Each component of the proposed framework is described as follows.

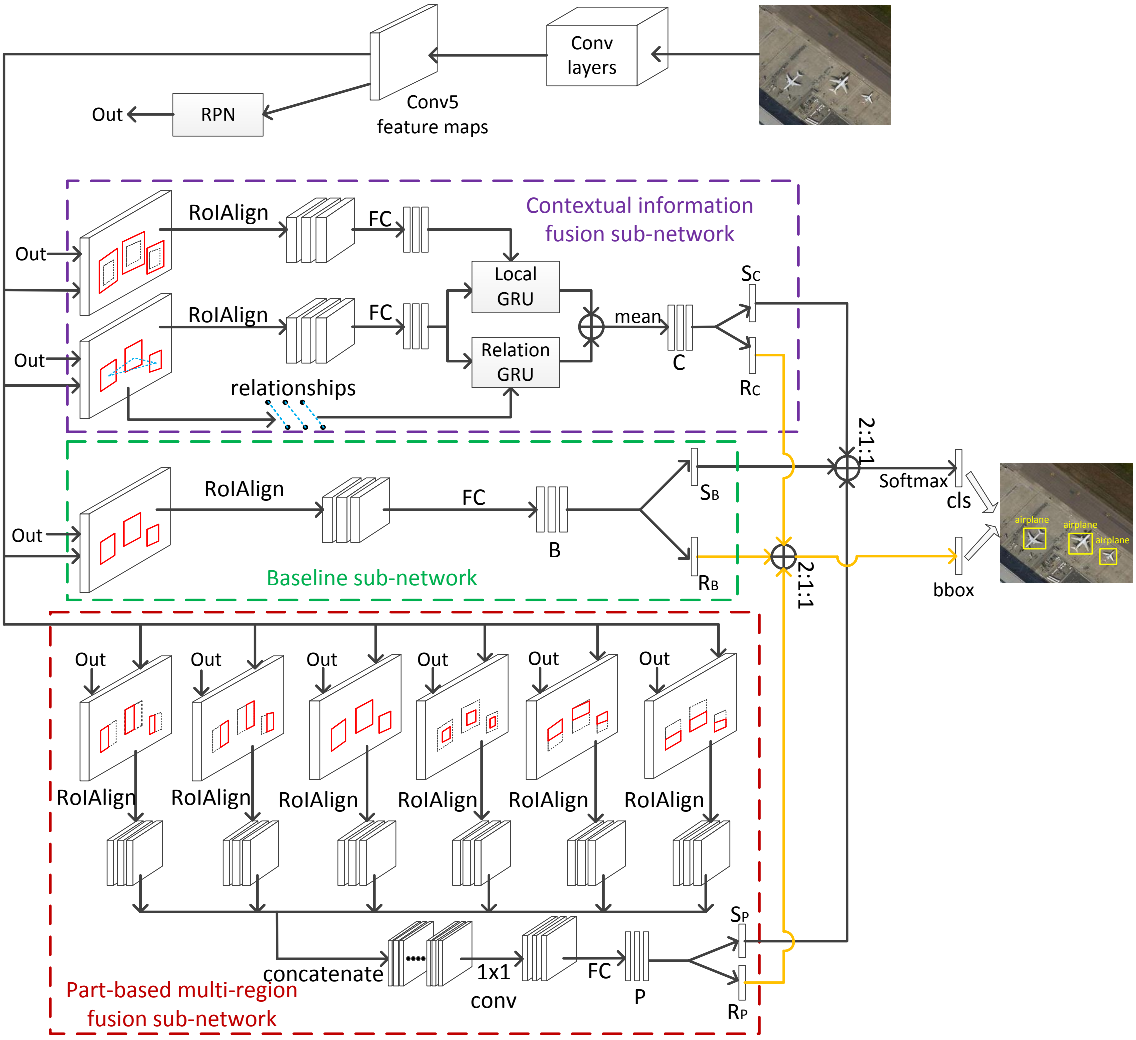

Figure 3. The proposed framework, which is made up of four parts. (1) A contextual information fusion sub-network; (2) a part-based multi-region fusion sub-network; (3) a baseline sub-network; (4) the last multi-model decision fusion part.

3.1. Local Contextual Information and Object-Object Relationship Contextual Information Fusion Sub-Network

Many works show the effectiveness of investigating features surrounding the regions of interest or relationships among objects [20,51]. Therefore, for object detection in remote sensing images, inspired by the work in [51], we construct our local contextual information and object-object relationship contextual information fusion sub-network. Different from [51] using global contextual information for the entire image, we employ local contextual features around objects. For some objects in remote sensing images, scenes far from them are more diverse, resulting in unstable contexts which are likely to be noise that affects the detection result. That is the reason we choose to exploit local contextual information for geospatial object detection. In addition, we replace RoI pooling with RoIAlign because of there existing a lot of dense and small objects in remote sensing images. The features to be fused in the sub-network consist of three parts: local contextual information, features in original candidate regions, and object-object relationship contextual information.

First, in conv5 layer, we extract the features from original proposal boxes and the 1.8× of original proposal boxes. The features in 1.8× of original proposal boxes are used as local contextual information. The RoIAlign layer and the fully connected layer act on the two types of features in succession. Second, we build relationships among objects [as illustrated in Figure 4]. The process is the same as [51]. There we set V to represent the collection of candidate boxes generated by RPN. The term v_i indicates the i-th candidate box. We calculate the relationship between v_i and v_j by:

$$e_{j\to i} = relu(W_p R^p_{j\to i}) * tanh(W_v[f^v_i, f^v_j]) \tag{5}$$

where $e_{j\to i}$ represents the influence of v_j on v_i and it is a scalar weight. W_p and W_v are weight matrices which are learned. The visual relationship vector is formed by concatenating visual feature f^v_i and f^v_j, indicated by $[f^v_i, f^v_j]$. The term $R^p_{j\to i}$ denotes the spatial position relationship. Visual feature f^v_i and f^v_j are results after $relu$, which are sparse. A lot of information will be lost if $relu$ is used again. So $tanh$ is applied to activate $W_v[f^v_i, f^v_j]$. $R^p_{j\to i}$ is obtained by:

$$\begin{aligned} R^p_{j\to i} = [&w_i, h_i, s_i, w_j, h_j, s_j, \frac{(x_i - x_j)}{w_j}, \frac{(y_i - y_j)}{h_j}, \\ &\frac{(x_i - x_j)^2}{w_j^2}, \frac{(y_i - y_j)^2}{h_j^2}, log(\frac{w_i}{w_j}), log(\frac{h_i}{h_j})] \end{aligned} \tag{6}$$

where (x_i, y_i) means the center of RoI b_i. w_i and h_i are the width and height of b_i. s_i is the area of b_i. The final object-object relationship contextual information m_i is calculated by:

$$m_i = \max_{j\in V} pooling(e_{j\to i} * f^v_j) \tag{7}$$

It represents that we choose the box which has the greatest impact on v_i as the final relationship contextual message to be integrated. Then, we exploit GRUs to merge the three features gained in the previous operation, taking the processed features from original proposed boxes as the initial hidden states, both the relationship contexts and the processed features (local contextual information) which stem from 1.8× of original proposed boxes as inputs related to two GRUs. Afterwards, we average the outputs of the two GRUs and denote the final feature as C. Finally, we apply C to gain the class scores S_C and the predicted boxes R_C.

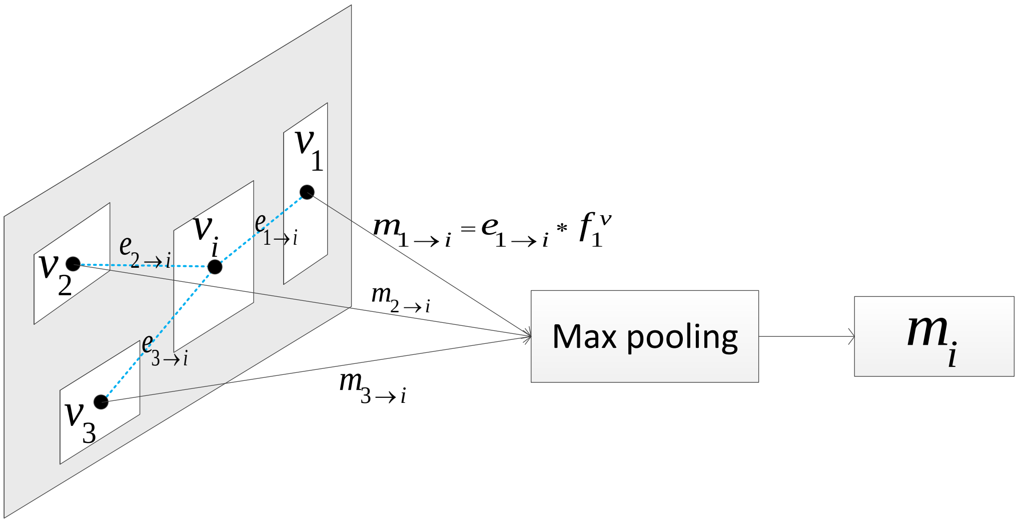

Figure 4. An illustration of building object-object relationship. The process is the same as [51]. For object v_i, the message $m_{1\to i}$ from object v_1 to object v_i is controlled by $e_{1\to i}$.

For large optical remote sensing images, it is necessary to use object-object relationship contextual information within meaningful limited regions in images instead of the entire images. That is because the effect of object-object relationship contextual information on the detection result is very little if the distance between two objects is too long. The images used in this paper are 400 pixels wide and 400 pixels high, just like limited regions cropped from large remote sensing images. Therefore we can obtain object-object relationship contextual information in the entire images.

3.2. Part-Based Multi-Region Fusion Sub-Network

For a specific object proposal, paying attention to each part of the object in it can help to obtain much useful spatial structure information about the object, so we can obtain more semantic information for better object detection performance. We use multiple parts of each object to acquire more local visual properties and geometric information, providing an enhanced feature representation.

The parts used include the original proposal box, the left-half part of the proposal box, the right-half part of the proposal box, the up-half part of the proposal box, the bottom-half part of the proposal box, and the inner part obtained by scaling the proposal box by a factor of 0.7 (see Figure 5). First, we gain those parts of each candidate region produced by RPN and perform the RoIAlign operation soon after. Second, we concatenate the pooled features along the channel axis. Then, a 1 × 1 convolution is implemented to reduce the dimension of the concatenated feature, which makes the feature adapt to the input shape of the fully connected layer. Later, the feature is fed into a fully connected layer to generate the final feature representation with more semantic information. We denote the final feature representation as P. Finally, we utilize P to gain the class scores S_P and the predicted boxes R_P.

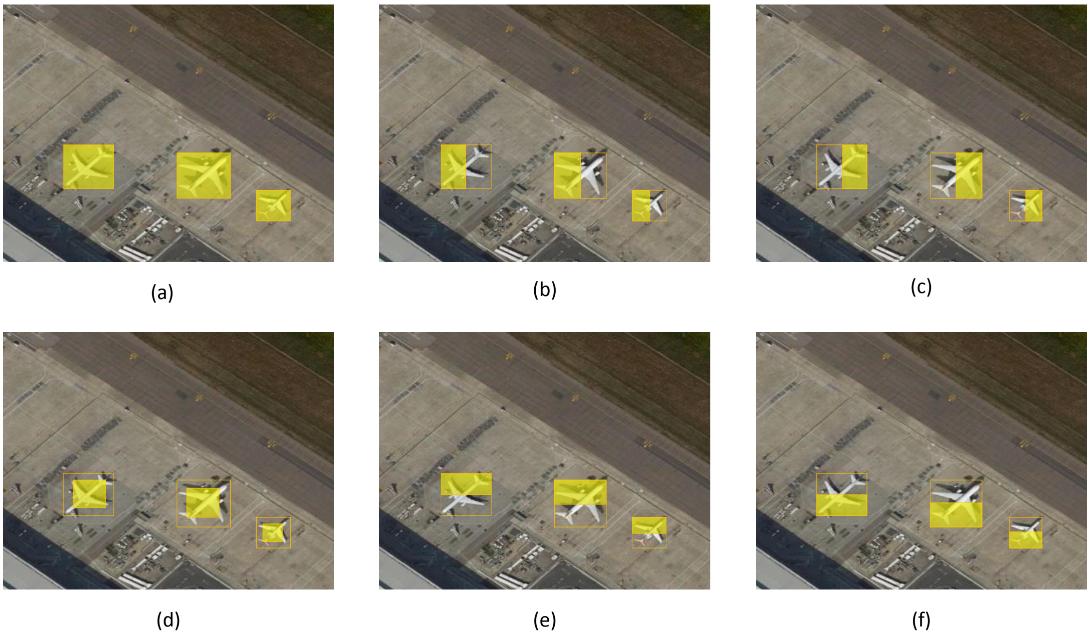

Figure 5. Illustration of object parts used in the proposed framework. (**a**) Original candidate boxes. (**b**) Left-half part of candidate boxes. (**c**) Right-half part of candidate boxes. (**d**) Inner part obtained by scaling candidate boxes by a factor of 0.7. (**e**) Up-half part of candidate boxes. (**f**) Bottom-half part of candidate boxes.

3.3. Multi-Model Decision Fusion Strategy

The multi-model decision fusion strategy, relying on several detection results, is more robust compared to the single model which may cause much false detection. In addition to exploiting the contextual information fusion sub-network and the part-based multi-region fusion sub-network, we also utilize a baseline sub-network that only uses the original proposal regions for object detection. In the baseline sub-network, we perform the RoIAlign operator as same as the two aforementioned sub-networks. Then we employ a fully connected layer to obtain the final feature denoted as B. Finally, we use B to gain the class scores S_B and the predicted boxes R_B.

After obtaining the three types of class scores S_C, S_P, S_B and predicted boxes R_C, R_P, R_B, we make a decision fusion on them. The decision fusion ratio of S_C, S_P, and S_B is 2:1:1, so do R_C, R_P, and R_B, which can provide better detection results in experiments. Then, we use a softmax layer to get the final class labels of all predicted boxes. The loss function employed in this paper is as same as that in Faster R-CNN [10].

4. Experiments and Results

In this part, we first introduce the data set and evaluation metrics used for the experiments. Then, we describe the implementation details and parameter settings of the proposed method. The results and some comparisons to other methods are discussed afterward. The models were trained on a computer with two Intel Xeon E5-2630 v4 CPUs and two NVIDIA GeForce GTX 1080 GPUs. The operating system and deep learning platform used were Ubuntu 16.04 and TensorFlow 1.3.0, respectively.

4.1. Data Set

We evaluate the performance of the proposed object detection method on a publicly available data set: NWPU VHR-10-v2 data set [20]. The data set stems from the positive image set of the original NWPU VHR-10 data set [31] and still contains ten classes of geospatial objects, including airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle. There are 1172 images (400 × 400 pixels) in the data set we use. The data set is challenging, because the objects are multi-category and multi-scale and the backgrounds are complex. In all experiments, the training data and test data we employ are the same as that in [20], 879 (75% of the data set) remote sensing images in the training data and 293 images in the test data.

4.2. Evaluation Metrics

Here, we evaluate the performance of object detection methods through two standard, universally agreed and widely used measures illustrated in [7], namely precision-recall curve (PRC) and average precision (AP).

4.2.1. Precision-Recall Curve (PRC)

The Precision metric measures the fraction of detections which are true positives, and the Recall metric weighs the fraction of positives which are correctly recognized. The number of true positives, the number of false positives, and the number of false negatives are denoted as TP, FP, and FN, respectively. Therefore, the Precision and Recall metrics can be obtained by:

$$Precision = \frac{TP}{(TP + FP)} \tag{8}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{9}$$

The PRC metric is based on the overlapping area between the detection and the ground truth object. A detection is considered to be a true positive if the intersection over union (IoU) between the detection and the ground truth box exceeds a predetermined threshold; otherwise, the detection is marked as a false positive. What is more, if several detections overlap with a same ground truth bounding box, only one is regarded as the true positive, and others are labeled as false positives. The intersection over union IoU is formulated as:

$$IoU = \frac{area(detection \cap groundtruth)}{area(detection \cup groundtruth)} \tag{10}$$

4.2.2. Average Precision (AP)

The AP calculates the average value of Precision over the interval from Recall = 0 to Recall = 1, namely the area under the PRC. Therefore, the higher the AP value, the better the performance, and vice versa.

4.3. Implementation Details and Parameter Settings

The proposed model is based on the successful VGG16 network [56] that was pretrained on ImageNet [57]. To augment the training data, we flip all the training images horizontally. For training our model, we utilize the stochastic gradient descent with 0.9 momentum. The learning rate is initialized to 0.001 and we use it for 20 k iterations; then we continue training for 10k iterations with 0.0001. The last fully connected layers for classification and bounding box regression are randomly initialized with zero-mean Gaussian distributions with standard deviations of 0.001, simultaneously other fully connected layers and the 1 × 1 convolutional layer with standard deviations of 0.01. Biases are initialized to 0. For training RPN, each mini-batch arises from a single image which includes many positive and negative example anchors, and we randomly sample 128 anchors in an image to calculate the loss function of a mini-batch. The sampled positive and negative anchors have a ratio of up to 1:1. If there are fewer than 64 positive samples in an image, we pad the mini-batch with negative ones. The entire model is trained end-to-end. Furthermore, we consider a detection to be correct if the IoU between the predicted bounding box and the ground truth bounding box exceeds 0.5. Otherwise, the detection is considered as a false positive. In the implementation of the test, we employ Soft-NMS to reduce redundancy for better detection performance.

4.4. Evaluation of Local Contextual Information and Object-Object Relationship Contextual Information Fusion Sub-Network

To evaluate the efficiency of our local contextual information and object-object relationship contextual information fusion sub-network, we designed a basic set of experiments. First, we run the standard Faster R-CNN model as a benchmark experiment. Then, on the basis of the baseline sub-network, we incorporate the proposed sub-network which fuses both local contextual information and object-object relationship contextual information. In the experiments, we find that using the features extracted from the 1.8× of the original proposal boxes as local contextual features leads to better detection performance. In the field of remote sensing image object detection, some works [18,20,27] take local contextual information into account and therefore obtain good results. However, the object-object relationship contextual information has not been proven to be beneficial for detecting geospatial objects. To illustrate the usefulness of the object-object relationship contextual information, we implement an experiment in which we incorporate the sub-network only containing local contextual information into the baseline sub-network. The detailed experimental results are summarized in Table 1. As shown in Table 1, an improvement of 4.24 percent points in mean average precision (mAP) can be seen by adding the local contextual information and object-object relationship contextual information fusion sub-network compared to the Faster R-CNN baseline network. This validates that our local contextual information and object-object relationship contextual information fusion sub-network has a strong discriminating ability to represent features of geospatial objects, providing useful contextual cues for better detection performance. In addition, Table 1 shows the mAP improves from 92.42% (only using local contextual information) to 94.04% (using both local contextual information and object-object relationship contextual information), demonstrating that the object-object relationship contextual information plays an important role in achieving better detection performance for geospatial object detection. Furthermore, we execute an experiment to illustrate that local contextual information is more useful than global contextual information for the entire image in remote sensing image object detection. In the experiment, we replace local contextual information with global contextual information for the entire remote sensing image in the overall proposed framework. The results are shown in Table 1. As we can see, in terms of mAP over all ten object categories, applying local contextual information outperforms the use of global contextual information for the entire image by 2.4%. This demonstrates that the use of local contextual information is critical, leading to better detection results than using global contextual information for the entire remote sensing image.

4.5. Evaluation of Part-Based Multi-Region Fusion Network

To verify that the part-based multi-region fusion sub-network has a positive effect on geospatial object detection, we compared the overall proposed model (including the part-based multi-region fusion sub-network) with the previous variant where the framework only merges the baseline sub-network and the local contextual information and object-object relationship contextual information fusion sub-network. As can be seen from Table 1, incorporating the part-based multi-region fusion sub-network offers a further performance increase of 1.0 percent point. This demonstrates that fusing multiple parts of each geospatial object can investigate more spatial structural information about objects, which helps to diversify object features and enhance semantic information for forming powerful feature representation.

4.6. Evaluation of Multi-model Decision Fusion Strategy

In the proposed approach, we make a decision fusion on the results of three sub-networks, which include the local contextual information and object-object relationship contextual information fusion sub-network, the part-based multi-region fusion sub-network, and the baseline sub-network. To evaluate the effectiveness of the decision fusion ratio of 2:1:1 corresponding to those three sub-networks, we set 25 different ratios for contrast. These ratios consist of 1:1:1, 1:1:2, 1:1:3, 1:2:1, 1:2:2, 1:2:3, 1:3:1, 1:3:2, 1:3:3, 2:1:1, 2:1:2, 2:1:3, 2:2:1, 2:2:3, 2:3:1, 2:3:2, 2:3:3, 3:1:1, 3:1:2, 3:1:3, 3:2:1, 3:2:2, 3:2:3, 3:3:1, 3:3:2. The experimental results are illustrated in Table 2. As we can see, using the fusion ratio of 2:1:1 achieves the best result among all the experimental results, gaining a mAP value 95.04%. This indicates that the set fusion ratio of 2:1:1 is beneficial to the detection.

Table 1. Detection results of using sub-networks. C-Gl: Incorporate the Contextual Information Fusion Sub-network only containing global contextual information for the entire image. C-Lo: Incorporate the Contextual Information Fusion Sub-network only containing local contextual information. C-Re: Incorporate the Contextual Information Fusion Sub-network only containing object-object relationship contextual information. P: Incorporate the Part-based Multi-region Fusion Sub-network.

	C			P	mAP	Airplane	Ship	Storage Tank	Baseball Diamond	Tennis Court	Basketball Court	Ground Track Field	Harbor	Bridge	Vehicle
	C-Gl	C-Lo	C-Re												
Faster R-CNN (Baseline)					0.8980	1.0000	0.9225	0.9415	0.9521	0.9267	0.8429	1.0000	0.8788	0.6899	0.8254
ours		√			0.9242	1.0000	0.9106	0.9523	0.9593	0.9554	0.9116	1.0000	0.9235	0.7419	0.8873
ours		√	√		0.9404	0.9999	0.9184	0.9898	0.9757	0.9545	0.9484	0.9994	0.9497	0.7605	0.9072
ours		√	√	√	0.9504	0.9934	0.9227	0.9918	0.9668	0.9632	0.9756	1.0000	0.9740	0.8027	0.9136
ours	√		√	√	0.9264	0.9999	0.9139	0.9618	0.9630	0.9493	0.9424	1.0000	0.9172	0.7051	0.9115

Table 2. Comparison detection results of 25 different decision fusion ratios.

Fusion Ratio	mAP	Airplane	Ship	Storage Tank	Baseball Diamond	Tennis Court	Basketball Court	Ground Track Field	Harbor	Bridge	Vehicle
1:1:1	0.9386	1.0000	0.9303	0.9741	0.9740	0.9439	0.9506	1.0000	0.9689	0.7406	0.9029
1:1:2	0.9337	1.0000	0.9104	0.9616	0.9617	0.9421	0.9471	1.0000	0.9686	0.7421	0.9032
1:1:3	0.9345	1.0000	0.9061	0.9557	0.9748	0.9420	0.9459	1.0000	0.9715	0.7414	0.9079
1:2:1	0.9416	1.0000	0.9142	0.9921	0.9758	0.9557	0.9631	1.0000	0.9381	0.7603	0.9169
1:2:2	0.9320	1.0000	0.8993	0.9756	0.9528	0.9422	0.9506	1.0000	0.9601	0.7359	0.9033
1:2:3	0.9313	1.0000	0.9107	0.9696	0.9438	0.9414	0.9471	1.0000	0.9693	0.7285	0.9031
1:3:1	0.9403	1.0000	0.9131	0.9717	0.9774	0.9628	0.9500	1.0000	0.9607	0.7656	0.9012
1:3:2	0.9339	1.0000	0.9304	0.9762	0.9512	0.9412	0.9500	1.0000	0.9535	0.7345	0.9022
1:3:3	0.9330	1.0000	0.9315	0.9752	0.9583	0.9413	0.9462	1.0000	0.9576	0.7174	0.9028
2:1:1	**0.9504**	**0.9934**	**0.9227**	**0.9918**	**0.9668**	**0.9632**	**0.9756**	**1.0000**	**0.9740**	**0.8027**	**0.9136**
2:1:2	0.9391	1.0000	0.9204	0.9623	0.9743	0.9445	0.9495	1.0000	0.9705	0.7641	0.9053
2:1:3	0.9356	1.0000	0.9103	0.9564	0.9748	0.9440	0.9495	1.0000	0.9716	0.7476	0.9015
2:2:1	0.9379	0.9999	0.8866	0.9680	0.9661	0.9599	0.9512	1.0000	0.9598	0.7814	0.9059
2:2:3	0.9355	1.0000	0.9357	0.9710	0.9373	0.9436	0.9495	1.0000	0.9685	0.7465	0.9032
2:3:1	0.9352	1.0000	0.9136	0.9762	0.9621	0.9430	0.9512	1.0000	0.9502	0.7536	0.9025
2:3:2	0.9363	1.0000	0.9306	0.9762	0.9646	0.9426	0.9500	1.0000	0.9567	0.7398	0.9029
2:3:3	0.9337	1.0000	0.9281	0.9727	0.9436	0.9429	0.9506	1.0000	0.9598	0.7356	0.9032
3:1:1	0.9381	0.9934	0.9468	0.9768	0.9660	0.9798	0.9512	1.0000	0.9326	0.7674	0.8671
3:1:2	0.9405	1.0000	0.9325	0.9615	0.9741	0.9456	0.9495	1.0000	0.9705	0.7704	0.9014
3:1:3	0.9314	1.0000	0.8675	0.9605	0.9735	0.9447	0.9495	1.0000	0.9714	0.7459	0.9016
3:2:1	0.9383	1.0000	0.9142	0.9704	0.9734	0.9453	0.9500	1.0000	0.9662	0.7611	0.9023
3:2:2	0.9399	1.0000	0.9309	0.9699	0.9746	0.9447	0.9500	1.0000	0.9705	0.7554	0.9029
3:2:3	0.9405	1.0000	0.9344	0.9659	0.9740	0.9445	0.9495	1.0000	0.9707	0.7636	0.9029
3:3:1	0.9361	1.0000	0.9216	0.9762	0.9601	0.9445	0.9506	1.0000	0.9486	0.7565	0.9027
3:3:2	0.9371	1.0000	0.9298	0.9762	0.9676	0.9443	0.9500	1.0000	0.9591	0.7408	0.9033

4.7. Comparisons with Other Detection Methods

We compared the proposed approach with five state-of-the-art methods, including the collection of part detector (COPD) [31], a transferred CNN model from AlexNet [58], the rotation-invariant convolutional neural network (RICNN) [17], the rotation-insensitive and context-augmented object detector (RICAOD) [20], and Faster R-CNN [10]. In the implementation of the ten-class object detection task, the COPD is made up of 45 seed-based part detectors. Each part detector is a linear support vector machine (SVM) classifier and corresponds to a particular viewpoint of an object class, therefore the collection of them providing a solution for rotation-invariant detection of multi-class objects. Exploited as a common CNN feature extractor, the transferred CNN model has shown great success for PASCAL Visual Object Classes object detection. For dealing with the problem of object rotation variations, the RICNN is designed to introduce and learn a new rotation-invariant layer on the basis of the existing CNN architecture, AlexNet. The RICAOD utilizes multiangle anchors for rotation-invariant object detection and combines local and contextual features to address the problem of appearance ambiguity. The quantitative comparison results of the six different methods are shown in Table 3 and Figure 6, representing the AP values and PRCs, respectively. As can be observed in Table 3, in terms of mean AP over all ten object categories, the proposed approach outperforms the COPD method [31], the transferred CNN method [58], the RICNN method [17], the RICAOD method [20], and the Faster R-CNN method [10] by 40.15%, 35.43%, 21.93%, 7.92%, and 5.24%, respectively. In addition, we also obtain good detection accuracy in each category, especially airplane, storage tank, basketball, ground track field, and harbor, with very high AP values. Those fully demonstrate that the proposed method achieves much better performance compared to the existing state-of-the-art methods. Table 3 also shows the average running time of each image for the six different approaches. We can observe that the proposed method costs less computation time than other methods except Faster R-CNN.

For all results, it can be easily illustrated: due to the use of the contextual features containing local contextual features and object-object relationship contextual features, the proposed method obtains a discriminative feature representation ability to effectively recognize objects in spite of the diversity and complexity of object appearance, such as storage tank, bridge, and so on; the part-based multi-region fusion sub-network provides more spatial structural information about objects, so that more semantic information can be obtained to enhance the feature representation; the multi-model decision fusion strategy makes the algorithm more robust and provides better detection performance, because it acts like operating on three different single CNN-based models, each of which generates representative characteristics that describe the object.

Figure 7 shows a lot of geospatial object detection results. The green boxes denote true positives; the red boxes denote false positives; the yellow boxes indicate false negatives.

Table 3. Comparison detection results of six different methods.

	mAP	Airplane	Ship	Storage Tank	Baseball Diamond	Tennis Court	Basketball Court	Ground Track Field	Harbor	Bridge	Vehicle	Time per Image (Second)
COPD [31]	0.5489	0.6225	0.6937	0.6452	0.8213	0.3413	0.3525	0.8421	0.5631	0.1643	0.4428	1.16
Transferred CNN [58]	0.5961	0.6603	0.5713	0.8501	0.8093	0.3511	0.4552	0.7937	0.6257	0.4317	0.4127	5.09
RICNN [17]	0.7311	0.8871	0.7834	0.8633	0.8909	0.4233	0.5685	0.8772	0.6747	0.6231	0.7201	8.47
RICAOD [20]	0.8712	0.9970	0.9080	0.9061	0.9291	0.9029	0.8031	0.9081	0.8029	0.6853	0.8714	2.89
Faster R-CNN [10]	0.8980	**1.0000**	0.9225	0.9415	0.9521	0.9267	0.8429	1.0000	0.8788	0.6899	0.8254	**0.09**
ours	**0.9504**	0.9934	**0.9227**	**0.9918**	**0.9668**	**0.9632**	**0.9756**	**1.0000**	**0.9740**	**0.8027**	**0.9136**	0.75

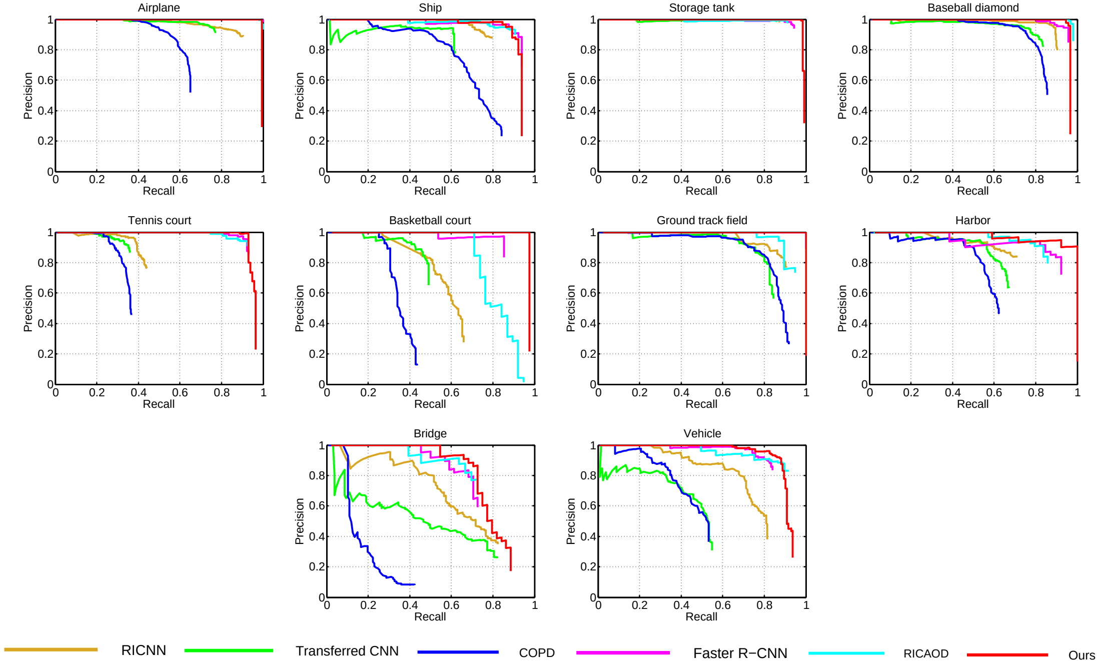

Figure 6. Precision-recall curves (PRCs) of the proposed method and other state-of-the-art methods for airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, vehicle classes, respectively.

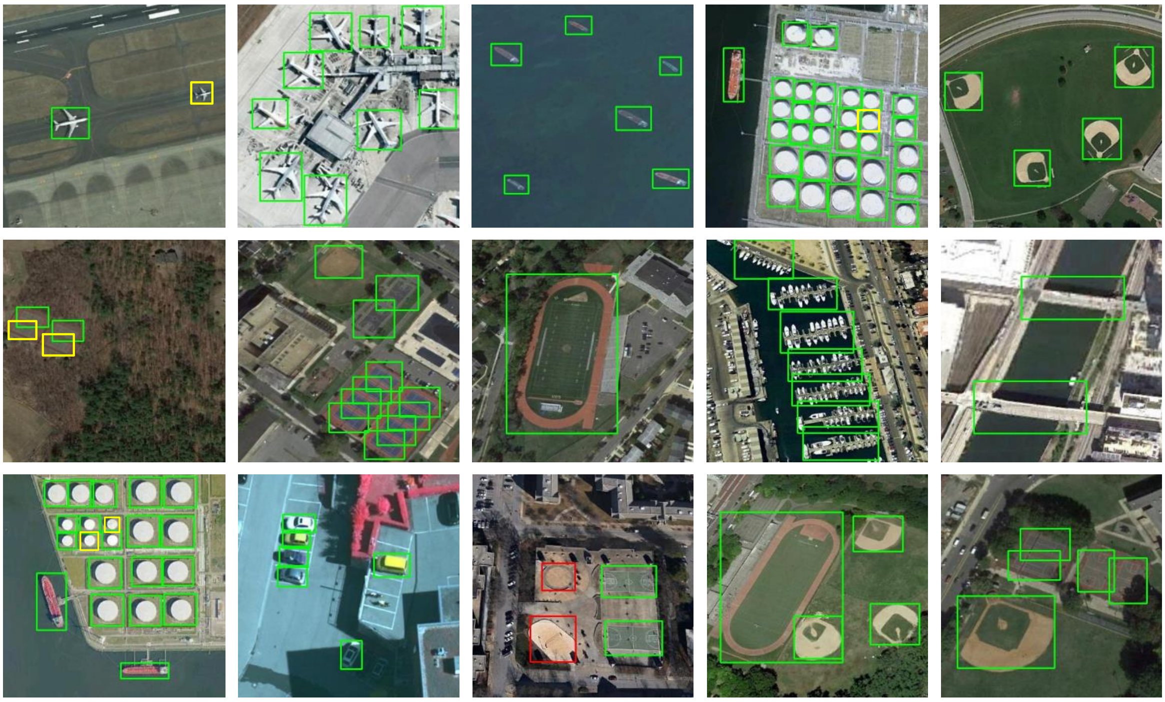

Figure 7. Some object detection results obtained by using the proposed method. The true positives, false positives, and false negatives are denoted by green, red, and yellow rectangles, respectively.

5. Conclusions

In this paper, we proposed a multi-model decision fusion framework for geospatial object detection. The framework combines a contextual information fusion sub-network, a part-based multi-region fusion sub-network, and a baseline sub-network to recognize and locate geospatial objects. The final detection results are obtained by way of making a decision fusion on the results of the three sub-networks. The proposed model presents a remarkable performance on the publicly available data set NWPU VHR-10-v2. All experiments show that: (1) local contextual information and object-object relationship contextual information are beneficial to effectively recognizing objects and alleviating the mis-detection between different types of objects with similar appearance; (2) the part-based multi-region fusion sub-network can provide more details of objects to alleviate the insufficient understanding of geospatial object spatial structure information; (3) the multi-model decision fusion strategy can lead to a more stable and robust model and achieve better algorithm performance; (4) the proposed framework can produce more accurate object detection results than other previous methods. In future work, for better detection performance, we will continue to improve the proposed framework. Many fine details of some small objects are lost due to the implementation of pooling, which can lead to the inability to identify the objects. Therefore, we will consider the use of features from lower convolutional layers. In addition, we will consider designing an operator to obtain more accurate localization of detected objects.

Author Contributions: Investigation, W.M., Q.G., Y.W. and W.Z.; Methodology, W.M. and Y.W.; Supervision, L.J.; Validation, X.Z.; Writing—original draft, W.M. and Q.G.; Writing—review and editing,Y.W. and W.Z.

Funding: The research was jointly supported by the National Natural Science Foundations of China (Nos. 61702392, 61671350, 61772400), and the China Postdoctoral Science Foundation (Nos. 2018T111022, 2017M623127).

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

References

1. Ahmad, K.; Pogorelov, K.; Riegler, M.; Conci, N.; Halvorsen, P. Social media and satellites. *Multimed. Tools Appl.* **2019**, *78*, 2837–2875. [CrossRef]
2. Ahmad, K.; Pogorelov, K.; Riegler, M.; Ostroukhova, O.; Halvorsen, P.; Conci, N.; Dahyot, R. Automatic detection of passable roads after floods in remote sensed and social media data. *arXiv* **2019**, *arXiv:1901.03298*.
3. Sirmacek, B.; Unsalan, C. Urban-Area and Building Detection Using SIFT Keypoints and Graph Theory. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 1156–1167. [CrossRef]
4. Tang, T.; Zhou, S.; Deng, Z.; Zou, H.; Lei, L. Vehicle Detection in Aerial Images Based on Region Convolutional Neural Networks and Hard Negative Example Mining. *Sensors* **2017**, *17*, 336. [CrossRef]
5. Tuermer, S.; Kurz, F.; Reinartz, P.; Stilla, U. Airborne Vehicle Detection in Dense Urban Areas Using HoG Features and Disparity Maps. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2013**, *6*, 2327–2337. [CrossRef]
6. 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.* **2014**, *52*, 4511–4523.
7. Cheng, G.; Han, J. A Survey on Object Detection in Optical Remote Sensing Images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
8. 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, 23–28 June 2014; pp. 580–587.
9. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1440–1448.
10. 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]
11. 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, 26 June–1 July 2016; pp. 779–788.
12. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.
13. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
14. 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, 11–18 December 2015; pp. 1134–1142.
15. Yang, X.; Sun, H.; Fu, K.; Yang, J.; Sun, X.; Yan, M.; Guo, Z. Automatic Ship Detection in Remote Sensing Images from Google Earth of Complex Scenes Based on Multiscale Rotation Dense Feature Pyramid Networks. *Remote Sens.* **2018**, *10*, 132. [CrossRef]
16. 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.* **2017**, *55*, 2486–2498. [CrossRef]
17. 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]
18. Xiao, Z.; Gong, Y.; Long, Y.; Li, D.; Wang, X.; Liu, H. Airport Detection Based on a Multiscale Fusion Feature for Optical Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1469–1473. [CrossRef]
19. Ren, Y.; Zhu, C.; Xiao, S. Deformable Faster R-CNN with Aggregating Multi-Layer Features for Partially Occluded Object Detection in Optical Remote Sensing Images. *Remote Sens.* **2018**, *10*, 1470. [CrossRef]
20. Li, K.; Cheng, G.; Bu, S.; You, X. Rotation-Insensitive and Context-Augmented Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2337–2348. [CrossRef]
21. Yang, Y.; Zhuang, Y.; Bi, F.; Shi, H.; Xie, Y. M-FCN: Effective Fully Convolutional Network-Based Airplane Detection Framework. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1293–1297. [CrossRef]
22. Xu, Y.; Zhu, M.; Li, S.; Feng, H.; Ma, S.; Che, J. End-to-End Airport Detection in Remote Sensing Images Combining Cascade Region Proposal Networks and Multi-Threshold Detection Networks. *Remote Sens.* **2018**, *10*, 1516. [CrossRef]

23. 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]
24. Chen, S.; Zhan, R.; Zhang, J. Geospatial Object Detection in Remote Sensing Imagery Based on Multiscale Single-Shot Detector with Activated Semantics. *Remote Sens.* **2018**, *10*, 820. [CrossRef]
25. Liu, Y.; Zhang, Z.; Zhong, R.; Chen, D.; Ke, Y.; Peethambaran, J.; Chen, C.; Sun, L. Multilevel Building Detection Framework in Remote Sensing Images Based on Convolutional Neural Networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 3688–3700. [CrossRef]
26. Wang, G.; Wang, X.; Fan, B.; Pan, C. Feature Extraction by Rotation-Invariant Matrix Representation for Object Detection in Aerial Image. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 851–855. [CrossRef]
27. Zhang, L.; Shi, Z.; Wu, J. A Hierarchical Oil Tank Detector With Deep Surrounding Features for High-Resolution Optical Satellite Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 4895–4909. [CrossRef]
28. Zhang, W.; Sun, X.; Wang, H.; Fu, K. A generic discriminative part-based model for geospatial object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2015**, *99*, 30–44. [CrossRef]
29. Zhang, W.; Sun, X.; Fu, K.; Wang, C.; Wang, H. Object Detection in High-Resolution Remote Sensing Images Using Rotation Invariant Parts Based Model. *IEEE Geosci. Remote Sens. Lett.* **2013**, *11*, 74–78. [CrossRef]
30. Cheng, G.; Han, J.; Guo, L.; Qian, X.; Zhou, P.; Yao, X.; Hu, X. Object detection in remote sensing imagery using a discriminatively trained mixture model. *ISPRS J. Photogramm. Remote Sens.* **2013**, *85*, 32–43. [CrossRef]
31. 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.* **2014**, *98*, 119–132. [CrossRef]
32. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Scalable multi-class geospatial object detection in high-spatial-resolution remote sensing images. In Proceedings of the IEEE Geoscience and Remote Sensing Symposium, Quebec City, QC, Canada, 13–18 July 2014; pp. 2479–2482.
33. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–26 June 2005; Volume 1, pp. 886–893.
34. 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.* **2015**, *53*, 3325–3337. [CrossRef]
35. Wang, Q.; He, X.; Li, X. Locality and Structure Regularized Low Rank Representation for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 911–923. [CrossRef]
36. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
37. Fei-Fei, L.; Perona, P. A Bayesian hierarchical model for learning natural scene categories. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–26 June 2005; Volume 2, pp. 524–531.
38. Zhang, D.; Han, J.; Cheng, G.; Liu, Z.; Bu, S.; Guo, L. Weakly Supervised Learning for Target Detection in Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2014**, *12*, 701–705. [CrossRef]
39. 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.
40. Han, J.; Zhou, P.; Zhang, D.; Cheng, G.; Guo, L.; Liu, Z.; Bu, S.; Wu, J. 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]
41. Li, Z.; Itti, L. Saliency and Gist Features for Target Detection in Satellite Images. *IEEE Trans. Image Process.* **2011**, *20*, 2017–2029. [PubMed]
42. Ma, W.; Zhang, J.; Wu, Y.; Jiao, L.; Zhu, H.; Zhao, W. A Novel Two-Step Registration Method for Remote Sensing Images Based on Deep and Local Features. *IEEE Trans. Geosci. Remote Sens.* **2019**, 1–10. [CrossRef]
43. Wang, Q.; Yuan, Z.; Du, Q.; Li, X. GETNET: A General End-to-End 2-D CNN Framework for Hyperspectral Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 3–13. [CrossRef]
44. Wang, Q.; Liu, S.; Chanussot, J.; Li, X. Scene Classification With Recurrent Attention of VHR Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 1155–1167. [CrossRef]

45. Nogueira, K.; Fadel, S.G.; Dourado, Í.C.; Werneck, R.D.O.; Muñoz, J.A.V.; Penatti, O.A.B.; Calumby, R.T.; Li, L.T.; dos Santos, J.A.; Torres, R.D.S. Exploiting ConvNet Diversity for Flooding Identification. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1446–1450. [CrossRef]
46. Wiewiora, E.; Galleguillos, C.; Vedaldi, A.; Belongie, S.; Rabinovich, A. Objects in Context. In Proceedings of the IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–20 October 2007; pp. 1–8.
47. Chen, Q.; Song, Z.; Dong, J.; Huang, Z.; Hua, Y.; Yan, S. Contextualizing Object Detection and Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 13–27. [CrossRef]
48. Li, J.; Wei, Y.; Liang, X.; Dong, J.; Xu, T.; Feng, J.; Yan, S. Attentive Contexts for Object Detection. *IEEE Trans. Multimed.* **2017**, *19*, 944–954. [CrossRef]
49. Bell, S.; Zitnick, C.L.; Bala, K.; Girshick, R. Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2874–2883.
50. Chen, X.; Gupta, A. Spatial Memory for Context Reasoning in Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4086–4096.
51. Liu, Y.; Wang, R.; Shan, S.; Chen, X. Structure Inference Net: Object Detection Using Scene-Level Context and Instance-Level Relationships. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6985–6994.
52. Hu, H.; Gu, J.; Zhang, Z.; Dai, J.; Wei, Y. Relation Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3588–3597.
53. Marcu, A.; Leordeanu, M. Dual Local-Global Contextual Pathways for Recognition in Aerial Imagery. *arXiv* **2016**, arXiv:1605.05462.
54. Cho, K.; van Merrienboer, B.; Gülçehre, Ç.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv* **2014**, arXiv:1406.1078.
55. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial Transformer Networks. In Proceedings of the Neural Information Processing Systems 2015, Montreal, QC, Canada, 7–12 December 2015; pp. 2017–2025.
56. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
57. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
58. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the International Conference on Neural Information Processing Systems, Lake Tahoe, ND, USA, 3–8 December 2012; pp. 1097–1105.