TARDet: Two-stage Anchor-free Rotating Object Detector in Aerial Images

Longgang Dai †, Hongming Chen †, Yufeng Li *, Caihua Kong, Zhentao Fan, Jiyang Lu, Xiang Chen
College of Electronic and Information Engineering, Shenyang Aerospace University

Abstract

Detection of rotating object in aerial images is a practical and challenging task. Nowadays, most detectors rely on anchor boxes with different scales, aspect ratios and angles for aerial objects that are usually distributed in arbitrary directions and show huge variations in scale and aspect ratios. However, the detection performance of these detectors is very sensitive to the anchoring hyperparameters. To address this issue, in this paper, we propose a Two-stage Anchor-free Rotating object Detector (TARDet). Our TARDet first aggregates feature pyramid context information by a feature refinement module, and generates rough localization boxes in an anchor-free manner by a directed generation module (DGM) in the first stage, and then refines it to a higher quality localization scheme. Furthermore, we design an alignment convolution module to extract alignment features and introduce RiRoI to adaptively extract rotationally invariant features from isovariant features. Finally, we apply a modified fast R-CNN head to generate the final detection results. Our approach achieves state-of-the-art performance on two popular aerial objects datasets, DOTA and HRSC2016.

1. Introduction

Aerial image detection aims to identify the position and class of object objects such as ships and vehicles. However, it is also a challenging task because aerial image objects have different scales and aspect ratios [37]. In addition, objects are usually displayed in arbitrary orientations and are densely arranged. The ability of directional detection methods for profiling in remote sensing image targets [13] has attracted a lot of attention from researchers.

In recent years, aerial image rotation detection has made rapid progress due to the rapid development of convolutional neural networks [3, 6, 19, 26, 35, 38]. However, most of the rotation detection methods are generally based on the horizontal bounding box (HBB) by adding angular dimensions to this fixed pattern. For example, S^2aNet [10] proposes an anchoring refinement network to generate high-quality anchoring and alignment features. In the ReDet [11] network, rotational isovariant features are extracted by adding a rotational isovariant network to the detector, which in turn accurately predicts the orientation. RoI Trans learns the spatial transformation from HBB to orientation bounding box (OBB). R^3Det [36] accurately detects objects by using a stepwise regression approach from coarse to fine granularity. The Transformer scheme provides promising orientation schemes that greatly reduce the number of rotation anchors, but also entail expensive computational costs. Although these methods have achieved better performance, these detectors are very sensitive to anchor frame hyperparameters and also limit the scenarios in which the detectors can be used.

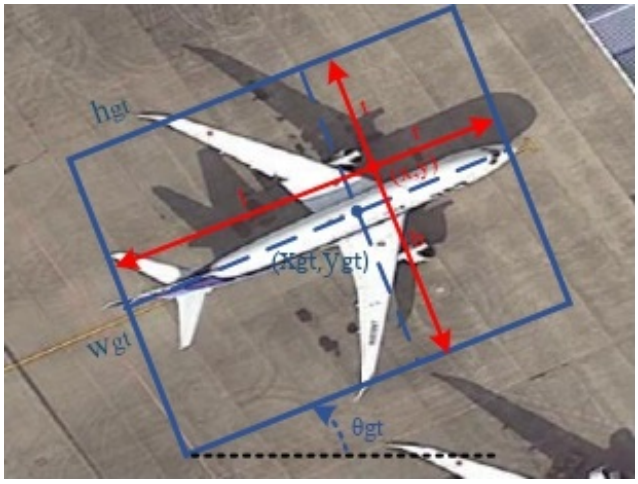

Figure 1. Oriented box definition. The blue box is ground truth.

Anchor-free detectors can avoid the hyperparameters associated with anchor boxes by eliminating the predefined anchor boxes. Currently, several anchor-free rotational detectors have been proposed. For example, FCOS [33] solves object detection in a per-pixel prediction manner, completely avoiding the complex computation associated with anchor boxes, and DARDet directly predicts the parameters of the rotation box at each foreground pixel of the feature map. DAFNet [16] introduces a new center-oriented ambiguity and a new divide-and-conquer corner point pre-

*Corresponding author. † Authors contributed equally to this work.

diction strategy. The work of [24] builds on a detector with per-pixel prediction, using a new geometric transformation to better represent oriented objects in angle prediction, and then develops a branch interaction module with a self-attentive mechanism to fuse features from classification and box regression branches. VCSOP [31] uses one subnetwork to search for centroids and the remaining three subnetworks to predict other parameters. However, these methods usually present rotated boxes in a complex form to solve the boundary discontinuity problem. The features used in these methods are not aligned with the rotated boxes. These drawbacks hinder the accuracy of the detector.

In this paper, we provide a new two-stage anchor-free rotating network. In the first stage, inspired by deformable convolution [5], we first aggregate the contextual information of the neighboring layers of the feature pyramid network (FPN) [21] while learning the offsets of the sampled points and use the aggregated information to refine the content of each sampled point. Instead of heuristically defining various anchors with different scales, angles, and aspect ratios, the network uses an anchor-free scheme to generate roughly oriented frames to predict targets. Then, the Align-Conv module is designed to adaptively align features according to the oriented frames. After that, the coarse frame is refined to the exact position by refinement network to generate a high-quality scheme. In the second stage, we use a modified Fast R-CNN [8] head for regression and classification to generate the final predictions.

To summarize, the main contributions of this paper are summarized as follows:

- We propose a two-stage anchor-free detector for aerial image orientation detection, which generates coarse positioning boxes directly in an anchor-free manner, avoiding the problems caused by horizontal frames, and then refines to a high-quality solution.

- We design the spatial refinement module, which is used to aggregate contextual information and refine features by embedding them in the FPN, and the alignment convolution module to extract alignment features for accurate object detection in aerial images. We also introduce RiRoI to adaptively extract rotationally invariant features from the equivalent features and modify Fast R-CNN to obtain the final object detection results.

- Extensive experiments are conducted on the DOTA and HRSC2016 datasets and its generalization capability is verified on the UG2+ Challenge dataset. The results show that our TARDet helps to improve the detection accuracy, and achieve the best performance on the DOTA and HRSC2016 datasets.

2. Related Work

In recent years, with the continuous development of deep learning, the object detection performance has been significantly improved. Existing object detection methods can be mainly classified into two modes, one-stage and two-stage, according to their structures.

2.1. One-Stage Detector

One-stage detectors [23, 25, 34] directly predict object classes and locations without any refinement steps. One-stage detectors can also be broadly classified into anchor-based and anchor-free methods. In the anchor-based approach, a large number of pre-defined anchor points are first flattened on the image, then the class of these anchor points is predicted and the refinement of the anchor coordinates is performed, and finally the refined anchor frame is output as the detection result. In the recently proposed [36] , the center and corner point information in the features are directly encoded to obtain more accurate positions. Sliding vertices predict more explicit four-point polygons in the image. However, in most anchor-frame-based methods, the HBB is still predicted by simply adding the angle of the anchor and then converting it to OBB. for example, in [20], a single-stage detector for oriented scene text detection is proposed, where the HBB is directly used as an anchor to regress the OBB and obtain state-of-the-art results on text detection.

However, this still requires a significant computational effort. In contrast, the anchor-free detector is predicted on a per-pixel basis, which would free the model from the highly intensive computation of learning anchor matching. Aerial object detection faces different challenges compared to text scene detection as mentioned in Section 1. Many detectors downsample the image to match the feature map size [7, 17, 40] and construct the final predicted object by resizing the output object, however, this may increase the error in object detection, especially when detecting objects from aerial images with a dense distribution of small targets. The FCOS method proceeds on a per-pixel basis, with the output feature map on of key points corresponding to the pixel coordinates in the input image, thus avoiding detection errors due to image resizing. However, the one-stage detector often suffers from too many negative samples and has low accuracy. In contrast, the two-stage method is more accurate in terms of precision.

2.2. Two-Stage Detector

The two-stage detector in object detection is implemented by examining the image twice, where the first examination is done by detecting regions of interest to generate region proposals, and then extracting features using the backbone feature map of each region proposal and passing

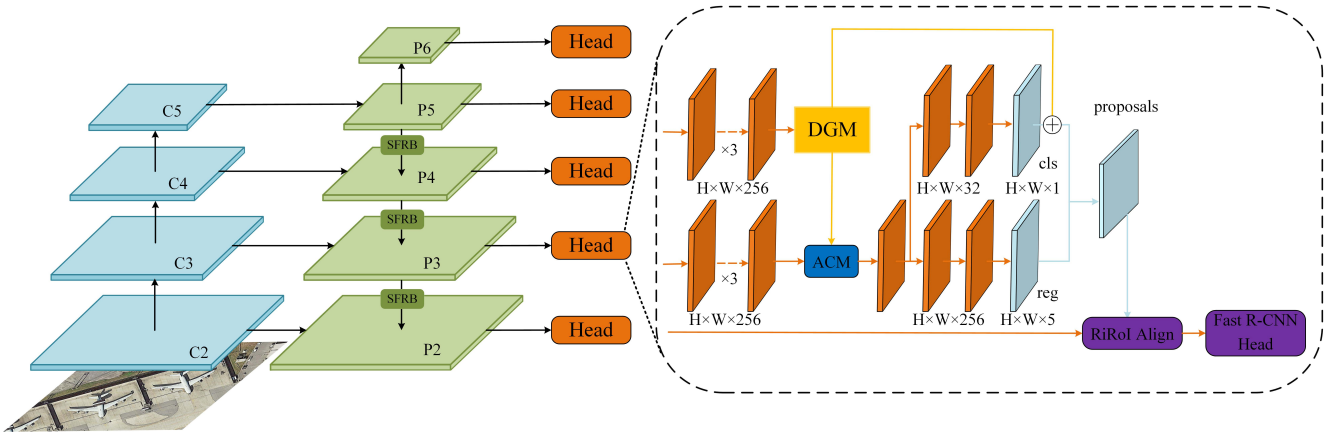

Figure 2. The overall architecture of TARDet. The Spatial Feature Refinement Block (SFRB) is embedded in the feature pyramid to refine the contextual information, and for the output feature maps, we use the Directed Generation Module (DGM) to generate coarse directed boxes. Finally, a fully convolutional network is used to refine the coarse frames to generate high-quality proposals.

these features to the classifier to identify object classes. The first two-stage approach was designed using R-CNN [30], then to a route by [8,9,12], where the general idea is to generate high-quality regions of interest (RoI) from horizontal anchors by RPN, and then to extract accurate features from the RoI using the RoI pooling operator. Finally, the boundary boxes are regressed using R-CNN and classified to improve the R-CNN.

Most of the state-of-the-art aerial object detectors are based on a two-stage framework. Two-stage direction detectors such as [39] handle directional regression by adding anchors with different angles to the region suggestion and RoI regression steps, which allows existing R-CNN-based methods to generate directional bounding boxes by identifying object directional angles. Afterwards, many algorithms have been proposed to improve its performance, including architectural redesign and reform [2,4,18,24], context and attention mechanisms [1, 29, 32], multiscale training and testing [21, 28], training strategies and loss functions [14, 23, 27], feature fusion and enhancement [15, 22]. Today, two-stage anchoring methods based on standard detection benchmarks still maintain state-of-the-art results.

However, it is worth noting that the feature extraction layer of most two-stage algorithms still uses horizontal bounding boxes to extract the corresponding object features, which is limited in predicting oriented bounding boxes, since HBB contains more background information than OBB. Therefore, this will lead to difficulties in extracting overlapping features between objects. Horizontal RoI usually leads to severe misalignment between bounding boxes and oriented objects. For example, horizontal RoI usually contains multiple instances due to the directional and dense objects in the aerial image. A natural solution to alleviate this problem is to use oriented bounding boxes as anchors [26]. Therefore, well-designed anchors with different angles, scales and aspect ratios are needed, but this leads to a large amount of computation and memory usage. In addition, most methods usually present rotated boxes in a complex form to solve the boundary discontinuity problem. The features used in these methods are not aligned with the rotated boxes. These drawbacks hinder the accuracy of the detector.

Therefore, to solve these problems in the two-level detector, we propose a TARDet that generates directed boxes directly by an anchor-free scheme, avoiding the problems caused by horizontal boxes and reducing computational overhead and memory usage.

3. Proposed Method

In this section, the design of the entire TARDet network architecture is described. Each sub-section describes the key modules designed in the network and the refinement process of the network.

3.1. The Framework of TARDet

In this paper, we propose an effective TARDet based on Faster R-CNN, which is an end-to-end network for object detection of remote sensing images. As shown in Fig. 2, our TARDet consists of a feature extraction model and a TARDet head. The feature extraction model consists of a backbone network and a FPN that uses three spatial feature refinement blocks on a top-down path to refine the contextual features between adjacent layers. the TARDet head applies three 3×3 convolutional layers to generate a 256-channel feature mapping, followed by DGM to generate coarse localization boxes in an anchor-free manner, which is refined to a high-quality localization scheme in the second stage, and RiRoI is introduced to adaptively extract

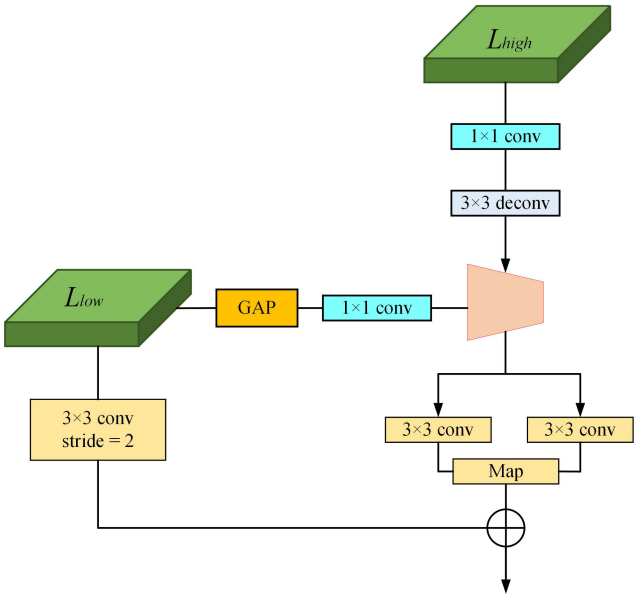

Figure 3. The details of Spatial Feature Refinement Block (SFRB).

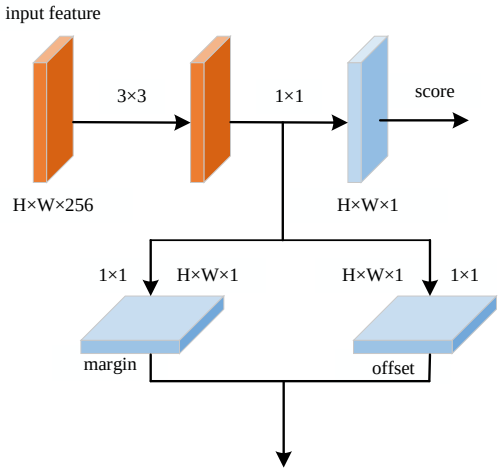

Figure 4. The details of the Directed Generation Module (DGM).

rotationally invariant features from the isovariant features.

3.2. Spatial Feature Refinement Block

The overall architecture of SFRB consists of both sampling point offset learning and global information refinement. The problems of inaccurate sampling of location information on feature maps with high resolution and the inclusion of different semantic information should have different weights are addressed. Inspired by deformable convolution [15], we first perform aggregation of contextual information of adjacent layers and, at the same time, learn the offsets of sampling points and use the aggregated information to refine the content of each sampling point. In our implementation, as shown in Fig. 3, given two adjacent level feature maps L_f and L_{f-1}, we first obtain the weight bootstrap information $\omega \in \mathbb{R}$ of the channels from the adjacent lower level feature map L_{f-1} by a global average pooling layer. Then, the mapping capability is increased by a 1×1 convolution. Also, a 1×1 convolution layer is used to compress the channels of L_f to reduce the computational cost, and L_f is upsampled to the same size as L_{f-1} by a deconvolution layer. Next, we cascade them and use the cascaded feature map as the input of the subnet to generate feature maps with further refinement using global information, with a convolution kernel size of 3×3. Meanwhile, a 3×3 convolution with a step size of 2 is used to reduce the size of the low-level features and add the corresponding elements.

To facilitate model convergence, we use the learned offsets to represent the locations of the sampling points, and the output of the subnetwork is

$$\begin{aligned} \delta &= \mathrm{conv}_1\left(\mathrm{cat}\left(\mathit{deconv}\left(L_f\right), \mathit{GAP}\left(L_{f-1}\right)\right)\right) \\ \omega &= \mathrm{conv}_2\left(\mathrm{cat}\left(\mathit{deconv}\left(L_f\right), \mathit{GAP}\left(L_{f-1}\right)\right)\right) \end{aligned} \tag{1}$$

where $GAP(\cdot)$ denotes the average pooling layer, $cat(\cdot)$ represents the cascade operation, $conv_1(\cdot)$ is the 3×3 convolutional layer with 2 channels, and deconv is the 3×3 deconvolutional layer. We sample from the location of the upsampled feature mapping, and then we add the offset. For training stability, we divide the offset by the average of the L_f length and width.

To solve the quantization problem caused by floating-point offset, we use a bilinear interpolation algorithm [22], which is a mechanism that uses four points adjacent to L_f to obtain a new output $\widetilde{L}_{f-1}$.

Next, we use the bootstrap information ω to further refine the generated feature map $\widetilde{L}_{f-1}$. Specifically, the outputs are added to the underlying feature map by weighting, using a 3×3 convolution layer, and then using a 3×3 convolution with a step size of 2 to reduce the size of the lower-level features and adding the corresponding elements to obtain the final output, mathematically,

$$P_l = \mathit{conv}_2\left(\omega \odot \widetilde{L}_{f-1}\right) + \mathit{conv}_{\mathit{down}}\left(L_{f-1}\right), \tag{2}$$

where $conv_2(\cdot)$ is the 3×3 convolutional layer, $conv_{down}(\cdot)$ is a 3×3 stride convolution.

3.3. Directed Generation Module

The anchor-based object detector uses IOU to separate positive and negative samples between the anchor and

ground truth. However, our DGM regresses the orientation frame directly from the point. As shown in Fig. 1, we define the oriented ground truth frame as $(x_{\rm gt}, y_{\rm gt}, w_{\rm gt}, h_{\rm gt}, \theta_{\rm gt})$. where $\theta_{\rm gt}$ denotes the clock angle between its side and the x-axis satisfying $\theta_{gt} \in [-\pi/4, \pi/4]$. Given a positive sample, its ground truth distance vector with respect to the left, top, right, and bottom sides of the ground truth frame is defined as $t_{\rm gt} = (l, t, r, b)$. Therefore, we use a region assignment scheme instead of the traditional IoU distinction. In the training phase, each ground truth box will first project the feature mapping onto the image according to its box size, and the points located in the center region of the ground truth box will be selected as positive samples, and the other points as negative samples.

We use five levels of feature maps defined as {P2, P3, P4, P5, P6}, where their span steps {s2, s3, s4, s5, s6} are 4, 8, 16, 32, and 64, respectively. after assigning each ground truth to its corresponding feature map, we mark these points as positive if they lie in the center region of the ground truth. The central region of the ground truth can be denoted as , where σ is the central rate. First we transform from the image coordinate system to its ground truth coordinate system by

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta_{\rm gt} & -\sin\theta_{\rm gt} \\ \sin\theta_{\rm gt} & \cos\theta_{\rm gt} \end{pmatrix} \begin{pmatrix} x - x_{\rm gt} \\ y - y_{\rm gt} \end{pmatrix}. \quad (3)$$

If the coordinates of the transformed points simultaneously satisfy then it is proved that the given sample point lies in the center region of the ground truth. Therefore, this is a positive sample.

$$|x'| < \sigma w_{\rm gt}/4 \quad (4)$$

$$|y'| < \sigma h_{\rm gt}/4 \quad (5)$$

As shown in Fig. 4, our DGM goes through three branches to generate scores, margins and offsets, respectively. We only train the margins branch on the positive sample. Since the ground truth coordinate system is not parallel to the image coordinate system, we need to convert each point on the feature map to its corresponding ground truth coordinate system in the same way as the region assignment. Next, the distance vector $t_{\rm gt} = (l, t, r, b)$ can be expressed as

$$l = w_{\rm gt}/4 + x', r = w_{\rm gt}/4 - x' \quad (6)$$

$$t = h_{\rm gt}/4 + y', b = h_{\rm gt}/4 - y' \quad (7)$$

In the definition of a directed box, the angle is in the symmetry interval where $\theta_{gt} \in [-\pi/4, \pi/4]$. Therefore, we directly use the ground truth angle as the target to train the angle branch. By combining margins and offsets, DGM can generate an orientation frame at each location.

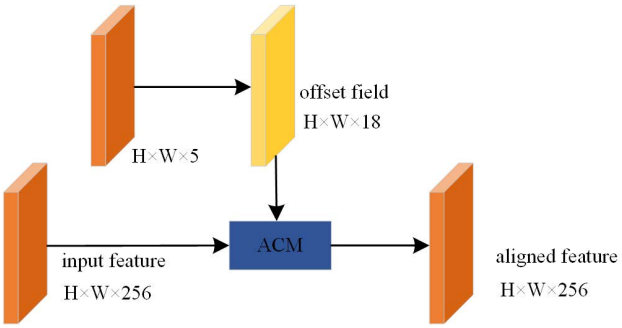

Figure 5. Alignment convolution module takes the input feature and the regression map as inputs and produces aligned features.

3.4. Refine Network

Next, we apply a small fully convolutional network to identify the foreground and refine the roughly oriented boxes precisely. However, the anchoring is consistent across the element map, and they have the same shape and scale at each location. Our coarsely oriented boxes differ at different locations, and these locations have misalignment problems with the consistent feature map. In this work, we design a new efficient ACM that captures the geometric information of the generated coarse OBB and its nearby contextual information to reduce the deviation between the predicted OBB and the ground truth OBB. The module uses a deformable convolutional representation to align features with the OBB. Specifically, given the sampled locations x, y on the feature map, we first regress the initial OBB vector x, y, h, w. Using this initial OBB, we tentatively select 9 sampling points (i.e., 3 rows and 3 columns). Horizontal and vertical offsets are obtained, and an offset field of channel 18 is obtained. These 9 localizations are then mapped onto the feature map and the features at the projection points are convolved by deformable convolution to extract aligned features as shown in Fig. 5. ACM is a light convolution module and the additional computational speed delay is negligible.

After feature alignment, the foreground is distinguished and the oriented boxes are refined by two 1×1 convolutions, respectively. We assign labels to the rough boxes by calculating their IOUs. Here we define the boxes with IOU higher than 0.7 as positive and those lower than 0.3 as negative. Thus, the coarse boxes are refined to the exact position to generate a high-quality detection scheme. We predict the classification scores in the second stage by adding an angle parameter on the regression branch for predicting the angle deviation, and finally, a modified fast R-CNN head is used to predict the classification scores and regress the directed bounding boxes.

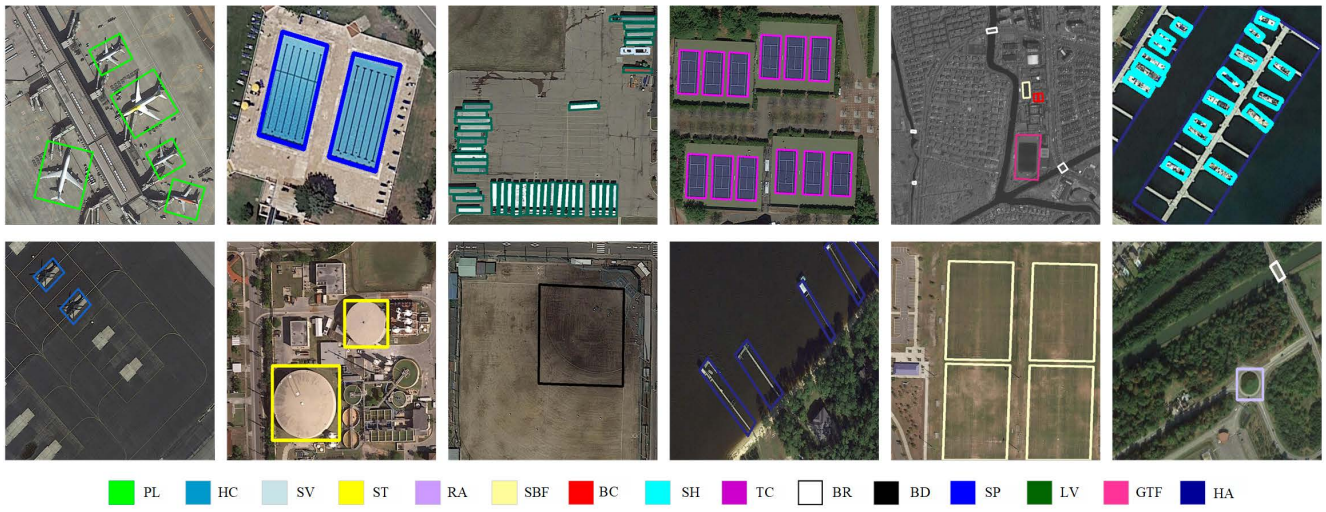

Figure 6. Visualization of detection results on the DOTA dataset.

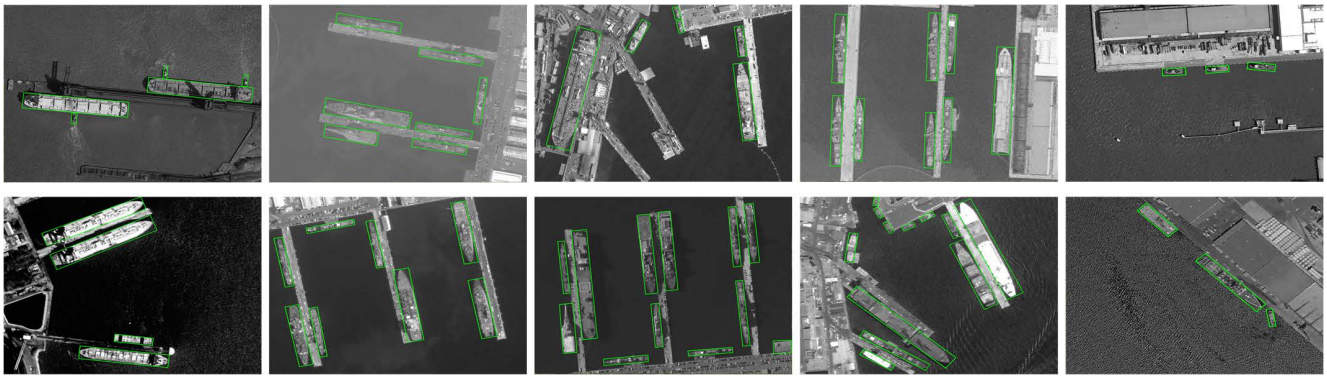

Figure 7. Visualization of detection results on the HRSC2016 dataset.

4. Experiments

In this section, we first describe the implementation details. To validate the effectiveness of the proposed TARDet, we conducted experiments on DOTA and HRSC2016 to evaluate the performance of our approach with several state-of-the-art algorithms. The experiments are also performed on the UG2+ challenge dataset. The details and experimental analysis are described as follows.

4.1. Experimental Settings

Dataset setup. To confirm our results, we conducted experiments on two typical remote sensing datasets, DOTA [38] and HRSC2016 [7]. Specifically, DOTA is a large aerial image dataset for remote sensing object detection, which includes 2806 images as well as 15 object categories. The abbreviations of the categories are defined as: storage tank (ST), swimming pool (SP), traffic circle (RA), helicopter (HC), tennis court (TC), baseball field (BD), small vehicle (SV), large vehicle (LV), soccer field (SBF), basketball court (BC), ship (SH), aircraft (PL), harbor (HA), bridge (BR), and ground athletic field (GTF). In the ablation study, we used the training set for training and the validation set for evaluation. Considering the GPU memory limitation, we cropped the original images to 1024×1024. and performed random flipping to increase the data. For comparison with other methods, we use only one scale for training and testing. The HRSC2016 dataset is a challenging dataset for single ship detection in aerial images, which contains a total of 1061 images in two scenes. The image sizes range from 300×300 to 1500×900. We use the training and validation sets of 617 images for training and evaluate the performance on the test set. All images were resized to (800, 500). The UG2+ challenge dataset is for vehicle detection on aerial images. We use the training set of 240 images for training and test the generalization performance of the model on the test set, where all images are cropped to patches of size 1024×1024.

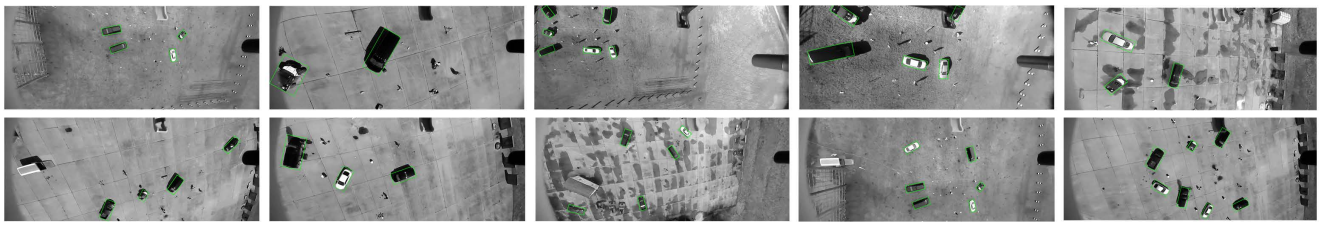

Figure 8. Visualization of detection results on the UG2+ Challenge dataset.

Table 1. Comparison of quantitative results on DOTA datasets. * means multi-scale training and testing.

	Methods	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
Two-stage	FR-O	ResNet101	79.42	77.13	17.70	64.05	35.30	37.16	37.16	89.41	69.64	59.28	50.30	52.91	47.89	47.40	46.30	54.13
	ICN	ResNet101	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
	ROI-Transfomer	ResNet101	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
	CAD-Net	ResNet101	87.80	82.42	49.40	73.50	71.10	63.50	76.70	**90.90**	79.20	73.30	48.40	60.90	62.00	67.00	62.20	69.90
	SCRDet	ResNet101	89.98	80.65	52.09	68.36	68.36	60.32	72.41	90.85	**87.94**	**86.86**	65.02	**66.68**	66.25	68.24	65.21	72.61
	FADet	ResNet101	**90.21**	79.58	45.49	76.41	73.18	68.27	79.56	90.83	83.40	84.68	53.40	65.42	74.17	69.69	64.86	73.28
	Gliding Vertex	ResNet101	89.64	**85.00**	52.26	**77.34**	73.01	73.14	86.82	90.74	79.02	86.81	59.55	70.91	72.94	70.86	57.32	75.02
Single-stage	RetinaNet	ResNet101	88.82	81.47	44.44	65.72	67.11	55.82	72.77	90.55	82.83	76.30	54.19	63.64	63.71	69.73	53.37	68.72
	P-RSDet	ResNet101	89.02	73.65	47.33	72.03	70.58	73.71	72.76	90.82	80.12	81.32	59.45	57.87	60.79	65.21	52.59	69.82
	O^2-DNet	Hourglass104	89.31	82.14	47.33	61.21	71.32	74.03	78.62	90.76	82.23	81.36	60.93	60.17	58.21	66.98	61.03	71.04
	DAL	ResNet101	88.61	79.69	46.27	70.37	65.89	76.10	78.53	90.84	79.98	78.41	58.71	62.02	69.23	71.23	60.65	71.78
	R^3Det	ResNet101	88.76	83.09	50.91	67.27	76.23	**80.39**	86.72	90.78	84.68	83.24	61.98	61.35	66.91	70.63	53.94	73.79
	S^2aNet	ResNet50	89.10	82.85	48.37	71.11	78.15	78.35	87.15	90.93	84.60	85.44	60.56	62.90	65.26	69.13	57.94	74.12
	TARDet (Ours)	ResNet50	88.89	80.35	49.67	69.92	74.05	81.37	86.33	90.87	86.56	83.63	61.46	67.02	75.42	67.88	55.17	74.57
	TARDet (Ours)	ResNet101	90.02	81.81	51.86	71.86	76.52	82.85	87.65	90.56	86.23	86.35	60.26	66.35	75.48	68.28	58.66	75.66
	TARDet* (Ours)	ResNet101	89.70	85.41	**58.28**	79.55	**78.24**	85.54	**89.04**	90.68	85.76	86.33	**69.03**	70.70	**82.16**	**73.37**	69.86	**79.57**

Table 2. Comparison of quantitative results on HRSC2016 datasets.

Methods	Backbone	Image Size	mAP	Speed
R^2CNN	ResNet101	800*800	73.07	5fps
RC2	VGG16	-	75.7	-
RRPN	ResNet101	800*800	79.08	1.5fps
R^2PN	VGG16	-	79.6	-
RoI-Transformer	ResNet101	800*512	86.20	6fps
R^3Det	ResNet101	800*800	89.26	12fps
LARSD	ResNet101	-	90.3	-
TARDet (Ours)	ResNet50	800*800	90.12	18.8fps
	ResNet50	600*600	89.52	23.5fps
	ResNet50	300*300	88.59	33.7fps
	ResNext50	800*800	**92.48**	15.2fps
	ResNext50	600*600	89.53	21.9fps
	ResNext50	300*300	88.62	31.5fps

Implementation Details. We performed a pre-training of ResNet-50 on ImageNet to initialize our backbone. In the next experiments, we trained the model for 24 iterations by setting the initial learning rate of the SGD optimizer to 0.001 and dividing by 10 in each decay step. The weight decay and momentum were 0.0001 and 0.9, respectively. All training was performed on a server with a Tesla V100 GPU (32G), inference was performed with an RTX 2080 Super (8G), and our model was based on the mmdetection library, an open source object detection toolkit based on Pytorch.

Evaluation Metrics. For evaluating our model quantitatively, average precision (AP) and mean AP (mAP) are adopted as the metrics for object detection, and accuracy is used for classification. The larger these metrics are, the better the result are.

4.2. Experimental Results

Results on DOTA Datasets. DOTA: We compared our results in DOTA, as shown in Tab. 1. Compared with the single-stage approach, the two-stage detector model still dominates in DOTA, although its structure is a bit more complex. And the performance is good. Compared with other two-stage methods, our TARDet achieves 74.57% mAP based on ResNet-50 and 75.66% mAP with ResNet-101 backbone. In addition, our method achieves the best performance of 79.57% in the backbone with ResNet-101 using multi-scale training and testing. It surpasses all previous algorithms on the DOTA dataset by testing. Our model also achieves better results in some very challenging categories, such as ships, small vehicles, and bridges. We also visualize some of the results, as shown in Fig. 6.

Results on the HRSC2016 Datasets. The objects in HRSC2016 have arbitrary orientations and large aspect rations. Previous algorithms set more rotation anchors for better performance leading to speed loss. Compared with the previous best result 90.3% by LARSD and 89.26% by R^3Det, we improve 2.18% and 3.22% respectively with only using ResNext50 backbone in 15.2 FPS. Note that our detection strategy can achieve a real-time level while maintaining high accuracy.

Table 3. Ablation study on the HRSC2016 and DOTA datasets.

	Baseline	Different Setting of TARDet		
SFRB		√		√
DGM			√	√
HRSC2016	84.12	85.32	87.12	**90.12**
DOTA	65.02	68.54	69.16	**75.66**

Results on the UG2+ Challenge Datasets. Vehicles in the UG2+ dataset are characterized by arbitrary orientation and small targets. We use the mAP metric to validate the generalization capability of our model. 96.42 mAP was achieved on the UG2+ challenge dataset, and the results show that the model has good generalization performance. We display the results in Fig. 8.

4.3. Ablation Study

To further validate the importance of SFRB and DGB in this method, an ablation study was performed. The experiments were performed on the DOTA and HRSC2016 validation sets, and the results are shown in Tab. 3. By adding SFRB, a 3.52% improvement over the baseline method was achieved on the DOTA dataset, which indicates that refining local contextual information can enhance the target features. After adding DGB, we achieve a 4.14% mAP enhancement over the baseline method on the HRSC2016 dataset, which indicates that DGB can generate directional frames well to cover the target and avoid the problems caused by horizontal frames. In addition, we found that the combination of SFRB and DGB can achieve better performance with 90.12% mAP and 75.66% mAP, respectively. thus, the involvement of both enhances the ability of CNN to extract rotation information, which leads to better regression and classification results.

5. Conclusion

In this paper, we propose a two-stage anchor-free detection framework for localization object detection in aerial images. Our method generates coarse localization boxes in an anchor-free manner and then refines them into a high-quality localization scheme. Among them, a feature refinement module is designed to aggregate feature pyramid context information, and an alignment convolution module is employed to extract alignment features, and RiRoI is introduced to adaptively extract rotationally invariant features from the equivalent features. The ablation study demonstrates the excellent performance of each component of TARDet. Experimental results on the DOTA, HRSC2016 datasets show that our proposed TARDet improves in accuracy and efficiency compared to other detectors.

References

[1] Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2874–2883, 2016. 3

[2] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *European conference on computer vision*, pages 354–370. Springer, 2016. 3

[3] Gong Cheng, Peicheng Zhou, and Junwei Han. Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12):7405–7415, 2016. 1

[4] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems*, 29, 2016. 3

[5] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 2

[6] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning roi transformer for oriented object detection in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2849–2858, 2019. 1

[7] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578, 2019. 2, 6

[8] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2, 3

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 3

[10] Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia. Align deep features for oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2021. 1

[11] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. Redet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2786–2795, 2021. 1

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[13] Shitian He, Huanxin Zou, Yingqian Wang, Runlin Li, Fei Cheng, Xu Cao, and Meilin Li. Enhancing mid–low-resolution ship detection with high-resolution feature distillation. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021. 1

[14] Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang. Bounding box regression with uncertainty for accurate object detection. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 2888–2897, 2019. 3

[15] Tao Kong, Anbang Yao, Yurong Chen, and Fuchun Sun. Hypernet: Towards accurate region proposal generation and joint object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 845–853, 2016. 3, 4

[16] Steven Lang, Fabrizio Ventola, and Kristian Kersting. Dafne: A one-stage anchor-free deep model for oriented object detection. *arXiv preprint arXiv:2109.06148*, 2021. 1

[17] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018. 2

[18] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-aware trident networks for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6054–6063, 2019. 3

[19] Yufeng Li, Caihua Kong, Longgang Dai, and Xiang Chen. Single-stage detector with dual feature alignment for remote sensing object detection. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021. 1

[20] Minghui Liao, Baoguang Shi, and Xiang Bai. Textboxes++: A single-shot oriented scene text detector. *IEEE transactions on image processing*, 27(8):3676–3690, 2018. 2

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2, 3

[22] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3, 4

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2, 3

[24] Youtian Lin, Pengming Feng, Jian Guan, Wenwu Wang, and Jonathon Chambers. Ienet: Interacting embranchment one stage anchor free detector for orientation aerial object detection. *arXiv preprint arXiv:1912.00969*, 2019. 2, 3

[25] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37, 2016. 2

[26] Zikun Liu, Hongzhen Wang, Lubin Weng, and Yiping Yang. Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds. *IEEE Geoscience and Remote Sensing Letters*, 13(8):1074–1078, 2016. 1, 3

[27] Mahyar Najibi, Mohammad Rastegari, and Larry S Davis. G-cnn: an iterative grid based object detector. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2369–2377, 2016. 3

[28] Mahyar Najibi, Bharat Singh, and Larry S Davis. Autofocus: Efficient multi-scale inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9745–9755, 2019. 3

[29] Zheng Qin, Zeming Li, Zhaoning Zhang, Yiping Bao, Gang Yu, Yuxing Peng, and Jian Sun. Thundernet: Towards real-time generic object detection on mobile devices. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6718–6727, 2019. 3

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3

[31] Furong Shi, Tong Zhang, and Tao Zhang. Orientation-aware vehicle detection in aerial images via an anchor-free object detection approach. *IEEE Transactions on Geoscience and Remote Sensing*, 59(6):5221–5233, 2020. 2

[32] Abhinav Shrivastava and Abhinav Gupta. Contextual priming and feedback for faster r-cnn. In *European conference on computer vision*, pages 330–348, 2016. 3

[33] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 1

[34] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 2

[35] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018. 1

[36] Xue Yang, Qingqing Liu, Junchi Yan, Ang Li, Zhiqiang Zhang, and Gang Yu. R3det: Refined single-stage detector with feature refinement for rotating object. *arXiv preprint arXiv:1908.05612*, 2(4):2, 2019. 1, 2

[37] Feng Zhang, Xueying Wang, Shilin Zhou, Yingqian Wang, and Yi Hou. Arbitrary-oriented ship detection through center-head point extraction. *arXiv preprint arXiv:2101.11189*, 2021. 1

[38] Gongjie Zhang, Shijian Lu, and Wei Zhang. Cad-net: A context-aware detection network for objects in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 57(12):10015–10024, 2019. 1, 6

[39] Zenghui Zhang, Weiwei Guo, Shengnan Zhu, and Wenxian Yu. Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks. *IEEE Geoscience and Remote Sensing Letters*, 15(11):1745–1749, 2018. 3

[40] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 2