Multi-level Feature Selection for Oriented Object Detection

Chen Jiang[1], Yefan Jiang[1], Zhangxing Bian[3], Fan Yang[2] and Siyu Xia[1]
[1]*School of Automation, Southeast University, Nanjing, China*
[2]*College of Telecommunications and Information Engineering, NJUPT, Nanjing, China*
[3]*Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, U.S.A.*

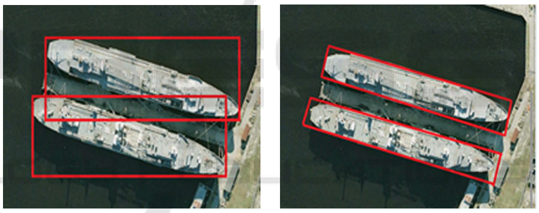
Keywords: Object Detection, Rotation Detection, Feature Selection, Remote Sensing, Path Aggregation.

Abstract: Horizontal object detection has made significant progress, but the representation of horizontal bounding box still has application limitations for oriented objects. In this paper, we propose an end-to-end rotation detector to localize and classify oriented targets precisely. Firstly, we introduce the path aggregation module, to shorten the path of feature propagation. To distribute region proposals to the most suitable feature map, we propose the feature selection module instead of using selection mechanism based on the size of region proposals. What's more, for rotation detection, we adopt eight-parameter representation method to parametrize the oriented bounding box and we add a novel loss to handle the boundary problems resulting from the representation way. Our experiments are evaluated on DOTA and HRSC2016 datasets.

1 INTRODUCTION

Object detection which benefits from the development of deep learning methods, especially in deep convolution neural networks, has made significant breakthroughs. The progress expands the applications scenarios of object detection, such as in security system, text detection and aerial images.

Current popular detectors can be divided into two types by different output representations: horizontal and oriented bounding box. Horizontal bounding box, which is always represented by (x,y,w,h) (the coordinate of center point, width, height), is difficult to locate multi-oriented objects. With the ratio of rotated objects enlarging, horizontal bounding boxes will include more background noise to the detriment of model training. For example, Fig. 1 shows, when rotated objects are close-packed, it will bring difficulties for detection and terrible visual experience. To address the limitation of horizontal detectors, numerous rotated object detection methods have been proposed and achieve considerable progress. However, rotation detectors are still faced with several problems.



(a) Horizontal bounding box (b) Oriented bounding box

Figure 1: Different representation ways on small and cluttered object detection. Horizontal bounding box (HBB) includes extra target while single target is framed by oriented bounding box (OBB).

The rotated objects are usually labeled by five parameters including an additional parameter θ to original horizontal bounding box representation (x,y,w, h). In the five-parameter representation method from OpenCV, when the parameter θ reaches its range boundary, such as 1° and −89°, two bounding boxes will be very approximate through exchanging width and height. The loss of these three parameters might be huge, even though the position of boxes almost doesn't change at all. Moreover, IOU is sensitive to minor angle fluctuation, leading to the decline of object detection performance. Some other methods label rotated boxes through recording the coordinates of four vertices. However, the coordinates are disordered and ambious—the same rotated object can be represented by several sets of values—resulting in abnormal loss. Another method (Xu et al., 2020) glides each vertex of the horizontal bounding box to form a quadrangle to represent the rotated object, which also has problems when the object is nearly horizontal. Moreover, it is difficult for rotated bounding box regression when the prediction of horizontal bound-

Jiang, C., Jiang, Y., Bian, Z., Yang, F. and Xia, S.
Multi-level Feature Selection for Oriented Object Detection.
DOI: 10.5220/0010213000360043
In *Proceedings of the 10th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2021)*, pages 36-43
ISBN: 978-989-758-486-2

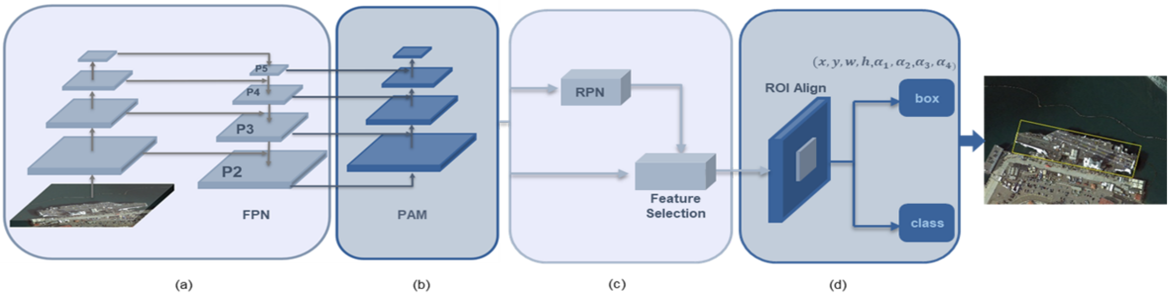


Figure 2: Architecture of our method. Our network consists of four modules: (a) backbone of network for feature extraction (using ResNet50 and FPN), (b) path aggregation module (PAM) for shortening the information path, (c) feature selection module (FSM) for selecting the optimal feature map, (d) ROI Align module and branches of regression and classification.

ing box exists deviations.

In this paper, we propose an effective and fast framework for multi-oriented object detection. We adopt the gliding vertex method (Xu et al., 2020) to label the rotated objects and attempt to solve the fundamental problems arising from this way. In gliding vertex method, there are eight parameters, through adding four gliding offset variables on the basis of classic horizontal bounding box representation, as is shown in Fig. 3. The representation way can prevent the sequential problems arising from directly regressing four vertices, because each offset is corresponding to the relative side of horizontal bounding boxes. However, it shares the same boundary problem with five-parameter representation method when the angle and offset variables reach the range boundary. We propose a new modulated loss in view of this situation to ensure that the prediction result can be obtained most simply and directly. In addition, the final prediction quadrangle depends on the accuracy of horizontal bounding box regression. We design an adaptive feature selection module to improve the regression results of horizontal bounding boxes. Region of interest generated from region proposal network (Ren et al., 2015) will be distributed to best feature layer through feature selection module. We also add bottom-up path augmentation module to preserve low-level localization signals, which is beneficial to following regression and classification. In summary, the main contributions of this paper include:

- We design a novel modulated loss function to solve the boundary problems based on gliding vertex method, to improve the accuracy of prediction results when the detection object is nearly horizontal.
- We propose a novel FSM to flexibly distribute multi scale ROI to the most suitable feature map and add PAM to our model, increasing the accuracy of the horizontal bounding box prediction results.
- We propose an effective multi-oriented object detection network, and reach substantial gains on DOTA and HRSC2016.

2 RELATED WORK

2.1 Horizontal Region Object Detection

Since (Girshick et al., 2014) proposed R-CNN, classic object detection has made considerable breakthrough. Based on this seminal work, Fast R-CNN (Girshick, 2015), Faster R-CNN (Ren et al., 2015) and R-FCN (Dai et al., 2016) are proposed subsequently, which improve the accuracy and efficiency of detection. Faster R-CNN includes object detection network and region proposal network (RPN), is the representative two-stage method. On the other hand, one-stage detectors, which predict bounding boxes directly from feature maps, are proposed to improve the speed for the simple architecture, such as SSD (Liu et al., 2016), YOLO (Redmon et al., 2016) and RetinaNet (Lin et al., 2017b). In particular, YOLO series have shown great performance through optimization of several versions. RetinaNet along with focal loss function is presented to handle class imbalance of samples. (Lin et al., 2017a) considers the scale variance in images and proposes Feature Pyramid Network (FPN) to address the problems of multi-scale objects. To get rid of the disadvantages of anchor-based networks, anchor-free detectors become the research focus in recent years. CenterNet (Duan et al., 2019), FCOS (Tian et al., 2019), and ExtremeNet (Zhou et al., 2019) are prototypical one-stage detectors. (Cai and Vasconcelos, 2018) introduce the idea of cascade and propose a multi-stage detector called Cascade R-CNN, that achieve high performance in both localization and classification. However, the above detectors merely generate horizontal bounding boxes, and still have application limitations in several

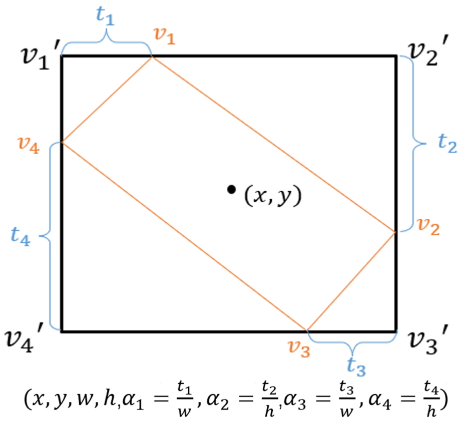

t_1
v_1'
v_1
v_2'
t_2
v_4
(x, y)
t_4
v_2
v_4'
v_3
v_3'
t_3
$(x, y, w, h, \alpha_1 = \frac{t_1}{w}, \alpha_2 = \frac{t_2}{h}, \alpha_3 = \frac{t_3}{w}, \alpha_4 = \frac{t_4}{h})$

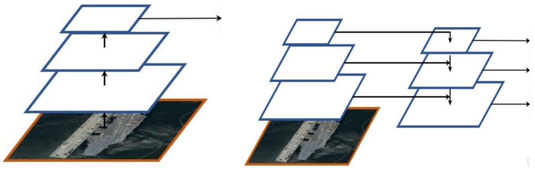
Figure 3: Representation of oriented bounding box. We adopt eight parameters *(x,y,w,h,*α_1,α_2,α_3,α_4*)* to label the oriented object, where *(x,y,w,h)* represents the horizontal bounding box.

real-world scenarios, especially in aerial images and multi-oriented scene texts.

2.2 Oriented Region Object Detection

Since horizontal detectors are not suitable for detection tasks in aerial images and scene texts, more detectors for rotated objects spring up. In scene text detection, RRPN (Ma et al., 2018) is improved in the framework of Faster R-CNN and propose rotated region proposal network (RRPN). TextBox++ (Liao et al., 2018a) adopts quadrilateral prediction based on SSD. RRD (Liao et al., 2018b) decouples classification and bounding box regression on rotation-invariant and rotation sensitive features to further improve TextBox++. R^{2}CNN (Jiang et al., 2017) also generates rotated bounding boxes to perform fast and effective text detection.

For object detection in aerial images, the complexity of the remote sensing background, the multi scales of samples and the huge number of dense, cluttered and rotated objects are extremely difficult, which calls for robust detectors. RoI Transformer (Ding et al., 2019) extracts rotated region of interest to locate and classify. SCRDet (Yang et al., 2019b) combines multi-dimensional attention network and refined sampling network and achieve state-of-the-art performance. R3Det (Yang et al., 2019a) proposes a refined single-stage detector with feature refinement to solve the feature misalignment problem. Gliding Vertex (Xu et al., 2020) and RSDet (Qian et al., 2019) reach SOTA performance on DOTA datasets by quadrilateral regression.



(a) Convolutional network (b) Feature pyramid network

Figure 4: Illustrations of two feature extraction network structures. (a) use the top feature map for fast prediction. (b) use feature pyramid network for more accurate prediction.

2.3 Multi-level Features

Features from different layers include distinct semantic information, which are useful for multi-scale object detection. SharpMask (Pinheiro et al., 2016), LRR (Ghiasi et al., 2016) and (Peng et al., 2017) use feature fusion to obtain more details. FCN (Long et al., 2015) and U-Net (Ronneberger et al., 2015) fused features from lower layers by skip-connections. FPN (Lin et al., 2017a) introduce a top-down path and combine semantic features from top layers and high-resolution information from lower layers for segmentation. PANet (Liu et al., 2018) firstly augments an additional down-top path in the framework of FPN and achieves better prediction. ASFF (Liu et al., 2019), NAS-FPN (Ghiasi et al., 2019) and BiFPN (Tan et al., 2020) adopt complex two-path integration for further improvement.

3 PROPOSED METHOD

3.1 Overview

The architecture of our network is shown in Fig. 2. It includes four modules: (a) the backbone of our network. We adopt FPN in the framework. (b) the path aggregation module (PAM), which is proposed by PANet(Liu et al., 2018) to improve the results of multi-scale feature extraction. (c) the feature selection module (FSM), which is designed to distribute ROI to the optimal feature map. (d) the full connection layer. We first introduce PAM in Sec. 3.2. Next, the detail of FSM will be explained in Sec. 3.3. Finally, we will introduce the boundary problem of gliding vertex representation method and propose a novel loss to handle it.

3.2 Path Aggregation Module

CNN usually applied the structure of Fig. 4a in the past, which used the final feature map for prediction.

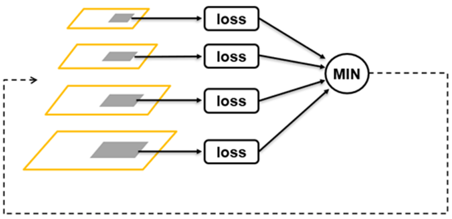


Figure 5: Illustration of feature selection module. Region proposals from RPN will be mapped to all feature maps and the optimal level is selected through calculating the layer with minimum IOU-loss.

However, pooling and other operations reduce the resolution of the convolution feature map, which cannot meet the needs of small target detection, resulting in low positioning accuracy and missing detection. As is shown in Fig. 4b, FPN (Lin et al., 2017a), which adds a top-down path, is proposed to ameliorate the problem. The up-bottom path transfers deep semantic information to shallow layers combined with low-level features, which enhances the robustness of the network to objects with different scales.

Meanwhile, the low-level features of the target, such as the edge, play a crucial role in the positioning. In order to enhance the results of localization, we need to make full use of low-level local feature information with high spatial resolution. PAM (Fig. 2(b)) introduces an additional bottom-up path aggregation path, which greatly shortens the propagation path of local feature information. Since the network aggregates the propagation path, it can better extract the local feature information, such as texture and edge of the target, and semantic feature information, and improve the ability of the network to detect multi-scale targets in remote sensing images. Ablation study is given in Sec. 4.4.

3.3 Feature Selection Module

After obtaining the multi-scale region proposals generated from RPN(Ren et al., 2015), the problem to be solved is how to allocate the ROI of different scales to the corresponding feature map. The common distribution method is that ROI will be mapped to different layers according to the size. Assuming the size of ROI is $w \times h$, according to Formula (1), it would be mapped to feature map P_k:

$$k = \left[k_0 + \log_2 \frac{\sqrt{wh}}{224} \right] \tag{1}$$

k_0 corresponds to the feature layer of the ROI with 224×224 area. It illustrates that the ROI with smaller size will be mapped to the feature map with higher spatial resolution, and the larger ROI will be mapped to the feature map with lower resolution. However, there are limitations of this simple method, which might not be the optimal plan. For example, two region proposals with a difference of about 10 pixels in size may be assigned to different feature layers under this method, but in fact the two regions may be very similar. Feature maps from PAM combine low-level feature information with high-level feature information, and it is vital to select the most suitable feature map, which is beneficial to final regression and classification. The architecture of FSM is illustrated in Fig. 5. Firstly, each region proposal generated from RPN (the grey regions) will be mapped to different feature levels. Next, the ground truth (horizontal bounding box) will be mapped to these feature maps and calculate IOU-Loss individually on each feature map. Finally, we will compare all loss values and choose the minimum to decide the optimal level to be pooled through ROIAlign. Better than proposal size, the smallest IOU-Loss value includes the most feature information. Through FSM, we transform the simple feature selection mechanism to IOU-based adaptive method. Ablation study is given in Sec. 4.4.

3.4 Loss Functions

We adopt a simple representation for rotated object, which is intuitively displayed in Fig. 3. The black horizontal bounding box B_h denoted by $(v_1^{'},v_2^{'},v_3^{'},v_4^{'})$, has four sliding vertices on each edge (v_1,v_2,v_3,v_4), which construct a quadrilateral B_r to represent the rotated object, the orange one in the figure. (v_1,v_2,v_3,v_4) are corresponding to the top, right, bottom and left side of B_h. The horizontal bounding box is represented by (x,y,w,h), where (x,y) is the center, w is width and h is height. The oriented bounding box is denoted by $(x,y,w,h,\alpha_1,\alpha_2,\alpha_3,\alpha_4)$, where the extra variables are defined as follows:

$$\alpha_1 = \frac{t_1}{w} \quad \alpha_2 = \frac{t_2}{h} \tag{2}$$

$$\alpha_3 = \frac{t_3}{w} \quad \alpha_4 = \frac{t_4}{h} \tag{3}$$

where $t_i = ||v_1^{'} - v_1||, i \in \{1,2,3,4\}$ represents the distance between $v_1^{'}$ and v_1.

However, both α_i=0 or 1 can represent horizontal bounding box, which is confused for bounding box regression. For example, the ground truth is horizontal, whose α_i are set to 1, and the predicted offset α_i are all nearly 0. Two regions are highly coincident but the loss would be far more than 0. In addition, when objects are similar to horizontal, it might be simpler to regress to near 0 rather than 1. Clearly, the final prediction result can be obtained in the simpler and more direct way when the target is nearly horizontal.

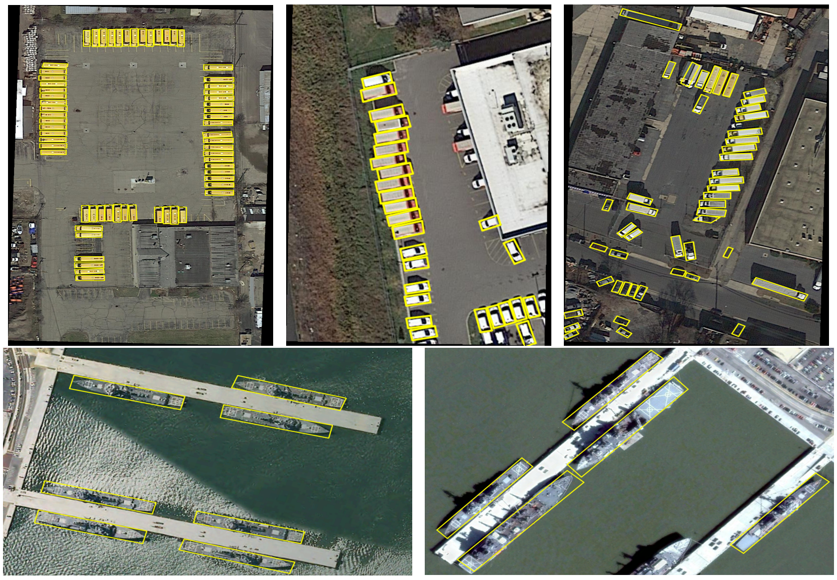


Figure 6: Example results of our method. The top row is from DOTA and the bottom row is from HRSC2016.

We propose a simple and novel loss function to handle this problem. When all the $\alpha_i^{'}$ of ground truth are nearly 0 or 1, we take it as the horizontal object and decide which regression direction to choose according to the smaller loss value. The regression loss function is defined as follows:

$$L_{reg} = L_h + L_r \tag{4}$$

$$L_r = \begin{cases} \sum_{i=0}^{4} \text{smooth}_{L1}\left(\alpha_i, \tilde{\alpha}_i\right), & l_1 \leq \tilde{\alpha}_i \leq l_2 \\ L_{\min} & \text{else} \end{cases} \tag{5}$$

$$L_{\min} = \min \begin{cases} \sum_{i=0}^{4} \text{smooth}_{L1}\left(\alpha_i, 0\right) \\ \sum_{i=0}^{4} \text{smooth}_{L1}\left(\alpha_i, 1\right) \end{cases} \tag{6}$$

where L_h is the regression loss function for horizontal box, the same as that in Faster R-CNN, and l_1, l_2 are the threshold to determine whether to be taken as horizontal boxes. The total loss function is given by

$$L = \frac{1}{N_{cls}} \sum_i L_{cls} + \frac{\lambda}{N_{reg}} \sum_i p_i^* L_{reg} \tag{7}$$

where N_{cls}, N_{reg} indicate the number of mini-batch size and positive targets in ground truth respectively, i denotes the index of a proposal in a mini-batch. p_i^* is a binary value ($p_i^* = 0$ for background and $p_i^* = 1$ for foreground). The hyper-parameter λ controls the trade-off, which is set to be 1 by default. The loss of RPN follows the Faster R-CNN (Ren et al., 2015).

4 EXPERIMENTS

We evaluate our proposed method on the challenge datasets DOTA and HRSC2016 for object detection in remote sensing, which both include a mass of arbitrary-oriented objects. The ablation study is conducted on HRSC2016, which is full of multi-oriented ships with different scales. Some qualitative results on HRSC2016 and DOTA are shown in Fig. 6

4.1 Datasets

DOTA is one of the largest and most challenging datasets in aerial image detection with quadrangle annotations. DOTA contains 2,806 aerial images from different sensors and platforms including 15 object categories with 188,182 instances, the size of which ranges from around 800 × 800 to 4,000 × 4,000 pixels. It is split into training, validation and testing sets, accounting for 1/2, 1/6, 1/3 of the whole data set, respectively. The categories are: Plane (PL), Swimming pool (SP), Baseball diamond (BD), Ground field track (GTF), Large vehicle (LV), Ship (SH), Tennis court (TC), Soccer-ball field (SBF), Basketball court (BC), Storage tank (ST), Bridge (BR), Roundabout (RA), Harbor (HA), Small vehicle (SV) and Helicopter (HC).

HRSC2016 is a dataset in ship detection with large range of aspect ratio and wide variety of orientations, which contains 1061 images with 29 categories. The size of each image in HRSC2016 is various, ranging

Table 1: Comparisons with other methods on DOTA.

Methods	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
FR-O(Xia et al., 2018)	79.09	69.12	17.17	63.49	34.2	37.16	36.2	89.19	69.6	58.96	49.4	52.52	46.69	44.8	46.3	52.93
R-DFPN(Yang et al., 2018)	80.92	65.82	33.77	58.94	55.77	50.94	54.78	90.33	66.34	68.66	48.73	51.76	55.1	51.32	35.88	57.94
R^{2}CNN(Jiang et al., 2017)	80.94	65.67	35.34	67.44	59.92	50.91	55.81	90.67	66.92	72.39	55.06	52.23	55.14	53.35	48.22	60.67
RRPN(Ma et al., 2018)	88.52	71.2	31.66	59.3	51.85	56.19	57.25	**90.81**	72.84	67.38	56.69	52.84	53.08	51.94	53.58	61.01
ICN(Azimi et al., 2018)	81.4	74.3	**47.7**	70.3	64.9	67.8	70	90.8	**79.1**	78.2	53.6	62.9	67	64.2	50.2	68.2
RADet(Li et al., 2020)	79.45	76.99	48.05	65.83	65.46	**74.4**	68.86	89.7	78.14	74.97	49.92	64.63	66.14	**71.58**	**62.16**	69.09
RoI-Transformer(Ding et al., 2019)	88.64	78.52	43.44	**75.92**	68.81	73.68	**83.59**	90.74	77.27	81.46	58.39	53.54	62.83	58.93	47.67	69.56
Ours	**89.2**	**86.5**	46.7	72.2	**69.5**	68	76.2	90.5	78.6	**84.4**	**58.5**	**69.9**	**67.4**	70.6	53.2	**72.1**

from 300×300 to 500×500. The training, validation and testing sets contain 436, 181 and 444 images.

4.2 Implementation Details

Network Setting. The proposed method is implemented on the framework of "maskrcnn benchmark" with Ubuntu 16.04, NVIDIA GTX 1080, and 8G Memory. We adopt ResNet50 as backbone and the batch size is set to 2 because of the limited memory. Stochastic gradient descent (SGD) is used in all experiments with weight decay and momentum set to 0.0001 and 0.9, respectively. The base learning rate is set to 0.0025 and is divided by a factor of 10 at each decay step. The threshold l_1, l_2 in Formula (5) is set to 0.1 and 0.9, respectively.

Dataset Setting. For DOTA, the model is trained for $80k$ iterations and the learning rate decays by 10 after $54k$ and $64k$ steps from an initial value of $2.5e^{-4}$ to $2.5e^{-6}$. We use random flipping and random rotate from (0,90,180,270) degree for data augmentation in training. For HRSC2016, the model is trained for $40k$ iterations and the learning rate decays at $28k$. Multi scale testing is applied with (0.5,0.8,1.0).

4.3 Comparisons with the State-of-the-Art Methods

We compare our proposed method with the state-of-the-art algorithms on DOTA and HRSC2016. The compared results of DOTA are depicted in Table 1. The results reported are achieved from the official DOTA evaluation server. The compared method include R^{2}CNN and RRPN, the methods for scene text detection, along with RoI-transformer and ICN, methods for aerial image detection. The results show that, our method obtains 72.1% mAP of OBB task and 2.54% better than the state-of-the-art method, RoI-Transformer. The compared results of HRSC2016 are shown in Table 2. Some qualitative results are selected for comparison, and we obtain 93.16% mAP in this dataset with ResNet50 as backbone.

4.4 Ablation Study

Effects of Path Aggregation Module. We use the ResNet50 as backbone and regress ($x, y, w, h, \alpha_1, \alpha_2, \alpha_3, \alpha_4$) with $smooth - l1$ loss for baseline. It shows that by adding PAM, it can improve performance by 1.04% mAP on the HRSC2016 in Table 3. PAM aims to shorten the propagation path of local feature information and achieve better results than independent FPN. It also reveals that PAM works on oriented object detection.

Effects of Feature Selection Module. The purpose of feature selection module is to help region proposals to select the most suitable feature map for further regression and classification, which substitutes the selection mechanism based on the size of ROI. The performance is improved from 89.88% mAP to 90.21% mAP in Table 3, which increases by 0.33% and is not significant. The feature maps from FPN include less information than those from PAM, which leads to the limited effect of FSM. When we combine PAM with FSM, the result increases from 89.88% mAP to 91.43% mAP, which increases by 1.55% and is more than the sum of two modules improvement.

Effects of $\mathbf{L}_{\min}$ Loss. The compared results in Table 3 shows the performance of $smooth - l1$ loss and the performance by adding our $L_{\min}$ loss in the fifth row. The $L_{\min}$ loss is proposed to detect nearly horizontal objects more effectively and directly. The performance increases by 0.99% through adding this loss function and improves more than 2% by adding three modules.

5 CONCLUSIONS

In this paper, we present an effective framework for oriented and multi-scale object detection. We introduce the PAM to shorten the feature propagation path from low level to high level and enhance the localization capability. To allocate ROI of different scales to the corresponding feature map, we propose FSM instead of the allocation scheme anchored in the size

Table 2: Comparisons with other methods on HRSC2016.

Method	Backbone	mAP
R^{2}CNN(Jiang et al., 2017)	ResNet101	79.73
RRPN(Ma et al., 2018)	ResNet101	85.64
RetinaNet-H(Yang et al., 2019a)	ResNet101	89.27
DRN(Pan et al., 2020)	Hourglass104	92.70
Ours	ResNet50	**93.16**

Table 3: Ablation study on HRSC2016.

FPN	PAM	FSM	$L_{\min}$	AP
✓				89.88
✓	✓			90.92
✓		✓		90.21
✓			✓	90.87
✓	✓	✓		91.43
✓	✓	✓	✓	**91.91**

of region proposals. In addition, we adopt the eight-parameter method to represent the oriented object and propose a novel modulated loss function to address the problem of the representation way when the target is nearly horizontal. We conduct extensive experiments to illustrate the achievements of our method across two datasets DOTA and HRSC2016 in comparison with several qualitative approaches and the effect of each module is shown by ablation study.

REFERENCES

Azimi, S. M., Vig, E., Bahmanyar, R., Körner, M., and Reinartz, P. (2018). Towards multi-class object detection in unconstrained remote sensing imagery. In *Asian Conference on Computer Vision*, pages 150–165. Springer.

Cai, Z. and Vasconcelos, N. (2018). Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162.

Dai, J., Li, Y., He, K., and Sun, J. (2016). R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387.

Ding, J., Xue, N., Long, Y., Xia, G.-S., and Lu, Q. (2019). Learning roi transformer for oriented object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2849–2858.

Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., and Tian, Q. (2019). Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6569–6578.

Ghiasi, G., Fowlkes, C. C., et al. (2016). Laplacian reconstruction and refinement for semantic segmentation. *arXiv preprint arXiv:1605.02264*, 4(4).

Ghiasi, G., Lin, T.-Y., and Le, Q. V. (2019). Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7036–7045.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587.

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

Li, Y., Huang, Q., Pei, X., Jiao, L., and Shang, R. (2020). Radet: Refine feature pyramid network and multi-layer attention network for arbitrary-oriented object detection of remote sensing images. *Remote Sensing*, 12(3):389.

Liao, M., Shi, B., and Bai, X. (2018a). Textboxes++: A single-shot oriented scene text detector. *IEEE transactions on image processing*, 27(8):3676–3690.

Liao, M., Zhu, Z., Shi, B., Xia, G.-s., and Bai, X. (2018b). Rotation-sensitive regression for oriented scene text detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5909–5918.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017a). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017b). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Liu, S., Huang, D., and Wang, Y. (2019). Learning spatial fusion for single-shot object detection. *arXiv preprint arXiv:1911.09516*.

Liu, S., Qi, L., Qin, H., Shi, J., and Jia, J. (2018). Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.

Ma, J., Shao, W., Ye, H., Wang, L., Wang, H., Zheng, Y., and Xue, X. (2018). Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia*, 20(11):3111–3122.

Pan, X., Ren, Y., Sheng, K., Dong, W., Yuan, H., Guo, X., Ma, C., and Xu, C. (2020). Dynamic refinement network for oriented and densely packed object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11207–11216.

Peng, C., Zhang, X., Yu, G., Luo, G., and Sun, J. (2017). Large kernel matters–improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4353–4361.

Pinheiro, P. O., Lin, T.-Y., Collobert, R., and Dollár, P. (2016). Learning to refine object segments. In *European conference on computer vision*, pages 75–91. Springer.

Qian, W., Yang, X., Peng, S., Guo, Y., and Yan, C. (2019). Learning modulated loss for rotated object detection. *arXiv preprint arXiv:1911.08299*.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Tan, M., Pang, R., and Le, Q. V. (2020). Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10781–10790.

Tian, Z., Shen, C., Chen, H., and He, T. (2019). Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 9627–9636.

Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., and Zhang, L. (2018). Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3974–3983.

Xu, Y., Fu, M., Wang, Q., Wang, Y., Chen, K., Xia, G.-S., and Bai, X. (2020). Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yang, X., Liu, Q., Yan, J., Li, A., Zhang, Z., and Yu, G. (2019a). R3det: Refined single-stage detector with feature refinement for rotating object. *arXiv preprint arXiv:1908.05612*.

Yang, X., Sun, H., Fu, K., Yang, J., Sun, X., Yan, M., and Guo, Z. (2018). Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks. *Remote Sensing*, 10(1):132.

Yang, X., Yang, J., Yan, J., Zhang, Y., Zhang, T., Guo, Z., Sun, X., and Fu, K. (2019b). Scrdet: Towards more robust detection for small, cluttered and rotated objects. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8232–8241.

Zhou, X., Zhuo, J., and Krahenbuhl, P. (2019). Bottom-up object detection by grouping extreme and center points. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 850–859.