Neural Computing and Applications (2020) 32:14549–14562
https://doi.org/10.1007/s00521-020-04893-9

S.I. : DEEP LEARNING APPROACHES FOR REALTIME IMAGE SUPER RESOLUTION (DLRSR)

Arbitrary-oriented object detection via dense feature fusion and attention model for remote sensing super-resolution image

Fuhao Zou[1] · Wei Xiao[1] · Wanting Ji[2] · Kunkun He[1] · Zhixiang Yang[3] · Jingkuan Song[4] · Helen Zhou[5] · Kai Li[1]

Received: 12 April 2019 / Accepted: 25 March 2020 / Published online: 7 May 2020

Abstract

In this paper, we aim at developing a new arbitrary-oriented end-to-end object detection method to further push the frontier of object detection for remote sensing image. The proposed method comprehensively takes into account multiple strategies, such as attention mechanism, feature fusion, rotation region proposal as well as super-resolution pre-processing simultaneously to boost the performance in terms of localization and classification under the faster RCNN-like framework. Specifically, a channel attention network is integrated for selectively enhancing useful features and suppressing useless ones. Next, a dense feature fusion network is designed based on multi-scale detection framework, which fuses multiple layers of features to improve the sensitivity to small objects. In addition, considering the objects for detection are often densely arranged and appear in various orientations, we design a rotation anchor strategy to reduce the redundant detection regions. Extensive experiments on two remote sensing public datasets DOTA, NWPU VHR-10 and scene text dataset ICDAR2015 demonstrate that the proposed method can be competitive with or even superior to the state-of-the-art ones, like R2CNN and R2CNN++.

Keywords Object detection · Arbitrary oriented · Rotation proposals · Remote sensing image · Attention model · Dense feature pyramid network · Super-resolution

1 Introduction

Automatically object detection for remote sensing image is usually a significant prerequisite for the visual recognition tasks, such as object coarse or fine-grained classification, object attribute learning, object counting and analysis of battle-field situation. Thus, object detection in remote sensing image has attracted a large amount of attentions in past decades. This phenomenon is further pushed to a new height by the success of deep convolutional network (DCNN) [1] in various computer vision tasks.

Strongly promoted by the advance in DCNN, a large body of object detection methods have been springed up, which mainly contain horizontal and rotation region-based methods. The representative horizontal region-based object detection method contain the RCNN [2], spatial pyramid pooling network (SSP-Net) [3], fast RCNN [4], faster RCNN [5], YOLO [6], SSD [7], R-FCN [8] and Mask RCNN [9], etc. However, this kind of methods only performs well in natural scene images but poor in the case of remote sensing image. With respect to remote sensing image scenario, the object detection methods will confront with the challenges of light variation, blur, imaging

✉ Fuhao Zou
fuhao_zou@hust.edu.cn

[1] School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China

[2] School of Natural and Computational Science, Massey University, Auckland, New Zealand

[3] Wuhan Digital Engineering Research Institute, Wuhan, China

[4] Innovation Center, University of Electronic Science and Technology of China, Chengdu, China

[5] School of Engineering, Manukau Institute of Technology, Auckland, New Zealand

perspective and dense arrangement, etc. To handle such problems, a series of rotation region-based methods have been proposed, such as R-DFPN [10], PDDP [11], R2-CNN [12], R2-CNN++ [13], AOSTD [14]. In contrast to the former, the later can generate more fitting bounding box in aspect of arbitrary-oriented object detection, having more accurately object localization and classification for scene text detection and object detection in remote sensing image. However, objects such as cars and ships in satellite imagery have a small spatial extent (as low as 10 pixels) and are often densely clustered. These methods did not solve this problem, so the results of these methods are still unsatisfactory. Recent studies have shown that the use of SR as a pre-processing step can yield significant improvements to the detection of small objects [15, 16] because the super-resolution methods increase the resolution of images, which add more distinguishable features that an object detection algorithm can use for discrimination.

In this paper, we focus on the arbitrary-oriented object detection in remote sensing image. Although many arbitrary-oriented object detection methods have been proposed before, this task still poses a great challenge resulting from the image sensing variance such as light, blur, intensive arrangement and image sensing perspective. As is well known, network engineering is increasingly more important for computer vision task [17, 18]. Inspired by these, we aim to develop new arbitrary object detection architectures to further push the frontier of object detection for sensing image. In the proposed architecture, we comprehensively take into account multiple strategies, such as feature fusion, attention model, rotation region proposal, rotation ROI pooling, super-resolution pre-processing simultaneously to boost the performance in terms of localization and classification under the faster RCNN-like framework. Functionally, the proposed architecture comprises five module including dual path network (DPN) [17] backbones module, SE [19] attention module, rotation region proposal (RRPN) [14] module and RRPN-based fast RCNN module [12, 14].

It is worthwhile highlighting the properties of the proposed method as follows.

1. We integrate dense FPN into the DPN as backbone network. Dense FPN enhances feature propagation and encourages feature reuse and DPN presents a new topology of connection paths and enables new features exploration which are both important for learning good representations. So this backbone can produce informative feature and discriminative multi-scale feature maps which ensures the effectiveness of detecting multi-scale objects;
2. SE attention model is leveraged to activate the channels useful to object detection while suppressing the channel closely related to the noise;
3. We adopt rotation anchors and rotation ROI pooling strategies to produce minimum circumscribed rectangle bounding box and overcome the difficulty of detecting densely arranged objects and eventually get a higher accuracy.
4. Extensive experiments on DOTA dataset are implemented to justify the rationality of combinations of five core modules of the proposed architecture and simultaneously show it is competitive with or even superior to the state-of-the-art ones, like R2CNN and R2CNN++.

2 Related works

Here, we review the representative object detection methods from comprehensive perspective which contain horizontal and rotation region proposal-based methods.

2.1 Horizontal region proposal-based object Detections

With the widespread use of DCNN in object detection, more and more efficient region-based object detection algorithms are proposed, such as region proposals with CNNs (RCNN) [2], spatial pyramid pooling network (SSP-Net) [3], fast-RCNN [5] and R-FCN [8]. RCNN adopts a multistage detection network structure strategy which first uses selective search [20] to generate a set of proposals followed by classifying each proposal with combination of ConvNet feature extractor and SVM classifier. SPPnet [3] demonstrated that such region-based detectors could be applied much more efficiently on feature maps extracted on a single image scale. Fast RCNN encourage using features computed from a single scale, because it maintains a good trade-off between accuracy and speed. Faster-RCNN [5] unifies RPNs with fast RCNN object detection networks, which adopt a training scheme that alternates between fine-tuning for the region proposal task and then fine-tuning for object detection, while keeping the proposals fixed. This scheme converges quickly and produces a unified network with convolutional features that are shared between both tasks. Region-based fully convolutional network (R-FCN) [8] builds a fully convolution network, which greatly reduces the number of parameters, improves the detection speed and has a good detection effect. Apart from efficient object detection like faster RCNN, Mask RCNN [9] simultaneously produces a high-quality segmentation mask for each instance.

Instead of depending on regional proposals, You Only Look Once (YOLO) [6] and Single Shot MultiBox Detector (SSD) [7] are regression-based object detection methods, which directly estimate objects region and truly enable real-time detection. Moreover, feature pyramid network (FPN) [17] adopts the multi-scale feature pyramid form and makes full use of the feature map to achieve better detection results.

2.2 Rotation region proposal-based object Detection

These approaches mentioned above are also called as horizontal region proposal-based object detection. However, for sensing image scenario, the object with a large range of aspect ratio, once the angle of proposal is inclined, the redundant region will be relatively large, vulnerably resulting in missing detection due to bad favorableness for the operation of non-maximum suppression. In order to handle such problem, a series of arbitrary-oriented object detection are proposed in the field of scene text detection (e.g., R2CNN [12], AOSTD [14]), ship detection (e.g., PDDP [11], R-DFPPN [10]), as well as object detection in remote sensing image like R2CNN++ [13]. For example, in the field of scene text detection, R2CNN [12] proposes a rotational region CNN based method, achieving outstanding results on scene text detection. However, since R2CNN still uses horizontal anchors at the first stage, the negative effects of non-maximum suppression still exist. To mitigate the shortcoming, a few rotation region proposal-based methods are proposed such as AOSTD [14]), PDDP [11], R-DFPPN [10]), R2CNN++ [13], which effectively improve the quality of the proposal. What is more, recent some studies have shown that the use of super resolution can yield improvements for remote sensing object detection or segmentation [21–24].

By contrast, object detection in remote sensing image is more difficult than text detection and ship detection. The detail reasons are as follows. Firstly, scene text detection and ship detection only focus on single-object detection, which cannot be directly applied to multi-class object detection scenario. Second, the arrangement of scene text is usually more sparse than that of remote sensing image. In the end, it is required to be taken into account the impact resulting from factors such as scale, angle, density and scene complexity. This paper considers these factors comprehensively and proposes a general algorithm for multi-categories arbitrary-oriented object detection in aerial images.

3 The proposed method

To handle the problem mentioned above, we propose an arbitrary-oriented end-to-end training and testing object detection method which takes scale variance, rotation factor and feature engineering into account jointly. The architecture of the proposed is illustrated in Fig. 1, which is composed of five parts including DPN backbone module, attention module, dense FPN module, rotation region proposal networks (RRPN) module and rotation-based fast RCNN module. In addition, to mitigate the noise impact, we adopt super-resolution processing before object detection. The implement detail and its motivation of each module will be described in the following.

3.1 DPN backbone module

It is well known that the ResNet [1], ResNeXt [25] and DenseNet [26] make a significant success in various computer vision tasks, such as image classification, segmentation and object detection. In principle, the improvements of ResNet [1], ResNeXt [25] and DenseNet [26] owe to the subtle usage of residual path and densely connected paths, enabling effective feature re-usage and re-exploitation, respectively.

Inspired by these, Chen [17] proposes a novel dual path architecture, called the dual path network (DPN). The DPN inherits the advantages of residual and densely connected paths simultaneously, possessing higher parameter efficiency, lower computational cost and lower memory consumption, and being regarded as the state-of-the-art one in the family of DCNN. The DPN is built by stacking multiple mirco-blocks as shown in Fig. 2, in which the structure of each micro-block is designed with a bottleneck style which begins by a 1 × 1 convolutional layer followed by a 3 × 3 convolutional layer, and finalizes with a 1 × 1 convolutional layer. The output of the last 1 × 1 convolutional layer is partitioned into two parts: The first part is added to the residual path in element-wise way, and the second part is concatenated with the densely connected path. To enhance the learning capacity of each micro-block, DPN adopts the grouped convolution layer in the second layer like the ResNeXt [25].

Specifically, the implementation of DPN is:

$$x_{\text{dense}}, x_{\text{residual}} = \text{Split}(f(\text{Add}(f_{conv1\times1}(x_1), f_{\text{conv}1\times1}(x_2)))); \tag{1}$$

$$O_{\text{dense}} = \text{Concat}(x_{\text{dense}}, x_1); \tag{2}$$

$$O_{\text{residual}} = \text{Add}(x_{\text{residual}}, x_2) \tag{3}$$

while x_1, x_2 denotes the feature from individual path, namely DenseNet path and ResNet path, respectively, *Split*

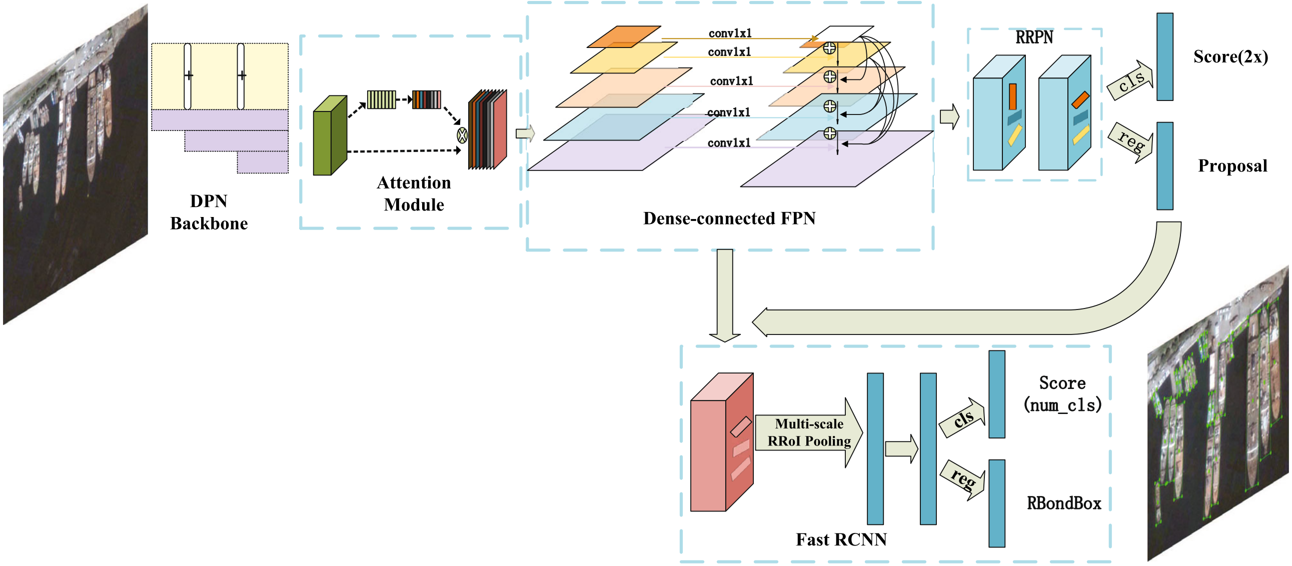
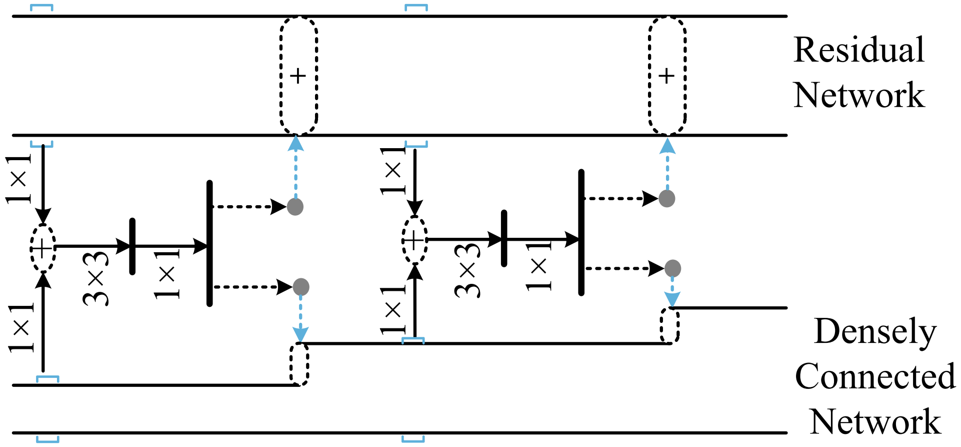


Fig. 1 The architecture of the proposed method

Fig. 2 Architecture of DPN



means split operation, O_{dense}, O_{residual} denotes the output of DenseNet path and ResNet path.

For object detection task, the selection of backbone network is the basis for designing a new method successfully. Since the DPN has the above-mentioned merits, the DPN is selected as the backbone network of the proposed new model. The input of the DPN backbone module the image after super-resolution processing and its output are then fed into the attention model module. Specially, the model complexity and computational complexity of DPN are competitive lower, since DPN-92 (145 MB, 6.5GFLOPs) costs about 15% fewer parameters, consumes about 19% FLOPs than ResNeXt-101 (32x4D) (170 MB), while the DPN-98 (236 MB, 11.7GFLOPS) costs about 26% fewer parameters and consumes about 25% FLOPs than ResNeXt-101 (64x4D) (320 MB). In addition, the training of DPN-98 is 15% faster and uses 9% less memory than the best performing ResNeXt. When meeting very high-resolution images, the DPN model has lower model complexity and higher training speed, which makes more efficient.

3.2 Attention module

Here, the Squeeze and Excitation (SE) [19] network is chosen as visual attention module to boost the performance of the object detection, which is an embedding composite block and can be integrated with the almost all DCNN network, such as ResNeXt and DenseNet. As illustrated in Fig. 3, the attention module comprises two parts: Squeeze block and Excitation block. The Squeeze block is used to transform C feature maps of size $H \times W$ into C feature maps of size 1×1 via Global Average Pooling operation. Specifically, a statistic $Z \in \mathbb{R}^C$ is generated by shrinking U

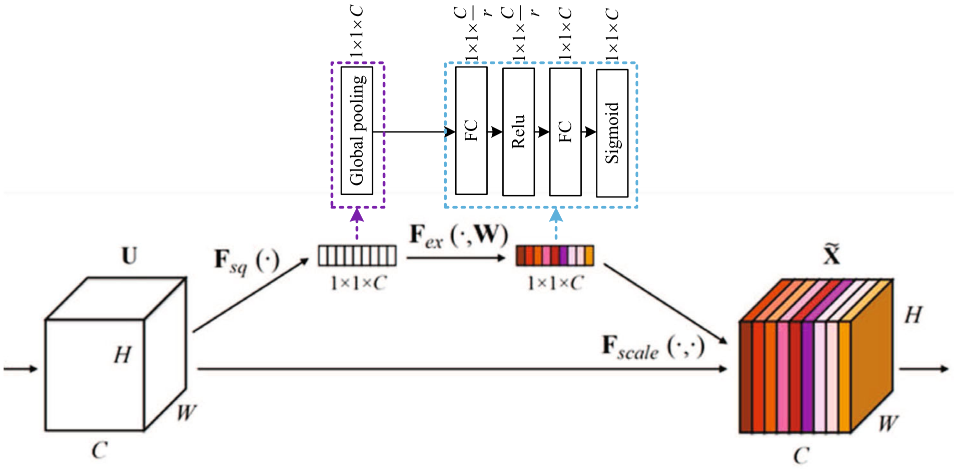


Fig. 3 Pipeline of the SE-NET Attention module

through its spatial dimensions $H \times W$, such that the c-th element of Z is calculated by

$$z_c = F_{sq}(U_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \tag{4}$$

As to the Excitation block, it is a combination multiple operations of $1 \times 1 \times \frac{C}{r}$ FC, $1 \times 1 \times \frac{C}{r}$ Relu, $1 \times 1 \times C$ FC and $1 \times 1 \times C$ Sigmoid. Specifically, the output of Excitation block is calculated by

$$s = F_{ex}(z, W) = \sigma(W_2 \delta(W_1 z)) \tag{5}$$

where δ refers to the Relu function, σ refers to the Sigmoid function, $W_1 \in \mathbb{R}^{\frac{c}{r} \times c}$ and $W_2 \in \mathbb{R}^{c \times \frac{c}{r}}$. Finally, the output of Excitation block comprises C feature maps of size 1×1. It is worth noting that the resulting feature maps are sparse vector.

As illustrated in Fig. 3, the resulting sparse feature maps are exploited as convolution kernels to perform convolution operation over the original feature maps. That is to say, the original feature maps is imposed sparse processing. By the sparsification, informative feature maps are selectively emphasized and less useful ones are suppressed in channel-wise direction. Then, the feature maps closely related to object detection are activated and the others are prohibited. The weight of SE-NET is obtained with automatical training way.

3.3 Dense FPN module

As we all know, low-level feature has relatively few semantic information, but the object location is accurate. In contrast, high-level feature semantic information is rich, but the object location is relatively coarse. The feature pyramid is an effective way to fuse different level information. Dense feature pyramid network (DFPN) [10] has got very good results in small object detection tasks. It exploits the feature pyramid, which is connected via top-down pathway, lateral connection and dense connections. Aerial object detection in remote sensing image can be considered a task to detect objects range from small size to large one. Meanwhile, considering the complexity of background in remote sensing images, there are a lot of interferences in the image. Therefore, the feature information obtained through the DFPN may enhance feature propagation and encourages feature reuse similar to DPN [17]. Dense Feature Pyramids Network is a significant component for detecting objects *at different scales*. Intuitively, this property enables a model to detect objects across a large range of scales by scanning the model over both positions and pyramid levels.

Figure 4 shows the architecture of DFPN based on ResNets [1]. In the bottom-up feedforward network, we still choose multi-level feature maps as C_2, C_3, C_4, C_5, corresponding to the last layer of each residual block which have strong semantic features. Note that they have strides of 4, 8, 16, 32 pixels. In the top-down network, we get higher-resolution features by lateral connections and dense connections as P_2, P_3, P_4, P_5. For example, in order to get P_2, we first reduce the number of C_2 channels by using a 1×1 convolutional layer, and then we use nearest neighbor upsampling for all the preceding feature maps. We merge them by concatenating rather than simply adding. Finally, we eliminate the aliasing effects of upsampling through a 3×3 convolutional layer, while reducing the number of channels. After the iteration above, we get the final feature maps P_2, P_3, P_4, P_5. Since we do not add much learnable parameters (only few 1×1 convs), the training memory cost will not be much higher and the training speed will not be much lower than the original FPN, which keeps the efficiency. In fact, our model achieves 80–100 samples per second training speed, which is competitive to the complex ResNeXt.

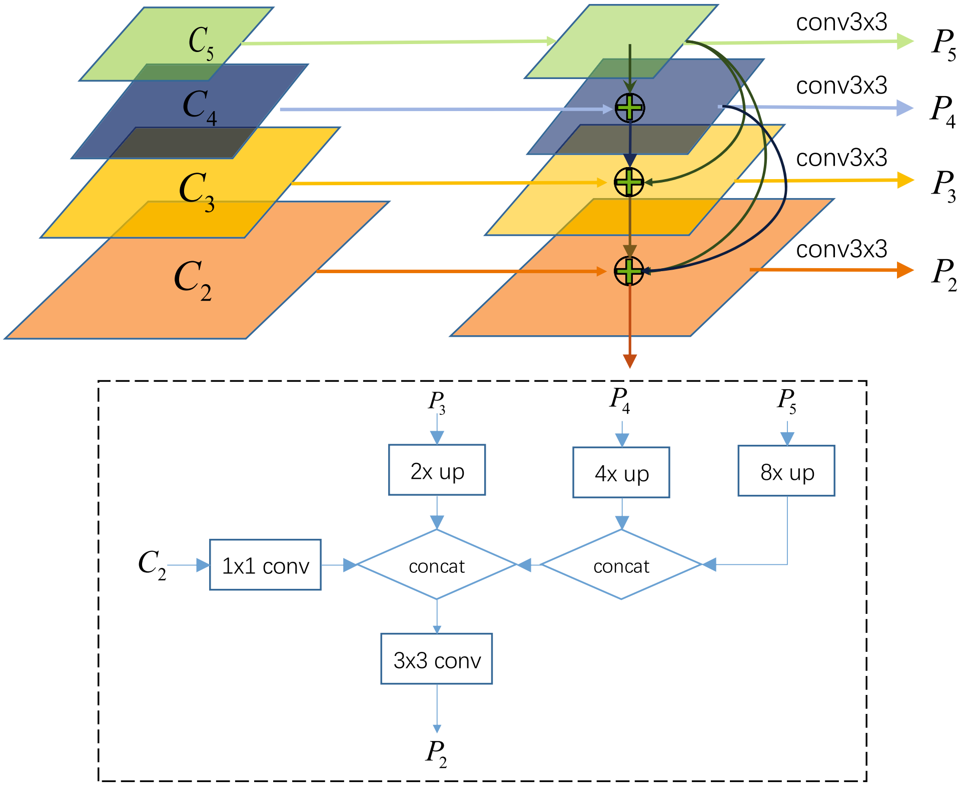


Fig. 4 Pipeline of the dense FPN

3.4 Rotated region proposal network (RRPN) module

The traditional bounding box is a horizontal rectangular box, so its representation is relatively simple, using four variables $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$ to represent a bounding box. $(x_{\min}, y_{\min})$ and $(x_{max}, y_{\max})$ represents the coordinates of the upper left and lower right corners of the bounding box, respectively. But this representation is obviously not suitable for representing a rotation bounding box. In order to represent the rotation bounding box better, we use five variables (x, y, w, h, θ) to determine a rotation bounding box. As shown in Fig. 5 where the (x, y) denotes the center coordinate of the rotation bounding box, and the orientation θ is the angle at which the horizontal axis (x-axis) rotates counterclockwise to the first edge of the encountered rectangle. At the same time, we define this side as width and the other as height.

RPN is proposed to accelerate the process of horizontal proposals generation. The multi-scale anchor boxes are generated by sliding over the last convolutional layer. Each anchor produces 2 classification scores and 5 coordinates output. To fit the objects of different sizes, the RPN adopts two parameters, scale and aspect ratio, which control size and shape of anchors. The scale parameter determines the size of the anchor, and the aspect ratio determines the ratio of the width to the height. The parameters setting of scale and aspect ratio is closely dependent on the scenario of task and dataset. Here, the DOTA dataset is selected as benchmark dataset. With respect to the DOTA dataset, targets usually have unnatural shape with arbitrary orientations, and the horizontal proposals generated by RPN are not robust for DOTA. So we adopt RRPN to encode rotation information and generate rotated proposals. The orientation parameter θ is to control the orientation of a proposal, i.e., $-\pi/6$, 0, $\pi/6$, $\pi/3$, $\pi/2$, and $2\pi/3$. Due to small targets with a majority in DOTA dataset, we set smaller anchor scales such as 16, 32, 64, 128 and 256.

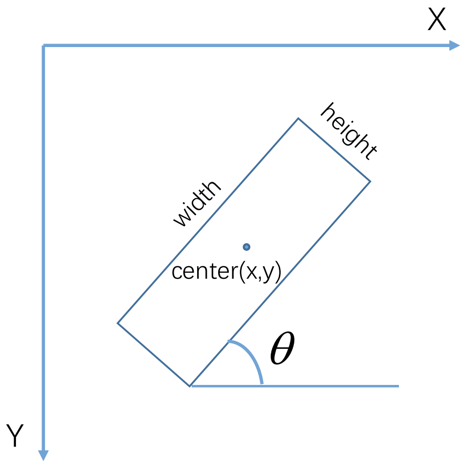


Fig. 5 General representation of rotation bounding box

Then, we assign a single scale to each feature map, and the size of the scale is $\{16, 32, 64, 128, 256\}$ pixels on $\{P_2, P_3, P_4, P_5, P_6\}$, respectively. In addition, the aspect ratios set {1:1, 1:2, 2:1, 1:4, 4:1, 1:9, 9:1} is assigned to cover a wide range of objects. For each point on the feature map, 42 rotation anchors (6 orientations, 7 aspect ratios and 1 scales) are generated, as well as 210 outputs (5 × 42) for the regression branch and 84 score outputs (2 × 42) for the classification branch.

After the rotation anchors are generated, a sampling strategy for the rotation anchors is needed to train the network. First, we define the intersection-over-union (IOU) overlap as the overlap between the ground truth and rotation anchor. Then, we define positive and negative samples according to the following rules. Positive rotation anchors feature the following: (1) the highest IOU or an IOU larger than 0.7 with respect to the ground truth and (2) an intersection angle with respect to the ground truth less than $\frac{\Pi}{12}$. Negative rotation anchors feature the following: (1) an IOU lower than 0.3, or (2) an IOU large than 0.7 but with an intersection angle with respect to the ground truth larger than $\frac{\Pi}{12}$. Anchors that are neither positive nor negative do not contribute to the training objective.

3.5 Rotation-based fast RCNN module

The module is the detection head that uses the rotation proposals. It is quite similar to fast RCNN. The main differences lie in twofold. One is the input proposals are rotation region proposals rather than the horizontal ones, which are yielded by RRPN Module. Another is that the ROI pooling layer is replaced with rotation ROI pooling like in the literature [10, 14]. And then, we adopt multi-task loss to minimize the objective function defined as follows:

$$L(p_i, l_i, t_i^*, t_i) = \frac{1}{N_{\text{cls}}}\sum_i L_{\text{cls}}(p_i, l_i) + \lambda\frac{1}{N_{\text{reg}}}\sum_i p_i L_{\text{reg}}(t_i^*, t_i) \tag{6}$$

where l_i denotes the label of the detected object, p_i is the probability distribution of detected object classes evaluated by the softmax function, t_i denotes the predicted five parameterized coordinate vectors, and t_i^* denotes the offset of ground-truth and positive anchors. The hyper-parameter λ in Eq. 6 determines the balance between the two task losses and the λ is set to 1 in this paper. Besides, the functions L_{cls} and L_{reg} are defined as:

$$L_{\text{cls}}(p, l) = -\log pl \tag{7}$$

$$L_{\text{reg}}(t_i^*, t_i) = smooth_{L_1}(t_i^* - t_i) \tag{8}$$

$$smooth_{L_1} = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{9}$$

The parameterized coordinate regression mode is as follows:

$$\begin{cases} t_x = \frac{x - x_a}{w_a}, t_y = \frac{y - y_a}{h_a} \\ t_w = \log\frac{w}{w_a}, t_h = \log\frac{h}{h_a} \\ t_\theta = \theta - \theta_a + k\frac{\pi}{2} \end{cases} \tag{10}$$

$$\begin{cases} t_x^* = \frac{x^* - x_a}{w_a}, t_y^* = \frac{y^* - y_a}{h_a} \\ t_w^* = \log\frac{w^*}{w_a}, t_h = \log\frac{h^*}{h_a} \\ t_\theta^* = \theta^* - \theta_a + k\frac{\pi}{2} \end{cases} \tag{11}$$

where x, y, w and h denote the center coordinates of bounding box and its width and height. Variables x, x_a and x^* are for the predicted bounding box, anchor bounding box, and ground-truth bounding box, respectively (so do for y, w, h). The parameter $k \in Z$ to keep θ in the range $[-90, 0)$. In order to keep the bounding box in the same position, w and h need to be swapped when k is an odd number.

As described in the previous section, we give rotation anchors fixed orientations within the range $[-90, 0)$, and each of the 6 orientations can fit the ground truth that has an intersection angle of less than $\frac{\Pi}{12}$. Thus, every rotation anchor has its fitting range, which we call its fit domain. When an orientation of a ground truth box is in the fit domain of an rotation anchor, this rotation anchor is most likely to be a positive sample of the ground truth box. As a result, the fit domains of the 6 orientations divide the angel range $[-90, 0)$ into 6 equal parts. Thus, a ground truth in any orientation can be fitted with a rotation anchor of the appropriate fit domain.

4 Experimental results

Experiments are performed on the deep learning framework MXNet on a server with GeForce GTX 1080 Ti and 11G memory. We perform experiments on both remote sensing image dataset and scene text dataset to verify the effectiveness and generality of our approach.

4.1 Dataset and setting

DOTA is a large scale dataset for arbitrary-oriented object detection in optical remote sensing images provided by Xia and Bai [27]. It contains 2806 images from different sensors, and each image is of the size in the range from about 800 × 800 to 4000 × 4000 pixels. What is more, the instances in images exhibit a wide variety of scales, orientations and shapes. These images are annotated by experts using 15 categories, containing Plane, baseball diamond, bridge, ground-track-field, small vehicle, large vehicle, ship, tennis-court, basketball court, storage tank, soccer-ball-field, roundabout, harbor, swimming pool, helicopter. DOTA dataset contains 188282 instances, which is labeled by an arbitrary quadrilateral, such as x_0, y_0, x_1, y_1, x_2, y_2, x_3, y_3. Due to significant progress in horizontal bounding-box detection task (HBB), we just evaluate our methods in oriented bounding-box detection task (OBB). We use the scripts called DOTA_devkit to split the images into 1024 × 1024. In the end, we have 14,348 train images and 4871 test images. We trained 60 epochs totally on DOTA. The base learning rate is 5×10^{-4}, and the learning rate changed during 45 and 52 epochs from 5×10^{-4} to 5×10^{-6}.

The public benchmark NWPU WHR-10 [28] contains 10-class geospatial object for detection. These ten classes of objects are airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge and vehicle. This dataset contains totally 800 very-high-resolution (VHR) remote sensing images that were cropped from Google Earth and Vaihingen dataset and then manually annotated by experts. We train the model with 5×10^{-4} learning rate for the first 10 epochs and then 5×10^{-5} for the last 10 epochs.

ICDAR2015 is used in challenge 4 of ICDAR 2015 Robust Reading Competition. It includes 1500 natural images in total, 1000 of which are used for training and the remaining are for testing. The text regions are annotated by 4 points of the quadrangle. We used the image's original resolution 1280 for training and testing. We trained 40 epochs totally on ICDAR2015 and changed the learning rate in 15 epochs and 30 epochs, respectively.

We use the pretrained model DPN-92 to initialize the backbone network. Besides, weight decay and momentum are 1×10^{-4} and 9×10^{-1}, respectively. We employ SGD Optimizer with momentum over 4 GPUs with a total of 4 images per minibatch (1 images per GPU). The anchors have areas of 16^2 to 256^2 on pyramid levels P_2 to P_6, respectively. Furthermore, we just use random flipping as data augmentation.

4.2 Evaluation and ablation study

4.2.1 Baseline setting

In our experiments, faster-RCNN(ResNet)-based detection pipeline is used as the baseline of the ablation experiments. All experiments data and parameter settings are strictly consistent for the fairness and accuracy of the experiments. We use mean average precision(mAP) as a measure of model accuracy performance. The results of DOTA dataset reported here were obtained by submitting our detections to the official DOTA evaluation server. Our method is called AOOD, which uses DPN-92 as backbone and incorporates attention module and dense connected FPN structure.

4.2.2 The effect of backbone

The original faster-RCNN uses VGG-16/ZF as the backbone. Since He proposed the residual neural network [1], it has been widely used as the backbone network for visual tasks such as image recognition, object detection and semantic segmentation. Compared with traditional network, ResNet has deeper network layers, lighter parameters, faster convergence and stronger feature representation. As we all know, more densely and complicated network structure learns more details and more discriminating features. Based on ResNet, we compared some improved networks such as ResNeXt [25], DenseNet [26] and DPN [17]. We all build the detection framework in strict accordance with the faster-RCNN's pattern, which uses the last convolutional feature to generate proposals and feed into regression and classification branches. It is evident from Table 1 that the detection results have been improved after using more feature engineering backbone, and total mAP has increased by about 0.31–1.44. Especially, we can find out that the performance improves greatly when using DPN as the backbone.

4.2.3 The effect of attention

As discussed above, the attention model is beneficial to weaken the interference of the noise and enhance the object feature. Squeeze and Excitation Network (SENet) has proved to be an effective learnable channel attention mechanism. For different backbone networks, we all embed the attention module at the end of each convolutional stage. We think that it will maximize the attention signal. As shown in Table 1, the attention model helps to improve the total detection mAP obviously. Especially, attention module improves the detection accuracy of multiscale and small objects. Compared to no-attention methods, SENet increases mAP by 1.17–1.72.

4.2.4 The effect of dense FPN

Low-level feature contains less semantic information, but the location information is accurate; conversely, high-level features have rich semantic information but coarse location information. It is widely recognized that multi-scale feature

Table 1 Ablative study of each module in our proposed method on DOTA dataset

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
baseline	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
ResNeXt	80.99	66.59	35.44	67.59	60.01	50.71	55.88	90.89	66.51	73.38	55.23	52.32	56.52	54.01	49.56	61.04
DN	80.95	66.98	35.37	68.21	60.05	50.93	56.02	90.89	66.56	72.65	55.41	52.66	55.54	53.65	48.85	60.98
DPN	81.10	69.31	38.73	69.01	61.12	51.68	56.88	90.84	67.12	74.11	55.96	53.60	57.02	55.12	50.15	62.11
DPN + AM	86.25	74.31	39.21	69.11	62.71	51.86	57.52	90.80	68.02	74.01	56.69	56.31	58.20	56.11	53.58	63.65
DN+AM	85.20	72.11	37.75	68.51	61.81	51.64	56.79	90.80	67.78	73.58	55.87	54.71	56.98	55.31	51.68	62.70
ResNeXt + AM	84.92	69.91	37.71	68.44	61.87	51.77	56.01	90.78	66.21	73.47	55.58	53.48	57.01	55.67	50.31	62.21
DPN + DFPN	87.46	75.60	42.41	69.48	63.11	53.32	58.98	90.86	71.93	75.69	57.67	57.17	63.99	66.77	57.43	66.12
DN + DFPN	86.87	75.34	40.25	69.13	62.45	52.69	58.66	90.71	69.92	74.54	55.63	57.51	62.70	62.24	52.78	64.76
RX + DFPN	84.89	74.66	39.86	68.95	62.55	52.93	57.13	90.70	68.54	75.28	56.53	57.27	60.28	63.87	52.84	64.42
AOOD	89.77	80.20	43.89	69.48	67.52	59.03	66.11	90.84	79.55	85.44	62.94	61.54	64.19	67.22	62.00	69.98
AOOD + SR(P)	**89.99**	**81.25**	**44.50**	**73.20**	**68.90**	**60.33**	**66.86**	**90.89**	**80.99**	**86.23**	**64.98**	**63.88**	**65.24**	**68.36**	**62.13**	**71.18**

fusion and context information embedding are very helpful for improving the performance of small targets detection. Dense FPN is selected as the another feature engineering strategy to boost the performance. As shown in Table 1, dense FPN helps improve the small objects detection performance greatly by about 0.47–11.65 in mAP.

4.2.5 The effect of rotated RPN

Some methods use original horizontal regions proposal network, but regress to oriented bounding box, such as R^2CNN and faster-RCNN-for-DOTA. R^2CNN regresses the horizontal proposal to the coordinate representation of 5 values (x, y, w, h, θ), but the latter regresses to 8 values $(x_0, y_0, x_1, y_1, x_2, y_2, x_3, y_3)$. The regression from horizontal proposal to oriented detection box is inefficient and not robust, which often causes a large coordinate offset. In addition, the 8-value representation of the regression targets even leads to irregular and non-rectangle detection box. Besides, (x, y, w, h, θ) is a rotation-friendly representation for the angle regression, and it is easy to calculate the angle offset between two different rotated boxes. So we compare the RRPN-based method with the traditional horizontal RPN-based structure. It is obvious to see from Fig. 6 that RRPN-based method generates more robust bounding boxes, which have more standardized shape. Also, the mAP is increased by about 1.56 as shown in Fig. 7.

4.2.6 The effect of image super-resolution and image pyramid

Super-resolution is a very important image quality enhancement technology. Although the picture quality of the DOTA dataset is not bad, the resolution of the image after cropping needs to be improved. We use the RCAN's [29] pretrained model on DIV2K dataset [30] to fine-tune the split DOTA. Considering that the depth of the convolutional neural network is critical to the image SR effect, simply splicing the residual modules together to build a deeper network does not result in better improvements. Therefore, we have improved on the basis of the pretraining model and changed the RCAB structure to densely connected, so that the low-level features can be better propagated in the network, making full use of the low-frequency information of the image, thus making the network more focused on learn high-frequency information. While fine-tuning, we use the ADAM optimizer and set $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$. The initial learning rate is set to 1×10^{-4}. Furthermore, image pyramid training and testing is an effective method to gain improvement. In our experiments, we scale the original

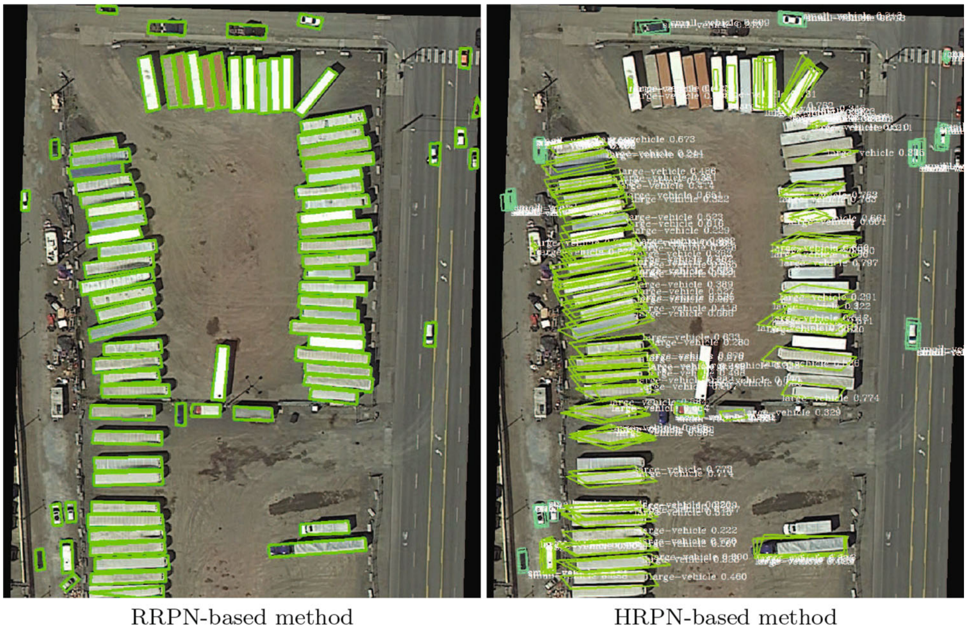


Fig. 6 Comparison between RRPN and HRPN

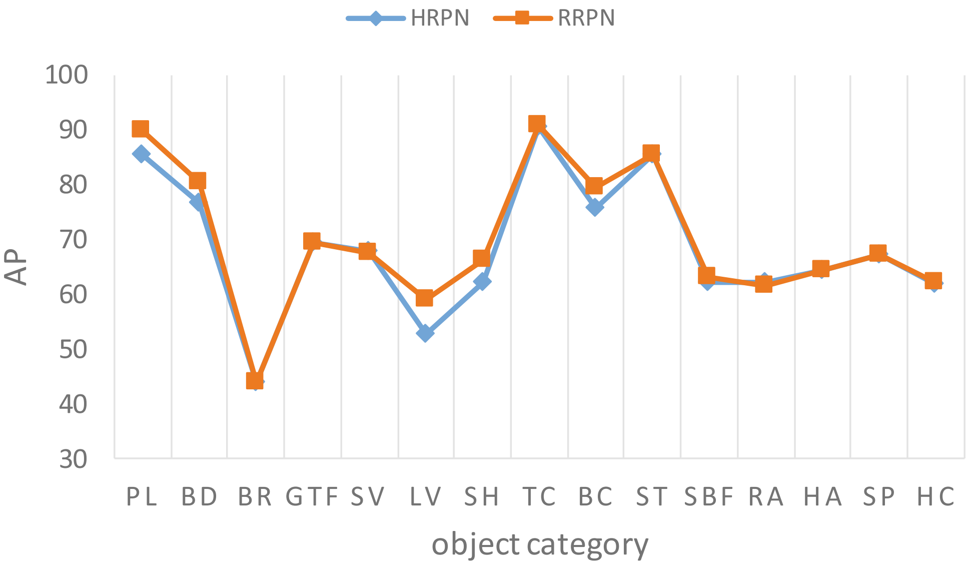


Fig. 7 Ablative study on the effect of Rotated RPN. HRPN and RRPN denote original horizontal region proposal network and oriented region proposal network, respectively. Our method is called AOOD, which uses the DPN as backbone and adds attention module and dense connected FPN (DPN + AM + DFPN)

spitted image (1024 × 1024) to [800 × 800, 1024 × 1024, 1280 × 1280] and then send it to train and test. Note that our final detection results are generated by R-NMS. As shown in Table 1, super-resolution and image pyramid, i.e., SR(P), can improve performance steadily and get 71.18 mAP finally.

4.3 Performance on benchmark

The proposed method is compared to the state-of-the-art object detectors on three benchmarks: DOTA, NWPU VHR-10 and ICDAR2015. Our model achieves competitive performances in all three benchmarks.

4.3.1 DOTA

To verify the superiority of our method, we compare with AOVD [31], R-DFPN [10], ICN [32], R2CNN++ [13] and so on, which are all enable to detect multi-class arbitrary orientation objects. Table 2 shows the performance of these methods. Because of the feature fusion and attention, R2CNN++ and our method get excellent detection performance in small objects. Our approach focuses on enhancing the informative information and robustness of features by introducing densely connected FPN and attention module. The experiments show that our method reaches 71.18 mAP, achieving the best performance.

Table 2 Comparative experiment on DOTA dataset

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
SSD [7]	39.83	9.09	0.64	13.18	0.26	0.39	1.11	16.24	27.57	9.23	27.16	9.09	3.03	1.05	1.01	10.59
YOLOv2 [6]	39.57	20.29	36.58	23.42	8.85	2.09	4.82	44.34	38.35	34.65	16.02	37.62	47.23	25.5	7.45	21.39
R-FCN [8]	37.80	38.21	3.64	37.26	6.74	2.60	5.59	22.85	46.93	66.04	33.37	47.15	10.60	25.19	17.96	26.79
FR-H [5]	47.16	61.00	9.80	51.74	14.87	12.80	6.88	56.26	59.97	57.32	47.83	48.70	8.23	37.25	23.05	32.29
FR-O [34]	79.09	69.12	17.17	63.49	34.20	37.16	36.20	89.19	69.60	58.96	49.4	52.52	46.69	44.80	46.30	52.93
R-DFPN [10]	80.92	65.82	33.77	58.94	55.77	50.94	54.78	90.33	66.34	68.66	48.73	51.76	55.10	51.32	35.88	57.94
R2CNN [12]	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
AOVD [31]	88.52	71.20	31.66	59.30	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 [32]	81.40	74.30	47.70	70.30	64.90	67.80	70.00	90.80	79.10	78.20	53.60	62.90	67.00	64.20	50.20	68.20
RoITransformer [35]	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
R2CNN++ [13]	89.66	81.22	**45.50**	75.10	68.27	60.17	66.83	**90.90**	80.69	86.15	64.05	63.48	**65.34**	68.01	62.05	71.16
AOOD+SR(P)	**89.99**	**81.25**	44.50	73.20	**68.90**	60.33	66.86	90.89	**80.99**	**86.23**	**64.98**	**63.88**	65.24	**68.36**	**62.13**	**71.18**

Visualized presentation of object detection on the DOTA dataset is shown in Fig. 8.

4.3.2 NWPU VHR-10

NWPU VHR-10 contains 10-class geospatial object for detection. We compare it with seven methods and achieve the best detection performance, at 89.10. Our model achieves the best performance in more than half of the categories. The specific results are shown in Table 3.

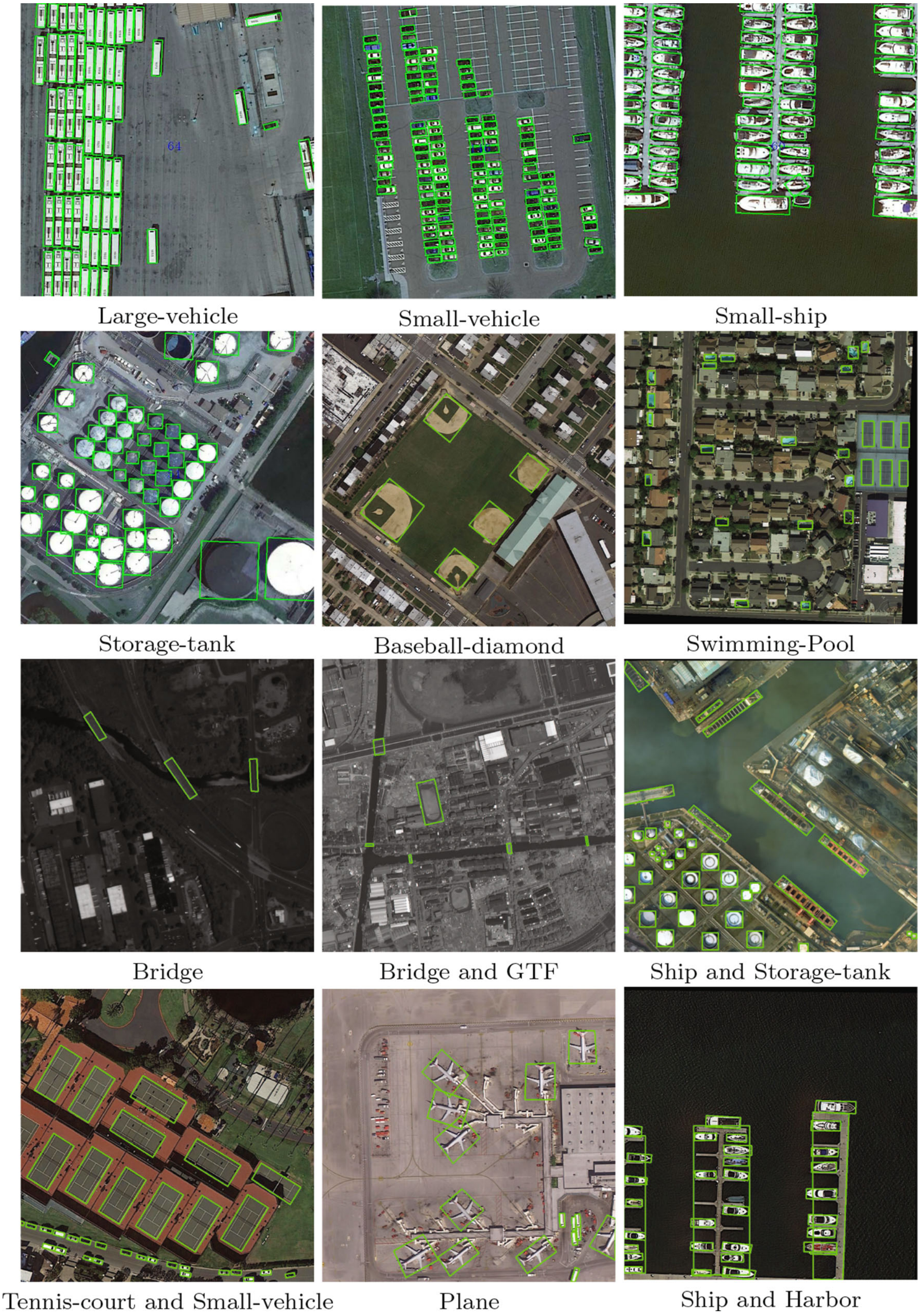
4.3.3 ICDAR2015

Scene text detection is also a main application scenario of rotation detection. We used EAST [33], RRPN [14] and R2CNN [12] for comparative experiments. Table 4 shows the performance of these methods, our method achieves 82.64% in the ICDAR2015 dataset, better than most mainstream algorithms. The precision–recall curves of AOOD on the ICDAR2015 dataset is illustrated in Fig. 9. It proves that the proposed method is useful for both remote sensing images and scene texts.

4.4 computational cost analysis

The proposed method comprises multiple modules, such as Dual Path Network (DPN) backbones module, dense FPN module, rotation region proposal module and rotation fast RCNN module under the faster RCNN like framework. So, we compared our proposed model with another classic architecture faster RCNN on FPN, which backbone is ResNet-101 network, and also have FPN module, region proposal module and fast RCNN module. The DPN-92 costs about 15% fewer parameters than ResNeXt-101 which is more complicated than ResNet-101. In terms of computational complexity, DPN-92 consumes about 19% less FLOPs than ResNeXt-101. Then, according to the analysis in the SENet [19], adding the SE module to the original network will only introduce less than 1% of additional calculations, but will bring a significant increase in network accuracy.

Compared to FPN, dense FPN has feature propagation between each layer. But only the upsampling operation with low computational complexity is added. So the dense FPN module does not take more time. The difference between rotation RPN network and RPN network is that an angle parameter is introduced to the anchors whose time cost is almost negligible. Moreover, the input of the rotation fast RCNN module are proposals. Others are consistent with the original fast RCNN module. Finally, SR is only used as a pre-processing step. If you want to get the detection results quickly, you do not have to do the pre-



Large-vehicle

Small-vehicle

Small-ship

Storage-tank

Baseball-diamond

Swimming-Pool

Bridge

Bridge and GTF

Ship and Storage-tank

Tennis-court and Small-vehicle

Plane

Ship and Harbor

Fig. 8 Visualized presentation of object detection in sensing image

Table 3 Comparative experiment on NWPU VHR-10 dataset

Detection method	mAP
R-P-faster RCNN [36]	76.50
SSD512 [37]	78.40
DSSD321 [38]	78.80
DSOD300 [39]	79.80
Deformable R-FCN [40]	79.10
Deformable faster RCNN [34]	84.40
RICADet [41]	87.12
AOOD (proposed)	**89.10**

Table 4 Comparative experiment on ICDAR2015 dataset

Method	Recall	Precision	F-measure	Res.
CTPN [42]	51.56	74.22	60.85	–
SegLink [43]	76.80	73.10	75.00	–
EAST [33]	78.33	83.27	80.72	720P
RRPN [14]	82.17	73.23	77.44	–
R2CNN [12]	79.68	**85.62**	82.54	720P
AOOD (proposed)	**82.64**	85.35	**83.56**	720P

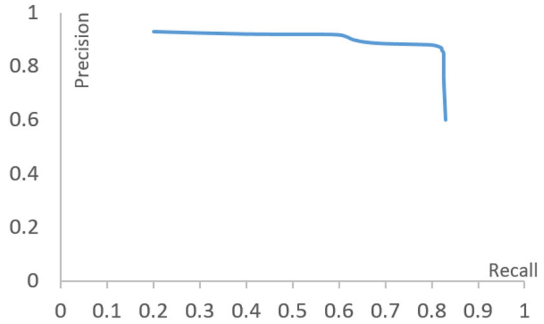


Fig. 9 Precision/recall curve

processing step. But if you want better results, you need to use it. In conclusion, we focused on the problems of object detection for remote sensing image and made many improvements on the basis of faster RCNN on FPN network, and these operations are not time-consuming.

5 Conclusion and future plan

In summary, this paper proposes a arbitrary-oriented object detection method, which has the following property: (1) To enhance the feature re-usage and new features exploration, the DPN and dense FPN are simultaneously exploited to act as backbone network and generate feature pyramid feature map, which will produce informative feature and discriminative multi-scale feature maps by introducing residual path and densely connected paths; (2) SE attention model is leveraged to activate the channels useful to object detection while suppressing the channel closely related to the noise; (3) rotation region proposal and rotation ROI pooling strategies are integrated into the architecture to produce minimum circumscribed rectangle bounding box, efficiently reducing the redundant detection region. In spite of this, some performance boosting strategies such as dilated convolution, smaller orientation interval and context information are not considered, which will be exploited in the future work.

Acknowledgements This work is supported in part by the National Natural Science Foundation of China under Grant Nos. 61672254, 61672246, 61572221 and 61300222, Program for Hust Academic Frontier Youth Team, Key project of National Natural Science Foundation of China Grant No. U1536203, Natural Science Foundation of Hubei Province Grant No. 2015CFB687 and the Fundamental Research Funds for the Central Universities, HUST: 2016YXMS088 and 2016YXMS018. The authors appreciate the valuable suggestions from the anonymous reviewers and the Editors.

References

1. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition (CVPR 2016), 2016, pp 770–778
2. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the 2014 IEEE conference on computer vision and pattern recognition (CVPR), 2014, pp 580–587
3. He K, Zhang X, Ren S, Sun J (2014) Spatial pyramid pooling in deep convolutional networks for visual recognition. In: Proceedings of the 13th European conference on computer vision (ECCV 2014), 2014, pp 346–361
4. Girshick R (2015) Fast R-CNN [region-based Convolutional Neural Network]. In: Proceedings of the 2015 IEEE international conference on computer vision (ICCV), 2015, pp 1440–1448
5. Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149
6. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified real-time object detection. In: Proceedings of the 2016 IEEE conference on computer vision and pattern recognition (CVPR), 2016, pp 779–788
7. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC (2016) SSD: single shot multibox detector. In: Proceedings of the 14th European conference computer vision (ECCV2016), 9905, pp 21–37
8. Dai J, Li Y, He K, Sun J (2016) R-FCN: object detection via region-based fully convolutional networks. In: Proceedings of the 2016 conference on advances in neural information processing systems (NIPS), pp 379–387
9. He K, Gkioxari G, Dollar P, Girshick R (2017) Mask R-CNN. In: Prcoeedings of the 2017 IEEE international conference on computer vision (ICCV), 2017, pp 2980–2988
10. Yang X, Sun H, Kun F, Yang J, Sun X, Yan M, 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 Sens 10(1):132–146

11. Yang X, Sun H, Sun X, Yan M, Zhi G, Kun F (2018) Position detection and direction prediction for arbitrary-oriented ships via multitask rotation region convolutional neural network. IEEE Access 6:50839–50849
12. Jiang Y, Zhu X, Wang X, Yang S, Li W, Wang H, Fu P, Luo Z, R2CNN: rotational region CNN for orientation robust scene text detection. arXiv:1706.09579
13. Yang X, Fu K, Sun H, Yang J, Guo Z, Yan M, Zhang T, Xian S, R2CNN++: multi-dimensional attention based rotation invariant detector with robust anchor strategy. arXiv:1811.07126
14. Ma J, Shao W, Ye H, Wang L, Wang H, Zheng Y, Xue X (2018) Arbitrary-oriented scene text detection via rotation proposals. IEEE Trans Multimedia 20(11):3111–3122
15. Shermeyer J, Van Etten A (2019) The effects of super-resolution on object detection performance in satellite imagery. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops
16. Haris M, Shakhnarovich G, Ukita N (2018) Task-driven super resolution: object detection in low-resolution images. arXiv preprint arXiv:1803.11316
17. Chen Y, Li J, Xiao H, Jin X, Yan S, Feng J (2017) Dual path networks. In: Proceedings of the 2017 conference on advances in neural information processing systems, 2017, pp 4468–4476
18. Lin T-Y, Dollar P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: Proceedings of the 2017 IEEE conference on computer vision and pattern recognition (CVPR), 2017, 936–944
19. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation Networks. In: Proceedings of the 2018 IEEE conference on computer vision and pattern recognition, 2018, pp 7132–7141
20. Uijlings JRR, van de Sande KEA, Gevers T, Smeulders AWM (2013) Selective search for object recognition. Int J Comput Vis 104(2):154–171
21. Shamsolmoali P, Zareapoor M, Wang R et al (2019) A novel deep structure U-net for sea-land segmentation in remote sensing images. IEEE J Sel Top Appl Earth Observ Remote Sens 12(9):3219–3232
22. Shamsolmoali P, Zareapoor M, Wang R et al (2019) G-GANISR: gradual generative adversarial network for image super resolution. Neurocomputing 366:140–153
23. Li F et al (2017) Super-resolution for GaoFen-4 remote sensing images. IEEE Geosci Remote Sens Lett 15(1):28–32
24. Wu W et al (2016) A new framework for remote sensing image super-resolution: sparse representation-based method by processing dictionaries with multi-type features. J Syst Archit 64:63–75
25. Xie S, Girshick R, Dollar P, Tu Z, He K (2017) Aggregated residual transformations for deep neural networks. In: Proceedings of the 2017 IEEE conference on computer vision and pattern recognition (CVPR), 2017, pp 5987–5995
26. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the 2017 IEEE conference on computer vision and pattern recognition (CVPR), 2017, pp 2261–2269
27. Xia G-S, Bai X, Ding J, Zhu Z, Belongie S, Luo J, Datcu M, Pelillo M, 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 (CVPR) 2018, pp 3974–3983
28. Cheng G, Zhou P, Han J (2016) Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. IEEE Trans Geosci Remote Sens 54(12):7405–7415
29. Zhang Y, Li K, Li K, Wang L, Zhong B, Yun F (2018) Image super-resolution using very deep residual channel attention networks. In: Proceedings of the IEEE conference on ECCV 2018, pp 1–16
30. Eirikur A, Radu T (2017) NTIRE 2017 challenge on single image super-resolution: dataset and study. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) 2017, pp 1–10
31. Tang T, Zhou S, Deng Z, Lei L, Zou H (2017) Arbitrary oriented vehicle detection in aerial imagery with single convolutional neural networks. Remote Sens 9(11):1170
32. Azimi SM, Vig E, Bahmanyar R, Körner M, Reinartz P (2018) Towards multi-class object detection in unconstrained remote sensing imagery. arXiv preprint, arXiv:1807.02700
33. Zhou X et al (2017) EAST: an efficient and accurate scene text detector. In: Proceedings of the IEEE conference on computer vision and pattern recognition
34. Ren Y, Zhu C, Xiao S (2018) Deformable faster R-CNN with aggregating multi-layer features for partially occluded object detection in optical remote sensing images. Remote Sens 10(9):1470
35. Ding J et al (2018) Learning ROI transformer for detecting oriented objects in aerial images. arXiv preprint arXiv:1812.00155
36. Han X, Zhong Y, Zhang L (2017) An efficient and robust integrated geospatial object detection framework for high spatial resolution remote sensing imagery. Remote Sens 9(7):666
37. Liu W et al (2016) SSD: single shot multibox detector. In: European conference on computer vision. Springer, Cham
38. Fu C-Y et al (2017) DSSD: deconvolutional single shot detector. arXiv preprint arXiv:1701.06659
39. Shen Z et al (2017) DSOD: learning deeply supervised object detectors from scratch. In: Proceedings of the IEEE international conference on computer vision
40. Xu Z et al (2017) Deformable convnet with aspect ratio constrained NMS for object detection in remote sensing imagery. Remote Sens 9(12):1312
41. Li K et al (2017) Rotation-insensitive and context-augmented object detection in remote sensing images. IEEE Trans Geosci Remote Sens 56(4):2337–2348
42. Tian Z et al (2016) Detecting text in natural image with connectionist text proposal network. In: European conference on computer vision. Springer, Cham
43. Shi B, Bai X, Belongie S (2017) Detecting oriented text in natural images by linking segments. In: Proceedings of the IEEE conference on computer vision and pattern recognition

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.