Dynamic Pyramid Attention Networks for multi-orientation object detection

Hongchun Yuan[1,2*], Hui Zhou[1,2], Zhenyu Cai[1,2], Shuo Zhang[1,2], Ruoyou Wu[1,2]
[1]College of Information Technology, Shanghai Ocean University, China
[2]Key Laboratory of Fisheries Information, Ministry of Agriculture, China
hcyuan@shou.edu.cn, 506014862@qq.com, czy96@outlook.com, 1042785279@qq.com, ternencewu@163.com

Abstract

The objects in remote sensing images often appeared in any direction, and thus multi-orientation object detection has received considerable attention. However, most existing oriented object detection methods rely on increasing the network layers, which wastes many computing resources while only bringing a slight improvement. We find that a few pixels around the convolution kernel participate in the calculation when extracting image features in the convolution network. If we can incorporate the global information into the feature map, the model's performance will be significantly improved. In this paper, we proposed the dynamic pyramid attention network (DPANet) for remote sensing images, which consists of the self-attention feature pyramid network (SAFPN) and the dynamic feature map selection module (DFMS). The SAFPN employs the self-attention mechanism to learn the correlation between each pixel value and the global pixel in different feature layers by downsampling the upper feature layers to the lower one. Furthermore, the DFMS module dynamically selects feature maps to further expand the receptive field by weighing the effectiveness of different feature layers and reducing the interference of unnecessary feature maps. The remote sensing datasets HRSC2016, UCAS_AOD and NWPU_VHR are used to evaluate the performance of DPANet and the experiment results show that the proposed network outperforms the benchmark models significantly.

Keywords: Oriented object detection, Remote sensing images, Self-attention, Feature layer selection

1 Introduction

Object detection is a fundamental and difficult challenge of remote sensing research and receives a lot of interest due to the development of deep convolution neural networks. The performance of object detection has been dramatically improved in recent years [1-3], particularly in terms of speed and accuracy. Most object detection methods can be divided into two major types. The first is Faster-RCNN [4], which represents the two-stage object detection methods. It focuses on the extraction of region proposals and fine-tunes the location of the detection bounding box. The single-stage detection networks, such as YOLO（You Look Only Once）[5], SSD (Single Shot MultiBox Detector) [6], and RetinaNet [7] are the second type. Specifically, YOLO completes feature extraction, object classification, and coordinate regression simultaneously, which makes it faster but less accurate than the two-stage network. However, methods such as Faster-RCNN cannot be directly applied to multi-orientation object detection because the orientation will cause the objects and detection bounding boxes to be misaligned. As shown in Figure 1, the horizontal bounding box contains complex background and loses the shape information, whereas the oriented bounding box can preserve the target's actual size. The adjacent horizontal bounding boxes typically have considerable overlap, and the bounding box with lower confidence will be easily suppressed during the process of non-maximum suppression, resulting in missed detection.

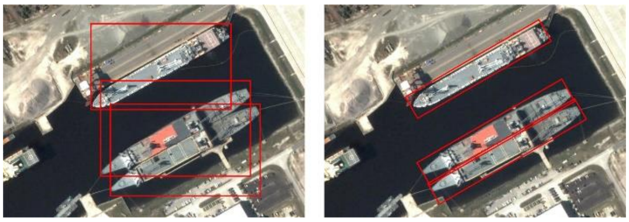


Figure 1. Comparison of HBBs and OBBs. For adjacent ships, HBBs in image (a) have a larger overlap area than OBBs in image (b), and boxes with lower confidence scores will be suppressed.

To address the shortcomings of horizontal bounding boxes, many researchers have proposed oriented bounding boxes methods for object detection in remote sensing images. Jiang et al. [8] detected the text content by incorporating anchor and multi-scale ROI-Pooling, and adding the loss of angle regression into the loss function (R2CNN). Ma et al. [9] proposed a region extraction-based method that combines RROI and ROI learning to improve the performance of rotated text detection (RRPN). Yang et al. [10] proposed a refined one-stage rotation detector (R3Det), which combines the advantages of horizontal anchor's high recall rate and rotation anchor's adaptability to dense scenes. Zhang et al. [11] designed a spatial and scale-aware attention module that makes the network focus on the areas with more information at an appropriate feature scale and suppresses irrelevant information simultaneously. Zhou et al. [12] proposed replacing the traditional five-coordinate method with the polar coordinate system and regressing the coordinates using a new loss function. A feature fusion structure designed by X. Yang et al. [13] can solve the small target problem from the perspective of feature fusion and anchor sampling (SCR-Det). They proposed a supervised multi-dimensional attention network that could mitigate the effects of background noise and solve the dense permutation problem.

In the object detection network, the attention mechanism in the feature extraction part makes the deep neural network to learn the areas of interest in each image. According to

*Corresponding Author: Hongchun Yuan; E-mail: hcyuan@shou.edu.cn
DOI: 10.53106/160792642022012301008

previous research, there are primarily three attention mechanisms. The first one is spatial attention, as illustrated in Figure 2. The Spatial Transformer Network proposed by Jaderberg et al. [14] learns a set of radical transformation parameters, allowing the input image to adjust the object in the image to a more appropriate position, completing the preprocessing work suitable for various tasks. The second one is channel attention, as illustrated in Figure 3. SENet proposed by J. Hu et al. [15] focuses on the channel features with the most information and suppresses the unimportant ones. The third mechanism is the attention mechanism that combines channel and space. S. Woo et al. [16] proposed a CBAM module that combines space and channel attention to improve performance. As a result, the application of the attention mechanism for feature extraction yielded promising results. In the past two years, the proposed Transformer [17], particularly the self-attention module, has achieved excellent results in the area of Natural language processing (NLP) and has a wide range of application in the field of computer vision.

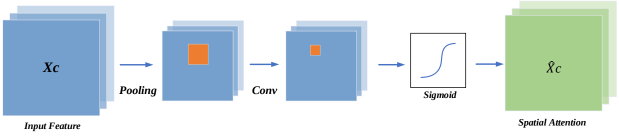


Figure 2. The structure of Spatial attention Module

The self-attention mechanism has been first used in the field of NLP. As shown in Figure 4, it is an attention mechanism with different positions of a single sequence used to calculate the interactive representation of sequences. In machine reading, Cheng et al. [18] utilize the self-attention mechanism that allows the network to learn the relationship between the current word and the previous part of the sentence. Xu et al. [19] describe the generation task in the image with the self-attention module. The attention weights visualization clearly shows which areas of the image the model is concerned with when output the results. Nowadays, an increasing number of researchers are focusing on the application of self-attention in other fields. Carion et al. [20] proposed a set-based global loss function and use Transformer to achieve end-to-end object detection. It uses binary matching and a transformer encoder-decoder architecture to compute unique predictions. A small set of fixed learning object queries is provided and the relationship between the target object and the global image context is then considered, with the final prediction set being output directly and in parallel. Wang et al. [21] divided a two-dimensional self-attention module into two one-dimensional self-attention modules. Simultaneously, it establishes an attention module with a global sense field while reducing the calculation amount. It adds a position-sensitive attention layer, allowing it to use position information better and become an instance segmentation backbone network. According to the findings of the preceding studies, self-attention structure plays a critical role in learning global information. In multi-orientation object detection networks, it is difficult to accurately regress the angle of bounding boxes. Self-attention can help the network obtain richer structural characteristics of the objects, and the model can learn the correlation of global information and regress the coordinate information better.

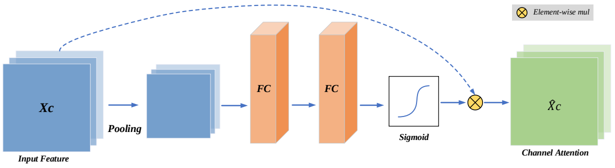


Figure 3. The structure of Channel attention Module

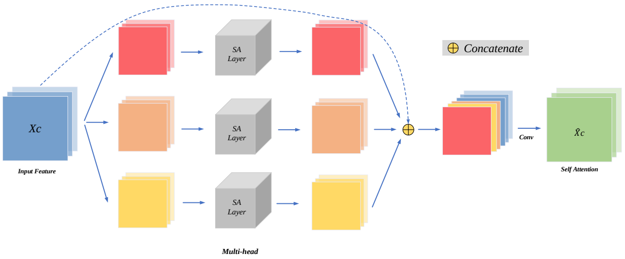


Figure 4. The structure of Self-attention Module

In this work, we propose the Dynamic Pyramid Attention network as a brand new and effective multi-scale feature extraction network to improve the global feature learning problem for multi-orientation object detection. The self-attention module for learning global information embedded on the multi-scale feature network FPN is introduced in the area of remote sensing for oriented object detection. In the second part, the DPANet further enhances network learning through the self-attention module. Simultaneously, it dynamically selects feature maps to reduce the unnecessary interference of low-quality feature maps and reduces the calculation of the model. The experience results demonstrate that our method can achieve consistent and significant improvements in the detection of multi-directional objects with the experiments on various datasets, including the remote sensing dataset HRSC2016 and UCAS_AOD. Experiments on the dataset NWPU_VHR also show that our method is universal for object detection with the horizontal bounding box. Our proposed method is simple to incorporate into existing object detection pipelines.

Our primary contribution includes the introduction of self-attention to orientation object detection and a method for evaluating the efficiency of feature layers. We analyze the distribution of positive proposals from the different feature layers and highlight the different results brought by different distributions. In order to reduce feature dilation in the top-down pathway, we proposed the SAFPN module to expand and change the receptive field of the feature map to the global receptive field. It is able to increase the information representation ability of feature maps with different sizes and improve the whole model's high-quality feature extraction ability.

2 DPA Network

The architecture of our method is shown in Figure 5. The DPANet extends the framework of FPN in RetinaNet and consists of two stages: dynamic feature map selection (referred to as DFMS) and self-attention FPN (referred to as SAFPN) to generate proposals and refine the outputs. We dynamically evaluate the quality of feature maps generated from FPN in the DFMS stage and fine-tune the high-quality ones by inputting them into the self-attention module. The low semantic feature map and high semantic feature map obtained from the backbone network are convoluted to the same scale and used for self-attention calculation in the SAFPN stage.

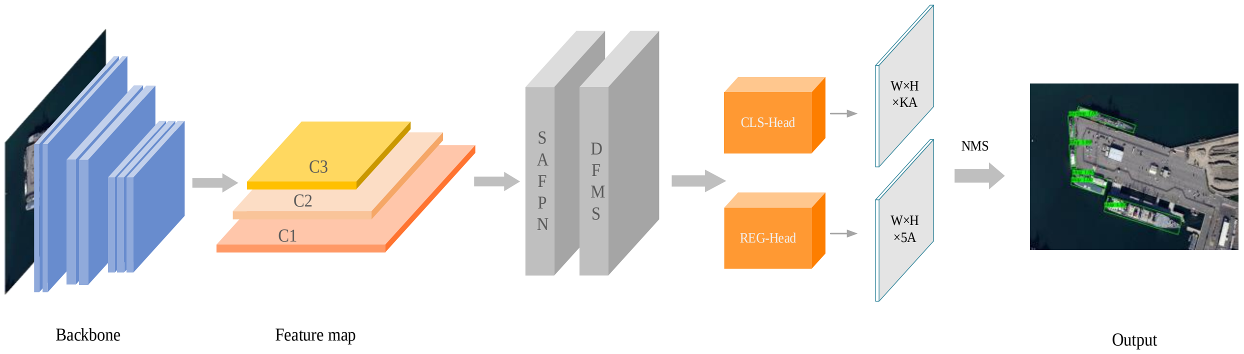


Figure 5. Architecture of the proposed method (using RetinaNet as an embodiment). 'C1', 'C2', 'C3' are the feature maps extracted from ResNet's layer2, layer3 and layer4. The SAFPN module and DFMS module enhance the learning of global features through the self-attention module to improve the generation of positive proposals.

2.1 Multi-orientation Detector on RetinaNet

In actual application scenarios, orientation object detection pays more attention to the real-time scene. Therefore, we use RetinaNet, a single-stage object detection network, as the backbone of our model. It uses ResNet [22] as the feature extraction network and an architecture similar to FPN to construct the multi-scale feature pyramid. Predefined horizontal anchors are set at levels P3, P4, P5, P6, P7. We denote the bounding box with (x, y, w, h, θ) due to the additional angle parameters. For bounding box regression, we have:

$$t_x = \frac{x - x_a}{w_a}, t_y = \frac{y - y_a}{h_a} \tag{1}$$

$$t_w = log\frac{w}{w_a}, t_h = log\frac{h}{h_a}, t_\theta = \tan(\theta - \theta_a) \tag{2}$$

where the x, y, w, h, θ are the center coordinates, width, height and angle. x_a denotes the anchor value. Given ground truth offset $t^* = (t_x^*,\ t_y^*,\ t_w^*,\ t_h^*,\ t_\theta^*)$. The definition of the multitask loss formula is as follows:

$$L = L_{cls}\big(\,p, p^* + L_{reg}(t, t^*)\,\big) \tag{3}$$

in which p denotes the score of prediction classification, t denotes the offset of prediction box, and p^* denotes the class label of anchor ($p^* = 1$ denotes positive sample, $p^* = 0$ denotes negative sample).

2.2 Dynamic Feature Map Selection

It is well known that the low-level feature map contains less semantic information but an accurate target location. In contrast, the high-level feature map contains more semantic information but a rough target location. After extracting features, most object detection models directly input the feature maps to the decoder to output proposals for classification and regression. However, the quality of generated proposals is frequently deficient, complicating the training of the subsequent network. In this part, the self-attention module is used to improve feature extraction capabilities, and experiments show that feature extraction capabilities for small feature maps via self-attention are insufficient. In the process of training, the network spatially matches the preset anchor and the real-labeled frame based on IoU and outputs an appropriate number of proposals as positive samples for regression by limiting a certain threshold (such as IoU>0.5). Figure 6 shows that the proportion of positive suggestions generated by different layers of feature maps in different datasets is different and unbalanced. We could better weigh the ability to vary scale feature maps to generate positive sample proposals through designing the efficiency φ of proposals. By dynamically evaluating the effectiveness of different feature layers and performing self-attention operations on them separately, we can get better results than by operating directly on all feature maps. φ can be calculated as follows.

$$\varphi_i = \frac{\omega_i}{\sum_k^n \omega_k}, \qquad \omega = \frac{x_i}{y_i} \tag{4}$$

where x is the number of positive proposals in the i -th feature layer and y is the total number of proposals generated in the i -th feature layer.

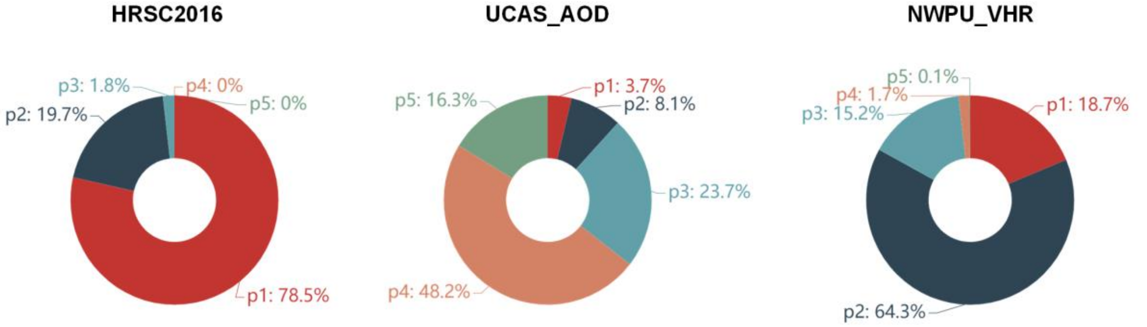


Figure 6. The proportion of positive proposals generated from each feature layer in datasets HRSC2016, UCAS_AOD and NWPU_VHR. The proportion of feature layers in different datasets varies due to the different size of target objects in different datasets. The main feature layers in HRSC2016 are p1, p2 and p3, the main feature layers in UCAS_AOD are p3, p4 and p5, and the main feature layers in NWPU_VHR are p1, p2 and p3.

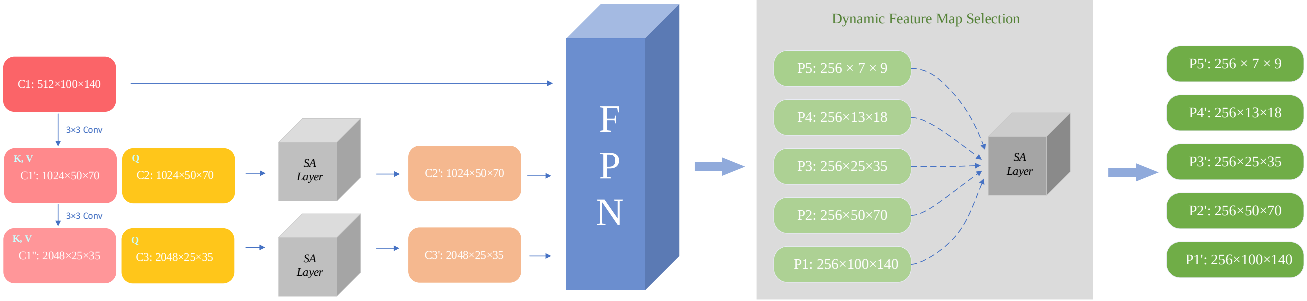


Figure 7. Pipeline of SAFPN and DFMS. The feature maps C1, C2 and C3 are produced by layer 2, layer 3 and layer 4 generated from ResNet. We downsample C1 to get the same size as the high semantic information feature maps C2 and C3, which we call C1' and C1". The feature map C1' serves as Key and Value and C2 serves as Query. Then, we added them into the self-attention module. Next we repeat the process for feature map C1". The new fusion feature map is acquired and loaded into the FPN. The feature layer selected by the DFMS module is used as K, V and Q, respectively calculated in the self-attention module.

2.3 Self-Attention FPN

The FPN baseline can only generate rough feature maps without any further attention modules [23]. In FPN architecture, there are two unavoidable issues. The first is when high-level information is gradually diminished as it is integrated into different feature levels of up-down paths [24]. The other issue is that extraneous information may influence this architecture and lower the model's overall performance [25]. To address these two shortcomings of baseline FPN, we proposed a novel self-attention feature pyramid network (SAFPN). The goal of SAFPN is to improve high-level semantic features and then transmit the improved semantic information down to lower feature levels. Context information can be better utilized when using other feature pixels in the feature map to enhance the representation of target pixels. The proposed architecture is based on a traditional feature pyramid network (FPN) [26] and uses ResNet as the backbone. The basic FPN architecture is well known for its widespread application in a wide range of computer vision tasks, particularly detection tasks, where detection results may be more accurate due to the durability and rationality of its structure. As shown in Figure 7, we keep the basic frame but add two practical modules to improve performance. The self-attention module is composed of three steps: similarity calculation, softmax, and weighted average. The following are the calculation formulas.

$$Q = conv1(x_1), K = conv1(x_2), V = conv3(x_3) \tag{5}$$

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}V}\right) \tag{6}$$

$$output = \gamma * Attention(Q,K,V\ ) + x_3 \tag{7}$$

where x_1, x_2, x_3 are the input feature maps. Q and K denote the query and key. Query can be thought of as the concatenation of M vectors with dimension d, and Key as the concatenation of N vectors with dimension d, γ is a hyper parameter.

3 Experiments

3.1 Datesets

In this section, we evaluate the proposed method with quite a few comparative experiments on remote sensing image datasets, known as the HRSC2016, UCAS_AOD and NWPU_VHR datasets.

HRSC2016 is a challenging dataset in the field of ship detection. It contains 1061 labelled images. The image sizes range from 300×300 to 1500×900 pixels. It is divided into the training set, validation set and test set, each with 436, 181 and 444 images.

UCAS_AOD is a dataset for detecting aerial aircraft and cars. The aircraft dataset contains 1000 images of 3210 aircraft, and the vehicle dataset contains 510 images of 2819 vehicles. All images are carefully chosen to ensure that the target direction is distributed evenly throughout the dataset. We randomly divide it into the training set, validation set and test set as 2:1:1.

NWPU_VHR includes 10 different geospatial object classes. It contains 757 airplanes, 159 basketball courts, 302 ships, 390 baseball squares, 524 tennis courts, 124 bridges, 655 storage tanks, 477 cars, 163 track and field grounds and 224 ports in the dataset. It is made up of 715 RGB images with spatial resolutions, ranging from 0.5m to 2m collected from Google Earth, as were 85 pan-sharpened infrared images with spatial resolutions of 0.08m.

3.2 Evaluation Metrics

We use the precision-recall curve (PRC) and average precision (AP) as the evaluation metrics, which are widely applied to measure the performance of models.

The PRC is a measure of accuracy at different recall rates. The precision and recall are calculated by the number of true positive (referred to as TP), the number of false positive (referred to as FP), and the number of false negative (referred to as FN). The precision and recall are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

where the ratio of overlap area between the predicted box and ground truth box over 0.5, we recognized the box as TP; otherwise, it is FP.

The AP metric is the value of the area under the PRC, and the mean average precision (mAP) is the mean AP value of all object classes. The higher the value of mAP, the better performance of the model. They are defined as the following formulas:

$$AP = \int_0^1 P(R)dR \times 100\% \quad \textbf{(10)}$$

$$mAP = \frac{1}{N_{cls}} \sum_{i=1}^{N_{cls}} AP_i \quad \textbf{(11)}$$

where $P(R)$ denotes the $P\text{-}R$ function, N_{cls} represents the number of categories.

3.3 Implementation details

The baseline network that we use is DAL [27], which uses RetinaNet as the backbone network and adds a matching matrix to balance the positive and negative samples. In the dataset HRSC2016, UCAS_AOD and NWPU_VHR, the ratio of anchor we use are (0.5, 1, 2). All the pictures are resized to 800×800 in the data preprocessing part. The data is enhanced by randomly flipping and rotating the images to boost the robustness of the model. The optimizer we use in the network is Adam, and the learning rate is set to 0.0001. We trained our model on an NVIDIA GTX1080Ti for 100 epochs with a batch size of 2. All the experimental results are evaluated under PASCAL VOC 2012 metric.

3.4 Ablation study

3.4.1 Effect of DFMS

To illustrate the efficacy of our method, we conducted a series of comparative experiments. We find that the different choices of feature layers, which are selected by DFMS and sent to the self-attention module, will produce different results. The feature layer with a higher proportion will outperform the feature layer with a lower proportion. Table 1, Table 2, and Table 3 show that the selection of feature maps improves the mAP significantly compared to the baseline method. The PRC of various methods in datasets HRSC2016, UCAS_AOD and NWPU_VHR are depicted in Figure 8, Figure 9 and Figure 11.

Table 1. Detection results of different layers in DFMS module in dataset HRSC2016

Methods	Layers	mAP
Baseline	-	88.60
Baseline+SAFPN	-	90.19
Baseline+SAFPN(DFMS)	p2	90.42
Baseline+SAFPN(DFMS)	p2+p3	**91.52**
Baseline+SAFPN(DFMS)	p4+p5	90.90
Baseline+SAFPN(DFMS)	p2+p3+p4+p5	91.39

Table 2. Detection results of different layers in DFMS module in dataset UCAS_AOD

Methods	Layers	mAP
Baseline	-	89.87
Baseline+SAFPN	-	92.24
Baseline+SAFPN(DFMS)	p2+p3	92.74
Baseline+SAFPN(DFMS)	p4+p5	92.22
Baseline+SAFPN(DFMS)	p2+p3+p4	**93.36**
Baseline+SAFPN(DFMS)	p2+p3+p4+p5	92.26

Table 3. Performance evaluation of HBB task on NWPU_VHR

Methods	Layers	mAP
Baseline	-	88.30
Baseline+SAFPN	-	89.42
Baseline+SAFPN(DFMS)	p2	89.50
Baseline+SAFPN(DFMS)	p2+p3	89.36
Baseline+SAFPN(DFMS)	p2+p3+p4	**90.74**
Baseline+SAFPN(DFMS)	p2+p3+p4+p5	89.72

3.4.2 Effect of SAFPN

SAFPN is a brand-new designed structure that enhances the semantic representation of each feature layer by using low-level feature maps. Due to the intervention of feature maps with high semantic information, combined with the self-attentive module, the confidence level of the proposals output from each feature layer is increased. Thus, the number of positive proposals output from each feature layer is increased, which provides a good guarantee for coordinate position regression. Figure 10 shows that positive proposals from HRSC2016 increased by 4.6%, UCAS_AOD increased by 17.5%, and NWPU_VHR increased by 5.3%.

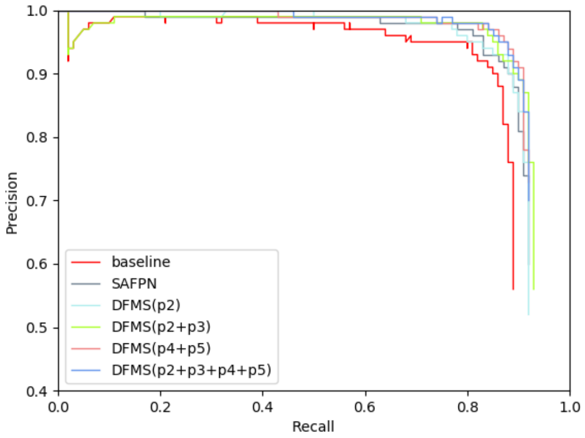


Figure 8. Precision-recall curves of test results by different methods in HRSC2016

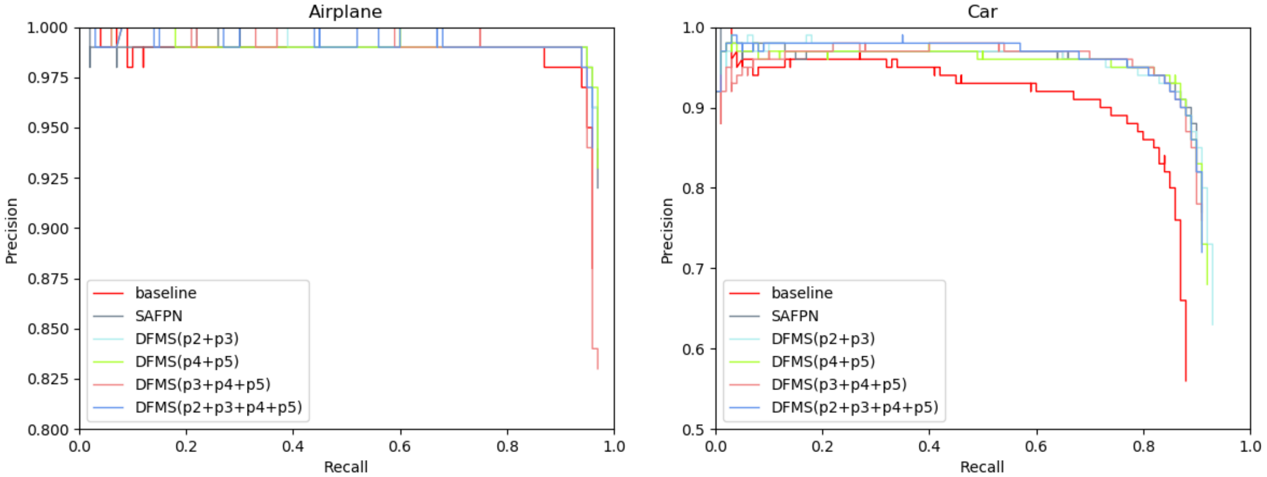


Figure 9. Precision-re call curves of different methods for each category in UCAS_AOD

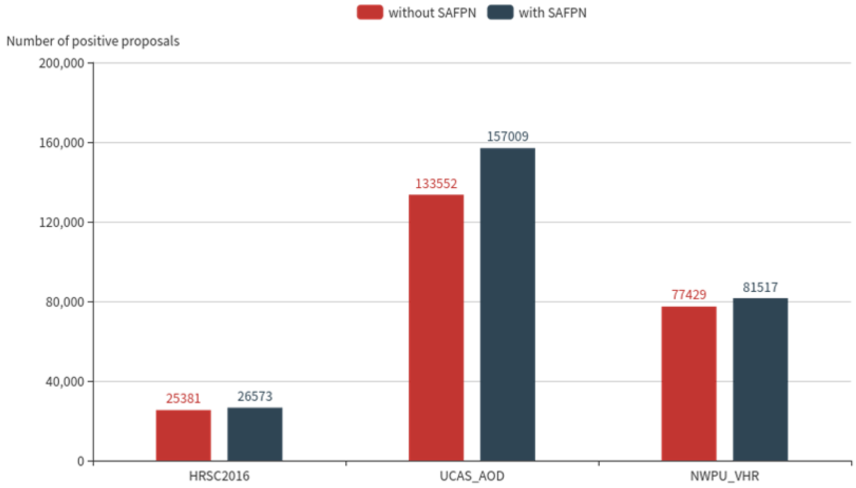


Figure 10. Comparison of the number of positive proposals among the three dataset

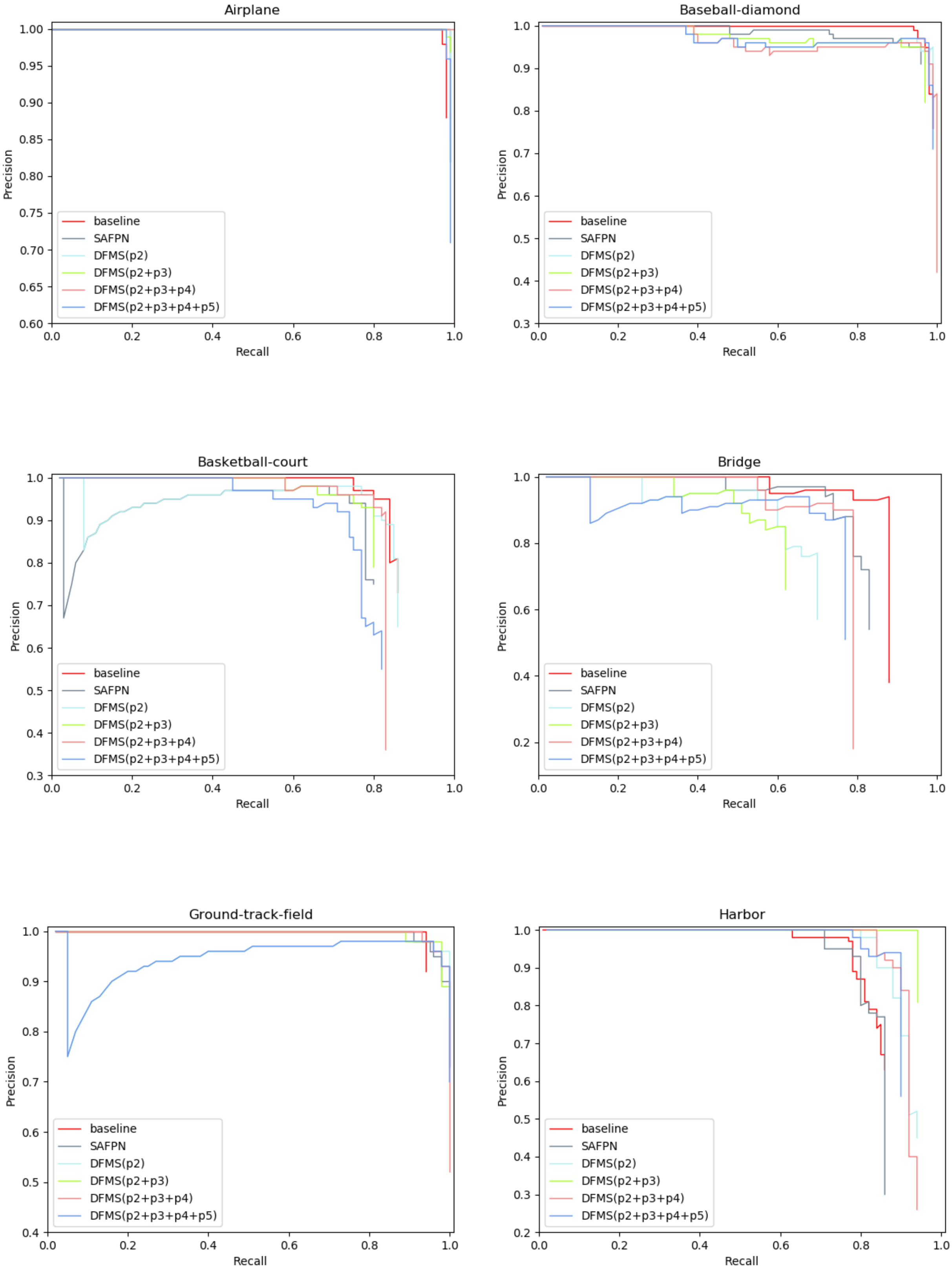
4 Results

4.1 Results on HRSC2016

The objects in HRSC2016 are with high aspect ratios and multiple directions. Table 4. shows an experiment comparison of our method with other methods. The comparison shows that our method achieves the most advanced performance. We use ResNet-50 as the backbone network and resize the input image to 800×800. Furthermore, our model is a single-stage detection network, and its performance is not inferior to some two-stage detection networks. The comparison results of the dataset are shown in Figure 12.

Table 4. Comparisons with state-of-the-art detectors on HRSC2016

Methods	Backbone	Size	mAP
Two-stage			
R2CNN [8]	ResNet101	800×800	90.19
RRPN [9]	ResNet101	800×800	90.42
RoI Trans [28]	ResNet101	800×800	91.52
Gliding Vertex [29]	ResNet101	512×800	90.90
CSL [30]	ResNet101	-	91.39
One-stage			
RRD [31]	VGG16	384×384	84.30
ROPDet [32]	ResNet101	800×800	89.16
GWD [33]	ResNet101	512×512	89.85
Baseline	ResNet50	800×800	88.60
DPANet(Ours)	ResNet50	800×800	**91.52**


Airplane
Precision
Recall
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.0
0.2
0.4
0.6
0.8
1.0
Baseball-diamond
Basketball-court
Bridge
Ground-track-field
Harbor
0.3
0.4
0.5
0.6
0.7
0.8
0.9
baseline
SAFPN
DFMS(p2)
DFMS(p2+p3)
DFMS(p2+p3+p4)
DFMS(p2+p3+p4+p5)

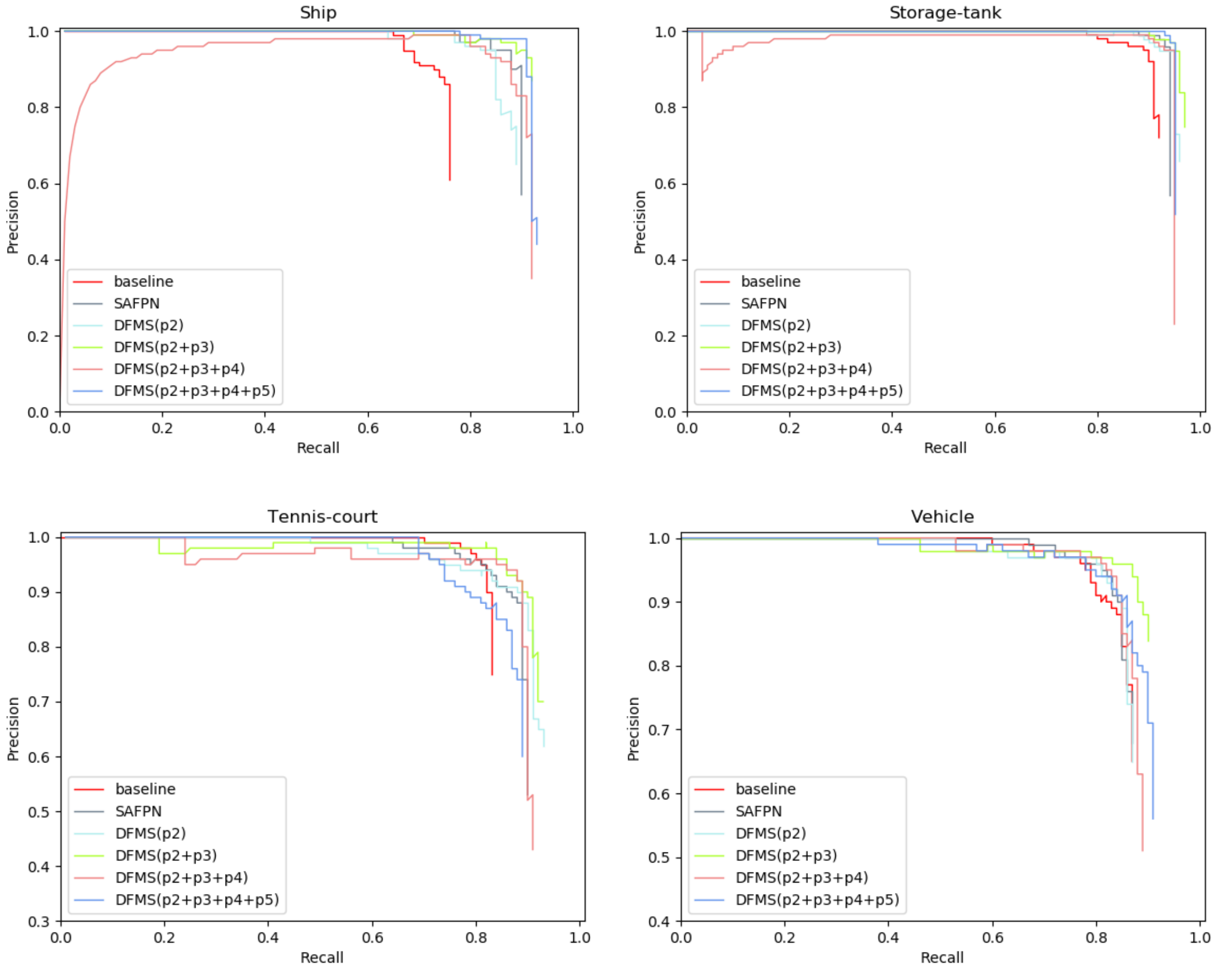


Figure 11. Precision-recall curves of different methods for each category in NWPU_VHR

4.2 Results on UCAS_AOD

The experiment results in Table 5 show that the result of our method has been improved by 3.49% compared to the baseline, and 3.33% higher than PRSdet, whose backbone is much deeper. It also indicates that the confidence of the detection target has also been significantly improved. The detection results are compared in Figure 13.

Table 5. Comparisons with state-of-the-art detectors on UCAS_AOD

Methods	Backbone	Size	mAP
FR-O [34]	ResNet101	800×800	88.30
RoI Trans [28]	ResNet101	800×800	88.95
PRSDet [12]	ResNet101	512×512	90.03
Baseline	ResNet50	800×800	89.87
DPANet(Ours)	ResNet50	800×800	**93.36**

4.3 Results on NWPU_VHR

The results show that our method is both effective for target detection whether it use horizontal frames or rotating frames. As shown in Table 6, the detection results of storage tank, harbor, vehicle, ground-track-field and baseball-diamond have achieved the best performance. The addition of DPANet can improves the object detection performance of horizontal bounding box. The significant improvement demonstrates the adaptability of our method. Figure 14 depicts a comparison of the experiment results.

5 Conclusion

In this paper, we proposed a dynamic pyramid attention network for oriented objects detection in remote sensing images, which consists of SAFPN and DFMS. SAFPN is a self-attention-based improved FPN that can improve the inherent disadvantages of FPN through enhancing high-level semantic features and transmitting the enhanced semantic information to different feature levels. The DFMS module is a dynamic selection of training feature map layers based on proposal distribution. It can achieve much better results with less computation with the help of self-attention. The experiment results show that our method can detect rotating objects in complex scenes. The shortcoming of our method is that the addition of two modules greatly lengthens the training time. However, the advantage of DPANet is that it could be added to any other one-stage object detection network, whether OBB or HBB, to improve detection results.

Table 6. Performance evaluation of HBB task on NWPU_VHR

Class	SSD[6]	RI-CAO[35]	Multi-scale FFN[36]	ODNN[37]	DPANet (ours)
Airplane	95.7	**99.70**	93.4	93.0	99.59
Ship	**92.8**	90.80	77.1	84.5	89.46
Storage tank	85.6	90.61	87.5	87.1	**94.24**
Baseball-diamond	96.6	92.91	93.0	92.8	**97.42**
Tennis Court	82.1	**90.29**	82.7	82.0	88.10
Basketball Court	86.0	80.13	83.8	**89.0**	82.23
Ground-track-field	58.2	90.81	83.7	78.0	**99.74**
Harbor	54.8	80.29	82.5	76.0	**92.16**
Bridge	41.9	68.53	72.5	**81.0**	76.89
Vehicle	75.6	87.14	82.3	81.5	**87.57**
mAP	75.9	87.12	83.8	84.8	**90.74**

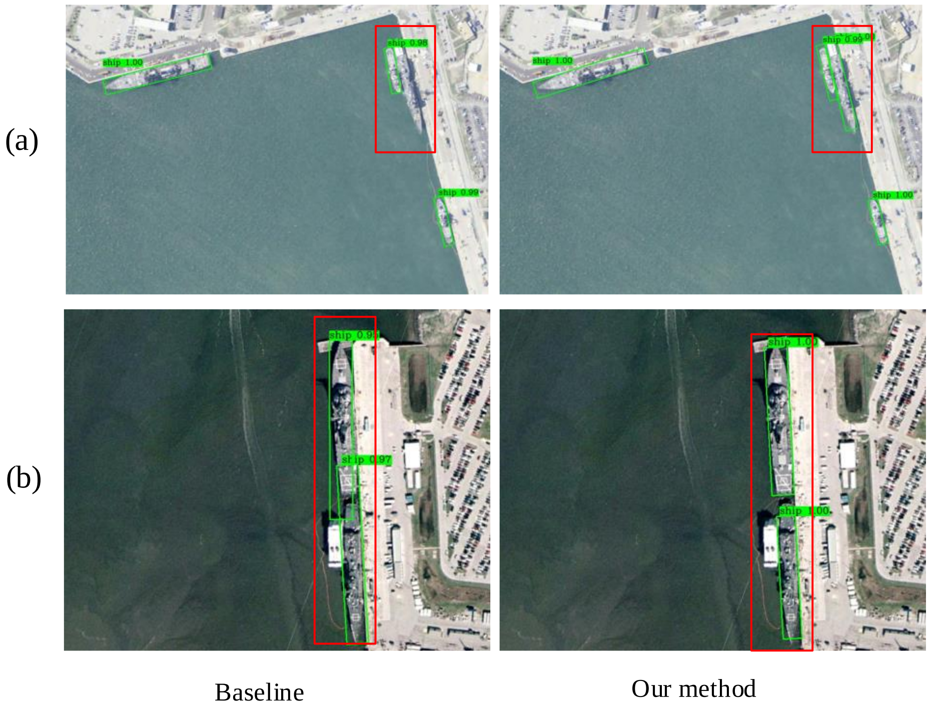


Figure 12. Comparison of visualization results in dataset HRSC2016. The proposed method presents a better detection accuracy than baseline method. In group (a), the baseline method missed an object. The proposed method obtains better regression coordinates in group(b).

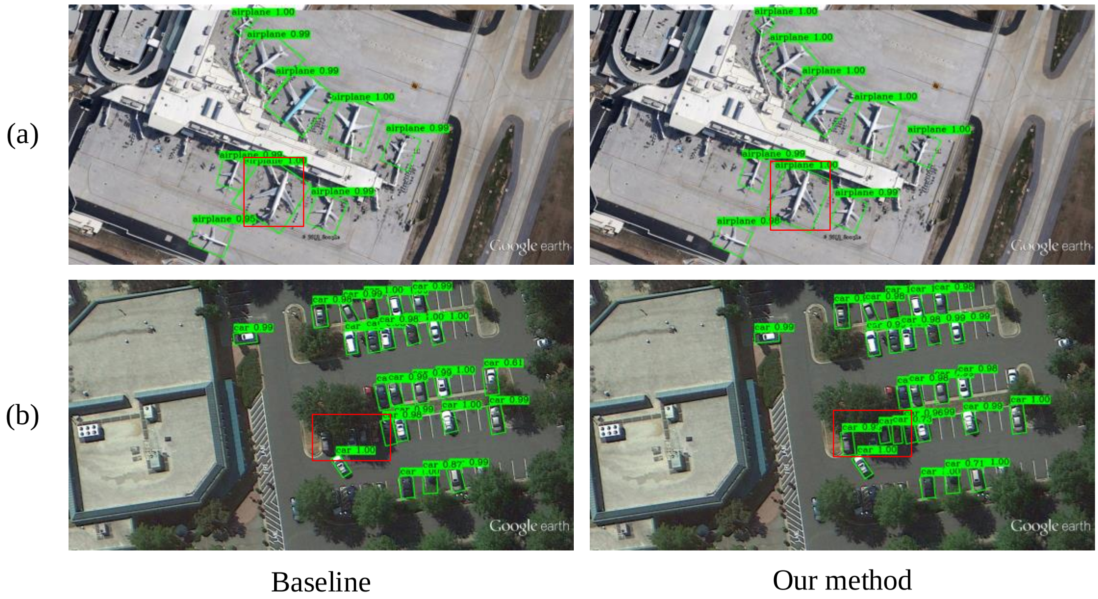


Figure 13. Comparison of visualization results in dataset UCAS_AOD. In group (a), the proposed method performed better in directional regression. In group (b), the baseline method missed some objects.

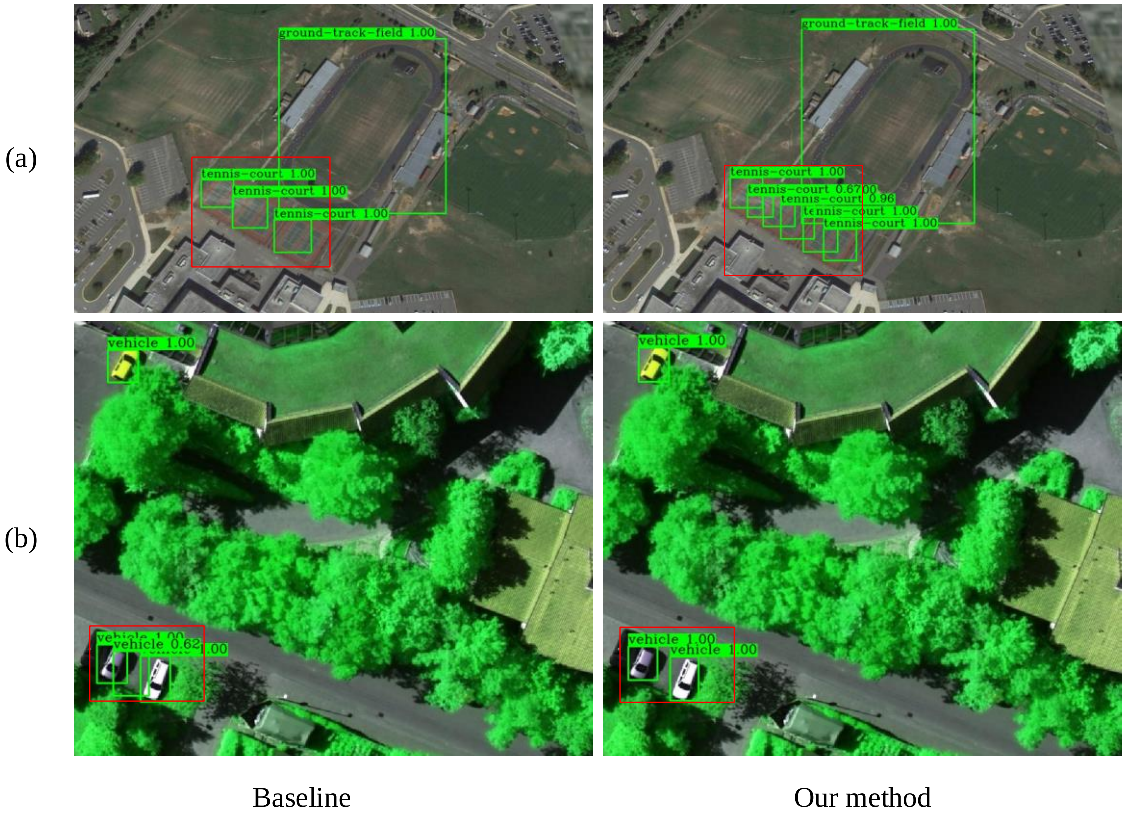


Figure 14. Comparison of visualization results in dataset NWPU_VHR. In group (a), the proposed method can distinguish each nearby target one by one. In group (b), a wrong object is identified in the baseline method.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grant 41776142.

References

[1] Q. Fan, W. Zhuo, C.-K. Tang, Y.-W. Tai, Few-shot object detection with attention-RPN and multi-relation detector, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 4012-4021.

[2] Y.-T. Chang, W. K. T. M. Gunarathne, T. K. Shih, Deep Learning Approaches for Dynamic Object Understanding and Defect Detection, *Journal of Internet Technology*, Vol. 21, No. 3, pp. 783-790, May, 2020.

[3] Y. Ye, H. Chen, C. Zhang, X. Hao, Z. Zhang, Sarpnet: Shape attention regional proposal network for lidar-based 3d object detection, *Neurocomputing*, Vol. 379, pp. 53-63, February, 2020.

[4] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, *IEEE Transactions on Pattern Analysis & Machine Intelligence*, Vol. 39, No. 6, pp. 1137-1149, June, 2017.

[5] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 779-788.

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, 2016 *European Conference on Computer Vision (ECCV)*, Amsterdam, Netherlands, 2016, pp. 21-37.

[7] T-Y Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, *IEEE Transactions on Pattern Analysis & Machine Intelligence*, Vol. 42, No. 2, pp. 318-327, Febuary, 2020.

[8] Y. Jiang, X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, P. Fu, Z. Luo, R2cnn: rotational region cnn for orientation robust scene text detection, *arXiv preprint arXiv: 1706.09579*, June, 2017.

[9] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, X. Xue, Arbitrary-oriented scene text detection via rotation proposals, *IEEE Transactions on Multimedia*, Vol. 20, No. 11, pp. 3111-3122, November, 2018.

[10] X. Yang, J. Yan, Z. Feng, T. He, R3det: Refined single-stage detector with feature refinement for rotating object, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, No. 4, pp. 3163-3171, May, 2021.

[11] G. Zhang, S. Lu, W. Zhang, CAD-Net: A context-aware detection network for objects in remote sensing imagery, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 57, No. 12, pp. 10015-10024, December, 2019.

[12] L. Zhou, H. Wei, H. Li, W. Zhao, Y. Zhang, Y. Zhang, Arbitrary-Oriented Object detection in remote sensing images based on polar coordinates, *IEEE Access*, Vol. 8, pp. 223373-223384, November, 2020.

[13] X. Yang, J. Yang, J. Yan, Y. Zhang, T. Zhang, Z. Guo, X. Sun, K. Fu, SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects, *2019 IEEE/CVF International Conference on*

Computer Vision (ICCV), Seoul, Korea, 2019, pp. 8231-8240.

[14] M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu, Spatial transformer networks, *Advances in neural information processing systems*, Montreal, Canada, 2015, pp. 2017-2025.

[15] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 42, No. 8, pp. 2011-2023, August, 2020.

[16] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, *European Conference on Computer Vision*, Munich, Germany 2018, pp. 3-19.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, *Advances in neural information processing systems*, Long Beach, CA, USA, 2017, pp. 5998-6008.

[18] J. Cheng, L. Dong, M. Lapata, Long short-term memory-networks for machine reading, *arXiv preprint arXiv:1601.06733*, September, 2016.

[19] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, *International conference on machine learning*, Lille, France, 2015, pp. 2048-2057.

[20] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, 2020 *European Conference on Computer Vision (ECCV)*, online, 2020, pp. 213-229.

[21] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, L.-C. Chen, Axial-deeplab: Stand-alone axial-attention for panoptic segmentation, *European Conference on Computer Vision*, online, 2020, pp. 108-126.

[22] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2016 *IEEE conference on computer vision and pattern recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770-778.

[23] R. Yin, R. Zhang, W. Zhao, F. Jiang, DA-Net: Pedestrian Detection Using Dense Connected Block and Attention Modules, *IEEE Access*, Vol. 8, pp. 153929-153940, August, 2020.

[24] Y. Gong, X. Yu, Y. Ding, X. Peng, J. Zhao, Z. Han, Effective fusion factor in FPN for tiny object detection, *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, 2021, pp. 1159-1167.

[25] H. Kuang, B. Wang, J. An, M. Zhang, Z. Zhang, Voxel-FPN: Multi-scale voxel feature aggregation for 3D object detection from LIDAR point clouds, *Sensors*, Vol. 20, No. 3, Article No. 704, February, 2020.

[26] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, *2017 IEEE conference on computer vision and pattern recognition (IEEE/CVF)*, Honolulu, Hawaii, USA, 2017, pp. 936-944.

[27] Q. Ming, Z. Zhou, L. Miao, H. Zhang, L. Li, Dynamic anchor learning for arbitrary-oriented object detection, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, No. 3, pp. 2355-2363, May, 2021.

[28] J. Ding, N. Xue, Y. Long, G.-S. Xia, Q. Lu, Learning roi transformer for oriented object detection in aerial images, *2019 IEEE Conference on Computer Vision and Pattern Recognition (IEEE/CVF)*, Long Beach, CA, USA, 2019, pp. 2849-2858.

[29] Y. Xu, M. Fu, Q. Wang, Y. Wang, K. Chen, G.-S. Xia, X. Bai, Gliding vertex on the horizontal bounding box for multi-oriented object detection, *IEEE transactions on pattern analysis and machine intelligence*, Vol. 43, No. 4, pp. 1452-1459, April, 2021.

[30] X. Yang, J. Yan, Arbitrary-oriented object detection with circular smooth label, 2020 *European Conference on Computer Vision (ECCV)*, online, 2020, pp. 677-694.

[31] M. Liao, Z. Zhu, B. Shi, G.-S. Xia, X. Bai, Rotation-sensitive regression for oriented scene text detection, *2018 IEEE conference on computer vision and pattern recognition (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 5909-5918.

[32] Z. Yang, K. He, F. Zou, W. Cao, X. Jia, K. Li, C. Jiang, ROPDet: real-time anchor-free detector based on point set representation for rotating object, *Journal of Real-Time Image Processing*, Vol. 17, No. 6, pp. 2127-2138, December, 2020.

[33] X. Yang, J. Yan, Q. Ming, W. Wang, X. Zhang, Q. Tian, Rethinking rotated object detection with gaussian wasserstein distance loss, *arXiv preprint arXiv:2101.11952*, November, 2021.

[34] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, L. Zhang, DOTA: A large-scale dataset for object detection in aerial images, 2018 *IEEE Conference on Computer Vision and Pattern Recognition (IEEE/CVF)*, Salt Lake City, UT, USA, 2018, pp. 3974-3983.

[35] K. Li, G. Cheng, S. Bu, X. You, Rotation-insensitive and context-augmented object detection in remote sensing images, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 56, No. 4, pp. 2337-2348, April, 2018.

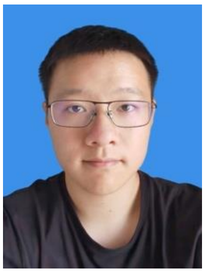
[36] S. Zhuang, P. Wang, B. Jiang, G. Wang, C. Wang, A single shot framework with multi-scale feature fusion for geospatial object detection, *Remote Sensing*,Vol. 11, No. 5, Article No. 594, March, 2019.

[37] S. Jiang, W. Yao, M. S. Wong, G. Li, Z. Hong, T.-Y. Kuc, X. Tong, An Optimized Deep Neural Network Detecting Small and Narrow Rectangular Objects in Google Earth Images, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 13, pp. 1068-1081, March, 2020.

Biographies

Hongchun Yuan received the B.S. and master’s degrees from Anhui Agricultural University and the Ph.D. degree in pattern recognition and intelligent system from University of Science and Technology of China, Anhui, China. He is currently the Vice chairman of agriculture and Forestry Committee of National Computer Basic Education Research Association of colleges and Universities. His research interests include application of artificial intelligence, computer vision and image processing.

Hui Zhou was born in 1996. He received the B.S. degrees in software engineering from the Xuzhou University of Technology, Jiangsu, China, in 2019. He is currently pursuing the M.S. degree at the college of Information Technology, Shanghai Ocean University. His research interests include object detection and monocular depth estimation.



Zhenyu Cai was born in 1996. He received the B.S. degrees in spatial information and digital computing technology from the Shanghai Ocean University, in 2018. He is currently pursuing the M.S. degree at the college of Information Technology, Shanghai Ocean University. His research interests include image anomaly detection and video processing.

Shuo Zhang received the B.S. in computer science and technology from North China Institute Of Aerospace Engineering, Langfang, China, in 2018. He is currently pursuing the M.S. degree at the College of Information Technology, Shanghai Ocean University, Shanghai, China. His current research interests include image enhancement, and artificial intelligence.

Ruoyou Wu received the B.E. degree in software engineering from China University of Petroleum (East China), Qingdao, China, in 2018. He is currently pursuing the M.S. degree with School of Information, Shanghai Ocean University, Shanghai, China. His research interests include image processing, espe- cially on underwater image enhancement.