remote sensing

MDPI

Article

Structured Object-Level Relational Reasoning CNN-Based Target Detection Algorithm in a Remote Sensing Image

Bei Cheng [1], Zhengzhou Li [1,2,*], Bitong Xu [1], Xu Yao [1], Zhiquan Ding [3] and Tianqi Qin [3]

[1] College of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China; chengbei@cqu.edu.cn (B.C.); 202012021023t@cqu.edu.cn (B.X.); 202012131079@cqu.edu.cn (X.Y.)
[2] Key Laboratory of Beam Control, Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu 610209, China
[3] Sichuan Institute of Aerospace Electronic Equipment, Chengdu 610100, China; 13350314996@163.com (Z.D.); tianqiqin2008@163.com (T.Q.)
* Correspondence: lizhengzhou@cqu.edu.cn; Tel.: +86-132-0601-5717

Abstract: Deep learning technology has been extensively explored by existing methods to improve the performance of target detection in remote sensing images, due to its powerful feature extraction and representation abilities. However, these methods usually focus on the interior features of the target, but ignore the exterior semantic information around the target, especially the object-level relationship. Consequently, these methods fail to detect and recognize targets in the complex background where multiple objects crowd together. To handle this problem, a diversified context information fusion framework based on convolutional neural network (DCIFF-CNN) is proposed in this paper, which employs the structured object-level relationship to improve the target detection and recognition in complex backgrounds. The DCIFF-CNN is composed of two successive sub-networks, i.e., a multi-scale local context region proposal network (MLC-RPN) and an object-level relationship context target detection network (ORC-TDN). The MLC-RPN relies on the fine-grained details of objects to generate candidate regions in the remote sensing image. Then, the ORC-TDN utilizes the spatial context information of objects to detect and recognize targets by integrating an attentional message integrated module (AMIM) and an object relational structured graph (ORSG). The AMIM is integrated into the feed-forward CNN to highlight the useful object-level context information, while the ORSG builds the relations between a set of objects by processing their appearance features and geometric features. Finally, the target detection method based on DCIFF-CNN effectively represents the interior and exterior information of the target by exploiting both the multiscale local context information and the object-level relationships. Extensive experiments are conducted, and experimental results demonstrate that the proposed DCIFF-CNN method improves the target detection and recognition accuracy in complex backgrounds, showing superiority to other state-of-the-art methods.

Keywords: target detection; remote sensing image; local context; object-level relationship; attention mechanism

Citation: Cheng, B.; Li, Z.; Xu, B.; Yao, X.; Ding, Z.; Qin, T. Structured Object-Level Relational Reasoning CNN-Based Target Detection Algorithm in a Remote Sensing Image. *Remote Sens.* **2021**, *13*, 281. https://doi.org/10.3390/rs13020281

Received: 17 December 2020
Accepted: 11 January 2021
Published: 14 January 2021

1. Introduction

The target detection and recognition in remote sensing images facilitates a wide range of applications such as airplane detection [1–3], road detection [4], building detection [5], land planning [6], and urban monitoring [7]. However, the remote sensing image contains diverse scenes, including man-made targets with drastic boundaries and a large number of landscape objects with similar characteristics to the background. Meanwhile, the target in the remote sensing image is usually small in size, which is easy to change with other objects in different environments. In addition, the appearance and size of the target may vary according to the viewpoint, lighting, and weather. Therefore, it is challenging to detect and recognize targets in remote sensing images due to various scenes with different objects and diverse targets with different features.

Efficient and accurate target detection in remote sensing images has attracted much attention, and a variety of traditional target detection algorithms have been developed. Generally, these algorithms can be divided into four categories, namely, the template-based method, the target-based image analysis method, the knowledge-based method, and the machine learning-based method [5,8–10]. The template-based method could effectively detect targets in a single and simple background, but its performance will be greatly decreased when the object varies in size, density distribution, and direction. The target-based method [11,12] firstly segments the image into relatively uniform pixel groups, and the pixel groups are then divided into different categories according to the multi-feature association criteria. The performance of this method is determined by segmentation algorithm and image complexity. Furthermore, the contour quality of the target directly affects the performance of subsequent image classification to a large extent. With respect to the knowledge-based method [13–15], it translates the implicit knowledge into the explicit detection rules. Then, it determines whether the target satisfies these rules. Usually, the prior knowledge and detection rules are defined specifically, and the accuracy of this method is greatly limited by the complex and changing scenes and changing target. The machine learning-based method usually extracts multiple features, such as histogram of oriented gradient (HOG) [16–18], bag-of-words feature [19–21], scale-invariant feature transform (SIFT) [22], and texture features [23–25]. Then, a classifier is learned for target detection, such as support vector machines (SVM) [20,26–28], k-nearest neighbor (kNN) [16,29,30], AdaBoost [25,31–33], and so on [34–38]. This method has advantages in scalability and compatibility, and it can establish the target detector automatically via machine learning techniques. However, the selection of manual features and training data has an obvious influence on the detection result of this method. Overall, these traditional methods rely heavily on the manual features, leading to poor performance of target detection and recognition, as well as insufficient adaptability to different situations and various targets.

Recently, the deep learning technique can extract explicit and implicit features through multi-scale and multi-level convolution layers, and it can approach arbitrary data with fully connected non-linear networks. The deep-learning-based method achieves end-to-end processing, and it shows great potential in remote sensing target detection and recognition. The general deep-learning-based target detection method such as region-CNN (RCNN) [39], Fast RCNN [40] and Faster RCNN [41], you only look once (YOLO) [42], single shot multiBox detector (SSD) [43], and region-based fully convolutional networks (R-FCN) [44] have been widely used in many applications. Wang et al. [45] propose a multiscale attention network to highlight the useful features and suppress the cluttered background. Cheng et al. [46] present a rotation-invariant CNN (RICNN) model to detect the targets with various degrees of rotation. Wang et al. [47] put forward a feature-merged single-shot detection (FMSSD) network, which is trained to fuse the contextual information in multiple scales. This method is robust for targets in a small size. Zheng et al. [48] propose a hyper-scale object detection network (HyNet), and the network is trained to solve the scale-variation problem for the geospatial target detection. Han et al. [49] adopt a Bayesian framework combined with weakly supervised learning for geospatial target detection.

However, the methods mentioned above do not consider the fusion and utilization of contextual information. Researches on visual perception systems show that the objects and specific environment are interdependent, providing rich context associations. The contextual information can indirectly model the relationship between the target and environment, indicating a direction to feature extraction and feature fusion. In the field of visual cognition, it is generally acknowledged that the target in consistent backgrounds or a familiar scene context can be detected more accurately and faster than that in inconsistent scenes. Furthermore, evidence is shown by a large body of empirical researches [50–54] that proper context modeling can improve the efficiency of target search and recognition. Sean Bell et al. [53] incorporate external contextual information by adding a recurrent neural network to the CNN. Nevertheless, this work only focuses on the internal characteristics of these region proposals, and ignores the explicit effect of the object on the target. In [55], the

authors achieve a top-down contextual priming by augmenting a semantic segmentation network in Faster RCNN. However, the semantic segmentation is defined as a separate network that cannot model the relationship between object and target. Mottaghi [56] presents a deformable part-based model by combining the region-level local context the with the scene-level global context. This model is robust to the target detection and recognition in natural image scenes with small fields of view. However, the remote sensing image usually has a large field of view with complex scenes, making it difficult to describe the context information of target detection in a single term. For example, a satellite remote sensing image often contains airports, ports, sea, and even cities, as shown in Figure 1. Zeng et al. [57] proposes a gated bi-directional CNN, which is trained to fuse local contextual information from the candidate areas. However, the explicit relationship between object and target is ignored by this network. In [52], a structural inference network detector is proposed to combine the scene context information and the target relationship within a single image. Nevertheless, the method does not exploit interior features about the target itself. Furthermore, the information fusion method is rather rough, and the meaningless background noise of the whole scene may be incorporated into the contextual information. Li et al. [58] developed a context convolution neural network model based on attention mechanism, which utilizes both global and local contextual information. However, the definition of the global context is ambiguous. At the same time, the contextual information around the target does not rise to the object level, leading to the lack of top-to-down supervised learning for target detection and recognition. Although these studies deal with contextual information under the deep learning framework, they have made little progress in exploring the integration of contextual information, especially in the establishment of explicit relationships between the object and target. Additionally, these methods usually choose a rough and simple pooling method for fusing context information, taking the useless context information as useful information. Overall, it is still challenging to make effective use of contextual information and integrate the information into current deep learning frameworks.

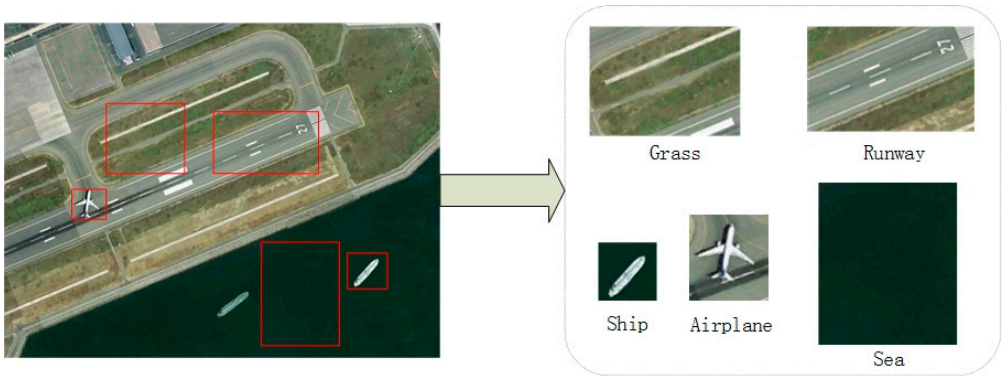

Figure 1. The contextual information in a remote sensing image.

Motivated by the above observations, a novel diversified context information fusion framework for convolutional neural network (DCIFF-CNN) is presented in this paper. The DCIFF-CNN utilizes both local context around the region proposal and the object-level relationship outside the contextual information. Meanwhile, the gated recurrent unit (GRU) is applied to generate a structured object relation graph. The GRU can iteratively highlight some context slices that are conducive to the detection task, and provide powerful feature representation for target detection. Similar to Faster-RCNN, DCIFF-CNN can be divided into two sub-networks, namely, a multi-scale local context region proposal network (MLC-RPN) and an object-level relationship contextualized target detection network (ORC-TDN). The MLC-RPN mergers the convolution-layer features of various scales to capture more fine-grained details. The object relational structured graph (ORSG) and the attentional message integrated module (AMIM) are established in the ORC-TDN to infer contextual

object instances and obtain more meaningful contextual information. First, the MLC-RPN extracts multi-scale candidate regions to get the object of interest in the image. Then, those regions are delivered to the ORC-TDN to infer contextual object relationships and detect targets. In the ORC-TDN, the ORSG is devised to build the relationship between a set of objects by processing their appearance feature and geometric feature; the established AMIM integrates the message passed from the object relational graph by designing a multi-dimensional attention model. After that, the object relationship and target feature are fed into the GRU to encode the contextual information. Finally, these features are fed into two sibling fully-connected layers, namely, a box-classification layer and a bounding box regression layer.

Recently, the recurrent neural network has been widely used in the field of target detection. Meanwhile, several interesting works combine convolution neural networks and recurrent neural networks to model target relationships. Song et al. [59] exploit recurrent neural networks to encode the co-occurring frequency of object-to-object relation to the features feeding into the classifiers. As a simplified form of recurrent neural network, Gated Recurrent Unit (GRU) is first proposed by [60]. Liu et al. [52] propose scene GRU and edge GRU to encode the message in the whole image and the regions. In [61], the edge GRU and node GRU are proposed to generate a scene graph, which is used as a platform to model target relationships. In this work, the GRU layer is applied to exploit the contextual information of the remote sensing image.

Structured graphs have been used extensively to solve the problems in visual recognition. A structure inference machine is designed to model the relationship between the group activity in [62]. Hu et al. [63] propose a generic structured model to encode the relationship between scene and object, and the model is employed to improve the image classification accuracy. Graph Convolutional Recurrent Network (GCRN) [64] combines the recurrent neural network and structured graph to identify spatial relationships and find dynamic patterns. In [65], image classification is optimized by the structured prior knowledge in the form of a knowledge map. Liu et al. [52]. incorporated a Structure Inference Network (SIN) into a universal target detection framework to infer object state. Though this work is inspired by [52], which exploits graph structure inference to model the target relationships, there are essential differences between the two approaches, i.e., the information acquisition mode and the information fusion mode.

Attention mechanism is an essential information processing method in human perception. There are many attempts to combine attention mechanism with deep neural networks to improve the ability of image processing. Li et al. [58] presented attention-based Long Short-Term Memory (LSTM) layers to exploit global context for object detection. Fan et al. [66] propose an Attention-RPN to detect objects of unseen categories with only a few annotated examples. Attention-CoupleNet in [67] is proposed to incorporate the attention-related information of objects by designing a cascade attention structure. Song et al. [68] propose a multi-scale attention deep neural network (MSA-DNN) for object detection, which uses multiple attention blocks with different scales to exploit the global semantic information. In [69], Convolutional Block Attention Module (CBAM) is proposed to infer attention maps along two separate dimensions. Inspired by [69], AMIM is presented to integrate the message passed from the object relational graph by designing a multi-dimensional attention model.

This paper makes the following four contributions: (1) this paper proposes a novel target detection framework that makes effective use of contextual information. Most target detection methods only focus on the interior features of targets, and the exterior context information is usually ignored. The DCIFF-CNN integrates diversified contextual information to capture more fine-grained details and infer contextual object instances. (2) A novel ORSG module is designed to model the relations between a set of objects by processing their appearance features and geometric information. The contextual information in the scene is extremely complex, especially for the remote sensing image. (3) The AMIM is integrated into the feed-forward convolutional neural networks with negligible overheads

to highlight the useful object-level contextual information. (4) An end-to-end deep learning multi-target detection framework integrating MLC-CRPN, ORC-TDN, ORSG and AMIM is explored for remote sensing images. Over all, the use of interior local contextual information and external object-level contextual information in the proposed method contributes to an enhanced feature representation scheme for target detection. Several state-of-the-art target detection methods are adopted to evaluate the proposed method. The experimental results indicate that the proposed method improves the target detection performance with more desirable and reasonable outputs.

The contents of this paper are organized as follows. The target detection scheme based on interior local contextual information and external object-level contextual information for remote sensing image is discussed in Section 2. Section 3 describes the experimental results on various remote sensing image datasets, and discusses the different experimental parameters. Section 4 concludes the full paper.

2. Materials and Methods

The target detection framework of DCIFF-CNN is exhibited in Figure 2, which is mainly composed of two parts, namely, MLC-RPN and ORC-TDN. The MLC-RPN fuses the multi-scale convolution-layer features to capture the region proposals; the ORC-TDN models the relationship between targets and objects explicitly. The ORSG modules and the AMIM module are embedded into ORC-TDN to infer contextual object instances and obtain more meaningful contextual information. The ORSG models the explicit relationship between a set of objects by processing their appearance feature and geometric feature, and the AMIM integrates the message passed from the object relational graph by designing a multi-dimensional attention model. The contextual information inferred by the two modules is then fed into GRU for target detection. Compared with traditional CNN-based target detection methods, DCIFF-CNN shows an advantage in exploiting the relationship of context to reason the target. The modules in the DCIFF-CNN framework will be described in the following sections.

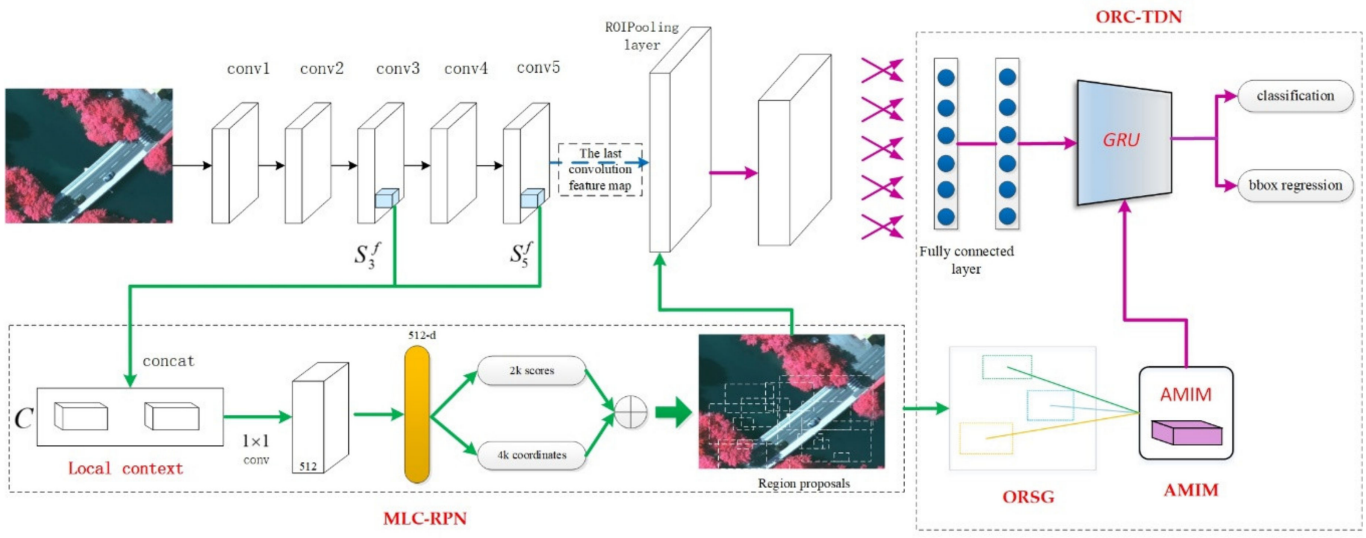

Figure 2. The overall framework of diversified context information fusion framework based on convolutional neural network (DCIFF-CNN).

2.1. Multi-Scale Local Context Region Proposal Network (MLC-RPN)

Similar to RPN [58], an image of any size is fed into the MLC-RPN, and a group of region proposals with classification probabilities is then output from the MLC-RPN. The general framework of the MLC-RPN is displayed in Figure 3. A pre-trained convolutional neural network VGG [70] is employed in the MLC-RPN. Meanwhile, features maps of

different levels are integrated to extract the region proposal. The scale of targets in remote sensing images with various resolutions is different. Generally, the scale of targets in low-resolution images is smaller; the scale of targets in high-resolution images is larger. Additionally, the information contained in the small-scale target is more macroscopic and abstract, whereas the information contained in the large-scale target is richer and more detailed. One issue in traditional convolutional networks is that the detection performance for small-size instances is unstable. This is because the feature map of the last layer in the networks is too coarse to make an accurate classification. For instance, an object with size of 64 × 64 going to the last convolutional layer of a VGG16 network will have a size of 4 × 4. Inspired by the structured object modeling network in natural images [52], a multi-scale local contextual feature fusion method is proposed in this paper that can take full advantage of different information to represent rich features of ground objects. Different from the method in SSD [43], the proposed MLC-RPN focuses on the multi-resolution region, whereas the method in SSD extracts regions of different scales. In the CNN network, a certain feature scale changes with the iteration of the network. When the scale change is unreasonable, it may be necessary to use the feature of multiple scales for auxiliary processing. Furthermore, the features of different scales are fused to realize feature complementation. In the CNN network, layer 3 and layer 5 are selected for context feature fusion at different scales. The sampled feature representations of f with various scales is denoted as $\left\{S_i^f | i = 3, 5\right\}$. In the MLC-RPN, the feature map of the last layer must have a shape of 512 × 7 × 7, so that it can match the dimension and amplitudes of the fully-connected layer. To meet this requirement, the local response normalization (LRN) is applied to normalize multiple feature maps f to match the original amplitudes. Then, these features are concatenated into $C = concat\left\{S_i^f\right\}\{i = 3, 5\}$, where *concat* indicates that the feature maps are spliced in the channel dimension. Ultimately, a 1 × 1 convolution layer is adopted to compress the sampled features into a uniform space.

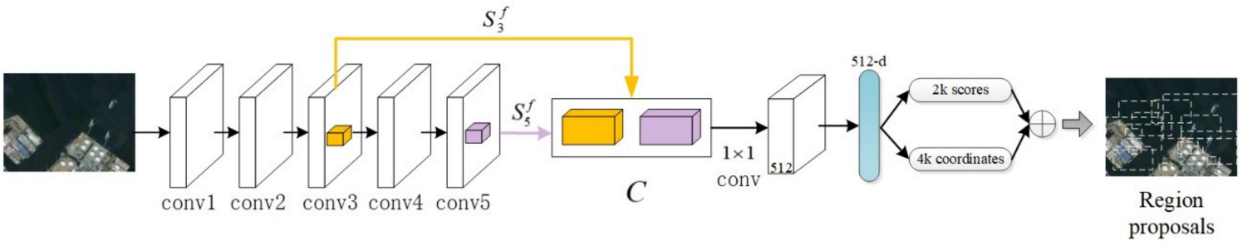

Figure 3. The general framework of the multi-scale local context region proposal network (MLC-RPN).

The resulting feature map is sent to two side-by-side fully-connected layers, i.e., the box-classification layer and box-regression layer. Denoting the number of region proposals as k, the regression layer will have $4k$ outputs, which are encoded to represent the coordinates of k boxes. Additionally, $2k$ scores will output from the box-classification layer, which is used to estimate the probability of each region proposal.

2.2. Object-Level Relationships Context-Based Target Detection Network (ORC-TDN)

The final target detection results cannot be directly obtained from the region proposals output by MLC-RPN because they only show the area where the suspected target may appear. To further detect the target accurately, ORC-TDN is added behind the MLC-RPN to extract complete target features. The establishment of the proposed ORC-TDN method adds the attentional message integrated module (AMIM) and the object relational structured graph (ORSG) into the existing CNN architecture. The AMIM merges contextual information and aggregate information from a set of elements. The ORSG processes a set of objects simultaneously through encoding the appearance feature and geometric feature, so that the contextual object instances can be inferred.

2.2.1. Problem Statement

In the real world, targets always appear in a specific scene, and some connection exists between targets and other objects. In remote sensing images, context refers to any information in the image that directly or indirectly affects the perception of the target in the scene. For example, the basketball court may look like a tennis court, but it is definitely ships, not vehicles, that can appear on the water's surface.

Inspired by the structured object modeling network in natural images [52], the structured graph can be deployed to adaptively aggregate information from contexts. To learn latent representations of objects, a visually-grounded graph most accurately related to the image is generated [61]. Shown in Figure 4, a relationship graph $R = (T, O, E)$ is proposed to model the graphical problem between target and context information. The node $t \in T$ represents the target, while node $o \in O$ is the contextual area (object), and $e \in E$ denotes the edge (relationship) between each pair of nodes.

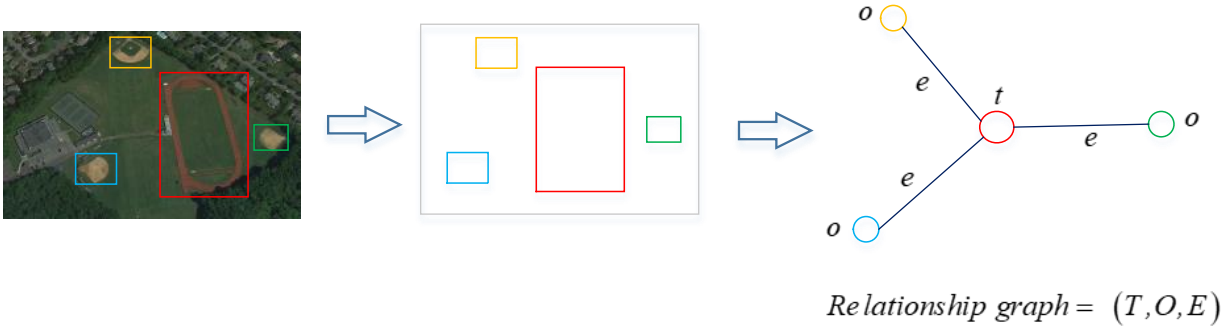

Figure 4. The graphical problem between target and objects.

GRU is a lightweight and effective recurrent neural network, which takes its previous node information and incoming messages as input, and generates an updated node information as output. It can be seen that passing messages among the GRU units along the context areas is feasible. The structure of GRU is illustrated in Figure 5, and there are two gate functions to regulate the flow of information, namely the update gate z and the reset gate r. The update of hidden state is decided by the update gate, and the status message of the previous moment is related to the value of the update gate. A large value of the update gate means that more state information from the previous moment is introduced, and vice versa. The reset gate is applied to restrain the amount of state information from the previous state that needs to be ignored. A small value of the reset gate means that more state information from the previous state is ignored, and vice versa. The update gate and the reset gate are computed as follows:

$$z_t = \sigma(W_z[h_{t-1},\ x_t]) \tag{1}$$

$$r_t = \sigma(W_r[h_{t-1},\ x_t]) \tag{2}$$

where W_r and W_z are the weight matrixes that can be learned; $[,]$ represents the concatenation of vectors, and σ denotes the logistic sigmoid function. h_{t-1} and x_t represent the previous hidden state and the current input, respectively. The output of GRU is denoted as

$$h_t = z_t h_{t-1} + (1 - z_t)\widetilde{h} \tag{3}$$

where

$$\widetilde{h} = \phi(W_h[h_{t-1} * r_t,\ x_t]) \tag{4}$$

$\widetilde{h}_t$ denotes the new hidden state; ϕ is activation function; W_h is a learnable weight matrix, and $*$ represents the element-wise multiplication. As shown in the above formulations, the updated information of z_t can be used for both forgetting information and selecting

information. Generally, GRU can be applied to remember long-term information, and the input of GRU is a symbol sequence, where the initial state of GRU can be set as null or a random vector. In this paper, the GRU is utilized to encode various contextual information of the target state.

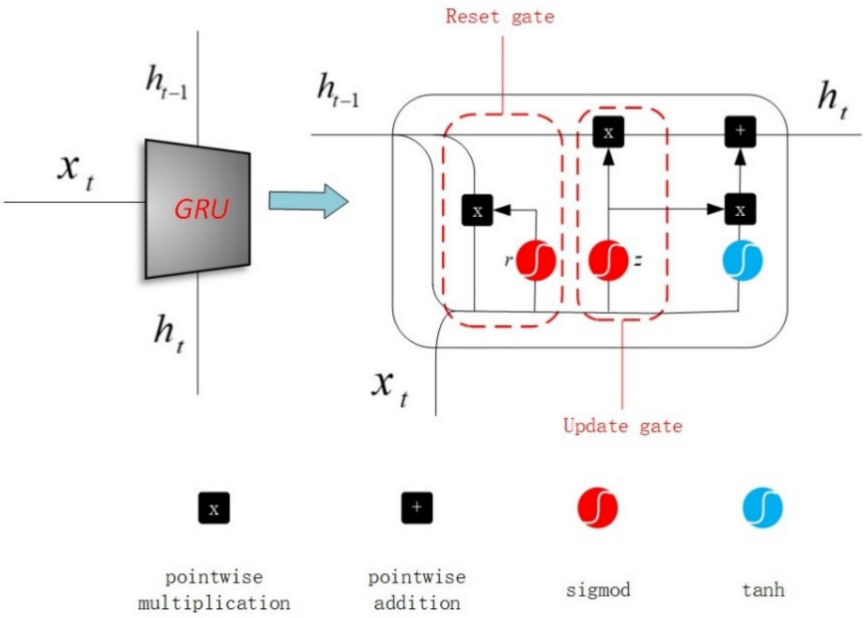

Figure 5. Illustration for gated recurrent unit (GRU).

2.2.2. Attentional Message Integrated Module (AMIM)

In the relationship graph, it is important to encode and integrate the messages transmitted by the node. Since each node needs to receive multiple incoming messages and these messages can be updated at any time, it is essential to exploit a polymerized function that can merge the incoming messages into meaningful representations and remember the updated node information at the same time. Different from the context information in natural images, the features of target in remote sensing images are spare and the background is complex, causing much useless background information in the context information. In this case, an AMIM is deployed to find and fuse the useful contextual information. Inspired by [69], each attention module in this paper predicts an attention map M, which contains both channel dimension and spatial dimension. Meanwhile, AMIM differs from the attention mechanism in [69] in that it processes and incorporates multiple contextual information instead of dealing with one piece of information separately. The feature map of input is denoted as $\mathbf{f} \in \mathbb{R}^{C\times H\times W}$, which is compressed through spatial dimension and channel dimension, respectively. In this way, a multi-dimensional calculation can be carried out to obtain more accurate and effective features. The feature map $\mathbf{F}^c = [\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_i, \cdots, \mathbf{f}_C]$ and $\mathbf{F}^s = [\mathbf{f}_{(1,1)}, \mathbf{f}_{(1,2)}, \cdots, \mathbf{f}_{(j,k)}, \cdots, \mathbf{f}_{(H,W)}]$ are given as the combination of channels $\mathbf{f}_i \in \mathbb{R}^{1\times H\times W}$ and spatial $\mathbf{f}_{(j,k)} \in \mathbb{R}^{C\times 1\times 1}$, respectively. The channel attention map $\mathbf{A}^c \in \mathbb{R}^{C\times 1\times 1}$ is obtained by performing average pooling and max pooling on the features, and it is derived as follows.

$$\mathbf{A}^c(\mathbf{f}) = \sigma\left(\mathbf{W}_1\left(\delta\left(\mathbf{W}_2\mathbf{z}^c_{avg}\right)\right)\right) + \sigma(\mathbf{W}_1(\delta(\mathbf{W}_2\mathbf{z}^c_{max}))) \tag{5}$$

$$\mathbf{z}^c_{avg}(i) = \frac{1}{H\times W}\sum_{j=1}^{H}\sum_{k=1}^{W}\mathbf{f}_i(j,k) \tag{6}$$

$$\mathbf{z}^c_{max}(i) = \max_{\substack{j\in\{1,2,\cdots,H\}\\ k\in\{1,2,\cdots,W\}}}\{\mathbf{f}_i(j,k)\} \tag{7}$$

where $\mathbf{W}_1 \in \mathbb{R}^{C\times\frac{C}{2}}$ and $\mathbf{W}_2 \in \mathbb{R}^{\frac{C}{2}\times C}$, σ and δ stand for the sigmoid function and the ReLU operator, respectively. Note that the weights $\mathbf{W}_1$ and $\mathbf{W}_2$ come from two fully connected layers. $\mathbf{z}^c_{avg} \in \mathbb{R}^{C\times1\times1}$ and $\mathbf{z}^c_{max} \in \mathbb{R}^{C\times1\times1}$ denote features from average pooling and max pooling, respectively. Similarly, the spatial attention map $\mathbf{A}^s \in \mathbb{R}^{1\times H\times W}$ is calculated by

$$\mathbf{A}^s(\mathbf{f}) = \sigma(\mathbf{W}_3 * (\mathbf{z}^s_{avg}; \mathbf{z}^s_{max})) \tag{8}$$

$$\mathbf{z}^s_{avg}(j,k) = \frac{1}{C}\sum_{i=1}^{C}\mathbf{f}_{(j,k)}(i) \tag{9}$$

$$\mathbf{z}^s_{max}(j,k) = \max_{i\in\{1,2,\cdots,C\}}\left\{\mathbf{f}_{(j,k)}(i)\right\} \tag{10}$$

where σ and $*$ represent the sigmoid function and the convolution operation, respectively. $\mathbf{W}_3 \in \mathbb{R}^{1\times1\times C\times1}$ stands for the weight of the convolution layer $\mathbf{z}^s_{avg} \in \mathbb{R}^{1\times H\times W}$ and $\mathbf{z}^s_{max} \in \mathbb{R}^{1\times H\times W}$ denotes features from average pooling and average pooling, respectively.

The structure of AMIM is illustrated in Figure 6, which includes attention modules from both the channel dimension and the spatial dimension. The attention module of the channel dimension and spatial dimension derives an attention map with one-dimension and two-dimension, respectively. In this paper, the two modules work in parallel to extract important information efficiently. If the input feature map values highly on both channel re-scaling and spatial re-scaling, the feature will be given higher activation.

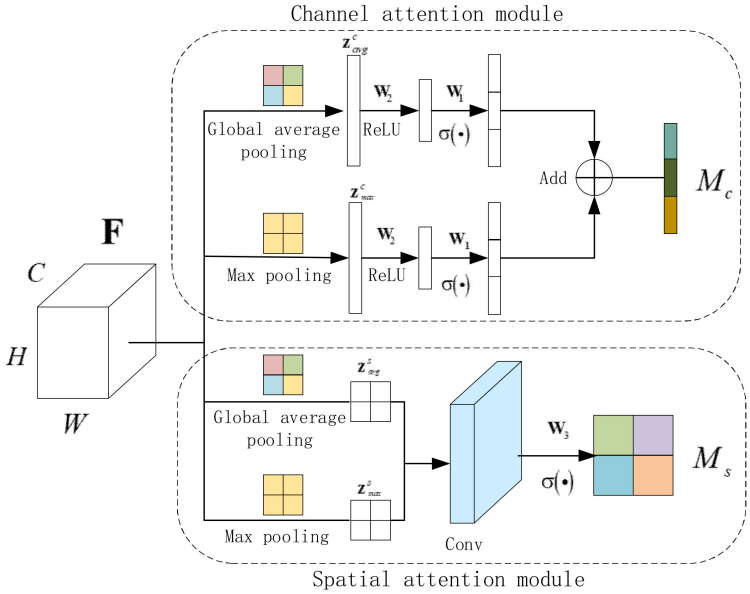

Figure 6. Illustration of the attentional message integrated module (AMIM).

2.2.3. Object Relational Structured Graph (ORSG)

The purpose of using GRU in this paper is to effectively transfer messages from targets and objects to nodes. It is important to devise a transmission function that can transfer information from all areas into a meaningful representation. The feature maps of object are taken as the initial value for the GRU, and the input value comes from the integrated message of other nodes. The relationship between the object and the target changes with the state update of objects, and the updated time step makes the model more stable. In this paper, a structured graph of object relationship is proposed to update the message of nodes, as shown in Figure 7. In ORSG, the deep neural network is combined with the graph model for a structured prediction task that is solved by structural reasoning technology. The GRU in Figure 7 is applied to encode messages from objects. Since there are multiple objects, an integrated message m_i needs to be pre-calculated. If a long sequence of messages from

every object is taken as input, a lot of time will be consumed and there will be much useless information. With the addition of AMIM, the GRU updates the node with some selected important integrated messages. It is reasonable that different objects contribute different messages to the node, and the target-object relationship $e_{j\to i}$ is denoted as the influence of object on target. Obviously, the influence of an object on the target consists of both the appearance feature and the geometric feature. As shown in Figure 7, the integrated message to the target node is calculated as

$$\mathbf{m}_i = \mathbf{A}^c\left(\mathbf{F}_{o_j\to t_i}\right) * \mathbf{F}_{o_j\to t_i} + \mathbf{A}^s\left(\mathbf{F}_{o_j\to t_i}\right) * \mathbf{F}_{o_j\to t_i} \tag{11}$$

$$\mathbf{F}_{o_j\to t_i} = e_{j\to i} * \mathbf{f}_{o_j\to t_i} \tag{12}$$

$$e_{j\to i} = \sum_{j:j\to i} \tanh\left(W_f\left[f_{t_i}, f_{o_j}\right]\right) * relu\left(W_s S_{j\to i}\right) \tag{13}$$

where $\mathbf{A}^c$ and $\mathbf{A}^s$ are attention weight coefficients in Section 2.2.2; $\mathbf{F}_{o_j\to t_i}$ denotes the message from object o_j to target t_i; $\mathbf{f}_{t_i}$ and $\mathbf{f}_{o_j}$ represent the appearance feature of target t_i and object o_j, respectively; f_{t_i} and f_{o_j} are scalar forms of $\mathbf{f}_{t_i}$ and $\mathbf{f}_{o_j}$, respectively; $e_{j\to i}$ is a scalar weight; W_f and W_s are learnable weight matrixes; $[,]$ denotes the concatenation of vectors; $*$ represents the element-wise multiplication. The geometric feature $S_{j\to i}$ is defined as follows

$$S_{j\to i} = \left[\log\left(\frac{|x_i - x_j|}{w_j}\right), \log\left(\frac{|y_i - y_j|}{h_j}\right), \log\left(\frac{w_j}{w_j}\right), \log\left(\frac{h_i}{h_j}\right)\right] \tag{14}$$

where x_j and y_j represent the top-left coordinates of bounding-box for the context area. w_j and h_j stand for the width and the height of bounding-box for the context area, respectively. The first two elements are converted by $\log(.)$ to calculate more nearby objects.

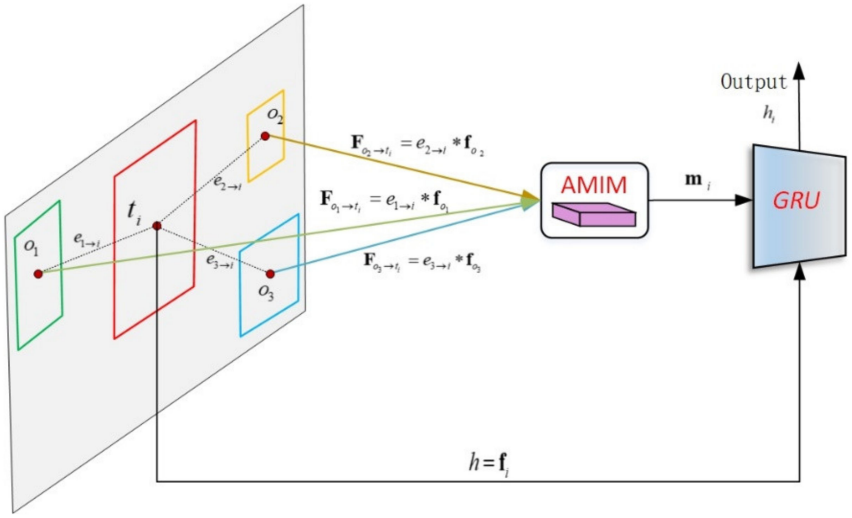

Figure 7. Object relational structured graph (ORSG).

Figure 7 shows the object-level relational structured graph. For target t_i, the visual feature $\mathbf{f}_i$ of the node t_i is taken as the initial hidden state of GRU. $e_{1\to i}$ is applied to calculate the message $\mathbf{F}_{j\to i}$ from node o_1 to node t_i. The AMIM integrates all the messages from the object into $\mathbf{m}_i$, and the result is input to the GRU. The h_i output by GRU is then used as the final updated node state. Based on the description of GRU in Section 2.2.1, Equation (4) can be expressed as follows:

$$h_i = \phi(W_h[\mathbf{f}_i * r_i, \mathbf{m}_i]) \tag{15}$$

In the following iterations, the new target-object message is taken as the new input of GRU, and the next hidden state is further calculated.

2.2.4. Target Detection Process

As mentioned earlier, the multi-class remote sensing target detection method proposed in this paper involves many contents, and it seems to be a little complex. To facilitate an easy understanding, the target detection flow path of the proposed method is briefly described in Algorithm 1. In the ORC-TDN, the final integrated node representations of GRU are used to predict target category and bounding box regression. Supposing that b and b' denote the predicted bounding box and the true bounding box, respectively, p and p' represent the predicted class probability and the true class probability. The loss function of each target is optimized as follows.

$$L = L_{cls}(p, p') + \lambda L_{reg}(b, b') \tag{16}$$

$$L_{cls}(p, p') = -\log\left[p'p + (1 - p')(1 - p)\right] \tag{17}$$

$$L_{reg}(b, b') = smooth_{L_1}(b - b') \tag{18}$$

where λ is the balancing parameter, and $smooth_{L_1}$ is the smooth L_1 loss proposed in [40]. To verify the proposed method, corresponding experiments are given in the next section.

Algorithm 1: The target detection method for remote sensing images based on structured object-level relational reasoning.

Input: Remote sensing image dataset
Output: Bounding boxes and target category of multi-class targets
1. Get ROIs (region proposals) through MLC-RPN.
1.1. **Set:** The feature map from conv3 S_3^f, The feature map from conv5 S_5^f
1.2. Get the fused features $C = concat\left\{S_i^f\right\}\{i = 3, 5\}$
1.3. Perform 1×1 convolution on C and send it to the full connection layer.
1.4. Get ROIs
2. The ROIs are fed to ORC-TDN.
2.1. Establish the ORSG and AMIM
2.2. ORSG includes node t_i, node o_j and edge $e_{j \to i}$, edge is determined by relative object feature and position in ROIs
2.3. Calculate the message $\mathbf{F}_{j \to i}$ from node o_j to node t_i through Formula (12)–(14)
2.4. Obtain the channel attention map $\mathbf{A}^c$ and spatial attention map $\mathbf{A}^s$ of $\mathbf{F}_{j \to i}$ through Formula (5)–(10).
2.5. Obtain the integrated message $\mathbf{m}_i$ of the object context through Formula (11)
2.6. The context information $\mathbf{m}_i$ and appearance feature $\mathbf{f}_i$ of target are taken as the input of GRU. Obtain the output of GRU through (15).
2.7. The output of GRU is fed into the full connection layer.
3. Obtain bounding boxes and target category of multi-class targets from the full connection layer of ORC-TDN.

3. Results

Sufficient experiments have been undertaken in this section to demonstrate the capability of the proposed DCIFF-CNN. The image data used for experiments are introduced first. Then, some common evaluation criteria are listed. Finally, the effectiveness and robustness of the proposed target detection method is validated.

3.1. Dataset Description and Experimental Settings

A multi-class target detection dataset, NWPU VHR-10 [8,46,71], is used to validate the proposed method. This dataset includes 650 optical remote sensing images, and they are divided into ten different types of target. At least one kind of target exists in each image, and most images contain more than one type of targets. For these images, 565 images were

obtained from Google Earth, and the other 85 color images were obtained from Vaihingen data. In this paper, 20% of the dataset is applied to train and adjust the model; another 20% to verify the model, and the last 60% to test the model. Since most of the methods adopt this ratio, in order to reflect the applicability of the proposed method, this ratio is adopted for experiments in this paper. At the same time, in order to make the results more convincing, the comparison methods used in this paper also use this ratio for experiments. The detailed information of the dataset, i.e., target sizes and target numbers are listed in Table 1.

Table 1. A detailed introduction to NWPU VHR-10 dataset.

Target Classes	Target Numbers (Pixels)	Target Sizes (Pixels)
Airplane	757	50 × 77–104 × 117
Ship	302	20 × 40–30 × 52
Storage tank	655	27 × 22–61 × 51
Baseball diamond	390	66 × 70–109 × 129
Tennis court	524	45 × 54–122 × 127
Basketball court	159	52 × 52–179 × 179
Ground track field	163	195 × 152–344 × 307
Harbor	224	95 × 32–175 × 50
Bridge	124	88 × 90–370 × 401
Vehicle	477	20 × 41–45 × 80

To improve quantity and diversity of the images for testing, another real dataset collected by ourselves is also used in the experiments. This dataset consists of three types of targets, i.e., airplane, ship, and car. In order to diversify the context information, ships exist in two kinds of scenes: river and sea. The airplane and ship in the dataset are collected from Google Earth, and car is photographed by an unmanned aerial vehicle of DJI M100, which is equipped with a DJI Zenith Z3 camera. Since this dataset contains real shooting images, it can better demonstrate the robustness and practicability of the proposed method. The detailed information of the collected data is listed in Table 2.

Table 2. The detailed information of the collected dataset.

Target	Target Context	Image Size (Pixel)	Number of Targets
airplane	runway	877 × 768	1500
Ship	sea	1104 × 740	1000
Ship	river	1104 × 740	1000
car	road	1280 × 720	1500

A PC equipped with CPU of Intel Core i7, random access memory with capacity of 16GB, and a GPU of NVIDIA GTX-1080 is used to perform the experiment. Meanwhile, this PC runs the operating system of Ubuntu 14.04.

3.2. Evaluation Metrics

To verify the performance of the proposed target detection framework, two widespread criteria are adopted to estimate the detection results quantitatively, namely, precision rate and recall rate. The precision rate denotes the ratio of predicted positive targets to all actual positive targets, and the recall rate indicates the ratio of predicted positive targets to all actual positive targets. They are calculated as follows

$$precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{19}$$

$$recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{20}$$

where TP (true positives) represents the number of predicted positive targets, FP (false positives) denotes the number of falsely detected targets and FN (false negatives) is the number of falsely detected backgrounds. When the overlapping area between the ground truth and the bounding box is greater than 0.5, the area is defined as TP. On the contrary, when the overlapping area between the ground truth and the bounding box is less than 0.5, the area is defined as FP.

The AP represents the mean value of precision within the range from recall 0 to recall 1 and it is obtained by computing the area under the precision-recall curve. The mean AP (mAP) indicates the mean value of AP across several classes.

AC and PR are adopted to quantify the target detection result of the collected dataset. AC denotes the accuracy of detection result and PR indicates the precision ratio.

$$\mathrm{AC} = \frac{Number\ of\ detected\ target}{Number\ of\ target} \tag{21}$$

$$\mathrm{PR} = \frac{Number\ of\ detected\ target}{Number\ of\ detected\ target + Number\ of\ detected\ background} \tag{22}$$

3.3. Target Detection Results on NWPU VHR-10 Dataset

To evaluate the robustness and effectiveness of the proposed method, another eleven methods are taken for performance comparison. Among these methods, there are four traditional methods, which are widely used in target detection. The remaining seven methods are based on deep learning, which have made a great breakthrough in the field of target detection. SSCBow is chosen as the representative bow-based method. COPD is taken to represent the SVM-based method, and FDDL is employed as a typical method based on sparing coding. In this paper, YOLO1 [72], YOLO2 [72], YOLO3 [42], YOLO4 [73], RICNN [46], FRCNN [41], MSCNN [74], and SSD [43] are selected as the deep learning approaches. FRCNN and YOLO are representative CNN-based methods for the target detection; MSCNN and SSD focus on the multiple scales, and RICNN is widely used to assess new methods, especially the target detection method in remote sensing images.

Figure 8 shows the detection results of these methods. It can be seen that the methods based on deep learning exhibit obvious advantages compared with traditional methods. The traditional manual methods have a limitation of only extracting the artificial features of the bottom layer, whereas the method based on deep learning can extract both the explicit features from the bottom layer and abstract features from the top layer. Furthermore, the proposed DCIFF-CNN achieves the highest mean average precision (mAP) among all the methods, and it obtains a higher AP among all targets. Furthermore, it can be seen that YOLO1 gets a slightly lower mAP value. This is because the input image in YOLO1 is divided into a S $\times$ S grid which can only deal with two targets simultaneously, leading to its weak ability in detecting small and dense targets such as storage tanks. The mAP value of YOLO2 and YOLO3 are both higher than that of YOLO1. In addition, the detection performance measured in mAP of DCIFF-CNN is slightly higher than that of RICNN. This is because the RICNN is implemented in multiple pipelines, whereas the proposed DCIFF-CNN is an end-to-end network that can skip intermediate steps. Furthermore, the performance of SSD is better than other baseline methods due to the multiple feature maps and pixel resampling stages employed by SSD. However, the DCIFF-CNN obtains higher AP values in all kinds of targets. For the target of ground track fields, MSCNN obtains a satisfying AP value. However, the performance of this method is unstable for detecting small targets. Consequently, the experiment results in Figure 8 illustrate that the proposed DCIFF-CNN obtains better robustness and effectiveness for target detection, which benefits from the multi-scale local context and object-level relationships context.

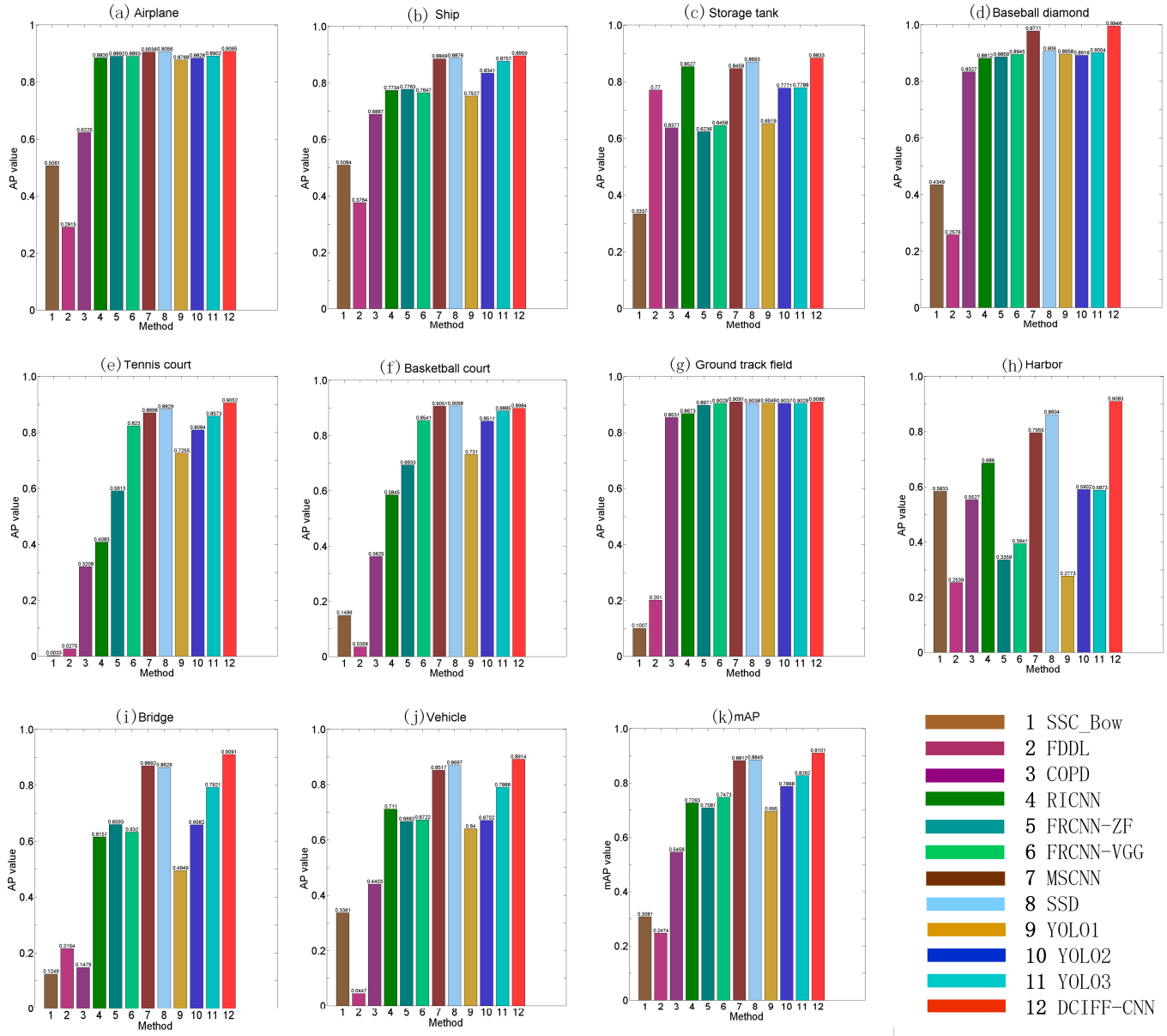

Figure 8. Performance comparisons of twelve different methods in terms of average precision (AP) values. (**a**) The AP value of an airplane; (**b**) The AP value of a ship; (**c**) The AP value of a storage tank; (**d**) The AP value of a baseball diamond; (**e**) The AP value of a tennis court; (**f**) The AP value of a basketball court; (**g**) The AP value of a ground track field; (**h**) The AP value of an harbor; (**i**) The AP value of a bridge; (**j**) The AP value of a vehicle; (**k**) The mAP value of ten classes of targets. All the methods are carried out with the same dataset and the same data ratio (Dataset: NWPU-VHR (train:20%, val:20%, test:60%)).

Some typical detection results from the proposed DCIFF-CNN method are exhibited in Figure 9, where DCIFF-CNN achieves an outstanding performance in all ten categories of targets. Figure 9a shows the detected airplane with different colors and various directions. Figure 9b,i display the detected targets in complex backgrounds. Figure 9c,d show the detected harbor and bridge, respectively, and the detection accuracy of these targets can reach 99%. The detected targets in Figure 9e,f conform to the same category, and they are similar in color, shape, and size. Figure 9j–l show the detection results of vehicles in various backgrounds. It is demonstrated that the proposed method has excellent target detection capability under various complex backgrounds. In Figure 9b,g,h the storage tank crowds together, but they are detected and recognized efficiently. In conclusion, the proposed

method obtains accurate and stable results for target detection and recognition in different categories and different scenarios.

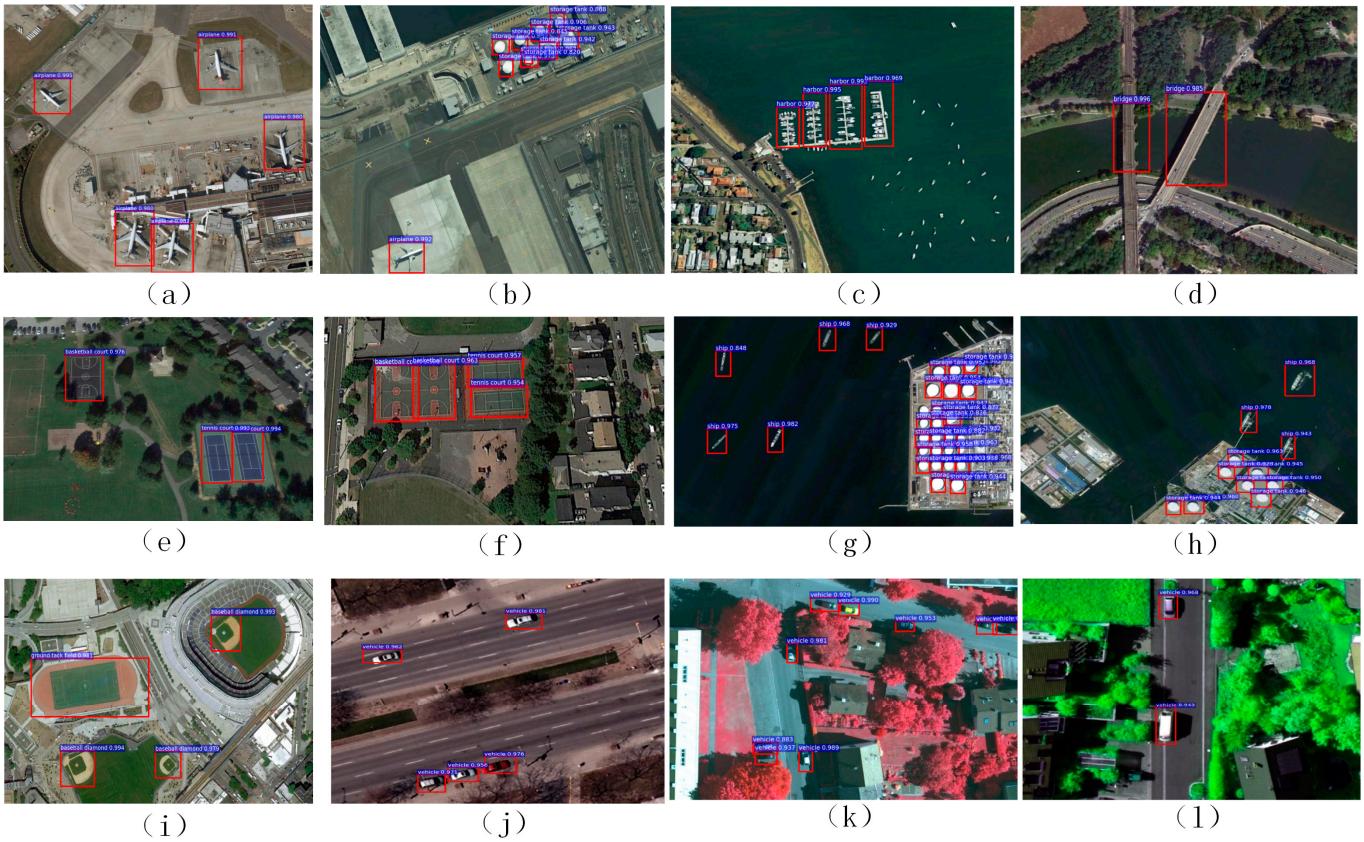

Figure 9. Some target detection results with the proposed approach.

The detection results of some different methods are displayed in Figure 10. The size of the storage tank in the image is extremely small, i.e., 27 × 27 to 30 × 30 pixels. Furthermore, the targets in the image are crowded together. In Figure 10, the red boxes, the blue boxes and the green boxes denote the correct detected targets, the incorrect detected targets, and the missing targets, respectively. It is verified that the proposed DCIFF-CNN method is more robust to the detection of small targets than other methods.

The precision-recall curves (PRCs) of eight methods are displayed in Figure 11. For a better comparison and visualization, the range of different targets in the Y-axis is adjusted according to the actual detection results. It can be seen that most methods achieve stable results for detecting airplane, ground track field, and baseball diamon. This is because these targets have special features in this dataset, which are of great help to target detection. For example, the appearance feature of the airplane in this dataset is unique; the ground track field has an obviously larger size than any other targets, and the shape of the baseball diamond is unique. Meanwhile, the proposed method is excellent in both precision and recall for the target detection of ship, tennis court, harbor, bridge and vehicle. The PRCs of these targets have little fluctuation, and an effective balance is achieved between the detection rate and the recall rate. It is indicated that the proposed explicit context model can effectively exploit rich contextual information to detect different targets more accurately. For the target of a storage tank, the variation range of PRC is slightly larger, and the recall rate decreases obviously with the increase of precision rate. This is because the size of this target is relatively small, and its context environment is more complex, as this target may appear in the sea or an urban environment at the same time. However, the proposed method is still superior to other methods. For the target of a basketball court, the

proposed method is somewhat inferior to the MSCNN method. This is because the similar characteristics of basketball courts and tennis courts bring some difficulties to the target detection. However, the proposed method is still superior to most methods. Overall, these results show that the added AMIN and ORSG can integrate useful contextual information, and generate object-level graphs through the background information.

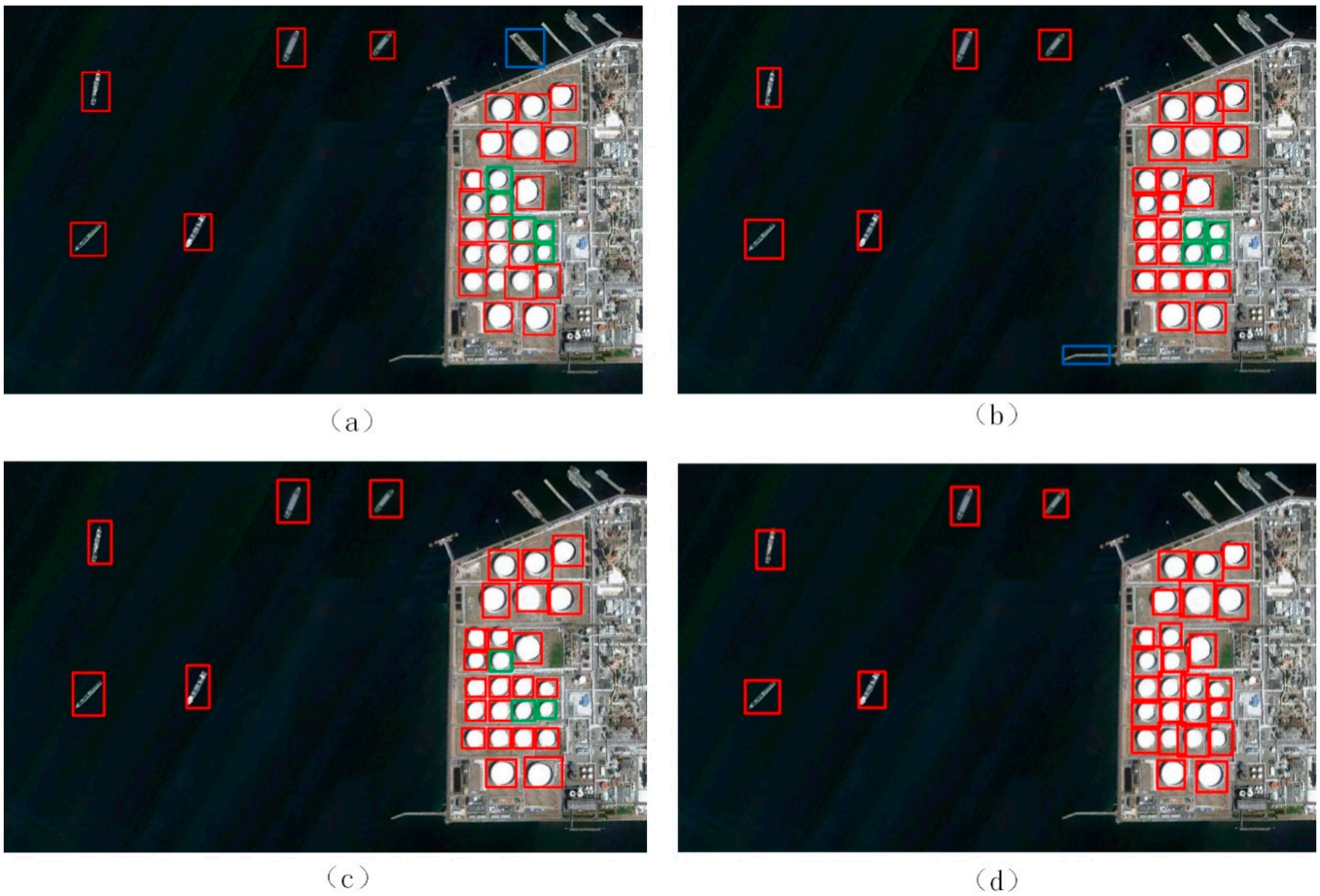

Figure 10. Some target detection results of some different methods. (**a**) FRCNN-VGG; (**b**) YOLO3; (**c**) SSD; (**d**) DCIFF-CNN.

The computational complexity analysis between these methods is provided in Table 3. As can be seen in Table 3, YOLO-based methods are more efficient than other methods, but with some compromise of detection accuracy. SSD gets faster speed due to the single shot multi-box detector. It can be seen that the proposed DCIFF-CNN achieves a tradeoff of detection accuracy and computation efficiency. This demonstrates that the proposed method is robust in both detection accuracy and speed. This might benefit from the use of attentional contextual information, which discards useless contextual information.

Table 3. The computational complexity analysis for different methods in terms of NWPU VHR-10 dataset.

Method	FRCNN-ZF	FRCNN-VGG	MSCNN	SSD	YOLO1	YOLO2	YOLO3	DCIFF-CNN
Mean times(s) (Testing for per image)	1.31	1.55	1.62	1.22	0.94	1.03	1.15	1.20

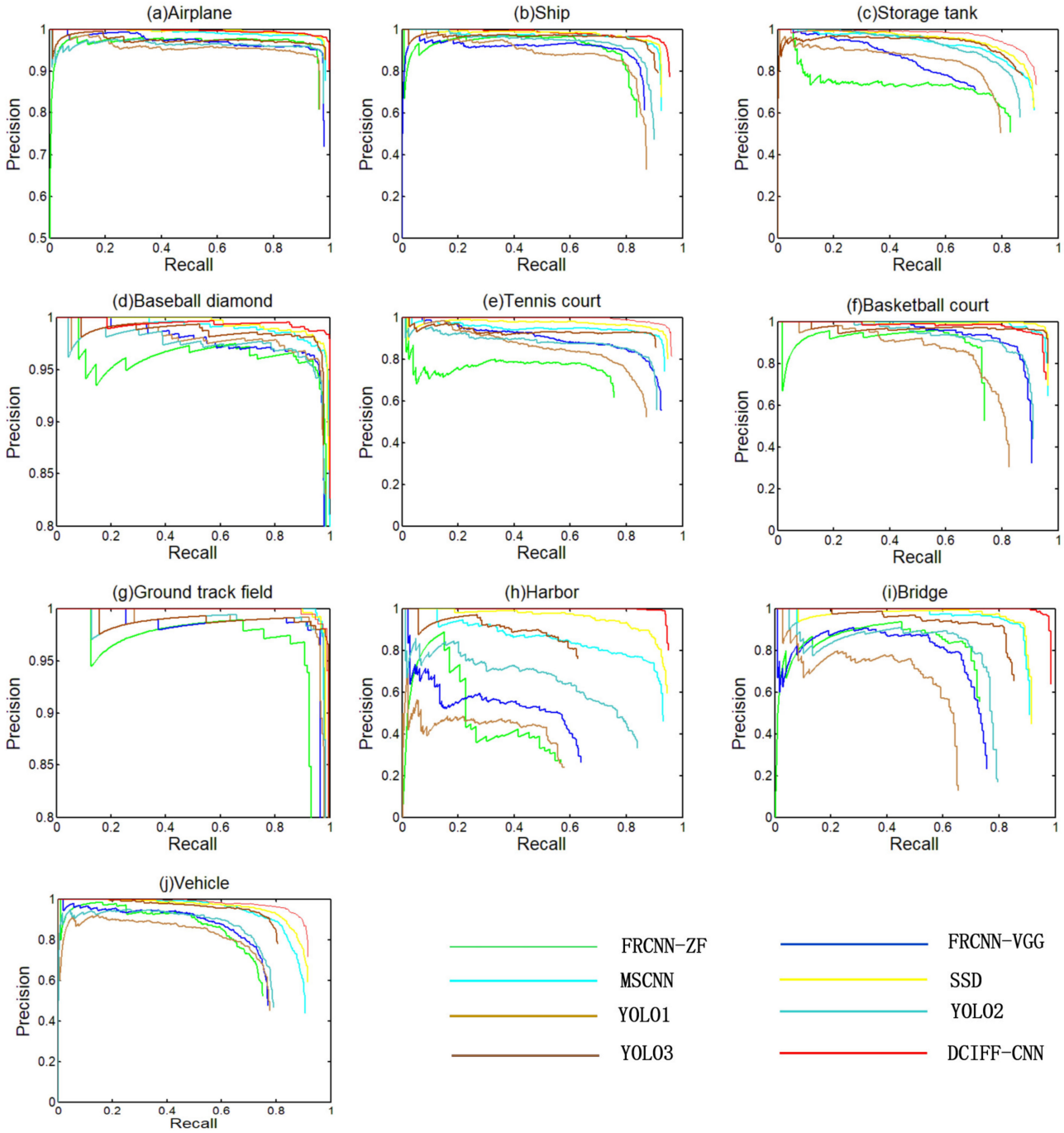

Figure 11. The precision-recall curves (PRCs) of the proposed method and other compared methods. (**a**) Airplane; (**b**) Ship; (**c**) Storage tank; (**d**) Baseball diamond; (**e**) Tennis court; (**f**) Basketball court; (**g**) Ground track field; (**h**) Harbor; (**i**) Bridge; (**j**) Vehicle.

3.4. Target Detection Results of the Collected Dataset

Some heat maps of the collected dataset are displayed in Figure 12, where the first row and the second row denote the original image and the corresponding heat map, respectively. For the target of airplane, it can be seen that both the target and the parking apron are emphasized. For the target of ship, the urban area and the sea surface are clearly demarcated. For the target of car, the urban area and the runway area are divided by the proposed DCIFF-CNN, showing that the proposed DCIFF-CNN integrating contextual information through AMIM and ORSG is robust for complicated scene clutter.

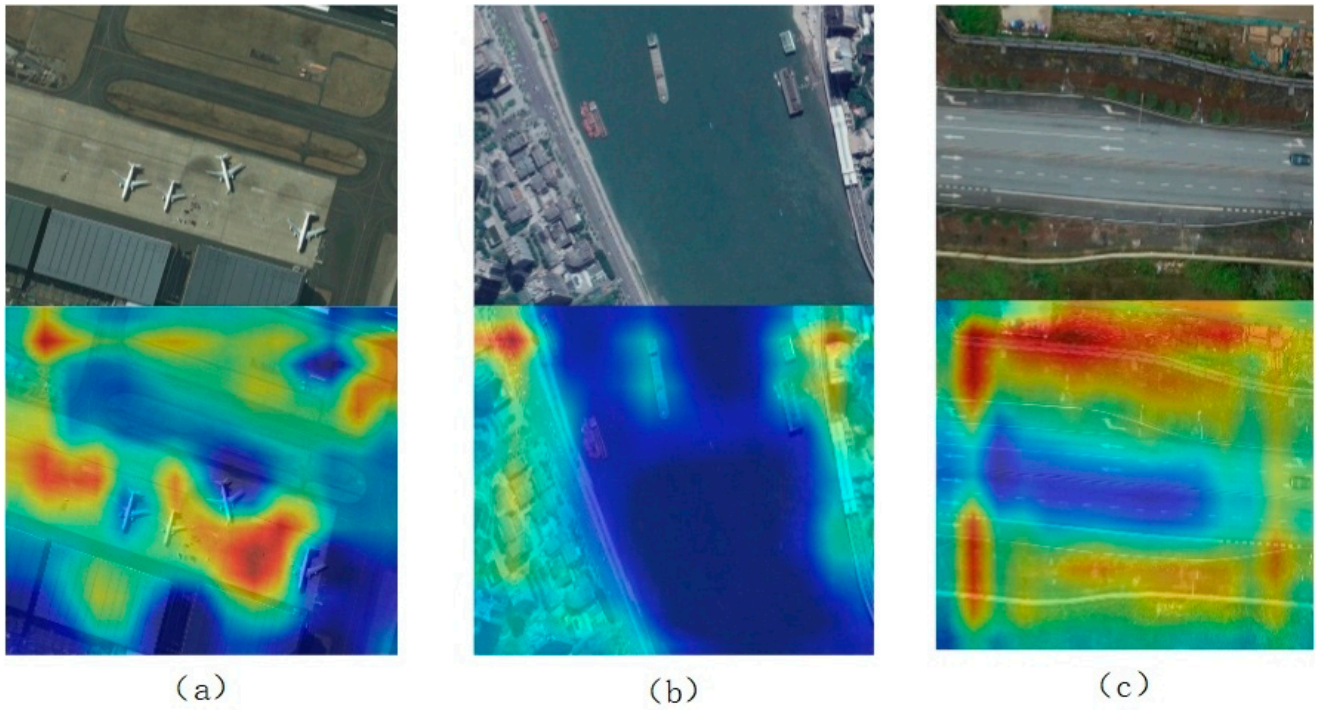

Figure 12. The heat maps of the collected dataset. (**a**) Airplane; (**b**) Ship; (**c**) Car.

Table 4 shows the AC value and PR value of the collected dataset. It can be seen from Table 4 that the proposed DCIFF-CNN obtains better detection results than other methods. For the target of airplane, the DCIFF-CNN achieves 2.85% better AC and 5.25% better PR than the YOLO4. Furthermore, the proposed method obtains a value of 90% in both AC and PR, demonstrating that the proposed DCIFF-CNN can maintain a low false alarm rate with a high recall rate. For the target of ship, there are two different backgrounds, and the number of targets is more than that of other targets. In this situation, the proposed DCIFF-CNN is superior to other methods, showing that the proposed ORSG and AMIM are robust to various complex backgrounds. For the target of car, the AC value obtained by YOLO4 is slightly higher than that of the proposed method. However, the proposed method achieves the highest PR value. This demonstrates that the proposed DCIFF-CNN is stable for detecting small and dense targets, due to the efficient use of context information by the added ORSG and AMIM.

Table 4. Performance comparisons of various methods on the collected dataset.

Target	Index	FRCNN-ZF	FRCNN-VGG	YOLO4	DCIFF-CNN
Airplane	AC	76.93% (1154/1500)	85.00% (1275/1500)	87.20% (1308/1500)	90.05% (1358/1500)
	PR	76.88% (1154/1501)	85.40% (1275/1493)	87.32% (1308/1498)	92.57% (1358/1467)
Ship	AC	78.55% (1571/2000)	81.05% (1621/2000)	83.60% (1672/2000)	88.60% (1772/2000)
	PR	76.90% (1571/2043)	84.25% (1621/1924)	84.32% (1672/1983)	91.10% (1772/1945)
Car	AC	76.13% (1142/1500)	83.07% (1246/1500)	89.40% (1341/1500)	89.07% (1336/1500)
	PR	70.06% (1142/1630)	83.01% (1246/1501)	89.88% (1341/1492)	92.27% (1336/1488)

Figure 13 shows the AP value of different methods. The results in Figure 13 indicate that the proposed DCIFF-CNN achieves better detection performance than other methods, owing to the explicit modeling of the target context. For the target of airplane, the YOLO-based methods obtain better AP value than the FRCNN-based method, reflecting that the YOLO-based method is more suitable for detecting small targets. However, the proposed DCIFF-CNN obtains the highest AP value, which indicates that the proposed explicit context model can effectively utilize the information around the target to improve the target detection accuracy. For the target of ship, FRCNN-ZF obtains the lowest AP value, and YOLO4 and DCIFF-CNN obtain high AP values. This demonstrates that the detection performance is improved with the depth of the model. For the target of car, YOLO4 gets a slightly higher AP value than the proposed DCIFF-CNN since the background and context information for the target of ship is more complex than that of other targets. In particular, the context information of the ship outside the harbor may contain complex environments, such as cities. This leads to a slight decrease in the performance of our proposed method. However, compared with YOLO4, the proposed DCIFF-CNN still achieves better target detection results than the other two types of targets. This indicates that the proposed method can process a variety of contextual information in a balanced manner, and obtain relatively good detection accuracy in various environments. Overall, according to the AP value, the proposed DCIFF-CNN achieves better robustness and effectiveness in target detection than other methods.

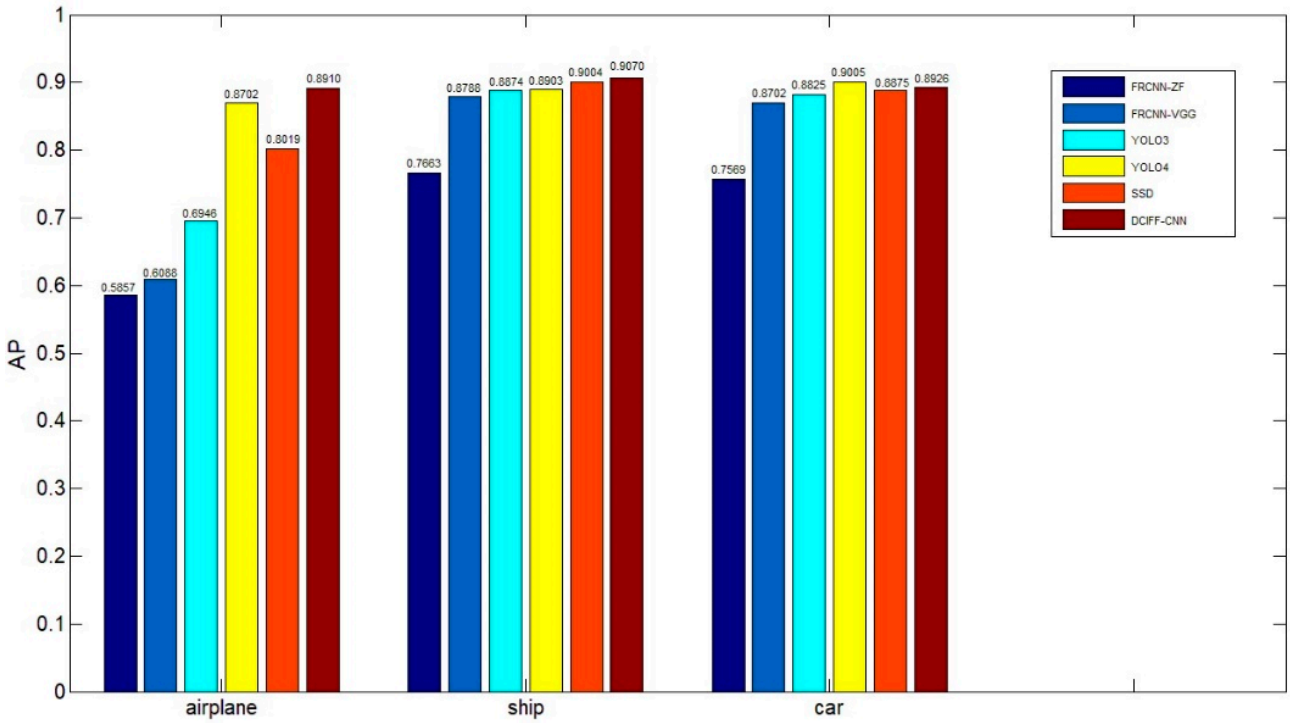

Figure 13. The average precision of target detection on the collected dataset.

4. Discussion

4.1. Analysis of Multi-Scale Feature Settings

In this paper, multiple scales of features are employed to explore the inside contextual information for region proposal. The NWPU-VHR dataset and AP (including mAP) value are applied in this section for evaluating the robustness of multi-scale feature setting. To demonstrate the effectiveness of these features, multi-scale features are used in the MLC-RPN for performance comparison. Generally, with the deepening of network layers, the feature map changes greatly, and conv4 is closer to conv5 than conv3. The distinction between the feature maps of conv3 and the feature maps of conv5 is even more obvious; it

will play a greater role in feature fusion. As shown in Table 5, the adding of layer conv3 and conv4 contributes to a 40.8% and 1.25% increase in mAP, respectively, confirming that the application of multi-scale features is beneficial to target detection. Furthermore, it can be seen from Table 5 that the increase of conv3 layer and conv4 layer with the multi-scale features brings no further performance improvement, but more complexity of the model is caused and more GPU memory is needed. When there are too many layers fused together and close to each other, the increase of distinguishing features will not be too much, and the auxiliary effect will be limited. Moreover, the amount of computation will increase with the increase of fusion features, which is not cost-effective for experiments. Therefore, we chose to add the conv3 according to the experiment results.

Table 5. Comparison of different multi-scale features utilized in the MLC-RPN. (The evaluation metric: AP; Dataset: NWPU-VHR (train:20%, val:20%, test:60%)).

Multi-Scale Settings	MLC-RPN (Conv5)	MLC-RPN (Conv4 + Conv5)	MLC-RPN (Conv3 + Conv5)	MLC-RPN (Conv3 + Conv4 + Conv5)
Airplane	0.9086	0.9083	0.9065	0.9063
Ship	0.8756	0.8754	0.8959	0.8954
Storage tank	0.8035	0.8010	0.8833	0.8866
Baseball diamond	0.9954	0.9959	0.9946	0.9946
Tennis court	0.9020	0.8992	0.9052	0.9051
Basketball court	0.8962	0.8970	0.8984	0.8984
Ground track field	0.9091	0.9972	0.9086	0.9086
Harbor	0.9047	0.9030	0.9083	0.9083
Bridge	0.8953	0.8994	0.9091	0.9091
Vehicle	0.8846	0.8791	0.8914	0.8893
mAP	0.8976	0.9056	0.9101	0.9102

4.2. Analysis of MLC-RPN

In this paper, the MLC-RPN is proposed to exploit multi-scale contextual information. The proposed MLC-RPN is analyzed by comparing it to the multi-scale network of SSD, which extracts feature maps of different scales for target detection. The large-scale feature map (front feature images) is extracted to detect small targets, while the small-scale feature map (back feature images) is used to detect large targets. Figure 14 shows the detection results of various targets. It can be seen that the proposed MLC-RPN is robust for targets in various scales, especially the target of bridge and tennis court. The size of bridge in the image varies with the length of the bridge, so the target of bridge has multiple scales in this dataset. The result in Figure 14 indicates that the proposed multi-scale contextual information fusion method achieves better performance when the scale of a certain target varies greatly. Meanwhile, the target of tennis court has a similar size to the basketball court, but its local characteristics are different from those of a basketball court. Therefore, it is necessary to fuse these multi-scale features to obtain more differentiated feature information. The proposed multi-scale contextual information fusion method shows advantages by integrating multiple scale information of the same target and providing more distinguishable information for similar targets. Overall, the proposed MCL-RPN achieves better robustness in detecting multi-class and multi-scale targets.

4.3. Analysis of Structured Object-Level Relational Reasoning

The ORSG is employed in this paper to explore the positive contextual information from the object-level relationship. Additionally, the AMIM is added in ORSG for integrating the contextual information. To demonstrate the robustness and effectiveness of the ORC-TDN, especially the ORSG and AMIM, another two pooling methods of generating contextual information are adopted for comparison. The NWPU-VHR dataset is applied in this section for evaluating the robustness of ORSG and AMIM. As shown in Table 6, the network with ORSG and AMIM achieves 4.19% better mAP than that without ORSG

and AMIM. Compared with AMIM, the performance of integrating contextual information by max-pooling and average pooling is reduced by 1.92% and 1.22%, respectively. These results validate the superiority of the proposed AMIM in exploiting contextual information. Furthermore, the ORC-TDN with ORSG and AMIM achieves at least 1.22% better performance in mAP compared with the network without ORSG or AMIM, and the ORC-TDN obtains the best detection results in most categories. This demonstrates the robustness and effectiveness of the proposed ORC-TDN. The effectiveness of the attentional integration strategy and object-level reasoning are demonstrated by the gains of AP score, as shown in Figure 15. The results in Figure 15 indicate that the proposed method achieves balanced results in multi-class target detection.

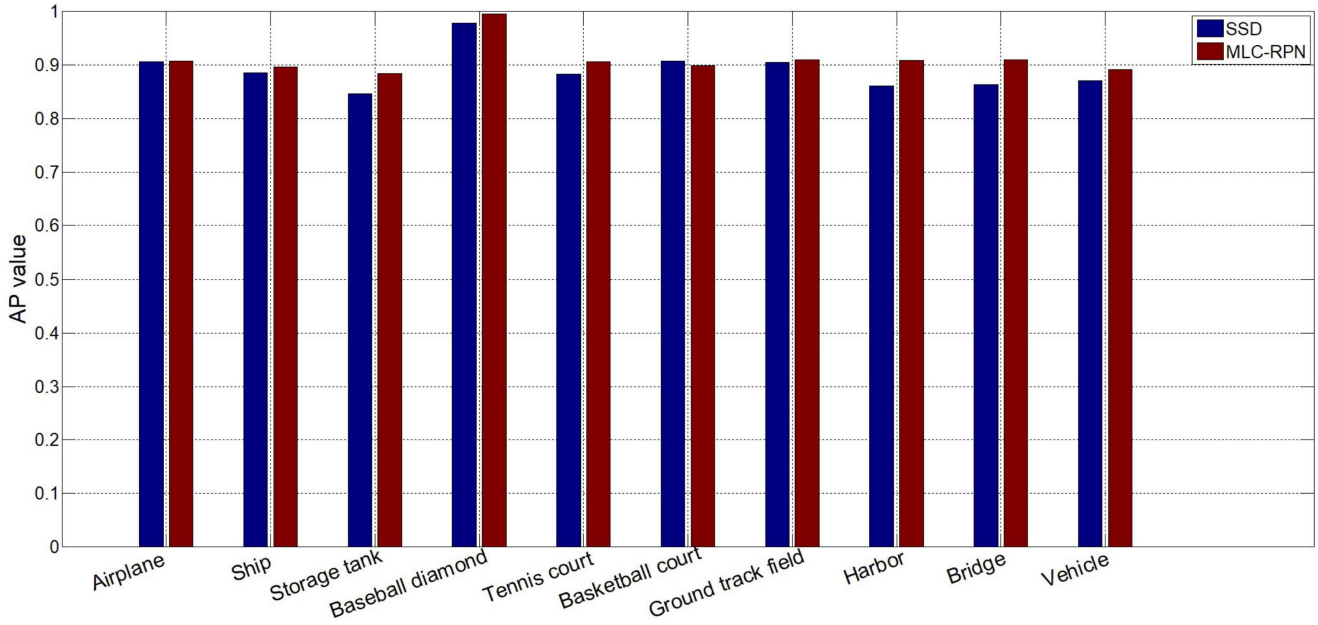

Figure 14. The detection results of ten categories of targets by MLC-PRN and SSD. (The evaluation metric: AP; Dataset: NWPU-VHR (train: 20%, val: 20%, test: 60%)).

Table 6. mAP value of context fusion network using different methods. (Dataset: NWPU-VHR (train:20%, val:20%, test:60%)).

Method	Without ORSG and AMIM	With ORSG and Max-Pooling	With ORSG and Average Pooling	With ORSG and AMIM
mAP	0.8682	0.8909	0.8979	0.9101

Some feature maps of the conv5 layers from the three networks are displayed in Figure 16. Comparing the results shown in Figure 16b–e, it can be seen that the object-level contextual information fusion enhances the response of the target-like regions in the feature map, demonstrating that the proposed ORSG is conducive to enhancing the feature recognition capability of the network. For all targets, the features in Figure 16c–e are brighter than those in Figure 16b. This indicates that the network without ORSG and AMIM fails to distinguish the foreground features from the background features. Moreover, the features of the target area in Figure 16e are more highlighted that that in Figure 16b–d whereas the features of background area are weakened. This indicates that the AMIM significantly improves the features representation. This is because the environmental information beneficial to the target is effectively exploited by the AMIM and ORSG, which further enhances the regional feature of the target.

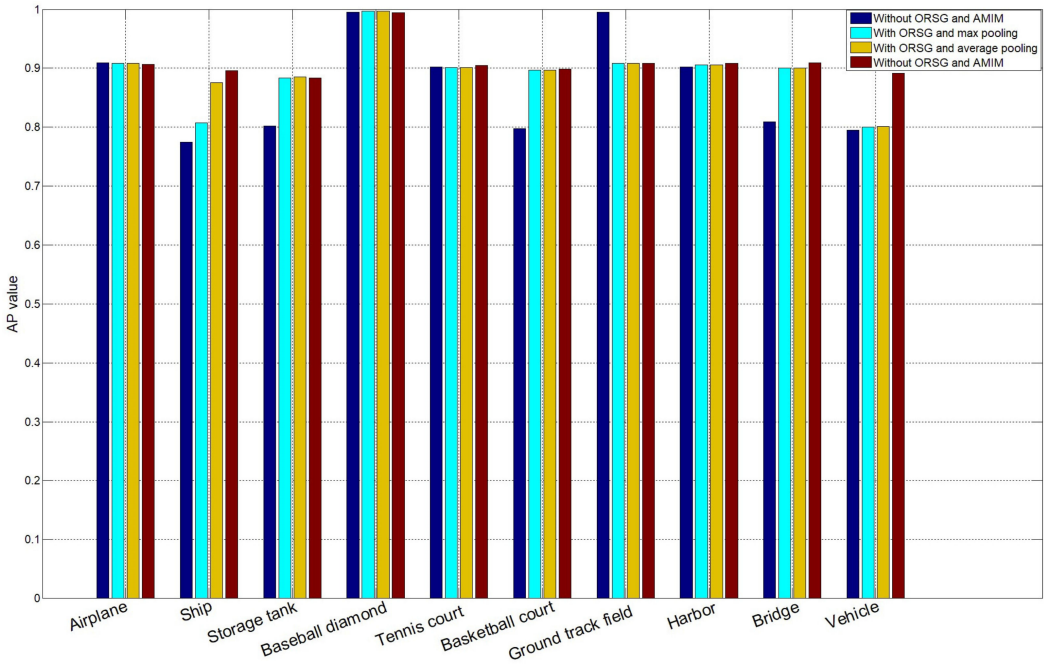

Figure 15. The AP value for different targets under various context fusion networks. (The evaluation metric: AP; Dataset: NWPU-VHR (train: 20%, val: 20%, test: 60%)).

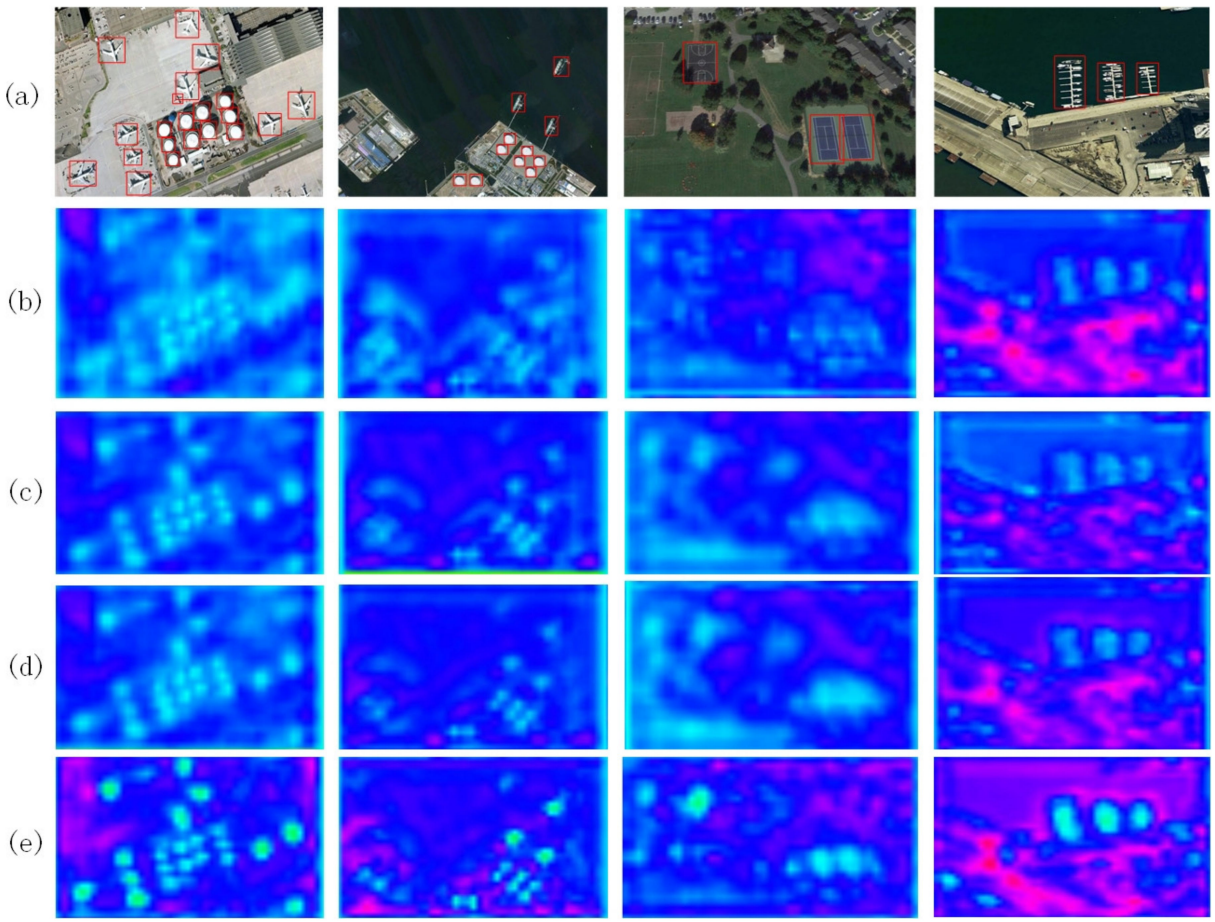

Figure 16. Feature maps of each conv5 in three networks. (**a**) Input images. (**b**) Without ORSG and AMIM. (**c**) With ORSG and max-pooling. (**d**) With ORSG and average pooling. (**e**) With ORSG and AMIM.

5. Conclusions

An end-to-end DCIFF-CNN target detection method for remote sensing images is proposed in this paper. The method exploits the inside context information and object-level relationship through MLC-RPN and ORC-TDN. The MLC-RPN fuses the multi-scale convolution layer features to capture the candidate regions. In ORC-TDN, the ORSG module models the explicit relationship between a set of objects by processing their appearance feature and geometric feature, and the AMIM module integrates the message passed from the object relational graph by designing a multi-dimensional attention model. A large number of experiments were carried out on the NWPU-VHR dataset and the collected dataset. The experimental results show that the proposed DCIFF-CNN method achieves better target detection performance than the state-of-the-art methods. Moreover, the proposed method can maintain stable performance for targets with varied scales, varied scenes, and varied quantities.

Author Contributions: B.C. and Z.L. conceived of the study; B.C. wrote the code, performed the analysis and wrote the article; Z.L. analyzed the results; B.X. and X.Y. collected the dataset; Z.D. and T.Q. revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the National Natural Science Foundation of China under Grant No. 61675036, 13th Five-year Plan Equipment Pre-research Fund under Grant No. 6140415020312, and Chinese Academy of Sciences Key Laboratory of Beam Control Fund under Grant No. 2017LBC006.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: http://pan.baidu.com/s/1hqwzXeG.

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

Abbreviations

CNN	convolutional neural network
DCIFF-CNN	diversified context information fusion framework based on convolutional neural network
MLC-RPN	multi-scale local context region proposal network
ORC-TDN	object-level relationships context target detection network
AMIM	attentional message integrated module
ORSG	object relational structured graph
TP	true positives
FP	false positives
FN	false negatives
AP	average precision
mAP	mean average precision
AC	accuracy ratio

References

1. Hu, Y.; Li, X.; Zhou, N.; Yang, L.; Peng, L.; Xiao, S. A Sample Update-Based Convolutional Neural Network Framework for Object Detection in Large-Area Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 947–951. [CrossRef]
2. 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]
3. Zhu, M.; Xu, Y.; Ma, S.; Li, S.; Ma, H.; Han, Y. Effective Airplane Detection in Remote Sensing Images Based on Multilayer Feature Fusion and Improved Nonmaximal Suppression Algorithm. *Remote Sens.* **2019**, *11*, 1062. [CrossRef]
4. Cheng, H.Y.; Weng, C.C.; Chen, Y.Y. Vehicle detection in aerial surveillance using dynamic Bayesian networks. *IEEE Trans. Image Process.* **2012**, *21*, 2152–2159. [CrossRef]

5. Stankov, K.; He, D.-C. Detection of Buildings in Multispectral Very High Spatial Resolution Images Using the Percentage Occupancy Hit-or-Miss Transform. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 4069–4080. [CrossRef]
6. Georganos, S.; Grippa, T.; Vanhuysse, S.; Lennert, M.; Shimoni, M.; Wolff, E. Very High Resolution Object-Based Land Use–Land Cover Urban Classification Using Extreme Gradient Boosting. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 607–611. [CrossRef]
7. Zhang, T.; Huang, X. Monitoring of Urban Impervious Surfaces Using Time Series of High-Resolution Remote Sensing Images in Rapidly Urbanized Areas: A Case Study of Shenzhen. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 2692–2708. [CrossRef]
8. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
9. Chaudhuri, D.; Kushwaha, N.K.; Samal, A. Semi-Automated Road Detection From High Resolution Satellite Images by Directional Morphological Enhancement and Segmentation Techniques. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2012**, *5*, 1538–1544. [CrossRef]
10. Stankov, K.; He, D.-C. Building Detection in Very High Spatial Resolution Multispectral Images Using the Hit-or-Miss Transform. *IEEE Geosci. Remote Sens. Lett.* **2013**, *10*, 86–90. [CrossRef]
11. Blaschke, T.; Hay, G.J.; Kelly, M.; Lang, S.; Hofmann, P.; Addink, E.; Queiroz Feitosa, R.; van der Meer, F.; van der Werff, H.; van Coillie, F.; et al. Geographic Object-Based Image Analysis—Towards a new paradigm. *ISPRS J. Photogramm. Remote Sens.* **2014**, *87*, 180–191. [CrossRef] [PubMed]
12. Li, X.; Cheng, X.; Chen, W.; Chen, G.; Liu, S. Identification of Forested Landslides Using LiDar Data, Object-based Image Analysis, and Machine Learning Algorithms. *Remote Sens.* **2015**, *7*, 9705–9726. [CrossRef]
13. Leninisha, S.; Vani, K. Water flow based geometric active deformable model for road network. *ISPRS J. Photogramm. Remote Sens.* **2015**, *102*, 140–147. [CrossRef]
14. Ok, A.O.; Senaras, C.; Yuksel, B. Automated Detection of Arbitrarily Shaped Buildings in Complex Environments From Monocular VHR Optical Satellite Imagery. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 1701–1717. [CrossRef]
15. Ok, A.O. Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts. *ISPRS J. Photogramm. Remote Sens.* **2013**, *86*, 21–40. [CrossRef]
16. Cheng, G.; Guo, L.; Zhao, T.; Han, J.; Li, H.; Fang, J. Automatic landslide detection from remote-sensing imagery using a scene classification method based on BoVW and pLSA. *Int. J. Remote Sens.* **2012**, *34*, 45–59. [CrossRef]
17. Cheng, G.; Han, J.; Guo, L.; Liu, Z.; Bu, S.; Ren, J. Effective and Efficient Midlevel Visual Elements-Oriented Land-Use Classification Using VHR Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 4238–4249. [CrossRef]
18. 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 2014 IEEE Geoscience and Remote Sensing Symposium, Quebec City, QC, Canada, 13–18 July 2014.
19. Yu, L.; Xian, S.; Hongqi, W.; Hao, S.; Xiangjuan, L. Automatic Target Detection in High-Resolution Remote Sensing Images Using a Contour-Based Spatial Model. *IEEE Geosci. Remote Sens. Lett.* **2012**, *9*, 886–890.
20. Bai, X.; Zhang, X.; Zhou, J. VHR Object Detection Based on Structural Feature Extraction and Query Expansion. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 6508–6520.
21. Dingwen, Z.; Junwei, H.; Gong, C.; Zhenbao, L.; Shuhui, B.; Lei, G. Weakly Supervised Learning for Target Detection in Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 701–705. [CrossRef]
22. 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]
23. Tao, C.; Tan, Y.; Cai, H.; Tian, J. Airport Detection from Large IKONOS Images Using Clustered SIFT Keypoints and Region Information. *IEEE Geosci. Remote Sens. Lett.* **2011**, *8*, 128–132. [CrossRef]
24. 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]
25. Aytekin, Ö.; Zongur, U.; Halici, U. Texture-Based Airport Runway Detection. *IEEE Geosci. Remote Sens. Lett.* **2013**, *10*, 471–475. [CrossRef]
26. De Morsier, F.; Tuia, D.; Borgeaud, M.; Gass, V.; Thiran, J.-P. Semi-Supervised Novelty Detection Using SVM Entire Solution Path. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 1939–1950. [CrossRef]
27. Das, S.; Mirnalinee, T.T.; Varghese, K. Use of Salient Features for the Design of a Multistage Framework to Extract Roads From High-Resolution Multispectral Satellite Images. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 3906–3931. [CrossRef]
28. Fukun, B.; Bocheng, Z.; Lining, G.; Mingming, B. A Visual Search Inspired Computational Model for Ship Detection in Optical Satellite Images. *IEEE Geosci. Remote Sens. Lett.* **2012**, *9*, 749–753. [CrossRef]
29. Ma, L.; Crawford, M.M.; Tian, J. Local Manifold Learning-Based *k*-Nearest-Neighbor for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 4099–4109. [CrossRef]
30. Blanzieri, E.; Melgani, F. Nearest Neighbor Classification of Remote Sensing Images with the Maximal Margin Principle. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 1804–1811. [CrossRef]
31. 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]
32. Zhenwei, S.; Xinran, Y.; Zhiguo, J.; Bo, L. Ship Detection in High-Resolution Optical Imagery Based on Anomaly Detector and Local Shape Feature. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 4511–4523. [CrossRef]

33. Leitloff, J.; Hinz, S.; Stilla, U. Vehicle Detection in Very High Resolution Satellite Images of City Areas. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 2795–2806. [CrossRef]
34. Wegner, J.D.; Hansch, R.; Thiele, A.; Soergel, U. Building Detection from One Orthophoto and High-Resolution InSAR Data Using Conditional Random Fields. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2011**, *4*, 83–91. [CrossRef]
35. Li, E.; Femiani, J.; Xu, S.; Zhang, X.; Wonka, P. Robust Rooftop Extraction from Visible Band Images Using Higher Order CRF. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 4483–4495. [CrossRef]
36. Lei, Z.; Fang, T.; Huo, H.; Li, D. Bi-Temporal Texton Forest for Land Cover Transition Detection on Remotely Sensed Imagery. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 1227–1237. [CrossRef]
37. Benedek, C.; Shadaydeh, M.; Kato, Z.; Szirányi, T.; Zerubia, J. Multilayer Markov Random Field models for change detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2015**, *107*, 22–37. [CrossRef]
38. Dong, Y.; Du, B.; Zhang, L. Target Detection Based on Random Forest Metric Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 1830–1838. [CrossRef]
39. 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.
40. Girshick, R. Fast R-CNN. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
41. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN-Towards Real-Time Object Detection with Region Proposal Networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2015; pp. 91–99.
42. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. In *Computer Vision and Pattern Recognition*; IEEE Press: Salt Lake City, UT, USA, 2018.
43. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot Multi Box Detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016.
44. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN:Object Detection via Region-based Fully Convolutional Networks. In *Computer Vision and Pattern Recognition*; IEEE Press: Las Vegas, Nevada, USA, 2016.
45. Wang, C.; Bai, X.; Wang, S.; Zhou, J.; Ren, P. Multiscale Visual Attention Networks for Object Detection in VHR Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 310–314. [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. Wang, P.; Sun, X.; Diao, W.; Fu, K. FMSSD: Feature-Merged Single-Shot Detection for Multiscale Objects in Large-Scale Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 3377–3390. [CrossRef]
48. Zheng, Z.; Zhong, Y.; Ma, A.; Han, X.; Zhao, J.; Liu, Y.; Zhang, L. HyNet: Hyper-scale object detection network framework for multiple spatial resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2020**, *166*, 1–14. [CrossRef]
49. 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]
50. 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] [PubMed]
51. Choi, M.J.; Lim, J.J.; Torralba, A.; Willsky, A.S. Exploiting Hierarchical Context on a Large Database of Object Categories. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; Volume 1, pp. 129–136.
52. Liu, Y.; Wang, R.; Shan, S.; Chen, X. Structure Inference Net: Object Detection Using Scene-Level Context and Instance-Level Relationships. In *Conference on Computer Vision and Pattern Recognition*; IEEE Press: Salt Lake City, UT, USA, 2018; pp. 6985–6994.
53. Bell, S.; Zitnick, C.L.; Bala, K.; Girshick, R. Inside-Outside Net-Detecting Objectsin Contextwith Skip Poolingand Recurrent Neural Networks. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
54. Galleguillos, C.; Belongie, S. Context based object categorization: A critical survey. *Comput. Vis. Image Underst.* **2010**, *114*, 712–722. [CrossRef]
55. Shrivastava, A.; Gupta, A. Contextual Priming and Feedback for Faster R-CNN. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016.
56. Mottaghi, R.; Chen, X.; Liu, X.; Cho, N.-G.; Lee, S.-W.; Fidler, S.; Urtasun, R.; Yuille, A. The Role of Context for Object Detection and Semantic Segmentation in the Wild. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 891–898.
57. Zeng, X.; Ouyang, W.; Yang, B.; Yan, J.; Wang, X. Gated Bi-directional CNN for Object Detection. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016.
58. 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]
59. Song, X.; Jiang, S.; Wang, B.; Chen, C.; Chena, G. Image Representations with Spatial Object-to-Object Relations for RGB-D Scene Recognition. *IEEE Trans. Image Process.* **2019**, *29*, 525–537. [CrossRef]

60. Cho, K.; van Merrienboer, B.; Bahdanau, D. On the Properties of Neural Machine Translation: Encoder–Decoder approaches. In Proceedings of the Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8), Doha, Qatar, 25 October 2014.
61. Xu, D.; Zhu, Y.; Chozy, C.B.; Fei-Fei, L. Scene Graph Generation by Iterative Message Passing. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
62. Deng, Z.; Vahdat, A.; Hu, H.; Mori, G. Structure Inference Machines: Recurrent Neural Networks for Analyzing Relations in Group Activity Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4772–4781.
63. Hu, H.; Zhou, G.-T.; Deng, Z.; Liao, Z.; Mori, G. Learning Structured Inference Neural Networks with Label Relations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA,, 27–30 June 2016.
64. Seo, Y.; Defferrard, M.E.; Vandergheynst, P.; Bresson, X. Structured Sequence Modeling with Graph Convolutional Recurrent Networks. *Proceedings of the International Conference on Learning Representations*, Toulon, France, 24–26 April 2017.
65. Marino, K.; Salakhutdinov, R.; Gupta, A. The More You Know: Using Knowledge Graphs for Image Classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 20–28.
66. Fan, Q.; Zhuo, W.; Tang, C.-K.; Tai, Y.-W. Few-Shot Object Detection with Attention-RPN and Multi-Relation Detector. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020.
67. Zhu, Y.; Zhao, C.; Guo, H.; Wang, J.; Zhao, X.; Lu, H. Attention CoupleNet: Fully Convolutional Attention Coupling Network for Object Detection. *IEEE Trans. Image Process.* **2019**, *28*, 113–126. [CrossRef]
68. Song, K.; Yang, H.; Yin, Z. Multi-Scale Attention Deep Neural Network for Fast Accurate Object Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *29*, 2972–2985. [CrossRef]
69. Sanghyun, W.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018.
70. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Computer Vision and Pattern Recognition. In Proceedings of the Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
71. 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]
72. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
73. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-J.M. YOLOv4 Optimal Speed and Accuracy of Object Detection. In Proceedings of the Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
74. Cai, Z.; Fan, Q.; Feris, R.S.; Vasconcelos, N. A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016.