remote sensing

MDPI

Article

An Efficient and Robust Integrated Geospatial Object Detection Framework for High Spatial Resolution Remote Sensing Imagery

Xiaobing Han [1,2], Yanfei Zhong [1,2,*] and Liangpei Zhang [1,2]

[1] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China; whu_hxb@163.com (X.H.); zlp62@whu.edu.cn (L.Z.)
[2] Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China
* Correspondence: zhongyanfei@whu.edu.cn; Tel.: +86-27-6877-9969

Academic Editors: Lizhe Wang, Liping Di, Qian Du, Peng Liu and Prasad S. Thenkabail
Received: 30 April 2017; Accepted: 23 June 2017; Published: 28 June 2017

Abstract: Geospatial object detection from high spatial resolution (HSR) remote sensing imagery is a significant and challenging problem when further analyzing object-related information for civil and engineering applications. However, the computational efficiency and the separate region generation and localization steps are two big obstacles for the performance improvement of the traditional convolutional neural network (CNN)-based object detection methods. Although recent object detection methods based on CNN can extract features automatically, these methods still separate the feature extraction and detection stages, resulting in high time consumption and low efficiency. As a significant influencing factor, the acquisition of a large quantity of manually annotated samples for HSR remote sensing imagery objects requires expert experience, which is expensive and unreliable. Despite the progress made in natural image object detection fields, the complex object distribution makes it difficult to directly deal with the HSR remote sensing imagery object detection task. To solve the above problems, a highly efficient and robust integrated geospatial object detection framework based on faster region-based convolutional neural network (Faster R-CNN) is proposed in this paper. The proposed method realizes the integrated procedure by sharing features between the region proposal generation stage and the object detection stage. In addition, a pre-training mechanism is utilized to improve the efficiency of the multi-class geospatial object detection by transfer learning from the natural imagery domain to the HSR remote sensing imagery domain. Extensive experiments and comprehensive evaluations on a publicly available 10-class object detection dataset were conducted to evaluate the proposed method.

Keywords: geospatial object detection; high spatial resolution (HSR) remote sensing imagery; integration; pre-training mechanism; feature sharing

1. Introduction

Geospatial object detection from remote sensing imagery is an important tool when analyzing object-related information [1–3]. High spatial resolution (HSR) remote sensing imaging sensors can now acquire aerial and satellite images with abundant detail and complex spatial structural information, which can be used in a wide range of civil and engineering applications, such as segmentation [4], scene annotation [5], object detection [6–22] (e.g., airplane detection [6,12], urban area detection [13], vehicle detection [21,22]), scene classification and recognition [23–27], etc. Differing from natural imagery obtained by the camera on the ground from a horizontal view, HSR remote sensing imagery is obtained by satellite-borne or space-borne sensors from a top-down view, which is an approach that can be easily influenced by weather and illumination conditions. In addition, differing from the anteroposterior

object position from natural imagery, the position of the objects in HSR remote sensing imagery is mostly left–right. Before executing the object detection task, the term "object" for HSR remote sensing imagery should be defined. Specifically, objects in HSR remote sensing imagery include not only man-made objects (e.g., vehicles, ships, buildings, etc.) with sharp boundaries that are independent of the background environment, but also landscape objects, such as land-use/land-cover (LULC) parcels with vague boundaries [28]. As HSR remote sensing imagery contains various geospatial objects, the accurate detection of multi-class geospatial objects is of vital importance. However, multi-class geospatial object detection from HSR remote sensing imagery is a significant and challenging task for three main reasons. The first reason is the imaging conditions of HSR remote sensing imagery, which include large variations in the visual appearance of objects, caused by viewpoint variation, occlusion, background clutter, illumination, shadow, etc. The second reason is the small-size and scale-variable properties of the multi-class geospatial objects compared with the large-scale complex backgrounds in HSR remote sensing imagery. The third reason is the relative dearth of manually annotated samples for the geospatial object training data. Because of the challenging nature of multi-class geospatial object detection from HSR remote sensing imagery, a large amount of effort has been devoted to detecting and localizing geospatial objects [29].

Most of the traditional object detection methods regard the object detection problem as a classification problem, which consists of feature extraction and feature classification stages. For the remote sensing imagery object detection methods, the spectral-based object detection methods treat the detection as a two-class classification task, namely, the object and the background. The spectral-based detection methods includes the spectral matched filter (SMF), the matched subspace detector (MSD), the adaptive coherence/cosine detectors (ACDs), the sparse-representation based detectors, etc. These methods mainly focus on the differences of the target and the background [10]. There are four kinds of object detection methods, namely, template matching based methods, knowledge-based methods, OBIA-based methods, and machine learning methods [15]. Template matching-based methods can be divided into two classes-rigid template matching and deformable template matching, which involve two main procedures, namely, template generation and similarity measurement. The knowledge-based object detection methods use prior knowledge, including geometric information and contextual information, which generally translates the object detection problem into a hypothesis testing problem. OBIA-based object detection methods involve two main steps—image segmentation and object classification—where the appropriate segmentation scale is the key factor influencing the object detection result. For the machine learning based methods, they typically include feature extraction, optional feature fusion, dimension reduction, and classifier training stages. The feature extraction stage relying on the proposals generated with selective search (SS) [30] usually involves extracting handcrafted features such as spectral features, texture features, and local image features (e.g., scale-invariant feature transform (SIFT) or histogram of oriented gradients (HOG) [16]). The feature classification stage mainly deals with training a classifier, such as support vector machine (SVM) [31], conditional random fields [10], sparse coding based classifiers [14,32], bag-of-words (BoW) classifiers [23,27,31], etc. The core idea of these methods is to train a classifier to discriminate the predicted labels, i.e., object or not. In summary, these methods are heavily reliant on the manually designed feature descriptors and human-labeled training samples, and perform well when there is a large amount of training data and the feature representation is efficient. In addition to the involvement of human ingenuity in the feature design for specific object detection tasks, these approaches separate the object detection tasks into non-related region proposal generation and object localization stages, which greatly increases the training load of the algorithm.

Recent developments in deep learning [29,33,34] have provided an automatic feature extraction and feature representation framework for various tasks, including classification and object detection [23,26,32]. Due to the recent development of large public natural image datasets such as ImageNet [30], and high-performance computing systems such as graphics processing units (GPUs), CNN-based algorithms have achieved great success in large-scale visual recognition tasks. The CNN

is an efficient hierarchical feature representation framework, in which the higher layers demonstrate semantic abstraction properties [18,35–37]. Recent advances in object detection with deep learning techniques have been driven by the success of the region proposal method, namely, region-based CNN (R-CNN) [38], which is an effective and efficient solution. Based on the powerful feature extraction ability of deep learning, replacing the inexpensive SS-based region proposal methods with deep and powerful CNN-based methods is an important development. R-CNN [38], fast region-based convolutional neural network(Fast R-CNN) [39], and Faster R-CNN [40] are typical deep learning based object detection algorithms, which are a series of object detection solutions by respectively solving the corresponding problems. R-CNN transfers the object detection problem from the traditional shallow SVM algorithm to the more expressive CNN classifier. Fast R-CNN is an improvement based on R-CNN, which improves the object detection procedure by outputting the bounding boxes and the corresponding labels at the same time from the CNN classifier. However, Fast R-CNN is still hindered by the time consumption of the proposal generation procedure and the detection procedure. In order to avoid utilizing the time-consuming SS strategy when generating the region proposals, the region proposal network (RPN) has been proposed in the Faster R-CNN object detection algorithm. Faster R-CNN further improves Fast R-CNN by sharing features between the region proposal generation procedure and the detection procedure, which can greatly reduce the time consumption for computing the proposals.

Although object detection algorithms have been developed in natural imagery object detection fields, high-efficiency multi-class geospatial object detection for HSR remote sensing imagery has not yet been achieved. Based on the feature-sharing and time-saving properties of the Faster R-CNN algorithm, developing a highly efficient and robust integrated multi-class geospatial object detection framework for HSR remote sensing imagery is significant and necessary. The RPN is a kind of fully convolutional network (FCN) used to generate region proposals, and is designed to efficiently predict proposals with a wide range of scales and aspect ratios using "anchor" boxes. Compared with the traditional region proposal generation methods, the RPN considers the multi-scale properties and the rotation properties during the region generation procedure, which increases the accuracy of the object location and helps improve the detection efficiency. Faster R-CNN integrates the region proposal generation procedure and the detection procedure by sharing features, which can greatly reduce the time consumption for computing the proposals. In order to realize the joint optimization between region proposal generation and detection, an alternating training algorithm [40] is utilized for collaborative optimization of the region proposal generation and detection procedures. However, the Faster R-CNN based object detection algorithm still faces difficult convergence problems when the number of annotated training samples is limited for HSR remote sensing imagery.

To tackle the problem of limited annotated samples for HSR remote sensing imagery objects, a novel object detection framework, namely, R-P-Faster R-CNN, is proposed here for multi-class geospatial object detection from HSR remote sensing imagery. R-P-Faster R-CNN adequately utilizes a pre-training mechanism to increase the robustness when the number of annotated samples is limited. It is noted that ImageNet is a large natural image dataset which contains various categories and large quantities of images. Training the deep network on ImageNet can help to obtain a good convergence value for the algorithm. Transferring the pre-trained network parameters from the large-scale ImageNet to quantity-limited HSR remote sensing imagery has been demonstrated to be highly efficient [41]. In order to effectively detect the multi-class geospatial objects from HSR remote sensing imagery, a pre-training mechanism based on transfer learning is introduced to the multi-class geospatial object detection for HSR remote sensing imagery. The main contributions of this paper are summarized as follows:

(a) An Effective Integrated Region Proposal Network (RPN) and Object Detection Strategy for HSR Remote Sensing Imagery. Considering the feature extraction advantages of the deep learning based methods, we propose a learning-based RPN which effectively integrates the region proposal generation procedure and the object detection procedure by sharing the convolutional features of

these two stages. To make the integrated object detection framework more efficient, the network adopts an alternating training strategy. The integrated strategy makes the proposed object detection framework an end-to-end object detection framework for HSR remote sensing imagery.

(b) A Robust and Efficacious Compensation Strategy for the Lack of Labeled Samples for HSR Remote Sensing Imagery Object Detection. There are currently very few multi-class geospatial object detection datasets available. However, there are a lot of similarities between the large natural image datasets and the quantity-limited HSR remote sensing imagery datasets. Pre-training the large-scale deep learning based object detection framework on a natural imagery dataset, and then transferring the pre-trained network parameters for the HSR remote sensing imagery, can provide good initial values and ensure the convergence for the HSR remote sensing imagery object detection.

(c) An Efficient Training Time Conservation Strategy for HSR Remote Sensing Imagery Object Detection. To improve the time efficiency of HSR remote sensing imagery object detection, a pre-training mechanism and a transfer mechanism are conducted on the HSR remote sensing imagery, which gradually provides more appropriate initial values for the HSR remote sensing imagery object detection. In addition, the integration of the region proposal generation procedure and detection procedure also saves a lot of training time for the HSR remote sensing imagery.

The proposed R-P-Faster R-CNN algorithm was evaluated and compared with the conventional HSR remote sensing imagery object detection methods, as well as the current non-end-to-end CNN-based object detection methods. For the experiments, we adopted the NWPU VHR-10 dataset, which is a 10-class HSR remote sensing imagery geospatial object detection dataset. The experimental results confirmed that the proposed method can achieve a satisfactory detection result with limited labeled training samples.

The rest of this paper is organized as follows. Section 2 presents the related object detection works. The proposed highly efficient and robust integrated multi-class geospatial object detection algorithm—R-P-Faster R-CNN—is described in detail in Section 3. Section 4 presents a description of the dataset and the experimental settings. Sections 5 and 6 present the analysis of the experimental results and a discussion of the results, respectively. Finally, the conclusions are drawn in Section 7.

2. Related Works

Geospatial object detection from remote sensing imagery has been extensively studied during the past years. A number of handcrafted feature based object detection methods and automatic feature learning based object detection methods have been studied with natural image datasets [37]. Object detection based on remote sensing imagery has also been studied [12,16,42]. The spectral-based object detection methods treat the detection as a two-class classification task, namely, the object and the background. The spectral-based detection methods include the SMF, the MSD, the ACDs, the sparse representation based detectors, etc. These methods mainly focus on the differences between the target and the background [10]. OBIA-based object detection involves classifying or mapping remote sensing imagery into meaningful objects (i.e., grouping relatively local homogeneous pixels). OBIA involves two steps: image segmentation and object classification. To obtain a satisfactory OBIA object detection result, the core task is to obtain a proper segmentation scale to represent the objects. For the OBIA-based object detection methods, the object features, such as spectral information, size, shape, texture, geometry, and contextual semantic features, can be extracted [15]. For example, Liu et al. [43] detected inshore ships in optical satellite images by using the shape and context information that was extracted in the segmented image. Liu et al. [44] presented robust automatic vehicle detection in QuickBird satellite images by applying morphological filters for separating the vehicle objects from the background. However, all these methods are performed in an unsupervised manner, and they are effective only for detecting the designed object category in simple scenarios.

With the development of remote sensing imagery techniques and machine learning techniques, researchers have addressed multi-class geospatial object detection from complex-background remote

sensing imagery. The conventional object detection methods for HSR imagery are stage-wise and depend on handcrafted features by experience. Most of these methods treat the object detection problem as a classification problem, where the classification is performed using the handcrafted features and a predefined classifier [12]. For example, Han et al. [45] proposed to detect multi-class geospatial objects based on visual saliency modeling and discriminative learning of sparse coding. Cheng et al. [16] used HOG features and latent SVM to train deformable part based mixture models for each object category. However, all these methods are based on the use of prior information to design the handcrafted features, which usually requires a large number of human-labeled training examples. For these handcrafted feature based object detection algorithms, the main problem is their non-automatic properties.

The advanced machine learning techniques have made geospatial object detection easier with the automatic feature learning framework. Deep learning is now recognized as a good choice for remote sensing imagery object detection. However, the limited annotated samples for object detection has motivated some researchers to develop weakly supervised learning frameworks. Han et al. [42] proposed a weakly supervised learning framework based on Bayesian principles and an unsupervised feature learning method via deep Boltzmann machines (DBMs) to build a high-level feature representation for various geospatial objects. Zhang et al. [12] undertook aircraft detection from large-scale remote sensing imagery with a coupled CNN model by sharing features between CRPNet and LOCNet to reduce the time consumption, perfectly combining the coupled CNN and the weakly supervised learning framework. In addition, the rotation-invariant properties of objects has also been studied with deep learning. Cheng et al. [46] improved the CNN-based object detection algorithm by adequately considering the rotation-invariant properties of the images. In summary, the traditional CNN-based HSR remote sensing imagery object detection framework usually consists of several common stages, i.e., convolutional layers, nonlinear layers, pooling layers, and the corresponding loss function. Although these geospatial object detection frameworks can perform well in multi-class or single-class remote sensing imagery object detection, a unified framework for multi-class HSR imagery geospatial object detection is still needed.

3. Overview of the Proposed R-P-Faster R-CNN Framework

The proposed R-P-Faster R-CNN framework consists of three main procedures, namely, the effective Faster R-CNN procedure, the robust and efficacious pre-training procedure to compensate for the deficiency of labeled training samples, and the effective time conservation procedure. The effective Faster R-CNN procedure consists of two stages, namely, the RPN generation stage and the Fast R-CNN detection and location stage. The RPN realizes three main functions, namely, outputting the locations and scores of the region proposals, transforming the different-scale and different-ratio proposals into low-dimensional feature vectors, and outputting the classification probability of a region proposal and the regression values of the locations. Fast R-CNN takes the convolutional features and the predicted bounding boxes as the input, which is a location refinement stage on the basis of the RPN. The robust and efficacious compensation for the deficiency of labeled samples for Faster R-CNN works by first transferring the pre-trained network parameters from the ImageNet dataset, and then from the PASCAL VOC dataset, which can not only alleviate the deficiency of the labeled samples, but can also provide good initialization values for the Faster R-CNN object detection framework. The effective time conservation procedure of the proposed R-P-Faster R-CNN framework refers to the specific network structure and the network training mechanism.

3.1. Effective Integrated Region Proposal Network and Object Detection Faster R-CNN Framework

Faster R-CNN includes two stages, namely, the RPN stage and the Fast R-CNN detection stage. Faster R-CNN integrates the RPN and Fast R-CNN by sharing the convolutional features, and optimizes the whole network with a multi-task loss function in an alternating training manner. This process is described as follows.

3.1.1. Overall Architecture

The overall architecture of Faster R-CNN is composed of two components, namely, the RPN and Fast R-CNN. The RPN is the core innovation of the Faster R-CNN based object detection framework. It is a kind of FCN that deals with the arbitrary-size input images and generates a set of rectangular object proposals. The outstanding characteristic of the RPN is the utilization of anchors. Anchors are the centers of the sliding windows, and they ensemble the different-ratio and multi-scale region proposals to import into the RPN. With the anchors, the RPN can realize multi-scale information incorporation. For every location of the image, there are nine possible region proposals, namely, areas of $\{128^2, 256^2, 512^2\}$ and length-to-width ratios of {1:1, 1:2, 2:1}. The framework of Faster R-CNN is shown in Figure 1.

In order to generate the convolutional feature maps in the last shared convolutional layer, the convolutional features are imported into two sibling fully connected layers, namely, the box-regression layer (reg) and the box-classification layer (cls). Suppose that the number of maximum possible proposals for each location is denoted as k, then there will be $4k$ outputs encoding the coordinates of k boxes for the regression layer and $2k$ scores for the classification layer, estimating the probability of object or not. The principles of the anchors and the RPN are shown in Figure 2.

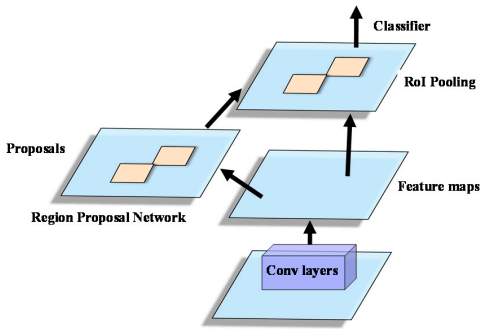


Figure 1. The framework of Faster R-CNN.

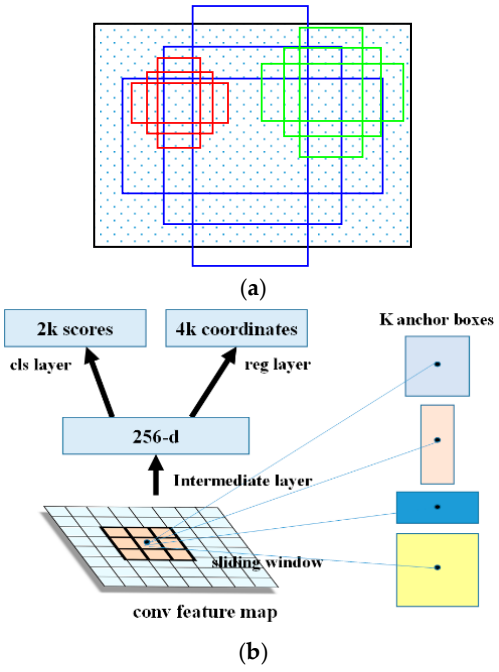


Figure 2. The principles of anchors and the RPN: (**a**) The principle of anchors; and (**b**) the principle of the RPN.

For the RPN, the judgement condition of whether the extracted region proposal is the required bounding box depends on the value of the intersection-over-union (IoU). When the IoU is larger than 0.7, it is considered as a foreground region proposal, and when the IoU is smaller than 0.3, it is considered as a background region proposal. During the region proposal generation procedure, the FCN-based RPN generates a large number of cross-boundary proposal boxes. To alleviate the redundancy phenomenon of the region proposals, non-maximum suppression (NMS) [45] is utilized to select the most useful region proposals. In addition, the RPN also has two other advantages. One advantage is the translation-invariant property of the anchors, which are constructed based on the assumption that if one translates an object in an image, the proposal should translate and the same function should be able to predict the proposal in either location. Specifically, the translation-invariant property also reduces the model size as the number of anchors is a fixed small value. The other advantage is that the multi-scale anchors act as regression references. Differing from the time-consuming image pyramids for processing multi-scale features, the RPN processes the multi-scale feature maps by sliding windows of multiple scales on the feature maps.

The loss function for an image is defined as:

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_t, t_i^*) \quad (1)$$

where i is the index of an anchor in a mini-batch, and p_i is the predefined probability of anchor i being an object. The ground-truth label is 1 if the anchor is positive, and is 0 if the anchor is negative. t_i is a vector representing the four parameterized coordinates of the predicted bounding box, and t_i^* is that of the ground-truth box associated with a positive anchor. The classification loss L_{cls} is the log loss over the two classes (object vs. not object). For the regression loss, we use $L_{reg}(t_i, t_i^*) = R(t_i - t_i^*)$, where R is the robust loss function. The classification loss and the regression loss are weighted by a balancing parameter λ. Usually, the cls term is normalized by the mini-batch size, and the reg term is normalized by the number of anchor locations. By default, the value of λ is set as 10, and thus both the cls and reg terms are roughly equally weighted.

Differing from the RPN stage, Fast R-CNN is a location refinement procedure. Fast R-CNN takes as input an entire image and a set of object proposals to score. The network first processes the whole image with several convolutional (conv) and max pooling layers to produce a conv feature map. Then, for each object proposal, a region of interest (RoI) pooling layer extracts a fixed-length feature vector from the feature map. Fast R-CNN adopts either the Zeiler and Fergus (ZF) model or the visual geometry group (VGG) model to realize the detection procedure. The ZF model and the VGG model are typical deep network based recognition models, which include several convolutional layers, pooling layers, and nonlinear layers. The structures of the ZF and VGG models are shown in Figure 3. Suppose that the size of the spatial window is $n \times n$, then, after the sliding window operation of the RPN, a lower-dimensional feature vector is obtained of 256-d for ZF and 512-d for VGG. After the sliding window operation, the features of the RPN are fed into two sibling fully connected layers: a box-regression layer and a box-classification layer. In the Faster R-CNN procedure, the value of n is equal to 3, and the effective receptive fields on the input image are 171 and 228 pixels for ZF and VGG, respectively.

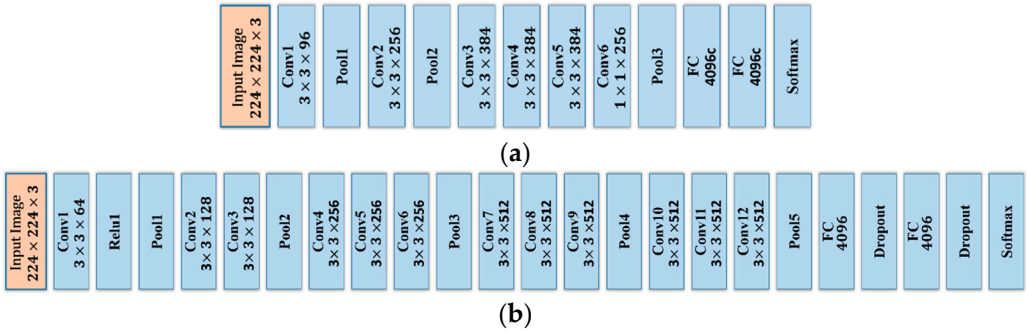


Figure 3. The structures of the ZF and VGG models: (**a**) the structure of the ZF model; and (**b**) the structure of the VGG model.

3.1.2. The Integration Strategy for the RPN and Fast R-CNN—Sharing Convolutional Features

Both the RPN stage and the Fast R-CNN stage can be trained separately, but each stage will consume a lot of time. In order to conserve the running time as much as possible, integrating the RPN stage and the Fast R-CNN stage for Faster R-CNN with a convolutional feature-sharing strategy rather than learning two separate networks can greatly reduce the running time consumption of the proposed algorithm. To realize the integration of the RPN and Fast R-CNN, alternating optimization is utilized to learn shared convolutional features between the region proposal generation stage and the object detection stage.

Faster R-CNN adopts a four-step alternating training procedure to realize the convolutional feature sharing via alternating optimization. In the first step, the network is initialized with an ImageNet-pre-trained model and fine-tuned end-to-end for the region proposal task. In the second step, a separate detection network trained by Fast R-CNN uses the proposals generated by the step-1 RPN. In this step, the network is also initialized by the ImageNet-pre-trained model. At this time, these two steps do not share convolutional layers. In the third step, the detector network is utilized to initialize the RPN training, but the shared convolutional layers are fixed and only the layers unique to the RPN are fine-tuned. At this time, the two networks share convolutional layers. At the last step, keeping the shared convolutional layers fixed, the unique layers of Fast R-CNN are fine-tuned. Through these four steps, the two networks share the same convolutional layers and an integration of the RPN and Fast R-CNN is achieved.

3.1.3. The Training Procedure of the Faster R-CNN Integrated Framework

During the training procedure of the Faster R-CNN integrated framework, the images are usually re-scaled, with the shorter side as $s = 600$ pixels. On the re-scaled images, the total stride for both the ZF and VGG nets on the last convolutional layer is 16 pixels. For the anchors, there are three scales with box areas of 128^2, 256^2, and 512^2 pixels, and three aspect ratios of 1:1, 1:2, and 2:1. During the training stage, the anchor boxes crossing the image boundaries are all ignored so that they do not contribute to the total loss. During the test stage, the fully convolutional RPN is applied to the entire image, which may generate cross-boundary proposal boxes, which are clipped to the image boundary. During the training stage, the number of proposals is also an important factor influencing the detection accuracy. If the proposals are highly overlapped with each other, the redundant computation is high. NMS on the proposal regions based on the cls scores is utilized to reduce the number of region proposals.

3.2. Robust and Efficacious Pre-Training Framework for Compensation of the Deficiency of Labeled Training Samples for HSR Remote Sensing Imagery Object Detection

The overall structure of the effective integrated region proposal network and object detection Faster R-CNN framework was introduced in the previous section. This structure is effective for natural image object detection especially when there is a large amount of labeled training samples.

However, for the HSR remote sensing imagery object detection task, the limited annotated samples is a significant factor influencing the detection performance. It is noted that the ImageNet dataset is a large and complicated natural image dataset, which contains 1000 categories with abundant information. Compared with this large and complicated natural image dataset, the current multi-class geospatial object detection datasets of HSR remote sensing imagery have the characteristics of small quantities, simple categories, complicated backgrounds, variable objects, etc. As the labeled samples of the multi-class geospatial object datasets of HSR remote sensing imagery are always deficient, to make up the defects for the HSR remote sensing imagery object detection, a compensation strategy—robust and efficacious pre-training mechanism—is needed to improve the performance of the multi-class geospatial object datasets of HSR remote sensing imagery.

Transfer learning is an effective technique in the deep learning research field for solving the problem of limited annotated samples in the target domain. Through learning the parameters of the deep network, a pre-training mechanism helps the deep network for object detection to quickly obtain the optimal values. To realize the multi-class geospatial object detection of HSR remote sensing imagery, a double pre-training mechanism is subsequently utilized on the ImageNet dataset and PASCAL VOC dataset, and then the pre-trained network parameters are transferred to the HSR remote sensing imagery. Similar to the four-step training stage of Faster R-CNN, the proposed R-P-Faster R-CNN realizes the optimization procedure with double pre-training. For the optimization procedure of the multi-class geospatial object detection stage, the multi-class geospatial object detection of HSR remote sensing imagery also adopts the four-step alternating training to obtain the detection results based on the PASCAL VOC dataset. Training and optimizing the network parameters on a large natural imagery dataset is a robust and efficacious approach for deep network based transfer learning.

Although the natural imagery dataset and the HSR remote sensing imagery dataset has some dissimilarities in the imaging mode and shooting angles, the categories of the HSR remote sensing imagery are similar or contained within the natural imagery dataset. Numerous experiments have verified that the effectiveness of the pre-training mechanism and transferring learning for the image recognition tasks. This similarity ensures the efficacious pre-training mechanism robust to the HSR remote sensing imagery object detection.

The specific procedure of the proposed R-P-Faster R-CNN is shown in Figure 4.

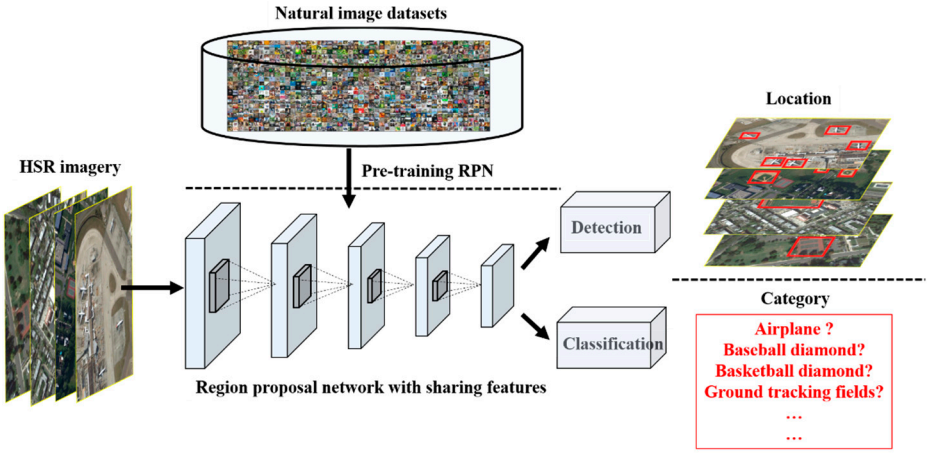


Figure 4. The flowchart of the proposed R-P-Faster R-CNN framework.

3.3. *Effective Training Time Conservation Framework*

Compared with the conventional stage-wise object detection algorithms, the training time conservation of the proposed R-P-Faster R-CNN framework can be illustrated from two aspects. The first aspect is the convolutional feature-sharing mechanism of Faster R-CNN, which reduces the time consumption of the proposed R-P-Faster R-CNN framework by sharing the convolutional

features between the RPN procedure and the detection procedure with a four-step optimization strategy. Fast R-CNN saves the time consumption at the detection stage by transferring the region proposal generation after the convolutional feature maps are generated. However, Faster R-CNN saves the time consumption at both the region proposal generation procedure and the detection procedure, and it saves the time consumption of the region proposal generation procedure by introducing the RPN. The time consumption of the convolutional feature-sharing strategy mainly reflects the test period.

The second aspect is the pre-training mechanism for multi-class geospatial object detection of HSR remote sensing imagery. It is noted that deep network needs a large amount of data to fit the complicated and nonlinear data distribution. Both gathering and constructing the large imagery dataset is tough for current HSR remote sensing imagery object detection task. However, the category and image similarities between HSR remote sensing imagery and natural imagery dataset provides the possibilities for cross-domain transferring learning. Transfer learning is an effective measure in the deep learning area especially where there are huge amount of data to train and complex network structures to model. Transferring the optimized network parameters from the natural imagery dataset to the HSR remote sensing imagery and pre-training the proposed object detection framework for HSR remote sensing imagery is easy for a network to quickly reach its optimal solutions, which guarantees the effectiveness of the proposed object detection framework in time conservation.

Based on the above two time conservation strategies of the proposed R-P-Faster R-CNN for HSR remote sensing imagery object detection, both the network structure and the pre-training strategy are optimal comparatively, which can provide effective time conservation measures for HSR remote sensing imagery object detection.

4. Dataset and Experimental Settings

In order to evaluate and validate the effectiveness of the proposed R-P-Faster R-CNN algorithm on HSR remote sensing imagery, the utilized dataset, the experimental settings, and the corresponding analysis of the experimental results are described in this section.

4.1. Description of the Dataset and Experimental Settings

The performance of the proposed algorithm was tested on a multi-class object detection dataset: the Northwestern Polytechnical University very high spatial resolution-10 (NWPU VHR-10) dataset [47]. The NWPU VHR-10 dataset is a multi-source and multi-resolution object detection dataset, which not only includes optical remote sensing images, but also includes pan-sharped color infrared images. There are 10 different types of objects contained within the NWPU VHR-10 dataset, i.e., airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle. The NWPU VHR-10 dataset contains 800 optical remote sensing images, where 715 images were acquired from Google Earth with a 0.5 m–2.0 m spatial resolution, and 85 pan-sharpened color infrared images were acquired from the Vaihingen data with a 0.08 m spatial resolution. The NWPU VHR-10 dataset contains two sub datasets, namely, a positive dataset of 650 images, with each image containing at least one target to be detected, and a negative dataset of 150 images, without any targets to be detected. For the positive image set, 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 have been manually annotated with bounding boxes, and were utilized as the ground truth. For the positive dataset, the split ratios of the training dataset, validation dataset, and test dataset were 20%, 20%, and 60%, respectively, comprising 130 images, 130 images, and 390 images. The object categories of the NWPU VHR-10 dataset are shown in Figure 5.

The performance of the proposed algorithm was compared with the handcrafted feature based methods of the BoW feature [14], spatial sparse coding BoW (SSCBoW) [48], Fisher discrimination dictionary learning (FDDL) [45], the collection of part detectors (COPD) [28] method, and the automatic deep learning feature based methods of transferred CNN [46], newly trained CNN [46], rotation-invariant CNN (RICNN) without fine-tuning [46], and RICNN with fine-tuning. The detection

accuracy, computational time, and the precision-recall curves (PRCs) are taken as the evaluation indexes. For the proposed algorithm—R-P-Faster R-CNN—three recognition networks were adopted, namely, ZF, VGG-16 fine-tuned on ImageNet, and VGG-16 fine-tuned on ImageNet and PASCAL VOC. For a fair comparison, the same training datasets and test datasets were adopted for the proposed algorithm and the other comparison methods. For the ZF model, the parameters were set as listed below. The initial learning rate of the first-stage RPN was set as 0.001, with an incremental ratio of gamma as 0.1 and the step size as 60,000; the initial learning rate of the first-stage Fast R-CNN was set as 0.001, with a step strategy of gamma as 0.1 and the step size as 30,000; the initial learning rate of the second-stage RPN was set as 0.001, with a step strategy of gamma as 0.1 and the step size as 60,000; the initial learning rate of the second-stage Fast R-CNN was set as 0.001, with a step strategy of gamma as 0.1 and the step size as 30,000. The momentum and weight decay were set to 0.9 and 0.0005, respectively. The total iteration number of the two-stage RPN was set as 80,000, and the total iteration number of the two-stage Fast R-CNN was set as 40,000.

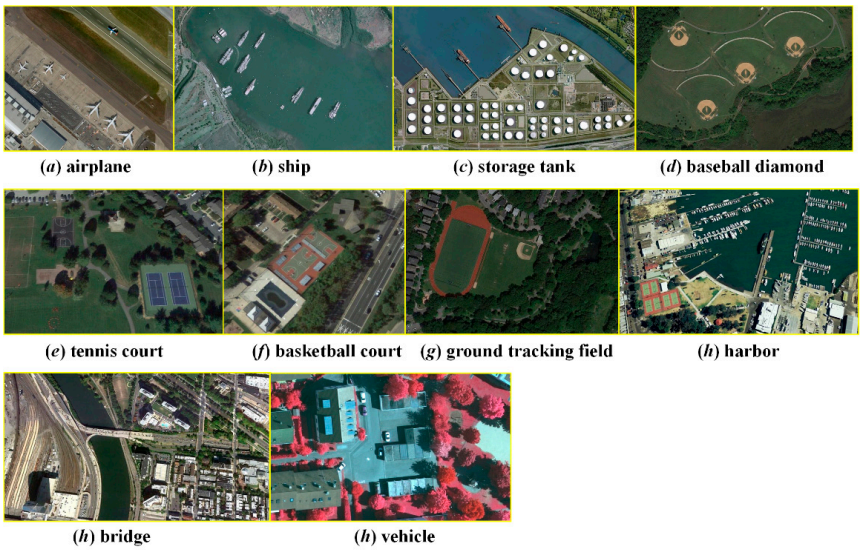


Figure 5. Object categories of the NWPU VHR-10 dataset.

4.2. Evaluation Indicators

In order to quantitatively evaluate the performance of the proposed algorithm, the widely utilized evaluation indicators of average precision (AP), Accuracy, Kappa, true negative rate (TNR), negative predictive value (NPV), and precision recall curves (PRCs) are adopted for the object detection framework [12,15,16,28,42,45,46]. The evaluation indicators are used in two stages, namely, the object category recognition stage and the location regression stage.

4.2.1. Average Precision

The AP computes the average value of the precision over the interval from recall = 0 and recall = 1, i.e., the area under the PRC; hence, the higher the AP, the better the performance. In addition, mean AP (mAP) computes the average value of all the AP values for all the classes.

4.2.2. Accuracy, Kappa, TNR and NPV

To evaluate the performance of the proposed detection algorithm, a number of indexes are needed, namely, true positive (TP), false positive (FP), false negative (FN), and true negative (TN). Accuracy, Kappa, TNR, and NPV are utilized for the assessment, and the definitions are given in Equations (2)–(4):

$$\text{Accuracy} = \frac{TP + TN}{(TP + FP + FN + TN)} \tag{2}$$

$$\text{TNR} = \frac{TN}{(TN + FP)} \tag{3}$$

$$\text{NPV} = \frac{TN}{(TN + FN)} \tag{4}$$

4.2.3. Precision–Recall Curve

The positive and negative samples recognized by the object detection framework are evaluated by the precision indicator, which measures the proportion of detections that are TP, and the recall indicator measures the proportion of positives that are correctly identified. The precision and recall indicators are formulated as follows:

$$\text{Precision} = \frac{TP}{(TP + FP)} \tag{5}$$

$$\text{Recall} = \frac{TP}{(TP + FN)} \tag{6}$$

A detection map is considered to be a TP if the area overlap ratio a_0 between the predicted bounding box and the ground-truth bounding box exceeds 0.5. Otherwise, the detection is considered a FP. In addition, if several detections overlap with the same ground-truth bounding box, only one is considered as a TP, and the others are considered as FN. In the SS algorithm, the number of object proposals is also a very important parameter, which directly influences the performance of the recall rate of each object category, and hence indirectly determines the final object detection performance.

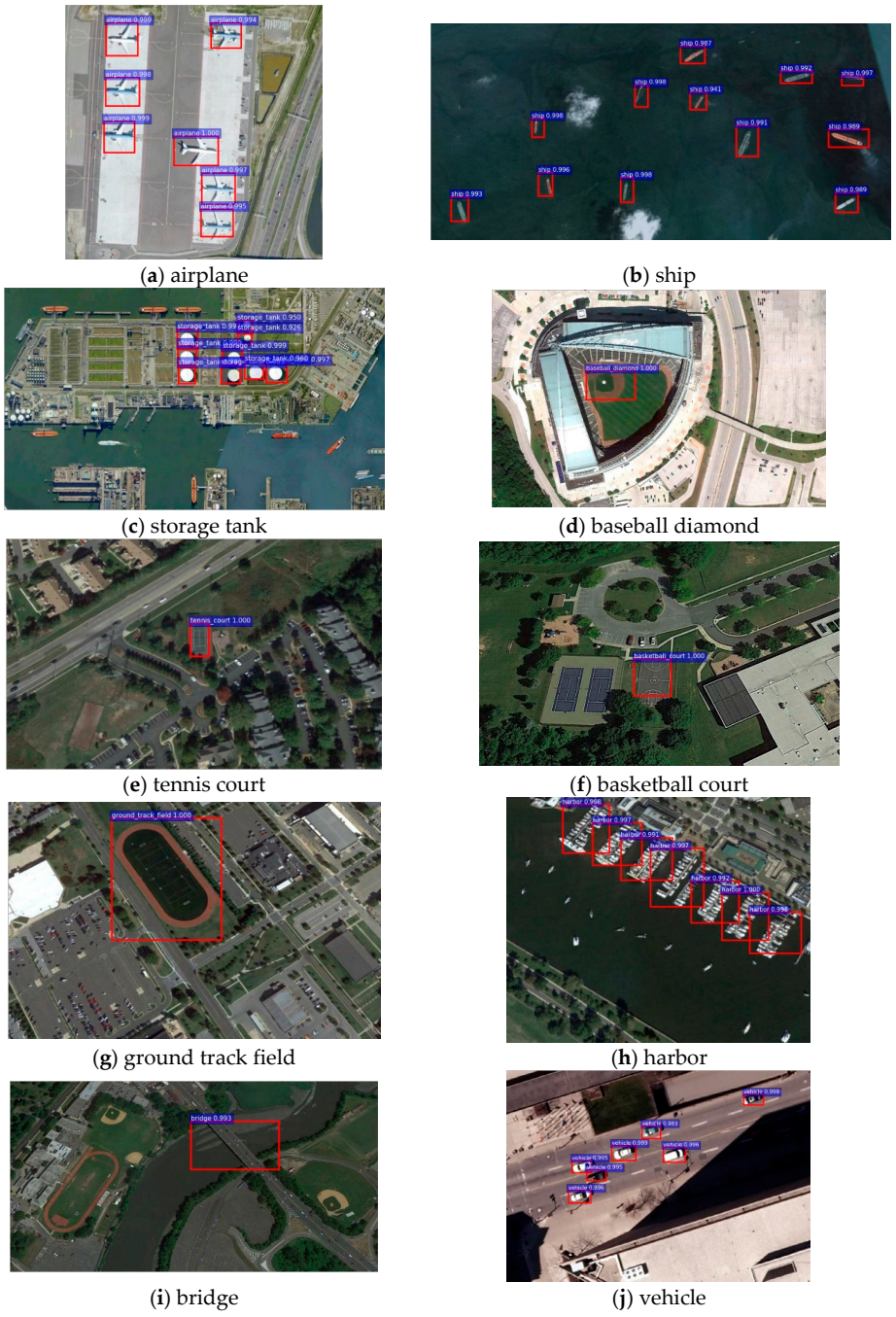
5. Results

Detection examples for the NWPU VHR-10 dataset with the proposed R-P-Faster R-CNN algorithm are shown in Figure 6.

Figure 6 shows the qualitative detection results of airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle for the proposed R-P-Faster R-CNN (single) (VGG16) algorithm, separately. In Figure 6, it can be seen that the proposed R-P-Faster R-CNN algorithm demonstrates better detection performance on the classes of airplane, baseball diamond, and ground track field. Figure 6 also shows that the proposed R-P-Faster R-CNN (single) (VGG16) shows a better detection performance on the small vehicle objects. However, the proposed R-P-Faster R-CNN demonstrates a less satisfactory location detection performance on the object class of storage tank.

Quantitative comparisons of the 10 different methods are shown in Tables 1–3, and Figures 7 and 8, as measured by AP values, Accuracy, Kappa, average running time per image, and PRCs, respectively. For the proposed R-P-Faster R-CNN algorithm, two pre-training approaches were adopted for the VGG16 architecture, namely, a single fine-tuning mechanism and a double fine-tuning mechanism. In Table 1, it can be seen that the proposed R-P-Faster R-CNN fine-tuned once on the ImageNet dataset obtains the best mean AP value of 76.5% among all the object detection methods. To make a concrete analysis, in Table 1, it can be seen that the proposed R-P-Faster R-CNN algorithm obtains better AP values for the classes of airplane, tennis court, basketball court, harbor, bridge, and vehicle. For the storage tank class, the RICNN with fine-tuning algorithm shows a much better detection performance than the other algorithms. In Table 2, it can be seen that the proposed R-P-Faster R-CNN algorithm also obtains the best Accuracy and Kappa values among comparison algorithms, which confirms the overall superior performance. After comparing the AP values of the different detection methods, the recall values of the proposed R-P-Faster R-CNN algorithm should also be compared. Figure 7 shows the recall values of the proposed R-P-Faster R-CNN algorithm with the ZF model and VGG-16 model.

From the overall view, it can be seen that the recall value of the proposed R-P-Faster R-CNN with VGG model is higher than with the ZF model. In Figure 7, it can be seen that the classes of airplane, baseball diamond, ground track field, and harbor obtain high recall values of greater than 90%, but the classes of storage tank, basketball court, and bridge present worse recall values. The curve at the top of the PRCs indicates a better performance. In Figure 8, it can be seen that most of the classes show a better detection performance, but the classes of airplane, baseball diamond, tennis court, ground track field, harbor, and vehicle demonstrate the best tendency. By jointly analyzing the AP values, the recall rate, and the PRCs, it can be seen that the proposed R-P-Faster R-CNN algorithm shows a superior detection performance for the classes of airplane, baseball diamond, ground track field, and harbor.



(**a**) airplane (**b**) ship

(**c**) storage tank (**d**) baseball diamond

(**e**) tennis court (**f**) basketball court

(**g**) ground track field (**h**) harbor

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

Figure 6. Detection examples for the NWPU VHR-10 dataset with the proposed R-P-Faster R-CNN algorithm.

Table 1. The AP values of the object detection methods.

	BoW	SSC BoW	FDDL	COPD	Transferred CNN	Newly Trained CNN	RICNN without Fine-Tuning	RICNN with Fine-Tuning	R-P-Faster R-CNN (ZF)	R-P-Faster R-CNN (Double) (VGG16)	R-P-Faster R-CNN (Single) (VGG16)
Airplane	0.025	0.506	0.292	0.623	0.661	0.701	0.860	0.884	0.803	**0.906**	0.904
Ship	0.585	0.508	0.376	0.689	0.569	0.637	0.760	**0.773**	0.681	0.762	0.750
Storage tank	0.632	0.334	0.770	0.637	0.843	0.843	0.850	**0.853**	0.359	0.403	0.444
Baseball diamond	0.090	0.435	0.258	0.833	0.816	0.836	0.873	0.881	0.906	**0.908**	0.899
Tennis court	0.047	0.003	0.028	0.321	0.350	0.355	0.396	0.408	0.715	**0.797**	**0.797**
Basketball court	0.032	0.150	0.036	0.363	0.459	0.468	0.579	0.585	0.677	0.774	**0.776**
Ground track field	0.078	0.101	0.201	0.853	0.800	0.812	0.855	0.867	**0.892**	0.880	0.877
Harbor	0.530	0.583	0.254	0.553	0.620	0.623	0.665	0.686	0.769	0.762	**0.791**
Bridge	0.122	0.125	0.215	0.148	0.423	0.454	0.585	0.615	0.572	0.575	**0.682**
Vehicle	0.091	0.336	0.045	0.440	0.429	0.448	0.680	0.711	0.646	0.666	**0.732**
Mean AP	0.246	0.308	0.245	0.546	0.597	0.618	0.710	0.726	0.702	0.743	**0.765**

Table 2. The Accuracy values of the object detection methods.

	BoW	SSCBoW	COPD	Transferred CNN	RICNN with Fine-Tuning	R-P-Faster R-CNN (ZF)	R-P-Faster R-CNN (VGG16)
Accuracy	0.524	0.696	0.763	0.780	0.784	**0.789**	**0.789**

Table 3. Computation time comparisons for the 10 different methods.

Average Running Time Per Image (Seconds)	BoW	SSC BoW	FDDL	COPD	Transferred CNN	Newly Trained CNN	RICNN without Fine-Tuning	RICNN with Fine-Tuning	R-P-Faster R-CNN (ZF)	R-P-Faster R-CNN (VGG16)
	5.32	40.32	7.17	1.07	5.24	8.77	8.77	8.77	**0.04**	**0.15**

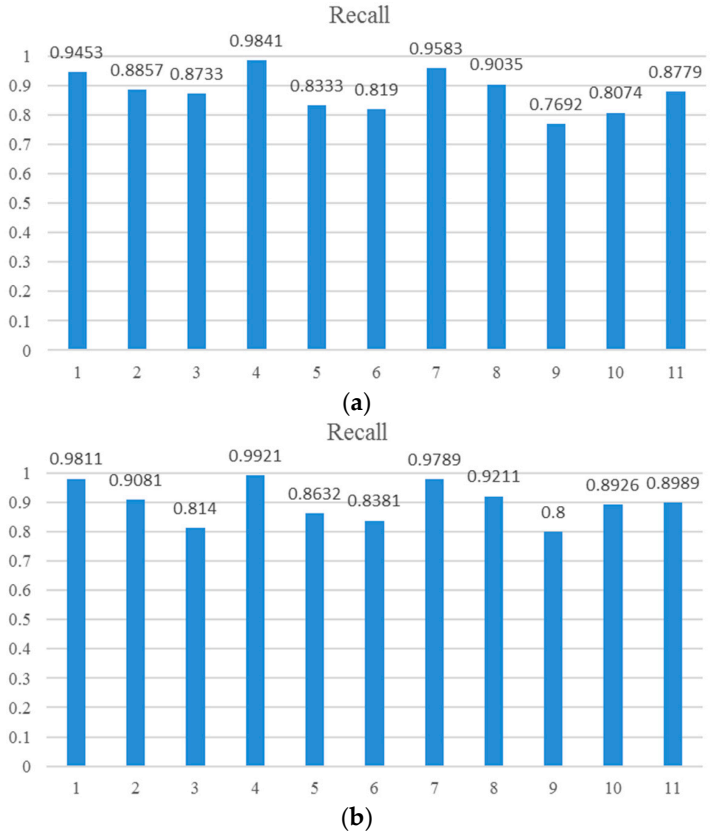
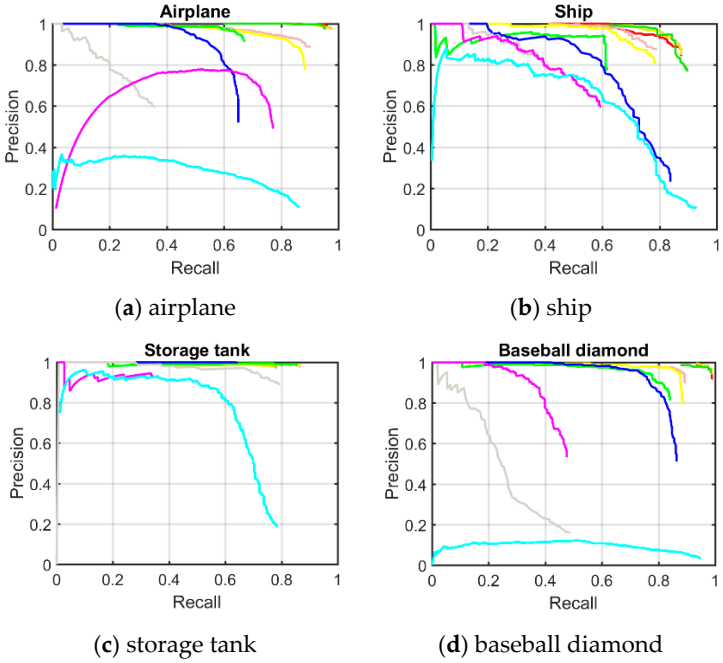


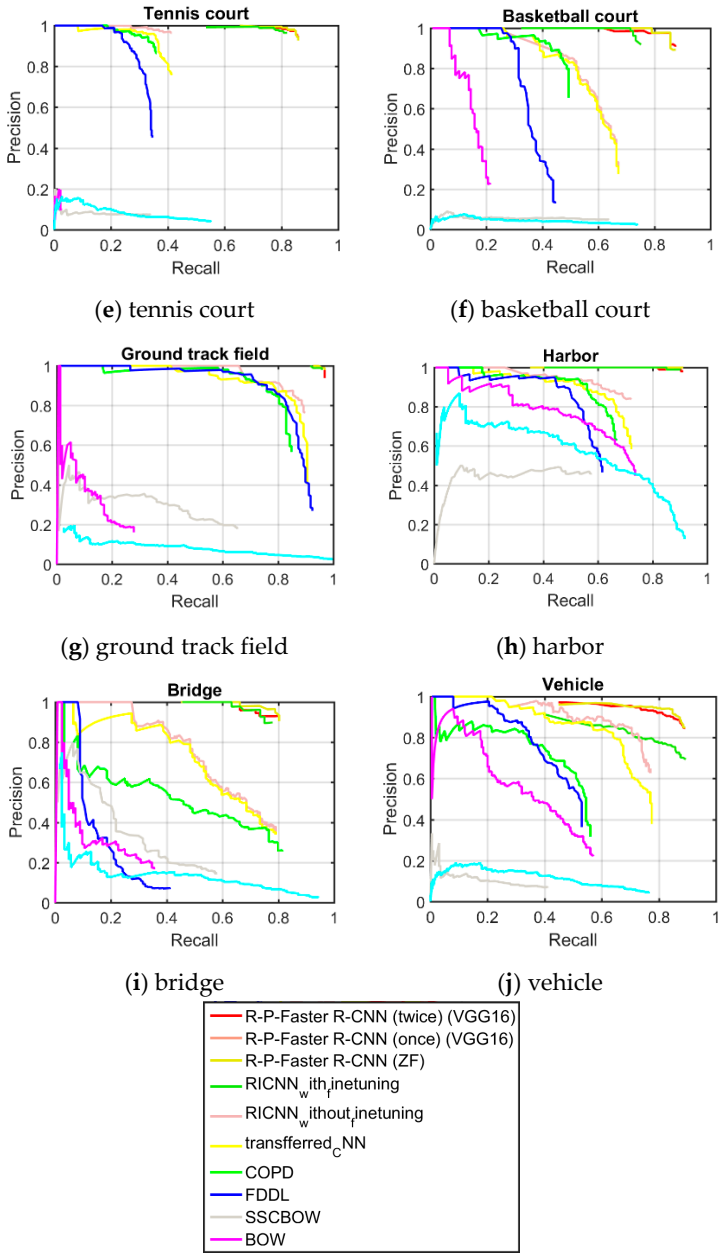
Figure 7. Quantitative evaluation results measured by recall rate for all 10 object categories (1, airplane; 2, ship; 3, storage tank; 4, baseball diamond; 5, tennis court; 6, basketball court; 7, ground trackfield; 8, harbor; 9, bridge; 10, vehicle; and 11, average). The numbers on the bars denote the recall rates for each object category. (**a**) The recall values obtained with the ZF model for fine-tuning. (**b**) The recall value obtained with the VGG16 model for fine-tuning.



(**a**) airplane (**b**) ship

(**c**) storage tank (**d**) baseball diamond

Figure 8. *Cont.*



(**e**) tennis court (**f**) basketball court

(**g**) ground track field (**h**) harbor

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

Figure 8. The PRCs for the proposed R-P-Faster R-CNN algorithm, as well as the comparison methods.

In addition to the evaluation indexes, the computational efficiency is also an important factor for evaluating the performance of the proposed R-P-Faster R-CNN algorithm. In Table 3, it can be seen that the proposed R-P-Faster R-CNN (VGG16) algorithm with the best detection performance takes about 0.15 s, which confirms that it is an efficient detection method.

6. Discussion

The number of bounding boxes plays a significant role in the region proposal generation stage of the proposed R-P-Faster R-CNN algorithm in the HSR remote sensing imagery object detection task. Hence, a sensitivity analysis between the recall rate/AP value and the number of bounding boxes for the ZF model and VGG16 model of R-P-Faster R-CNN is provided in the following sections.

6.1. Sensitivity Analysis of the Bounding Box Number for the ZF Model of the Proposed R-P-Faster R-CNN Algorithm

In order to demonstrate the performance of the proposed R-P-Faster R-CNN algorithm with the ZF model, the tradeoff between the recall rate and the number of object proposals and the tradeoff between the AP value and the number of object proposals are shown in Figures 9 and 10, respectively. The detection performance of the proposed R-P-Faster R-CNN is measured by the recall rate and AP value, which implies that the larger the values, the better the detection performance. However, the recall rate and AP value cannot both be large, as expected in a real detection situation. Figures 9 and 10 demonstrate oscillating curve trends, which are mainly related to the real object size for each object category when detecting with the proposed R-P-Faster R-CNN algorithm. For example, the best recall rates of the proposed R-P-Faster R-CNN (ZF) algorithm, for most of the classes, are obtained when the number of region proposals is 2000, but a satisfactory AP value for the proposed R-P-Faster R-CNN (ZF) is obtained when the number of region proposals is 1000 for ground track field, 1400 for storage tank and vehicle, and 1800 for airplane, baseball diamond, and harbor. Figures 11 and 12 show the relationship between TNR and NPV values with the number of region proposals. In these two figures, it can be seen that the TNR and NPV values are not sensitive to the number of region proposals with the ZF model.

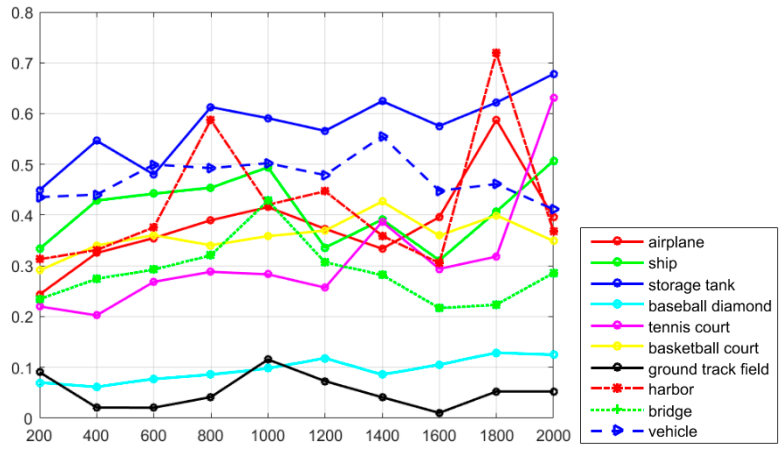


Figure 9. Tradeoff between the recall rate and the number of object proposals with the proposed R-P-Faster R-CNN (ZF) algorithm.

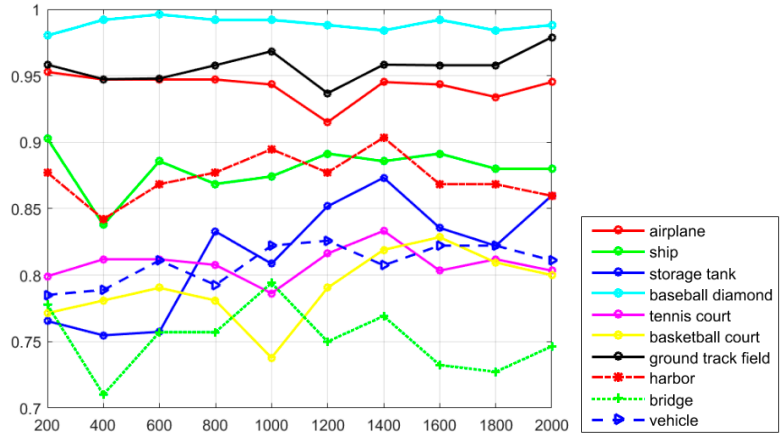


Figure 10. Tradeoff between the AP value and the number of object proposals with the proposed R-P-Faster R-CNN (ZF) algorithm.

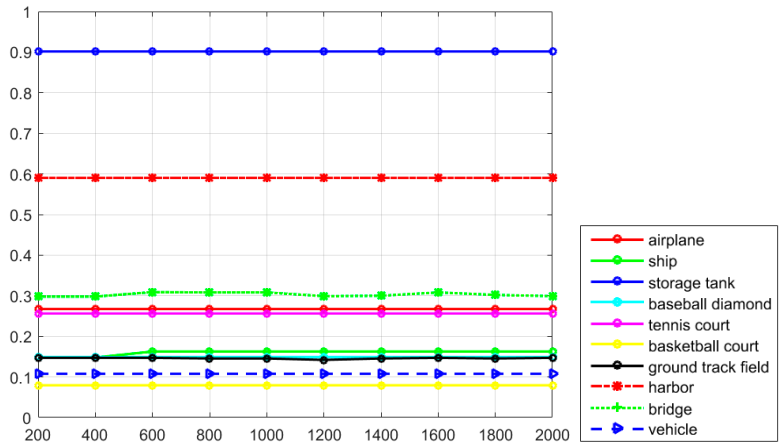


Figure 11. Tradeoff between the TNR and the number of object proposals with the proposed R-P-Faster R-CNN (ZF) algorithm.

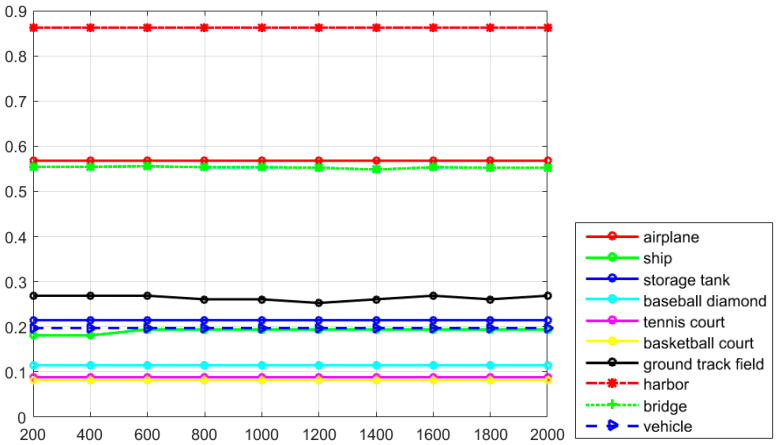


Figure 12. Tradeoff between the NPV and the number of object proposals with the proposed R-P-Faster R-CNN (ZF) algorithm.

6.2. Sensitivity Analysis of the Bounding Box Number for the VGG16 Model of the Proposed R-P-Faster R-CNN Algorithm

In the region proposal generation stage, the number of object proposals is a very important parameter, which directly affects the recall rate of each object category, as well as the final detection performance. Figures 13 and 14 explore the tradeoff between the recall rate and the average number of object proposals per image, for all 10 object categories, with the proposed R-P-Faster R-CNN (VGG16) algorithm on the test dataset. In Figures 13 and 14, it can be seen that the recall rate and the AP values both demonstrate relative oscillating trends, which are mainly due to the different sensitivities of the different categories with the proposed R-P-Faster R-CNN algorithm on the HSR remote sensing imagery. For example, the best recall rates of the proposed R-P-Faster R-CNN (VGG16) algorithm for most of the classes are obtained when the number of region proposals is 1800, but a satisfactory AP value for the proposed R-P-Faster R-CNN (VGG16) is obtained when the number of region proposals is 600 for airplane and ship, 800 for storage tank, harbor, baseball diamond, and ground track field, and 1200 for vehicle. Figures 15 and 16 show the influence of the number of region proposals on the TNR and NPV values with the VGG16 model, both of which show a stable trend. In Figures 15 and 16, it can be seen that the TNR and NPV values are not sensitive to the number of region proposals.

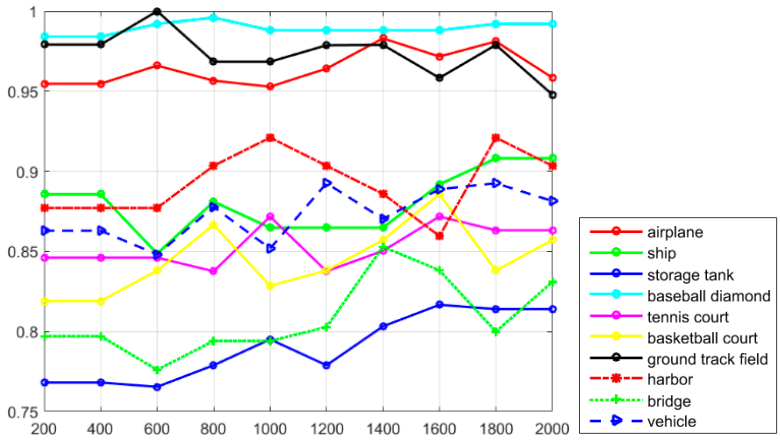


Figure 13. Tradeoff between the recall rate and the number of object proposals with the proposed R-P-Faster R-CNN (single) (VGG16) algorithm.

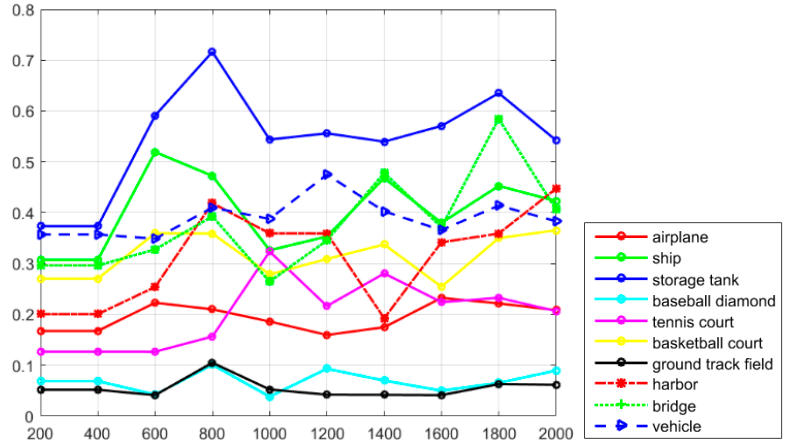


Figure 14. Tradeoff between the AP value and the number of object proposals with the proposed R-P-Faster R-CNN (single) (VGG16) algorithm.

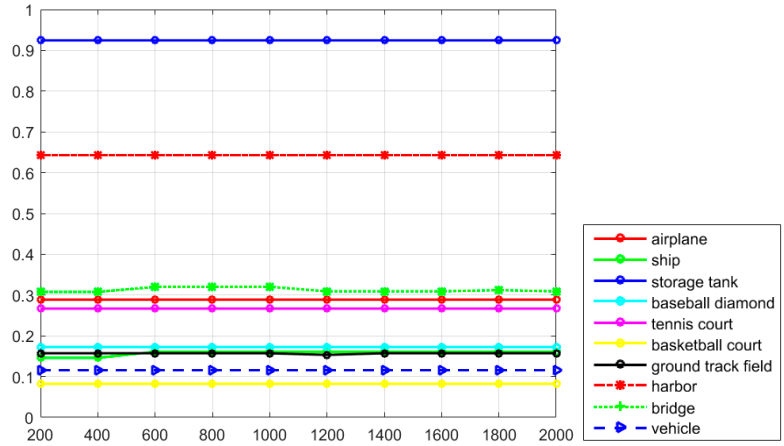


Figure 15. Tradeoff between the TNR and the number of object proposals with the proposed R-P-Faster R-CNN (VGG16) algorithm.

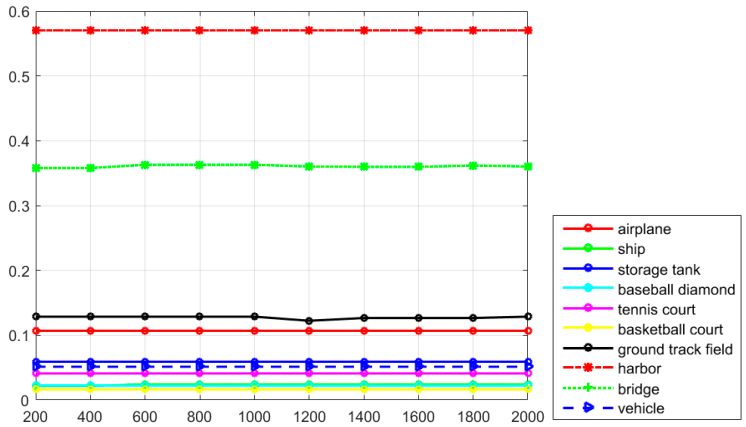


Figure 16. Tradeoff between the NPV and the number of object proposals with the proposed R-P-Faster R-CNN (VGG16) algorithm.

7. Conclusions

In this paper, an effective R-P-Faster R-CNN object detection framework has been proposed for HSR remote sensing imagery. Considering the complex distribution of geospatial objects and the low efficiency of the current object detection methods for HSR remote sensing imagery, the robust properties of a transfer mechanism and the sharable properties of Faster R-CNN are considered and combined in the R-P-Faster R-CNN object detection framework. The transfer mechanism provides the deep learning based object detection algorithm with good initial network parameters. The sharable properties of the R-P-Faster R-CNN object detection framework help save the time consumption of re-training the neural network. The combination of these properties results in the proposed R-P-Faster R-CNN algorithm performing better than the current object detection methods.

In contrast to the CNN-based HSR imagery object detection methods, the region proposal generation stage and the object recognition stage are separated, which improves the time consumption for efficiently training the object detection framework. The feature-sharing mechanism of the proposed R-P-Faster R-CNN framework can effectively solve this problem and improve the performance of HSR imagery geospatial object detection. The experimental results obtained with the NWPU VHR-10 geospatial object detection dataset confirm that the proposed R-P-Faster R-CNN framework is efficient and effective. In our future work, a more effective object recognition framework will be considered for HSR imagery geospatial object detection.

Acknowledgments: This work was supported by National Key Research and Development Program of China under Grant No. 2017YFB0504202, National Natural Science Foundation of China under Grant Nos. 41622107 and 41371344, and Natural Science Foundation of Hubei Province under Grant No. 2016CFA029.

Author Contributions: All the authors made significant contributions to the work. Xiaobing Han and Yanfei Zhong designed the research and analyzed the results. Liangpei Zhang provided advice for the preparation and revision of the paper.

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

References

1. Blaschke, T. Object based image analysis for remote sensing. *ISPRS J. Photogramm. Remote Sens.* **2010**, *65*, 2–16. [CrossRef]
2. Blaschke, T.; Hay, G.J.; Kelly, M. Geographic object-based image analysis-towards a new paradigm. *ISPRS J. Photogramm. Remote Sens.* **2014**, *87*, 180–191. [CrossRef] [PubMed]
3. Bontemps, S.; Bogaert, P.; Titeux, N.; Defourny, P. An object-based change detection method accounting for temporal dependencies in time series with medium to coarse spatial resolution. *Remote Sens. Environ.* **2008**, *112*, 3181–3191. [CrossRef]

4. Aksoy, S.; Yalniz, I.Z.; Tasdemir, K. Automatic detection and segmentation of orchards using very high resolution imagery. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 3117–3131. [CrossRef]
5. Yao, X.; Han, J.; Cheng, G.; Qian, X.; Guo, L. Semantic Annotation of High-Resolution Satellite Images via Weakly Supervised Learning. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 3660–3671. [CrossRef]
6. Liu, L.; Shi, Z. Airplane detection based on rotation invariant and sparse coding in remote sensing images. *Optik* **2014**, *125*, 5327–5333. [CrossRef]
7. Bai, X.; Zhang, H.; Zhou, J. VHR object detection based on structural feature extraction and query expansion. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 6508–6520.
8. Yu, Y.; Guan, H.; Ji, Z. Rotation-invariant object detection in high resolution satellite imagery using super pixel-based deep Hough forests. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 2183–2187. [CrossRef]
9. 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.* **2015**, *12*, 701–705. [CrossRef]
10. 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]
11. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE International Conference, Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; pp. 886–893.
12. Zhang, F.; Du, B.; Zhang, L.; Xu, M. Weakly Supervised Learning Based on Coupled Convolutional Neural Networks for Aircraft Detection. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 5553–5563. [CrossRef]
13. 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]
14. 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.
15. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
16. 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]
17. Lei, Z.; Fang, T.; Huo, H.; Li, D. Rotation-invariant object detection of remotely sensed images based on Texton forest and Hough voting. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 1206–1217. [CrossRef]
18. LeCun, Y. Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information*; Morgan Kaufmann: San Francisco, CA, USA, 1990.
19. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. *arXiv* **2016**, preprint arXiv:1612.03144.
20. Grabner, H.; Nguyen, T.T.; Gruber, B.; Bischof, H. On-line boosting based car detection from aerial images. *ISPRS J. Photogramm. Remote Sens.* **2008**, *63*, 382–396. [CrossRef]
21. Eikvil, L.; Aurdal, L.; Koren, H. Classification-based vehicle detection in high-resolution satellite images. *ISPRS J. Photogramm. Remote Sens.* **2009**, *64*, 65–72. [CrossRef]
22. Lazebnik, S.; Schmid, C.; Ponce, J. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In Proceedings of the IEEE Computer Society Conference, Computer Vision and Patter Recognition, New York, NY, USA, 17–22 June 2006; pp. 2169–2178.
23. Cheriyadat, A. Unsupervised Feature Learning for Aerial Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 439–451. [CrossRef]
24. Zhao, B.; Zhong, Y.; Xia, G.S.; Zhang, L. Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 2108–2123. [CrossRef]
25. Han, X.; Zhong, Y.; Zhang, L. Scene classification based on a hierarchical convolutional sparse auto-encoder for high spatial resolution imagery. *Int. J. Remote Sens.* **2017**, *38*, 514–536. [CrossRef]
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Learn.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
27. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th Sigspatial International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
28. 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]

29. Bai, X.; Zhang, H.; Zhou, J. VHR object detection based on structural feature extraction and query expansion. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 1–13.
30. 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.
31. Li, F.F.; Perona, P. A Bayesian hierarchical model for learning natural scene categories. In Proceedings of the IEEE International Conference, Computer Vision and Pattern Recognition, Washington, DC, USA, 20–26 June 2005; pp. 524–531.
32. Hu, F.; Xia, G.X. Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification. *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.* **2015**, *8*, 2015–2030. [CrossRef]
33. LeCun, Y.; Bengio, Y.; Hinton, G.E. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
34. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Nevada, NV, USA, 3–6 December 2012; pp. 1106–1114.
35. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
36. Uijlings, J.R.; van de Sande, K.E.; Gevers, T.; Smeulders, A.W. Selective search for object recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
37. Simard, P.Y.; Steinkraus, D.; Platt, J.C. Best practices for convolutional neural networks applied to visual document analysis. In IEEE International Conference on Document Analysis and Recognition, Edinburgh, UK, 3–6 August 2003; pp. 958–962.
38. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 142–158. [CrossRef] [PubMed]
39. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
40. 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]
41. Hu, F.; Xia, G.; Hu, J.; Zhang, L. Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery. *Remote Sens.* **2015**, *7*, 14680–14707. [CrossRef]
42. Han, J.; Zhang, D.; Cheng, G. 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]
43. Liu, G. A new method on inshore ship detection in high-resolution satellite images using shape and context information. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 617–621. [CrossRef]
44. Liu, W.; Yamazaki, F.; Vu, T.T. Automated vehicle extraction and speed determination from QuickBird satellite images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2011**, *4*, 75–82. [CrossRef]
45. 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]
46. 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]
47. NWPU VHR-10 Dataset. Available online: http://www.escience.cn/people/gongcheng/NWPU-VHR-10.html (accessed on 26 June 2017).
48. 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]