Article

CISPNet: Automatic Detection of Remote Sensing Images from Google Earth in Complex Scenes Based on Context Information Scene Perception

Wenxu Shi ^{1,2}, Jinhong Jiang ^{1,2}, Shengli Bao ^{1,2,*} and Dailun Tan ^{3,*}

¹ Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu 610081, China; shiwenxu17@mails.ucas.ac.cn (W.S.); jiangjinhong17@mails.ucas.ac.cn (J.J.)

² School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 10049, China

³ School of Mathematics and Information, China West Normal University, Nanchong 637009, China

* Correspondence: baohigh@casit.com.cn (S.B.); dailun_tan@cwnu.edu.cn (D.T.)

Received: 25 September 2019; Accepted: 7 November 2019; Published: 12 November 2019

Abstract: The ability to detect small targets and the speed of the target detector are very important for the application of remote sensing image detection, and in this paper, we propose an effective and efficient method (named CISPNet) with high detection accuracy and compact architecture. In particular, according to the characteristics of the data, we apply a context information scene perception (CISP) module to obtain the contextual information for targets of different scales and use k-means clustering to set the aspect ratios and size of the default boxes. The proposed method inherits the network structure of Single Shot MultiBox Detector (SSD) and introduces the CISP module into it. We create a dataset in the Pascal Visual Object Classes (VOC) format, annotated with the three types of detection targets, aircraft, ship, and oil tanker. Experimental results on our remote sensing image dataset as well as the Northwestern Polytechnical University very-high-resolution (NWPU VRH-10) dataset demonstrate that the proposed CISPNet performs much better than the original SSD and other detectors especially for small objects. Specifically, our network can achieve 80.34% mean average precision (mAP) at the speed of 50.7 frames per second (FPS) with the input size 300×300 pixels on the remote sensing image dataset. On extended experiments, the performance of CISPNet in fuzzy target detection in remote sensing image is better than that of SSD.

Keywords: remote sensing images; deep learning; deep convolutional neural network; real-time target detection; context feature scene perception

1. Introduction

With the rapid development of remote sensing spaceborne technologies, such as Sentinel-1, TerraSAR-X, and RADARSAT-2 [1–3], target detection on remote sensing images has been playing an important position in the field of civil areas and defense security [4–6]. However, target detection in remote sensing images remains a great challenge due to the complex background and existence of speckle noises in remote sensing images. Therefore, it is worthwhile to develop a detector with strong feature extraction capabilities to obtain better target detection performance of remote sensing images.

Over the past decade, remote sensing images have provided abundant shape structure and texture information of landscape targets, and the 2-dimensional target detection algorithms have been widely studied in remote sensing images. Cheng et al. [7], developed a discriminatively trained mixture model for extracting feature pyramids from multi-scale layers using a histogram of oriented gradient (HOG) [8], and then threshold operation was performed on the response of the model to judge the

presence of the target in the remote sensing image. Bai et al. [9] used the ranking support vector machine (SVM) [10] to identify the existence of the target in the remote sensing image. In addition, the methods [11–13] extracted semantic information from texture and shape features, and used the machine learning algorithms such as the contrast box algorithm [14] and semi-supervised hierarchical classification [15] to get the final detection results. Though these conventional methods are effective in specific scenes, the methods are based on hand-crafted features with poor generalization abilities, which makes it difficult and time-consuming to detect targets on large complex remote sensing image data sets.

Recently, the convolutional neural network (CNN) based methods have achieved encouraging results for general target detection and classification problems [16–19]. In particular, target detection using CNN has achieved remarkable successes, which can be classified into two types: The first is the two-stage detector, such as Region-CNN [20], Fast R-CNN [21], Faster R-CNN [22], and feature pyramid networks (FPNs) [23], and the other is the single-stage detector, such as the you only look once (YOLO) series [24–26], Single Shot MultiBox Detector (SSD) [27], deconvolutional single shot detector (DSSD) [28]. The former type relies on a series of candidate region suggestions as samples, and then classifies the samples based on a CNN. The latter type directly estimates the target region without generating candidate regional proposals, and directly transforms the target region positioning problem into a regression processing problem.

The above-mentioned visual detection algorithms have also been widely used in remote sensing image target detection. Zhu et al. [29] proposed a new method of airport target detection based on a convolution neural network, combined with a cascaded area recommendation network. Based on the SSD [21] detection framework, Chen et al. [30] presented a detection method for airplane detection that attempted to improve the detection accuracy of multi-scale remote sensing image targets. Yang et al. [31] proposed a multi-scale remote sensing image target detection method based on the feature pyramid networks (FPN) [23] detection framework. Although target detection in remote sensing images may be relatively accurate using the CNN, there are usually no effective approaches to utilize contextual information fully, which makes it difficult to understand the complex scenarios in remote sensing images, which may result in inaccurate detection of the small targets and cause the problems named as “box-in-box” [32] as shown in Figure 1a. In the figure, we can see that the SSD detects a single target with two overlapping boxes. The smaller box has partial image such as the part of the ship. In order to solve the same problem, Wang et al. [33] has taken the objects proposals generated by SSD and replaces the original visual geometry group 16 (VGG16) [18] with a densely connected network [34] as the backbone network to improve the relevance of contextual information. However, this method increases the computational cost of the network, and the detection accuracy of the method for small-scale targets is poor.

Figure 1. Conventional single shot multibox detector (SSD) vs. The proposed context information scene perception (CISP)Net. (a) SSD detects ship targets with overlapping boxes and the small aircraft targets that cannot be located. (b) Ship and small aircraft targets that can be located by the proposed CISPNet.

To address the above problem, especially for small target detection, in this paper, we present a single-stage detector named context information scene perception (CISP)Net for target detection on remote sensing images, which is based on the Single Shot MultiBox Detector (SSD) [27] and apply a context information scene perception (CISP) module to obtain the context information for targets of different scales. Compared with other detection methods such as YOLO [24], SSD [27], DSSD [28], and RSSD [32], our framework is more suitable for target detection in remote sensing images, and has achieved the relatively advanced performance. The contributions of our work can be summarized as follows:

- (1) Different from previous detection models, we built a remote sensing image target detection framework based on SSD that can improve the relevance of contextual information and handle different complex scenes.
- (2) To improve the relevance of contextual information, we propose a context information scene perception (CISP) module to obtain the context information for targets of different scales.
- (3) We show that CISPNet can implement relatively advanced and reliable performance on remote sensing image dataset and the NWPU VRH-10 dataset and verify the effectiveness of the optimization for its architecture.

The rest of this paper is organized as follows. Section 2 introduces the basis of the proposed method. Section 3 introduces the details of the proposed CISPNet framework. Section 4 presents experiments conducted on remote sensing dataset and the NWPU VRH-10 dataset to validate the effectiveness of the proposed framework and discusses the results of the proposed method. Finally, Section 5 concludes this paper.

2. Architecture of SSD

Here, we briefly review the most widely used detector, SSD [27], which uses a single deep neural network to detect targets in images, which is the basis of the proposed method CISPNet.

As shown in Figure 2, SSD is based on the VGG16 [18] and an auxiliary feature pyramid structure is added to the end of the base network. Note that the SSD makes use of multi-scale feature maps generated from the network to detect objects in various sizes, specifically, the layer Conv4_3 and FC7 are used to detect targets of the smallest size and the deeper layers are adopted for detecting targets of bigger sizes. Although SSD can alleviate the problems arising from object scale variation, it has limitations for detecting small targets. The main reason is the shallower layers lack the semantic information and each layer in the network is used independently as an input to the classifier network, thus they cannot reflect appropriate contextual information from different scales, which makes it difficult to understand complex scenarios in images. Hence, in this work, we attempt to improve SSD by improving the relevance of contextual information.

Figure 2. Architecture of the Single Shot MultiBox Detector with input size 300×300 .

3. Proposed Method

In this section we will detail the architecture of the proposed CISPNet framework. As shown in Figure 3a, the CISPNet assembles four context information scene perception (CISP) modules and two feature fusion modules (FFM) into a conventional SSD. The structure of these additional modules is simple and can be easily combined with conventional detection networks.

Figure 3. The architecture of CISPNet and its novel module CISP. (a) The architecture of the CISPNet framework; (b) The layer settings of the CISP module, where each box represents a conv+bn+relu group.

In the context information scene perception (CISP) module, as shown in Figure 3b, multiple dilated rate convolution layers are used in parallel. Each dilated convolution layer has a different dilated rate, and the size of the dilated rate reflects its corresponding receptive field size. The context information scene perception module uses convolution kernels with different receptive fields to extract features from Conv4_3 and FC7, so that the model can perceive the changes of context information in different scales and sub-domains. In this way, the loss of semantic information can be reduced and make the feature map understand more contextual information from different scales. The inner structure of context information scene perception module is shown in Figure 3b. Firstly, the number of channels of the feature map $U \in \mathbb{R}^{W \times H \times 512}$ is reduced by using a 1×1 convolution to obtain a feature map $X \in \mathbb{R}^{W \times H \times 256}$. Then, three kinds of dilated convolutions with different dilated rates $(d_1, d_2, d_3) = (1, 2, 4)$ are used for feature sampling on the feature map X in parallel, and the feature map $X_1 \in \mathbb{R}^{W \times H \times 256}$, $X_2 \in \mathbb{R}^{W \times H \times 256}$, and $X_3 \in \mathbb{R}^{W \times H \times 256}$ are obtained. Finally, the feature map is concatenated to obtain the final feature map $N \in \mathbb{R}^{W \times H \times 1024}$, $N = [X, X_1, X_2, X_3]$.

Based on the CISP module, we propose a single-stage detector CISPNet, which can detect small targets more effectively. More specifically, we first assemble two CISP modules between features of the

layers Conv4_3 and FC7 and the layers FC7 and Conv8_2 respectively. In addition, we connect another two separate CISP modules to Conv4_3 and FC7 detection branches respectively, and then generate the new layers Conv4_3 and FC7, respectively. As the layers Conv4_3 and FC7 in the backbone are relatively shallow, and the feature semantic information extraction by a shallow network is less and might not have enough capability to detect small targets, we use CISP modules to enhance Conv4_3 and FC7 features. Step forward, feature fusion strategies always contribute to learning better features from the combination of original features [28,32]. We also applied this method in CISPNet. In detail, with the help of two FFMs, the new Conv4_3 and FC7 are generated by feature fusion.

In order to more clearly show the influence of the proposed CISP module on the image feature extraction, we select a remote sensing image containing a large ship target and three small ship targets in a wide sea area, as the input image and extract the features through SSD and the proposed CISPNet. Figure 4 shows a qualitative comparison of the feature mapping results from our proposed CISPNet and the SSD. With the increase of resolution, the feature map becomes smaller and more abstract, and small-sized objects hardly have a response on the deeper layers. As shown in Figure 4a₁,a₂,b₁,b₂, the semantic information of a ship from the feature map of CISPNet's Conv4_3 and FC7 layers is clearer than that extracted by the Conv4_3 and FC7 layers of SSD, in particular, the features extracted by CISPNet can highlight the smaller ship targets. At the same time, the above comparison also shows that the CISP module could reduce the loss of semantic information in the process of feature extraction, and make the feature map understand more contextual information from different scales. Figure 4a₃,a₄ have more highlighted feature information and stronger semantic information than those in Figure 4b₃,b₄. In theory, the prediction results on Figure 4a₁,a₂ are better than those on Figure 4b₁,b₂, especially for the detection of small targets.

Figure 4. CISPNet and SSD illustrate the image and its corresponding multi-scale feature map; (a₁)–(a₆) are the characteristic maps of the FaSSD output at the multi-scale feature layer; (b₁)–(b₆) are the characteristic maps of the SSD output at the multi-scale feature layer.

The CISP module could also be placed in other positions, and more CISP modules will bring more context information to the conventional network. Considering the trade-off between the improvement of precision and the increase of inference time, we have experimented and finally select the version shown in Figure 3a.

4. Experiments and Results

In this section, we first introduce the construction of the dataset for target detection in a remote sensing image, and then illustrate the evaluation metrics, training strategies, and implementation details. Then, we compare the proposed CISPNet with state-of-the-art methods to demonstrate its advantage. Finally, some ablation studies are discussed to verify the role of each component,

in addition, the expanded experiment shows that the proposed CISPNet performance over SSD in fuzzy target detection.

4.1. Benchmark Dataset

Similar to the related works [29–31], the remote sensing images used in this paper were collected from Google Earth with a resolution of 500×375 pixels and a spatial resolution of 0.5 m to 2 m. Compare with the targets in nature scene images, like ImageNet [35], MS COCO [36], Pascal VOC [37], the targets in remote sensing images (such as aircraft, ship, and oiltanker) usually have complex backgrounds. In addition, the remote sensing images used in this paper are different from the existing remote sensing images databases, such as the NWPU VHR-10 dataset [38], and (Aerial Image Dataset)AID dataset [39], which include more small targets (the area for small targets: $\text{area} < 32^2$ pixels) and medium targets (the area for medium targets: $32^2 < \text{area} < 96^2$ pixels), which further increases the difficulty of target detection in remote sensing images.

We collected 1500 remote sensing images to construct a dataset for tiny target detection, each image is required to be marked with detection targets: aircraft, ship, oiltanker, and the manually pixel-wise annotated in Pascal VOC format. Each detection object is marked with a ground truth box as shown in Figure 5.

Figure 5. Sample images from the constructed remote sensing image dataset. The detection targets are annotated with ground truth boxes.

Table 1 gives more details about this data set and Figure 6 shows the aspect ratio and area size distribution. It is obvious that the area of bounding boxes is generally small in Figure 6, which means that the architecture of our model should pay more attention to small detection targets and the set size of the default box can be further reduced. It is noteworthy that the aspect ratio is not symmetrically distributed around 1 and its distribution is also relatively concentrated as shown in Figure 6.

Table 1. The remote sensing image dataset.

Dataset	Class	Images	Total Instances	Target		
				Small	Medium	Large
Train	aircraft	374	4281	3191	1090	0
	oiltanker	320	2725	1225	1500	0
	ship	526	1418	1100	315	3
Test	aircraft	87	918	698	220	0
	oiltanker	76	886	401	485	0
	ship	137	408	300	108	0

Figure 6. The histograms of the distribution of the area size and aspect ratio of the bounding boxes.

4.2. Evaluation Metrics

The mean average precision (mAP) and the precision-recall curves (PRC) commonly used in the field of target detection are used to compare the detection performance of different methods. The precision-recall curves and the mean average precision are described in detail below.

The precision indicator can be seen as a measure of exactness or fidelity, and the recall indicator is a measure of completeness. The precision and recall indicators are formulated as follows:

$$precision = \frac{TP}{(TP + FP)}, recall = \frac{TP}{(TP + FN)} \quad (1)$$

where, TP, FP, and FN represent the number of true positives, false positives, and false negatives respectively. The precision-recall curve commonly takes the recall as the transverse coordinates and the precision as the vertical coordinates. If the detection method can keep a high precision with the increasing of recall, it means that the detection method has good performance.

The average precision (AP) calculates the average value of the precision over the interval from recall = 0 to recall = 1, which can be formulated as:

$$AP = \int_0^1 p(r) dx \quad (2)$$

where p is the value of precision and r denotes the value of recall. Hence, the AP is equal to the area under the curve, the higher the AP, the better the performance of the detection method, and the mean average precision (mAP) represents the average value of all category AP.

4.3. Training Strategies and Implementation Details

We implemented the proposed CISPNet with PyTorch v0.4.1 on a PC with an NVIDIA GeForce GTX 1080Ti graphics processing units (GPUs), CUDA8.0, and cuDNN 6.0.21. In this experiment, we perform and evaluate all experiments on our remote sensing image dataset benchmark. In order to obtain a more robust model for the size and shape of various remote sensing image targets, data enhancement operations should be carried out for each training image, such as random cropping and flipping of the input image. Considering the memory limitations of GPU, when the input size is 300, the batch size is set to 16. Following [27], we train a total of 120 k iterations, with a learning rate of 10^{-3} for the first 80 k iterations, 10^{-4} for the next 20 k iterations, then continue training for 20 k iterations with a 10^{-5} learning rate. We fine-tune the entire model with a weight decay of 0.0005 and a momentum of 0.9. The optimizer chosen is the stochastic gradient descent.

The aspect ratios of the images set the aspect ratios of the default boxes, as shown in Figure 4a, we set different aspect ratios for the default boxes according to the length-width ratio distribution of the image and denote them as $a_r = \{1/3, 1/2, 2/3, 3/2, 1, 2, 3\}$, in addition, we use K-means to

calculate the three cluster centers of the area size as the scale of the default boxes. During the training process, we need to confirm which default boxes conform to ground truth box detection and train the corresponding network. For box matching, we match default boxes to any ground truth box with Jaccard overlap higher than a threshold of 0.5.

4.4. Experimental Results and Comparisons

In order to comprehensively evaluate the superiority and effectiveness of the proposed CISPNet model, we compared it with the RCNN based algorithms, YOLO, and SSD methods. For a fair comparison, the same training data set and testing data set are used for the proposed CISPNet method and other comparison methods.

Table 2 and Figure 7 show the quantitative comparison results of ten different methods, measured by AP values and PRC, respectively. R-CNN [20] is the first algorithm to use CNN for target detection, and its obvious disadvantage is that it generates about 2000 regional proposals generated by selective search to hypothesize object locations, which requires the computer to have a lot of memory space. In addition, normalized operation of regional proposals makes the algorithm lose a lot of contextual information and semantic information, resulting in positioning accuracy of 61.48% mAP and detection efficiency of 0.07 frames per second (FPS). In order to solve the problem of losing context information and semantic information in the process of image normalization by R-CNN, Fast R-CNN [21] inputs an entire image into the network and extracts a fixed-length feature vector from the feature map through the region of interest (RoI) pooling layer, which reduces the running time of the detection networks and exposes the bottleneck of regional proposal computation. Therefore, Faster R-CNN [22] introduces a region proposal network (RPN) that shares entire image convolutional features with the detection network, thus enabling nearly cost-free region proposals and achieves 68.75% mAP at the speed of 10.2 FPS. YOLO [24] and YOLO v2 [25] convert the detection problem into a single regression problem, straight from image pixels to bounding box coordinates and class probabilities. YOLO can identify targets in images more quickly than the RCNN based algorithms due to the simple network structure, however, it has great restrictions when precisely localizing certain targets, especially small ones. The detection efficiency achieves 64.2 FPS, while the detection accuracy of mAP is only 61.61%. In SSD [27], multiple feature layers are used to improve the detection performance. However, the low-level layers have the shortcoming of insufficient semantic information, and the high-level layer has the problem of large loss of feature information due to the down sampling, which may result in inaccurate detection of targets. The mAP on the testing data set achieves 74.01%. Recently, various methods have attempted to improve the accuracy of SSD, especially for small targets. DSSD [28] uses ResNet-101 [19] instead of VGG16 [18] to achieve higher accuracy. RSSD [34], DOSD [40], and ESSD [41] have improved the detection accuracy by fusing the low-level and high-level features.

Table 2. Comparisons of speed and accuracy based on remote sensing image dataset tests. (For fair comparison, as introduced in Section 3, we have improved on the basis of SSD and only improved the network structure and the size of the default boxes.).

Method	Backbone	Input	Aircraft	Oiltanker	Ship	mAP	FPS
R-CNN [20]	AlexNet	1000 × 600	61.67	62.59	60.17	61.48	0.07
Fast R-CNN [21]	VGG16	1000 × 600	64.21	65.18	62.22	63.87	2.1
FasterR-CNN [22]	VGG16	1000 × 600	69.77	69.59	66.88	68.75	10.2
YOLO [24]	GoogleNet	448 × 448	60.26	58.47	59.14	59.29	68.4
YOLO v2 [25]	Darknet	544 × 544	62.79	61.25	60.78	61.61	64.2
SSD [27]	VGG16	300 × 300	73.91	74.73	73.39	74.01	71.0
DSSD [28]	ResNet101	321 × 321	75.75	73.11	74.32	74.39	20.6
RSSD [34]	VGG16	300 × 300	77.26	77.18	77.06	77.17	40.6
DSOD [40]	DenseNet	300 × 300	74.56	74.89	73.58	74.34	34.2
ESSD [41]	VGG16	300 × 300	78.76	79.64	78.09	78.83	50.1
CISPNet (ours)	VGG16	300 × 300	80.28	81.56	79.17	80.34	50.7

Figure 7. Precision-recall curves (PRCs) of the proposed CISPNet and other state-of-the-art approaches for airplane, ship, oiltanker, respectively.

In CISPNet, we use the CISP module to make the feature map understand more contextual information from different scales, and use the FFM module to fuse the features of the shallower layers to enhance the semantic information of the shallower layers. It solves the problem that SSD lacks understanding of context information and semantic information. From Table 2 and Figure 7, the proposed CISPNet method outperforms the other methods. Specifically, our CISPNet obtained 6.37%, 6.83%, and 5.78% performance gains in terms of mean AP over the aircraft, oiltanker, and ship, compared with the SSD model, respectively. Our CISPNet performs much better than the original SSD and other detectors. Specifically, our network can achieve 80.34% mAP at the speed of 50.7 FPS.

4.5. Detection Examples

In order to compare the detection performance of CISPNet and SSD for the remote sensing images in an intuitive manner, as shown in Figure 8, we visualized some of the images. From the figure, we can clearly see that our proposed CISPNet is beneficial to the detection of targets in the remote sensing images, especially for small targets.

In the upper two rows of images, we can see that SSD detection results show a single object with various overlapping boxes and the smaller box has partial image such as the part of a ship or the part of an aircraft. However, the detection results of the proposed CISPNet algorithm in the same picture did not show the box-in-box status. In the lower three rows of images, the dimensions of the aircrafts, ships and oiltankers are small. SSD cannot accurately detect the position of aircrafts, ships and oiltankers because of the lack of understanding of the scene. CISPNet can more fully understand the contextual information via the CISP models and FFMs so that it can better distinguish the background and the detected targets, and determine the location of the aircrafts, ships and oiltankers through the contextual information.

Figure 8. Detection examples on the remote sensing image test dataset with SSD/CISPNet model. (a) The results of the SSD. (b) The results of CISPNet.

4.6. Ablation Study

In this section, we set up different models and test them on the remote sensing image test dataset to verify the impact of each module on the detection performance. At the same time, we also discuss the effect of different aspect ratios of default boxes on detection accuracy. The results are shown in Tables 3 and 4.

Table 3. Ablation study of CISPNet on the remote sensing image dataset tests.

SSD	+CISP	+FFM	Metric/%			mAP	FPS
			Aircraft	Oiltanker	Ship		
✓			73.91	74.73	73.39	74.01	71.0
✓	✓		79.12	80.77	78.25	79.38	63.1
✓		✓	75.58	77.24	75.63	76.15	68.4
✓	✓	✓	80.28	81.56	79.17	80.34	50.7

Table 4. Ablation study: Effects of different aspect ratios on the remote sensing image dataset tests.

Aspect Ratios	Metric/%			mAP
	Aircraft	Oiltanker	Ship	
1	59.24	65.26	51.47	58.66
1/3,1,3	64.44	69.75	75.25	69.81
1/3,1/2,1,2,3	78.63	79.24	77.63	78.50
1/3,1/2,2/3,1,3/2,2,3	80.28	81.56	79.17	80.34

4.6.1. SSD with Context Information Scene Perception Modules and the Feature Fusion Modules

To estimate the contribution of different components of CISPNet, we further constructed three variants and tested them on the remote sensing image dataset tests to verify the impact of each module on the detection performance. The results are shown in Table 3. The first step is to validate the effect of CISP module. We insert four CIPs at positions shown in Figure 3a, and the first two rows of Table 3 illustrate that the CISP module has a higher promotion, and the mAP on the testing data set achieves 79.38%, which is improved by 5.37% compared with the conventional SSD. Second, we insert two FFMs at bottom positions shown in Figure 3a, the mAP is 76.15%, which is better than the conventional SSD. Finally, we insert four CIPs and two FFMs at positions shown in Figure 3a and the last row of Table 3, and the mAP is increased to 80.34%, which shows the effectiveness of the added four CIPs and two FFMs for enhancing the overall performance. These experiments have proved the importance of each component of CISPNet.

4.6.2. SSD with the Different Aspect Ratios of Default Boxes

In order to handle different target scales, we set default boxes with different aspect ratios to process images with different scales, and combined the results afterwards. We specified different aspect ratios for the models and recorded their evaluated results in Table 4.

As shown in Table 4, based on the experiments, we gradually increased the features of the different aspect ratios of the default box for detection. We started with a single default box with an aspect ratio of 1 and gradually denoted the default box with different aspect ratios. From the presented results, it can be inferred that the performance impacts of the large-scale different aspect ratios pairs were relatively large. This proved that the default boxes of different aspect ratios were beneficial for the improvement of the detection performance that in turn justified the effectiveness of our model.

4.7. Fuzzy Target Detection

Optical remote sensing image is usually photographed outdoors in a high angle shot via satellite and aerial sensors, and thus, there may be fuzzy targets, which are caused by cloud cover, shadow noise, and cluttered backgrounds. In addition to comparing the performance of the algorithms in the remote sensing image train and test dataset, the CISPNet and SSD algorithms are used to detect fuzzy images. The detection results are shown in Figure 9. We can clearly see that our proposed CISPNet is beneficial to the detection of fuzzy targets. This is because that the context information scene perception (CISP) module and feature fusion module (FFM) are added in CISPNet, which can more comprehensively understand the feature information from context information so that it can better distinguish between the background and the fuzzy targets, and determine the location of fuzzy targets through the contextual information.

Figure 9. Fuzzy targets detection examples on the remote sensing image test dataset with SSD and CISPNet models. (a) The results of the SSD. (b) The results of CISPNet.

4.8. Experiments on the NWPU VRH-10 Dataset

NWPU VRH-10 [32] is a well annotated dataset that includes remote sensing target detection, which contains 2D bounding boxes annotated on 650 remote images for airplane (757 instances), ship (302 instances), storage tank (655 instances), ballpark (390 instances), tennis court (524 instances), basketball court (159 instances), ground track filed (163 instances), harbor (224 instances), bridge (124 instances), and vehicle (447 instances), with 10 categories in total. The split ratio of training and testing set is 6:4, and the Intersection over Union(IoU) threshold of the evaluated is 0.5 on the testing.

As shown in Table 5, we evaluate our proposed method on the NWPU VRH-10 dataset to compare with the other popular architectures reviewed and re-implemented in [42]. Our CISPNet model achieved the highest mAP score (0.917) compared to others. Obviously, CISPNet is better than SSD in detecting small and medium targets such as aircraft, ship, storage tankers and so on. The detailed average precision of each category is listed in Table 5, and our proposed CISPNet performs well compared to most of them.

Table 5. Comparison of different detection methods based on NWPU VHR-10 benchmarks.

Methods	Faster R-CNN	YOLO	SSD	ESSD	CISPNet
Airplane	0.899	0.874	0.956	0.973	0.975
Ship	0.679	0.847	0.937	0.882	0.936
Storage tanker	0.643	0.427	0.617	0.807	0.804
Ballpark	0.932	0.931	0.995	0.974	0.982
Tennis court	0.697	0.658	0.860	0.849	0.871
Basketball court	0.529	0.870	0.944	0.917	0.931
Ground track filed	0.946	0.975	0.987	0.978	0.980
Harbor	0.616	0.800	0.950	0.936	0.944
Bridge	0.218	0.903	0.966	0.937	0.941
Vehicle	0.542	0.704	0.745	0.806	0.810
mAP	0.669	0.799	0.894	0.907	0.917

5. Conclusions

In this paper, we have proposed an effective single-stage framework, CISPNet, which is based on SSD and the context information scene perception (CISP) module, to improve the detection accuracy of small and dense targets in optical remote sensing. On the one hand, it proves that the CISPNet achieves better performance on the remote sensing image benchmark dataset than SSD and other advanced detectors, and the best performance is achieved with the most compact structure; on the other hand, it proves that CISPNet introduces the context information scene perception (CISP) module and this feature fusion module (FFM) can effectively improve performance. Moreover, experimental results on the NWPU VRH-10 dataset reveal that CISPNet has significantly outperformed the original SSD, especially for detecting small targets. In addition, in extended experiments, the performance of CISPNet in fuzzy target detection is better than the conventional SSD.

In the future, we expect research will be able to demonstrate the proposed method is not restricted to SSD based methods but also applicable to other structures utilizing multi-scale features.

Author Contributions: W.S. and J.J. conceived, designed and performed the algorithm and experiments; As the supervisor of W.S., he proofread the paper several times and provided guidance throughout the whole preparation of the manuscript; S.B. and D.T. provided advice for the preparation and revision of the paper. All authors read and approved the final manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (No. 11871059). The Natural Science Foundation of Sichuan Education Department (No. 18ZA0469, No. 15ZA0152), The New Generation of Artificial Intelligence Major Program in Sichuan Province (No. 2018GZDZX0036), The Key Program of Sichuan Science and Technology Department (No. 2018SZ0040), Supported by Sichuan Science and Technology Program (No. 2019YFG0299).

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

References

1. Torres, R.; Snoeij, P.; Geudtner, D.; Bibby, D.; Davidson, M.; Attema, E.; Potin, P.; Rommen, B.; Floury, N.; Brown, M.; et al. GMES Sentinel-1 mission. *Remote Sens. Environ.* **2012**, *120*, 9–24. [\[CrossRef\]](#)
2. Brusch, S.; Lehner, S.; Fritz, T.; Soccorsi, M. Ship surveillance with TerraSAR-X. *IEEE Trans. Geosci. Remote Sens.* **2010**, *49*, 1092–1103. [\[CrossRef\]](#)
3. Crisp, D.J. A ship detection system for RADARSAT-2 dual-pol multi-look imagery implemented in the ADSS. In Proceedings of the IEEE International Conference on Radar (2013), Adelaide, Australia, 9–12 September 2013; pp. 318–323.
4. Marino, A.; Sugimoto, M.; Ouchi, K.; Hajnsek, I. Validating a Notch Filter for Detection of Targets at Sea with ALOS-PALSAR Data: Tokyo Bay. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 74907–74918. [\[CrossRef\]](#)
5. Zhao, J.; Guo, W.; Zhang, Z.; Yu, W. A coupled convolutional neural network for small and densely clustered ship detection in SAR images. *Sci. China Inf. Sci.* **2019**, *62*, 42301. [\[CrossRef\]](#)
6. Ma, L.; Chen, L.; Zhang, X.; Chen, H.; Soomro, N.Q. A waterborne salient ship detection method on SAR imagery. *Sci. China Inf. Sci.* **2015**, *58*, 1–3. [\[CrossRef\]](#)
7. 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. Photo Grammetry Remote Sens.* **2013**, *85*, 32–43. [\[CrossRef\]](#)
8. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; pp. 886–893.
9. 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.
10. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [\[CrossRef\]](#)
11. Leng, X.; Ji, K.; Zhou, S.; Zou, H. An adaptive ship detection scheme for spaceborne SAR imagery. *Sensors* **2016**, *16*, 1345. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Wang, C.; Bi, F.; Zhang, W.; Chen, L. An Intensity-Space Domain CFAR Method for Ship Detection in HR SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 529–533. [\[CrossRef\]](#)

13. Atteia, G.E.; Collins, M.J. On the use of compact polarimetry SAR for ship detection. *ISPRS J. Photogramm. Remote Sens.* **2013**, *80*, 1–9. [[CrossRef](#)]
14. Yu, Y.D.; Yang, X.B.; Xiao, S.J.; Lin, J.L. Automated Ship Detection from Optical Remote Sensing Images. *Key Eng. Mater.* **2012**, *500*, 785–791. [[CrossRef](#)]
15. Zhu, C.; Zhou, H.; Wang, R.; Guo, J. A Novel Hierarchical Method of Ship Detection from Spaceborne Optical Image Based on Shape and Texture Features. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 3446–3456. [[CrossRef](#)]
16. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
17. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
18. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
20. 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, 2 July 2004; pp. 580–587.
21. Girshick, R. Fast R-Cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
22. Ren, S.; He, K.; Girshick, R.; Jian, S. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *arXiv* **2015**, arXiv:1506.01497. [[CrossRef](#)] [[PubMed](#)]
23. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Belongie, S. Feature Pyramid Networks for Object Detection. *arXiv* **2016**, arXiv:1612.03144.
24. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. *arXiv* **2015**, arXiv:1506.02640.
25. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
26. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
27. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
28. Fu, C.-Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. Dssd: Deconvolutional single shot detector. *arXiv* **2017**, arXiv:1701.06659.
29. Zhu, M.; Xu, Y.; Ma, S.; Tang, H.; Xin, P.; Ma, H. Airport Detection Method with Improved Region-Based Convolutional Neural Network. *Acta Opt. Sin.* **2018**, *38*, 0728001.
30. Chen, Z.; Zhang, T.; Ouyang, C. End-to-End Airplane Detection Using Transfer Learning in Remote Sensing Images. *Remote Sens.* **2018**, *10*, 139. [[CrossRef](#)]
31. Yang, X.; Sun, H.; Fu, K.; Yang, J.; Sun, X.; Yan, M.; Guo, Z. Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks. *Remote Sens.* **2018**, *10*, 132. [[CrossRef](#)]
32. Jeong, J.; Park, H.; Kwak, N. Enhancement of SSD by concatenating feature maps for object detection. *arXiv* **2017**, arXiv:1705.09587.
33. Wang, J.; Li, J.; Zhou, X.; Zhang, X. Improved SSD Algorithm and Its Performance Analysis of Small Target Detection in Remote Sensing Images. *Acta Opt. Sin.* **2019**, *39*, 0628005. [[CrossRef](#)]
34. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
35. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [[CrossRef](#)]

36. Lin, T.Y.; Maire, M.; Belongie, S.; Bourdev, B.; Girshick, B.; Hays, J.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; et al. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*; Zurich, Switzerland, 6–12 September 2014, pp. 740–755.
37. Everingham, M.; Eslami, S.M.A.; Gool, L.V.; Williams, C.K.L.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes Challenge: A Retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136. [[CrossRef](#)]
38. 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](#)]
39. Xia, G.S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. Aid: A benchmark dataset for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [[CrossRef](#)]
40. Shen, Z.; Liu, Z.; Li, J.; Jiang, Y.-G.; Chen, Y.; Xue, X. Dsod: Learning deeply supervised object detectors from scratch. In *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 22–29 October 2017; pp. 1919–1927.
41. Leng, J.; Liu, Y. An enhanced SSD with feature fusion and visual reasoning for object detection. *Neural Comput. Appl.* **2018**, *31*, 6549–6558. [[CrossRef](#)]
42. Yao, L.Q.; Hu, X.; Lei, H. A Study on Object Detection for Remote Sensing Images Based on Multi-scale Convolutional Neural Networks. *Acta Opt. Sin.* **2019**, *39*, 1128002.

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).