Applied Intelligence (2022) 52:15278–15294
https://doi.org/10.1007/s10489-022-03393-8

RAOD: refined oriented detector with augmented feature in remote sensing images object detection

Qin Shi[1] · Yu Zhu[1] · Chuantao Fang[1] · Nan Wang[1] · Jiajun Lin[1]

Accepted: 15 February 2022 / Published online: 13 March 2022

Abstract

Object detection is a challenging task in remote sensing. Aerial images are distinguished by complex backgrounds, arbitrary orientations, and dense distributions. Considering those difficulties, this paper proposes a two-stage refined oriented detector with augmented features named RAOD. First, a novel Augmented Feature Pyramid Network (A-FPN) is built to enhance fusion both in spatial and channel dimensions. Specifically, it mainly consists of three modules: Scale Transfer Module (STM), Feature Aggregate Module (FAM) and Feature Refinement Module (FRM). STM reduces information loss when fusing features in the top-down pathway. FAM aggregates features from different scales. FRM aims to refine the integrated features using a lightweight attention module. Then, we adopt a two-step processing, which consists of a coarse stage and a refinement stage. In the coarse stage, deformable RoI pooling is adopted to improve the network's ability of modeling spatial transformations and then horizontal proposals are transformed into oriented ones. In the refinement stage, Rotated RoI align (RRoI align) is used to extract rotation-invariant features from rotated RoIs and further optimize the localization. To enhance stability and robustness during training, smooth Ln is chosen as regression loss as it has better ability in terms of robustness and stability than smooth L_1 loss. Extensive experiments on several rotation detection datasets demonstrate the effectiveness of our method. Results show that our method is able to achieve 79.78%, 74.7% and 94.82% on DOTA-v1.0, DOTA-v1.5 and HRSC2016, respectively.

Keywords Remote sensing image · Oriented object detection · Augmented feature pyramid · Deformable RoI pooling · Rotated RoI align

1 Introduction

Object detection aims to localize the objects and identify their categories. As a significant task in remote sensing image processing, object detection has wide applications in civil and military fields. Recently, object detection based on convolutional neural networks (CNNs) has made great progress. Many detection algorithms such as Faster RCNN [1], YOLO [2], SSD [3] and RetinaNet [4] achieve promising performances in natural image scenes. Compared with natural images, objects in aerial images are often densely packed and have arbitrary orientations, various appearances and complex backgrounds. General object detectors based on horizontal bounding boxes report worse performance when directly applying to aerial images. Current popular oriented object detection algorithms employ different strategies to achieve better detection results. For example, R^{3}Det [5], S^{2}A-Net [6] focus on designing feature alignment modules. SCRDet [7], RSDet [8] explore new loss functions and CSL [9], DCL [10] employ label techniques. Extracting accurate features plays a crucial role in object detection and recognition, such as SIFT [11], HOG [12] and SDD [13]. Traditional feature extraction methods are difficult to be applied effectively in aerial images due to the complex backgrounds, large

✉ Yu Zhu
zhuyu@ecust.edu.cn

Qin Shi
sq15052502008@126.com

Chuantao Fang
feraint@outlook.com

Nan Wang
wangnan@ecust.edu.cn

Jiajun Lin
jjlin@ecust.edu.cn

[1] School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China

aspect ratios and dense distributions. Inappropriate feature extraction is one of the most important reasons for low performance when detecting rotated instances in aerial images. To achieve high accuracy in aerial object detection, a refined two-stage oriented detector with augmented features (RAOD) is proposed.

First, it is noticed that FPN [14] suffers from some limitations, which hinder its ability to extract representative visual features. FPN is an effective component in object detection frameworks, which extracts multi-scale features. Specifically, FPN adopts 1×1 convolution to reduce channel dimensions of feature maps from the backbone. Then, FPN upsamples higher level feature maps by a factor of 2 and then merges feature maps of the same resolution by element-wise addition in a top-down pathway. FPN-based methods detect objects at different scales, alleviating the conflicts between the spatial resolution and receptive fields to some extent. However, there are some intrinsic flaws in FPN: 1) The reduction of channels at each level leads to information loss. The outputs of last four residual blocks of ResNet [15] have channels of $\{256, 512, 1024, 2048\}$, respectively. The channels of higher-level feature maps are reduced to a smaller constant of 256. The decay of channel-wise information degrades the feature representation of networks to a certain extent. 2) Inadequate cross-scale feature fusion. The nearest neighbor interpolation operation obtains the features of floating-point coordinate points by adjacent pixels and lacks global semantic information. FPN directly sums up feature maps after nearest neighbor upsampling without considering the semantic gap between different feature maps, resulting in aliasing effects. 3) Lack of communications between non-adjacent levels. The high-level semantic features and low-level content features are complimentary for the task of object detection [16]. Nevertheless, deep semantical information is weaken gradually in the top-down pathway and non-adjacent feature maps do not interact with each other well. To this end, this paper devises a novel feature pyramid network that can tackle the above problems and boost the detection accuracy effectively.

Second, horizontal proposals can achieve higher recall and speed while rotated proposals perform better in oriented and densely packed scenes [17]. Thus, this paper adopts a coarse-to-fine manner that uses horizontal anchors in the coarse stage and rotated proposals in the refined stage. In order to extract a fixed-size feature map (e.g., 7×7) from each proposal generated by Region Proposal Networks (RPNs), RoI pooling and RoI align are commonly used in two-stage object detectors [1, 18]. However, the regular operators have limitations in modeling geometric transformations [19], leading to poor performance in detecting various objects in aerial images. To this end, common RoI align is replaced with deformable RoI pooling [20] in the coarse stage and Rotation RoI align (RRoI align) [19] is adopted in the refined stage. Deformable RoI pooling adds 2D offsets to the sampling points in regular RoI pooling, enabling to adaptively localize objects with different shapes. RRoI align produces horizontal fixed-size features maps from regions with different scales, aspect ratios and angles, enabling the network to obtain rotation-invariant features for more robust detection of oriented objects.

In addition, a smooth Ln loss [21] is adopted to regress the position of arbitrarily rotated objects to enhance the robustness and stability of training. Our main contributions are summarized as follows:

- To address limitations in original FPN [14], a simple yet effective feature pyramid network named A-FPN is devised. We design three modules which are tailored to obtain augmented multi-scale features and can be easily plugged into FPN-based models.
- Towards high-quality aerial object detection, we develop a coarse-to-fine oriented detector. In the coarse stage, geometry-robust features are extracted to facilitate the transformation from horizontal bounding boxes to oriented bounding boxes. In the refinement stage, rotation-invariant features are obtained for better detection of arbitrarily rotated objects.
- For more accurate localization of oriented objects in aerial images, we choose smooth Ln loss [21] in the regression branch. Compared with smooth L_1 loss [22], it has better performance in robustness and stability.
- Our proposed method achieves state-of-the-art performances on three public large-scale datasets for aerial object detection, including DOTA [23] and HRSC2016 [24].

The overview of the paper is organized as follows. Section 2 introduces the related work in oriented object detection in deep neural networks, multi-scale feature fusion and feature modeling. Then, the proposed RAOD is described in Section 3. Next, the comprehensive experiments are conducted in Section 4. Finally, Section 5 concludes the whole work.

2 Related work

CNNs have been widely used in remote sensing image object detection. Some representative approaches based on CNNs are introduced in Section 2.1. Then we focus on multi-scale feature fusion in Section 2.2 and feature modeling in Section 2.3.

2.1 Oriented object detection

General object detection methods based on horizontal bounding box often suffer from misalignment between objects and RoIs. For example, Faster R-CNN [1], YOLO [2], RetinaNet [4] and SSD [3]. In recent years, many well-designed methods adapt general object detectors to aerial images domain and achieve promising performance on the challenging aerial object detection benchmarks (e.g., DOTA [23] and HRSC2016 [24]). RoITransformer [19] learns the transformation from horizontal bounding boxes to rotated ones. SCRDet [7] proposes a novel loss to solve the boundary problem caused by the periodicity of angle. RSDet [8] proposes a modulated rotation loss to address the loss discontinuity. Gliding Vertex [25] employs quadrilateral regression prediction to detect multi-oriented objects more accurately. MRDet [26] decouples detection into different subtasks. ReDet [27] proposes Rotation-invariant RoI Align to extract rotation-equivariant features from the backbone. [7, 8, 19, 25–27] are representative two-stage methods which achieve high detection accuracy. To further improve detection speed, many one-stage methods are proposed. R^{3}Det [5] extracts the features from corners and centers of the anchors and then reconstructs the feature map to solve the inconsistency in existing single-stage detectors. S^{2}A-Net [6] not only realizes the alignment between anchors and convolutional features but also alleviates the misalignment between classification and localization. SCRDet++ [28] designs instance level feature denoising module to improve detection for small and cluttered objects. Considering the complexity of pre-defined anchors, some anchor-free algorithms have been devised. BBAvectors [29] proposes an anchor-free detector that regresses the box boundary-aware vectors based on the center keypoints of arbitrarily oriented objects. Oriented reppoints [30] employs a set of adaptive points as the representation of oriented objects, in order to capture the geometric and semantic information for robust detection. O^{2}DETR [31] designs an efficient rotated detector based on transformer [32] by replacing the original self-attention mechanism in DETR [33] with depthwise separable convolutions, which implements an end-to-end detection framework in oriented object detection. Among the existing detection methods in aerial images, two-stage oriented object detectors based on anchors enjoy relatively higher accuracy, thus are still in a dominant position. The above methods show excellent performance, but most of them do not make full use of features which are beneficial to enhance detection accuracy. Therefore, this paper proposes a method for oriented detection in remote sensing images, which can extract discriminative features and achieve advanced performance.

2.2 Multi-scale feature fusion

FPN [14] develops an effective framework by fusing features from different scales. It is well-known that FPN significantly improves the performance of object detection and instance segmentation and is further studied in many works. PANet [34] shortens the pathway between the highest level and lower levels by adding an extra bottom-up pathway. AugFPN [35] proposes three sub-modules for FPN and improves the fusion. NAS-FPN [36] automatically explores feature framework topology in a scalable search space. BiFPN [37] creates a weighted bi-directional FPN-based structure while applying feature fusion repeatedly. AC-FPN [38] introduces context and content mechanisms into the FPN framework to alleviate the conflict between receptive fields and resolution. CE-FPN [39] enhances channel information and makes full use of semantical features from the topmost level. DRFPN [40] plugs two attention modules into FPN to relax the discrepancy from feature map level and pixel level. FPT [41] adopts three kinds of transformers, enabling them to interact features across scales and space. We argue that the detection accuracy can be further improved by introducing an effective feature pyramid network. Therefore, this paper devises a novel feature pyramid network named A-FPN which contributes to the detection performance.

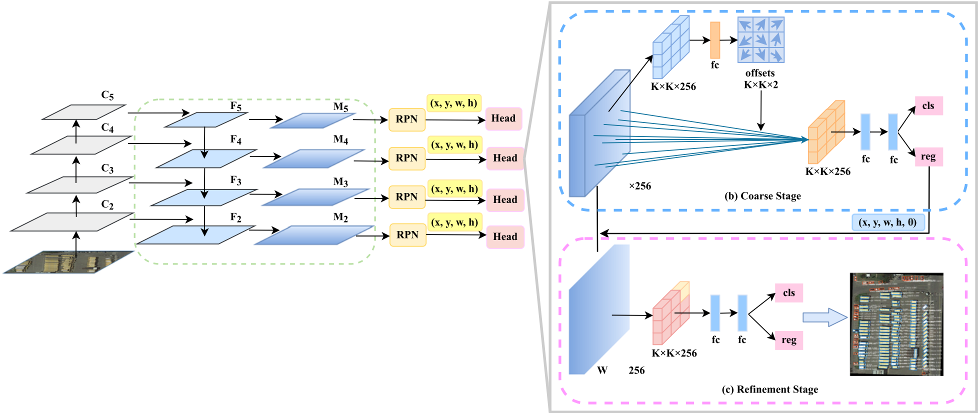
2.3 Feature modeling

Objects in aerial images often have arbitrary orientations and various appearances. There are inherent limitations in CNNs when modeling scale and rotation transformations [20]. STN [42] inserts an additional learnable module which performs spatial transformation. DCN [20] explicitly model the deformation at the image level. ORNs [43] builds a CNN framework with orientation information encoded explicitly by active rotating filters. Common RoI operators (e.g., RoI pooling [1] and RoI align [18]) are widely used to extract fixed-size features at the instance level, which show poor performance in handling geometric variations in aerial images. To enhance the capability of feature modeling, this paper adopts deformable RoI pooling [20] to extract more robust features in the coarse stage and RRoI align [19] to maintain rotation invariance in the refined stage.

3 Method

In this section, we firstly describe the pipeline of our proposed RAOD in Section 3.1. Secondly, the framework of our proposed A-FPN is presented in Section 3.2. Detailed introductions of the coarse adjustment stage and refinement

Fig. 1 An overview of RAOD



stage can be found in Section 3.3 and Section 3.4, respectively. Finally, the matching strategy and loss function are introduced in Section 3.5.

3.1 Architecture network

The architecture of our proposed RAOD is illustrated in Fig. 1. RAOD mainly consists of a backbone network, an Augmented Feature Pyramid Network (A-FPN), a region proposal network (RPN) and a detection head. Specifically, we firstly design an Augmented Feature Pyramid Network (A-FPN) to enhance the representation ability by augmenting and refining the multi-scale features from the CNN backbone. The detection head consists of two stages: a coarse stage and a refinement stage. In the coarse stage, this paper performs deformable RoI pooling [20] on the horizontal proposals to obtain geometric robust features and then learns the transformation from horizontal RoIs to rotated RoIs. In the refinement stage, RRoI align [19] is adopted to extract rotation-invariant features from rotated RoIs.

3.2 Augmented feature pyramid network

To address the limitations in FPN [14], this paper proposes a novel Augmented Feature Pyramid Network (A-FPN). The overall framework of A-FPN is illustrated in Fig. 2. Basically, three modules are designed: Scale Transfer Module (STM), Feature Aggregate Module (FAM) and Feature Refinement Module (FRM). According to the setting of FPN, the multi-scale feature maps from the backbone used to build the feature pyramid network are denoted as $\{C_2, C_3, C_4, C_5\}$,which have channels of $\{256, 512, 1024, 2048\}$ and strides of $\{4, 8, 16, 32\}$ pixels with respect to the input image. $\{F_2, F_3, F_4, F_5\}$ represent the feature maps with the same channels of 256.

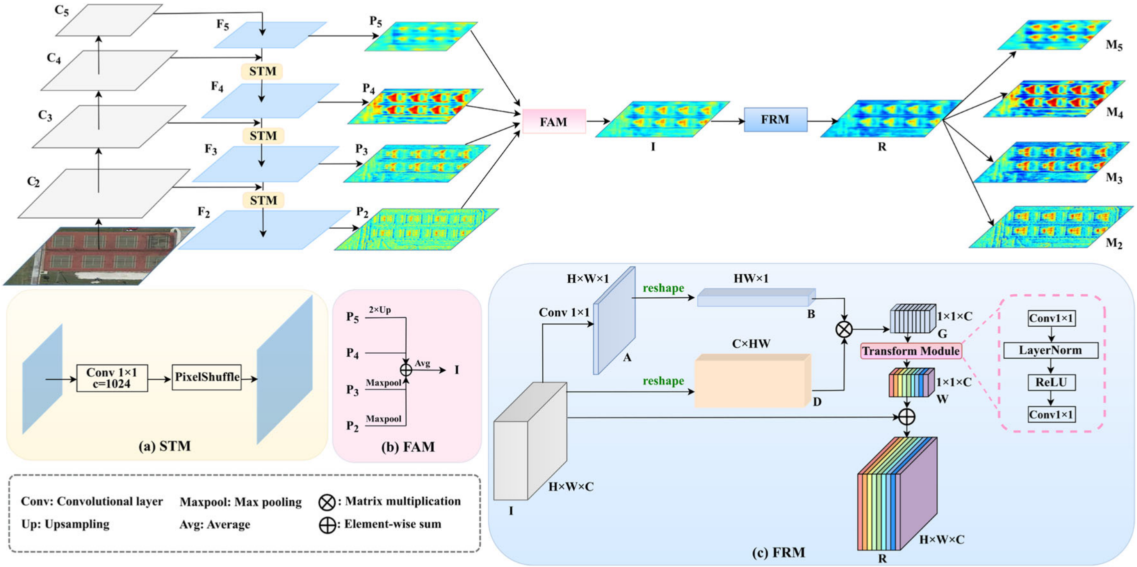


Fig. 2 The framework of A-FPN and visualizations of heatmaps

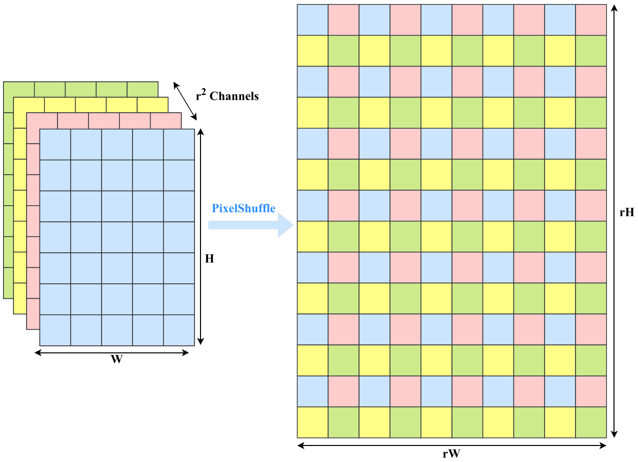


Fig. 3 Illustration of sub-pixel convolution layer. W and H represent width and height of the feature map, respectively. r denotes upscale factor and equals 2 in the graph

$\{P_2, P_3, P_4, P_5\}$ are the fused feature maps after the 3 × 3 convolution layer. The final outputs of our proposed A-FPN are denoted as $\{M_2, M_3, M_4, M_5\}$.

3.2.1 Scale Transfer Module (STM)

FPN [14] reduces the channels of C_i to the same constant 256 through a 1 × 1 convolution layer and then upsamples the higher-level feature map (using nearest neighbor upsampling) before element-wise addition in the top-down path. This process leads to the loss of channel information and aliasing effect, degrading the ability of feature representation. This paper introduces a new fusion module which consists of two steps. First, the channel dimension of F_i is extended to 1024 by applying a 1 × 1 convolution layer. Second, sub-pixel convolution [44] is performed to upsample the feature map with lower resolution. Then, F_{i+1} and F_i are merged through element-wise summation. Sub-pixel convolution transforms a $H \times W \times C \times r^2$ feature map to a feature map of shape $rH \times rW \times C$. Mathematically, the operation of the sub-pixel convolution layer can be described as follow:

$$PixelShuffle(I)_{x,y,c} = I_{\lfloor x/r \rfloor, \lfloor y/r \rfloor, c+C\cdot \bmod (x,r)+C\cdot r\cdot \bmod (y,r)} \tag{3.1}$$

where I represents the input feature map, (x, y) and c indexes the spatial location and channel of the output feature map. The hyperparameter r indicates upscale factor which is set to 2 by default, indicating that every 4 channels are used to expand into a single channel feature with spatial size doubled (as shown in Fig. 3). STM enables the network to shuffle the feature in the channel dimension and augment the spatial information.

3.2.2 Feature aggregate module (FAM)

High-level feature maps contain strong semantic meanings to detect small objects. Low-level feature maps are rich in detailed content and are more suitable to detect large objects. Features from different levels can facilitate each other. Nevertheless, the deep semantic feature is mitigated gradually in the top-down information flow and the feature map on each level lacks attention to the non-adjacent level. To make better use of features from each level, this paper proposes Feature Aggregate Module (FAM). Firstly, the features $\{P_5, P_3, P_2\}$ are resized to the same resolution as P_4. More specifically, we perform the nearest interpolation on P_5 and adaptive max pooling on P_3, P_2, respectively. Then the integrated feature map P is obtained by calculating averege of $\{P_2, P_3, P_4, P_5\}$:

$$P = \frac{1}{4}\sum_{i=2}^{5} P_i \tag{3.2}$$

3.2.3 Feature refinement module (FRM)

After obtaining the coarse integrated feature map from FAM, the Feature Refinement Module (FRM) is incorporated to reduce the aliasing effect and to further strengthen the representation ability of the model. The lightweight global context (GC) block [45] is chosen as the attention module because it can effectively model the global context

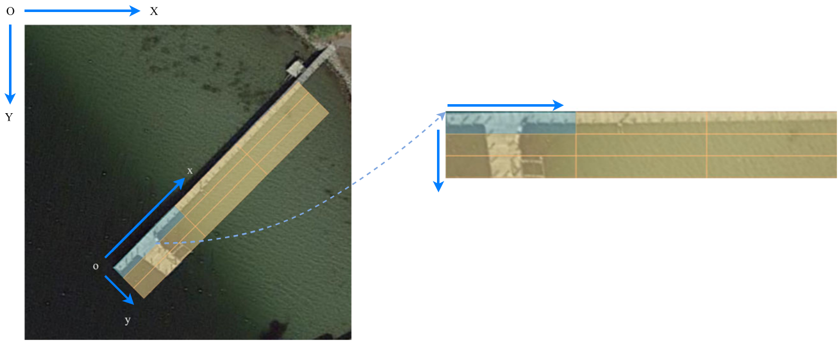
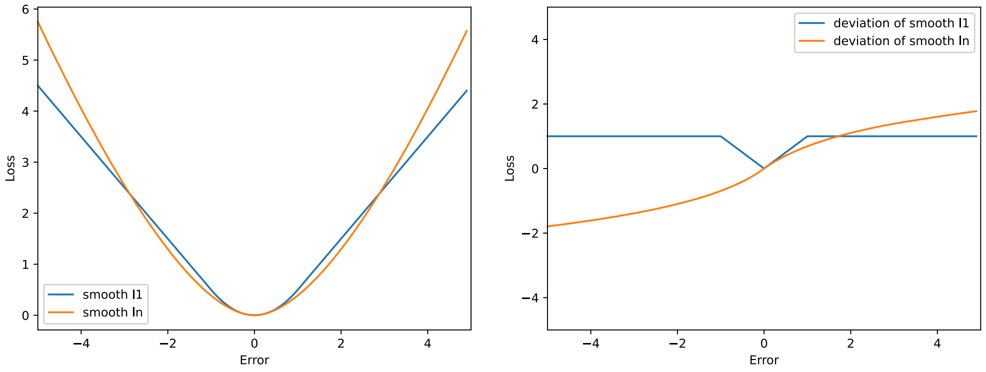


Fig. 4 Illustration of global horizontal coordinate system XOY bounding to the feature map and local oriented coordinate system xOy bounding to the rotated RoI

Fig. 5 Differences between smooth L_1 and smooth Ln loss functions. γ in smooth L_1 is set as 1.0



(a) forward loss functions.

(b) backward deviation functions.

and capture channel-wise interdependencies. As illustrated in Fig. 2, given the integrated feature map $I \in \mathbb{R}^{C\times H\times W}$, we reduce the number of channels to 1 through a 1 × 1 convolution layer and reshape it to $B \in \mathbb{R}^{HW\times 1\times 1}$. Then softmax function is applied to calculate the attention weights:

$$S_j = \frac{e^{x_j}}{\sum_{k=1}^{H\times W} e^{x_k}} \tag{3.3}$$

Meanwhile, feature I is reshaped to generate feature $D \in \mathbb{R}^{C\times HW}$. After that, a matrix multiplication is performed between B and D to obtain global context features $G \in \mathbb{R}^{1\times 1\times C}$. Next, G is put into the transform module to calculate the importance of each channel, which contains two steps: 1) 1×1 convolution to reduce the channels from C to C/r, following by layernorm and ReLU; 2) 1×1 convolution to increase the channels to C. The hyper-parameter r is set to 4 by default. Then, feature map W $\in \mathbb{R}^{1\times 1\times C}$ is obtained. In addition, broadcast element-wise sum is applied between W and I to obtain the refined feature $R \in \mathbb{R}^{H\times W\times C}$. This process can be formulated as:

$$W = Conv_{1\times 1}(LayerNorm(ReLU(Conv_{1\times 1}(G)))) \tag{3.4}$$

$$R = I + W \tag{3.5}$$

Next, R is rescaled to the original resolution corresponding to each level using the same but reverse operation like FAM. Specifically, adaptive max pooling is performed on R to obtain M_5' and nearest interpolation is used to get M_3' and M_2'. M_4' is directly copied from R. Finally, the enhanced feature maps $\{M_2', M_3', M_4', M_5'\}$ are merged with the original feature maps $\{F_2, F_3, F_4, F_5\}$ to generate the final feature maps $\{M_2, M_3, M_4, M_5\}$. For simplicity, the process of obtaining $\{M_2', M_3', M_4', M_5'\}$ from R is not illustrated in the graph. FRM is capable of exploiting global context information while modeling interdependencies between channels. It facilitates improving feature discriminability.

3.3 Coarse adjustment stage

As shown in Fig. 1, the detection head is conducted in a coarse-to-fine manner, which consists of two stages: a coarse adjustment stage and a refinement stage. We introduce the coarse adjustment stage in this section and the refinement stage in the next section. Commonly used RoI operations (e.g., RoI pooling [1] and RoI align [18]) divide proposals into equally sized sub-regions and then average four sampling points with each sub-region, which inherently have limitations in modeling various transformations of objects [19]. Therefore, regular RoI operations have poor generalization when dealing with objects of different variations in remote sensing images. In the coarse stage, deformable RoI pooling [20] is adopted to extract fixed size features from horizontal proposals from RPN. Given an input proposal of size W and H, it is divided into $K \times K$ bins (e.g., 7×7). The (i, j)-th bin $(0 \leqslant i, j \leqslant K)$ is denoted as bin_{ij}. Firstly, the corresponding feature of the proposal is obtained through the standard RoI pooling. Then, the offsets Δp_{ij} for each bin are learned through a fully connected

Table 1 Effectiveness of each component in our proposed method

	FPN	RoI pooling	A-FPN	Deformable RoI pooling	Smooth Ln	mAP(%)
baseline	✓	✓				76.14
		✓	✓			76.74
			✓	✓		77.01
our method			✓	✓	✓	**77.04**

Results are reported on DOTA-v1.0 test set. The best mAP is shown in bold

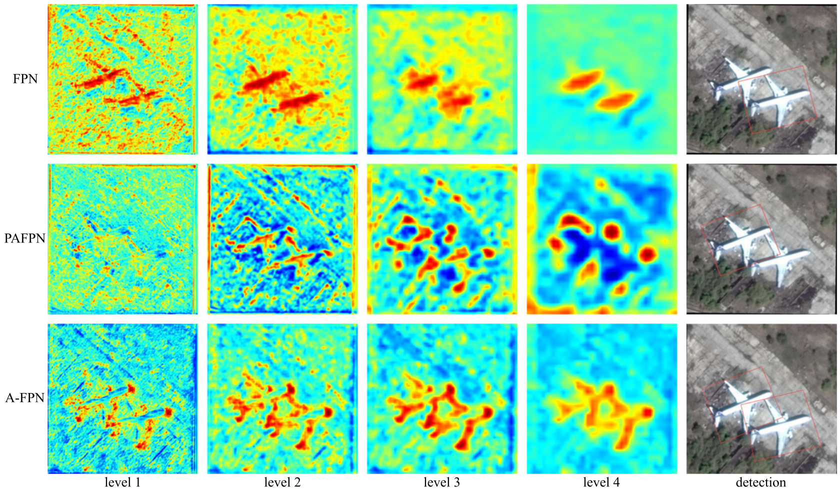


Fig. 6 Heatmaps of the final output feature maps at differernt levels in FPN [14], PANet [34] and A-FPN. The corresponding detection results are provided in last column

layer. The deformable RoI pooling operation for bin_{ij} is defined as follows:

$$y(i,j) = \sum_{p \in bin(i,j)} F\left(p_0 + p_n + \Delta p_{ij}\right)/n_{ij} \tag{3.6}$$

where p_n, p_0, n_{ij} represents the spatial positions, top-left corner and the number of pixels in the bin, respectively. And (3.6) is implemented by bilinear interpolation. Deformable RoI pooling [20] enables the network to adapt its feature representation to the configuration of different objects by deforming the pooling patterns. Then, feature maps of horizontal RoIs are input into two fully connected layers to learn the corresponding rotated bounding boxes. Specifically, we use (x, y, w, h) to represent a horizontal bounding box and (x, y, w, h, θ) to represent a rotated bounding box where (x, y), w, h, θ denote the bounding box's center coordinates, width, height and angle, respectively. Ranging in $[-3\pi/4, \pi/4)$, θ denotes the angle between the x-axis and the long side h of the bounding box. This paper adopts rotation and scaling transformation to convert the horizontal bounding box to the rotated bounding box. The offsets of the rotated bounding box can be calculated as follows:

$$\begin{bmatrix} t_x \\ t_y \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_t - x_a \\ y_t - y_a \end{bmatrix} \begin{bmatrix} \frac{1}{w_a} & 0 \\ 0 & \frac{1}{h_a} \end{bmatrix} \tag{3.7}$$

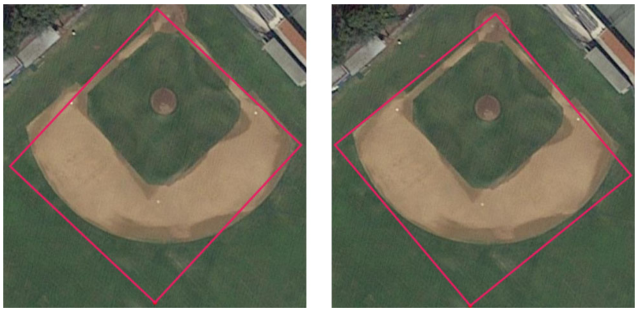
$$t_w = \log\left(\frac{w_t}{w_a}\right), t_h = \log\left(\frac{h_t}{h_a}\right) \tag{3.8}$$

$$t_\theta = (\theta_t - \theta_a) \bmod 2\pi \tag{3.9}$$

where x_t, x_a represents the ground-truth box and anchor box, respectively (likewise for y, w, h, θ). The operation mod adjusts the target of angle offset t_θ in $[0, 2\pi)$.

3.4 Refinement stage

The rotated bounding boxes generated in the coarse stage are not robust enough to the arbitrary orientations in aerial images. In the refinement stage, RRoI align [19] extracts rotation-invariant features for the final classification and localization. Given a feature map $F \in \mathbb{R}^{H \times W \times C}$ and a rotated RoI $(x_r, y_r, w_r, h_r, \theta_r)$, F is divided into $K \times$



(a) RoI align (b) deformable RoI pooling

Fig. 7 Detection results of using RoI align and deformable RoI pooling in the coarse stage

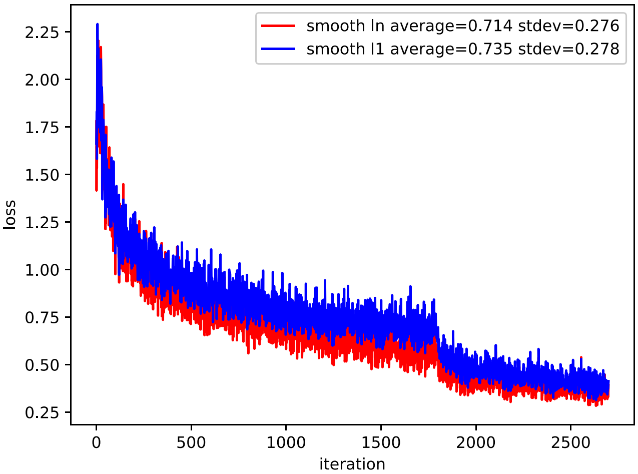


Fig. 8 Training loss curves after using smooth Ln loss and smooth L_1 loss

Table 2 Result comparisons without any data augmentation on DOTA-v1.0 test set

Method	Backbone	**mAP(%)**	PL	BD	BR	GTF	SV	LV	SH
CADNet [48]	R101	69.90	87.80	82.40	49.40	73.50	71.10	63.50	76.60
SCRDet [7]	R101	69.83	89.41	78.83	50.02	65.59	69.96	57.63	72.26
R^3Det [5]	R101	71.69	**89.54**	81.99	48.46	62.52	70.48	74.29	77.54
R^3Det [5]	R152	73.74	89.49	81.17	50.53	66.10	70.92	78.66	78.21
DRN [49]	H104	70.70	88.91	80.22	43.52	63.35	73.48	70.69	84.94
CenterMap [50]	R50	71.74	88.88	81.24	53.15	60.65	78.62	66.55	78.10
BBAVectors [29]	R101	72.32	88.35	79.96	50.69	62.18	78.43	78.98	87.94
SCRDet++ [28]	R152	74.41	89.20	83.36	50.92	68.17	71.61	80.23	78.53
F-O^2DETR [31]	R50	74.47	88.76	81.91	51.20	72.18	77.64	80.47	87.84
MEAD [51]	R101	74.80	88.42	79.00	49.29	68.76	77.41	77.68	86.60
ReDet [27]	ReR50-ReFPN	76.25	88.79	82.64	53.97	74.00	78.13	**84.06**	**88.04**
Oriented RepPoints [30]	R101	76.21	89.21	**84.22**	**58.42**	72.05	**79.81**	77.66	87.35
(Ours)	R50	77.04	88.79	82.02	54.10	77.13	79.08	82.90	87.67
(Ours)	R101	**77.13**	89.01	80.84	53.78	**79.05**	79.15	82.98	87.46
Method	Backbone	TC	BC	ST	SBF	RA	HA	SP	HC
CADNet [48]	R101	**90.90**	79.20	73.30	48.40	60.90	62.00	67.00	62.20
SCRDet [7]	R101	90.73	81.41	84.39	52.76	63.62	62.01	67.62	61.16
R^3Det [5]	R101	90.80	81.39	83.54	61.97	59.82	65.44	67.46	60.05
R^3Det [5]	R152	90.81	85.26	84.23	61.81	63.77	68.16	69.83	**67.17**
DRN [49]	H104	90.14	83.85	84.11	50.12	58.41	67.62	68.60	52.50
CenterMap [50]	R50	88.83	77.80	83.61	49.36	66.19	72.10	72.36	58.70
BBAVectors [29]	R101	90.85	83.58	84.35	54.13	60.24	65.22	64.28	55.70
SCRDet++ [28]	R152	90.83	86.09	84.04	**65.93**	60.80	68.83	71.31	66.24
F-O^2DETR [31]	R50	90.85	84.56	81.68	61.42	64.61	67.50	64.28	62.15
MEAD [51]	R101	90.78	85.55	84.54	62.10	66.57	72.59	72.84	59.83
ReDet [27]	ReR50-ReFPN	90.89	**87.78**	86.75	61.76	60.39	75.96	68.07	63.59
Oriented RepPoints [30]	R101	90.87	87.10	84.80	61.79	**67.76**	73.89	**73.38**	54.75
(Ours)	R50	90.81	87.54	86.15	65.29	66.81	76.84	70.40	60.06
(Ours)	R101	90.85	87.02	**86.79**	62.65	62.86	**77.15**	72.91	64.49

'R' and 'H' in the Backbone denotes the ResNet [15] and the Hourglass network [47], respectively. The best mAP and APs are highlighted in bold

K (e.g.,7 × 7) bins. The orientation of each bin is the same as the feature map. The width and height of each $bin_{jj}(0 \leqslant i, j < K)$ is $\frac{w_r}{K}$ and$\frac{h_r}{K}$ respectively. The number of sampling points in each bin is $N \times N$. As shown in Fig. 4, this paper defines a global horizontal coordinate system XOY bounding to the feature map and a local oriented coordinate system xOy bounding to the rotated RoI. The local top-left coordinate is denoted as $\left(\frac{iw_r}{K}, \frac{jh_r}{K}\right)$. The local coordinates of sampling points in bin_{ij} are denoted as $\left(\frac{jh_r}{K} + \frac{(j'+0.5)h_r}{KN}, \frac{iw_r}{K} + \frac{(i'+0.5)w_r}{KN}\right)$ $\left(i', j' = 0, 1, \ldots N-1\right)$. Each rotated bin in the feature map is converted into the axis-aligned region using affine transformation (scale, shift and rotate). And the local coordinate (x_o, y_o) in the bin is transformed to the corresponding global coordinate (x_h, y_h). This process can be formulated as the following equation:

$$\begin{pmatrix} x_h \\ y_h \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x_0 - w_r/2 \\ y_0 - h_r/2 \end{pmatrix} + \begin{pmatrix} x_r \\ y_r \end{pmatrix} \quad (3.10)$$

Then average pooling is performed in each bin_{ij} and the output feature is calculated by bilinear operation I as follows:

$$y(i, j) = \frac{1}{N \times N} \sum_{(x_0, y_0) \in bin_{(i,j)}} I\left(F, \tau\left(x_o, y_0\right)\right) \quad (3.11)$$

RRoI align warps the rotated RoIs with arbitrary orientations, sizes and aspect ratios into horizontal feature maps with a fixed size of 7 × 7 and produces rotation-invariant features in the spatial dimension. Similar to the

Table 3 Result comparisons with data augmentations on DOTA-v1.0 test set

Method	Backbone	**mAP(%)**	PL	BD	BR	GTF	SV	LV	SH
RoI Trans [19]	R101	69.56	88.64	78.52	43.44	75.92	68.81	73.68	83.59
SCRDet [7]	R101	72.61	89.98	80.65	52.09	68.36	68.36	60.32	72.41
R^3Det [5]	R101	73.79	88.76	83.09	50.91	67.27	76.23	80.39	86.72
R^3Det [5]	R152	76.47	89.80	83.77	48.11	66.77	78.76	83.27	87.84
DRN [49]	H104	73.23	89.71	82.34	47.22	64.10	76.22	74.43	85.84
CenterMap [50]	R101	76.03	89.83	84.41	54.60	70.25	77.66	78.32	87.19
BBAVectors [29]	R101	75.36	88.63	84.06	52.13	69.56	78.26	80.40	88.06
CSL [9]	R152	76.17	90.25	85.53	54.64	75.31	70.44	73.51	77.62
SCRDet++ [28]	R152	76.56	88.68	85.22	54.70	73.71	71.92	84.14	79.39
OWSR [52]	R101	76.36	**90.41**	85.21	55.00	78.27	76.19	72.19	82.14
CFA [53]	R152	76.67	89.08	83.20	54.37	66.87	81.23	80.96	87.17
F-O^2DETR [31]	R50	79.66	88.89	83.41	56.72	79.75	79.89	85.45	**89.77**
ReDet [27]	ReR50-ReFPN	**80.10**	88.81	82.48	**60.83**	**80.82**	78.34	**86.06**	88.31
Oriented RepPoints [30]	R101	78.12	88.72	80.56	55.69	75.07	**81.84**	82.40	87.97
(Ours)	R50	78.76	88.63	86.49	56.74	76.64	77.71	84.76	88.09
(Ours)	R101	79.78	87.49	**86.74**	60.51	76.51	77.10	85.03	88.72

Method	Backbone	TC	BC	ST	SBF	RA	HA	SP	HC
RoITrans [19]	R101	90.74	77.27	81.46	58.39	53.54	62.83	58.93	47.67
SCRDet [7]	R101	90.85	87.94	86.86	65.02	66.68	66.25	68.24	65.21
R^3Det+ [5]	R101	90.78	84.68	83.24	61.98	61.35	66.91	70.63	53.94
R^3Det [5]	R152	90.82	85.38	85.51	65.67	62.68	67.53	78.56	72.62
DRN [49]	H104	90.57	86.18	84.89	57.65	61.93	69.30	69.63	58.48
CenterMap [50]	R101	90.66	84.89	85.27	56.46	69.23	74.13	71.56	66.06
BBAVectors [29]	R101	90.87	87.23	86.39	56.11	65.62	67.10	72.08	63.96
CSL [9]	R152	90.84	86.15	86.69	69.60	68.04	73.83	71.10	68.93
SCRDet++ [28]	R152	90.82	87.04	86.02	67.90	60.86	74.52	70.76	72.66
OWSR [52]	R101	90.70	87.22	86.87	66.62	68.43	75.43	72.70	57.99
CFA [53]	R152	90.21	84.32	86.09	52.34	**69.94**	75.52	80.76	67.96
F-O^2DETR [31]	R50	90.84	86.15	**87.66**	69.84	68.97	78.83	78.19	70.38
ReDet [27]	ReR50-ReFPN	90.87	**88.77**	87.03	68.65	66.90	**79.26**	79.70	74.67
Oriented RepPoints [30]	R101	90.80	84.33	87.64	62.80	67.91	77.69	**82.94**	65.46
(Ours)	R50	90.85	88.27	85.12	64.90	65.79	78.43	76.41	72.50
(Ours)	R101	**90.88**	88.76	85.74	**69.85**	64.87	78.67	78.26	**77.51**

'R' and 'H' in the Backbone denotes the ResNet [15] and the Hourglass network [47]. The best mAP and APs are highlighted in bold

coarse adjustment stage, then two fully connected layers are added followed by two branches for final classification and regression.

3.5 Oriented object detection

3.5.1 Matching strategy

This paper adopts the IoU as the criteria when matching between the rotated bounding box and the ground truth. The rotated bounding box can be assigned to be positive if its IoU is over the threshold of 0.5. We calculate the IoU within polygons.

3.5.2 Loss function

Multi-task loss is adopted both in the coarse adjustment stage and refinement stage, which consists of classification loss and regression loss. For each RRoI (x, y, w, h, θ), the loss function is defined as:

$$L\left(p, u, t^*, t\right) = L_{cls}(p, u) + \lambda u L_{reg}\left(t^*, t\right) \tag{3.12}$$

Table 4 Result comparisons DOTA-v1.5 test set without any data augmentation

Method	mAP(%)	PL	BD	BR	GTF	SV	LV	SH	TC
FR-O+RT [54]	65.0	71.9	76.1	51.9	69.2	52.1	75.2	80.7	90.5
(Ours)	**65.5**	71.5	75.5	50.0	69.6	52.0	75.1	80.3	89.7
Method	-	BC	ST	SBF	RA	HA	SP	HC	CC
FR-O+RT [54]	-	78.6	68.3	49.2	71.7	67.5	65.5	62.2	10.0
(Ours)	-	77.5	68.7	51.5	70.0	72.8	64.1	59.7	19.7

The best mAP is shown in bold

where u indicates the class label ($u = 1$ for object and $u = 0$ for background), $t = \left(t_x, t_y, t_h, t_w, t_\theta\right)$ represents the predicted rotated bounding box and $\left(t_x^*, t_y^*, t_h^*, t_w^*, t_\theta^*\right)$ denotes the ground-truth. The hyperparameter λ controls the trade-off and is set to 1 by default. The classification loss L_{cls} is implemented by cross-entropy loss:

$$L_{cls}(p, u) = -\log p_u \tag{3.13}$$

where p_u denotes the probability over classes. And smooth Ln loss [21] is adopted for the regression:

$$L_{reg}\left(t^*, t\right) = \sum_{i \notin \{x,y,h,w,\theta\}} smooth_{Ln}\left(t_i^*, t_i\right) \tag{3.14}$$

in which

$$smooth_{Ln}(x) = (|x| + 1)\ln(|x| + 1) - |x| \tag{3.15}$$

And the deviation function of smooth Ln is calculated as follows:

$$\frac{\partial smooth_{Ln}(x)}{\partial x} = sign(x) \cdot \ln(sign(x) \cdot x + 1) \tag{3.16}$$

Note that both (3.15) and (3.16) are continuous functions. Figure 5 illustrates the differences between smooth L_1 and smooth Ln loss functions. The curve of smooth Ln function has better smoothness. Smooth Ln loss function has more resistance to outliers and can adjust regressive steps better.

Table 5 Result comparisons on DOTA-v1.5 test set with data augmentations

Method (Team Name)	**mAP(%)**	PL	BD	BR	GTF	SV	LV	SH	TC
peijin	71.6	80.9	83.6	55.1	70.7	59.9	76.4	88.3	90.9
CSULQQ	72.3	87.8	83.6	56.7	74.4	63.2	71.0	87.8	90.8
AICyber	74.7	88.4	85.4	56.7	74.4	63.9	72.7	87.9	90.9
OWSR [52]	74.9	-	-	-	-	-	-	-	-
FR-O+RT [54]	**77.6**	87.5	84.3	62.2	79.8	67.3	83.2	89.9	90.9
FR-O+RT* [54]	73.6	80.6	84.4	57.3	76.8	52.7	81.4	89.2	90.9
(Ours)	74.7	80.9	86.5	61.1	74.3	58.0	83.1	89.6	90.9
Method (Team Name)	-	BC	ST	SBF	RA	HA	SP	HC	CC
peijin	-	79.2	78.3	59.1	74.8	74.1	74.9	59.8	39.5
CSULQQ	-	84.6	84.0	67.8	75.5	67.4	71.2	68.8	22.5
AICyber	-	86.3	85.0	68.9	76.0	74.1	72.9	73.4	37.9
OWSR [52]	-	-	-	-	-	-	-	-	-
FR-O+RT [54]	-	83.9	77.7	73.9	75.3	78.6	77.1	75.2	54.8
FR-O+RT* [54]	-	83.3	73.4	69.1	73.0	77.8	75.6	74.5	37.6
(Ours)	-	86.4	73.7	66.6	72.7	78.2	75.4	76.1	41.4

The best mAP is shown in bold. Note that we report the results of single model of OWSR [52] for fair comparisons. * means the results of using our experimental settings and the same data augmentation strategies as ours

4 Experiments and analysis

4.1 Datasets

To comprehensively evaluate the effectiveness of our proposed oriented object detector, extensive experiments are conducted on DOTA [23] and HRSC2016 [24].

DOTA has two released version: DOTA-v1.0 and DOTA-v1.5. DOTA-v1.0 is a large-scale dataset for object detection in aerial images, which is comprised of 15 categories, 12,806 images and 188,282 instances labeled by an arbitrary quadrilateral. The short names for categories are defined as (abbreviation-full name): PL-Plane, BD-Baseball diamond, BR-Bridge, GTF-Ground field track, SV-Small vehicle, LV-Large vehicle, SH-Ship, TC-Tennis court, BC-Basketball court, ST-Storage tank, SBF-Soccer-ball field, RA-Roundabout, HA-Harbor, SP-Swimming pool, and HC-Helicopter. Objects in the dataset exhibit various shapes, orientations and scales. The training set, validation set and test set account for 1/2, 1/6 and 1/3 of the whole dataset, respectively. DOTA-v1.5 is an upgraded version of DOTA-v1.0 which is released for DOAI Challenge 20193. DOTA-v1.5 contains 16 categories and 402,089 instances. DOTA-v1.5 adds a new category of Container Crane (CC) and additionally annotates rather small objects about or below 10 pixels, which increase the difficulty of detection. The training set and validation set are used for training and the test set is adopted for testing. Due to the size of images in DOTA-v1.0 and DOTA-v1.5 ranges from around 800×800 to $4,000 \times 4,000$ pixels, we crop the images into $1,024 \times 1,024$ patches with a stride of 500. For multi-scale data augmentation, multi-scale data are prepared at three scales {0.5, 1.0, 1.5} and rotation augmentation is performed randomly from 4 angles {0, 90, 180, 270}.

HRSC2016 is a challenging dataset for ship detection, which contains 1061 images with the size ranging from 300×300 to $1,500 \times 900$ and over 20 categories of ships in diverse appearances. The training, validation and test set include 436 images, 181 images and 444 images, respectively. This paper uses training set and validation set for training, test set for testing. All the images are resized to $(800, 512)$. Random horizontal flipping is adopted during training.

4.2 Implementation details

Experiments are implemented on Pytorch and MMDetection [46]. All the models are trained for 12 epochs on DOTA-v1.0 and DOTA-v1.5, 36 epochs on HRSC2016. The general training schedule in MMdetection is adopted. The stochastic gradient descent (SGD) optimizer is used with an initial learning rate of 0.01, and the learning rate decreases by 0.1 after 8 and 11 epochs. The momentum and weight

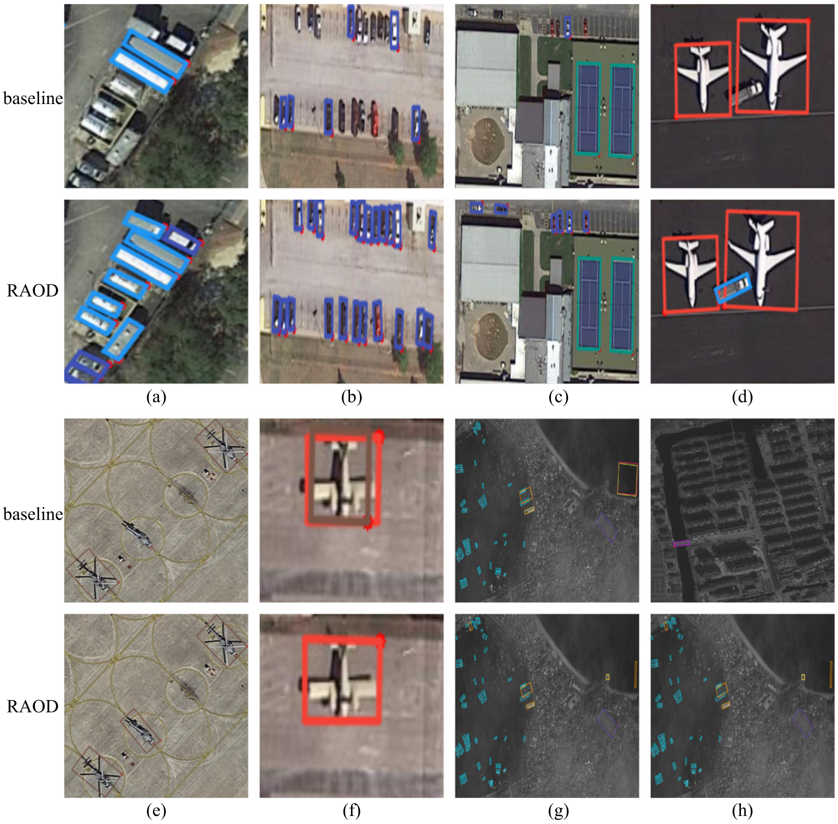


Fig. 9 Qualitative comparisons between our method and the baseline on the test set of DOTA-v1.0

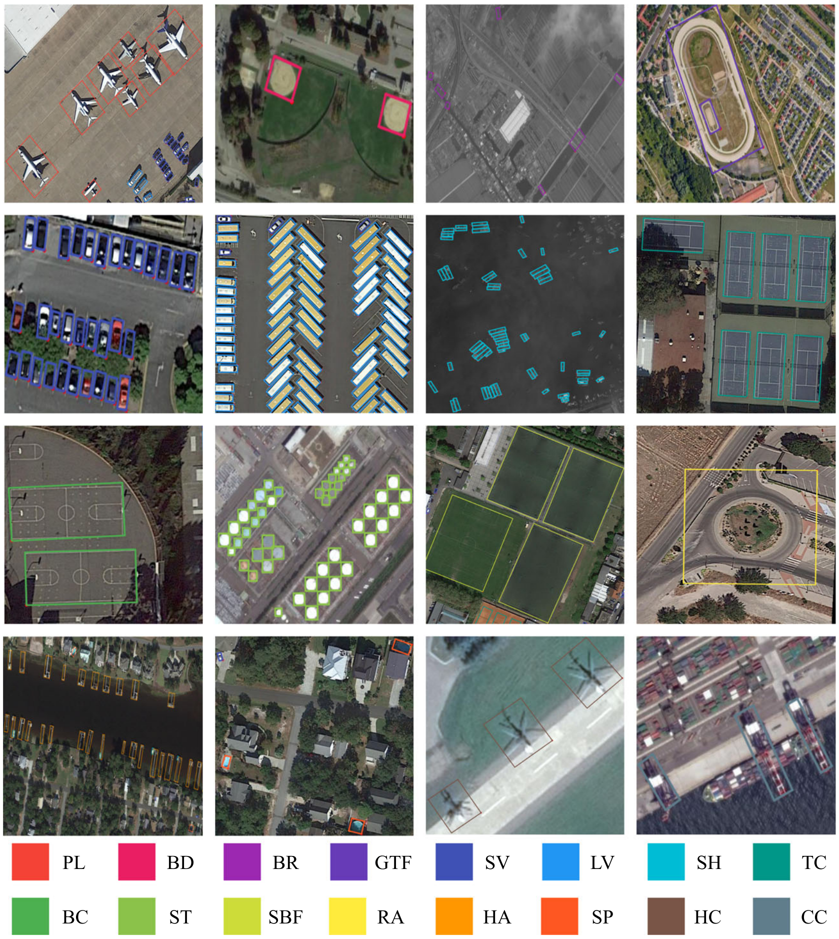


Fig. 10 Visualization results on the test set of DOTA-v1.5

decay are 0.9 and 0.0001, respectively. We use 1 RTX Titan GPU with a batch size of 2 for training. It is worth noting that only three horizontal anchors with aspect ratios of $\{1/2, 1, 2\}$ are set, avoiding a large number of anchors caused by adding preseted angles. For training our model, 512 proposals are sampled randomly with a $1:3$ positive to negative ratio. For testing, there are 10,000 proposals (2,000 at each feature level) before NMS and 2,000 proposals after NMS. For evaluation, the mean Average Precision (mAP) is adopted as the primary metric. The results of DOTA-v1.0 and DOTA-v1.5 reported are obtained by submitting predictions to the official DOTA evaluation server [1].

4.3 Ablation studies

To evaluate the effectiveness of our proposed method, a series of ablation experiments are performed on DOTA-v1.0 test set. This paper uses RoI-Transformer [19] as the baseline. RoI-Transformer is a two-stage method for oriented detection. For a fair comparison, we reproduce RoI-Transformer and obtain 76.14% mAP, which is higher than the official one. ResNet-50 [15] is used as the backbone and all the experimental settings are consistent as those reported in Section 4.2. No data augmentation is used in this section. Comparing with the baseline method, RAOD is able to improve the expressiveness of multi-scale features and model geometric transformations more effectively, which are important for detecting arbitrary-oriented objects in remote sensing images. Effectiveness of each component is detailed in Sections 4.3.1, 4.3.2 and 4.3.3, respectively.

4.3.1 Effectiveness of A-FPN

As shown in Table 1, the mAP is 76.14% with FPN and 76.74% with our proposed A-FPN. Figure 6 shows the heatmaps of the final output feature maps at different levels in FPN [14] , PANet [34] and A-FPN, respectively. It can be observed that A-FPN reduces much redundant information brought by direct fusion in the top-down pathway and strengthens the features at each level, increasing the saliency of the object areas. Compared with FPN and PANet, the boundaries are clearer in the heatmaps of A-FPN. This also validates that extracting discriminative features can help to improve the final detection results effectively.

[1] https://captain-whu.github.io/DOTA/

4.3.2 Effectiveness of deformable RoI pooling

Table 1 shows the results of using RoI align [18] and deformable RoI pooling [20] in the coarse stage, respectively. It can be seen that deformable RoI pooling gains about 0.27% improvement in mAP and enables the network to capture the boundary of the object more precisely compared to RoI align (as shown in Fig. 7). And the results indicate that geometry transformation modeling contributes to the localization performance.

4.3.3 Effectiveness of smooth *Ln*

As shown in Table 1, with the participation of smooth Ln loss, mAP achieves 77.04%. Figure 8 shows the training loss after adopting smooth Ln loss and smooth L_1 loss. Both the mean and variance of the smooth Ln loss are lower than the smooth L_1 loss, which demonstrates that smooth Ln loss achieves a more stable training and enables the model to better converge.

4.4 Comparisons with state-of-the-art methods

In this section, we further compare our proposed method with the state-of-the-art algorithms on three benchmark datasets DOTA-v1.0, DOTA-v1.5 and HRSC2016. First, experiments without any data augmentations are conducted. In addition, experiments on data augmentations of multi-scale training and rotation training are also carried out for further comparisons. Tables 2 and 3 shows comparisons without any augmentation and with augmentation on test set of DOTA-v1.0, respectively. Similarly, the corresponding comparisons on test set of DOTA-v1.5 are shown in Tables 4 and 5.

4.4.1 Visualizations

Figure 9 compares the visual results between the baseline and our proposed method on the test set of DOTA-v1.0. It is observed that RAOD achieves better performance in detecting densely packed small objects and precisely locate the instances with sharp variety on orientation and aspect ratio.

4.4.2 Results on DOTA-v1.0

Table 2 compares our proposed single-scale RAOD with other state-of-the-art algorithms. As for the single-scale training, RAOD obtains 77.04% mAP with ResNet-50 backbone, which outperforms the previous best single-scale model by 0.37% and most multi-scale methods. With a stronger ResNet-101 backbone, RAOD achieves the best result of 77.13%, which surpasses the previous best single-scale model by 0.46%. In addition, RAOD obtains the best results on ground field track (GTF), storage tank (ST) and harbor(HA). The results on plane (PL), small vehicle (SV), ship (SH), tennis court (TC) and basketball court (BC) are silghtly lower than the best results and outperform most of other methods. In the case of multi-scale training and testing (as shown in Table 3), RAOD obtains advanced performance of 78.76% mAP with ResNet-50 backbone and second-best result of 79.78% mAP with ResNet-101 backbone. For objects with a large aspect ratio, such as bridge (BR) and harbor (HA), RAOD achieves the second-best and the third-best performance with 60.51% and 78.67%, respectively. For the helicopter (HC) with various appearances and irregular shapes, our method achieves the best result of 77.51%, which surpasses the second place over 2.84%. Those comparisons verify the superiority of RAOD.

Table 6 Result comparisons with the state-of-the-art methods on the test set of HRSC2016

Method	Backbone	mAP(%)
RRPN [56]	R101	79.08
R^2PN [57]	V16	79.60
RoI Trans [19]	R101	86.20
Gliding Vertex [25]	R101	88.20
R^3Det [5]	R101	89.26
RetinaNet-DAL [58]	R101	89.77
OPLD [59]	R50	88.44
DRN [49]	H104	92.70*
BBAVectors [29]	R101	88.60
MEAD [51]	R101	89.83
Ours	R50	89.71/**94.82***
Ours	R101	**89.92**/94.28*

‘V’ in the Backbone denotes the VGG [55]. * indicates the results evaluated under VOC2012 metrics while other methods are evaluated under the VOC2007 metrics. The best mAP is shown in **bold**

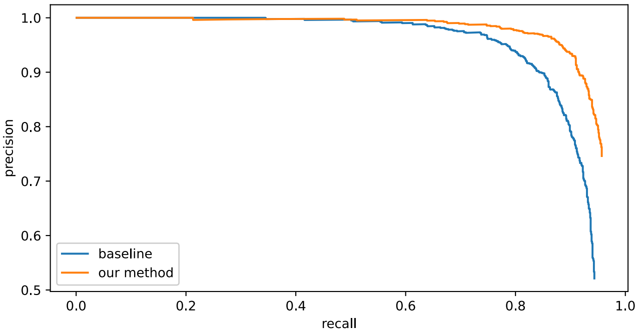


Fig. 11 Precision-recall curves of our method and baseline on the testing set of HRSC2016

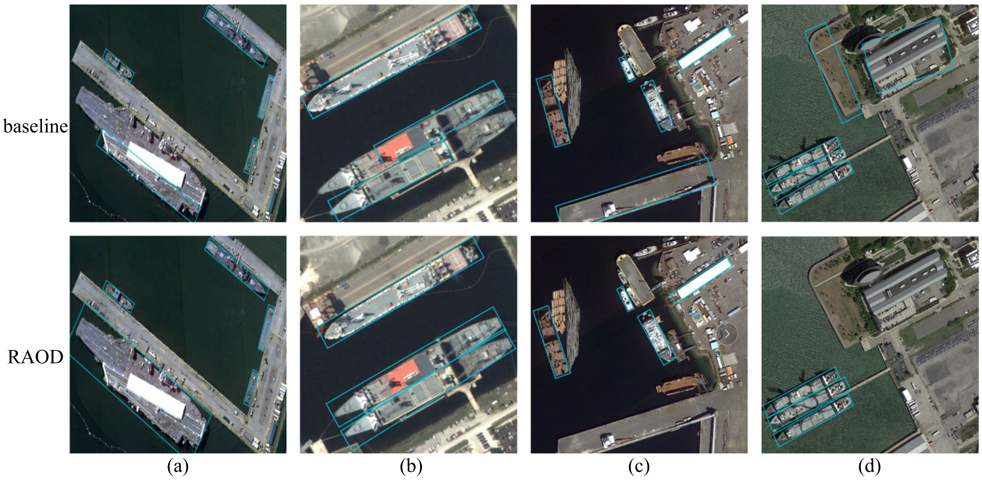


Fig. 12 Qualitative comparisons between the baseline and the proposed RAOD on HRSC2016 test set

4.4.3 Results on DOTA-v1.5

DOTA-v1.5 is employed for the evaluation of detection performance in Detecting Objects in Aerial Images Challenge 2019 (DOAI2019), which is a much more challenging task in oriented object detection of aerial images. For fair comparisons, all the results are quoted in one decimal place. Compared with the baseline method FR-O+RT reported in [54], our model with the single-scale setting has improved by 0.5% in mAP (as shown in Table 4), which demonstrates the effectiveness of our method. Besides, Table 5 shows the comparisons in the setting of data augmentations. To our best knowledge, most of the competitors in DOAI2019 adopt multiple augmentation strategies for better performance, such as multi-scale training and testing, rotation training and testing, class balance resampling and model ensembling. Most of them do not report detection results without data augmentations. However, this paper only adopts multi-scale training and rotation training for data augmentations. So the comparisons between our RAOD and other competitors are unfair. For fair comparisons, we reimplement FR-O+RT [54] under our experimental setting and use the same data augmentation strategies as ours. As shown in Table 5, FR-O+RT [54] obtains 73.6% on mAP, which is lower than our proposed RAOD by 1.1%. This paper pays more attention to the effectiveness of the model itself and these comparisons validate that our method can achieve better or similar accuracy without many data augmentations. When checking the AP on each category, RAOD ranks first in baseball diamond (BD), tennis court (TC), basketball court (BC) and helicopter (HC). It demonstrates that RAOD outperforms other methods on categories with irregular layouts and complex backgrounds. However, our model fails to achieve high performance on container (CC) and small vehicle (SV) whose mAP is 41.4% and 58.0%, respectively. We consider it is due to dense distributions and small sizes in

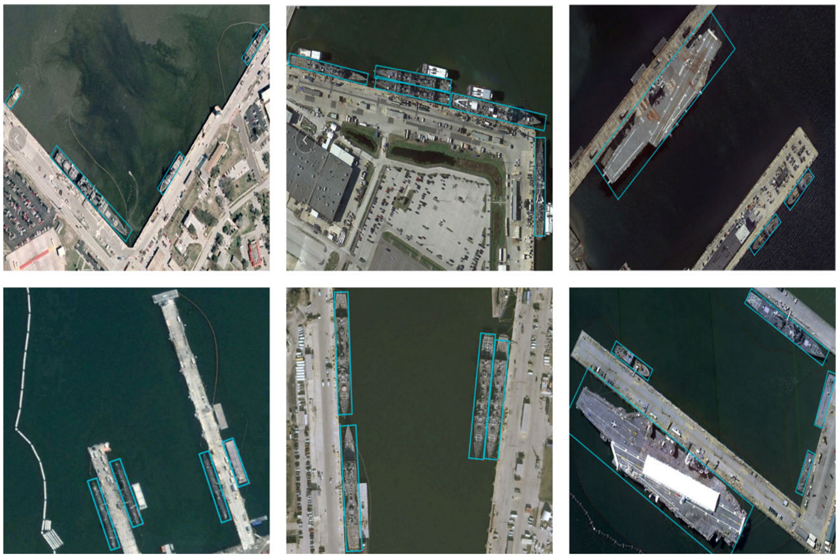


Fig. 13 Ship detection results on the HRSC2016 test set

those categories. Therefore, there is still room for progress in our method, which is one part of our follow-up research. Figure 10 visualizes the detection results on DOTA-v1.5.

4.4.4 Results on HRSC2016

Table 6 compares our proposed method with other state-of-the-art methods on the test set of HRSC2016. To make a comprehensive comparison, our method is evaluated under the VOC2007 metric and the VOC2012 metric. Our method achieves leading performance in mAP by 89.71% under VOC2007 metric and 94.82% under VOC2012 metric with ResNet-50 backbone. With a stronger backbone ResNet-101, our method obtains 89.92% and 94.28% under VOC2007 metric and VOC2012 metric, respectively. Both the results outperform other methods. To further verify the effectiveness of our method, we plot the precision-recall curve in Fig. 11. Figure 12 shows visual comparisons between the baseline and our proposed method. Some detection results are shown in Fig. 13. These results demonstrate that our proposed method localizes objects more accurately, especially for strip-like instances with arbitrary orientations despite the low luminosity and resolution. And RAOD detects fewer false positive boxes.

5 Conclusion

This paper presents a refined two-stage detector with augmented features named RAOD for oriented object detection. The key idea of our proposed A-FPN is to better generate hierarchical discriminative features and enhance the representation ability of the model. Then we adopt a coarse-to-fine manner in the detection head. In the coarse stage, geometry-robust features are extracted from the horizontal bounding boxes and then are transformed into oriented ones. In the refinement stage, rotation-invariant features are obtained for detecting rotated objects more accurately in remote sensing images. Extensive experiments on DOTA-v1.0, DOTA-v1.5 and HRSC2016 verify the superiority of our method in the task of oriented object detection in remote sensing images. Our proposed method not only shows advanced performances on the aforementioned popular datasets in aerial images but also outperforms on some categories with arbitrary orientations, large aspect ratios and complex backgrounds. In future work, we aim to reduce the parameters of the network and achieve a better trade-off between the detection speed and accuracy. In addition, we will explore some strategies to boost the performance on those categories with relatively low precisions (e.g., container and small vehicle).

Acknowledgements The authors greatly appreciate the financial supports of the Shanghai Association for Science and Technology under Grant 17DZ1100808.

Declarations

Conflict of Interests The author(s) declared no conflicts of interest with respect to the research, authorship, and publication of this paper.

References

1. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. Adv Neural Inf Process Syst 28:91–99
2. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 779–788
3. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC (2016) Ssd: Single shot multibox detector. In: European conference on computer vision. Springer, pp 21–37
4. Lin T-Y, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp 2980–2988
5. Yang X, Liu Q, Yan J, Li A, Zhang Z, Yu G (2019) R3det: Refined single-stage detector with feature refinement for rotating object. arXiv:190805612 2(4)
6. Han J, Ding J, Li J, Xia G-S (2021) Align deep features for oriented object detection. IEEE Trans Geosci Remote Sens
7. Yang X, Yang J, Yan J, Zhang Y, Zhang T, Guo Z, Sun X, Fu K (2019) Scrdet: Towards more robust detection for small, cluttered and rotated objects. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 8232–8241
8. Qian W, Yang X, Peng S, Guo Y, Yan J (2019) Learning modulated loss for rotated object detection. arXiv:1911.08299
9. Yang X, Yan J (2020) Arbitrary-oriented object detection with circular smooth label. In: European Conference on Computer Vision. Springer, pp 677–694
10. Yang X, Hou L, Zhou Y, Wang W, Yan J (2021) Dense label encoding for boundary discontinuity free rotation detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 15819–15829
11. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
12. Shu C, Ding X, Fang C (2011) Histogram of the oriented gradient for face recognition. Tsinghua Sci Technol 16(2):216–224
13. Wang Z (2022) Automatic and robust hand gesture recognition by sdd features based model matching. Appl Intell:1–12
14. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2117–2125
15. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
16. Zhu D, Xia S, Zhao J, Zhou Y, Niu Q, Yao R, Chen Y (2021) Spatial hierarchy perception and hard samples metric learning for high-resolution remote sensing image object detection. Appl Intell:1–16
17. Zhang K, Zeng Q, Yu X (2021) Rosd: Refined oriented staged detector for object detection in aerial image. IEEE Access 9:66560–66569

18. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 2961–2969
19. Ding J, Xue N, Long Y, Xia G-S, Lu Q (2018) Learning roi transformer for detecting oriented objects in aerial images. arXiv:1812.00155
20. Dai J, Qi H, Xiong Y, Li Y, Zhang G, Hu H, Wei Y (2017) Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 764–773
21. Liu Y, Jin L (2017) Deep matching prior network: Toward tighter multi-oriented text detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1962–1969
22. Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 1440–1448
23. 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, pp 3974–3983
24. Liu Z, Yuan L, Weng L, Yang Y (2017) A high resolution optical satellite image dataset for ship recognition and some new baselines. In: International conference on pattern recognition applications and methods, vol 2. SCITEPRESS, pp 324–331
25. Xu Y, Fu M, Wang Q, Wang Y, Chen K, Xia G-S, Bai X (2020) Gliding vertex on the horizontal bounding box for multi-oriented object detection. IEEE Trans Pattern Anal Mach Intell 43(4):1452–1459
26. Qin R, Liu Q, Gao G, Huang D, Wang Y (2020) Mrdet: A multi-head network for accurate oriented object detection in aerial images. arXiv:2012.13135
27. Han J, Ding J, Xue N, Xia G-S (2021) Redet: A rotation-equivariant detector for aerial object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 2786–2795
28. Yang X, Yan J, Yang X, Tang J, Liao W, He T (2020) Scrdet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing. arXiv:2004.13316
29. Yi J, Wu P, Liu B, Huang Q, Qu H, Metaxas D (2021) Oriented object detection in aerial images with box boundary-aware vectors. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 2150–2159
30. Li W, Zhu J (2021) Oriented reppoints for aerial object detection. arXiv:2105.11111
31. Ma T, Mao M, Zheng H, Gao P, Wang X, Han S, Ding E, Zhang B, Doermann D (2021) Oriented object detection with transformer. arXiv:2106.03146
32. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
33. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: European Conference on Computer Vision. Springer, pp 213–229
34. Liu S, Qi L, Qin H, Shi J, Jia J (2018) Path aggregation network for instance segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8759–8768
35. Guo C, Fan B, Zhang Q, Xiang S, Pan C (2020) Augfpn: Improving multi-scale feature learning for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 12595–12604
36. Ghiasi G, Lin T-Y, Le QV (2019) Nas-fpn: Learning scalable feature pyramid architecture for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7036–7045
37. Tan M, Pang R, Le QV (2020) Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10781–10790
38. Cao J, Chen Q, Guo J, Shi R (2020) Attention-guided context feature pyramid network for object detection. arXiv:2005.11475
39. Luo Y, Cao X, Zhang J, Guo J, Shen H, Wang T, Feng Q (2021) Ce-fpn: Enhancing channel information for object detection. arXiv:2103.10643
40. Ma J, Chen B (2020) Dual refinement feature pyramid networks for object detection. arXiv:2012.01733
41. Zhang D, Zhang H, Tang J, Wang M, Hua X, Sun Q (2020) Feature pyramid transformer. In: European Conference on Computer Vision. Springer, pp 323–339
42. Jaderberg M, Simonyan K, Zisserman A et al (2015) Spatial transformer networks. Adv Neural Inf Process Syst 28:2017–2025
43. Zhou Y, Ye Q, Qiu Q, Jiao J (2017) Oriented response networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 519–528
44. Shi W, Caballero J, Huszár F, Totz J, Aitken AP, Bishop R, Rueckert D, Wang Z (2016) Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1874–1883
45. Cao Y, Xu J, Lin S, Wei F, Hu H (2019) Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp 0–0
46. Chen K, Wang J, Pang J, Cao Y, Xiong Y, Li X, Sun S, Feng W, Liu Z, Xu J et al (2019) Mmdetection: Open mmlab detection toolbox and benchmark. arXiv:1906.07155
47. Newell A, Yang K, Deng J (2016) Stacked hourglass networks for human pose estimation. In: European conference on computer vision. Springer, pp 483–499
48. Zhang G, Lu S, Zhang W (2019) Cad-net: A context-aware detection network for objects in remote sensing imagery. IEEE Trans Geosci Remote Sens 57(12):10015–10024
49. Pan X, Ren Y, Sheng K, Dong W, Yuan H, Guo X, Ma C, 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, pp 11207–11216
50. Wang J, Yang W, Li H-C, Zhang H, Xia G-S (2020) Learning center probability map for detecting objects in aerial images. IEEE Trans Geosci Remote Sens 59(5):4307–4323
51. He Z, Ren Z, Yang X, Yang Y, Zhang W (2021) Mead: a mask-guided anchor-free detector for oriented aerial object detection. Appl Intell:1–16
52. Li C, Xu C, Cui Z, Wang D, Jie Z, Zhang T, Yang J (2019) Learning object-wise semantic representation for detection in remote sensing imagery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp 20–27
53. Guo Z, Liu C, Zhang X, Jiao J, Ji X, Ye Q (2021) Beyond bounding-box: Convex-hull feature adaptation for oriented and densely packed object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 8792–8801
54. Ding J, Xue N, Xia G-S, Bai X, Yang W, Yang MY, Belongie S, Luo J, Datcu M, Pelillo M et al (2021) Object detection in aerial images: A large-scale benchmark and challenges. arXiv:2102.12219
55. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
56. 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 Multimed 20(11):3111–3122

57. Zhang Z, Guo W, Zhu S, Yu W (2018) Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks. IEEE Geosci Remote Sens Lett 15(11):1745–1749
58. Ming Q, Zhou Z, Miao L, Zhang H, Li L (2020) Dynamic anchor learning for arbitrary-oriented object detection. arXiv:2012.04150 1(2):6
59. Song Q, Yang F, Yang L, Liu C, Hu M, Xia L (2020) Learning point-guided localization for detection in remote sensing images. IEEE J Sel Top Appl Earth Observ Remote Sens 14:1084–1094

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

Qin Shi was born in Yangzhou, Jiang su, China. She has received the B.S. degree in East China University of Science and Technology. She is currently a postgraduate at the school of information science and engineering, East China University of Science and Technology in 2021. Her research interests include deep learning, object detection, image processing and pattern recognition.

Yu Zhu received the Ph.D. degree in optical engineering from Nanjing University of Science and Technology, China, in 1999. She is currently a professor in the department of electronics and communication engineering of East China University of Science and Technology. Her research interests include image processing, computer vision, multimedia communication and deep learning. She has published more than 80 papers in journals and conferences.

Chuantao Fang received his B.S. degree from East China University of Science and Technology in 2018. He is currently a postgraduate at the school of information science and engineering, East China University of Science and Technology. His research interests include deep learning, image super-resolution an pattern recognition.

Nan Wang received his B.S. degree from Nanjing University, Nanjing, China, in 2009, M.S. and Ph.D. from the Graduate School of IPS, Waseda University, Japan, in 2011, and 2014, respectively. He is currently an associate professor of electronics and communication engineering in East China University of Science and Technology, Shanghai, China. His current research interests include VLSI design automation, low power design techniques, network-on-chip and reconfigurable architectures. Dr. Wang is a member of IEEE.

Jiajun Lin obtained his PHD degree from TSINGHUA University, Beijing. He is a professor at School of Information Science and Engineering, East China University of Science and Technology. His research interests include Intelligent Information Processing and Security of Industry Control Systems.