SRPAR: anchor-free detector with aspect ratio priority for slender objects

Hong-Gang Xie, Ming Yang, Bo-Lun Yan, Kai-Yuan Hou, and Di Jiang

Abstract. Slender objects are more difficult to detect than conventional objects with regular shapes, and the rotation of these objects makes detecting them even more challenging. To address these challenges, we proposed SRPAR, an anchor-free detector with aspect ratio priority for slender objects. The aspect ratio priority factor is designed based on the object’s aspect ratio and rotation angle. The aspect ratio priority factor guides the regression of slender rotating objects and improves the regression accuracy. In addition, through the multi-level prediction of feature pyramid network, the range of bounding boxes and corresponding angular regressions at each level is limited, and the regression of overlapping object prediction boxes is accelerated. To better evaluate the SRPAR’s detection performance of slender rotating objects having an aspect ratio of at least 3:1, some images of sticks are supplemented into the baseball bat subset of the common objects in context (COCO) dataset to form a new self-made COCO-Stick dataset. Experimental results on the dataset of object detection in aerial images dataset and the self-made COCO-Stick dataset show that, compared with state-of-the-art detectors, the proposed method has some advantages in detection accuracy. © The Authors. Published by SPIE under a Creative Commons Attribution 4.0 International License. Distribution or reproduction of this work in whole or in part requires full attribution of the original publication, including its DOI. [DOI: 10.1117/1.JEI.31.4.043001]

Keywords: object detection; aspect ratio; anchor-free.

1 Introduction

With the rapid development of deep learning, many high-performance general object detection methods have been proposed. However, the current general detectors are no longer sufficient for detecting some specific types of objects, such as sticks and other slender objects with aspect ratios of at least 3:1. Figure 1 shows a stick-like object from the common objects in context (COCO) dataset, and the angle between the slender object and the horizontal ground is represented by $\theta$. Regarding slender objects with large aspect ratios, Tian et al. noted that, even with careful design, because the scales and aspect ratios of anchor boxes are kept fixed, detectors encounter difficulties when dealing with object candidates with large shape variations. Therefore, anchor-free object detectors may be better at detecting objects with large shape variations like slender objects. However, an anchor-free detector usually detects an object by locating the key points (corner points, extreme points, and center points) on that object. During the detection process, the key points on the object inevitably fall into multiple bounding boxes, and the network has difficulties in choosing which bounding box to return. The location that causes bounding box overlap is defined as a fuzzy sample. In addition, the aspect ratios of objects in datasets at home and abroad are mostly between 1:1 and 3:1, which does not match the requirement of slender object detection. As a result, the main problems of anchor-free slender object detection include large aspect ratios, overlapping bounding boxes, and a lack of suitable datasets.

*Address all correspondence to Hong-Gang Xie, xiehg@hbut.edu.cn; Ming Yang, ymut@hbut.edu.cn
Many anchor-free object detectors have been proposed in recent years. Law and Deng proposed CornerNet, which determines an anchor box by detecting and matching the two corner points of the object’s upper left corner and lower right corner and then detects the object using edge information. Zhou et al. proposed CenterNet, which shows the object through the object’s center point, returns some of the object’s information at the center point, and then completes the object detection. When an anchor-free object detector detects objects, however, fuzzy samples that cause bounding boxes to overlap will definitely arise, lowering regression accuracy. Tian et al. proposed fully convolutional one-stage object detector (FCOS), which reduces the number of fuzzy samples by utilizing multi-level prediction to limit the range of bounding box regression at each level.

When the slender object and horizontal ground are at a certain angle, the horizontal box cannot accurately indicate the position information of the object. Therefore, the rotating box may be better at revealing the slender object’s position information. Zhang et al. proposed R2PN, which is a rotating regional proposal network to generate multi-directional candidate proposals containing the object angle information, hence improving regional proposal quality. In addition, due to the slender object’s unique shape with a large aspect ratio, which causes a considerable variation in the decay rates of its long and short sides during regression, positioning of the slender object during regression becomes inaccurate. As a result, the slender object’s regression accuracy is reduced. To this end, Yang et al. proposed the derivable skew intersection-over-union (IoU) loss, which improves the accuracy of regression of slender objects to some extent.

In response to the above problems, an anchor-free detector with aspect ratio priority for slender rotation object (SRPAR) is proposed in this paper. First, to solve the problem that the long and short sides of a slender object do not decay at the same rate during regression, the aspect ratio priority factor, which improves the slender rotating object’s regression accuracy, was designed to guide the regression of the slender rotating object. Second, to reduce the impact of fuzzy samples on object detection accuracy, the feature pyramid network (FPN) multi-level prediction was employed to limit the range of bounding boxes and corresponding angle regressions at each level, thus significantly reducing the number of fuzzy samples. Finally, the self-made COCO-Stick dataset was created to better evaluate SRPAR’s detection performance, and a series of experiments were conducted on both the dataset of object detection in aerial images (DOTA) dataset and self-made COCO-Stick dataset.

In summary, the work of this article mainly has the following contributions:

1. The aspect ratio priority factor is designed. It is utilized to guide the regression of slender rotating objects and improve regression accuracy.
2. The multi-level prediction of FPN is employed. Through multi-level FPN prediction, the range of bounding boxes and corresponding angular regressions at each level are limited, and the number of fuzzy samples is greatly reduced.
3. The self-made COCO-Stick dataset is created. SRPAR’s detection performance of slender rotating objects having aspect ratios of at least 3:1 is better evaluated.

The remainder of the paper is structured as follows. First, related work is reviewed in Sec. 2. Second, in Sec. 3, the proposed SRPAR is introduced in detail. Then, in Sec. 4, related experiments are designed to evaluate SRPAR’s detection performance. Finally, in Sec. 5, the conclusions of this research are given.

2 Related work

The majority of the objects detected by the proposed SRPAR are stick-like objects with large aspect ratios. Due to the unique shapes of the stick-like objects with large aspect ratios, as well as the fact that the stick-like objects and horizontal ground are at certain angles, the detection of stick-like objects becomes even more challenging. Anchor-free object detectors may be better at detecting objects with large aspect ratios. In the remainder of this section, relevant journal articles are consulted about anchor-free object detection and rotating objects to further refine the proposed SRPAR by learning from the relevant algorithms, and thus SRPAR’s detection accuracy is improved.

2.1 Anchor-Free Object Detection

The anchor-free object detection algorithm minimizes the amount of calculation by avoiding complex calculations on anchor boxes. The anchor-free algorithm primarily predicts the object’s corner, extreme, or center points and then processes these key points to generate the object’s detection box.

The anchor-free detection algorithm, based on corner or extreme points prediction, generates the detection box of an object by processing the bounding points. By detecting and matching these two corner points in the top left and bottom right corners of the object, the CornerNet algorithm, which only uses edge information to detect objects, determines an anchor box. However, the object’s inside information is typically more recognized, making erroneous detection more likely. Zhou et al. proposed CenterNet, which adds the detection of the center point of the object to the CornerNet algorithm to overcome this problem, which improves the detection accuracy. Zhou et al. proposed ExtremeNet, which locates the anchor box by locating the four extreme points at the top, bottom, left, and right edges of the object, as opposed to the previous two algorithms that determine the anchor box by corner points. Yang et al. proposed RepPoints, which is based on the detection of key points. Unlike CornerNet and ExtremeNet, RepPoints does not have a matching problem between key points. Instead, it indicates the object with a novel point set; the object detection performance is promising.

The anchor-free algorithm, based on center prediction, locates the detection box by the distance from the center point to the four boundaries. FCOS is an object detection algorithm based on pixel-level prediction that locates the anchor box by predicting the distance from the center point to the object boundary. Furthermore, none of the positions far from the object center produced high-quality prediction boxes; thus center-ness was added to suppress these low-quality prediction boxes, hence improving detection accuracy.

2.2 Rotating Object Detection

Because of the unique shape of slender objects’ large aspect ratios, when rotations occur, significant variations in angle are caused, making detection of slender objects with angular rotations more challenging. R2PN is a rotating regional proposal network that generates multi-directional candidate proposals containing object angle information, hence improving regional proposal quality. However, if the slender object’s lean angle is relatively small, the slender object can easily be mistakenly detected as being horizontal. Xu et al. proposed a gliding vertex, which proposes a tilt factor that efficiently solves the problem of near-horizontal object detection.

The majority of boundary discontinuities in rotating object detectors based on regression are caused by angular periodicity. To this purpose, Yang et al. proposed the IOU smooth L1
loss, which solves the angular periodicity-induced boundary discontinuity problem. When the arrangement of slender rotating objects is denser, the object instances are highly overlapped during the detection process, resulting in unsatisfactory detection.\textsuperscript{18-21} To overcome this problem, Yang et al.\textsuperscript{10} proposed a single-stage end-to-end detector with a combination of rotating and horizontal anchor boxes to adapt to a dense arrangement of rotating objects.

In summary, a new object detection model (SRPAR) is proposed in this paper. To address the problem that the long and short sides of a slender object do not decay at the same rate during regression, the aspect ratio priority factor is designed to guide the regression of slender objects, thus improving the regression accuracy. Furthermore, to reduce the impact of fuzzy samples on object detection, FPN\textsuperscript{11} multi-level prediction is employed to limit the range of bounding boxes and corresponding angular regressions at each level, thus minimizing the number of fuzzy samples.

3 Our Approach
As shown in Fig. 2, the network structure is divided into four parts: the backbone network, the FPN, the feature selection module, and the multitasking subnetwork. Because the dataset to be trained is not complicated and a balance between detection accuracy and speed is needed, ResNet-101\textsuperscript{22} was chosen as the backbone network. After the fuzzy samples are reduced by FPN\textsuperscript{11} multi-level prediction, the feature selection module inputs the feature-selected multi-scale feature map into the multitasking subnetwork. Regression and classification subnetworks make up the multitasking subnetwork. The categories are predicted through the classification subnet. The regression subnet is responsible for predicting the distances $t$, $b$, $l$, and $r$ from the prediction point to the top, bottom, left, and right edges of the bounding box, respectively, as well as the corresponding rotation angle $\theta$ of the bounding box.

3.1 FPN Multi-Level Prediction
When an anchor-free object detector detects objects, however, fuzzy samples that cause bounding box overlap undoubtedly arises. A location will fall into multiple bounding boxes at the same time, and the network will have difficulty in deciding which bounding box to return. For the detection of horizontal objects, the range of the bounding box regression at each level is limited using FPN\textsuperscript{11} multi-level prediction, and thus the number of fuzzy samples is reduced. However, in this paper, the object to be detected has a rotation angle, and therefore, only limiting the range of the bounding box regression without considering the angle is inappropriate. Therefore,
through the multi-level prediction of FPN, the range of bounding boxes and corresponding angular regressions at each level are limited, and the number of fuzzy samples is significantly reduced. The effectiveness of this approach is also confirmed in the ablation study in Sec. 4.1.

The regression subnetwork is in charge of regressing the five variables $t, b, l, r$, and $\theta'$. $t, b, l, r$, and $r$ are the distances between the prediction point and the bounding box’s top, lower, left, and right edges, respectively, and $\theta'$ is the rotated object’s prediction angle. The mapping relationship between the five variables $t, b, l, r, \theta'$ and the ground-truth box denotes a prediction box, when the prediction box overlaps with the object bounding box. The corresponding bounding box is shown in Fig. 3. The coordinates of the upper left and bottom right corners of the bounding box are $B(x_B, y_B)$ and $A(x_A, y_A)$, respectively. The coordinates of the two points are denoted by the object’ five-parameter label coordinates of $(x_c, y_c, w, h, \theta)$. The bounding box’s rotation angle is $\theta$, which denotes the acute angle of the bounding box to the $x$ axis. According to the long-edge definition method of angles, $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$.

When any prediction point $p(x, y)$ falls in the object bounding box, the prediction point is considered to be a positive sample, and the regression subnetwork is in charge of regressing the five variables $t, b, l, r, \theta'$. The four variables $t, b, l, r$ of regression are denoted by the three coordinates $B(x_B, y_B), A(x_A, y_A)$, and $p(x, y)$, as shown in Eq. (1).

$$
\begin{align*}
    l &= \sqrt{(y_B - y)^2 + (x_B - x)^2} \times \sin \left( \arctan \frac{y_B - y}{x_B - x} - \theta' \right), \\
    t &= \sqrt{(y_B - y)^2 + (x_B - x)^2} \times \cos \left( \arctan \frac{y_B - y}{x_B - x} - \theta' \right), \\
    r &= \sqrt{(y - y_A)^2 + (x - x_A)^2} \times \sin \left( \arctan \frac{y - y_A}{x - x_A} - \theta' \right), \\
    b &= \sqrt{(y - y_A)^2 + (x - x_A)^2} \times \cos \left( \arctan \frac{y - y_A}{x - x_A} - \theta' \right). 
\end{align*}
$$

The range of bounding box regression at each level is restricted by FPN multi-level prediction. Specifically, five levels of feature maps defined as $\{P3, P4, P5, P6, P7\}$ are employed. $P3, P4$, and $P5$ are produced by the backbone convolutional neural network (CNN) feature maps. $C3, C4$, and $C5$ are followed by a $1 \times 1$ convolutional layer, as shown in Fig. 4. $P6$ and $P7$ are obtained by applying one convolutional layer with the stride being 2 on $P5$ and $P6$, respectively. The size of the convolution kernel is $3 \times 3$. $H \times W$ represents the height and width of feature maps, and the input images in the model are all $600 \times 400$ in size.

For the fuzzy sample problem, the specific approach chosen is to compute the regression variables $t, b, l, r$, which correspond to the positions of the predicted points on the feature map.
level. Limiting only the regression range of the bounding box without the corresponding angle is not suitable. If the position at feature level \( j \) (\( j \) refers to any level) satisfies \( \theta_j \leq \frac{\pi}{2} \) and \( m_j \leq \max(l, t, r, b) \leq m_j \), the position at the feature level is set as a positive sample, and a bounding box is regressed. Otherwise, no regression is required. In \( m_j \leq \max(l, t, r, b) \leq m_j \), \( m_j \) is the maximum distance to be regressed for a pixel at feature level \( j \). In this study, \( m_2, m_3, m_4, m_5, m_6, \) and \( m_7 \) are set as 0, 38, 75, 150, 300, and \( \infty \). Because the input image size in the model is 600 \times 400, the specific values of 38, 75, 150, and 300 are taken. And the maximum pixel distance that the points on the image need to be returned to is 600. The maximum pixel distance of 600 is divided by the number of steps in the convolution layer of 2. The value of 300 is taken as the maximum distance of \( m_6 \), which is then divided by the number of steps in the convolution layer of 2, and the maximum distance of \( m_5 \) is obtained as 150. In the same way, the values of \( m_4 \) and \( m_3 \) are obtained. The similar approach was used in Tian et al.’s work,\(^2\) which experimentally confirmed that the multi-level prediction of FPN\(^{11}\) had a better performance based on this approach. This means that the acceptable object regression sizes for feature maps \( P_3, P_4, P_5, P_6, \) and \( P_7 \) are within the ranges of \([0, 38]\), \([38, 75]\), \([75, 150]\), \([150, 300]\), and \([300, \infty]\), respectively. This ensures that objects at different scales always lie in the acceptable level of each feature level during regression, thus reducing the number of fuzzy samples to a certain extent and helping to improve regression accuracy.

### 3.2 Aspect Ratio Priority

Recently, anchor-free object detectors have attracted considerable attention because of being easy to operate and having superior performance. The anchor-free object detection algorithm minimizes the amount of calculation by avoiding complex calculations on anchor boxes. Furthermore, anchor-free object detectors can already achieve detection accuracy that is similar or even superior to traditional object detectors based on anchor boxes, under the equivalent testing conditions.\(^{10}\) Slender objects with large aspect ratios may be better indicated by anchor-free object detectors, which would improve slender object detection accuracy.

A slender object with a large aspect ratio decays faster on the long side and slower on the short side during regression, which results in inconsistent regression decay rates between the long and short sides. The larger the aspect ratio of the slender object is, the more obvious this difference in decay rates is, leading to inaccurate positioning of the slender object during regression, which ultimately reduces the regression accuracy of the slender object. This study proposes a new solution to the problem. Our study designs the aspect ratio priority factor to solve the problem of inconsistent decay rates between the long and short sides of slender objects during regression. The aspect ratio priority factor is represented by Eq. (2)

\[
\Delta_{\text{priority}} = \left( \frac{\min(l, r)}{\max(l, r)} \times \frac{\min(t, b)}{\max(t, b)} \times (\text{angle}^2) \right)^{\tau}.
\]
In Eq. (2), $\tau$ is the inverse of the aspect ratio of the object; it is represented by Eq. (3)

$$\tau = \min(w, h)/\max(w, h).$$  \hspace{1cm} (3)

In Eq. (3), $w$ and $h$ are the width and height of the ground-truth box, respectively.

In the SRPAR’s loss function, the aspect ratio priority factor, which is designed to be placed in the regression loss as a penalty weight during long-edge regression [this is reflected in Eq. (5)], is used to guide the regression of the slender object. The aspect ratio priority factor restrains the regression rate of the long side of the object during regression, thus reducing the discrepancy between the decay rates of the long and short sides of the object and ultimately improving the regression accuracy of the slender object.

### 3.3 Loss Function

Slender objects detected by the proposed SRPAR involve classification and localization in the prediction process. Thus, the loss function can be divided into two components: classification loss and regression loss.

The $L_{\text{cls}}$ is the classification loss, which is defined as shown in Eq. (4).

$$L_{\text{cls}} = \begin{cases} -(1 - p(x, y))^\beta \log(p(x, y)) & \beta = 1 \\ -p(x, y)^\beta \log(1 - p(x, y)) & \beta = 0 \end{cases}.$$  \hspace{1cm} (4)

The classification loss $L_{\text{cls}}$ is the focal loss.\(^{21}\) In Eq. (4), $p(x, y)$ indicates the prediction point; $\beta$ is the indicator function, which is 1 if the point $(x, y)$ falls into the corresponding real object box and 0 otherwise; $\xi$, which is larger than 0, is the adjustable factor.

When the object is being located, if the five variables $t, b, l, r, \theta'$ are regressed together, due to the angular periodicity, there is a tendency for angular loss discontinuities to occur at the boundary during regression. Therefore, $t, b, l, r$ and the predicted angle $\theta'$ are regressed separately.

For ResNet networks, $L2$ loss converges much faster than $L1$ loss,\(^{15}\) and hence, $L2$ loss is chosen as the loss function for $t, b, l, r$ during regression. In addition, an aspect ratio priority factor is designed as a penalty weight for long-edge regression. It helps to improve the accuracy of the regression for slender objects, as confirmed in the ablation study in Sec. 4.1.

The $L_{\text{reg}}$ is chosen as the loss of the four quantities $t, b, l, r$ during regression, which is defined as shown in Eq. (5)

$$L_{\text{reg}}((t, b, l, r), (w, h)) = \begin{cases} \Delta_{\text{priority}} \times (w - l - r)^2 + (h - t - b)^2 & \text{if } w > h \\ (w - l - r)^2 + \Delta_{\text{priority}} \times (h - t - b)^2 & \text{if } w < h \end{cases}.$$  \hspace{1cm} (5)

In Eq. (5), $\Delta_{\text{priority}}$ is the defined aspect ratio priority factor; $t, b, l, r$ are the distances between the prediction point and the bounding box’s top, lower, left, and right edges, respectively; and $w$ and $h$ are the width and height of the ground-truth box, respectively.

The $L_{\text{mr}}$ is chosen as the loss for the angular regression, which is defined in Eq. (6)

$$L_{\text{mr}}(\theta', \theta) = \min \left\{ \frac{|x_c - x|}{w} + \frac{|y_c - y|}{h} + |w - l - r| + |h - t - b| + |\theta - \theta'| \right\}.$$  \hspace{1cm} (6)

The angular regression loss $L_{\text{mr}}$ is $L_{\text{mr}}$ loss.\(^{22}\) In Eq. (6), $\theta'$ indicates the prediction of object’s angle, $\theta$ indicates the label of object’s angle, and $(x_c, y_c)$ indicates the center point of the ground-truth box.

The loss function is defined in Eq. (7)

$$L = \frac{\lambda_1}{N_{\text{pos}}} \sum_{x, y} L_{\text{cls}} + \frac{\lambda_2}{N_{\text{pos}}} \sum_{x, y} L_{\text{reg}}((t, b, l, r), (w, h)) + \frac{\lambda_3}{N_{\text{pos}}} \sum_{x, y} L_{\text{mr}}(\theta', \theta).$$  \hspace{1cm} (7)

In Eq. (7), $N_{\text{pos}}$ denotes the number of positive samples. A positive sample occurs when the point $p(x, y)$ falls into the corresponding ground-truth box. The hyper-parameters $\lambda_1, \lambda_2, \lambda_3$ control the trade-off and are all set to 1 by default.
4 Experiments

All experiments were conducted in the Ubuntu 18.04 operating system, with the Pytorch deep-learning framework and the Quadro p4000 graphics processing unit. This experiment was carried out on the DOTA dataset and the self-made COCO-Stick dataset, and ResNet-101\textsuperscript{22} was employed as the training network.

On the DOTA dataset, the network utilized in the experiments was trained with a method for stochastic optimization (Adam) for 6 k iterations, with the initial learning rate being $1.25 \times 10^{-3}$ and a mini-batch of 10 images. The learning rate was reduced by a factor of 10 at iteration 3.6 and 4.8 k, respectively. The weight decay was set as $1 \times 10^{-4}$. On the self-made COCO-Stick dataset, the network was trained with Adam for 3 k iterations with the initial learning rate being $1.25 \times 10^{-4}$. The learning rate was reduced by a factor of 10 at iteration 1.8 and 2.4 k, respectively, and the other training parameters were kept the same as before.

4.1 Ablation Study

To address the problem of fuzzy samples arising from overlapping object prediction boxes, FPN\textsuperscript{11} multi-level prediction was deployed to limit the range of bounding boxes and corresponding angle regressions at each level. To verify the effectiveness of our method, this experiment compared the detection results of the original network and the network after adopting FPN\textsuperscript{11} multi-level prediction on the DOTA\textsuperscript{12} dataset, which contains 15 categories. The model evaluation indicator is the fuzzy sample rate, which was quoted from Tian et al.’s work.\textsuperscript{2}

In Table 1, Dataset denotes the dataset employed, and With/FPN denotes FPN\textsuperscript{11} multi-level prediction. The experimental results are shown in Table 1. When FPN\textsuperscript{11} multi-level prediction was adopted in the original network, the fuzzy sample ratio for all objects in the DOTA\textsuperscript{12} dataset decreased from 13% to 3.3%. The experimental results show that the number of fuzzy samples is significantly reduced, after FPN\textsuperscript{11} multilevel prediction was utilized to limit the range of bounding boxes and corresponding angle regressions at each level.

To solve the problem of the long and short edges of slender objects not decaying at the same rate during regression, an aspect ratio priority factor was designed to guide the regression of slender rotating objects. To verify the effectiveness of the introduced aspect ratio factor, this experiment compared the detection results of the original network and the network after the addition of the aspect ratio factor on the DOTA\textsuperscript{12} dataset. The evaluation metrics of the model are the mean average precision (mAP) and recall.

The experimental results are shown in Table 2. Dataset denotes the dataset used, and With/priority denotes the aspect ratio priority factor introduced into the network. As can be seen from Table 2, when the aspect ratio factor was introduced into the original network, the mAP and recall of all objects in the DOTA\textsuperscript{12} dataset increased by 3.1% and 14%, respectively. The mAP and recall values of the objects in the network increased, after the introduction of the aspect ratio factor. The experimental results show that the introduction of an aspect ratio factor in the network can improve the regression accuracy of slender objects.

| Dataset  | With/FPN | Fuzzy sample (%) |
|----------|----------|------------------|
| DOTA     | —        | 13.0             |
|          | √        | 3.3              |

| Dataset  | With/priority | mAP (%) | Recall (%) |
|----------|---------------|---------|------------|
| DOTA     | —             | 64.1    | 70.7       |
|          | √             | 67.2    | 84.7       |
4.2 DOTA

The detection objects of the proposed model are slender rotating objects with aspect ratios of no less than 3:1. Ships with large aspect ratios met the detection needs in the DOTA dataset, and hence, the ships were detected in the DOTA dataset. Ship objects were extracted from 1441 raw images of the training set in the DOTA dataset and cropped to pieces of size $600 \times 400$, yielding 12,644 images containing only ships. The txt label file corresponding to these images was then transformed into the xml label file corresponding to the five-parameter method. Furthermore, these ships have different aspect ratios of 5:1 to 10:1, 3:1 to 5:1, and 1:1 to 3:1, and the corresponding number of the ships in each range accounts for 2.27%, 24.36%, and 73.37% of the total number of ships in the DOTA training set, respectively.

Experimental results on the DOTA dataset are shown in Table 3. In this paper, the model evaluation indicator is AP, the value of which is the area between the curves made with accuracy and recall as variables and the coordinate axes. Moreover, $AP_{sr}$ refers to the AP with aspect ratios between 1:1 and 3:1. $AP_{mr}$ and $AP_{lr}$ refer to the AP with aspect ratios that are within 3:1 to 5:1 and 5:1 to 10:1, respectively. SCRDet, RSDet, RRPN, and R3Det were selected as comparative algorithms. The comparison results of different algorithms are shown in Table 3.

The data in Table 3 show that the AP of the proposed SRPAR outperforms the other four detection algorithms for typical slender objects with large aspect ratios, such as ships. Furthermore, the $AP_{sr}$, $AP_{mr}$, and $AP_{lr}$ were also tested for ships with different aspect ratios on the four algorithmic models of SCRDet, RSDet, RRPN, and R3Det, and the results are shown in Table 3. The $AP_{sr}$ of the proposed SRPAR for ship detection is the second highest. In addition, the proposed SRPAR has the highest $AP_{mr}$ and $AP_{lr}$ for ship detection of all of the compared algorithms. The $AP_{lr}$ of ship detection for the proposed SRPAR and the other four algorithms are low due to the large variability in the decay rate of the long and short sides of ships with aspect ratios ranging from 5:1 to 10:1, which leads to inaccurate positioning of the ship objects in the regression and ultimately reduces the regression accuracy of the ship objects. These comparison results show that the proposed SRPAR effectively detects slender rotating objects with large aspect ratios in remote sensing images.

Compared with the number of ships with other aspect ratios, the number of ships in the DOTA dataset with an aspect ratio of 1:1 to 3:1 is higher. For objects with different aspect ratios, there is a sample imbalance problem, which should be studied further in the future.

Ship detection results of proposed SRPAR on the DOTA dataset are shown in Fig. 5. When the aspect ratio of the ship object is large, the trained model can still detect it, reflecting that the method proposed is effective in detecting slender rotating objects.

4.3 COCO-Stick

The model is designed to meet the detection needs of specific types of objects, such as sticks, which are slender objects with an aspect ratio of at least 3:1. Although experimental results on the DOTA dataset have shown that the proposed model is good at detecting slender rotating
objects, the model had not yet done any detection of stick objects. Therefore, a stick dataset that meets requirements of SRPAR’s detection needed to be created; in the dataset, the aspect ratios of objects should be mainly between 3:1 to 5:1 and 5:1 to 10:1 and the aspect ratios of objects between 1:1 and 3:1 should only account for a very small part of the dataset. At present, there is a lack of such datasets at home and abroad. For this purpose, 500 rotating stick images with a resolution of 800 × 800 to 1200 × 1200 were collected and cropped to pieces of size 600 × 400, yielding 2859 images. The five-parameter method was utilized to label them, resulting in a new dataset called Stick.

The Stick dataset has several characteristics. The objects have large aspect ratios. Objects in the Stick dataset have different aspect ratios of 5:1 to 10:1, 3:1 to 5:1, and 1:1 to 3:1, as shown in Table 4.

The baseball bat subset of the COCO1 dataset contains some baseball bats that meet the detection requirements of the proposed model for slender rotating stick objects. Baseball bat objects were extracted from raw images in the COCO1 dataset and cropped to pieces of size 600 × 400, yielding 2756 images, which were subsequently labeled using the five-parameter method. The Stick dataset was added to the baseball bat subset to form a new self-made COCO-Stick dataset. The self-made COCO-Stick dataset has 5615 images of stick-like objects.

The self-made COCO-Stick dataset has several characteristics as well. The objects have large aspect ratios. Furthermore, objects in the self-made COCO-Stick dataset have different aspect ratios of 5:1 to 10:1, 3:1 to 5:1, and 1:1 to 3:1, as shown in Table 5. Objects with an aspect ratio of at least 3:1 account for 79.69% in the dataset, the majority of the dataset, and satisfy the SRPAR’s detection requirements.

Table 4 Percentage distribution of sticks in the Stick dataset.

| Aspect ratio   | Number | Percentage (%) |
|----------------|--------|----------------|
| 1:1 to 3:1     | 107    | 2.67           |
| 3:1 to 5:1     | 2451   | 61.24          |
| 5:1 to 10:1    | 1444   | 36.09          |
| Total          | 4002   | 100            |
SCRDet, RSDet, RRPN, and R3Det were used as comparison algorithms for the proposed model, and the results of the different algorithms on the self-made COCO-Stick dataset are shown in Table 6. The AP, APsr, and APmr values of our model for stick detection are higher than the corresponding values of the other four algorithms. The APlr value for stick detection is the highest among these comparative algorithms, and this value best reflects the performance of

| Aspect ratio     | Number | Percentage (%) |
|------------------|--------|----------------|
| 1:1 to 3:1       | 1426   | 20.31          |
| 3:1 to 5:1       | 3056   | 45.53          |
| 5:1 to 10:1      | 2538   | 36.16          |
| Total            | 7020   | 100            |

| Method     | AP (%) | APsr (%) | APmr (%) | APlr (%) |
|------------|--------|----------|----------|----------|
| SCRDet16   | 57.2   | 78.1     | 68.7     | 33.0     |
| RSDet22    | 55.8   | 76.4     | 67.0     | 31.1     |
| RRPN23     | 50.0   | 72.9     | 60.2     | 26.3     |
| R3Det10    | 62.9   | 81.6     | 74.5     | 39.9     |
| Proposed   | 64.1   | 83.9     | 75.0     | 41.2     |

Note: Bold values also represent the optimal values of the five algorithms compared in the table for the detection of targets with different aspect ratios.

Table 5 Percentage distribution of sticks in the self-made COCO-Stick dataset.

Fig. 6 Detection results on the self-made COCO-Stick dataset.
the model in detecting objects with large aspect ratios. The experimental results show that the proposed model has some advantages on the self-made COCO-Stick dataset.

Detection results of the proposed SRPAR on the self-made COCO-Stick dataset are shown in Fig. 6. The top two images in Fig. 6 are detection images from the baseball bat set, and the bottom two images in Fig. 6 are detection images from the Stick dataset. When detecting stick-like objects having large aspect ratios, SRPAR detects them well, indicating that proposed the SRPAR performs better on the self-made COCO-Stick dataset.

5 Conclusion

When it comes to detecting specific types of slender objects, the anchor-free object detector faces some problems including large aspect ratio, overlapping of bounding boxes, and the lack of appropriate datasets. A new object detection method is proposed to address these problems. First, the aspect ratio priority factor was designed to solve the problem of inconsistent decay rates between the long and short sides of slender objects during regression. Second, FPN multi-level prediction was utilized to limit the range of bounding boxes and corresponding angle regressions at each level, the number of fuzzy samples was significantly reduced. Finally, the self-made COCO-Stick dataset was created to further evaluate the SRPAR’s performance in detecting slender rotating objects. To more intuitively verify SRPAR’s performance in detecting slender rotating objects with varying aspect ratios, these objects were classified into three categories, each of which has a different aspect ratio of 1:1 to 3:1, 3:1 to 5:1, and 5:1 to 10:1. Experiments were conducted with the DOTA and self-made COCO-Stick datasets. Experimental results on the DOTA dataset reveal that SRPAR outperforms several advanced object detection algorithms currently available in terms of detection accuracy. The performance of the method was then further validated on the self-made COCO-Stick dataset. However, imbalance in aspect ratios exists in the object training samples in the DOTA dataset employed by SRPAR. Further studies need to be conducted to improve SRPAR’s performance in detecting slender rotating objects.

References

1. T. Y. Lin et al., Microsoft COCO: Common Objects in Context, Springer International Publishing (2014).
2. Z. Tian et al., “FCOS: a simple and strong anchor-free object detector,” IEEE Trans. Pattern Anal. Mach. Intell. PP (99), 1–1 (2020).
3. H. Law and J. Deng, “CornerNet: detecting objects as paired keypoints,” Int. J. Comput. Vis. 128(3), 642–656 (2020).
4. X. Zhou, D. Wang, and P. Krhebnbh, “CenterNet: objects as points,” Computer Science, arXiv:1904.07850 (2019).
5. J. Ding et al., “Learning RoI transformer for detecting oriented objects in aerial images,” arXiv:1812.00155 (2018).
6. P. Shi et al., “Remote sensing image object detection based on angle classification,” IEEE Access 9, 118696–118707 (2021).
7. H. Zhang et al., “Cascade RetinaNet: maintaining consistency for single-stage object detection,” arXiv:1907.06881 (2019).
8. L. Lei et al., “Aircraft detection of remote sensing images based on faster R-CNN and Yolov3,” in Int. Conf. Culture-oriented Sci. & Technol., pp. 166–170 (2020).
9. Z. Zhang et al., “Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks,” IEEE Geosci. Remote Sens. Lett. 15(11), 1745–1749 (2018).
10. X. Yang et al., “R3Det: refined single-stage detector with feature refinement for rotating object(C),” in Thirty-Five AAAI Conf. Artif. Intell. (2021).
11. T. Y. Lin et al., “Feature pyramid networks for object detection,” in IEEE Conf. Comput. Vis. and Pattern Recognit., pp. 936–944 (2017).
12. G. Xia et al., “DOTA: a large-scale dataset for object detection in aerial images,” in IEEE/CVF Conf. Comput. Vis. and Pattern Recognit., pp. 3974–3983 (2018).
13. X. Zhou, J. Zhuo, and P. Krhenbhl, “ExtremeNet: bottom-up object detection by grouping extreme and center points,” in IEEE/CVF Conf. Comput. Vis. and Pattern Recognit. (2019).
14. Z. Yang et al., “RepPoints: point set representation for object detection,” in IEEE/CVF Int. Conf. Comput. Vis., pp. 9656–9665 (2019).
15. Y. Xu et al., “Gliding vertex on the horizontal bounding box for multi-oriented object detection,” in IEEE Trans. Pattern Anal. Mach. Intell. 43(4), 1452–1459 (2021).
16. X. Yang et al., “SCRDet: towards more robust detection for small, cluttered and rotated objects,” in IEEE/CVF Int. Conf. Comput. Vis., pp. 8231–8240 (2019).
17. X. Yang and J. Yan, Arbitrary-Oriented Object Detection with Circular Smooth Label, Springer International Publishing (2020).
18. A. Ghahremani, E. Bondarev, and P. With, “Toward robust multitype and orientation detection of vessels in maritime surveillance,” J. Electron. Imaging 29(3), 033015 (2020).
19. Y. K. Li, X. J. Wu, and J. Kittler, “L1-2DPCANet: a deep learning network for face recognition,” J. Electron. Imaging 28(2), 023016 (2019).
20. Z. Wang, S. Wan, and X. Ma, “Remote sensing image dense target detection based on rotating frame,” J. Phys. Conf. Ser. 2006(1), 012049 (2021).
21. T. Y. Lin et al., “Focal loss for dense object detection,” IEEE Trans. Pattern Anal. Mach. Intell. 42(2), 318–327 (2017).
22. W. Qian et al., “Learning modulated loss for rotated object detection,” Proc. AAAI Conf. Artif. Intell. 35(3), 2458–2466 (2019).
23. J. Ma et al., “Arbitrary-oriented scene text detection via rotation proposals,” in IEEE Trans. Multimedia 20(11), 3111–3122 (2018).

Hong-Gang Xie received his PhD in engineering from Wuhan University and has hosted a Natural Science Foundation of Hubei Province and participated in many National Natural Science Foundation of China. He is also responsible for a number of projects entrusted by enterprises and has published 15 SCI/EI papers in total. He is mainly engaged in video and image signal processing and target detection research.

Ming Yang is pursuing his master’s degree at Hubei University of Technology. He is mainly engaged in image signal processing and target detection research.

Bo-Lun Yan is pursuing his master’s degree at Hubei University of Technology. His research is in target detection and image processing.

Kai-Yuan Hou is pursuing his master’s degree at Hubei University of Technology. His research is in target detection.

Di Jiang is pursuing his master’s degree at Hubei University of Technology. His research is in image processing.