AROA: Attention Refinement One-Stage Anchor-Free Detector for Objects in Remote Sensing Imagery

Xu He[1], Shiping Ma[1], Linyuan He[1,2](✉), Fei Zhang[1], Xulun Liu[1], and Le Ru[1]

[1] Aeronautics Engineering College, Air Force Engineering University, Xi'an 710038, China
[2] Unbanned system Research Institute, Northwestern Polytechnical University, Xi'an 710072, China

Abstract. Object detection in remote sensing images is a typical computer vision application, which has broad requirements in practice. Recently, attention mechanisms have been widely utilized in a diverse range of visual tasks such as object detection and semantic segmentation. Aimed at the characteristics of remote sensing objects such as rotation variations and inter-class similarity, a question we ask is, what kind of attention mechanism do we really need? In this article, we propose a novel attention refinement one-stage anchor-free object detector (AROA) that leverages attention mechanisms to refine the performance of remote sensing object detection in a one-stage anchor-free network framework. Specifically, we first design an asymmetric spatial self-attention (AS^{2}A) mechanism to capture rich long-range spatial contexts and eliminate the rotate distortion. Then, to solve the issue of inter-class similarity and boost the multiclass identification capability, we propose a channel attention mechanism, named chain-connected channel attention (C^{3}A), which connects the adjacent attention blocks like a chain and dramatically mines the channel relationships. In addition, we also introduce an IoU-wise module (IM) to strengthen the correlation between localization and classification branches and filter out the detected boxes with low positioning quality. Extensive experimental results on the DOTA and NWPU VHR-10 datasets demonstrate the effectiveness of the proposed AROA.

Keywords: Object detection · Remote sensing images · Attention mechanisms · Anchor-free · One-stage

1 Introduction

With the fast-paced development of unmanned aerial vehicles (UAVs) and remote sensing technology, remote sensing image analysis has been increasingly applied in fields such as land surveying, environmental monitoring, and emergency relief. Object detection in remote sensing images is regarded as a high-level computer vision task which is purposed to pinpoint the targets in a remote sensing image. Due to the characteristics of remote sensing targets such as complex backgrounds, sparse texture, multiple scales, and variations of orientations, remote sensing object detection remains a challenging and significant research issue.

Y. Peng et al. (Eds.): ICIG 2021, LNCS 12888, pp. 269–279, 2021.
https://doi.org/10.1007/978-3-030-87355-4_23

In recent years, due to outstanding learning ability, the most advanced detection models have been developed by deep convolutional neural networks (DCNNs). Two-stage and one-stage are two basic research paradigms for remote sensing object detection. Two-stage detectors [1,2] resolve the task of object detection in two steps. The first step is to generate numerous region proposals as candidate regions of the objects. The second step is to extract the features of each region proposal from the backbone network and feed the features into the classifier and regressor for classification and position fine adjustment. Despite the two-stage detectors being effective and accurate, the computational cost is fairly high because of the complex traversal process of the region proposals. Compared with two-stage detectors, one-stage detectors [3–6] directly predict the classification and localization in a simple pipeline, which is more computationally efficient. Nevertheless, anchor-based one-stage detectors that rely on anchors mechanism would result in complicated computations related to anchor boxes. Therefore, research on one-stage anchor-free detectors that emancipate the model from massive computations on anchors has drawn much attention in recent years. To the best of our knowledge, many scholars have begun to design one-stage anchor-free detectors [7–17] used in remote sensing images. For instance, X-LineNet [7] detected the oriented aircraft in remote sensing images by predicting a pair of middle lines segments inside each bounding box. Combined with CornerNet [5] and CenterNet [18], Chen *et al.* [8] utilized an end-to-end fully convolutional network (FCN) to identity the targets according to the predicted corners, center, and angle of the ship. Shi *et al.* [17] transformed the vehicle detection task into the multitask learning issue of center keypoint and morphological sizes. These methods illustrated significant advantages for applying one-stage anchor-free detectors to extract remote sensing targets. Nevertheless, due to the cluttered arrangement, rotation variations, and inter-class similarity, conventional one-stage anchor-free detectors cannot be applicable to remote sensing object detection task directly.

Similar to the selective mechanism of human visual attention system, the attention in computer vision attends to the discriminative and valuable feature representation and restrain the information useless. Concretely, non-local neural network [19] (NLNN) is the most classical mechanism of spatial self-attention, which can capture long-range spatial contexts in order to obtain rich semantic information. Some works have made several attempts to incorporate NLNN in remote sensing object detection task. Li *et al.* [10] designed a cross-layer attention to obtain the non-local correlation between small objects. Wang *et al.* [11] inserted the NLNN block into the feature pyramid network to provide the richer spatial association information. However, the above-mentioned NLNN attention has serious rotate distortion, leading to performance degradation for multi-orientation remote sensing objects. Therefore, inspired by the asymmetric convolution, we proposed a novel asymmetric spatial self-attention (AS^{2}A) which strengthens the input of the NLNN with a 3×3 square convolution and two asymmetric central skeleton parts of the square convolution, i.e., 1×3 convolution and 3×1 convolution. AS^{2}A can be exploited to capture rich long-range spatial contexts and enhance the rotational robustness. Beyond spatial attention, squeeze and excitation network [20] (SENet), which can be served as the pioneer of channel attention for classification task, adaptively assign the attention weights of different channels. Wu *et al.* [12] made use

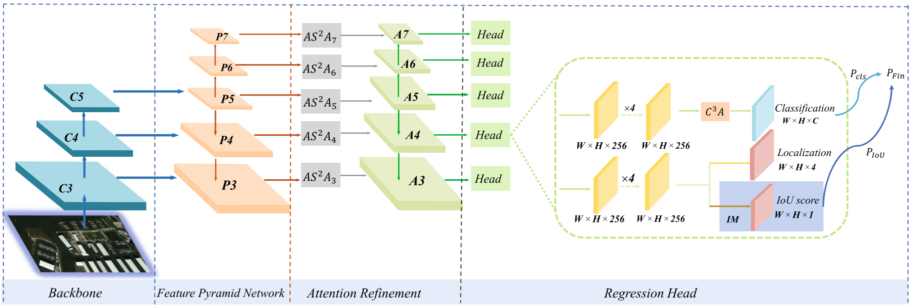


Fig. 1. The architecture of the proposed AROA, where C3, C4 and C5 represent the feature maps of the backbone network. P3 to P7 and A3 to A7 denote the feature levels of feature pyramid network (FPN) used for subsequent attention refinement and regression head, respectively. The shared regression head consists of three branches: classification, localization, and IoU score prediction, respectively. W, H, and C indicate the height, weight, and channel of the feature map, respectively.

of the SENet to enhance the small target capture capability. Zhang *et al.* [13] proposed a selective concatenation module to fuse the low- and high-level features by the SENet attention mechanism. However, as far as we know, these methods only integrate a single SE block into the network, which lacks of the adequate representation ability to identify different categories of targets with similar appearance. To solve this problem, we design a chain-connected channel attention (C^3A) to connect the multiple adjacent SE blocks for learning the deep channel relationships and discriminating the multiclass remote sensing objects with inter-class similarity. In addition, to filter out the low-quality detection boxes in densely arranged remote sensing objects, we propose an IoU-wise module (IM) to predict the Intersection-over-Union (IoU) for each detected box and the final bounding box confidence is set as the harmonic mean of the predicted IoU score and classification score. This can dramatically pull down the box probability as long as the localization or classification score has a rather small probability, so as to remove the low-quality bounding box. The experimental results have shown that the proposed attention mechanisms and IoU-wise module effectively improve the multiclass object detection performance.

2 Methodology

Figure 1 illustrates the architecture of our method. Based by FCOS [6], we implement a one-stage anchor-free detector as a fundamental framework. The input remote sensing images utilize a backbone network to extract features to FPN, and a shared regression head is used to perform the classification, localization, and IoU score prediction tasks. In this letter, the proposed AS^{2}A is employed behind the FPN for refining the feature representations of each level and the C^3A is merged before the final classification branch to boost the classification performance. Moreover, IM can be formulated as the harmonic mean of classification and localization confidences for the final non-maximum suppression post-processing and average precise computational.

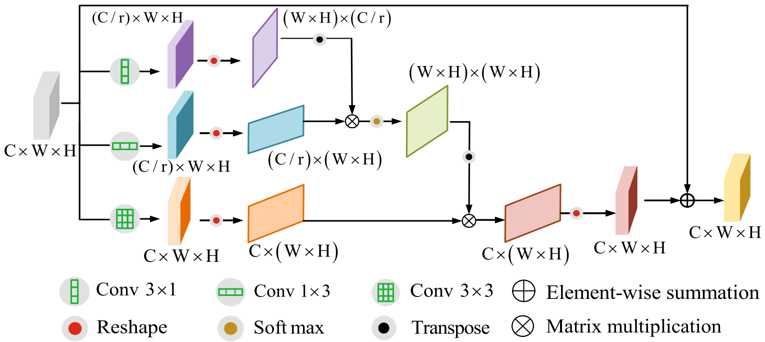


Fig. 2. Details of the position-wise spatial attention module inspired by [19]. C, W, H denote the channel, width, and height dimensions of the feature maps, respectively.

2.1 Asymmetric Spatial Self-attention

Observation shows that the targets in remote sensing images usually has a complicated background and multi-orientation patterns. Therefore, to obtain the long-range valuable context information and eliminate the rotate distortion, our method incorporates the asymmetric convolution which enhances the central cross section of square convolution in the input of the non-local network. As shown in Fig. 2, on the one hand, we first feed the given pyramid feature maps $P \in \mathbb{R}^{C \times W \times H}$ into two convolution units with the kernels of 1×3 and 3×1, resulting in two feature maps $P^{3\times1}, P^{3\times1} \in \mathbb{R}^{C/r \times W \times H}$, where r is set to 4 in our method. Next, $P^{3\times1}$ and $P^{3\times1}$ are reshaped to $\mathbb{R}^{(C/r)\times(W\times H)}$. In addition, $P^{3\times1}$ follows the operation of transpose, which results in $P^{3\times1} \in \mathbb{R}^{(W\times H)\times(C/r)}$. Then, we multiply the two feature maps and perform an operation of softmax on the calculated position-wise matrix to obtain the spatial attention maps $M \in \mathbb{R}^{(W\times H)\times(W\times H)}$:

$$m_{i,j} = \frac{\exp(P_i^{1\times3} \cdot P_j^{3\times1})}{\sum_{i=1}^{W\times H} \exp(P_i^{1\times3} \cdot P_j^{3\times1})} \tag{1}$$

where $m_{i,j}$ assesses the effect of the i^{th} position on the j^{th} position. On the other hand, we feed the pyramid feature maps P into a 3×3 convolution unit to obtain a feature map $P^{3\times3}$ of the same shape and reshape it to $\mathbb{R}^{C\times(W\times H)}$. Then, we multiply $P^{3\times3}$ by the transpose of spatial attention map M and reshape the output feature maps F to $\mathbb{R}^{C\times W\times H}$. Finally, we obtain the final output feature maps O by an element-wise summation operation between F and given feature map P as follows:

$$O_j = \delta F_j + P_j = \delta \sum_{i=0}^{W\times H} (m_{j,i} P_i^{3\times3}) + P_j \tag{2}$$

where δ is initialized to 0 and learnable to balance the weights between the features that traverse all positions and original features. Thus, the asymmetric spatial self-attention mechanisms adaptively enhance the long-range correlation of the spatial contexts and eliminate the rotate distortion, thus improving the representation capability for the subsequent per-pixel classification and regression.

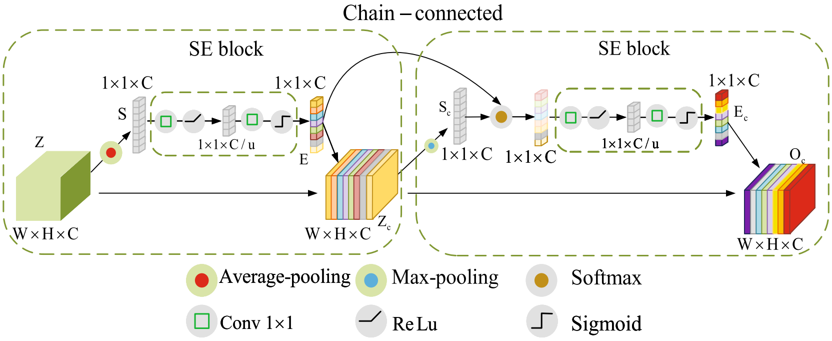


Fig. 3. Diagram of the proposed C^{3}A. In the first SE block, we use average pooling to squeeze the feature and max-pooling in the second SE block.

2.2 Chain-Connected Channel Attention

For reducing the effect of inter-class similarity in the multiclass remote sensing objects, it is necessary to adaptively rescale features across channels and boost the feature representation capability for classification task. Based on SENet, we design a chain-connected channel attention to learn the channel relationships and enhance category discrimination. For one SE attention block, we first squeeze the given feature maps $Z \in \mathbb{R}^{W\times H\times C}$ by a squeezer $s : S = s(Z)$, where $S \in \mathbb{R}^{1\times 1\times C}$ is the output vector and s is the squeeze operation, e.g., average pooling and max-pooling. Then, we feed the output S into an exciter $e : E = e(S,u) = \sigma(conv_1\rho(conv_2S))$, where $E \in \mathbb{R}^{1\times 1\times C}$ is the output of the excitation operation, u denotes the scaling factor of channels and is set to 4 in our method, ρ is a ReLU function, σ is a sigmoid function, and $conv_1$ and $conv_2$ indicate 1×1 convolution layers to rescale the channel number. Next, we can obtain the final output feature maps $U \in \mathbb{R}^{W\times H\times C}$ of one attention block by a fuser $f : U = f(E,Z) = (E \otimes Z) \oplus Z$, where $\otimes$ represents element-wise multiplication and $\oplus$ denotes the element-wise summation operation. In addition, to deeply mine the channel relationships, we connect the adjacent attention blocks like a chain by a connector $c : c(\alpha S, \beta \hat{E}) = softmax(\alpha S, \beta \hat{E})$, where $\hat{E}$ denotes the output vector of the exciter in the previous attention block, α and β are learnable parameters. Finally, we can obtain the final chain channel attention optimized output feature map $O_c \in \mathbb{R}^{W\times H\times C}$ as follows:

$$O_c = f(E_c, Z_c) = f((e(c(\alpha S_c, \beta \hat{E}), u), Z_c) \tag{3}$$

where O_c, Z_c, E_c, and S_c represent the feature maps and vectors used in the final attention block for the classification prediction. Figure 3 illustrates a two attention block connected structure. In our method, we adopt a triple attention blocks-connected pattern for best performance.

2.3 IoU-Wise Module

Due to the cluttered arrangement of remote sensing targets, the bad positioning detection boxes which have not well enclosed the grounding truth will generate a larger IoU with adjacent objects and severely hinders detection precision improvement. Therefore, the quality of localization is also an important index affecting the remote sensing object detection. Therefore, in this article, we propose an IoU-wise module to tackle this problem. As shown in Fig. 1, different from FCOS, we design an IoU-wise branch, in parallel with the localization regression branch to predict the IoU of each regressed box. To reduce the computational complexity, the IoU-wise module is only composed of a 1 × 1 convolution, batch normalization, and sigmoid activation layer. The predicted IoU score P_{IoU} falls in the range of 0 to 1 and is trained with binary cross entropy. Finally, we can obtain the final prediction confidence P_{Fin} for the NMS and AP computational processes as follows:

$$P_{Fin} = 2/(\frac{1}{P_{IoU}} + \frac{1}{P_{Cls}}) \tag{4}$$

where P_{Cls} represents the classification score. Therefore, IM will filter out the low-quality boxes and retain the detection boxes with high IoU score and classification score.

3 Experiments and Result Analysis

3.1 Datesets

In this article, we conduct experiments on two remote sensing datasets: DOTA [21] and NWPU VHR-10 [22].

DOTA. Comprised of 15 categories targets, it has 1,411 images and 188,282 instances annotated with horizontal and oriented bounding boxes. In our experiments, only the annotations of the horizontal bounding boxes are utilized. In addition, the validation, test, and training images have corresponding ratios of 1/6, 1/3, and 1/2. Each image has a size within 800 × 800 to 4,000 × 4,000 pixels.

NWPU VHR-10. These images contain 10 geospatial object classes, having 3,775 target instances. 715 RGB images and 85 sharpened color infrared images are included. Among them, 715 RGB images were collected from Google Earth with spatial resolutions ranging from 0.5 m to 2 m, while 85 pan-sharpened infrared images with a spatial resolution of 0.08 m were collected from Vaihingen data. These object instances were manually labelled with horizontal borders by experts.

3.2 Experimental Details

In the experiments, ResNet-100, which is initialized with the weights pretrained on ImageNet, is used as the backbone network. We use stochastic gradient descent (SGD) to optimize the network and set the initial learning rate to 0.001. The learning rate is reduced by a factor of 1.8 every 20k iterations. In addition, the weight

Table 1. Comparisons on DOTA with the State-of-the-Art Detectors. We choose the threshold of IoU which is 0.5 during calculating AP

Method	Pl	Bd	Br	Gft	Sv	Lv	Sh	Tc	Bc	St	Sbf	Ra	Ha	Sp	He	mAP	FPS
SSD [3]	76.84	57.26	27.09	30.73	23.96	56.91	41.73	80.83	37.53	36.01	33.85	32.98	63.19	29.89	20.48	43.25	15.65
Faster R-CNN [1]	75.23	56.07	39.94	33.71	26.29	47.43	29.64	81.24	27.01	31.56	28.50	32.10	53.10	45.73	35.04	42.84	7.13
CornerNet [5]	87.33	66.23	34.32	52.98	65.93	72.11	74.02	88.93	72.53	74.02	41.77	55.23	50.98	54.18	24.99	61.09	1.13
FCOS [6]	88.53	68.21	34.78	58.33	65.98	74.22	76.34	88.34	74.18	76.59	44.88	57.23	52.90	57.92	23.01	62.76	16.85
P-RSDet [14]	89.02	73.65	47.33	72.03	70.58	73.71	72.76	90.82	80.12	81.32	59.45	57.87	60.79	65.21	52.59	69.82	12.78
CAD-Net [15]	87.80	82.40	49.40	**73.50**	71.10	63.50	76.70	90.90	79.20	73.30	48.44	60.93	62.08	67.02	62.23	69.90	6.82
RetinaNet [4]	88.92	67.67	33.55	56.83	66.11	73.28	75.24	90.87	73.95	75.07	43.77	56.72	51.05	55.86	21.46	62.02	7.34
IENet [16]	80.20	64.54	39.82	32.07	49.71	65.01	52.58	81.45	44.06	78.51	46.54	56.73	64.40	64.24	36.75	57.14	**17.05**
O^2-DNet [17]	89.91	82.14	47.33	67.21	71.32	74.03	78.62	90.76	82.23	81.36	60.93	60.17	58.21	66.98	61.03	71.04	15.28
SCRDet [2]	**90.18**	81.88	55.30	73.29	72.09	77.55	78.06	**90.91**	82.44	86.39	64.53	63.45	75.77	78.21	60.11	75.35	3.37
SAOA	88.33	**82.73**	**56.02**	71.58	**72.98**	**77.59**	**78.29**	88.63	**83.33**	**86.61**	**65.93**	**63.52**	**76.03**	**78.43**	**61.33**	**75.41**	15.65

The abbreviations are defined as: Pl: Plane, Bd: Baseball diamond, Br: Bridge, Gft: Ground field track, Sv: Small vehicle, Lv: Large vehicle, Sh:Ship, Tc: Tennis court, Bc: Basketball court, St: Storage tank, Sbf: Soccer-ball field, Ra: Roundabout, Ha: Harbor, Sp: Swimming pool, and He: Helicopter.

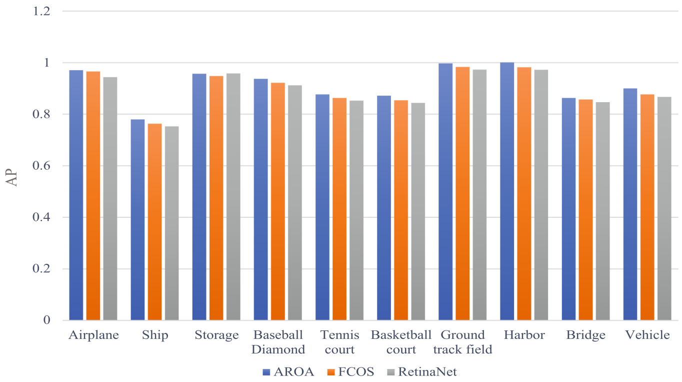


Fig. 4. Detection results of three one-stage detectors on NWPU VHR-10.

decay and momentum are set as 0.0001 and 0.9, respectively. The experiments are conducted in PyTorch1.1 using an NVIDIA GeForce RTX 2080 Ti 11 GB and CPUE52603v4@2.20 GHz CPU. We randomly divide the training and testing sets of NWPU VHR-10 according to the rate of 7:3. Additionally, we use the batch sizes of 10 for DOTA and 12 for the NWPU VHR-10 dataset. In our model, we set the vital parameters $\gamma = \mu = 4$ for best performance.

3.3 Experimental Results and Analysis

In this article, three frequently-used evaluation metrics, the average precision (AP) and the mean average precision (mAP), are adopted to evaluate the accuracy of the detectors in our experiments. In addition, we use a common metric, frames per second (FPS), to measure the speed of object detection.

Comparing with State-of-the-Art Detectors. To verify the effectiveness of our AROA, we compare it with several state-of-the-art detectors. As shown in Table 1,

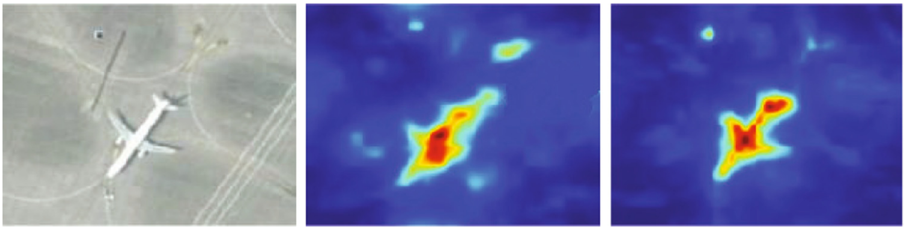
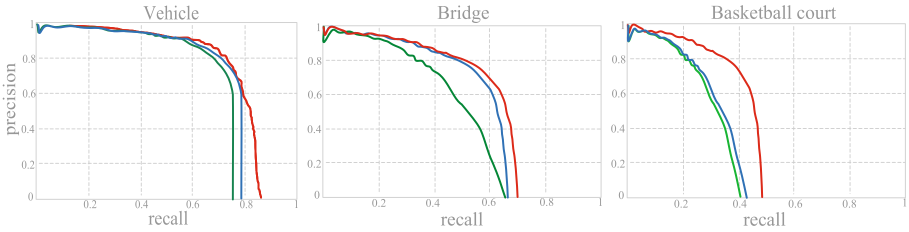


Fig. 5. Visualization of attention map of the AS^{2}A (middle) and C^{3}A (right).



(a) (b) (c)

Fig. 6. (a), (b), and (c) represent the precision-recall curves at IoU = {0.5, 0.75, 0.9} for vehicle, bridge, and basketball court. The red, blue, and green curves represent the results of AROA, FCOS, and RetinaNet, respectively. (Color figure online)

we compare our method with seven remote sensing detectors on DOTA. For a fair comparison, the backbone network of all compared methods is selected as ResNet-100 and the input image size of all methods is resized to a fixed scale (800 × 800). We also utilize data augmentation to enlarge the datasets such as flipping, rotation, and so on. The proposed AROA achieves 75.41% in mAP at a relatively fast speed with 15.65 FPS, outperforming all reported anchor-based detectors (SSD, Faster R-CNN, RetinaNet, CAD-Net, and SCRDet). It is about five as fast at almost the same accuracy compared with SCRDet, which manifests the high accuracy and efficiency of our method. In addition, Fig. 4 shows the detection results of FCOS, RetinaNet, and our AROA on NWPU VHR-10. The AP values of ten class objects on the proposed AROA are all higher than the other two one-stage detectors, which also demonstrates the advantage of our method for remote sensing object detection. Meanwhile, we perform the comparison experiments on three multi-orientation, high inter-class similarity, and low-discernibility targets (vehicle, basketball court, and bridge). Figure 6 shows the precision-recall curves of AROA, FCOS and RetineNet. We can see that our AROA achieves better performance than FCOS and RetinaNet. Moreover, it is remarkable that with a stricter IoU threshold, AROA takes a larger improvement over FCOS and RetinaNet, which indicates that AROA can locate the objects more accurately. As shown in Fig. 5, we present the partial attention maps of the AS^{2}A and the C^{3}A. We observe that the AS^{2}A effectively enhances the object information, especially for the orientation information. Meanwhile, the C^{3}A improves the classification ability by generating more discriminative feature maps.

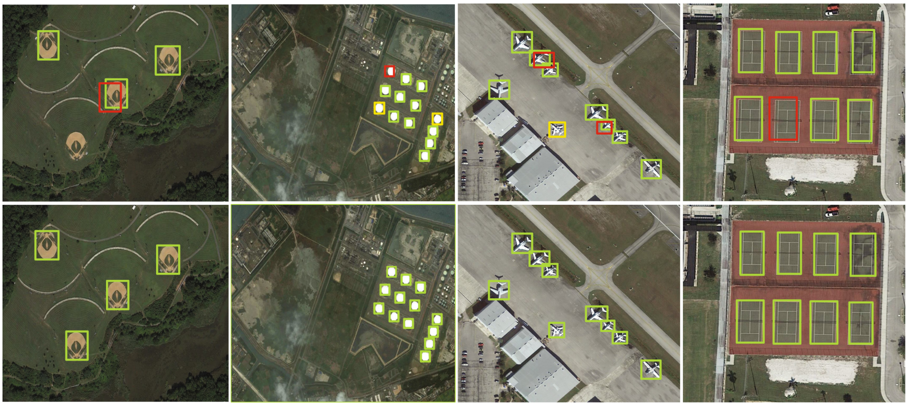


Fig. 7. The green, red, and yellow bounding boxes represent the true positives, false positives, and true negatives, respectively (Color figure online)

Table 2. Ablation study of the proposed AROA on NWPU VHR-10.

Setting	mAP	FPS
Baseline	89.21%	17.13
Baseline + C^{3}A	90.36%	16.98
Baseline + AS^{2}A	90.79%	16.71
Baseline + IM	90.56%	16.89
Baseline + C^{3}A + IM	92.91%	16.61
Baseline + C^{3}A + AS^{2}A	92.12%	15.83
Baseline + AS^{2}A + IM	92.65%	16.04
AROA	93.45%	15.65

Ablation Study. We perform an ablation study to study the effects of the proposed AS^{2}A, C^{3}A and IM over the NWPU VHR-10 dataset. Eight network models are trained for the ablation study. First, **Baseline:** The basic network structure in FCOS does not include the centerness branch. Next, three network models which add the components of AS^{2}A, C^{3}A, and IM, respectively. Then, **AROA:** The full implementation of the proposed AROA architecture in Fig. 1. In addition, three network models which equip with paired components are included to verify the complementarity of the three proposed modules. As shown in Table 2, the Baseline model can only achieve mAP of 89.21% at 17.13 FPS. After adding AS^{2}A, C^{3}A and IM to the Baseline model sequentially, the mAP is improved by 1.15%, 1.58%, and 1.35%, respectively. The same situation occurs in three models with different paired components, which illustrates that three improvement modules are complementary. Finally, the proposed AROA that combines the AS^{2}A, C^{3}A and IM achieves a 4.24% mAP improvement compared to the Baseline model, pushing the mAP to 93.45% with a relatively little additional computational cost (a decrease of 1.48 FPS). As shown in Fig. 7, some targets that are not detected or imprecisely located with the **Baseline** (the first row) are accurately detected with our **AROA** (the second row).

4 Conclusion

In this article, the challenge issues of complicated background, rotation variations, and interclass similarity are identified in remote sensing object detection. Aimed at these problems, we presents a novel attention refinement one-stage anchor-free detector. The core components contain asymmetric spatial self-attention and chain-connected channel attention which are designed to refine spatial and channel feature expression. Moreover, an IoU-wise module is proposed to strengthen the correlation between localization and classification for accurate object detection. Extensive experiments demonstrate the superiority of our method for object detection in remote sensing images. In future work, we will design more auxiliary modules to achieve the detection of oriented targets in remote sensing images.

Acknowledgment. This work was supported in part by the National Natural Science Foundation of China under Grant 61701524, 62006245 and in part by the China Postdoctoral Science Foundation under Grant 2019M653742. Thanks to the anonymous reviewers for their valuable suggestions.

References

1. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **2015**, 91–99 (2015)
2. Yang, X., et al.: SCRDet: towards more robust detection for small, cluttered and rotated objects. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 8232–8241 (2019)
3. Liu, W., et al.: SSD: Single Shot MultiBox Detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
4. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
5. Law, H., Deng, J.: CornerNet: detecting objects as paired keypoints. In: Proceedings of the European Conference on Computer Vision, pp. 734–750 (2018)
6. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9627–9636 (2019)
7. Wei, H., Zhang, Y., Wang, B., Yang, Y., Li, H., Wang, H.: X-LineNet: detecting aircraft in remote sensing images by a pair of intersecting line segments. IEEE Trans. Geosci. Remote Sens. **9**(2), 1645–1659 (2020)
8. Chen, J., Xie, F., Lu, Y., Jiang, Z.: Finding arbitrary-oriented ships from remote sensing images using corner detection. IEEE Geosci. Remote Sens. Lett. **17**(10), 1712–1716 (2019)
9. Shi, F., Zhang, T., Zhang, T.: Orientation-aware vehicle detection in aerial images via an anchor-free object detection approach. IEEE Trans. Geosci. Remote Sens. **59**(6), 5221–5233 (2020)

10. Li, Y., Huang, Q., Pei, X., Chen, Y., Jiao, L., Shang, R.: Cross-layer attention network for small object detection in remote sensing imagery. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. 14, 2148–2161 (2020)
11. Wang, J., Wang, Y., Wu, Y., Zhang, K., Wang, Q.: FRPNet: a feature-reflowing pyramid network for object detection of remote sensing images. IEEE Geosci. Remote Sens. Lett. (2020)
12. Wu, Y., Zhang, K., Wang, J., Wang, Y., Wang, Q., Li, Q.: CDD-Net: a context-driven detection network for multiclass object detection. IEEE Geosci. and Remote Sens. Lett. (2020)
13. Zhang, X., Wang, G., Zhu, P., Zhang, T., Li, C., Jiao, L.: GRS-Det: an anchor-free rotation ship detector based on Gaussian-mask in remote sensing images. IEEE Trans. Geosci. Remote Sens. **59**(4), 3518–3531 (2020)
14. Zhou, L., Wei, H., Li, H., Zhao, W., Zhang, Y., Zhang, Y.: Arbitrary-oriented object detection in remote sensing images based on polar coordinates. IEEE Access **8**, 223373–223384 (2020)
15. Zhang, G., Lu, S., Zhang, W.: CAD-Net: a context-aware detection network for objects in remote sensing imagery. IEEE Trans. Geosci. Remote Sens. **57**(12), 10015–10024 (2019)
16. Lin, Y., Feng, P., Guan, J.: IENet: interacting embranchment one stage anchor free detector for orientation aerial object detection (2019). arXiv:1912.00969
17. Wei, H., Zhang, Y., Chang, Z., Li, H., Wang, H., Sun, X.: Oriented objects as pairs of middle lines. ISPRS-J. Photogramm. Remote Sens. **169**, 268–279 (2020)
18. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points (2019). arXiv:1904.07850
19. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)
20. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
21. Xia, G., et al.: DOTA: a large-scale dataset for object detection in aerial images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3974–3983 (2018)
22. Cheng, G., Han, J.: A survey on object detection in optical remote sensing images. ISPRS-J. Photogramm. Remote Sens. **117**, 11–28 (2016)