Anchor-Free Oriented Proposal Generator for Object Detection

Gong Cheng, Member, IEEE, Jiabao Wang, Ke Li, Xingxing Xie, Chunbo Lang, Yanqing Yao, and Junwei Han, Fellow, IEEE

Abstract—Oriented object detection is a practical and challenging task in remote sensing image interpretation. Nowadays, oriented detectors mostly use horizontal boxes as intermediate to derive oriented boxes from them. However, the horizontal boxes are inclined to get small Intersection-over-Unions (IoUs) with ground truths, which may have some undesirable effects, such as introducing redundant noise, mismatching with ground truths, and detracting from the robustness of detectors. In this article, we propose a novel anchor-free oriented proposal generator (AOPG) that abandons horizontal box-related operations from the network architecture. AOPG first produces coarse oriented boxes by a coarse location module (CLM) in an anchor-free manner and then refines them into high-quality oriented proposals. After AOPG, we apply a Fast Region-based Convolutional Neural Network (R-CNN) head to produce the final detection results. Furthermore, the shortage of large-scale datasets is also a hindrance to the development of oriented object detection. To alleviate the data insufficiency, we release a new dataset on the basis of our DIOR dataset and name it DIOR-R. Massive experiments demonstrate the effectiveness of AOPG. Particularly, without bells and whistles, we achieve the accuracy of 64.41%, 75.24%, and 96.22% mAP on the DIOR-R, DOTA, and HRSC2016 datasets, respectively. Code and models are available at https://github.com/jbwang1997/AOPG.

Index Terms—Anchor-free oriented proposal generator (AOPG), oriented object detection, oriented proposal generation.

I. INTRODUCTION

ORIENTED object detection, which exhibits a strong ability to analyze objects in remote sensing images, has attracted extensive attention from researchers. Apart from the function of generic detection, oriented object detection also requires the model to localize objects with oriented boxes so that the predictions can more accurately represent the positions of arbitrary-oriented instances. For the past few years, oriented object detection has made extensive progress [1]–[4] by taking advantage of the deep learning algorithm. However, most of them are still along the route of Faster R-CNN [5] to utilize horizontal anchors and proposals as references to generate oriented bounding boxes.

Referring to horizontal boxes as anchors or proposals to detect oriented objects suffers heavily from several drawbacks. 1) Horizontal boxes usually include massive background regions and multiple objects due to the shape inconsistency with arbitrary-oriented ground truths. This will introduce irrelevant information into region features, severely disturbing the network to obtain good performance. 2) The Intersection-over-Unions (IoUs) between horizontal boxes and oriented ground truths are inclined to be small, which cannot explicitly reflect the matching degrees with ground truths. 3) There is a huge gap, measured in terms of regression target, between horizontal proposals and oriented proposals. Since the regression targets of horizontal proposals are usually extremely large, it will hurt the robustness of the model.

To illustrate the aforementioned problems, we show the horizontal and oriented proposals that are assigned with the same ground truth in Fig. 1. As we can see: 1) the horizontal proposal includes at least four vehicles; thus, it is very difficult for detectors to precisely classify and localize the targets; 2) even the horizontal proposal has been assigned as positive, and it only has a small IoU of 0.16 with the ground truth; in contrast, the oriented proposal has a big IoU of 0.87; and
3) the regression target from the horizontal proposal to the ground truth reaches 2.06, which is very large for training detectors. As shown in the right figure, the regression target of the oriented proposal is just 0.14, which is much smaller than the horizontal proposal.

In the past few years, many researchers have made tremendous effects to address those problems. For example, the RoI transformer [1] attempts to train a lightweight module, named RoI learner, to transform the original horizontal proposals into oriented ones. RoI learner is flexible to plug in all two-stage detectors and significantly increases the detection accuracy. Motivated to alleviate feature misalignment, S²ANet [4] first regresses oriented boxes based on horizontal anchors and then refines the features according to the predictions of the first stage. Next, more accurate oriented boxes are generated from the refined features. However, despite their success, those detectors exploit horizontal anchors as references to produce proposals. They still suffer heavily from the low IoUs and large shape distance problems brought by horizontal boxes. Although rotated RPN [6] straightly slides a bunch of oriented anchors on input images, it needs to place an excessive number of anchors in one pixel (a total of 54 oriented anchors with three scales, three ratios, and six angles) to cover all potential case of ground truths. The densely predefined anchors lead to big computation costs and detract from the model efficiency. What is more, manually designed oriented anchors still cannot match perfectly with ground truths.

To abolish all horizontal boxes in the model and generate high-quality oriented proposals, we deliver a novel anchor-free oriented proposal generator (AOPG) in this work. The main structure of our AOPG is shown in Fig. 2. Instead of sliding fixed-shape anchors on images, we adopt an anchor-free scheme to predict coarse oriented boxes at each position. Then, a novel technique named AlignConv is applied to align the features with coarse oriented boxes. After alignment, we further refine the coarse boxes into accurate locations to generate high-quality oriented proposals. In the second stage, we apply a Fast R-CNN head to predict the classification scores and regress the final oriented bounding boxes.

Furthermore, we observe that the shortage of large-scale datasets severely hinders the development of oriented object detection in remote sensing images. To alleviate the data insufficiency, we annotate the oriented bounding boxes for objects in our DIOR dataset [7] and term it DIOR-R. Up to 23,463 remote sensing images and 192,512 instances covering 20 common classes are included in it.

We conduct comprehensive experiments to demonstrate the flexibility and effectiveness of our AOPG. Attributed to the high-quality oriented proposals, our AOPG achieves 64.41%, 75.24%, and 96.22% mAP without any bells and whistles on the DIOR-R, DOTA, and HRSC2016 datasets.

In summary, our contributions lie in threefold as follows.

1) We analyze the drawbacks of using horizontal boxes in oriented object detection. To avoid those drawbacks, we propose an AOPG. This novel design can produce high-quality proposals without involving horizontal boxes, which largely boosts detection accuracy.

2) We release a new large-scale oriented object detection dataset termed DIOR-R. There are up to 23,463 remote sensing images and 192,512 instances covering 20 common classes in this dataset. We also test several advanced methods on it, and the results show that our DIOR-R is still challenging.

3) We conduct extensive experiments on the DIOR-R, DOTA, and HRSC2016 datasets. The results demonstrate that our AOPG contributes to the improvement of detection accuracy. Particularly, our method obtains large gains over the baseline on all DIOR-R, DOTA, and HRSC2016 datasets.

II. RELATED WORK

Recently, computer vision has rapidly grown with the help of deep learning technology [8]–[14]. As a crucial field of computer vision, object detection receives extensive attention from researchers and has made large progress in the past decades. Object detectors can be mainly separated into two paradigms based on their structures. Two-stage detectors, following the route of Faster R-CNN [5], [15], [16], first generate RoIs from a lightweight fully convolutional network and then extract the RoI features for precise classification and localization. Many latest works [17]–[26] are based on the two-stage framework. Despite leading in many detection benchmarks, two-stage detectors face the problem of high computational complexity and complicated structure. On the contrary, one-stage detectors desire to directly complete detection by one fully convolutional network, which is more efficient and easier to implement. One-stage detectors often suffer from excessive negative samples since there is no sampling operation as in Faster R-CNN. To address this problem, focal loss [27] dynamically gives a large weight for hard samples. With focal loss, one-stage detectors [27]–[30] achieve competitive results with two-stage detectors.

Benefiting from the development of generic detection, oriented object detection also has made great progress [1]–[3], [31]–[39] in recent years. SCRDet [3] alleviates the influence of angle periodicity by designing a novel IoU-smooth L1 loss. R3Det [33] encodes centers and corners information in the features to get a more accurate location. Instead of predicting the angle of oriented boxes, gliding vertex [2] predicts four-point polygons that are more explicit in images.

However, the majority of oriented detectors still use horizontal boxes as references. As discussed in the introduction, this scheme brings many drawbacks to oriented detectors. To avoid the problems brought by horizontal boxes, researchers have made lots of attempts. The RoI transformer [1] trains the RoI learner module to transform horizontal proposals into oriented proposals, which significantly increases the detection accuracy. S²ANet [4] uses the refinement stage to align oriented object features and produce more accurate predictions. Despite their success, those detectors exploit horizontal anchors as references to produce proposals. They still suffer heavily from the low IoUs and large shape distance problems brought by horizontal boxes. Some works [6], [40] directly apply oriented anchors to images
and then generate oriented RoIs or predict oriented objects. However, they need to put a massive number of oriented anchors on images to cover the huge oriented spatial domain. Thus, those detectors are more inefficient and face a serious positive and negative sample imbalance problem. What is more, as demonstrated in [41], anchors need to be specifically designed in different datasets to achieve good results. Even with careful design, anchors still cannot predict some extreme samples.

In recent years, some works attempt to get rid of man-made anchors. Guided anchoring [20] derives anchors from image features and then refines image features guided by the adaptive anchors. Cascade RPN [19], following the guided anchoring route, runs a cascade structure to adapt anchors gradually. FCOS [41] and FoveaBox [42] directly regress the boxes from points. They predict distances from a point to the right, top, left, and bottom sides of objects, which totally abolish anchors. Different from the above works, CornerNet [43] and CenterNet [44] regard object detection as point detection [45]. CornerNet [43] detects and pairs the left-top and the right-bottom points of objects. CenterNet [44] directly predicts the center points of objects. In the aspect of oriented object detection, DRN [32] also adopts the point detection strategy, which brings a new idea to oriented detection. BBAVector [46] adapts CenterNet [44] to the oriented object detector by taking advantage of box boundary-aware vectors.

In this work, we propose a novel AOPG by adopting the anchor-free scheme for generating high-quality oriented proposals. Instead of sliding horizontal anchors on the input images, we straightly generate oriented boxes from points. There is no horizontal box scheme in AOPG architecture, thus avoiding the problems brought by horizontal boxes. We will introduce the details of our model in Section III.

III. OUR APPROACH

In this section, we first investigate the drawbacks of using a horizontal box-based detection scheme for oriented object detection in Section III-A by comparing the proposals from the Faster R-CNN and our AOPG in two aspects. Next, we introduce the details of our AOPG. We adopt the feature pyramid network (FPN) [22] as the backbone. In each feature map, a coarse location module (CLM) is applied to generate...
oriented boxes from feature points. The details of CLM will be presented in section III-B1. After receiving the coarse oriented boxes from CLM, we use the AlignConv operation to eliminate the misalignment between the features and the oriented boxes and then refine the coarse boxes using a small fully convolutional network. We simply average the coarse scores from the CLM and the fine scores from the fully convolutional network in order to get more accurate proposal scores. We will introduce the refinement process in Section III-C.

A. Horizontal-Box Scheme Versus Oriented-Box Scheme

Before introducing our model, we compare the horizontal-box scheme and the oriented-box scheme for oriented object detection in two aspects, namely, IoU distribution and regression target distribution. To this end, we train Faster R-CNN and our AOPG on the DOTA dataset, which are based on horizontal scheme and oriented scheme, respectively. First, we randomly collect 5000 positive samples having extreme targets to regress. Those problems will hurt the final accuracy of detectors and small, which are much fitter for detectors to learn. The central regions of ground-truth boxes as shown in Fig. 3(c), are regular samples. Besides, we calculate the regression targets of these two detectors and give their distributions measured in terms of the height and width target values for horizontal and oriented proposals, respectively, as shown in Fig. 3(b) and (c). It is clear that the targets of horizontal proposals are not symmetrically distributed around the origin. In addition, there are many samples having extreme targets to regress. Those problems will increase the difficulty of training the detectors. In contrast, the targets of oriented proposals, as shown in Fig. 3(c), are regular and small, which are much fitter for detectors to learn.

According to the comparison, we think that involving horizontal box-related operations in the detectors. This investigation motivates us to design AOPG, a novel framework to generate high-quality oriented proposals.

B. Coarse Location Module

1) Oriented Box Definition: As shown in Fig. 4, we define an oriented ground-truth box as \((x_{\text{gt}}, y_{\text{gt}}, w_{\text{gt}}, h_{\text{gt}}, \theta_{\text{gt}})\). Here, \((x_{\text{gt}}, y_{\text{gt}})\) denotes its center coordinate, \(w_{\text{gt}}\) and \(h_{\text{gt}}\) represent its width and the height, and \(\theta_{\text{gt}}\) stands for the clockwise angle between its one side and the X-axis satisfying \(\theta_{\text{gt}} \in [-\pi/4, \pi/4]\). Given a positive sample indicated with \((x, y)\), its ground-truth distance vector with respect to the left, top, right, and bottom sides of ground-truth box is defined as \(t_{\text{gt}} = (l, t, r, b)\).

2) Region Assignment: Anchor-based detectors use IoUs between anchors and ground truths to assign samples as positives or negatives. However, our CLM regresses oriented boxes from points straightly, so there is no anchor in the model. Thus, we apply a region assignment scheme to replace the traditional IoU assignment. In the training phase, each ground-truth box will be first assigned to one feature map of FPN according to its box size. Next, we project the feature maps back onto the images and choose the points that locate in the central regions of ground-truth boxes as positive samples, and the others are negative samples.

Specifically, we make use of five levels of feature maps defined as \(\{P_2, P_3, P_4, P_5, P_6\}\), where their strides \(\{s_2, s_3, s_4, s_5, s_6\}\) are 4, 8, 16, 32, and 64, respectively. The ground truth with the size belonging in \([\alpha^2 s_i^2 / 2, 2 \alpha^2 s_i^2]\) is assigned to \(P_i\), where \(\alpha\) is a factor to scale up the range. Especially, we set the minimum ground-truth size assigned to \(P_2\) as 0 and the maximum ground-truth size assigned to \(P_6\) as 100000 in order to cover all different sizes of ground truths.

After assigning each ground truth to its corresponding feature map, we label the points as positives if they lie in the central regions of ground truths. The central region of a ground truth can be expressed as \(B_{\sigma} = (x_{\text{gt}}, y_{\text{gt}}, \sigma w_{\text{gt}}, \sigma h_{\text{gt}}, \theta_{\text{gt}})\), where \(\sigma\) is the central rate. In order to judge whether a point \((x_{\text{pt}}, y_{\text{pt}})\) on \(P_i\) locates in the \(B_{\sigma}\), we need to convert the point from the image coordinate system to its ground-truth coordinate system by

\[
\begin{pmatrix}
  x' \\
  y'
\end{pmatrix} = \begin{pmatrix}
  \cos \theta_{\text{gt}} - \sin \theta_{\text{gt}} \\
  \sin \theta_{\text{gt}} \cos \theta_{\text{gt}}
\end{pmatrix} \begin{pmatrix}
  x - x_{\text{gt}} \\
  y - y_{\text{gt}}
\end{pmatrix}.
\]

If the coordinate of a point after transformation satisfies \(|x'| < \sigma w_{\text{gt}}/2\) and \(|y'| < \sigma h_{\text{gt}}/2\), the point lies in the central region of the ground truth. Thus, it is a positive sample, as shown in Fig. 4.

3) Training: As shown in Fig. 2, our CLM has three branches, which produce coarse boxes’ scores, distances, and angles in each position, represented by \(c, t,\) and \(\theta\).

The score branch is a classification branch to indicate the central region of ground truths. We train this branch with all points obtained with the aforementioned region assignments.

![Fig. 4. Oriented box definition in our AOPG. The blue box is ground truth. The points located in the blue region are positives.](image-url)
The distance branch predicts the distances from each point on the feature map to the left, top, right, and bottom sides of ground-truth boxes. We only train the distance branch on positive samples. Because the ground-truth coordinate system is unparallel with the image coordinate system, we need to transform each point on the feature map to its corresponding ground-truth coordinate system in the same way as region assignment. Next, as we can see in Fig. 4, the ground-truth angle vector $t_{gt} = (l, r, b)$ of a point with coordinate index $(x, y)$ can be formulated as

\[
\begin{align*}
    l &= w_{gt}/2 + x' , r = w_{gt}/2 - x' \\
    t &= h_{gt}/2 + y' , b = h_{gt}/2 - y'.
\end{align*}
\]

(2)

We train the distance branch by the normalized distance vector $t'_{gt} = (l', r', b')$, where $t'_{gt} = \log(t_{gt}/z)$, and $z$ is a normalizing factor defined in each feature map.

In the definition of oriented boxes, the angle is in a symmetric interval, where $\theta_{gt} \in [-\pi/4 , \pi/4]$. Therefore, we directly use the ground-truth angles as the targets to train the angle branch. Thus, the loss function of CLM can be formulated as

\[
L_{CLM} = \frac{\lambda}{N} \sum_{(x,y)} L_{clt}(c, c^*_{gt}) + \frac{1}{N_{pos}} \sum_{(x,y)} c^*_{gt} L_{dist}(t, t^*_{gt}) + \frac{1}{N_{pos}} \sum_{(x,y)} c^*_{gt} L_{angle}(\theta, \theta_{gt})
\]

(3)

where $L_{clt}$ is the focal loss as in [27], and $L_{dist}$ and $L_{angle}$ are the smooth L1 loss as in [5]. $N$ denotes the number of all points on feature maps, and $N_{pos}$ is the number of positive points. $c^*_{gt}$ indicates the point labels obtained through the region assignment scheme that $c^*_{gt} = 1$ when a point is positive, otherwise 0.

By combining distances and angles, CLM can generate one oriented box in each location. The comparison of the coarse oriented boxes with the horizontal proposals generated by RPN is made in Section IV-C, which shows that the coarse oriented boxes are much better than human-made anchors.

### C. Oriented Box Refinement

Next, we apply a tiny fully convolutional network to identify the foreground and precisely refine the coarse oriented boxes. However, the anchors are uniform on the whole feature map, and they share the same shape and scale in each position. Our coarse oriented boxes vary across locations, which have a misalignment matter with consistent feature maps. In this work, we apply a feature alignment technique called AlignConv to align the features with the coarse oriented boxes. After alignment, high-quality oriented proposals are generated for precise classification and location.

1) Feature Alignment by AlignConv: Fig. 5 shows the framework of AlignConv. The main part of AlignConv is a deformable convolution layer, which aligns the input feature guided by the offset field. Different from normal deformable convolution that generates the offset field from a small network, AlignConv derives the offset field from oriented boxes. To be more specific, for a position vector $p \in [0, 1, \ldots, H-1] \times [0, 1, \ldots, W-1]$, a standard $3 \times 3$ deformable convolution operation can be represented as

\[
Y(p) = \sum_{r \in \mathcal{R}} W(r) \cdot X(p + r + o)
\]

(4)

where $X$ and $Y$ are the input and output features, $W$ is the kernel weights of deformable convolution, $r$ is a vector element from the regular grid $\mathcal{R} = \{(-1, -1), (-1, 0), \ldots, (1, 1)\}$, and $o$ denotes the position offsets. In AlignConv, we constrain the sampling points to obey a regular distribution in the coarse oriented box $B = (x, y, w, h, \theta)$, as shown in Fig. 5. The sampling position can be deduced by

\[
r_{box} = \frac{1}{s_i} ( (x, y) - p + \left( \frac{w}{2}, \frac{h}{2} \right) \cdot R^T(\theta) )
\]

(5)

where $R(\theta)$ is the rotation matrix, the same as [4], and $s_i$ is the stride of feature map $P_i$. We can calculate the offset field $O$ in position $p$ by

\[
O = \sum_{r \in \mathcal{R}} (r_{box} - p - r).
\]

(6)

2) High-Quality Oriented Proposals: After feature alignment, we apply two $1 \times 1$ convolution layers to generate classification and regression maps, separately, targeting to identify the foreground and refine the oriented boxes. For proposal refinement, we need to first calculate the coarse boxes’ IoUs with ground-truth boxes to assign their labels. Here, the coarse boxes are generated via CLM. The boxes with the IoUs higher than 0.7 are considered positives, and the ones lower than 0.3 are regarded as negatives. The others are ignored. With the definitions of positive and negative samples,
D. Oriented Object Detection

As shown in Fig. 6, we adopt a modified Fast R-CNN head in the second stage to predict the classification scores and regress the final oriented bounding boxes. Different from the original Fast R-CNN head, we add an angle parameter on the regression branch for angle bias prediction. The details of AOPG can be referred to in our released code.\footnote{https://github.com/jbwang1997/AOPG}

IV. EXPERIMENTS

A. Datasets and Training Details

1) Datasets: DOTA\footnote{https://captain-whu.github.io/DOTA/dataset.html} [35] is a large-scale aerial object detection dataset containing 2806 images and 188,282 instances of 15 common object classes: harbor (HA), storage tank (ST), large vehicle (LV), swimming pool (SP), soccer ball field (SBF), bridge (BR), ship (SH), plane (PL), baseball diamond (BD), tennis court (TC), ground track field (GTF), basketball court (BC), small vehicle (SV), roundabout (RA),
and helicopter (HC). We use both training and validation sets for training and evaluate the model on the testing set. All results are obtained through the DOTA evaluation server.

HRSC2016\(^3\) [40] is a challenging ship detection dataset, including 1061 images with the size ranging from 300 × 300 to 1500 × 900. In the experiments, both the training set (436 images) and the validation set (181 images) are used for training. We evaluate the models at the testing set (444 images) in terms of PASCAL VOC07 and VOC12 metrics.

In addition to the above datasets, we also release a new oriented object detection dataset termed DIOR-R.\(^4\) Our DIOR-R shares the same images with the previous DIOR dataset [7]. Differently, we annotate the oriented bounding boxes for every instance to adapt to the oriented detection task. There are 23,463 images and 192,518 instances in this dataset, covering a wide range of scenes and 20 common object classes. The categories of objects in DIOR-R include airplane (APL), airport (APO), baseball field (BF), basketball court (BC), bridge (BR), chimney (CH), expressway service area (ESA), expressway toll station (ETS), dam (DAM), golf field (GF), ground track field (GTF), harbor (HA), overpass (OP), ship (SH), stadium (STA), storage tank (STO), tennis court (TC), train station (TS), vehicle (VE), and windmill (WM). We count the instance numbers of different categories in Fig. 7 and visualize some examples in Fig. 8.

Fig. 10. Visualization of the detection results on the DOTA dataset.

Fig. 11. Visualization of the detection results on the HRSC2016 dataset.

2) Training Details: Unless specified, we adopt ResNet50 [47] with FPN [22] as the backbone and use the same hyperparameters with oriented Faster R-CNN. Note that we set the factors $\alpha$ and $\sigma$ in region assignment to 8 and 0.2, respectively. The normalizing factors $z$ are 16, 32, 64, 128, and 256, which correspond to the feature levels $P_2$, $P_3$, $P_4$, $P_5$, and $P_6$. The NMS threshold for proposal generation is set to 0.8. All experiments are conducted on one NVIDIA RTX 2080 Ti with a batch size of 2. In the training phase, we load the model pretrained on the ImageNet [48]. We optimize the model by the stochastic gradient descent (SGD) algorithm with the initial learning rate of 0.005, the momentum of 0.9, and the weight decay of 0.0001. Random horizontal and vertical flipping have been applied during training to avoid overfitting.

The image size of the DOTA dataset ranges from 800 × 800 to 4000 × 4000, which is too big for training. Following [1], we crop original images into a series of 1024 × 1024 patches with a stride of 824. For multiscale experiments, we first resize original images into three scales (0.5, 1.0, 2.0) and then crop all the images of these three scales into 1024 × 1024 with a stride of 524. We run 12 epochs and reduce the learning rate by a factor of 10 at the end of epoch 8 and epoch 11.

For the DIOR-R dataset, we keep the image size at the original size of 800 × 800. The same training schedule as the DOTA dataset is implemented. For the HRSC2016 dataset, we resize the images to (800, 1333) without changing the aspect ratios. We run 36 epochs and reduce the learning rate by a factor of 10 at epoch 24 and epoch 33.
TABLE I

| Methods          | R50-500PN | R101-500PN | R50-FPN | R101-FPN | Gliding Vertix | RoI Transformer | Faster R-CNN-O* |
|------------------|-----------|------------|---------|----------|----------------|----------------|----------------|
| Faster R-CNN-O*  | 75.39     | 75.39      | 75.39   | 75.39    | 75.39          | 75.39          | 75.39          |
| RetinaNet-O*     | 75.39     | 75.39      | 75.39   | 75.39    | 75.39          | 75.39          | 75.39          |
| DRN [32]         | 75.39     | 75.39      | 75.39   | 75.39    | 75.39          | 75.39          | 75.39          |
| DAL [51]         | 75.39     | 75.39      | 75.39   | 75.39    | 75.39          | 75.39          | 75.39          |
| RSDet [52]       | 75.39     | 75.39      | 75.39   | 75.39    | 75.39          | 75.39          | 75.39          |
| R3Det [33]       | 75.39     | 75.39      | 75.39   | 75.39    | 75.39          | 75.39          | 75.39          |
| S3ANet [4]       | 75.39     | 75.39      | 75.39   | 75.39    | 75.39          | 75.39          | 75.39          |
| AOPG             | 75.39     | 75.39      | 75.39   | 75.39    | 75.39          | 75.39          | 75.39          |

TABLE II

| Method          | Backbone | PL | BD | BR | GTF | SV | LV | SH | TC | BC | ST | SBF | RA | HA | SP | HC | mAP | FPS |
|-----------------|----------|----|----|----|-----|----|----|----|----|----|----|-----|----|----|----|----|-----|-----|
| One-stage       |          |    |    |    |     |    |    |    |    |    |    |     |    |    |    |    |    |
| APE [30]        |          | 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10|
| RetinaNet-O*    |          | 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10|
| DRN [32]        |          | 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10|
| DAL [51]        |          | 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10|
| RSDet [52]      |          | 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10|
| R3Det [33]      |          | 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10|
| S3ANet [4]      |          | 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10|
| AOPG            |          | 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10|
| Two-stage       |          |    |    |    |     |    |    |    |    |    |    |     |    |    |    |    |    |
| APE [30]        |          | 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10|
| RetinaNet-O*    |          | 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10|
| DRN [32]        |          | 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10|
| DAL [51]        |          | 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10|
| RSDet [52]      |          | 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10|
| R3Det [33]      |          | 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10|
| S3ANet [4]      |          | 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10|
| AOPG            |          | 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10| 10|

TABLE III

| Method          | RRPN [6] | R2CNN [53] | RT [1] | GV [2] | DRN [32] | PnPU [50] | DAL [51] | S3ANet [4] | AOPG |
|-----------------|----------|------------|--------|--------|----------|-----------|----------|------------|------|
| mAP (VOC 07)   | 79.08    | 73.07      | 86.20  | 88.20  | 89.20    | 89.77     | 87.54    | 90.17      | 90.34|
| mAP (VOC 12)   | 85.64    | 79.73      | -      | 92.70  | -        | 95.01     | 96.22    |            |      |
Fig. 12. Proposals generated with different methods, including RPN, CLM, and CLM+.

TABLE IV
COMPARISON OF DIFFERENT ORIENTED PROPOSAL GENERATORS (OPGS) ON THE DOTA VALIDATION SET

| Oriented Proposal Generators (OPGs) | AlignConv | Recall (%) | mAP (%) |
|-----------------------------------|-----------|------------|---------|
| RoI-Trans                         | ✓         | 87.86      | 73.12   |
| Anchor-based OPG                  | ✓         | 86.14      | 72.53   |
| FCOS-based OPG                    | ✓         | 86.96      | 73.09   |
| CLM                               | ✓         | 85.44      | 72.17   |
| CLM+                              | ✓         | 86.32      | 72.69   |
|                                  | ✓         | 89.82      | 73.69   |
|                                  | ✓         | 90.68      | 74.16   |

TABLE V
VALUATIONS OF AOPG WITH DIFFERENT FRAMEWORKS

|                   | Faster R-CNN | Cascade R-CNN | Double Head |
|-------------------|--------------|---------------|-------------|
| RPN               | 69.05        | 73.45         | 72.03       |
| AOPG              | 75.24        | 76.31         | 75.56       |

model. As shown, our AOPG achieves 90.34% and 96.22% mAP when using the VOC07 metric and the VOC12 metric, respectively, both of which are very competitive compared with current state-of-the-art methods. We visualize the detection results in Fig. 11.

C. Comparisons of Different Oriented Proposal Generators

We conduct experiments to compare our AOPG with other OPGs. The experiments are conducted on the DOTA validation set, and the results are reported in Table IV. Here, RoI-Trans indicates applying RoI learner in RoI transformer [1] as the proposal generator. Anchor-based OPG means using horizontal anchors and directly regressing the angles in the proposal generation phase. FCOS-based OPG means applying the FCOS regression method when generating oriented proposals.

As we can see in Table IV, RoI-Trans, anchor-based OPG, and FCOS-based OPG can only reach 73.12%, 72.53%, and 72.17% mAP, respectively. After applying AlignConv, their accuracies are boosted to 73.09% and 72.69%. In contrast, our CLM can achieve 73.69% mAP on its own and 74.16% with AlignConv, which surpasses other OPG methods. We also compare the recall of different proposal generation methods on the DOTA validation set. Our CLM achieves 89.82% on its own and up to 90.68% with AlignConv. Those results confirm the motivation of AOPG that relying on horizontal boxes will harm the oriented object prediction. Our AOPG is in an anchor-free manner and gets rid of horizontal boxes so that it outperforms other methods. What is more, the refinement module (AlignConv) boosts the recall gains of anchor-based OPG, FCOS-based OPG, and CLM with the margins of 0.82, 0.88, and 0.86 points, respectively, and also improves the mAP gains with the margins of 0.56, 0.52, and 0.47 points. The results demonstrate that the refinement module can indeed further improve the quality of the proposals and benefit the oriented detection.

To further illustrate the effectiveness, we visualize the top 300 proposals generated by traditional RPN and the different parts of AOPG. The comparison results are given in Fig. 12. Here, CLM+ indicates CLM with AlignConv. We can see that the horizontal proposals generated by RPN are confused in the places where the objects are densely packed. For instance, the horizontal proposals cannot separate harbors and large vehicles in the images. In contrast, the oriented proposals generated by the CLM can well localize them and also better fit the ground truths than horizontal proposals. Furthermore, the CLM+ produces more accurate oriented proposals and eliminates some false positive samples generated by CLM.

D. Evaluations of AOPG With Different Frameworks

To further boost the performance and evaluate the consistency of the effectiveness of AOPG, we supplement the
experiments based on more advanced frameworks, such as Cascade R-CNN [18] and double head [55]. The results are reported in Table V. As we can see, the Cascade R-CNN and double head can only achieve 73.45% and 72.03% mAPs. However, after replacing the original RPN with AOPG, their accuracies measured in terms of mAP increase to 76.31% (2.86↑) and 75.56% (3.53↑), which displays the consistent effectiveness of our work.

V. CONCLUSION

In this article, we analyzed the drawbacks of using horizontal boxes for oriented object detection and concluded that involving horizontal boxes in object detectors will hurt the final results. Taking this problem as a starting point, we proposed a novel AOPG that totally removes all horizontal boxes in the network. Moreover, we released a new large-scale oriented object detection dataset, named DIOR-R, to alleviate the shortage of data. Comprehensive experiments show that our AOPG contributes to the increase in detection accuracy. Specifically, our AOPG achieves the highest accuracy without any bells and whistles, and obtains large gains over the baseline on all DIOR-R, DOTA, and HRSC2016 datasets.

REFERENCES

[1] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, “Learning RoI transformer for oriented object detection in aerial images,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 2849–2858.
[2] Y. Xu et al., “Gliding vertex on the horizontal bounding box for multi-oriented object detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 4, pp. 1452–1459, Apr. 2021.
[3] X. Yang et al., “SCRDet: Towards more robust detection for small, cluttered and rotated objects,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 8232–8241.
[4] J. Han, J. Ding, J. Li, and G.-S. Xia, “Align deep features for oriented object detection,” IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–11, 2022.
[5] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[6] J. Ma et al., “Arbitrary-oriented scene text detection via rotation proposals,” IEEE Trans. Multimedia, vol. 20, no. 11, pp. 3111–3122, Nov. 2018.
[7] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, “Object detection in optical remote sensing images: A survey and a new benchmark,” ISPRS J. Photogramm. Remote Sens., vol. 159, pp. 296–307, Jan. 2020.
[8] X. Sun, B. Wang, Z. Wang, H. Li, H. Li, and K. Fu, “Research progress on few-shot learning for remote sensing image interpretation,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 14, pp. 2387–2402, 2021.
[9] Q. He, X. Sun, Z. Yan, and K. Fu, “DABNet: Deformable contextual and boundary-weighted network for cloud detection in remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–16, 2022.
[10] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, “When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs,” IEEE Trans. Geosci. Remote Sens., vol. 56, no. 5, pp. 2811–2821, May 2018.
[11] G. Cheng et al., “Prototype-CNN for few-shot object detection in remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–10, 2022.
[12] X. Feng, J. Han, X. Yao, and G. Cheng, “TCANet: Triple context-aware network for weakly supervised object detection in remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 8, pp. 6946–6955, Aug. 2021.
[13] G. Cheng, X. Sun, K. Li, L. Guo, and J. Han, “Perturbation-seeking generative adversarial networks: A defense framework for remote sensing image scene classification,” IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–11, 2022.
[39] C. Li, C. Xu, Z. Cui, D. Wang, T. Zhang, and J. Yang, “Feature-attentioned object detection in remote sensing imagery,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2019, pp. 3886–3890.

[40] Z. Liu, H. Wang, H. Weng, and L. Yang, “Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds,” IEEE Geosci. Remote Sens. Lett., vol. 13, no. 8, pp. 1074–1078, Aug. 2016.

[41] Z. Tian, C. Shen, H. Chen, and T. He, “FCOS: Fully convolutional one-stage object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 9627–9636.

[42] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi, “FoveaBox: Beyond anchor-based object detection,” IEEE Trans. Image Process., vol. 29, pp. 7389–7398, 2020.

[43] H. Law and J. Deng, “CornerNet: Detecting objects as paired keypoints,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 754–770.

[44] X. Zhou, D. Wang, and P. Krähenbühl, “Objects as points,” 2019, arXiv:1904.07850.

[45] J. Yang, Q. Liu, and K. Zhang, “Stacked hourglass network for robust facial landmark localisation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jul. 2017, pp. 79–87.

[46] J. Yi, P. Wu, B. Liu, Q. Huang, H. Qu, and D. Metaxas, “Oriented object detection in aerial images with box boundary-aware vectors,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Jan. 2021, pp. 2150–2159.

[47] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 630–645.

[48] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 248–253.

[49] K. Chen et al., “MMDetection: Open MMLab detection toolbox and benchmark,” 2019, arXiv:1906.07155.

[50] Z. Chen et al., “PBox loss: Towards accurate oriented object detection in complex environments,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 195–211.

[51] Q. Ming, Z. Zhou, L. Miao, H. Zhang, and L. Li, “Dynamic anchor learning for arbitrary-oriented object detection,” 2020, arXiv:2012.04150.

[52] W. Qian, X. Yang, S. Peng, Y. Guo, and J. Yan, “Learning modulated loss for rotated object detection,” 2019, arXiv:1911.08299.

[53] Y. Jiang et al., “R2CNN: Rotational region CNN for orientation robust scene text detection,” 2017, arXiv:1706.09579.

[54] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The Pascal visual object classes (VOC) challenge,” Int. J. Comput. Vis., vol. 88, no. 2, pp. 303–338, Sep. 2009.

[55] Y. Wu et al., “Rethinking classification and localization for object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 10186–10195.

Gong Cheng (Member, IEEE) received the B.S. degree from Xidian University, Xi’an, China, in 2007, and the M.S. and Ph.D. degrees from Northwestern Polytechnical University, Xi’an, in 2010 and 2013, respectively. He is currently a Professor with Northwestern Polytechnical University. His main research interests are computer vision, pattern recognition, and remote sensing image understanding.

Dr. Cheng is also an Associate Editor of IEEE Geoscience and Remote Sensing Magazine and a Guest Editor of IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING.

Jiabao Wang received the B.S. degree from Northwestern Polytechnical University, Xi’an, China, in 2019, where he is currently pursuing the master’s degree. His research interests include computer vision and pattern recognition, especially object detection.

Ke Li received the M.Sc. and Ph.D. degrees from the Zhengzhou Institute of Surveying and Mapping, Zhengzhou, China, in 2005 and 2008, respectively. He is currently an Associate Professor with the Zhengzhou Institute of Surveying and Mapping. His research interests include geographic information systems, deep learning, and remote sensing image analysis.

Xingxing Xie received the B.S. degree from Inner Mongolia University, Huhhot, China, in 2015, and the M.S. degree from Northwestern Polytechnical University, Xi’an, China, in 2018, where he is currently pursuing the Ph.D. degree. His main research interests are computer vision and pattern recognition.

Chunbo Lang received the B.Eng. degree from Northeast Forestry University, Harbin, China, in 2020. He is currently pursuing the Ph.D. degree with Northwestern Polytechnical University, Xi’an, China, under the supervision of Prof. Gong Cheng. His research interests include few-shot learning, segmentation, and remote sensing image understanding.

Yanqing Yao received the B.S. and M.S. degrees from the Inner Mongolia University of Science and Technology, Buotou, China, in 2016 and 2019, respectively. She is currently pursuing the Ph.D. degree. Her research interests include computer vision and remote sensing image analysis.

Junwei Han (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in pattern recognition and intelligent systems from Northwestern Polytechnical University, Xi’an, China, in 1999, 2001, and 2003, respectively. He was a Research Fellow with Nanyang Technological University, Singapore, The Chinese University of Hong Kong, Hong Kong, Dublin City University, Dublin, Ireland, and the University of Dundee, Dundee, U.K., from 2003 to 2010. He is currently a Professor with Northwestern Polytechnical University. His research interests include computer vision and brain imaging analysis.