An Object Detection Method for Remote Sensing Images Based on DA-YOLO

Ruizhe Hu and Rui Ting(✉)

Army Engineering University of PLA, Nanjing, China
516252190@qq.com

Abstract. Aiming at the difficulty of small-scale objects in high-resolution remote sensing images, this paper proposes a new detector DA-YOLO (dilation and attention YOLO) to locate objects quickly and accurately. Firstly, during the data preprocessing, the remote sensing images processed by "quadruple cropping" to adjust the original image size and enlarge the number of data instance. Then, the CSPDarknet53 backbone network is optimized: the dilated separable convolution (DSC) module is applied to enlarge the receptive range of feature maps without losing the resolution of feature maps. Then, the convolutional block attention module (CBAM) is introduced for feature enhancement, and finally, the last four stages of feature maps are used instead of three stages to obtain more contour details of small-scale objects. Extensive experiments show that DA-YOLO has good performance in DOTA, with a 2.36% increase in mAP compared to the original YOLOv4 without a significant decrease in detection speed.

Keywords: Dilated separable convolution · Attention mechanism · Multi-scale · Object detection · Remote sensing image

1 Introduction

Remote sensing technology is widely used in traffic monitoring, military reconnaissance and other fields, and remote sensing image object detection technology has gradually become a research hotspot in computer vision [1–3]. The basic task of object detection is to determine the class of each object and provide their boundaries. Due to the particularity of the location of the remote sensing observation platform, the remote sensing images often contain many complex backgrounds, and the wide variety of objects, different scales, and unstable shapes, which make the general object detectors unsatisfactory in remote sensing images. Furthermore, the scale problem results in poor representation of the features of objects in remote images by DNNs (Deep Neural Networks).

Facing the challenges of remote sensing imagery mentioned above, we propose DA-YOLO based on YOLOv4, which is used for multi-category, multi-scale, and multi-pose small-scale object detection in remote sensing images. The contributions of this paper mainly include the following four aspects:

P. You et al. (Eds.): ICIVIS 2022, LNEE 1019, pp. 119–130, 2023.
https://doi.org/10.1007/978-981-99-0923-0_13

(1) The data augmentation method of "quadruple cropping" is adopted. It is guaranteed that the small target information will not be lost when resizing the original image, and the number of instances is enlarged.
(2) DSC module is introduced. DSC module expands the receptive field and improves the feature extraction capability of small objects.
(3) CBAM [4] is introduced. CBAM aggregates global and local features and establishes long-term dependencies between channel attention, improving the representation of small objects.
(4) Experiments show that DA-YOLO outperforms the DOTA dataset [5], increasing mAP by 1.36% without a significant drop in speed.

2 Related Work

Affected by the acquisition method, the size of remote sensing image is far larger than the image size of general object detection dataset. The factors like density, scale and scene complexity should be considered in the impact results, causing the more difficulty for detection in remote sensing images. Especially for the detection of small-scale objects in high-resolution images, the accuracy is challenged. Relevant works with DNN have been devoted to the application of remote sensing images detection. Fan et al. [6] propose ClusDet that produces object cluster regions and estimates object scales for these regions, which greatly reduces the number of chips for final object detection. Yang et al. [7] adopt an objected feature fusion strategy which fully considers feature fusion, anchor sampling, and receptive field. Zhang et al. [8] introduces a feature enhancement method which learns global and local contexts together.

Input images are enormous while objects have less pixels. Resizing to the input size directly is not an optimal option. It causes an information loss if objects have only a few pixels. Mate Kisantal [9] used multiple copy and paste of small objects to enhance the image to reduce the loss of object information, but it still has limitations for dense objects. The algorithm in [10] proposes a scale adaptive proposal network to improve the precision of multi-object detection, but the detection efficiency of small objects is generally not high.

The receptive fields in deep learning refers to the area size of the pixels mapped on the original input image on each output characteristic map of CNN. In order to reduce the loss of effective object information and improve the detection accuracy of small objects, many scholars have done corresponding research. Gan et al. [11] applied FPN network to remote sensing image object detection, which improved the accuracy of the network for small-scale objects, but only detected specific objects. Qu et al. [12] used dilated convolution to enhance the receptive field of the third-level features in the network and enrich the detailed information of the object. Dilated convolution using sparse kernels is a better choice for alternating the convolutional and can flexibly aggregate context information while maintaining the same resolution [13]. Therefore, we use dilated convolution to maintain the size of feature map could improve the feature extraction ability of network to get more object information.

In addition, small-scale objects are more dependent on shallow level features. Fu et al. [14] proposed a feature fusion architecture to generate multi-scale feature levels

and combine features of different levels to form a powerful representation of object features. On the basis of YOLOv4 three-scale detection, we add a detection branch, and convey positioning and semantic information through the path aggregation network, so as to obtain richer texture and contour information, and improve the detection effect of small-scale objects.

3 Our Network

In this paper, we propose a fast and accurate small-scale object detection method for remote sensing images. As shown in Fig. 1, given an image as input, we first resize the image by "quadruple cropping". The cut images are then input into an improved backbone network based on CSPDarknet53 [15] to extract deep features. We add four DSC modules, which not only increase the size of the receptive field, but also preserve the resolution of the image without losing information. Then, we introduce the CBAM for feature enhancement. In addition, we also adopt the last four stages of feature maps instead of three stages to obtain more contour details of small-scale objects. Finally, the predicted bounding boxes are aggregated and redundant detections are removed by Non-Maximum Suppression (NMS) in the final detection results.

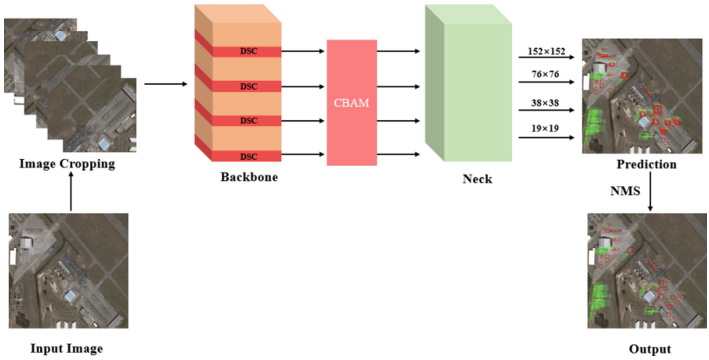
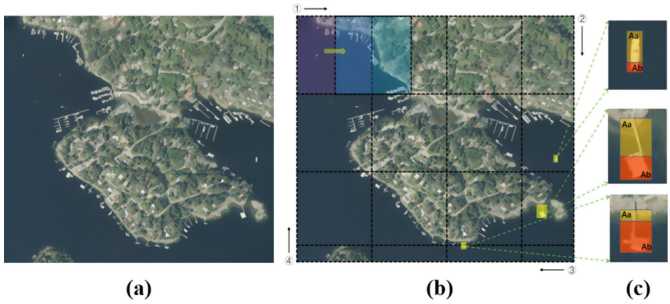


Fig. 1. Network architecture.

3.1 Quadruple Cropping

Sending the large-scale images directly for training to excessive compression and could not make the network training converge. Due to the objects in remote sensing images loss mostly information after over-scaling from the image preprocessing. Given this situation, this paper proposed a novel data cropping method named "quadruple cropping" by referring to References [5, 16]. The crop image is shown in Fig. 2.

Figure 2(a) is the original image. As shown in Fig. 2(b), the original image is cropped in four directions from ① to ④ with an overlap rate of 50%, and the image is cropped to a size of 800 × 800. If the width or the height of the original images is less than 800, there



(a) (b) (c)

Fig. 2. Quadruple cropping.

will not be cropping for the horizontal or vertical direction. Besides, if the remaining size is less than 800 after cropping, this part of the image will be abandoned.

In addition, for the object on the clipping boundary, the object will be damaged when clipping (Fig. 2(c)), if the object is abandoned, lots of object information will be lost, if all the object information is marked, it will cause missed detection and false detection due to insufficient object information. In this case, this paper uses the method of calculating the object incompleteness to determine whether to retain the object label, and the value calculation is shown in formula (1):

$$P = \frac{A_a}{A_b} \tag{1}$$

where A_a, A_b respectively represent the area of the labeled frame in the original image and the cropped image. If $P \geq 0.7$, it means that the sub-image contains more incomplete object information, so the coordinate will be completely retained. If $0.3 \leq P \leq 0.7$, the coordinate will be still retained and set 'difficult' in the label file to 1. If $P \leq 0.3$, we remove the annotation of the object. "Quadruple cropping" can ensure the retention of the object information in the greatest extent and increase the sample diversity after cropping, which eliminates the increase of detection error caused by less object and more background information.

3.2 Network Architecture

As one of the most advanced algorithms, YOLOv4 excellent for speed and precision. Its backbone network CSPDarknet53 adds CSPNet (cross-stage partial network) to each large residual block of Darknet53, and fuses it into the feature map through gradient changes. In the neck, PANet [17] (path aggregation network) with a more flexible ROI pooling is applied to shorten the path from up to bottom fusion. Inspired by this latest research algorithm, we employ an improved CSPDarket53 as the backbone, which has a good performance in extracting small-scale object features, especially suitable for object detection in remote sensing images.

DSC Module. Dilated convolution was first applied in the field of image segmentation, which can increase the perception range of the feature map without increasing the amount

of additional calculation. As shown in Fig. 3(a) is the receptive field of the standard convolution kernel, which is only 3 × 3; (b) is the receptive field when the expansion rate is 2, and the mapping range increases as the convolution kernel is zero-filled. The receptive field has changed from 3 × 3 to 7 × 7, and each convolution output contains a larger range of feature information. Inspired by the Inception network structure [18], we propose a dilated separable convolution module based on dilated convolution. As shown in Fig. 4, the DSC module first separates the channels into two groups, and combines dilated convolutions with different size dilated ratios (D_1, D_2) in each branch to effectively obtain multi-scale information. Finally, perform the *Add* operation on the original input and output to obtain the final output.

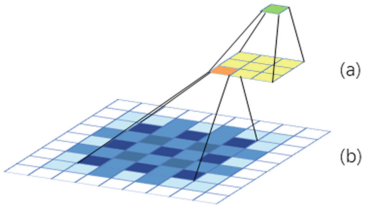


Fig. 3. Receptive field diagram.

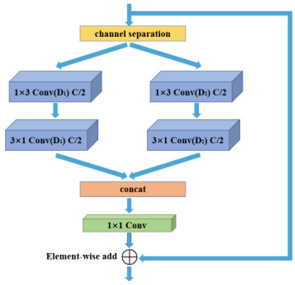


Fig. 4. DSC module.

Yolov4 integrates different scale features to detect objects. So, each layer of feature map will contain low-level features of the shallow layer and deep-seated high-level features, and the precision of prediction will be improved. But when the deep CNN extracts feature of different scales in remote sensing images, the pooling layer will reduce the resolution of the feature map, making it difficult for the receptive field that is too small to detect object information that only occupies a few dozen pixels. DSC module can expand the receiving field of the feature map while maintaining the resolution, so that each convolution output contains a csope range of feature information. Therefore, this paper applies the DSC module to the backbone network to expand the effective receptive field of the feature map while preserving the image resolution.

CBAM. The attention mechanism mainly emphasizes the importance of different characteristics by assigning weights to features. It mainly imitates the law of human brain

observation activities, and assigns more important attention to the target points on the image that need to be paid attention to through weight assignment. [19], which highlights the features of the image target and suppresses the feature information of other objects to achieve feature information enhancement. In order to better aggregate the local features and global features of the image, so that the extracted feature information can better characterize the location and category of the image target, CBAM [4] is introduced to enhance the feature maps of different scales extracted by the backbone network respectively. CBAM [4] is shown in Fig. 5. CBAM [4] is mainly includes channel attention module (CAM) and spatial attention module (SAM).

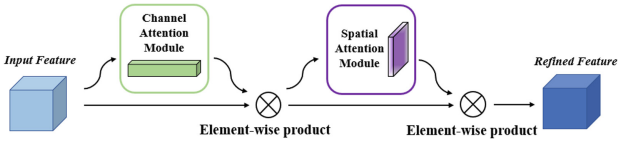


Fig. 5. The structure of CBAM.

CAM. After the image is subjected to the convolution operation, each channel of the feature map often expresses different features, and the features of each channel are different. However, each channel of the feature map maintains the same weight, and the importance relationship between each channel is not considered, which is not conducive to enhancing the feature information of the target. SAM can get the importance of different channels of the feature map, and assign corresponding weights to each channel, which can make better use of features with high weights and suppress features with low weights, and enhance the expression between features. The CAM structure is shown in Fig. 6.

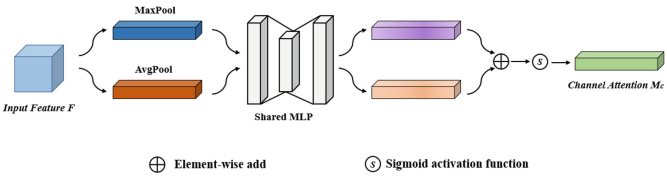


Fig. 6. The structure of CAM.

SAM. Different from CAM, SAM focuses more on spatial location information. SAM can obtain the importance of the spatial location information of the feature map and highlight the location of key features, thereby enhancing the representation capability of the feature map. The SAM structure is shown in Fig. 7.

Multi-scale Prediction. For the detection of small-scale objects, it is often more dependent on shallow features. The three-scale in YOLOv4 is not enough to fully obtain the subtle feature information of the object. Therefore, the original three-scale is expanded

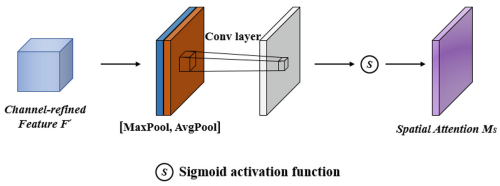


Fig. 7. The structure of SAM.

to four-scale, a more accurate anchor frame is assigned to the object on the larger feature map, and the feature maps from different information streams are effectively combined to gradually transfer the information in the low-level feature map to at the high level, the feature information of the object is continuously enriched and improved, the semantic information is more complete, and richer texture and contour information can be get, which effectively improves the detection effect of small-scale objects. The improved network structure is shown in Fig. 8.

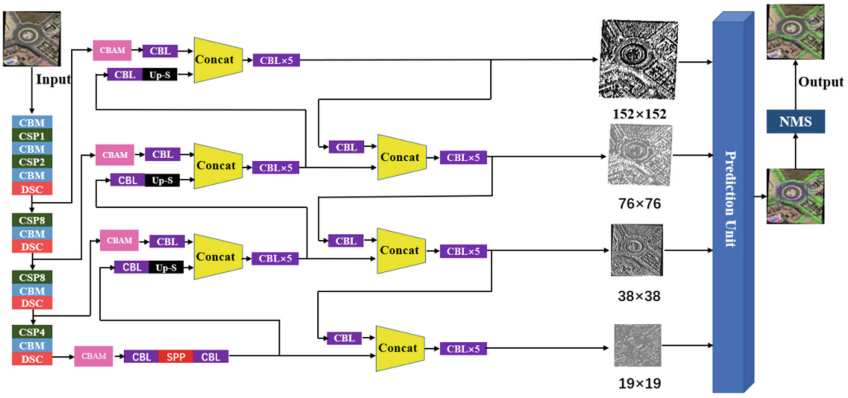


Fig. 8. This article network architecture.

4 Experiment and Analysis

Our detector is trained on 1 Titan 2080Ti GPU, optimized by Adam with the momentum of 0.0005 and weight decay of 0.9. The batch-size is 8 due to limitation of GPU memory. The learning rate is 0.001 initially. This article is trained and tested on the DOTA [5] dataset. DOTA [5] includes 2806 remote sensing images, with 1/6 is validation, and 1/3 is test. The image size is in the range of 800 × 800 to 4000 × 4000, of which 15 categories total 188282 instances. This paper uses the quadruple cropping method to crop 1411 training pictures into 30749 sub-images for network training.

4.1 Ablation Study

In this section, a series of ablation experiments were conducted on the validation set of DOTA [5] to evaluate the effect of each improvement. We apply the CSPDarknet53 as our baseline and modify components gradually to find the final appropriate settings. Table 1 shows the results of ablation studies.

Table 1. Comparison of ablation experiment results.

Component	CSPDarknet53				
Image cropping		√	√	√	√
DSC			√	√	√
CBAM				√	√
Four scale outputs					√
mAP (%)	73.14	73.52	74.01	75.13	75.50

In Table 1, while keeping other settings unchanged, only the data cropping method is introduced, which can effectively reduce the problem of information loss caused by excessive image compression, and mAP is increased by 0.38%. Based on data cropping, by adding the DSC module, the network can learn high-level semantic information more effectively and improve the recognition ability of small-scale objects. By introducing CBAM, the ability to recognize small targets is further improved. Through the feature maps of four scales, the texture and contour information of the objects are enriched, so that the mAP reaches 74.01%, 75.12% and 75.50%, which are increased by 0.49%, 1.12% and 0.37% respectively.

4.2 Comparison with Other Methods

We conduct a comparison experiment with SOTA object detectors in DOTA. Including two-stage algorithms FR-H [20], ICN [1], SCRDet [3], FADet [21] and single-stage algorithms SSD [22], EFR [23], etc. The proposed method achieves superior performances and outperforms other detectors. The effect comparison is shown in Table 2.

Table 2. Comparison of test results on the DOTA-v1.0 dataset.

Method	FR-H [20]	SSD [22]	FPN [24]	ICN [1]	EFR [23]	SCRDet [3]	FADet [21]	DA-YOLO
PL	80.32	57.85	88.70	90.00	88.36	90.18	90.15	**90.66**
BD	77.55	32.79	75.10	77.70	**83.90**	81.88	78.60	83.51
BR	32.86	16.14	52.60	53.40	45.78	**55.30**	51.92	51.42
GTF	68.13	18.67	59.20	73.30	67.24	73.29	**75.23**	69.20
SV	53.66	0.05	69.40	73.50	76.80	72.09	73.60	**78.40**

(*continued*)

Table 2. (*continued*)

Method	FR-H [20]	SSD [22]	FPN [24]	ICN [1]	EFR [23]	SCRDet [3]	FADet [21]	DA-YOLO
LV	52.49	36.93	78.80	65.00	77.15	77.65	71.27	**79.17**
SH	50.04	24.74	84.50	78.20	**85.35**	78.06	81.41	82.83
TC	90.41	81.16	90.60	90.80	90.77	90.51	**90.85**	87.94
BC	75.05	25.10	81.30	79.10	**85.55**	82.44	83.94	79.81
ST	59.59	47.47	82.60	84.80	75.77	86.39	84.77	**87.15**
SBF	57.00	11.22	52.50	57.20	54.64	**64.53**	58.91	61.32
RA	49.81	31.53	62.10	62.10	60.76	63.45	**65.65**	59.48
HA	61.69	14.12	76.60	73.50	71.40	75.77	**76.92**	75.34
SP	56.46	9.09	66.30	70.20	77.99	78.21	**79.36**	79.30
HC	41.85	0	60.10	58.10	60.94	60.11	**68.17**	67.05
mAP	60.46	29.86	72.00	72.50	73.49	75.32	75.38	**75.50**

As shown in Table 2, DA-YOLO shows competitive performance at 75.50%. For small objects in the image, such as small vehicle, storage tank, our method has a clear improvement in detection accuracy. The excellent performance for detecting small objects is attributed to the enlarged receptive field, the introduction of attention mechanism and the enrichment of texture and contour information. The table also shows that the two-stage detector still achieves superior performances in DOTA [5] research. However, they all use complex model structures to improve accuracy, which extremely slows down detection speed. The proposed single-stage detector achieves comparable performance with other two-stage detectors while keeping a fast detection speed.

4.3 Speed Experiment

In order to test the real-time performance of the DA-YOLO, we used the DOTA [5] dataset to compare with other algorithms in terms of inference speed. The comparison results are shown in Table 3.

Table 3. Comparison of the speed of different methods on the DOTA dataset.

Method	Backbone	Image size	Speed/fps
YOLO v2 [25]	Darknet19	416 × 416	19.8
SSD [22]	VGG16	512 × 512	24.6
RRPN [26]	Darknet19	800 × 800	3.5
RetinaNet-H [27]	ResNet101	800 × 800	14
R3Det [28]	ResNet101	800 × 800	13
R2CNN [29]	ResNet101	800 × 800	2
DA-YOLO	CSPDarknet53	800 × 800	17.2

As shown in the Table 3. Because multi-scale prediction will make the network more complicated, and the image cropping network will take more inference time, the detection speed is slightly lower than that of the one-stage detection algorithms such as SSD [22], but the fps has also reached 17.2. Compared with the two-stage detection method, our DA-YOLO still has great competitiveness in detection speed. In summary, the one-stage detection algorithm proposed in this paper has good detection performance while maintaining rapidity.

4.4 Detection Result

Figure 9 Respectively show the detection visualization results of DA-YOLO on the DOTA dataset [5].

Fig. 9. This article network architecture.

As shown in Fig. 9, No matter on the DOTA dataset [5], DA-YOLO can give better detection results and show better generalization ability. Especially for small-scale targets such as airplanes, oil tanks, and small cars, due to the full use of the shallow features of the network, the detection results can show better results.

5 Conclusion

Aiming at the problem of small-scale object detection in high-resolution remote sensing images, a novel and robust first-level target detector DA-YOLO is proposed. By cutting the aerial image, the image can retain most of the information of small objects; by introducing the DSC module, the receiving range of the feature map is expanded; by

introducing the CBAM [4] for feature enhancement, the expression of small objects is further improved. In addition, the last four scales of the feature map are used to obtain more texture and contour information of small objects. Experiments show that DA-YOLO has better detection performance, and can more accurately detect objects in aerial images for object detection without significantly reducing the detection speed.

References

1. Azimi, S.M., Vig, E., Bahmanyar, R., Körner, M., Reinartz, P.: Towards Multi-class Object Detection in Unconstrained Remote Sensing Imagery. In: Jawahar, C.V., Li, H., Mori, G., Schindler, K. (eds.) ACCV 2018. LNCS, vol. 11363, pp. 150–165. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20893-6_10
2. Dai, J., et al.: R-FCN: Object Detection via Region-based Fully Convolutional Networks. ArXiv abs/1605.06409 (2016)
3. Yang, X., et al.: SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 8231–8240. (2019)
4. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: Convolutional Block Attention Module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1
5. Guisong, X., et al.: DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3974–3983 (2018)
6. Yang, F., et al.: Clustered Object Detection in Aerial Images. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 8310–8319 (2019)
7. Yang, X., et al.: R2CNN++: Multi-Dimensional Attention Based Rotation Invariant Detector with Robust Anchor Strategy. ArXiv abs/1811.07126 (2018)
8. Gongjie, Z., et al.: CAD-Net: A Context-Aware Detection Network for Objects in Remote Sensing Imagery. IEEE Transactions on Geoscience and Remote Sensing 10015–10024 (2019)
9. Kisantal, M., et al.: Augmentation for small object detection. ArXiv abs/1902.07296 (2019)
10. Zhang, S., et al.: Scale Adaptive Proposal Network for Object Detection in Remote Sensing Images. IEEE Geoscience and Remote Sensing Letters, 864–868 (2019)
11. Gan, Y., et al.: Object Detection in Remote Sensing Images with Mask R-CNN (2020)
12. Junsuo, Q., et al.: Dilated convolution and feature fusion SSD network for small object detection in remote sensing images. IEEE Access, 82832–82843 (2020)
13. Yuhong, L., et al.: CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1091–1100 (2018)
14. Kun, F. et al.: Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. Isprs Journal of Photogrammetry and Remote Sensing 294–308 (2020)
15. Bochkovskiy, A., et al.: YOLOv4: Optimal Speed and Accuracy of Object Detection. ArXiv abs/2004.10934 (2020)
16. Etten, A.V.: You Only Look Twice: Rapid Multi-Scale Object Detection In Satellite Imagery. ArXiv abs/1805.09512 (2018)
17. Shu, L., et al.: Path aggregation network for instance segmentation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8759–8768 (2018)

18. Szegedy, C., Liu, W., Jia, Y., et al.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9 (2015)
19. Li, F., Feng, R., Han, W., et al.: Ensemble model with cascade attention mechanism for high-resolution remote sensing image scene classification. Opt. Express **28**(15), 22358–22387 (2020)
20. Shaoqing, R., et al.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 1137–1149 (2015)
21. Chengzheng, L., et al.: Feature-attentioned object detection in remote sensing imagery. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 3886–3890 (2019)
22. 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
23. Kun, F., et al.: Enhanced feature representation in detection for optical remote sensing images. Remote. Sens. 2095 (2019)
24. Lin, T.-Y., et al.: Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 936–944 (2017)
25. Redmon, J., Farhadi, A.: YOLO9000: Better, Faster, Stronger, pp. 6517–6525 (2017)
26. Nabati, R., Qi, H.: Rrpn: Radar region proposal network for object detection in autonomous vehicles. In: 2019 IEEE International Conference on Image Processing (ICIP). IEEE, pp. 3093–3097 (2019)
27. Lin, T.Y., Goyal, P., Girshick, R., et al.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, 2980–2988 (2017)
28. Yang, X., Liu, Q., Yan, J., et al.: R3det: Refined single-stage detector with feature refinement for rotating object. arXiv preprint arXiv:1908.05612 **2**(4), 2 (2019)
29. Jiang, Y., Zhu, X., Wang, X., et al.: R2CNN: Rotational region CNN for orientation robust scene text detection. arXiv preprint arXiv:1706.09579 (2017)