remote sensing

MDPI

Article

Enhanced Feature Representation in Detection for Optical Remote Sensing Images

Kun Fu [1,2,3,4,5,†], Zhuo Chen [1,3,4,*,†], Yue Zhang [2,4] and Xian Sun [2,3,4]

[1] School of Microelectronics, University of Chinese Academy of Sciences, Beijing 100049, China; fukun@mail.ie.ac.cn

[2] Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China; zhangyue@aircas.ac.cn (Y.Z.); sunxian@mail.ie.ac.cn (X.S.)

[3] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China

[4] Key Laboratory of Network Information System Technology, Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China

[5] Institute of Electronics, Chinese Academy of Sciences, Suzhou 215000, China

[*] Correspondence: chenzhuo@tju.edu.cn; Tel.: +86-10-5888-7208

[†] These authors contributed equally to this work.

Received: 19 June 2019; Accepted: 5 September 2019; Published: 8 September 2019

Abstract: In recent years, deep learning has led to a remarkable breakthrough in object detection in remote sensing images. In practice, two-stage detectors perform well regarding detection accuracy but are slow. On the other hand, one-stage detectors integrate the detection pipeline of two-stage detectors to simplify the detection process, and are faster, but with lower detection accuracy. Enhancing the capability of feature representation may be a way to improve the detection accuracy of one-stage detectors. For this goal, this paper proposes a novel one-stage detector with enhanced capability of feature representation. The enhanced capability benefits from two proposed structures: dual top-down module and dense-connected inception module. The former efficiently utilizes multi-scale features from multiple layers of the backbone network. The latter both widens and deepens the network to enhance the ability of feature representation with limited extra computational cost. To evaluate the effectiveness of proposed structures, we conducted experiments on horizontal bounding box detection tasks on the challenging DOTA dataset and gained 73.49% mean Average Precision (mAP), achieving state-of-the-art performance. Furthermore, our method ran significantly faster than the best public two-stage detector on the DOTA dataset.

Keywords: remote sensing; one-stage detector; top down module; receptive field

1. Introduction

Object detection in optical remote sensing images is widely applied into many key fields such as environmental monitoring, geological hazard detection, precision agriculture, etc. [1]. With the rapid development of remote sensing technology, the number of remote sensing images has been growing dramatically and the quality of images has been improved rapidly. In the meanwhile, the task of object detection is increasingly complicated, which may face a large amount of object categories and complex target scenes. The conventional analytical methods are hard to meet growing diversified needs. In this case, the introduction of Convolutional Neural Networks (CNN) [2], the performance of which has been widely proved on general object detection [3,4], attracts growing attention from remote sensing field. In general, these detectors can be divided into two main categories: the two-stage detection and the one-stage detection framework.

The two-stage detection method consists of two steps: the pre-processing step for region proposal and the classifying step. At the first step, region proposals that contain potential objects instead of background are generated from one image through an algorithm by using Selective Search (SS) [5]. The second step includes determination for the label and regression for the bounding box. Category-specific classifiers will be used to determine the labels of those region proposals. In addition, the bounding box will be corrected more precisely through box regression.

The two-stage detection and their extensions like Region-CNN (R-CNN) [6], Fast R-CNN [7], Faster R-CNN [8], and Feature Pyramid Networks (FPN) [9] make great progress on general object detection. However, two-stage detection has inherent defects: since numerous region proposals are required, training and testing are getting slow.

The one-stage detectors, such as You Only Look Once (YOLO) [10], Single Shot MultiBox Detector (SSD) [11], and Retina-Net [12], straightforwardly predict label confidences and locations of all possible objects on the feature maps from one image. After removing the region box, one-stage detection gains great improvement in detection speed but gets relatively poor detection accuracy on detecting multi-scale and multi-class targets. To address this issue, enhancing the capability of feature representations is effective to improve the performance of one-stage detector. SSD predicts objects from multiple layers with multi-scale features. This strategy neglects semantic information of features from shallow layers and does not solve the problem of poor performance for detecting small objects yet. FPN introduces a top-down structure to fusion features from different layers and brings about more semantic information of features. However, fusing features layer by layer is not effective enough to represent context information. Based on SSD, Receptive Field Block Net (RFBNet) [13] introduces receptive field Block (RFB) ahead of prediction module. RFB simulates the mechanism between the human population Receptive Field (pRF) size and eccentricity in the human visual system, which effectively produces multi-scale receptive fields. However, branches of RFB structures are independent, which results in the loss of correlation of features.

Compared with natural scenes, object detection in remote sensing images suffers from fundamental challenges like changeable-scale objects, small objects in dense-target scenes. Figure 1 illustrates the instance size distribution of typical remote sensing dataset DOTA [14] and natural dataset VOC [15]. It is clear that the DOTA dataset is dominated by instances under 100 pixels while the instances in VOC dataset are widely distributed between 0 to 500 pixels. In addition, there are both some extremely small and huge objects in remote sensing scenes. For example, a bridge can be as small as 15 pixels and as large as 1200 pixels, which is 80 times larger than the smaller one. The enormous differences among instances make the detection task more challenging because models must be flexible enough to handle extremely tiny and giant objects.

To handle these problems above, this paper attempts to enhance the ability of feature representation by proposing two novel structures: Dual Top-Down Module (DTDM) and Dense-connected Inception (Dense Inception) module. DTDM is merged into top layers of the backbone network, with limited extra computation under control, and replaces the top layers of the backbone network with two top-down modules branches. For one branch, the features from the last several layers are integrated from upper layers to lower layers and delivered directly to the prediction module. For the other, the top layers of the backbone network are combined to generate the fusion feature and replace the top layer. By combining different level features—low-level local detailed features and high-level global semantic features—the ability of feature representation is significantly enhanced and the performance on small objects is improved. Inspired by the dense structure of DenseNet [16], this paper proposes Dense Inception with dense-connected branches of different atrous layers, to combine features from branches with stronger correlation and more effective field, yielding more efficient integrated features. Dense Inception further enhances the ability of feature representation of each feature map, which is useful for detecting changeable-scale objects.

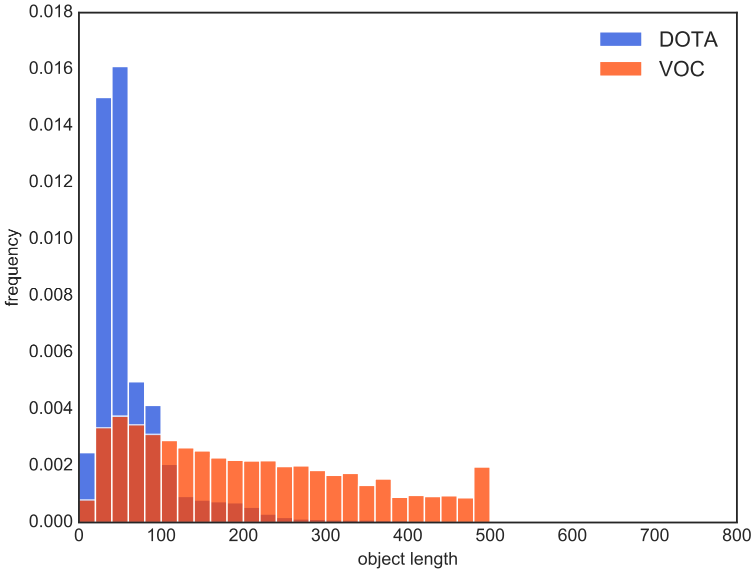


Figure 1. Comparison of instance size distribution of DOTA and VOC datasets.

Our detection framework effectively improves the model's ability of feature representation. With adding newly proposed structures, our detector achieves state-of-the-art performance among published articles on the public DOTA dataset. In this paper, our contributions can be briefly summarized as follows:

1. We propose a novel top-down module, DTDM, to combine multi-scale features from fusion layers of multiple layers. The multi-scale features from different layers can be effectively merged to detect multi-scale objects in complicated scenes.
2. We introduce a Dense Inception module to extract features more efficiently. The new structure can generate features that cover larger and denser receptive field size, as well as effectively integrating contextual information.

The rest of paper is organized as follows: Section 2 introduces proposed modules in detail. We compare DTDM modules with several relative modules. In Section 3, we conduct our experiment and list compared experiments. In addition, some experiment details are shown for further comprehension. The result of all experiments as well as their analysis are presented in Section 4. In Sections 5 and 6, we draw the conclusions from our work and make the plan for further work.

2. Methods

In this section, we describe the architecture of DTDM module and Dense Inception. Furthermore, we also compare proposed modules with relative structures. Finally, the whole pipeline structure of the model is presented in detail.

2.1. DTDM

CNN module consists of a series of convolution layers with filters, Pooling, Fully Connected layers (FC) and Softmax function. The backbone network of detectors is based on CNN modules, in which many convolution layers cascade and generate various feature maps with size descending. The features from high layers have large receptive fields and strong semantics, with great robustness to variations like illumination but at the cost of geometric details. In contrast, the features from lower layers have small receptive fields and rich geometric details, but possess much less semantic information. Classic detectors based on deep CNN like R-CNN, Fast R-CNN, Faster R-CNN and YOLO, merely make use of the top layer of CNN modules for feature representation. However, a single feature map is not enough to extract the various features with multi-scale and multi-level feature information.

To handle this contradiction, there are some explorations on multi-level object detection: detecting at multiple CNN layers or detecting with combined features of multiple CNN layers. The former, like FPN in Figure 2a, combines predictions from multiple feature layers to handle multi-scale objects. The latter, like Top-Down Module (TDM) [17] in Figure 2b, is generally implemented by skip connections: skip selected layers and feed the output of lower layers as the input to upper layers. It proves that both explorations have made progress to a large extent. In order to achieve the best combination of feature information, it is natural to implement both strategies. Since the two methods above separately work at different stages of detection: detecting at multiple CNN layers was applied in prediction or detecting with combined features in extracting semantic information. This paper proposed DTDM to detect with combined features and merged the structure into the backbone network Visual Geometry Group Net (VGG16) [18].

As is presented in Figure 2c, the input was fed into two branches: for one branch, feature maps of decreasing sizes from different layers were upsampled to the same size and concatenated together. For the other branch, it integrated a top-down network with lateral connections. High level semantic features were transmitted back by the top-down network and combined with the bottom-up features from intermediate layers. Then, the outputs of both branches were further processed via element-wise summation for prediction.

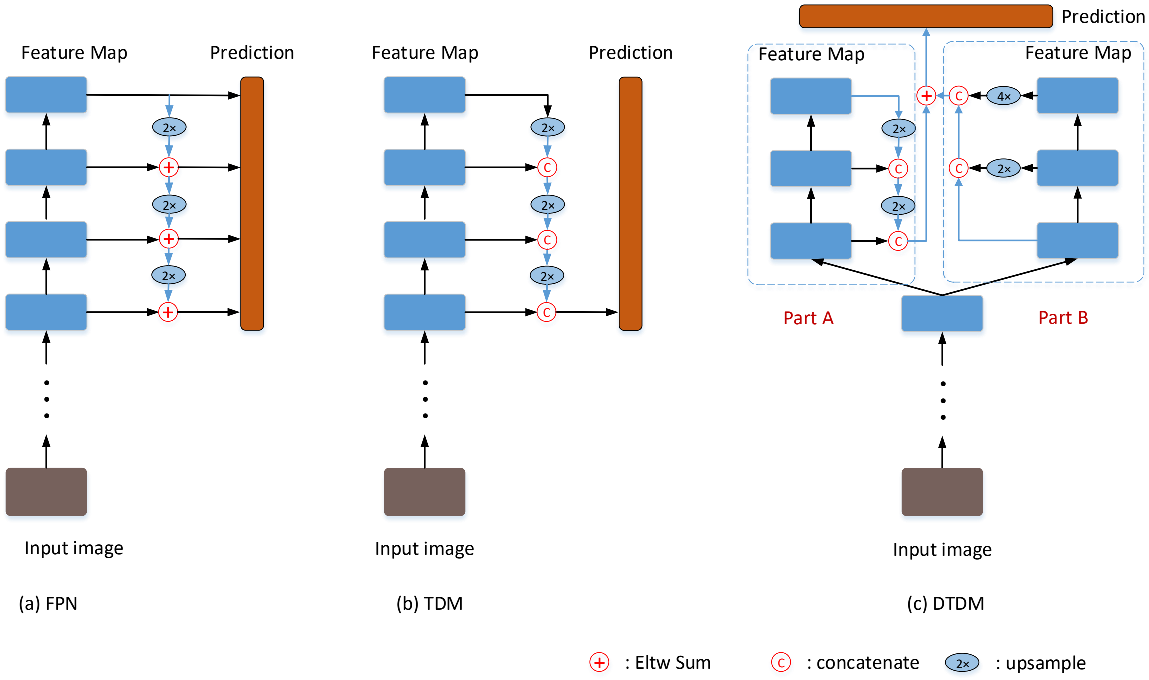
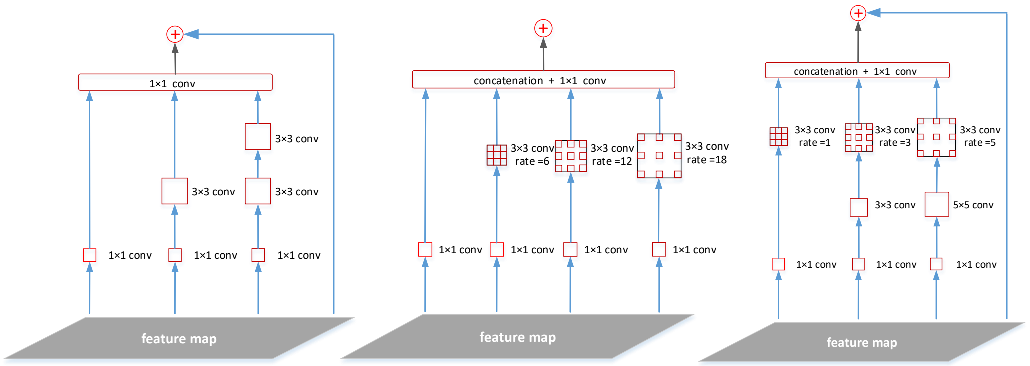


Figure 2. The top-down module in several structures. (**a**) Feature Pyramid Networks; (**b**) Top-Down Module; (**c**) Dual Top-Down Module.

2.2. Dense Inception

At the beginning of deep learning, CNN was used to extract feature from one image. Although CNN made a breakthrough on the performance of object detection, it still needs to be enhanced. In order to improve the image representation performance, very deep convolutional network have been proved efficient. However, the deepening of convolutional network brought about serious side effects: the high computational cost and huge amount of parameters.

One alternative is increasing the depth and width of the network with limited extra computational cost. Considering this motivation, the inception architecture [19] achieved great performance at relatively low computational cost. Figure 3 shows three typical multi-branch convolution layers. Inception architecture attempted to extract multi-scale feature information by launching multi-branch convolution layers with different convolution kernels. In the multi-branch CNN architecture of inception, all kernels were sampled at the same center. To further enhance the representation ability of CNN, the atrous convolution was first introduced [20] to apply multi-scale kernels, which is equivalent to convolving the input with up-sampled filters for a large receptive field. In many scenes like aerial image, there were various objects in multiple sizes. To deal with this situation, the feature maps must cover multi-scale receptive fields. Thus, RFBNet made use of multi-scale dilated convolution layers to adapt to multi-scale kernels of CNN, which covers multi-scale receptive fields and makes it significantly useful for extracting multi-level information. Furthermore, the introduction of residual connection can accelerate the training of inception network significantly [21]. However, each branch in RFBNet is independent of one another, resulting in the lack of correlation of semantic information from extracted features. To rethink the structure of RFBNet, there is another alternative to integrate multi-branch information ahead of delivery together to the prediction module. Inspired by the dense structure in densenet in semantic segmentation, this paper proposes a novel structure, called Dense Inception, in which the output of small-kernel CNN branch is concatenated to the input to the other bigger-kernel ones. In addition, there is a lateral connection from the original feature map to the output of the module. The combined features were further processed and delivered to the prediction module. The new module not only increased the receptive field of features but also made features of each branch more relational, which is greatly useful for detecting multi-scale objects. Figure 4 presents the structure of Dense Inception in detail.



(a) Inception resnet v2 (b) Aspp (c) RFB

Figure 3. Three typical structures of multi-branch convolution layers. (**a**) Inception resnet v2; (**b**) Aspp; (**c**) Receptive Field Block.

There are two strategies on Dense Inception: cascading and parallel modes. The cascading mode defines the cross-layer dense-connected structure. The lower atrous layers are connected to upper atrous layer, which can generate large receptive fields. As for parallel mode, multiple atrous layers accept the same input. It means that the output is indeed a sampling of the input with different scales of receptive fields. The module not only inherits the advantage of aspp by using multiple atrous-convolved feature representation, but also makes better use of inner relationship of multiple layers.

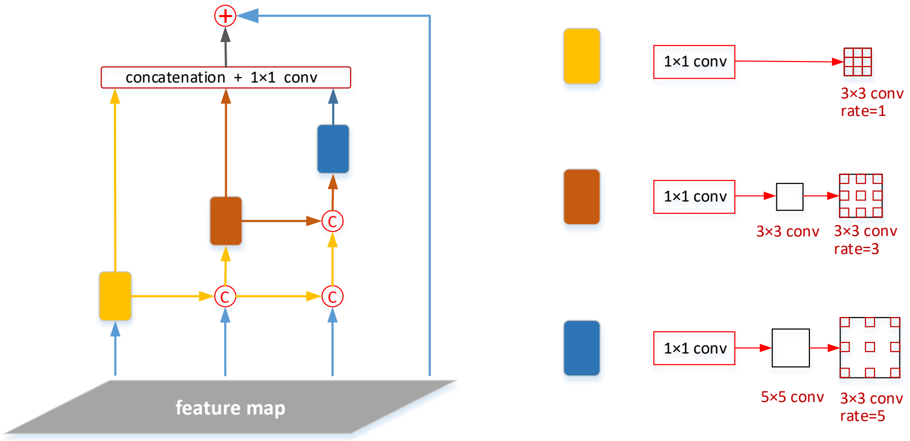


Figure 4. The structure of Dense Inception.

In the inception module, the branches were organized in parallel. Multiple atrous layers were fed with the same inputs and the inputs were indeed sampled with various scales of receptive fields. As for single branch, there exists one theoretical issue called 'gridding' [22]. During atrous convolutional operation, much feature information will be lost: for a feature map in dilated convolutional operation with kernel size k and dilation rate r, only $k \times k$ out of $(k + r) \times (k + r)$ in the region were used for the computation. When r became large, the information sampled by atrous convolution would be weakly correlated, which would cause severe loss of local information. The cascading atrous layers with increasing dilation rate can solve the problem. The progressive concatenation of atrous layers made the top atrous layers accept more complete information from the input.

In Dense Inception, the atrous layers are organized both in cascading and parallel modes. Since the atrous layer of larger dilation rate accepts the output of the other atrous layers of smaller dilation rate, it can efficiently produce larger receptive fields. The final output of Dense Inception is a feature map derived from multi-scale and multi-rate atrous convolutions. The details are presented as follows:

Firstly, in order to decrease and regularize the channel numbers of the input, we placed a 1×1 convolution layer between each branch and the original feature map, respectively. For one branch, the head was followed by different convolution layers of increasing kernel sizes. After the multi-size kernel convolution layer, we put the corresponding atrous convolution layer whose rate is identical with the kernel size of previous layer. Finally, we adopted the shortcut design from inception from Resnet and the original feature was fed for concatenation with outputs from each branch.

2.3. Structure of Detector

Considering the demand of both precision and speed, we chose lightweight network VGG16 as our backbone network. VGG16 is a classic lightweight network which is adopted by lots of state-of-the-art detectors, especially by those high-speed ones. However, lightweight detector is popular for its "light" but relatively weak ability of detecting multi-scale and multi-class targets. As is shown in Figure 5, we removed the last FC layer of VGG16 to decrease the parameters of original network as well. Layer conv1, conv2, conv3 of VGG16 were reserved and later layers were replaced by DTDM. In consideration of the resolution and feature representation of layers, we chose conv4_3 as a base feature map. Therefore, the conv4_3, conv7 and conv7_2 were combined via concatenation, and then the fusion feature map replaced original conv4_3 with the network once again. The corresponding feature size was 38 × 38, 19 × 19 and 10 × 10, respectively. In particular, the Dense Inception was placed after the fusion feature map from DTDM, con4_3 and conv7 to extract more effective features for prediction. In addition, the later layers were set the same as that in RFBNet.

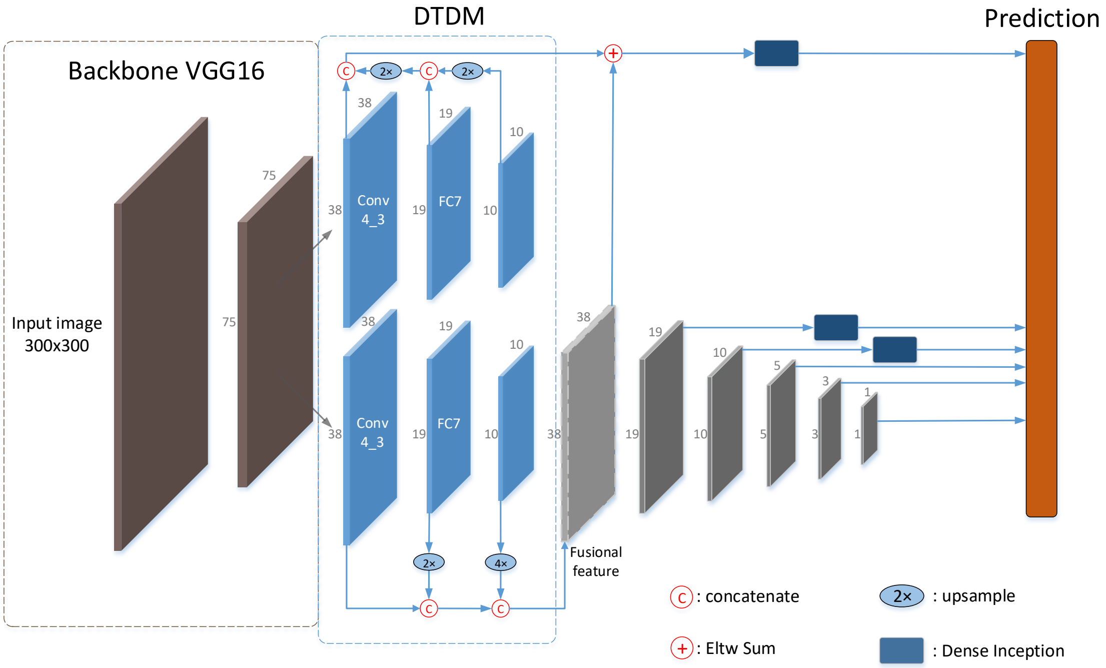


Figure 5. The pipeline of the whole detector.

3. Experiment Settings

3.1. Dataset

Remote sensing datasets differ from general datasets and possess their own characteristics. Multi-scale targets have a wide range: the size of objects varies from meters to hundreds of meters. Extremely uneven distribution of targets: the detected image may contain hundreds of objects or nothing valid.

Compared with other public optical remote image datasets, including UCAS-AOD [23], NWPU VHR-10 [24] and VEDAI [25], DOTA covers the most categories and the largest quantity scale. The other datasets have a lack of category or number of instances in Earth Vision, making themselves far from achieving simulation of the complicated realistic aerial scene. We conduct our experiment on the aerial images DOTA dataset. It contains 2806 aerial images and the size of images ranges from 800 × 800 pixels to 4000 × 4000 pixels. In detail, the dataset consists of 188,282 instances over 15 categories, including plane, ship, storage tank (ST), baseball diamond (BD), tennis court (TC), ground track field (GTF), harbor, bridge, large vehicle (LV), small vehicle (SV), helicopter (HC), roundabout (RA), swimming pool (SP), soccer ball field (SBF) and basketball court (BC).

The DOTA dataset has been divided into three subsets: training, validation and testing subsets, in which only the labels of training and validation subsets are public. Thus, we train and evaluate our detectors in the training and validation subset, respectively, and then submit our testing results on testing subsets to the public official website for comparing our detection performance with other published ones. In particular, it is difficult to directly train the network on the raw images with huge size. Therefore, the raw images in training and validation subsets are cropped at the size of 800 × 800 with the stride of 600.

3.2. Baseline

Considering the giant size of remote sensing images and some particular usage scenarios with real-time demand of detection, we preferentially choose a one-stage state-of-the-art detector with high detection speed. Therefore, we choose RFBNet as our baseline for its excellent balance between detection accuracy and computation. RFBNet is one of the best one-stage detectors which achieves the state-of-the-art performance among very deep detectors while keeping the real-time speed.

3.3. Evaluation Metric

For evaluation metrics, we adopt the same mean Average Precision (mAP) calculation as for PASCAL VOC. The most commonly used metric for evaluating the performance of detection algorithms is Average Precision (AP), derived from precision and recall. For a given annotated image, the output form a detector denotes $(b, l, p)_i$indexed by i, where b, l, and p separately represents predicted box location, label and confidence of target i. A predicted instance regarded as TP (True Positive) must meet the following conditions: the predicted label l is the same as the ground true label l_g; the iou (Intersection Over Union) between predicted bounding box b and the ground truth one b_g exceeds the preset threshold (commonly defined as 0.5):

$$IOU(b, b_g) = \frac{area(b \bigcap b_g)}{area(b \bigcup b_g)}. \tag{1}$$

Here, $area(b \bigcap b_g)$and $area(b \bigcup b_g)$ respectively denotes the intersection and union of the predicted and ground truth BBs. The other ones that do not meet above two conditions are regarded as False Positive (FP). In addition, if there are no more than one FP for one b_g, the b with max p will still be regarded as TP while the other ones will be regarded as FPs. For one special category, all relative TPs and FPs are counted from the whole testing images. Based on the TP and FP, the precision $P(\alpha)$ and recall $R(\alpha)$ are defined as follows:

$$P(\alpha) = \frac{TP}{TP + FP}, \tag{2}$$

$$R(\alpha) = \frac{TP}{TP + FN}, \tag{3}$$

where FN and α respectively denote the number of false negatives and the confidence threshold. The precision $P(\alpha)$ and recall $R(\alpha)$ can be determined as a function of the confidence threshold α. By computing the average value of Precision over Recall with different Recall value ranging from 0 to 1, the AP can be obtained. Finally, the mean AP (mAP) of all categories is defined as the final detection indicator.

3.4. Experiment Details

In our experiment, our training strategies mostly follow the RFBNet, including hard negative mining, data augmentation, some hyperparameters like default box scale and ratios. For better initialization for parameters, our model is trained from a pre-trained VGG model. In addition, all our training experiments are implemented on a Nvidia Tesla P100 GPU (Santa Clara, CA, USA) with batch size 16 for 240 k iterations. In particular, when compared with other published models, our detectors are evaluated on Nvidia 2080Ti GPU.

When testing our detector on a testing subset of the DOTA dataset, we directly set each raw image as our input. A sliding window of 800 × 800 scans the input image with a stride of 150 × 150 to generate a set of image patches. Then, the detector detects objects on each patch and outputs detecting results. The outputs are merged to make up the final results of raw whole image. The Non-maximum suppression (NMS) with IoU threshold is set as a 0.3 to filter overlapped boxes. In particular, we implement our detector based on the Pytorch framework and will open the code at https://github.com/pioneer2018/dtdm-di. For fair comparison, we preserve the most of original RFBNet.

4. Results

In this section, we set groups of ablation experiments to evaluate our detector and compare our detector with some state-of-the-art detectors. We first show the impact of each part of DTDM and evaluate the effectiveness of merged DTDM. Then, the evaluation of Dense Inception is presented. In particular, all ablation experiments are conducted via cropped val dataset, while the final detector is evaluated via official test dataset for comparison with state-of-the-art detectors. The results of DOTA dataset here are obtained by submitting our predictions to the official DOTA evaluation server. All results of experiments are shown and clarified in following subsections. Some visualization results of detection on DOTA dataset are shown in Figure 6. The figure shows that our model performs well on detecting multiple categories, especially in small and multi-scale objects in dense-targets scenes like ship, storage-tank and car.

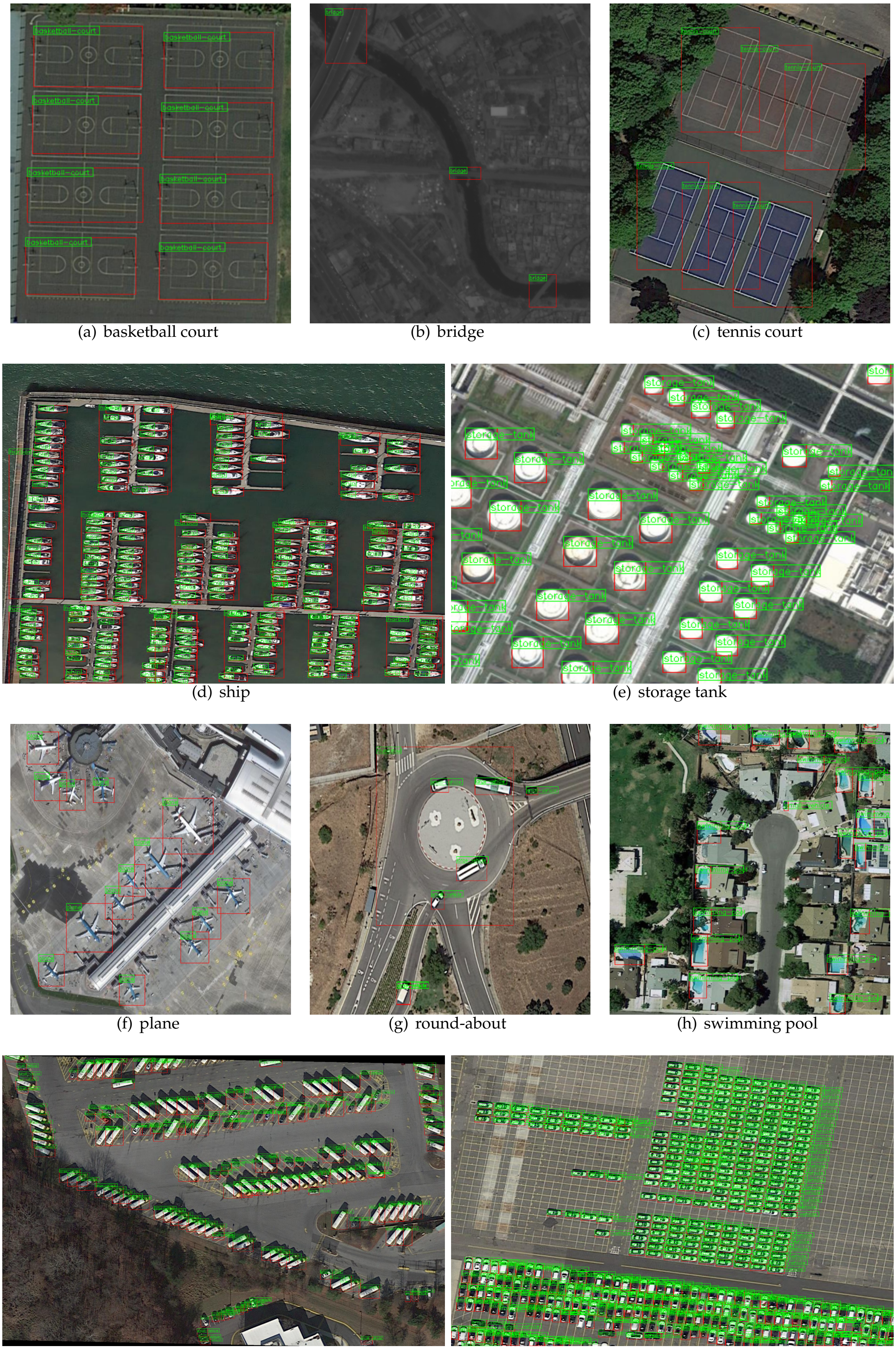
4.1. Ablation Experiments

4.1.1. DTDM

Since DTDM consists of two parts: we evaluate the impact of each part of DTDM. It can be shown in Table 1 that the baseline network RFBNet performs poorly on some categories like small object: SV (small vehicle), HC (helicopter), large object: GTF (ground track field), and various-scale object: bridge. When each part of DTDM is integrated into the model, respectively, there is obvious improvement on bridge, SV (small vehicle), and RA (Roundabout). In contrast, there is a much greater increase on HC (helicopter) for integration with part A, while a big improvement on bridge, RA (Roundabout) for integration with part B. When we apply merged DTDM into our model, we obtain an even better result than that of each part of DTDM. Finally, our model is 2.35% mAP better than a baseline network, which verifies the effectiveness of DTDM.

Table 1. Quantitative comparison of baseline and dual top-down module (DTDM) on the DOTA dataset. The images of subset are cropped as 800 × 800 size. The best result in each category is highlighted in bold.

baseline	val-clipped	plane	BD	bridge	GTF	SV	LV	ship	TC
		87.9	82.7	46.5	62.1	57.4	77.7	84.6	**90.9**
		BC	ST	SBF	RA	harbor	SP	HC	mAP
		76.9	54.8	61.2	54.8	76.8	48.2	50.3	**67.52**
baseline+ DTDM part A	val-clipped	plane	BD	bridge	GTF	SV	LV	ship	TC
		88.6	**82.9**	50.8	63.8	62.3	77.7	84.1	**90.9**
		BC	ST	SBF	RA	harbor	SP	HC	mAP
		75.3	55.5	60.3	60.8	79.3	48.9	53.8	**69.00**
baseline+ DTDM part B	val-clipped	plane	BD	bridge	GTF	SV	LV	ship	TC
		88.4	81.2	50.2	65.7	60.3	77.9	84.4	90.8
		BC	ST	SBF	RA	harbor	SP	HC	mAP
		74.8	55.9	59.9	66	80.4	53.2	51.1	**69.34**
baseline+ DTDM	val-clipped	plane	BD	bridge	GTF	SV	LV	ship	TC
		88.9	79.6	**56.1**	66.4	**63.1**	76.3	**85**	90.8
		BC	ST	SBF	RA	harbor	SP	HC	mAP
		74.9	56.3	59.5	**66.8**	80.8	**54.7**	49	**69.87**



(a) basketball court (b) bridge (c) tennis court

(d) ship (e) storage tank

(f) plane (g) round-about (h) swimming pool

(i) large car (j) small car

Figure 6. Some visualization results detected on the DOTA dataset. The red bounding boxes represent the location of objects and the green ones represent the label.

4.1.2. Dense Inception

Table 2 shows that there is obvious improvement among most categories. In particular, the improvement is concentrated mainly on large-scale objects (GTF) and changeable-scale ones (bridge, ST, RA). As has been analyzed in Section 2.2, atrous convolution causes some loss of local information and decreased correlation of adjacent pixels. We apply dense connection instead of larger atrous rate to atrous convolution. This structure yields a larger receptive field with more effective semantic information, which is helpful for detecting large-scale and changeable-scale objects. Furthermore, the integration of Dense Inception and DTDM leads to better performance at 70.37% mAP. Compared to a baseline network, there is a 2.87% mAP increase, while, compared to DTDM and RFBNet, there is a 0.5% mAP and 1.46% mAP increase, respectively.

Table 2. Quantitative comparison of baseline and Dense Inception on val subset of the DOTA dataset.The images of subset are cropped as 800 × 800 size. The best result in each category is highlighted in bold.

baseline	val-clipped	plane	BD	bridge	GTF	SV	LV	ship	TC
		87.9	**82.7**	46.5	62.1	57.4	77.7	**84.6**	**90.9**
		BC	ST	SBF	RA	harbor	SP	HC	mAP
		76.9	54.8	61.2	54.8	76.8	48.2	50.3	67.52
baseline+Dense Inception	val-clipped	plane	BD	bridge	GTF	SV	LV	ship	TC
		88.5	80.1	**49.2**	**66.3**	**59**	77.7	84.5	90.8
		BC	ST	SBF	RA	harbor	SP	HC	mAP
		74.9	**56.7**	**61.9**	**63.2**	**77.5**	**51.3**	**52.1**	**68.91**
baseline+DTDM+Dense Inception	val-clipped	plane	BD	bridge	GTF	SV	LV	ship	TC
		88.5	79.8	55.4	**67.6**	62.1	**78**	84.6	90.8
		BC	ST	SBF	RA	harbor	SP	HC	mAP
		75.5	**57.5**	**61.7**	65.3	**80.9**	52.2	**55.6**	**70.37**

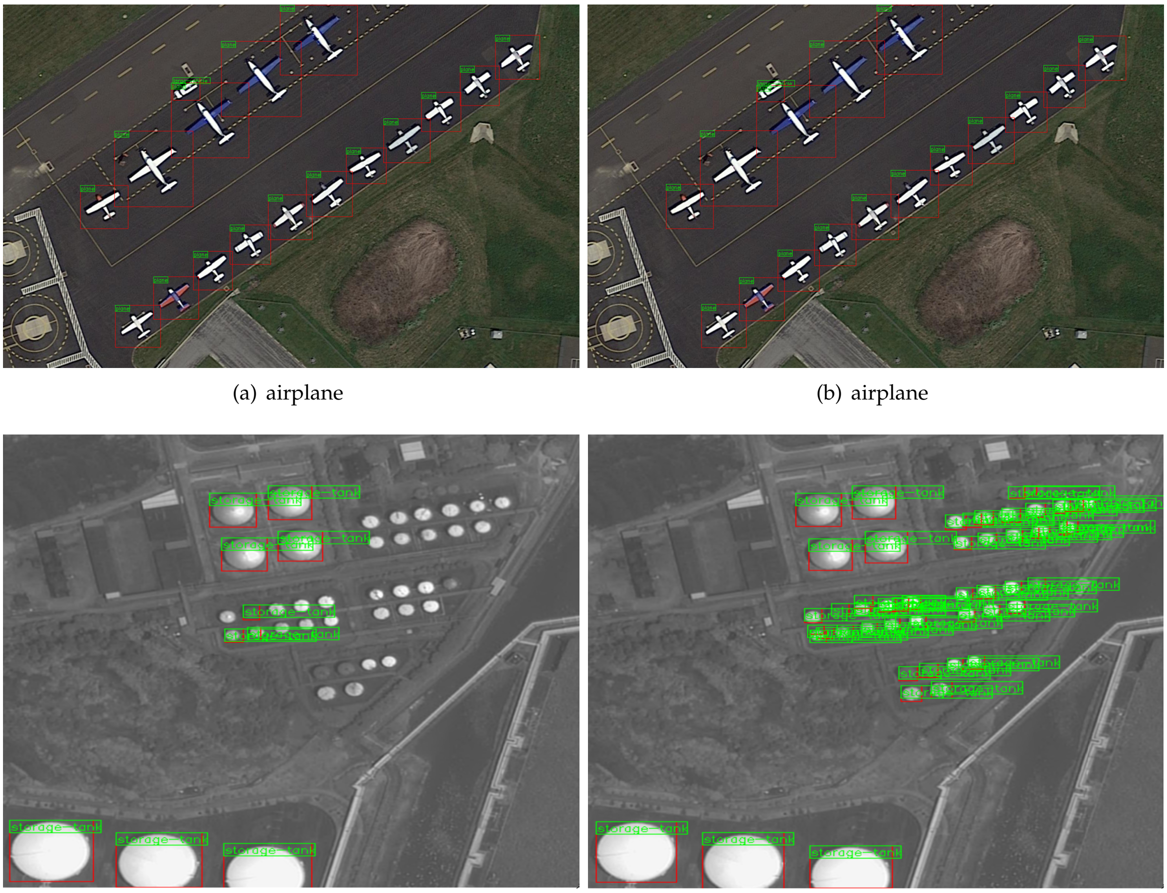
4.2. Comparison with the State-Of-The-Art

Besides the official baseline given by DOTA team, we also compare with You Only Look Once 9000 (YOLOv2) [26], Region-based Fully Convolutional Networks (R-FCN) [27], RFBNet, FPN, Image Cascade Network (ICN) [28], and IoU-Adaptive R-CNN [29], all of which once achieved excellent accuracy of detection. Table 3 presents detailed detection performance of those detectors of DOTA dataset both in detection accuracy and speed. We obtain first place among the published articles on DOTA at 73.49% mAP. In addition, our detector achieves best detection accuracy in about half of the categories. Compared with the baseline network, our model increases mAP by 2.66% for overall detection performance and we have achieved better results in all categories except BD. Compared with IoU-Adaptive R-CNN and ICN, which achieve best performance among all published methods, our detector gains better detection result. In the perspective of detection speed, our method gains much faster speed with 11 FPS than other two-stage detectors like IoU-Adaptive R-CNN with 5 FPS.

Figure 7 presents some visualization results of detection in different scenes where the figures of the left side and right side, respectively, represents the visualization of the baseline network and our model. It can be seen that both the baseline method and our method perform well in Figure 7a,b. Figure 7c–h present the fact that our method performs much better in detecting small-scale and changeable-scale objects. In addition, in target-dense scenes, our model can cover many more objects than the baseline network does.

Table 3. Comparative experiments of our method and some classic methods on the horizontal bounding boxes (HBB) task in original test subset of DOTA dataset. The best result in each category is highlighted in bold.

	SSD [11]	**YOLOv2 [26]**	**R-FCN [27]**	**RFBNet [13]**	**FPN [9]**	**ICN [28]**	**IoU-Adaptive R-CNN [29]**	**Ours**
plane	57.85	76.90	81.01	87.96	88.70	**89.97**	88.62	88.36
BD	32.79	33.87	58.96	84.46	75.10	77.71	80.22	**83.90**
bridge	16.14	22.73	31.64	38.50	52.60	**53.38**	53.18	45.78
GTF	18.67	34.88	58.97	66.31	59.20	**73.26**	66.97	67.24
SV	0.05	38.73	49.77	72.10	69.40	73.46	76.30	**76.80**
LV	36.93	32.02	45.04	73.31	**78.80**	65.02	72.59	77.15
ship	24.74	52.37	49.29	81.74	84.50	78.22	84.07	**85.35**
TC	81.16	61.65	68.99	90.64	90.60	90.79	90.66	**90.77**
BC	25.1	48.54	52.07	84.53	81.30	79.05	80.95	**85.55**
ST	47.47	33.91	67.42	64.69	**82.60**	84.81	76.24	75.77
SBF	11.22	29.27	41.83	55.94	52.50	**57.20**	57.12	54.64
RA	31.53	36.83	51.44	56.41	62.10	62.11	**66.65**	60.76
harbor	14.12	36.44	45.15	69.96	**76.60**	73.45	74.08	71.40
SP	9.09	38.26	53.3	75.23	66.30	70.22	66.36	**77.99**
HC	0.0	11.61	33.89	60.62	60.10	58.08	56.85	**60.94**
FPS	-	-	-	14	-	-	5	11
mAP	29.86	39.20	52.58	70.83	72.00	72.45	72.72	**73.49**



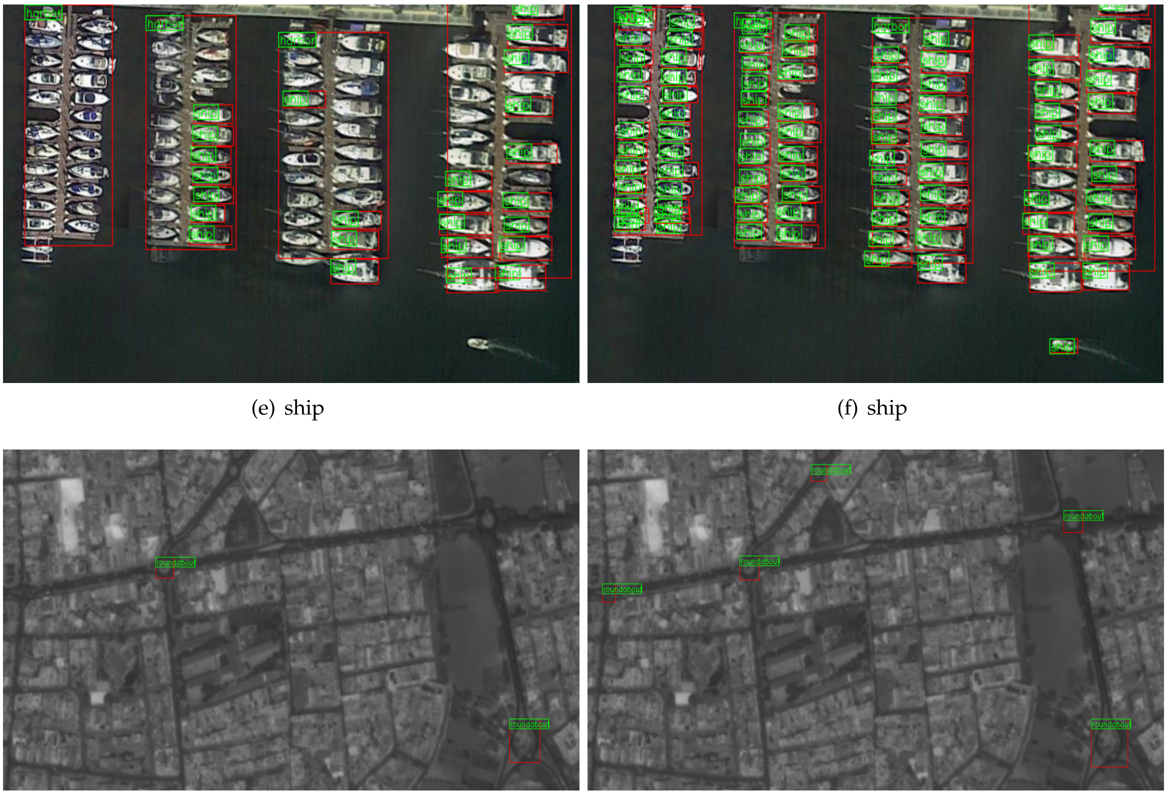
(a) airplane

(b) airplane

(c) storage tank

(d) storage tank

Figure 7. *Cont.*



(e) ship

(f) ship

(g) roundabout

(h) roundabout

Figure 7. Some results visualization detection. **The left side**: results of the baseline network. **The right side**: results of our model. The red bounding boxes represent the location of objects and the green ones represent the label.

5. Discussion

The above experiment results show that, compared with other published methods, our detector achieves higher mAP than other one-stage detectors and outperforms other two-stage detectors both in detection accuracy as well as speed on the DOTA dataset. From the perspective of both detection accuracy and speed, our detector achieves state-of-the-art detection performance. With the integration of DTDM and Dense Inception, we further enhance the baseline network, excellent one-stage detector RFBNet, by 2.66% mAP with a little loss in detection speed. In particular, DTDM enhances the ability of the feature representation of the backbone network to a large extent, which mainly improves the performance of detecting small and changeable-scale objects. In addition, Dense Inception further improves the detection accuracy by effectively increasing the receptive field, which is beneficial for detecting objects in complicated remote scenes. Compared to the work of Azimi et al. [28] and Yan et al. [29], which get the best detection performance among published methods, our method achieves even better detection accuracy and speed.

In the meantime, our detector performs relatively poorly in categories like bridge, GTF (ground track field), ST (storage tank) and RA (roundabout). We think there are two factors mainly resulting in this issue. On the one hand, Table 4 presents an extremely uneven number of the distribution of categories. There are numerous vehicles (SV, LV), but far less objects like GTF (ground track field), SBF (soccer ball field) and HC (helicopter). The sample lack of those categories makes it hard to fully train the model for detecting them. On the other hand, it is difficult to detect multi-scale objects that cover quite a wide range in size. Remote sensing images contain many special categories, which possess lots of large-scale objects, and these objects are easily cut to several slices when testing on cropped 800 × 800 images. Incomplete objects bring extra difficulty to our detection model.

Table 4. Labeled sample number distribution of all 15 categories from train and val subset of the DOTA dataset.

train + val DOTA subset	plane	BD	bridge	GTF	SV	LV	ship	TC
	10586	629	2511	469	31564	21356	37028	3127
	BC	ST	SBF	RA	harbor	SP	HC	All
	647	7917	479	578	8073	2176	703	127843

In addition, there are some misclassifications that cause some performance loss of detection. In Figure 8, the confusion matrix obtained by our model presents classification details among 15 categories and backgrounds. The misclassifications are located between some categories, such as GTF (ground track field) and SBF (soccer ball field), HC (helicopter) and plane, especially between objects and bg (background). The improvement of classification performance will be considered in the future.

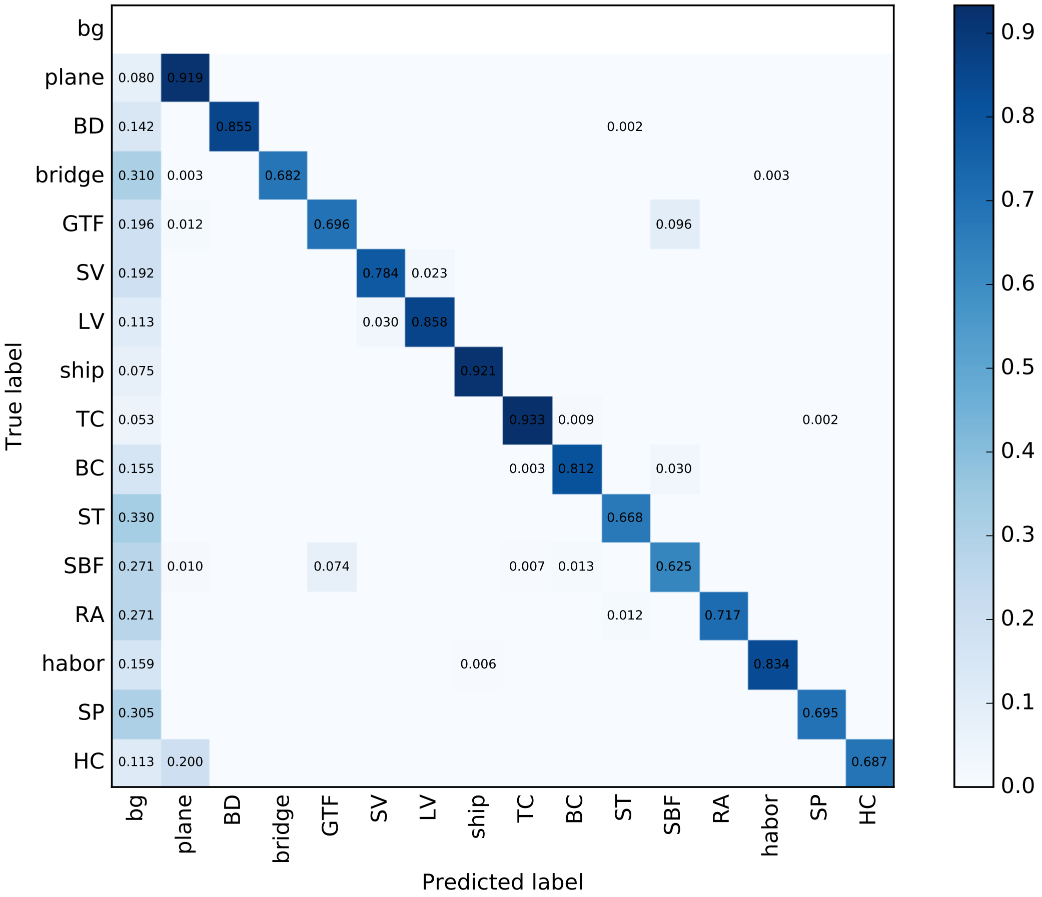


Figure 8. Confusion matrix obtained by our model on the DOTA validation subset.

In future study, we will change our loss function in response to an unbalanced distribution of samples and misclassifications among categories. The adaptive weight of each category will be added to the loss function to adapt to particular categories with a lack of samples. The strategy to learn semantic similarity directly from deep features [30] may be useful to decrease misclassifications. In addition, we will try to introduce an image pyramid into data preprocessing before training to better detect multi-scale objects. In addition, it has been proved that adding extra prior information to the model is the key to help improve the achievement of some tasks [31,32]. We will attempt to redesign our model to utilize other related prior information to enhance the detection.

6. Conclusions

In this paper, we propose a powerful one-stage object detector. In order to improve the performance while maintaining the high detection speed, we choose the lightweight network VGG16 as our backbone network. However, the performance of a lightweight network on feature representation remains to be enhanced and it is hard to adapt to the characteristics of remote sensing images. In order to improve the performance of detectors in remote sensing images, especially on detecting a small and changeable-scale, we choose to handle the problem by multi-layer feature fusion. In the meantime, a deeper network means better capability of feature representation. Furthermore, we introduce Dense Inception to deepen and widen the network with little computation cost. From all of the experiment results, both methods contribute a lot to the improvement of detecting performance. On the one hand, DTDM integrates multiple features from multiple layers via two branches. Each branch plays a different role in the fusion of features and the experiment shows that the impact of two branches are complementary for improving detection accuracy. On the other hand, Dense Inception adopts atrous convolution with an increasing atrous rate. The structure of dense connection yields a larger receptive field and avoids much loss of information, which brings an improvement in detecting multi-scale and changeable-shape categories. For further work, we plan to apply our model to other lightweight backbone networks like MobileNet-SSD to achieve more real-time detection speed.

Author Contributions: Formal analysis, Z.C.; Funding acquisition, K.F. and Y.Z. Investigation, Z.C.; Supervision, F.K., Y.Z. and X.S.; Resources, Y.Z. and X.S.; Visualization, Z.C.; Writing—original draft, Z.C.; Writing—review and editing, F.K., Z.C., Y.Z. and X.S.

Funding: This research was supported by the National Natural Science Foundation of China under Grants 61725105 and Grants 41801349. In addition, this research was supported by the Gusu Innovation Talent Foundation of Suzhou under Grant ZXT2017002.

Acknowledgments: The authors would like to thank all colleagues in the lab, who are willing to give useful opinions. The authors would like to express their sincere appreciation for the reviewers for their helpful comments and suggestions.

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

References

1. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
2. 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, 3–8 December 2012; pp. 1097–1105.
3. Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Chen, J.; Liu, X.; Pietikäinen, M. Deep learning for generic object detection: A survey. *arXiv* **2018**, arXiv:1809.02165.
4. Fu, K.; Zhang, T.; Zhang, Y.; Yan, M.; Chang, Z.; Zhang, Z.; Sun, X. Meta-SSD: Towards Fast Adaptation for Few-Shot Object Detection With Meta-Learning. *IEEE Access* **2019**, *7*, 77597–77606. [CrossRef]
5. Uijlings, J.R.R.; Sande, K.E.A.V.D.; Gevers, T.; Smeulders, A.W.M. Selective Search for Object Recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
6. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
7. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
8. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
9. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.

10. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
11. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.
12. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 21–26 July 2017; pp. 2980–2988.
13. Liu, S.; Huang, D. Receptive field block net for accurate and fast object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany; 8–14 September 2018; pp. 385–400.
14. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3974–3983.
15. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
16. 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. 4700–4708.
17. Shrivastava, A.; Sukthankar, R.; Malik, J.; Gupta, A. Beyond skip connections: Top-down modulation for object detection. *arXiv* **2016**, arXiv:1612.06851.
18. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556 .
19. 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 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
20. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.
21. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the Thirty-First, AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
22. Wang, P.; Chen, P.; Yuan, Y.; Liu, D.; Huang, Z.; Hou, X.; Cottrell, G. Understanding convolution for semantic segmentation. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1451–1460.
23. Zhu, H.; Chen, X.; Dai, W.; Fu, K.; Ye, Q.; Jiao, J. Orientation robust object detection in aerial images using deep convolutional neural network. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 27–30 September 2015; pp. 3735–3739.
24. 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]
25. Razakarivony, S.; Jurie, F. Vehicle detection in aerial imagery: A small target detection benchmark. *J. Vis. Commun. Image Represent.* **2016**, *34*, 187–203. [CrossRef]
26. 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. 7263–7271.
27. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 379–387.
28. Azimi, S.M.; Vig, E.; Bahmanyar, R.; Körner, M.; Reinartz, P. Towards multi-class object detection in unconstrained remote sensing imagery. *arXiv* **2018**, arXiv:1807.02700.
29. Yan, J.; Wang, H.; Yan, M.; Diao, W.; Sun, X.; Li, H. IoU-Adaptive Deformable R-CNN: Make Full Use of IoU for Multi-Class Object Detection in Remote Sensing Imagery. *Remote Sens.* **2019**, *11*, 286. [CrossRef]
30. Deng, C.; Yang, E.; Liu, T.; Li, J.; Liu, W.; Tao, D. Unsupervised semantic-preserving adversarial hashing for image search. *IEEE Trans. Image Process.* **2019**, *28*, 4032–4044. [CrossRef]

31. Peng, X.; Feng, J.; Xiao, S.; Yau, W.Y.; Zhou, J.T.; Yang, S. Structured autoencoders for subspace clustering. *IEEE Trans. Image Process.* **2018**, *27*, 5076–5086. [CrossRef]
32. Wang, Q.; Gao, J.; Yuan, Y. Embedding structured contour and location prior in siamesed fully convolutional networks for road detection. *IEEE Trans. Intell. Transp. Syst.* **2017**, *19*, 230–241. [CrossRef]