Arbitrary-Oriented Ship Detection through Center-Head Point Extraction

Feng Zhang, Xueying Wang, Shilin Zhou, Yingqian Wang, Yi Hou

Abstract—Ship detection in remote sensing images plays a crucial role in various applications and has drawn increasing attention in recent years. However, existing arbitrary-oriented ship detection methods are generally developed on a set of predefined rotated anchor boxes. These predefined boxes not only lead to inaccurate angle predictions but also introduce extra hyper-parameters and high computational cost. Moreover, the prior knowledge of ship size has not been fully exploited by existing methods, which hinders the improvement of their detection accuracy. Aiming at solving the above issues, in this paper, we propose a center-head point extraction based detector (named CHPDet) to achieve arbitrary-oriented ship detection in remote sensing images. Our CHPDet formulates arbitrary-oriented ships as rotated boxes with head points which are used to determine the direction. And rotated Gaussian kernel is used to map the annotations into target heatmaps. Keypoint estimation is performed to find the center of ships. Then, the size and head point of the ships are regressed. The orientation-invariant model (OIM) is also used to produce orientation-invariant feature maps. Finally, we use the target size as prior to finetune the results. Moreover, we introduce a new dataset for multi-class arbitrary-oriented ship detection in remote sensing images at a fixed ground sample distance (GSD) which is named FGSD2021. Experimental results on FGSD2021 and two other widely used data sets, i.e., HRSC2016, and UCAS-AOD demonstrate that our CHPDet achieves state-of-the-art performance and can well distinguish between bow and stern. Code and FGSD2021 dataset are available at https://github.com/f2020114/CHPDet.

Index Terms—Arbitrary-oriented ship detection, Remote sensing images, Keypoint estimation, Deep convolutional neural networks

I. INTRODUCTION

SHIP detection from high-resolution optical remote sensing images is widely applied in various tasks such as illegal smuggling, port management, and target reconnaissance. Recently, ship detection has received increasing attention and was widely investigated in the past decades [1]–[4]. However, ship detection in remote sensing images is a highly challenging task due to the arbitrary orientations, densely-parking scenarios, and complex backgrounds [5]–[7]. To handle the multi-orientation issue, existing methods generally use a series of predefined anchors [8], which has the following shortcomings.

Inaccurate angle regression. Fig. 1(a)-(d) illustrate four different representations of an arbitrary-oriented ship. Since ships in remote sensing images are generally in strips, the intersection over union (IoU) score is very sensitive to the angle of bounding boxes. As shown in Fig. 1(e), the ground truth box is the bounding box of a ship with an aspect ratio of 10:1. The red rotated box is generated by rotating the ground truth box with a small angle of 5°. It can be observed that such a small angle variation reduces the IoU between these two boxes to 0.63. Therefore, the anchor-based detectors which define the positive and negative anchors by IoU score usually suffer from an imbalance issue, and thus resulting in detection performance degeneration [9]. Moreover, the angle of the ship is a periodic function, and it is discontinuous at the boundary (0° or 180°), as shown in Fig. 1(f). This discontinuity will also cause performance degeneration.

Excessive hyper-parameters and high computational cost. Existing methods generally use oriented bounding boxes as anchors to handle rotated objects and thus introduce excessive hyper-parameters such as box sizes, aspect ratios, and orientation angles. Note that, these hyper-parameters have to be manually tuned for novel scenarios, which limits the generalization capability of these methods. Predefined anchor-
based methods usually require a large number of anchor boxes. For example, in R²PN [10], 6 different orientations were used in rotated anchor boxes, and there are a total of 24 anchors at each pixel on its feature maps. A large number of anchor boxes introduce excessive computational cost when calculating IoU scores and executing the non-maximum suppression (NMS) algorithm.

**Under-exploitation of prior information of ships.**

Most previous ship detectors adopted the commonly-used rotation detection algorithms in the area of remote sensing and scene text detection, while overlook the unique characteristics of ships in remote sensing images. That is, the position of the bow is relatively obvious and a certain category of the ship in remote sensing images has a relatively fixed size range by normalizing the ground sample distance (GSD) of images. The size of the ship and the position of the ship’s head are important clues for detection. However, these prior informations have been under-exploited by previous ship detection algorithms. These methods only model the ships as rotated rectangles to regress the parameters and do not use the obvious bow point to determine the direction of the ship. Due to the limitation of the effective receptive field of the network, the appearance information near the central point is mainly used in target classification. Size regression and target classification are obtained independently by two parallel branches. Therefore, the size of the target can not effectively assist target classification.

Motivated by the anchor-free detectors CenterNet [11] in natural scenes, in this paper, we propose a one-stage, anchor-free and NMS-free method for arbitrary-oriented ship detection in remote sensing images. We formulate ships as rotated boxes with a head point representing the direction. Specifically, orientation-invariant feature maps are first produced by an orientation-invariant model. Afterward, the peaks of the center feature map are selected as center points. Then, the offset, object sizes, and head positions are regressed on the corresponding feature maps at each center point. Finally, target size is used to adjust the classification score. The architecture of our CHPDet is shown in Fig. 2.

The major contributions of this paper are summarized as follows.

- We develop a one-stage, anchor-free ship detector CH-
A. Horizontal Object Detection

In recent years, deep convolutional neural networks (DCNN) have been developed as a powerful tool for feature representation learning [12], [13], and have achieved significant improvements in horizontal object detection [14]. Existing object detection methods generally represent objects as horizontal boxes, as shown in Fig. 1(a). According to different detection paradigms, deep learning-based object detection methods can be roughly divided into two-stage detectors, single-stage detectors, and multi-stage detectors. Two-stage detectors (e.g., RCNN [15], Fast-RCNN [16], Faster-RCNN [17], Mask-RCNN [18], R-FCN [19]) used a pre-processing approach to generate object proposals, and extract features from the generated proposals to predict the category. In contrast, one-stage detectors (e.g., YOLO [20], [21], SSD [22], RetinaNet [23]) do not have the pre-processing step and directly performed categorical prediction on the feature maps. Multi-stage detectors (e.g, cascade RCNN [24], HTC [25]) performed multiple classifications and regressions, resulting in notable accuracy improvements. In summary, two-stage and multi-stage detectors generally achieve better performance, but one-stage detectors are usually more time-efficient.

Compared to the above-mentioned anchor-based methods, anchor-free methods [26] [11] can avoid the requirement of anchors and have become a new research focus in recent years. For example, CornerNet [26] detected objects at each position of the feature map using the top-left and bottom-right corner points. CenterNet [11] modeled an object as a center point and performed keypoint estimation to find center points and regressed the object size. FCOS [27] predicted four distances, a center score, and a classification score at each position of the feature map to detect objects. The above-mentioned approaches achieve significant improvement in general object detection tasks. However, these detectors can only generate horizontal bounding boxes, which limits their applicability.

B. Arbitrary-oriented object detection

Arbitrary-oriented detectors are widely used in remote sensing and scene text images. Most of these detectors used rotated bounding boxes or quadrangles to represent multi-oriented objects, as shown in Fig. 1(b) (c). In R-RPN [28], rotated region proposal network was proposed to improve the quality of the region proposals. In R²CNN [29], a horizontal region of interest (RoI) was generated to simultaneously predict the horizontal and rotated boxes. RoI-Trans [30] transformed a horizontal RoI into a rotated RoI (RRoI). In SCRDet [31] and RSDet [9], novel losses were employed to address the boundary problem for oriented bounding boxes. In R³Det [32], a refined single-stage rotated detector was proposed for the feature misalignment problem. In CSL [33] and DCL [34], angle regression was converted into a classification task to handle the boundary problem. In S²A-Net [35], a fully convolutional layer was proposed to align features to achieve better performance. The aforementioned methods need a set of anchor boxes for classification and regression. These anchors introduce excessive hyper-parameters which limit the generalization capability and introduce an excessive computational cost. At present, several anchor-free arbitrary-oriented detectors (e.g., O²D-Net [36] and X-LineNet [37]) are proposed to detect oriented objects by predicting a pair of intersecting lines. However, The features used in these methods are not rotation-invariant and the performance still lags behind that of the anchor-base detectors.
C. Ship detection in remote sensing images

Different from other objects in remote sensing images, ships are in strips with a large aspect ratio. Generally, the outline of the ship is an approximate pentagon with two parallel long sides, and the position of the bow is relatively obvious. Consequently, a certain category of the ship in remote sensing images has a relatively fixed size range by normalizing the GSD of images.

Traditional ship detectors generally used a coarse-to-fine framework with two stages including ship candidate generation and false alarm elimination. For example, Shi et al. [38] first generated ship candidates by considering ships as anomalies and then discriminated these candidates using the AdaBoost approach [39]. Yang et al. [40] proposed a saliency-based method to generate candidate regions, and used a support vector machine (SVM) to further classify these candidates. Liu et al. [41], [42] introduced an RRoI pooling layer to extract features of rotated regions. In R²PN [10], a rotated region proposal network was proposed to generate arbitrary-proposals with ship orientation angle information. The above detectors are also based on a set of anchors and cannot fully exploit the prior information of ships.

III. PROPOSED METHOD

In this section, the architecture of CHDPDet is introduced in detail. Our method consists of 5 modules including an arbitrary-oriented ship representation module, a rotated Gaussian kernel module, a head point estimation module, an orientation-invariant module and a probability refinement module. All ships are represented by rotated boxes with a head point. We first detect centers of ships by extracting the peaks in heatmaps which are generated by rotated Gaussian kernels. Then, we locate the head points by two steps (directly regress from image features at the center location, and estimate head points from head heatmaps). We also extract orientation-invariant feature maps by the orientation-invariant model (OIM) to increased consistency between targets and corresponding features. Finally, we refine the detection results based on the prior information. The overall framework of CHDPDet is shown in Fig. 2.

A. Arbitrary-oriented ship representation

As shown in Fig. 1, the widely-used horizontal bounding boxes cannot be directly applied to the arbitrary-oriented ship detection task since excessive redundant background area is included. Moreover, since the arbitrary-oriented ships generally have a large aspect ratio and park densely, the NMS algorithm using a horizontal bounding box tends to produce missing detection. To this end, many methods represent ships as rotated bounding boxes, and these boxes are parameterized with 5 tuples \((x, y, w, h, \theta)\), where \((x, y)\) is the coordinate of the center of the rotated bounding box, \(w\) and \(h\) are the width and length of the ship, respectively. The angle \(\theta \in [0\degree, 180\degree)\) is the orientation of the long side with respect to the \(y\)-axis. This representation can result in regression inconsistency issue near the boundary case. Recently, some detectors represent objects by four clockwise vertices, which are parameterized by 8 tuples \((x_a, y_a, x_b, y_b, x_c, y_c, x_d, y_d)\). This representation can also introduce regression inconsistency due to the order of the four corner points.

To avoid the aforementioned inconsistency problem, we represent ships as two points and their corresponding size, which are parameterized by 6 tuples \((x_c, y_c, w, h, x_h, y_h)\). \((x_c, y_c)\) is the coordinate of the center of the rotated bounding box, \(w\) and \(h\) are the width and length of the ship, \((x_h, y_h)\) is the coordinate of the head point of the ship. The direction of the ship is determined by connecting the center and the bow. This representation of ships converts discontinuous angle regression to continuous keypoint estimation. This representation also extends the range of angle representation to \([0\degree, 360\degree)\) and enables the network to distinguish between bow and stern.

B. Rotated Gaussian Kernel

Our detectors use center heatmaps to classify and locate ships simultaneously. To adapt to the characteristics of the rotated target, we use the rotated Gaussian kernel (see Fig. 4) to map the annotations to target heatmaps in the training stage.

Specifically, given \(m^{th}\) annotated box \((x, y, w, h, \theta)\) belongs to \(c^m\) category, it is linearly mapped to the feature map scale. Then, 2D Gaussian distribution \(\mathcal{N}(m, \Sigma)\) is adopted to produce target heatmap \(C \in \mathbb{R}^{w \times h \times C}\). Here, \(m = (x, y)\) represents the probability density function of the rotated Gaussian distribution, and the probability density function can be calculated according to covariance matrix Eq. 1.
\[ \Sigma^{1/2} = \text{RSR}^T \]
\[ = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} \sigma_x & 0 \\ 0 & \sigma_y \end{pmatrix} \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix} \]
\[ = \begin{pmatrix} \sigma_x \cos^2 \theta + \sigma_y \sin^2 \theta & (\sigma_x - \sigma_y) \cos \theta \sin \theta \\ (\sigma_x - \sigma_y) \cos \theta \sin \theta & \sigma_x \cos^2 \theta + \sigma_y \sin^2 \theta \end{pmatrix}, \]

where \( s \) is a downsampling stride and \( \sigma_x = \frac{\sigma_{xw}}{\sqrt{w \times h}}, \sigma_y = \frac{\sigma_{yw}}{\sqrt{w \times h}} \), \( \sigma_p \) is a size-adaptive standard deviation [11], \( \alpha \) is set to 1.2 in our implementation, and its not carefully selected. Fig. 4 is a schematic diagram of mapping a rotated bounding box to a rotated Gaussian distribution.

If two Gaussian kernels belong to the same category with an overlap region, we take the maximum value at each pixel of the feature map. \( \tilde{C} \in \mathbb{R}^{\frac{w}{s} \times \frac{h}{s} \times C} \) is a prediction on feature maps produced by the backbones. Fig. 5(a) shows a visualization of the center heatmaps.

We extract locations with values larger or equal to their 8-connected neighbors as detected center points. The value of the peak point is set as a confidence measurement, and the coordinates in the feature map are used as an index to get other attributes. Therefore, the accurate location of the center point on the feature map is the key part of the whole detection.

The peaks of the Gaussian kernel, also the centers of rotated box, are treated as the positive samples while any other pixels are treated as the negative samples, which may cause a huge imbalance between positive and negative samples. To handle the imbalance issue, we use the variant focal loss as [11], [23]:

\[ \mathcal{L}_c = \frac{-1}{N} \sum_{xy} \begin{cases} (1 - C_{xy})^\gamma \log(C_{xy}) & \text{if } C_{xy} = 1 \\ (1 - C_{xy})^\beta \log(1 - C_{xy}) & \text{otherwise} \end{cases} \]

where \( \gamma \) and \( \beta \) are the hyper-parameters of the focal loss, \( N \) is the number of objects in image \( I \) which is used to normalize all positive focal loss instances to 1. We set \( \gamma = 2 \) and \( \beta = 4 \) in our experiments empirically as in [26].

To reduce the quantization error by the output stride, we produce local offset feature maps \( \mathbf{O} \in \mathbb{R}^{\frac{w}{s} \times \frac{h}{s} \times 2} \). Suppose that \( c = \{(\hat{x}_k, \hat{y}_k)\}^{n}_{k=1} \) is the set of detected center points, center point location is given by an integer coordinates \( c_k = (\hat{x}_k, \hat{y}_k) \) on feature map \( \mathbf{C} \). For each predicted center point \( c_k \), let the value on the offset feature maps \( f_k = (\delta \hat{x}_k, \delta \hat{y}_k) \) be the offset of center point \( c_k \).

The final center point location of \( \hat{c}_k \) is a prediction on feature maps \( \mathbf{R} \in \mathbb{R}^{\frac{w}{2s} \times \frac{h}{2s} \times 2} \) at each predicted center point \( c_k \in \text{center} \). The regression-based head point is \( \{(\hat{x}_k + \Delta \hat{x}_k, \hat{y}_k + \Delta \hat{y}_k)\}^{n}_{k=1} \), where \( (\Delta \hat{x}_k, \Delta \hat{y}_k) \) is the head point regression, and an L1 loss is used to optimized head regression feature maps.

\[ \mathcal{L}_{hr} = \frac{1}{N} \sum_{k=1}^{N} | \mathbf{R}_{c_k} - h_k |. \]

1) Regression-based head point estimation: Let head\(_k\) = \((h_x, h_y)\) be the \( k^{th} \) head point, we directly regress to the offsets \((\Delta \hat{x}_k, \Delta \hat{y}_k)\) on feature map \( \mathbf{R} \) \( \in \mathbb{R}^{\frac{w}{2} \times \frac{h}{2} \times 2} \) at each predicted center point \( c_k \in \text{center} \). The regression-based head point is \( \{(\hat{x}_k + \Delta \hat{x}_k, \hat{y}_k + \Delta \hat{y}_k)\}^{n}_{k=1} \), where \( (\Delta \hat{x}_k, \Delta \hat{y}_k) \) is the head point regression, and an L1 loss is used to optimized head regression feature maps.

\[ \mathcal{L}_{hr} = \frac{1}{N} \sum_{k=1}^{N} | \mathbf{R}_{c_k} - h_k |. \]

2) Bottom-up head point estimation: We use standard bottom-up-multi-human pose estimation [43] to refine the head points. A target map \( \mathbf{H} \in \mathbb{R}^{\frac{w}{S} \times \frac{h}{S} \times 1} \) is computed as described in Section III-B. A low-resolution equation is head = \( \hat{\text{head}} \).

Head point heatmap \( \mathbf{E} \in \mathbb{R}^{\frac{w}{2} \times \frac{h}{2} \times 1} \) and local offset heatmap \( \mathbf{HO} \in \mathbb{R}^{\frac{w}{2} \times \frac{h}{2} \times 2} \) are head maps produced by the backbones. These two head maps are trained with variant focal loss and an L1 loss.

\[ \mathcal{L}_{he} = \frac{-1}{N} \sum_{xy} \begin{cases} (1 - \mathbf{E}_{xy})^\gamma \log(\mathbf{E}_{xy}) & \text{if } \mathbf{H}_{xy} = 1 \\ (1 - \mathbf{H}_{xy})^\beta \log(1 - \mathbf{E}_{xy}) & \text{otherwise} \end{cases} \]

\[ \mathcal{L}_{ho} = \frac{1}{N} \sum_{k=1}^{N} | \mathbf{HO}_{c_k} - \left( \frac{\text{head}_k}{S} - \text{head} \right) |. \]

The bottom-up head point estimation is the same as the center point detection. Note that, in center point detection, each category has a center points heat map, while in head point estimation, all categories share one head points heatmap. We extract all peak point locations head = \( \{\hat{c}_k\} \) with a confidence \( \mathbf{HO}_{xy} > 0.1 \) as a potential head points set, and refine the potential head point locations by adding the offset \((\xi_x, \xi_y)\). Fig. 5(b) visualizes the head points heatmap.

We introduce a set of weighted factors to balance the contribution of these parts, and set \( \lambda_c = 0.1, \lambda_s = 0.1, \lambda_{hr} = 1, \lambda_{he} = 1 \), and \( \lambda_{ho} = 1 \) in all our experiments. We set \( \lambda_c = 0.1 \) since the scale of the loss is ranged from 0 to the output size \( h/S \). The overall training loss is

\[ \mathcal{L} = \mathcal{L}_c + \lambda_c \mathcal{L}_o + \lambda_s \mathcal{L}_s + \lambda_{hr} \mathcal{L}_{hr} + \lambda_{he} \mathcal{L}_{he} + \lambda_{ho} \mathcal{L}_{ho}. \]

In the testing phase, we first extracted the center points on the output center heatmaps \( \mathbf{C} \) for each category. We used a \( 3 \times 3 \) max-pooling layer to get the peak points and selected the top 100 peaks as potential center points. Each center point location is represented as an integer coordinates \( \hat{c} = (\hat{x}, \hat{y}) \). Take out the offsets \((\delta_x, \delta_y)\), size \((w, h)\), and head points regression \((\Delta x, \Delta y)\) on the corresponding feature map at the location of center points. We also picked all head peak points on the output center heatmaps \( \mathbf{E} \) with a scores head\(_i \in (x, y)\), if \( \mathbf{E}_{xy} > 0.1 \), and then assigned each regressed location head\(_i = (\hat{x} + \Delta x, \hat{y} + \Delta y)\) to its closest detected keypoint arg min_{\text{head}} \( (1 - \text{head})^2 \) as the head point \((\hat{h}_x, \hat{h}_y)\), then we add the head point offset \((\xi_x, \xi_y)\) to refine the head point estimation. Finally, we get the rotated boxes \((\hat{x} + \delta_x, \hat{y} + \delta_y, w, h, \hat{h}_x + \xi_x, \hat{h}_y + \xi_y)\). We use
Since ORPooling is introduced to extract the maximum response value for all ARF, the target features of different orientations at this location are identical. Based on the rotation invariance feature, six kinds of feature maps are got by convolution layers respectively. Moreover, OIM only introduces one convolution layer with a small number of parameters, which has little effect on the speed of training and inferencing.

The rotation-invariant feature is very important for detecting arbitrary oriented objects, which enhances the consistency of the feature. Our detectors extract locations with local maximum as detected center points, so at the object center, the rotation-invariant feature of arbitrary oriented objects are identical, which increases the generalization ability of the network. Otherwise, more parameters are needed to encode the orientation information.

E. Refine probability according to size

By normalizing the GSD of remote sensing images, objects of the same size on the ground have the same size in all images. The size of the target is an important clue to identify the target because a certain type of target in remote sensing images usually has a relatively fixed size range. We propose an approach to adjust the confidence score of targets according to the prior knowledge of ship size. As shown in Fig. 5(d), suppose that the category of the detected box is $a$, the original confidence score is $s_a$, assume that the length of the detected ship obeys a normal distribution, the mean and standard deviation of the length of category $a$ are $L_a$, $\delta_a$. Then the probability of the target belonging to $a$ is $p_a$, i.e.,

$$p_a = \frac{2}{\delta_a \sqrt{2\pi}} \int_{-\infty}^{-|l_a-l|} \exp\left(-\frac{(x-l_a)^2}{2\delta_a^2}\right) dx. \quad (10)$$

To reduce hyper-parameters, we assume that the standard deviation is proportional to the mean $\delta_a = L_a \times \lambda$ for all categories of ships. We multiply the two probabilities to obtain the final detection confidence, $\hat{p}_a = p_a \times s_a$.

IV. Experiments

We evaluate our method on our FGSD2021 dataset, the public HRSC2016 [45] and UCAS-AOD [46] dataset. In this section, we first introduce the datasets and implementation details, then perform ablation studies and compare our network to several state-of-the-art methods.

A. Datasets

1) HRSC2016: The HRSC2016 dataset [45] is a challenging dataset for ship detection in remote sensing images, which collected six famous harbors on Google Earth. The training, validation, and test sets include 436 images with 1207 samples, 181 with 541 samples, and 444 images with 1228 samples, respectively. The image size of this dataset ranges from 300 × 300 to 1500 × 900. This dataset includes three levels of tasks (i.e., L1, L2, and L3), and these three tasks contain 1 class, 4 classes, and 19 classes, respectively. Besides, the head point of ships is given in this dataset. Following [28] [35] [32], we evaluate our method on task L1. We used the training and validation set in the training phase and evaluated the detection performance on the test set.
2) **FGSD2021:** Existing ship datasets HRSC2016 have the following shortcomings. First, the GSD is unknown, so we cannot get the size of objects in the image by the actual size on the ground. Second, the size of the image is very small which is inconsistent with the actual remote sensing image detection task. To solve these problems, we propose a new ship detection dataset FGSD2021 which has a fixed GSD.

Our dataset is developed by collecting high-resolution satellite images from publicly available Google Earth, which covers some famous ports such as Dandiego, Kitsap-Bremerton, Norfolk, Pearl Harbor, and Yokosuka. We usually obtain multiple images of the same port on different days, and there are also some images from the HRSC2016 dataset. We collected 636 images with a normalized GSD, 1 meter per pixel. The images in our dataset are very large, usually, one image covers a whole port. The width of images is ranged from 157 to 7789 pixels, and the average width is 1202 pixels, the height is ranged from 224 to 6506 pixels, and the average height is 1205 pixels. Our FGSD2021 dataset is divided into 424 training images and 212 test images. The training set is used in the training phase. The detection performance of the proposed method is evaluated on the test set. FGSD2021 including 5274 labeled targets and 20 categories are chosen and annotated. We use the labelimg\(^1\) tools to label the ship, the angle range is \([0^\circ, 360^\circ]\), and the main direction is the direction of the bow. Some examples of annotated patches are shown in Fig. 7.

3) **UCAS-AOD:** The UCAS-AOD dataset \([46]\) contains 1510 aerial images of about 659 × 1280 pixels and 14596 instances of two categories including plane and car. The angle range of target in this dataset is \([0^\circ, 180^\circ]\), so we manually marked the direction of the head. We randomly sampled 1132 images for training and 378 images for testing. All images were cropped into patches of size 672 × 672.

**B. Implementation Details**

Our network was implemented in PyTorch on a PC with Intel Core i7-8700K CPU, NVIDIA RTX 2080Ti GPU. We used the Adam method \([47]\) as the optimizer, and the initial

---

\(^1\)https://github.com/chinakook/labelImg2
learning rate was set to $2.5 \times 10^{-4}$. We trained our network for 140 epochs with a learning rate being dropped at 90 epochs. During the training phase, we used random rotation, random flipping, and color jittering for data augmentation. To maintain the GSD of the image, we cropped all images into $1024 \times 1024$ slices with a stride of 820, resized them to $512 \times 512$. We merged the detection results of all the slices to restore the detecting results on the original image. Finally, we applied rotated-non-maximum-suppression (RNMS) with an IoU threshold of 0.15 to discard repetitive detections. The speed of the proposed network was measured on a single NVIDIA RTX 2080Ti GPU.

Several different backbones (e.g., deep layer aggregation (DLA) [48] and hourglass network (Hourglass) [49]) can be used to extract features from images. We followed CenterNet [11] to enhance DLA by replacing ordinary convolutions with deformable convolutions and add a 256 channel 3 $\times$ 3 convolutional layer before the output head. The hourglass network consists of two sequential hourglass modules. Each hourglass module includes 5 pairs of down and up convolutional networks with skip connections. This network generally yields better keypoint estimation performance [26].

C. Evaluation Metrics

The IoU between oriented boxes is used to distinguish detection results. The mean average precision (mAP) and head direction accuracy are used to evaluate the performance of arbitrary-Oriented detectors.

1) IoU: The IoU is the result of dividing the overlapping area by the union area of two boxes. We adopted the evaluation approach in DOTA [50] to get the IoU. If the IoU between a detection box and a ground-truth is higher than a threshold, the detection box is marked as true-positive (TP), otherwise false-positive (FP). If a ground-truth box has no matching detections, it is marked as false negative (FN).

2) mAP: The precision and recall are calculate by precision = $\frac{TP}{TP+FP}$, recall = $\frac{TP}{TP+FN}$. We first set a set of thresholds, and then we get a corresponding maximum precision for each recall threshold. AP is the average of these precisions. The mean average precision (mAP) is the mean of APs over all classes. The mAP$_{0.5}$-mAP$_{0.9}$ is computed under the IoU threshold of 0.5-0.8 respectively. PASCAL VOC2007 metric is used to compute the mAP in all of our experiments.

3) Head direction accuracy: The prediction angle range of the previous algorithm is $0^\circ$-180$^\circ$, which cannot distinguish between the bow and stern of the ship. The mAP base on the IoU between two rotated boxes is taken as the only evaluation criterion, which cannot reflect the accuracy of the bow direction. To solve this problem, we define bow direction accuracy as an additional evaluation. That is the proportion of the ships whose angle difference from the ground-truth less than 10 degrees in all TPs.

D. Ablation Study

In this subsection, we present ablation experiments to investigate our models.

### TABLE I: Results achieved on FGSD2021 with different ablation versions. ‘Baseline’ represents adding a branch to predict the angle based on CenterNet. ‘Head Point’ represents replacing the angle prediction branch to head point estimation module. ‘Rotate kernel’ represents generating center heatmap by rotated kernel in training. ‘OIM’ represents add orientation-invariant model behind the backbone. ‘Extra convolution’ represents replacing the OIM with two extra convolution layers. ‘Refine probability’ represents using the prior size information to adjust the confidence score of the detected boxes.

| Ablation Setting | baseline | Different Settings of CHPDet |
|------------------|----------|-----------------------------|
| Head Point       | ✓        | ✓ ✓ ✓ ✓ ✓ ✓                   |
| Rotate kernel    | ✓        | ✓ ✓ ✓ ✓ ✓ ✓ ✓                   |
| OIM              | ✓        | ✓ ✓ ✓ ✓ ✓ ✓ ✓                   |
| Extra convolution| ✓        | ✓ ✓ ✓ ✓ ✓ ✓ ✓                   |
| Refine Probability| ✓      | ✓ ✓ ✓ ✓ ✓ ✓ ✓                   |

1) CenterNet as baseline: As an anchor-free detector, CenterNet performs keypoint estimation to find the center point and regresses the object size at each center point position. To carry out arbitrary-oriented ship detection, we add an extra branch to predict the angle as a baseline which is named CenterNet-Rbb. CenterNet-Rbb uses a DLA34 as the backbone, and presents ships as rotated boxes with angle, and uses the L1 loss function to optimized angle regression feature maps. We set weighted factor $\lambda_{angle} = 0.1$ to balance the contribution of these parts since the scale of the loss is ranged from 0 to 180. As shown in Table I, CenterNet-Rbb achieves an mAP of 70.52% which demonstrates that our baseline achieves competitive performance.

2) Effectiveness of the head point estimation: When we replace the angle prediction branch with the head point estimation module, the overall mAP is improved from 70.52% to 82.96%. It is a significant improvement, which fully demonstrates the effectiveness of the head point estimation approach. This improvement mainly comes from two aspects. First, the algorithm makes full use of the prior knowledge of the bow point and improves the accuracy of angle regression. Second, since multi-task learning is performed, bow detection increases the supervision information and improves the accuracy of other tasks.

To further verify the promoting effect of head point estimation for center point detection and size detection, we set all angles of ground-truth and the detected box to $0^\circ$. Compared with the CenterNet-Rbb, The mAP of CHPDet has risen from 84.4% to 88.0%. This shows that the head point estimation is equivalent to multi-task joint training. It gives more supervision to the network and improves the performance of the network. Besides, the head point estimation only introduces 3 additional channels feature maps and 0.7 ms speed latency.

3) Effectiveness of the rotated Gaussian kernel: Our detector uses the rotated Gaussian kernel to map the annotations to target heatmaps and achieves an improvement of 0.6% in terms of nominal Gaussian kernels. This implies that rotated Gaussian kernel is a better representation for OBB in the aerial images.

The rotated Gaussian kernel can adjust its shape and di-
TABLE II: Performance of CHEDet achieved on FGSD2021 with different variance coefficient λ. ‘without refine’ represents using the original confidence without refinement. ‘Ground truth class’ represents using ground truth class label to eliminate the misclassification.

| Backbone        | Image Size | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 | 0.1 without refine | 0.1 Ground truth class |
|-----------------|------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-------------------|------------------------|
| DLA34           | 512 × 512  | 87.40| 87.91| 87.39| 87.45| 87.17| 87.20| 87.15| 81.10| 86.64| 89.39|                  |                        |
| DLA34 × 1024    | 86.37| 87.84| 89.28| 88.17| 88.68| 88.85| 88.47| 88.50| 88.39| 89.74|                  |                        |

TABLE III: Detection accuracy on different types of ships and overall performance with the state-of-the-art methods on FGSD. The short names for categories are defined as (abbreviation-full name): Air - Aircraft carriers, Was - Wasp class, Tar - Tarawa class, Aus - Austin class, Whi - Whidbey Island class, San - San Antonio class, New - Newport class, Tic - Ticonderoga class, Bur - Arleigh Burke class, Per - Perry class, Lew - Lewis and Clark class, Sup - Supply class, Kai - Henry J. Kaiser class, Hop - Bob Hope Class, Mer - Mercy class, Fre - Freedom class, Ind - Independence class, Ave - Avenger class, Sub - Submarine and Oth - Other. CHPDet† means CHPDet trained and detected with 1024 × 1024 image size.

| Method            | Backbone | Detection accuracy (%) |
|-------------------|----------|------------------------|
| R²CNN [29]        | DLA34    | 89.9 | 80.9 | 80.5 | 79.4 | 87.0 | 87.8 | 84.4 | 89.0 | 89.6 | 79.5 | 80.4 | 87.1 | 85.3 | 100  | 82.4 | 100  |
| Retinanet-Rbb [23]| DLA34    | 89.9 | 89.2 | 78.2 | 87.3 | 77.0 | 86.9 | 62.7 | 81.5 | 83.3 | 70.6 | 46.8 | 49.9 | 80.2 | 83.1 | 100  | 80.6 | 89.7 |
| ROI-Trans [30]    | DLA34    | 89.9 | 88.6 | 82.7 | 89.5 | 78.5 | 88.8 | 81.8 | 89.6 | 89.8 | 90.4 | 71.7 | 74.7 | 73.7 | 81.6 | 78.6 | 100  | 75.6 |
| SCRDet [31]       | DLA34    | 77.3 | 90.4 | 87.4 | 89.8 | 78.8 | 89.9 | 54.5 | 88.3 | 89.6 | 74.9 | 68.4 | 59.2 | 90.4 | 77.2 | 81.8 | 73.9 | 100  |
| CSL [33]          | DLA34    | 89.9 | 87.1 | 77.2 | 80.2 | 71.4 | 77.2 | 52.7 | 87.7 | 87.7 | 74.2 | 57.1 | 97.2 | 77.6 | 80.5 | 100  | 72.7 | 100  |
| DCL [34]          | DLA34    | 89.5 | 81.4 | 78.9 | 80.7 | 87.0 | 87.9 | 49.8 | 78.7 | 87.2 | 76.1 | 60.6 | 67.9 | 90.4 | 80.0 | 77.9 | 100  | 73.7 |
| R²Det [32]        | DLA34    | 90.9 | 80.9 | 81.5 | 90.1 | 79.3 | 87.5 | 29.5 | 77.4 | 89.4 | 69.7 | 63.7 | 80.3 | 76.8 | 72.7 | 83.3 | 90.9 | 80.4 |
| RSDet [33]        | DLA34    | 89.8 | 80.4 | 75.8 | 77.3 | 78.6 | 88.8 | 26.1 | 84.7 | 87.6 | 75.2 | 55.1 | 74.4 | 89.7 | 89.3 | 100  | 86.4 | 100  |
| S²A-Net [35]      | DLA34    | 89.1 | 81.3 | 73.3 | 89.1 | 80.9 | 89.9 | 81.2 | 89.2 | 90.7 | 88.9 | 60.5 | 75.9 | 81.6 | 89.2 | 100  | 68.6 | 90.9 |
| ReDet [51]        | DLA34    | 90.9 | 90.6 | 80.3 | 81.5 | 89.3 | 88.4 | 81.8 | 88.8 | 90.3 | 90.5 | 78.1 | 76.0 | 90.7 | 87.0 | 98.2 | 84.4 | 90.9 |
| Oriented R-CNN [52]| DLA34    | 90.9 | 89.7 | 81.5 | 81.1 | 79.6 | 88.2 | 98.9 | 89.8 | 90.6 | 87.8 | 60.4 | 73.9 | 81.8 | 86.7 | 100  | 60.0 | 79.4 |
| BBA Vectors [53]  | DLA34    | 99.5 | 90.9 | 75.9 | 94.3 | 90.9 | 52.9 | 88.5 | 90.0 | 80.4 | 72.2 | 76.9 | 88.2 | 99.6 | 100  | 94.0 | 100  |
| DARDet [54]       | DLA34    | 99.8 | 89.2 | 69.7 | 89.6 | 88.0 | 81.4 | 90.3 | 89.5 | 90.5 | 79.7 | 62.5 | 87.9 | 90.2 | 92.0 | 100  | 68.9 | 81.8 |
| CenterNet-Rbb [11]| DLA34    | 67.2 | 77.9 | 79.2 | 75.5 | 66.8 | 79.8 | 76.8 | 83.1 | 83.0 | 77.7 | 54.5 | 72.6 | 77.4 | 100  | 60.0 | 74.8 | 46.3 |
| CHPDet-DLA34†     | DLA34    | 90.9 | 90.4 | 89.6 | 89.3 | 89.6 | 99.1 | 99.4 | 90.2 | 90.0 | 90.3 | 70.7 | 87.9 | 89.2 | 96.5 | 100  | 85.1 | 100  |

5) Effectiveness of the Refine probability model: In the FGSD2021 dataset, the actual length of each category is determined. For example, the length of the Ticonderoga-class cruiser is 172.8 meters. In our designed network, the prior knowledge of ship length is used to refine the confidence of the detected ships belonging to a certain category. Table I shows the mAP values of different ablation versions on the test set. It can be observed that the baseline model achieves the lowest mAP. When the prior size information is incorporated, the performance has been improved. The accuracy improvement on low-resolution images is more obvious, e.g., from 86.61% to 87.91%, an increase of 1.3% in mAP. It demonstrates that the prior size information can improve classification accuracy.

We set a variance coefficient to adjust the influence of size on probability. Consequently, we use the length of this type of ship l_α multiplied by a coefficient r as the mean square error of this type δ_α, δ_α = l_α × r. The variance coefficient will affect classification accuracy. When the coefficient is large, the probability difference between different categories will be smaller, and the influence of the size on the confidence of the category will be smaller, and vice versa. As can be observed in Table II, when the coefficient is small, it is equivalent to use size as the main information to classify objects. Accuracy increases gradually as the coefficient increases, and when the coefficient is larger than 0.2, the coefficient has little impact on the accuracy. When we treat all categories as one category and remove the category influence on the detection results, the angle prediction, but also improves the accuracy of center point detection and size regression.
TABLE IV: Detection performance on the FGSD2021 at different IoU thresholds and the accuracy of bow direction. BDA presents bow direction accuracy

| Method          | Backbone | Image Size | mAP | mAP<sub>0.5</sub> | mAP<sub>0.6</sub> | mAP<sub>0.7</sub> | mAP<sub>0.8</sub> | BDA   | FPS  |
|-----------------|----------|------------|-----|-------------------|------------------|-------------------|-------------------|-------|------|
| R<sup>2</sup>CNN [29] | Resnet50 | 512 × 512  | 78.09 | 75.05            | 64.83            | 36.41             | -                 | 10.3  |
| Retinanet-Rbb [23] | Resnet50 | 512 × 512  | 73.49 | 69.17            | 62.82            | 45.00             | -                 | 35.6  |
| Rod-Trans [30]  | Resnet50 | 512 × 512  | 83.48 | 82.63            | 80.35            | 65.18             | -                 | 19.2  |
| SCRDet [31]     | Resnet50 | 512 × 512  | 75.90 | 70.98            | 61.82            | 35.12             | -                 | 9.2   |
| CSL [33]        | Resnet50 | 512 × 512  | 73.73 | 69.71            | 60.25            | 34.93             | -                 | 10.4  |
| DCL [34]        | Resnet50 | 512 × 512  | 73.34 | 69.19            | 57.80            | 28.54             | -                 | 10.0  |
| R<sup>2</sup>Det [32] | Resnet50 | 512 × 512  | 70.47 | 68.32            | 57.17            | 27.44             | -                 | 14.0  |
| RSDet [9]       | Resnet50 | 512 × 512  | 73.74 | 69.55            | 61.52            | 35.83             | -                 | 15.4  |
| S<sup>2</sup>A-Net [35] | Resnet50 | 512 × 512  | 80.19 | 79.58            | 75.65            | 58.82             | -                 | 33.1  |
| ReDet [51]      | ReResnet50 | 512 × 512 | 85.44 | 84.65            | 80.24            | 67.94             | -                 | 13.8  |
| Oriented R-CNN [52] | Resnet50 | 512 × 512  | 82.54 | 81.32            | 78.53            | 64.87             | -                 | 27.4  |
| BBAVectors [53] | Resnet50 | 512 × 512  | 83.59 | 82.74            | 78.55            | 62.48             | -                 | 18.5  |
| DARDet [54]     | Resnet50 | 512 × 512  | 80.31 | 79.62            | 74.77            | 59.21             | -                 | 31.9  |
| CenterNet-Rbb [11] | DLA34   | 512 × 512  | 70.52 | 69.34            | 65.52            | 45.33             | -                 | 48.5  |
| CHPDet(ours)    | DLA34   | 512 × 512  | 87.91 | 87.15            | 83.69            | 71.24             | 97.84             | 41.7  |
| CHPDet(ours)    | DLA34   | 1024 × 1024| 89.29 | 88.98            | 86.57            | 73.56             | 98.39             | 15.4  |

mAP is 89.33%, and 89.74%, respectively. At the same time, by incorporating prior information to adjust the classification confidence, the detection accuracy under 20 categories with an input image of size 1024x1024 achieved an mAP of 89.28% which shows that after incorporating the prior information, almost all categories are classified correctly.

6) Bow direction accuracy: It can be seen from Table III that the bow direction accuracy of our CHPDet is up to 97.84, 98.14, and 98.39, respectively. This shows that almost all bow directions of ships are correct. As shown in Fig. 9, the pink dots represent the correct head point and the green dots represent the wrong head point. Our detection algorithm can well detect the bow direction of all types of ships, including aircraft carriers, amphibious ships. Only a small number of ships or submarines whose bow and stern are similar from a bird-view perspective, the bow direction will be opposite.

E. Comparison with other methods

In this section, we compare our method with other representative ship detectors including RetinaNet-Rbb [23], R<sup>2</sup>CNN [29], SCRDet [31], S<sup>2</sup>A-Net [35], and CHPDet (ours), respectively. Different color of rotated boxes represents a different type of ships. The pink point represents the head point.

Fig. 8: Comparison of the detection results in FGSD2021 with different methods. The first column is the ground truth, and the second to the last columns are the results of Retinanet-Rbb [23], ROI-Trans [30], SCRDet [31], S<sup>2</sup>A-Net [35], and CHPDet (ours), respectively. Different color of rotated boxes represents a different type of ships. The pink point represents the head point.
TABLE V: Detection accuracy on the HRSC2016 dataset. 07 means using the 2007 evaluation metric.

| Method       | Backbone | mAP(07) |
|--------------|----------|---------|
| R²-CNN [29]  | ResNet101| 73.07   |
| RRPN [28]    | ResNet101| 79.08   |
| R²-PN [10]   | VGG16    | 79.6    |
| ROI-trans [30]| ResNet101| 86.20   |
| Gliding Vertex [55]| ResNet101| 88.20   |
| BBAVectors [53]| ResNet101| 88.6    |
| R²-Det [32]  | ResNet101| 89.26   |
| FPN-CSL [33] | ResNet101| 89.62   |
| R²-Det-DCL [34]| ResNet101| 89.46   |
| DAL [56]     | ResNet101| 89.77   |
| R²-Det-GWD [57]| ResNet101| 89.85   |
| RSDet [9]    | ResNet152| 86.5    |
| FR-Est [58]  | ResNet101| 89.7    |
| S²A-Net [35] | ResNet101| 90.2    |
| Oriented RepPoints [59]| ResNet50| 90.38   |
| ReDet [51]   | ReResNet50| 90.46   |
| Oriented R-CNN [52]| ResNet101| 90.50   |
| DARDet [54]  | ResNet50 | 90.37   |
| CHPDet(ours) | DLA34    | 88.81   |
| CHPDet(ours) | Hourglass104| 90.55   |

TABLE VI: Detection accuracy on the UCAS-AOD dataset.

| Method       | Backbone | car | airplane | mAP(07) |
|--------------|----------|-----|----------|---------|
| YOLOv3 [60]  | Darknet53| 74.63| 89.52    | 82.08   |
| RetinaNet [23]| ResNet101| 84.64| 90.51    | 87.57   |
| FR-O [50]    | ResNet101| 86.87| 89.86    | 88.36   |
| ROI-trans [30]| ResNet101| 87.99| 89.90    | 88.95   |
| FPN-CSL [33] | ResNet101| 88.09| 90.38    | 89.23   |
| R²-Det-DCL [34]| ResNet101| 88.15| 90.57    | 89.36   |
| DAL [56]     | ResNet101| 89.25| 90.49    | 89.87   |
| CHPDet(ours) | DLA34    | 88.58| 90.64    | 89.61   |
| CHPDet(ours) | Hourglass104| 89.18| 90.81    | 90.00   |

Fig. 9: Some bow direction detection result of CHPDet. The pink dots represent the correct head point and the green dots represent the wrong head point.

This improvement mainly comes from three aspects. First, the algorithm makes full use of the prior knowledge of the bow point and improves the accuracy of directional regression. Second, since multi-task learning is performed, bow detection increases the supervision information and improves the accuracy of other tasks. Last, the prior knowledge of ship length is used to refine the confidence of the detected ships belonging to a certain category. The usage of the prior knowledge of ships introduces significant performance improvements.

2) Results on HRSC2016: The HRSC2016 dataset contains plenty of ships with arbitrary orientations. We evaluate our method on task L1 which contains 1 class and report the results with VOC2007 metric. To demonstrate the performance of our detector, we compare it with other state-of-the-art methods, i.e., Det-DCL [34], Oriented R-CNN [52], and Oriented RepPoints [59]. The overall comparison performance is reported in Table V. Our method achieves the best performance over all the compared methods, at an accuracy of 90.55%. To further show the performance of CHPDet, the detection results are visualized in Fig. 10. As shown in the first two columns, the densely parked ships can be detected well. In the last two columns, there is a lot of background around ships,
which is a huge challenge for detectors. The results indicate that our proposed method can avoid false alarms in complex background.

3) Results on UCAS-AOD: The UCAS-AOD dataset contains a large mount of cars and planes, which are often overwhelmed by a complex background in aerial images. For a fair comparison, we only report the results under VOC2007 metric. Table VI shows the results with the recent methods on the UCAS-AOD dataset. It can be seen that our proposed method achieves the best performance (with an mAP of 90.00%). The CHPDet, which uses a larger output resolution (output stride of 4) compared to traditional object detectors (output stride of 8) and presents ship as the center and head points, can capture abundant information of small objects. Fig. 11 gives some example detection results on the UCAS-AOD dataset. We find that CHPDet performs well in a variety of challenging scenes, which demonstrates the generalization capability of the detector.

V. CONCLUSION

Our proposed approach converts discontinuous angle regression to continuous keypoint estimation by formulating ships as rotated boxes with a head point representing the direction. This design can incorporate the prior knowledge of the bow point, which not only improves the detection performance, but also expands the scope of predicted angle to $[0^\circ - 360^\circ]$). Our method can distinguish between bow and stern. CHPDet has simple structure. It has only one positive sample per annotation and simply extracts local peaks in the keypoint heatmap. It does not need Non-Maximum Suppression (NMS). This design ensures high time efficiency. The prior knowledge of ship length is also incorporated to refine the confidence of the detected ships belonging to a certain category.

Although our method achieves encouraging results on ship detection from remote sensing images, our method can not be directly used in normal object detection datasets in aerial images such as DOTA [50]. That is because, CHPDet needs more accurate annotations which mark the direction of the target head in the range of $360^\circ$. CHPDet is several times faster than most detectors in inference, but it suffers from a long training time. For future work, we will address this issue by encoding more training samples from annotated boxes.

In this paper, we proposed a one-stage anchor-free detection framework to detect arbitrary-oriented ships from remote sensing images by making full use of the prior of ships. Our method detects ships by extracting the center, head of ships, and regresses the size of ships at each center point with rotation-invariant features. And we refine the detection results based on the prior information. CHPDet avoids complex anchor design and computing relative to the anchor-based methods and can accurately predict angles in a large range $[0^\circ - 360^\circ]$. Experimental results demonstrate that our method achieves better accuracy and efficiency as compared with other ship detectors.

REFERENCES

[1] S. He, H. Zou, Y. Wang, R. Li, F. Cheng, X. Cao, and M. Li, “Enhancing mid-low-resolution ship detection with high-resolution feature distillation,” IEEE Geoscience and Remote Sensing Letters, 2021.
Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao, “Oriented response networks,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society, 2017, pp. 1302–1310.

Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao, “Oriented response networks,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society, 2017, pp. 4961–4970.

Z. Liu, L. Yuan, L. Weng, and Y. Yang, “A high resolution optical satellite image dataset for ship recognition and some new baselines,” in International Conference on Pattern Recognition Applications and Methods, vol. 2. SCITEPRESS, 2017, pp. 324–331.

C. Li, C. Xu, Z. Cui, D. Wang, T. Zhang, and J. Yang, “Feature-attentioned object detection in remote sensing imagery,” in 2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019, pp. 3886–3890.

D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, Y. Bengio and Y. LeCun, Eds., 2015.

F. Yu, D. Wang, E. Shelhamer, and T. Darrell, “Deep layer aggregation,” in 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. IEEE Computer Society, 2018, pp. 2403–2412.

A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” pp. 483–499, 2016.

G. Xia, X. Bai, J. Ding, Z. Zhu, S. J. Belongie, J. Lao, M. Datcu, M. Pelillo, and L. Zhang, “DOTA: A large-scale dataset for object detection in aerial images,” in 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 19-22, 2018. IEEE Computer Society, 2018, pp. 3974–3983.

J. Han, J. Ding, N. Xue, and G.-S. Xia, “Redet: A rotation-equivariant detector for aerial object detection,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2786–2795.

X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, “Oriented r-cnn for object detection,” arXiv preprint arXiv:2108.05699, 2021.

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 Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 2150–2159.

F. Zhang, X. Wang, S. Zhou, and Y. Wang, “Dardet: A dense anchor-free rotated object detector in aerial images,” arXiv preprint arXiv:2110.01025, 2021.

Y. Xu, M. Fu, Q. Wang, Y. Yang, K. Chen, G.-S. Xia, and X. Bai, “Gliding vertex on the horizontal bounding box for multi-oriented object detection,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 4, pp. 1452–1459, 2020.

Q. Meng, Z. Zhou, L. Miao, H. Zhang, and L. Li, “Dynamic anchor learning for arbitrary-oriented object detection,” arXiv preprint arXiv:2012.04150, vol. 1, no. 2, p. 6, 2020.

X. Yang, J. Yan, Q. Ming, W. Wang, X. Zhang, and Q. Tian, “Rethinking rotated object detection with gaussian wasserstein distance loss,” arXiv preprint arXiv:2101.11952, 2021.

K. Fu, Z. Chang, Y. Zhang, and X. Sun, “Point-based estimator for arbitrary-oriented object detection in aerial images,” IEEE Transactions on Geoscience and Remote Sensing, 2020.

W. Li and J. Zhu, “Oriented reppoints for aerial object detection,” arXiv preprint arXiv:2105.11111., 2021.

J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” arXiv preprint arXiv:1804.02767, 2018.

Feng Zhang received the B.E. degree in electronic information engineering from Harbin Institute of Technology (HIT), Harbin, China, in 2009, and the M.E. degree in information and communication engineering from National University of Defense Technology (NUDT), Changsha, China, in 2011. He is currently pursuing a Ph.D. degree from the College of Electronic Science and Technology, NUDT. His research interests focus on include remote sensing image processing, pattern recognition, and computer vision.

Xuying Wang received the B.S. degree in electronic information engineering from Beihang University, Beijing, China, in 2009, the M.S. and Ph.D. degrees in electronic science and technology from the National University of Defense Technology, Changsha, China, in 2011 and 2016. He is currently an Assistant Professor with the College of Electrical Science, National University of Defense Technology. His research interests include remote sensing image processing, pattern recognition.

Shilin Zhou received the B.S., M.S., and Ph.D. degrees in electrical engineering from Hunan University, Changsha, China, in 1994, 1996, and 2000, respectively. He is currently a Full Professor with the College of Electrical Science, National University of Defense Technology, Changsha. He has authored or co-authored over 100 referred papers. His research interests include image processing and pattern recognition.

Yingqian Wang received the B.E. degree in electrical engineering from Shandong University (SDU), Jinan, China, in 2016, and the M.E. degree in information and communication engineering from National University of Defense Technology (NUDT), Changsha, China, in 2018. He is currently pursuing a Ph.D. degree from the College of Electronic Science and Technology, NUDT. He has authored several papers in journals and conferences such as TPAMI, TIP, CVPR, and ECCV. His research interests focus on low-level vision, particularly on light field imaging and image super-resolution.

Yi Hou received the B.S. degree from Wuhan University, China, and the M.S. as well as a Ph.D. degree from the National University of Defense Technology, China. He held a visiting position with the Department of Computing Science, University of Alberta, Canada, from 2014 to 2016. His main research interests include robot visual SLAM, visual place recognition, time series classification, signal processing, computer vision, deep learning, pattern recognition, and image processing.