Locality-Aware Rotated Ship Detection in High-Resolution Remote Sensing Imagery Based on Multiscale Convolutional Network

Lingyi Liu, Yunpeng Bai, and Ying Li

Abstract—Ship detection has been an active and vital topic in the field of remote sensing for a decade, but it is still a challenging problem due to the large-scale variations, the high aspect ratios, the intensive and rotated arrangement, and the background clutter disturbance. In this letter, we propose a locality-aware rotated ship detection (LARSD) framework based on a multiscale convolutional neural network (CNN) to tackle these issues. The proposed framework applies a UNet-like multiscale CNN to generate multiscale feature maps with high-level semantic information in high resolution. Then, an anchor-based rotated bounding box regression is applied for directly predicting the probability, the edge distances, and the angle of ships. Finally, a locality-aware score alignment (LASA) is proposed to fix the mismatch between classification results and location results caused by the independence of each subnet. Furthermore, to enlarge the data sets of ship detection, we build a new high-resolution ship detection (HRSD) data set, where 2499 images and 9269 instances were collected from Google Earth with different resolutions. Experiments based on public data set high-resolution ship collection 2016 (HRSC2016) and our HRSD data set demonstrate that our detection method achieves the state-of-the-art performance.

Index Terms—Anchor-based rotated bounding box regression, locality-aware score alignment (LASA), multiscale convolutional neural network (CNN), optical remote sensing image, ship detection.

I. INTRODUCTION

SHIP detection has been a topic of interest in the field of remote sensing over the last decades and has made great progress in promoting national defense construction, harbor management, and cargo transportation. With the rapid growth of satellite technology and construction, high-resolution optical remote-sensing images can be easily obtained, which contain abundant details for classifying objects. Attributing to the advancement, ship detection from optical remote-sensing images is under more active research. Bi et al. [1] employed the bottom-up visual attention mechanism and top-down visual cues for candidate selection and discrimination, respectively. Yang et al. [2] integrated sea surface analysis into ship detection to make the detection method robust to the variation of sea surfaces. These ship detection methods implement ship detection by extracting and recognizing the features of shapes and textures from ships. Nonetheless, it is tricky to design the handcrafted features applied to all categories of ship.

Recently, the convolutional neural network (CNN) [3]–[6] have led to significant progress in object detection. These state-of-the-art object detectors for natural images have already achieved promising results. However, several challenges limit the application of the state-of-the-art object detectors in ship detection from high-resolution remote-sensing images.

1) The sizes of ships range widely in remote-sensing images due to the variety of ship categories and space resolutions, as shown in Fig. 1(a). The fixed single receptive field determined by the architecture of CNN models cannot match the scale variability of ships.

2) Ships are in stripe-like shapes and are often docked inshore or side-by-side intensively, as shown in Fig. 1(b). Thus, using horizontal bounding boxes for location may cover a relatively large redundancy region.

3) The remote-sensing images captured by satellites generally have large size and contains numerous ship-like object (e.g., containers and fish rafts). These background clutter disturbances easily lead to false alarms on the detectors.

4) The existing ship detection data sets in optical satellite images are scarce, especially for rotated ship detection data sets.

Inspired by the success from natural images, the deep learning-based ship detection methods on optical remote-sensing images have attracted more and more attention. These methods [7], [8] have already achieved encouraging performance in the field of remote-sensing detection. However, the limitation of horizontal bounding boxes mentioned above poses a great challenge for inshore and adjacent ship detection. Several rotated ship detectors have also been proposed to overcome this limitation. Liu et al. [9] proposed a rotated region-based CNN for ship detection, which uses a rotated region of interest (RRoI) pooling layer to extract rotation region features and directly regress rotation angles. Yang et al. [10] further proposed multiscale rotation dense feature pyramid networks (R-DFPN). They design a rotation anchor strategy with multiscale RoI align to improve the efficiency of the feature extract model for the rotated object.

Li et al. [11] proposed a rotated ship detector based on a fully convolutional network (FCN), which predicts the angle and the length of bunding boxes directly. Liao et al. [12] proposed a rotation-sensitive regression detector (RDD) which imple-
ments classification and regression using rotation-insensitive features and sensitive features, respectively. Ding et al. [13] proposed the RoI transformer to effectively extract features by spatial transformations. Qian et al. [14] proposed a rotation-sensitive detector (RSDet). They designed a modulated rotation loss to deal with the loss discontinuity caused by rotation and introduced the eight-parameter regression to improve regression inconsistency. Xu et al. [15] proposed to accurately describe a multioriented object by gliding the vertex of the horizontal bounding box on each corresponding side, instead of directly regressing the four vertices.

To date, several ship detection data sets have been released for research reference. Wang et al. [16] constructed the SAR ship detection data set (SSDD), which is a ship detection data set of SAR images and includes 1160 images and 2456 instances in total. Liu et al. [17] built a ship detection data set in optical satellite images named high-resolution ship collection 2016 (HRSC2016). It contains 1061 images and 2886 instances collected from Google Earth and is annotated with rotated bounding boxes. In general, these data sets promote the development of ship detection, but the limited number of data sets restricts the further improvement of ship detection methods.

In this letter, we propose a one-stage rotated object detection framework based on multiscale CNN. Specifically, our model first learns a UNet-like multiscale CNN to extract multilevel feature maps. Then, an anchor-based rotated bounding box regression is utilized to generate some candidate targets. Moreover, the scores of the candidate targets are refined by locality-aware score alignment (LASA). Finally, we adopt non-maximum suppression to merge overlapping bounding boxes.

The main contributions of this letter are as follows:

1) We built a novel ship detection framework based on a one-stage detection method and utilized an additional angle regression branch to predict rotated bounding boxes as object targets. Experiment results demonstrate that our proposed method outperforms the state-of-the-art methods.

2) We proposed LASA to refine the scores of bounding boxes according to their locations, which alleviate the problem of mismatch between classification results and location results.

3) We introduced a new ship data set labeled by oriented bounding boxes, which has 2499 images and 9269 instances. The data set contains various ships with different scenarios, scales and space resolutions to eliminate the data set bias.

II. PROPOSED METHOD

In this section, we provide details for the proposed detection framework, Fig. 2 shows the overall framework of our method.

A. UNet-Like MultiScale CNN

Extracting semantic features efficiently from high-resolution remote-sensing images is crucial to the deep learning-based ship detection methods. We use pretrained ResNet-101 in ImageNet classification data sets as our backbone. Although final feature maps of ResNet-101 have high-level semantic information, it loses more detailed information during several pooling operations. Inspired by the semantic segment task, we introduce the network structure of UNet. As shown in Fig. 3, (C2, C3, C4, C5) donates the feature maps of ResNet-101, the outputs of last residual blocks for every stage. While feature fusion processing, we double the size of the coarse-resolution feature map by nearest neighbor upsampling and concatenate the upsampling feature map with corresponded lower level feature map. A $3 \times 3$ convolution and a $1 \times 1$ convolution are used to fuse the feature maps and reduce the number of feature map channels. Undergoing three iterations, we merge high-level semantic information into the low-level feature map then expend the resolution of feature maps gradually. To recognize various scales of ships, we utilize multi-level feature maps as the input of anchor-based rotated bounding box regression and classification networks. The final set of feature maps is defined as {P2, P3, P4}, and we append a $1 \times 1$ convolution to equalize the channel numbers of the feature maps.
B. Rotated Bounding Box Regression

As mentioned above, large-scale variations and high aspect ratios are the main challenges for detecting ship location. Therefore, we adopt the anchor-based regression method and multiscale prediction. Our prediction module includes two task-specific subnetworks: the classification subnet and the location regression subnet.

1) Rotated Bounding Boxes: We use the five variables \((d_1, d_2, d_3, d_4, \theta)\) to uniquely determine the rotated bounding box. As shown in Fig. 4, \(x\) and \(y\) are the coordinates of the anchor point in input images, \(\{D_1, D_2, D_3, D_4\}\) donates the corner points of rectangles. We set the point with the lowest sum of \(x\) and \(y\) as the \(D_1\), and \(D_2, D_3, D_4\) follow clockwise. \(\{d_1, d_2, d_3, d_4\}\) are the distances from the anchor point to four sides of the rectangle. \(\theta\) is the horizontal angle of the rotated bounding box, and we convert the range of \(\theta\) from \([-90, 0]\) to \((-45, 45)\) for normalization.

2) Anchor-Based Regression: We use the horizontal anchor strategy to facilitate the regression of distance. The size of anchor priors is determined by k-means clustering. We divide the groundtruth into three groups according to their areas and choose five clusters at each group as the anchor priors.

The assignment rule is based on the combination of intersection-over-union (IoU) and the range of rotated bounding boxes. Specifically, we first compute the IoU between groundtruth and anchor priors with the same center. Then, a location is set as a positive sample if it is in the range of groundtruth and its corresponding IoU is greater than 0.5, as an ignored sample if it is in the range of groundtruth and its IoU is in \([0.2, 0.5]\), and as a background sample in other conditions.

The classification subnet applies three \(3 \times 3\) convolution layers and a \(1 \times 1\) convolution layer in sequence, followed by a sigmoid function as the normalized function. The output for each location in feature maps is a \(K\)-dimensional vector which predicts the scores of all \(K\) anchors. The location regression subnet also consists of three \(3 \times 3\) convolution layers and a \(1 \times 1\) convolution layer, and exports a 5-D vector \((t_1, t_2, t_3, t_4, t_5)\) for distance and angle regression. The distance and the angle are calculated in the formulas as below

\[
\begin{align*}
d_1 &= h_1e^{t_1}, & d_2 &= w_1e^{t_2}, \\
d_3 &= h_1e^{t_3}, & d_4 &= w_1e^{t_4}, \\
\theta &= \text{sigmoid}(t_5 \times 2 - 1) \times \pi/4
\end{align*}
\]

where \(h_i\) and \(w_i\) are the length and width of the \(i\)th anchor prior.

3) Loss Function: Our loss function consists of classification loss and location loss. We introduce the focal loss \([4]\) as the classification loss to reduce the influence of easy samples on the classification loss.

Location loss is split into distance loss and angle loss to simplify the location of rotated bounding boxes. Distance loss is defined by IoU loss in UnitBox \([18]\). The angle loss is consistent with the cosine loss in EAST \([19]\). Our loss function is represented by

\[
L = \frac{1}{N_{\text{cls}}} \sum_i L_{\text{cls}}(P_i, P_i^*) + \frac{1}{N_{\text{reg}}} \sum_i P_i^* (\lambda_d L_d(d_i, d_i^*) + \lambda_a L_a(\theta_i, \theta_i^*))
\]

where \(P_i^*\) refers to the label of the object, \(P_i\) is the predicted classification score, \(d_i^*\) represents the predicted distance, \(d_i\) represents the distance from position to four sides of groundtruth, \(\theta_i^*\) is the predicted rotated angle of the object, \(\theta_i\) is the predicted rotated angle of groundtruth, \(\lambda_d, \lambda_a\) are the weights to balance the importance among the losses, and they are all set to one in our experiments. \(L_{\text{cls}}, L_d, L_a\) are the loss of classification, distance, and angle.

C. Locality-Aware Score Alignment

We note that location regression and classification are independent for detection methods. This framework may lead to the dilemma that the predicted bounding boxes with high scores have a low overlap rate with groundtruth, which will filter out other superior bounding boxes during the nonmaximum suppression, as shown in Fig. 5(a). Inspired by the correspondence between the distribution of classification score map and the key characteristics (shape and location) of the ship, as shown in Fig. 5(b), we find that the scores within the bounding box can reflect the location performance. It is a simple but effective strategy to combine the results between classification and location by employing the scores within the bounding box to modify the score of bounding box. Therefore, we propose the LASA, which samples the scores at several positions from the score map to refine the score of the bounding box, the flowchart shown in Fig. 6. We first determine the locations of sampling points by the location of predicted bounding boxes and the sampling point distributions related to the rotated bounding box. Fig. 7 shows two different sampling point distributions of the LASA. Then we compute the value of each sampling point from the score map. The bilinear interpolation is adopted to overcome the misalignment caused by location quantization. The final score is the average of these computed scores at sampling points.

III. Experimental Results

Our experiments are undertaken on HRSC2016 and high-resolution ship detection (HRSD) data set to evaluate the performance of our algorithm. The experiment settings and analysis of the experiments are described as follows.
Fig. 6. Flowchart of the LASA.

Fig. 7. Sampling point distributions of the LASA. (a) Nine-point rectangular distribution. (b) Nine-point diamond distribution.

Fig. 8. Statistical information of the HRSD. (a) Sample density distribution of images. (b) Area distribution of instances. (c) Angle distribution of instances. (d) Aspect ratio distribution of instances.

TABLE I

| Datasets  | Number of images | Number of instances | Image size | Image type | Resolution |
|-----------|------------------|---------------------|------------|------------|------------|
| SSD+      | 1160             | 2456                | 500x200-500x500 | SAR        | 1-10m      |
| HRSC2016  | 1061             | 2886                | 300x300-1500x900 | Optical    | 0.4m, 2m   |
| HRSD      | 2499             | 9269                | 600x300-7500x3500 | Optical    | 0.4-4m     |

A. Data Set Setup

Our HRSD data set is collected at shore side, sea and harbor, obtained from Google Earth with several resolutions to ensure the diversity of imagery background. Considering the data set bias, we selected ships with different categories and scales as detected targets. In total, we collected 2499 images with 9269 instances. Table I shows the comparison between our HRSD data set and other ship detection data sets. In order to avoid the data set bias, our data set involves more images and instances with a variety of backgrounds than other data sets. Fig. 8 illustrates the statistical information of the HRSD data set. It indicates that the instances in different image vary greatly in both the number shown in Fig. 8(a) and the size represented by area in Fig. 8(b), and the amount of instances in various angles is well balanced shown in Fig. 8(c), which will enhance the robustness of detectors. Furthermore, as shown in Fig. 8(d), the aspect ratios of the rotated bounding boxes can fully fit the shape characteristics of instances at any angle, which lead to more robust performance in rotated and densely docked scenarios compared to the horizontal bounding boxes.

TABLE II

| Methods | Anchor-based regression | Multi-level prediction | Average Precision (%) |
|---------|-------------------------|------------------------|-----------------------|
| R2CNN [9] | ✓ | ✓ | 75.7 | 74.3 |
| R-DFPN [10] | ✓ | ✓ | 79.6 | 80.2 |
| Rotated FCN [11] | ✓ | ✓ | 82.3 | 78.3 |
| RRD [12] | ✓ | ✓ | 84.3 | 82.3 |
| RoI Transformer [13] | ✓ | ✓ | 86.2 | 84.8 |
| RSDet [14] | ✓ | ✓ | 86.5 | 85.3 |
| Gliding Vertext [15] | ✓ | ✓ | 88.2 | 80.0 |
| LASA | ✓ | ✓ | 90.3 | 88.3 |

We use θ-based oriented bounding boxes \((x, y, w, h, \theta)\) as the location method to adapt the shape of ships for the detectors, where \((x, y)\) is the center coordinate, \((w, h)\) is the width and height of bounding boxes, and \(\theta\) denotes the horizontal angle of the rotated bounding boxes.

B. Experiment Design and Implementation Details

We use the ResNet-101 pretrained on ImageNet to initialize our backbone. We train on two GPUs for 90k iterations with stochastic gradient descent. The initial learning rate is 0.001 and is divided by ten for every 10k iterations. The momentum is 0.9 and weight decay is 0.0001. The size of input images is fixed as 512 × 512 due to the limitation of GPU memory. The average precision (AP) is used as metrics to evaluate the performance of each detector.

C. Comparison to State-of-the-Art Rotated Ship Detection Methods

The performance of our proposed method is compared with five competitive methods: rotated region-based CNN(R2CNN)[9], R-DFPN [10], rotate ship detection based on FCN (rotated FCN) [11], RRD[12], RoI transformer [13], RSDet [14], and gliding vertex [15]. Table II summarizes the experiment results of these comparisons conducted both on HRSC2016 and our data set, our method achieves the state-of-the-art performance, 90.3% and 88.3% AP, respectively.

D. Experiment Analysis

1) Effect of LASA: We design two sampling point distributions: diamond and rectangle. Locality-aware rotated ship detection(LARSD)-1 is a nine-point rectangular distribution, as shown in Fig. 7. LARSD-2, LARSD-3, and LARSD-4 are diamond distributions, and the numbers of sampling points are 5, 9, and 13, respectively. To better analyze the effect of each proposed strategy, we build a baseline which only deploys UNet-like multiscale CNN, without the LASA and anchor-based regression. The experiment results in Table III show that LARSD with different distributions has a positive effect on AP. Compared to LARSD-1 and LARSD-3, we find diamond distribution has a better performance due to that the diamond arrangement is similar to the shape of ships. We notice that properly increasing the number of sampling points is a promising approach for higher performance. Notably, the superior number of sampling points is 9.

We further compare the effect of each component in our proposed method. The result in Table V demonstrates all components we proposed improve the performance of the baseline and LASA makes the best contribution to the AP. The final AP of our method reaches 90.3% in HRSC2016.
TABLE III
EFFECT OF LASA

| Methods | Distribution | Number of points | Average Precision(%) |
|---------|--------------|------------------|----------------------|
| Baseline | /            | /                | 84.3 82.3            |
| LARSD-1 | Rectangle    | 9                | 86.2 84.1            |
| LARSD-2 | Diamond      | 5                | 87.1 84.8            |
| LARSD-3 | Diamond      | 9                | 88.0 86.1            |
| LARSD-4 | Diamond      | 13               | 87.3 85.4            |

TABLE IV
EFFECT OF ANCHOR-BASED REGRESSION

| Methods | Number of anchors | Average Precision(%) |
|---------|-------------------|----------------------|
| Baseline | /                 | 84.3 82.3            |
| LARSD-5 | 3                 | 86.1 85.5            |
| LARSD-6 | 5                 | 86.9 86.3            |
| LARSD-7 | 7                 | 85.3 84.7            |

TABLE V
ABLATION STUDY OF EACH COMPONENT IN OUR PROPOSED METHOD

| Methods | LASA | Anchor-based regression | Average Precision(%) |
|---------|------|-------------------------|----------------------|
| Baseline | /    | /                       | 84.3 82.3            |
| LARSD-8 | √    | /                       | 88.0 86.1            |
| LARSD-9 | /    | √                       | 86.9 86.3            |
| LARSD-10| √    | √                       | 90.3 88.3            |

advanced rotated detection methods on HRSC2016 and our data sets. The experiment results show that our model LARSD has the state-of-the-art performance in ship detection. For the future work, we will extend our model application to the multiclass detection data set and combine the study on semantic segmentation with our feature extract network to design a more effective network.

REFERENCES

[1] F. Bi, B. Zhu, L. Gao, and M. Bian, “A visual search inspired computational model for ship detection in optical satellite images,” IEEE Geosci. Remote Sens. Lett., vol. 9, no. 4, pp. 749–753, Jul. 2012.
[2] G. Yang, B. Li, S. Ji, F. Gao, and Q. Xu, “Ship detection from optical satellite images based on sea surface analysis,” IEEE Geosci. Remote Sens. Lett., vol. 11, no. 3, pp. 641–645, Mar. 2014.
[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 779–788.
[4] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal loss for dense object detection,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2980–2988.
[5] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 91–99.
[6] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 2117–2125.
[7] G. Cheng, P. Zhou, and J. Han, “Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 54, no. 12, pp. 7405–7415, Dec. 2016.
[8] Z. Deng, H. Sun, S. Zhou, J. Zhao, L. Lei, and H. Zou, “Multi-scale object detection in remote sensing imagery with convolutional neural networks,” ISPRS J. Photogramm. Remote Sens., vol. 145, pp. 3–22, Nov. 2018.
[9] Z. Liu, J. Hu, L. Weng, and Y. Yang, “Rotated region based CNN for ship detection,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2017, pp. 900–904.
[10] X. Yang et al., “Automatic ship detection in remote sensing images from Google earth of complex scenes based on multiscale rotation dense feature pyramid networks,” Remote Sens., vol. 10, no. 1, p. 132, Jan. 2018.
[11] M. Li, W. Guo, Z. Zhang, W. Yu, and T. Zhang, “Rotated region based fully convolutional network for ship detection,” in Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS), Jul. 2018, pp. 673–676.
[12] M. Liao, Z. Zhu, B. Shi, G.-S. Xia, and X. Bai, “Rotation-sensitive regression for oriented scene text detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2017, pp. 5909–5918.
[13] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, “Learning RoI transformer for oriented object detection in aerial images,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 2849–2858.
[14] N. Sani et al., “Smoke sky–exploring new frontiers of unmanned aerial systems for wildland fire science and applications,” IEEE Trans. Pattern Anal. Mach. Intell., early access, Feb. 18, 2020, doi: 10.1109/TPAMI.2020.2974745.
[15] Y. Wang, C. Wang, H. Zhang, Y. Dong, and S. Wei, “A SAR dataset of ship detection for deep learning under complex backgrounds,” Remote Sens., vol. 11, no. 7, pp. 765, Mar. 2019.
[16] Z. Liu, L. Yuan, L. Weng, and Y. Yang, “A high resolution optical satellite image dataset for ship recognition and some new baselines,” in Proc. 6th Int. Conf. Pattern Recognit. Appl. Methods, 2017, pp. 324–331.
[17] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, “UnitBox: An advanced object detection network,” in Proc. 24th ACM Int. Conf. Multimedia, 2016, pp. 516–520.
[18] X. Zhou et al., “EAST: An efficient and accurate scene text detector,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 5551–5560.