PROCEEDINGS OF SCIENCE

Low-Quality and Multi-Target Detection in RSIs

Guiyang Liu[1]
University of Chinese Academy of Sciences
Key Laboratory of Space Utilization, Chinese Academy of Sciences
Technology and Engineering Center for Space Utilization
Beijing, 100094, China
E-mail: liuguiyang15@csu.ac.cn

Shengyang Li[2a]**; Yuyang Shao**[3b]
Key Laboratory of Space Utilization, Chinese Academy of Sciences
Technology and Engineering Center for Space Utilization
Beijing, 100094, China
E-mail: [a]shyli@csu.ac.cn; [b]shaoyy@csu.ac.cn

To improve the recognition and detection accuracy and recall of objects in unnatural images, we use remote sensing video image data to detect low-quality and small targets. Based on this, we propose the use of a deconvolution network and hyper features to control convolutional feature quality. We call this approach the Quality Deconvolution Single Shot Detector (QDSSD) detection model. Through the frame-by-frame annotation of the video data from Jilin Satellite No. 1, we propose a CSU-RSI-Video dataset, with no fewer than 30 targets per image frame. We published the data for researchers to do experiments. To achieve small target detection, we enrich the information by gradually adding the underlying detail features to the upper layers and deconvolve the high-layer information to obtain stable detailed features for target detection. The empirical results show that, among the detected problems of low-quality small targets, the improved QDSSD network has better detection capability, and the detection effect is the best for many small targets that are close to each other. In the CSU-RSI-Video dataset, the QDSSD model's mAP achieves 0.90227 for a single target. Comparing to the You Only Look Once (YOLO) model, the result is still superior in accuracy.

ISCC2017
16-17 December 2016
Guangzhou, China

[1]Speaker
[2]Correspondent Author
[3]Supported by the National Defense Science and technology innovation fund of Chinese Academy of Sciences (Project Number of Y6031511WY)

PoS(ISCC 2017)026

1.Introduction

A remote sensing image (RSI) is a type of digital image of the ground that is obtained by loading different types of camera equipment on a platform that is away from the ground. There are different types of RSIs, such as RSIs acquired from satellites, which is called satellite RSIs. Satellite RSIs are widely used in many research fields because of their wide shooting ranges and low prices. Therefore, the application of satellite RSIs are very extensive, such as environmental monitoring, crop yield estimation, fire detection, traffic navigation, and city and regional surveying and planning.

The target detection based on optical remote sensing image is consistent with the objective detection of natural images, which is to predict the position and corresponding probability of the object (e.g., vehicles, ships, and buildings) in the given image. Compared with the target detection of natural images, the target detection of RSIs still faces great challenges. The RSI has many features, such as a large angle of view, large background changes in the view, and poor image quality. These problems will pose great challenges to the accuracy and robustness of target detection. With the enhancement of space exploration ability and the explosive growth of remote sensing satellite images, there is an increasing demand for the application of RSIs in various fields. Early low-resolution remote sensing satellite image data cannot meet the need of target recognition and detection tasks in complex scenarios. With the continuous improvement of the aerospace industry and the improvement of probe payload design, RSI post-processing technology, the presence of high-resolution remote sensing (HRRS) satellites (e.g. GF1, Planet, ZY3) and unmanned aerial images have been providing highly detailed information about the structure of the target and the surrounding environment. Before the rapid development of deep learning, considerable efforts have been made to design various algorithms for the detection of different types of targets in RSIs and aerial images. Although there are a large number of studies researching on this topic, different approaches are applied in particular scenarios. As tremendous achievements have been made by the depth model in the field of classification and detection, we should focus on the design of general models to find a common method in the field of RSI.

In this work, we present a framework called Quality Deconvolution Single Shot Detector (QDSSD) to effectively achieve accurate representation in image detection, as illustrated in Figure 2. The core idea of QDSSD is to coordinate VGG and a deconvolution net to learn more about features that are complementary to each other, and thus, rich features will be extracted from the raw images. In addition, in response to the deficiencies of the existing RSI datasets, we propose an RSI dataset called CSUVideo, which contains low-quality and clear small targets.

2.Related Work

In this section, we mainly review the algorithm of object recognition and detection in the natural scene, and emphatically introduce the object detection model based on deep learning theory related to our work.

State-of-the-art CNNs-based object proposal and detection methods can be divided into two groups: (i) region proposal-based methods and (ii) proposal-free methods. Object proposals [1][2]considerably reduce the required amount of computation compared to sliding window methods[3] in detection frameworks. These frameworks can be classified into two general approaches: traditional frameworks and deep learning based models. Traditional algorithms mainly attempt to generate region proposals by merging multiple clusters or by scoring windows that are likely to be included in objects[4]. These algorithms usually adopt features

like geometric structures[5], color information[6], surface texture[7], surrounding environment characteristics[8], etc. Recently, many deep neural network-based detection models have achieved positive results. DeepBox[9] is trained with a novel four-layer CNNs to rerank region proposals through a bottom-up method, named EdgeBoxes[6]. Girshick et al. propose region-based convolutional network, named R-CNN[10]. In this framework, the convolution neural network is responsible for the feature extraction of the image, and a few thousand independent region proposals are adopted for object detection. Based on the Fast R-CNN[10], they replace part of the candidate area with a region proposal network[11] (RPN) and integrate the model into an end-to-end system for the overall training on the GPU. This network is called Faster R-CNN[11]. Different from others, Redmon et al. propose a YOLO framework which omitted the proposal-generated step, directly predicting candidate bounding boxes and class probabilities from images[12]. In practice, proposal-based models outperform proposal-free models regarding detection recall and accuracy. Another proposal-free method like YOLO is called SSD[13]. SSD improves YOLO in several ways: (i) through the use of fully connected layers to predict categories and anchor offsets for bounding box locations; (ii) through the use of convolutional feature maps in different sizes for prediction at different scales; (iii) through the use of default anchor boxes and different aspect ratios for adjusting varying object shapes[14].

As for image classification problems, scholars have designed many effective performing neural network architectures. Many different networks have emerged, such as AlexNet[7], VGG16[15], ResNet[16], InceptionX[17] and DenseNet[18]. Meanwhile, several regularization techniques, such as Dropout[19] and Batch Normalization[20], have also been proposed to further enhance model capabilities. These networks can well capture the features in an image and have the characteristic of strong feature expansiveness, which provides a good foundation for target detection.

3.Datasets

There are many public RSI datasets, but the sizes of the datasets are limited-- the content is not comprehensive, and annotation accuracy has yet to be verified. Therefore, it is necessary to collect and summarize the remote sensing data from China generated in recent years. In this section, we review several datasets commonly used for object detection and describe the proposed CSU Video Dataset.

3.1 Existing Remote Sensing Target Datasets

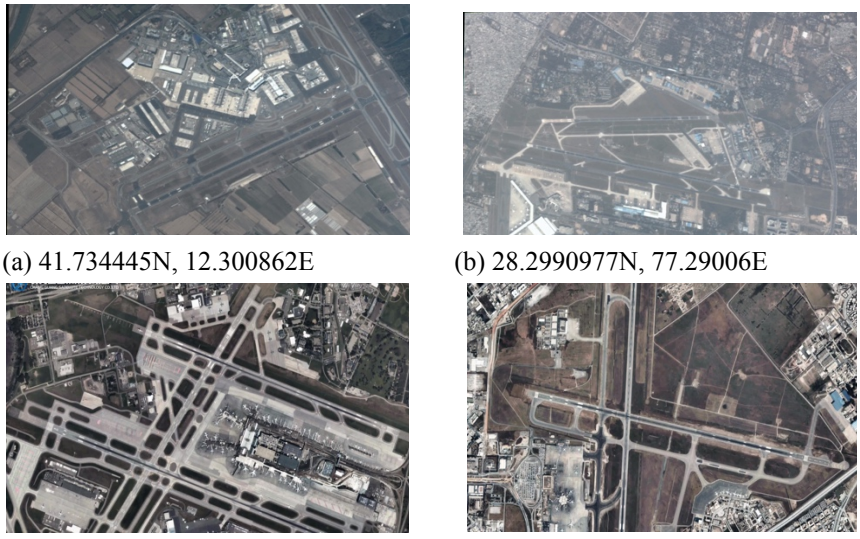
Datasets	Targets per Class	Class	Total Images	Spatial Resolution(m)	Year
TAS[21]	1319	1	30	0.3	2008
OIRDS[22]	1800	1	600	0.3	2009
WHU-RS19[23]	50	19	1005	~0.5	2012
SZTAK-INRIA[24]	665	1	30	2	2012
SIRI-WHU[25]	200	12	2400	2	2016
VEDAI[21]	~300	9	2731	2	2015
RSC11[26]	~100	11	1232	0.2	2016
RSSCN[27]	400	7	2800	---	2015
NWPU-RESISC[28]	700	45	31500	0.2~30	2016

NWPU-VHR[29]	80~200	10	800	0.2~30	2014
CSU-RSI-VIDEO[4]	60~85	1	3000	0.7~1.1	2017

Table 1: List of RSI datasets in recent years. The Data Utilization Center of Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences(CSU/CAS) annotates the Jilin No. 1 Star, which is a commercial video satellite and the International Space Station video data. Now, it is open to researchers worldwide.

3.2 CSUVideo: RSIs for Object Detection and Tracking

After analyzing the datasets, we find that many remote sensing datasets are now available from Google Maps, unmanned aerial vehicles (UAV), and high-resolution remote sensing imagery. Most of the data is processed to improve image quality. This data is a driving force for target detection in RSIs, but these algorithms tend to be unacceptable when dealing with low-quality, complex scenarios. Note that Google Earth images are post-processed using RGB rendering from the original optical aerial images. Jilin No. 1 is China's first self-developed commercial remote sensing video satellite system. It was developed by the Changchun Institute of Optics, Fine Mechanics and Physics. Jilin No. 1 Optical A Star is China's first self-developed high-resolution Earth observation optical imaging satellite. It has a ground-pixel resolution of 0.72 meters in full color and 2.88 meters in multi-spectrum. It has a conventional push-scan, large-angle side-swing, with three-dimensional, multi-band splicing and other imaging modes. No. 1 Jilin Video Star's ground-pixel resolution is 1.12 meters, mainly used to carry out high-resolution video imaging technology test verification. We use the Jilin No. 1 video satellite data to annotate and obtain the optical data in the real scene.



(a) 41.734445N, 12.300862E (b) 28.2990977N, 77.29006E

(c) 44.881944N, 93.221666E (d) 36.851111N, 10.22722E

Figure 1: Key Frame Image from Jilin No. 1 Commercial Satellite

We use the Jilin No. 1 Video satellite. Each video is about 30s, the video frame rate of 25 fps (frame per second) and each frame size is . Each frame of video becomes very vague compared to high-resolution image data. In many cases, the annotator can only infer whether a region is a potential target based on front-to-back frames information. Based on the above description, we classify the target according to the degree of cognition of the human eyesight. The target is divided into: the airplane that can be observed directly by the personnel and the airplane inferred from the surrounding context. The researchers can not only be part of the target

[4]https://pan.baidu.com/share/init?surl=qYkeefU

detection data, but also can split the dataset used for target tracking. We will open the dataset to global scientific research workers engaged in scientific research tasks in the near future.

4. Approaches

In this part, we compare the performance of the YOLO model to that of the QDSSD model using the CSU-RSI-Video dataset. For the following three aspects, we have made improvements to the SSD detection model. (i) The ideal of HyperNet[4] is introduced to the convolution feature extraction; (ii) using feature semantic information and the high-level information deconvolution for target detection; (iii) through the deconvolution of the characteristics of quality supervision and quality control.

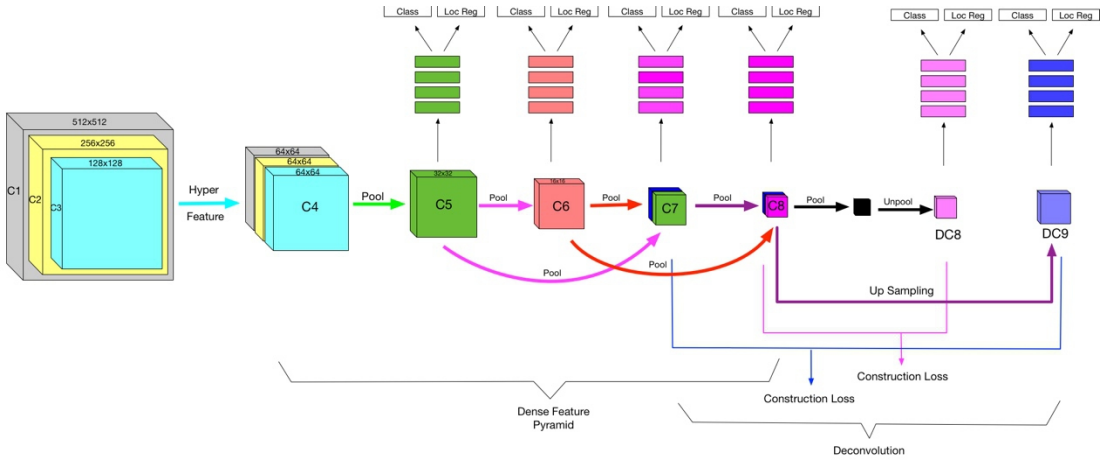


Figure 2: Architecture of the Quality Deconvolution SSD

4.1 Hyper Feature

Initially, a full image is forwarded through the convolutional layers, and the activation feature maps are produced. The hierarchical characteristics of different scales are integrated by stacking and then compressing them into a uniform space, named Hyper Feature[4]. To integrate multi-scale maps into the same resolution, we carry out different down-sampling strategies for different scale layers. This allows the details of the bottom layer to be added directly to the specified convolutional layer, providing detailed features for small target detection. In order to enrich the detailed characteristics of the raw data, we adopt different interpolation methods (Bilinear, Spline) and sampling strategies (MaxPool, AvgPool) to obtain the convolutional feature. Finally, we normalize hyper features using local response normalization[30](LRN), batch normalization (BN) and concatenate them into single output cube, which named Hyper Feature Cube.

4.2 High-Level Information Participation Detection

We detect that the target size is not large, ranging from to . The high-level semantic information of convolutional neural networks is only slightly affected by the bottom of the image, and it can provide invariant information. We will be the last layer network characteristic of convolutional deconvolution in the new scale on target detection.

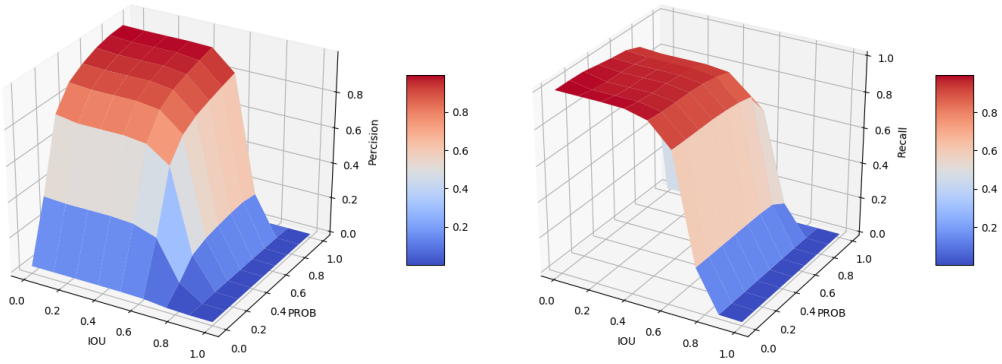
4.3 Control Convolution Feature Quality

Many researchers are constantly improving the network structure, using detection network in different layers for target detection. However, if the quality of a convolution network can not

be well supervised, then it cannot achieve better detection results in the complex detection network. We minimize the mean square error of the characteristics of the C8, DC8, C7, and DC9 layers and control the quality of the high-level deconvolution as a part of the global constraint.

5.Experiments and Results

The CSU-RSI-Video has four videos. Each one is about 30s. The video is taken at a specific airport where the scene changes slowly, and the scene contains static and moving objects. We take the first three videos as the training set, and the fourth video as the test set. According to the reservoir sampling algorithm, the corresponding length of 1/5 data is taken for each video in the training set. We use two algorithms to compare experiments in this data. One is the traditional YOLO[12] model, and the other is the QDSSD model that we proposed. In our experimental evaluation of QDSSD model, both training and testing were done on a Linux PC with Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.40GHz, a NVIDIA GeForce GTX TITAN X GPU, and 128GB of memory. After 100,000 generations, our mean Average Precision (mAP) is 0.90227 in the category of airplane detection. We arrive at the following results.



(a) Plane's Precision (b) Plane's Recall

Figure 3: The accuracy and recall rate of Tunisia Airport video test data (IOU means intersection-over-union, PROB means the probability values of corresponding classes).

Figure 4: Our QDSSD model results. Schematic diagram of detection and recognition results

The red squares show the ground truth of the target, and the green squares show that the model selected the box when the target recognition probability is 0.4.

Figure 5: The original YOLO model results that were fine-tuned from the ImageNet dataset.Schematic diagram of detection and recognition results

6.Discussion

From the above results, we can see that the effect of the QDSSD model is better than that of the original YOLO model. The YOLO model is very effective in target detection because it uses dimension clustering, fine grained features, and multi-scale training. However, it has weaknesses. First, the YOLO model controls the amount of segmentation of the image grid by adjusting the dimensions of the output vector of the fully connected layer through the prediction of the target probability of each grid and the regression of the corresponding position. This will limit the effect of detection of the YOLO in different scales, for different objects, and for different numbers of objects in the images. Second, the model does not have a single part of location regression. It stiffly maps the features in different locations through the final full convolution layer, which is not effective for multi-target detection. Third, to adapt to multi-scale objects, it has to transform the input data into multiple scales, which is a waste of resources. Compared to the YOLO model, the QDSSD model has obvious advantages. It is optimized in different feature layers, shares the parameters of the convolution layers, and can detect multiple targets better than the YOLO model. We also consider the scale of small targets as being mostly concentrated in the first few layers of convolution. To improve the robustness of detection, we use a deconvolution network and hyper connections to control quality. In our experiment, we implemented the QDSSD model. Through the global optimization of these three parts, we found that it can detect small targets better than the YOLO model.

References

[1] Carreira J, Sminchisescu C. *Constrained parametric min-cuts for automatic object segmentation[C]*// Computer Vision and Pattern Recognition. IEEE, 2010:3241-3248.

[2] Cheng M M, Zhang Z, Lin W Y, Torr P. BING*: Binarized Normed Gradients for Objectness Estimation at 300fps[C]*// Computer Vision and Pattern Recognition. IEEE, 2014:3286-3293.

[3] Felzenszwalb P F, Girshick R B, Mcallester D, Ramanan D. *Object detection with discriminatively trained part-based models.[J]*. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2010, 32(9):1627-1645.

[4] Kong T, Yao A, Chen Y, Sun F. *HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection[C]//* Computer Vision and Pattern Recognition. IEEE, 2016:845-853.

[5] Sande K E A V D, Uijlings J R R, Gevers T, Smeulders AWM. *Segmentation as selective search for object recognition[C]*// International Conference on Computer Vision. IEEE Computer Society, 2011:1879-1886.

[6] Zitnick C L, Dollár P. *Edge Boxes: Locating Object Proposals from Edges[C]*// European Conference on Computer Vision. Springer, Cham, 2014:391-405.

[7] Pont-Tuset J, Barron J, Marques F, Arbelaez P, Malik J. *Multiscale Combinatorial Grouping[C]*// Computer Vision and Pattern Recognition. IEEE, 2014:328-335.

[8] Alexe B, Deselaers T, Ferrari V. *What is an object?[C]*// Computer Vision and Pattern Recognition. IEEE, 2010:73-80.

[9] Kuo W, Hariharan B, Malik J. DeepBox*: Learning Objectness with Convolutional Networks[C]*// IEEE International Conference on Computer Vision. IEEE, 2015:2479-2487.

[10] Girshick R. *Fast R-CNN*[C]// IEEE International Conference on Computer Vision. IEEE Computer Society, 2015:1440-1448.

[11] Ren S, He K, Girshick R, Sun J. Faster R-CNN: *Towards Real-Time Object Detection with Region Proposal Networks.[J]*. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149.

[12] Redmon J, Divvala S, Girshick R, Farhadi A. *You Only Look Once: Unified, Real-Time Object Detection[J]*. 2015:779-788.

[13] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S. SSD*: Single Shot MultiBox Detector[J]*. 2015:21-37.

[14] Shen Z, Liu Z, Li J, Jiang Y, Chen Y, Xue X. *DSOD: Learning Deeply Supervised Object Detectors from Scratch[J]*. 2017.

[15] Simonyan K, Zisserman A. *Very Deep Convolutional Networks for Large-Scale Image Recognition[J]*. Computer Science, 2014.

[16] He K, Zhang X, Ren S, Sun J. *Deep Residual Learning for Image Recognition*[C]// Computer Vision and Pattern Recognition. IEEE, 2016:770-778.

[17] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. *Going deeper with convolutions[C]*// Computer Vision and Pattern Recognition. IEEE, 2015:1-9.

[18] Huang G, Liu Z, Weinberger K Q, Laurens VDM. *Densely Connected Convolutional Networks[J]*. 2016.

[19] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. *Dropout: a simple way to prevent neural networks from overfitting[J]*. Journal of Machine Learning Research, 2014, 15(1):1929-1958.

[20] Cooijmans T, Ballas N, Laurent C, Courville A. *Recurrent Batch Normalization[J]*. 2016.

[21] Razakarivony S, Jurie F. *Vehicle detection in aerial imagery : A small target detection benchmark[J]*. Journal of Visual Communication & Image Representation, 2016, 34:187-203.

[22] Tanner F, Colder B, Pullen C, Heagy D. *Overhead imagery research data set — an annotated data library & tools to aid in the development of computer vision algorithms[C]*// Applied Imagery Pattern Recognition Workshop. IEEE Xplore, 2009:1-8.

[23] Guofeng Sheng, Wen Yang, Tao Xu, Hong Sun. *High-resolution satellite scene classification using a sparse coding based multiple feature combination[J]*. International Journal of Remote Sensing, 2012, 33(8):2395-2412.

[24] Benedek C, Descombes X, Zerubia J. *Building Development Monitoring in Multitemporal Remotely Sensed Image Pairs with Stochastic Birth-Death Dynamics[J]*. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2012, 34(1):33.

[25] Zhao B, Zhong Y, Xia G S, Zhang L. *Dirichlet-Derived Multiple Topic Scene Classification Model for High Spatial Resolution Remote Sensing Imagery[J]*. IEEE Transactions on Geoscience & Remote Sensing, 2016, 54(4):2108-2123.

[26] Zhao L, Tang P, Huo L. *Feature significance-based multibag-of-visual-words model for remote sensing image scene classification[J]*. Journal of Applied Remote Sensing, 2016, 10(3):035004.

[27] Zou Q, Ni L, Zhang T, Wang Q. *Deep Learning Based Feature Selection for Remote Sensing Scene Classification[J]*. IEEE Geoscience & Remote Sensing Letters, 2015, 12(11):2321-2325.

[28] Chen C. *Remote Sensing Image Scene Classification Using Multi-scale Completed Local Binary Patterns and Fisher Vectors[J]*. Remote Sensing, 2016, 8(6):483.

[29] Cheng G, Han J, Zhou P, Guo L. *Multi-class geospatial object detection and geographic image classification based on collection of part detectors[J]*. Isprs Journal of Photogrammetry & Remote Sensing, 2014, 98(1):119-132.

[30] Jia Y, Shelhamer E, Donahue J, Karayev S, Jonathan L, Girshick R, Guadarrama S, Darrell T. Caffe*:Convolutional Architecture for Fast Feature Embedding[J]*. 2014:675-678.