Attentive Single Shot Quadrilateral Detector with Self-adaptive Anchor

Xingyi Yang[1],Hao Wang[1] ,Hao Zhu[2], Zheng Wang[2], Yanqing Zhu[2], Xiao XU[2] and Yining Hu*[2], Lizhe Xie*[3]

1. School of Computer Science and Engineering, Southeast University;

2. School of Cyber Science and Engineering, Southeast University;

3. Jiangsu Key Laboratory of Oral Diseases, Nanjing Medical University, Nanjing, China.

Yining Hu: hyn.list@seu.edu.cn

Lizhe Xie: xielizhe@njmu.edu.cn

In the DOAI Challenge, We design and adopt a one-stage object detector named Attentive Single Shot Rotational Detector (ASSRD).In addition to the circumscribed horizon bounding box regression, we simultaneously predict the relative vertex coordinates of the oriented bounding boxes using logistic regression. The network combines Feature Pyramid Network, Squeeze and Excitation module, and also Path Aggregation Network. Center Point Heat map is introduced as mask to raise the performance in detecting densely placed targets. To address the severe sample imbalance, we also exploit the strategy of Class-aware sampling. With data augmentation and Multi-scale testing and training, we achieve promising results on DOTA v1.5 Challenge.

I. METHODOLOGY

A. Network Architecture

We use a modified version of TextBoxes++ as the basic model. The overall network architecture of ASSRD is shown in Fig.1. The network receives images with size 512 as input. The network is composed of 3 parts: C-FPN, SE module and PA-Net. The C-FPN part is applied for multi-scale feature extraction. We propose to use Resnet101 pretrained on Imagenet as the encoder. We introduce the attention mechanism by using SE modules so as to enhance the channel relationship. Finally, the PA-net is applied to assemble the features from different level to conduct bottom-up path augmentation. The output contains class, box(horizontal bounding-box), and vertex (oriented bound-box).

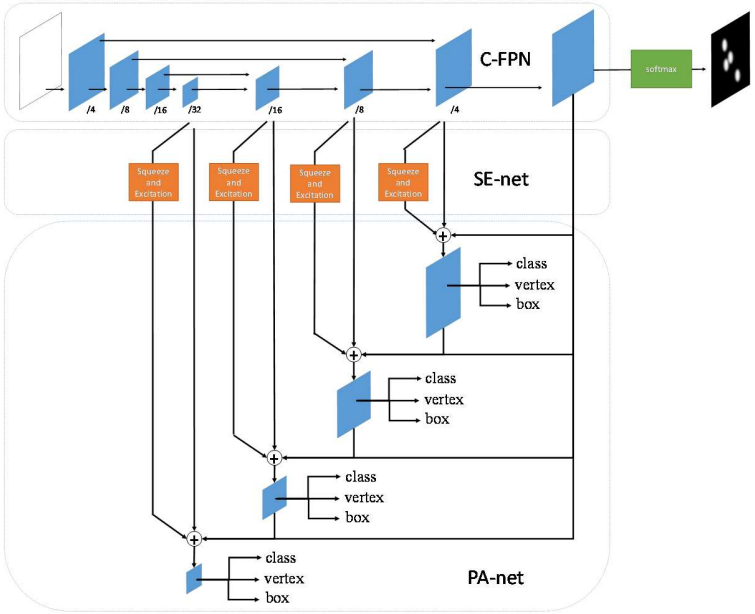

Fig.1 Network architecture of ASSRD

B. Dense object detection with center point heat map mask

One of the critical challenges in object detection in RS images is to the densely placed small targets. Previous one-stage detectors with dense anchor typically lack the ability to discriminate close objects. We address this problem inspired by the method of key-point detection using heat map, where an additional key-point branch is designed to extend the structure of FPN to complete the task of object center point detection. We use the multi-task learning to simultaneously predict the bounding boxes and the center heat map mask. Then we make use of the heat-map in two ways: on the one hand, we use the predicted object center to refine the bounding box regression result; on the other hand, the masks are also assembled with the multi-scale feature map to enhance the feature extractor with spatial attention mechanism.

C. Quadrilateral regression

Different from object detection tasks in ordinary scenes, in remote sensing images, target orientation should be annotated, otherwise, the commonly used horizontal bounding box may contain too many background pixels. We propose to perform quadrilateral vertex regression in two steps within one-stage.

In order to detect the bounding-box with arbitrary orientation, we carry out typical one-stage detection to predict the circumscribed horizontal bounding-box as the first step. First we find the grid cell which contains the target center. The prediction of the horizontal bounding-boxes(HBB) is to select from several anchors according to grid cell; regression is performed to select the anchor with best IOU with the horizontal bounding-box. The network predicts 4 variables for each bounding box(t_x, t_y, t_w, t_h). Presume the cell is offset from the top left corner of the image by (c_x, c_y) and the default anchor has width and height (w_d, h_d), then the HBB is defined as:

$$\begin{cases} x = \sigma(t_x) + c_x \\ y = \sigma(t_y) + c_y \\ W = w_d e^{t_w} \\ H = h_d e^{t_h} \end{cases} \tag{1}$$

The final step is to regress the 4 vertexes on the boundaries of the horizontal bounding-box according to the annotations. (x_0, y_0) is the top left corner of the HBB so each vertex (x_i, y_i) is expressed by the following equation:

$$\begin{cases} x_i = W\sigma(t_{x_i}) + x_0 \\ y_i = H\sigma(t_{y_i}) + y_0 \end{cases} (1 \le i \le 4) \tag{2}$$

In our approach, we use L2 loss for the prediction of horizontal bounding-boxes; cross entropy loss for the regression of vertexes coordinates.

D. Class Aware and Rivalry Non-Maximal Suppression

We adopt the Quadrilateral NMS to accurately repress the overlapping bounding-box with Class-Aware threshold and Rivalry mechanism.

1. Class-Aware NMS: not every class share the equal possible overlapping ratio. For example, the swimming pools and ground-track-fields seldom overlap with each other, whereas planes and large vehicles mutually occlude sometimes. With statistics analysis, we find that the IOU score approximately subjects to the exponential distribution.

   $$P(IOU > t) = e^{-\lambda t} \tag{3}$$

   We estimate the λ_i for each class with maximum likelihood estimation, then calculate a specific IOU threshold t_i for each class to set the $P(IOU > t_i) \ge 0.95$.

2. Rivalry NMS: In DOTA v1.5 and v1.0 dataset, some classes are quite confounding, therefore misclassification often happens between these pairs, such as small vs. large vehicle, harbor vs. bridge, basketball court vs. tennis court. Different from conventional NMS among the same class, we conduct the NMS in confusing pairs to reduce the false prediction mistakes.

E. Class-Aware Sampling

In order to solve the severe sample imbalance, we adopt the Class-Aware Sampling for training. Note N as the batch size we choose for training, for each batch, we randomly pick N classes from all 16 classes, and for each class among the N picked types, we choose one image which contains annotation of this type. Considering the annotation shortage for "bridge" and "container-crane", we intentionally increase the probability of picking the specific samples by 2.

II. EXPERIMENT

TABLE I COMPARATIVE EXPERIMENT OF OBB TASK1 ON DOTA1.5 DATASET

METHOD	AP	PLANE	BD	BRIDGE	GTF	SV	LV	SHIP	TC	BC	ST	SBF	RA	HARBOR	SP	HC	CC
R-RPN	**58.1**	**80.9.**	**75.9**	**37.5**	**57.2**	**48.0**	**60.2**	**69.7**	**91.1**	69.5	**65.3**	**40.8**	**39.8**	**52.7**	**61.3**	**44.7**	16.9
OURS	57.4	79.6	74.8	37.1	56.4	46.1	59.1	69.3	90.7	**75.6**	64.9	38.9	39.7	53.4	60.3	44.8	**28.0**

TABLE II COMPARATIVE EXPERIMENT OF OBB TASK1 ON DOTA1.0 DATASET

METHOD	AP	PLANE	BD	BRIDGE	GTF	SV	LV	SHIP	TC	BC	ST	SBF	RA	HARBOR	SP	HC
SSD	10.6	39.8	9.1	0.6	13.2	0.3	0.4	1.1	16.2	27.6	9.2	27.2	9.1	3.0	1.1	1.0
YOLOV2	21.4	39.6	20.3	36.6	23.4	8.9	2.1	4.8	44.3	38.4	36.7	16.0	37.6	47.2	25.5	7.5
R-FCN	26.8	37.8	38.2	3.6	37.3	6.7	2.6	5.6	22.9	46.9	66.0	33.4	47.2	10.6	25.2	17.7
FR-O	52.9	78.1	69.1	17.2	63.5	34.2	37.2	36.2	89.2	69.6	60.0	49.4	52.5	46.9	44.8	46.3
R^2CNN	60.8	80.9	65.7	35.3	**67.4**	**59.9**	60.9	55.8	90.7	66.9	**72.4**	55.1	52.2	55.1	53.4	48.2
R-RPN	**61.1**	**88.5**	**71.2**	31.7	59.3	51.9	56.2	57.3	**90.8**	72.8	67.4	**56.9**	**52.8**	53.1	51.9	**53.6**
OURS	60.2	87.7	66.5	**43.7**	47.8	52.0	**71.5**	**74.2**	90.6	**82.0**	69.1	41.3	24.8	**64.4**	**59.6**	27.2

The results in Table I and II demonstrate that the proposed ASSRD achieve competitive results on DOTA v1.5 and v1.0 OBB challenge against rotation-anchor-based R-RPN and two-stage method as R2CNN. But, our work runs at 9 fps on 512 input images, which is significantly faster than the best method so far.