Accurate Oriented Instance Segmentation in Aerial Images

ZhenRong Zhang and Jun Du(✉)

University of Science and Technology of China, Hefei, China
zzr666@mail.ustc.edu.cn, jundu@ustc.edu.cn

Abstract. The dominant instance segmentation methods first detect the object with an axis-aligned box, then predict the foreground mask on each proposal. While in aerial images, methods detecting objects with axis-aligned boxes are unsuitable, since the orientation of objects is arbitrary. What's more, the RoI pooling step existed in these systems results in the loss of spatial details due to the feature warping and resizing, which will degrade the segmentation quality, especially for large elongated objects. In this paper, we propose a novel accurate oriented instance segmentation method, named Rotated Blend Mask R-CNN. We perform mask prediction in oriented bounding boxes and predict the final mask by combining instance-level information with lower-level fine-granularity information. The proposed method is evaluated on the iSAID dataset, and competitive outcomes show that our model achieves state-of-the-art. Code will be made available at https://github.com/ZZR8066/RotatedBlendMaskRCNN

Keywords: Aerial images · Oriented instance segmentation

1 Introduction

Instance segmentation in aerial images is important as it can be applied in many areas, such as precision agriculture, security, military reconnaissance, etc. Instance segmentation aims at predicting category labels of all objects of interest and localizing them in pixel-level masks. Recently, many powerful instance segmentation systems [1–3] have been proposed, but most of them are researched on natural scene datasets, such as MSCOCO [4], PASCAL-VOC [5], Cityscapes [6] etc. Compared with the above datasets, objects in aerial images occur in high density, arbitrary orientation, large ratios, and huge scale variation. Most of the recent aerial images datasets focus on object detection [7,8], few datasets [9] provide annotations for instance segmentation and typically focus on a single object category annotation. A large-scale Instance Segmentation in Aerial Images Dataset (iSAID) [10], which is far more comprehensive and suitable for real-world applications in aerial scenes, was proposed just recently. Due to the above reasons, instance segmentation in aerial images has not been well researched.

© Springer Nature Switzerland AG 2021
Y. Peng et al. (Eds.): ICIG 2021, LNCS 12888, pp. 160–170, 2021.
https://doi.org/10.1007/978-3-030-87355-4_14

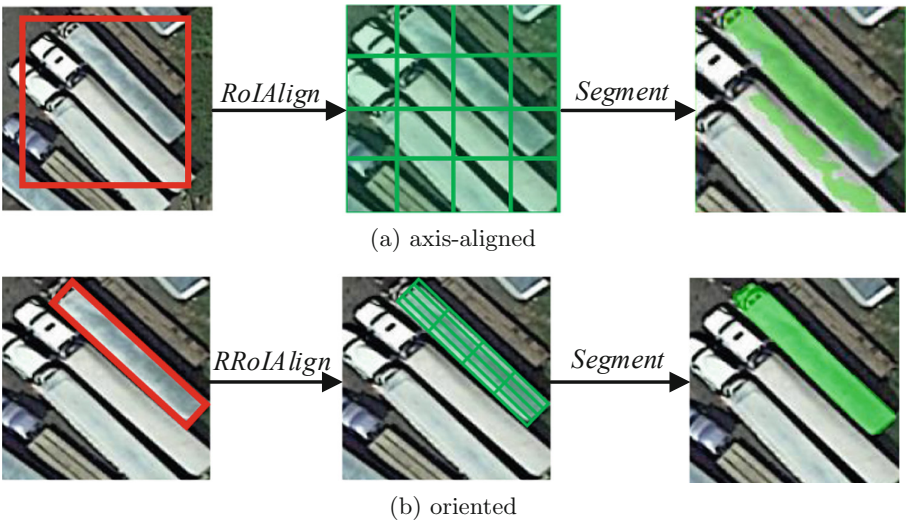


Fig. 1. The comparison between axis-aligned and oriented instance segmentation.

In this study, we take a full consideration on the aforementioned situation and propose a novel accurate oriented instance segmentation method, named Rotated Blend Mask R-CNN, which is based on the representative two-stage instance segmentation method Mask R-CNN [1]. Our method mainly consists of two parts, a detection network and a segmentation network. Specifically, in order to eliminate the ambiguity of axis-aligned boxes in densely packed objects, we achieve oriented bounding box regression in the detection network, which will generate more accurate mask prediction as shown in Fig. 1(b). As for the segmentation network, similar to [11], we improved mask prediction by effectively combining instance-level information with lower-level fine-granularity information, and we find that it can well process the situation for large elongate objects which are densely surrounded by objects of other categories.

The main contributions of our work are summarized as follows:

- We present a novel oriented instance segmentation method which predicts accurate instance masks based on oriented bounding boxes.
- Furthermore, we merge top-level coarse instance information with lower-level fine-granularity for describing the instance information within their best capacities.

2 Relate Work

2.1 Object Detection

The R-CNN [12] is a milestone for object detection method, many following methods [13–15] are based on it. SPPnet [13] removes crop/warp and other

operations on the original image and replaces it with a spatial pyramid pooling (SPP) layer on the convolutional features, which eliminates the requirement of a fixed-size input image and makes the system more robust to object deformations. Fast R-CNN [14] improves [12] training and testing speed by first processing the whole image with several convolutional and max pooling layers to produce a convolutional feature map and then extracting a fixed-length feature vector by a region of interest (RoI) pooling layer of each proposal. Faster R-CNN [15] introduces a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals. R-FCN [16] presents a position-sensitive RoI pooling to learn the location information of objects. Cascade R-CNN [17] increases the number of R-CNN to gradually generate better boxes.

The above detection methods are designed for the regression of axis-aligned bounding boxes, which are widely used in natural images. However, aerial images are taken from bird's-eye view, which implies that the orientation of objects is always arbitrary. Recently, many oriented bounding box regression methods have been proposed. Zhu et al. proposed Rotated Cascade R-CNN [18], which estimates the outline of the object in the first stage and regresses four vertices in the second stage. [19] proposed Adaptive Period Embedding (APE) to address the angular periodicity. Jian et al. proposed the RoI Transformer [20], which learns the spatial transformation parameters from the feature maps of axis-align RoIs and decodes them to generate oriented RoIs.

2.2 Instance Segmentation

He et al. proposed Mask R-CNN [1] which extends Faster R-CNN by adding a Fully Convolutional Network (FCN) [21] for predicting an object mask in parallel with the existing branch for bounding box recognition. The path aggregation network (PANet) [2], which won the COCO 2017 Challenge Instance Segmentation task, improves Mask R-CNN by bottom-up path augmentation, adaptive feature pooling and fully connected fusion. In order to calibrate the misalignment between the mask quality and the predicted score, the Mask Scoring R-CNN [3] proposed network block takes the instance feature and the corresponding predicted mask together to regress the mask IoU.

However, an underlying drawback in the above methods is that the RoI pooling step loses spatial details due to feature warping and resizing. Such distortion and fixed-size representation degrades the segmentation accuracy, especially for large objects. To address this issue, YOLACT [22] and YOLACT++ [23] accomplish this by breaking instance segmentation into two parallel subtasks, generating a set of prototype masks and predicting per-instance mask coefficient respectively, and producing instance masks by linearly combining the prototypes with the mask coefficients. Different from YOLACT, BlendMask [11], which outperforms Mask R-CNN in both mask AP and inference efficiency, merges top-level coarse instance information with lower-level fine-granularity to generate the final mask prediction.

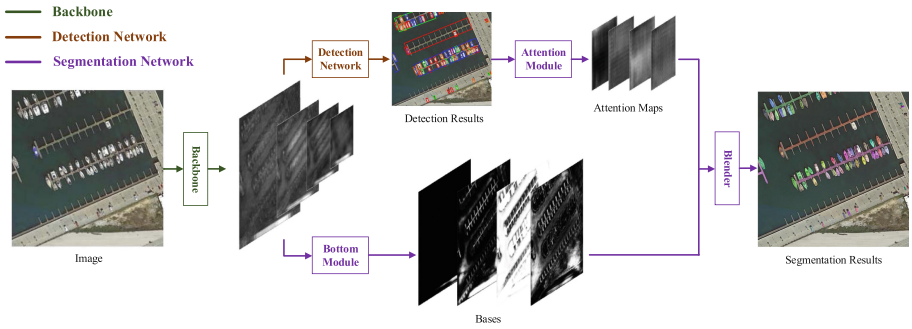


Fig. 2. The Rotated Blend Mask R-CNN architecture.

3 Method

In this section, we will present details of our proposed method, Rotated Blend Mask R-CNN, which is based on Mask R-CNN [1]. As shown in Fig. 2, the backbone is first applied to the input image to extract features that are used for the following detection and segmentation network. The detection network generates the oriented proposals, and the segmentation network will do the mask prediction of each proposal. Next we will elaborate on the above two networks respectively.

3.1 Detection Network

We detect objects in oriented boxes instead of axis-aligned boxes due to the orientation of densely packed objects as shown in Fig. 1. The detection network is illustrated in Fig. 3. We first obtain axis-aligned proposals from RPN [15], then for each proposal the RoIAlign [1] extracts a fixed-length feature vector from the feature map. Each feature vector is fed into a sequence of fully connected (FC) layers and outputs a five-dimensional vector $\boldsymbol{t}$ for each proposal. More specifically [20], each vector $\boldsymbol{t}$ consists of $(t_x, t_y, t_w, t_h, t_\theta)$ which are parameters of spatial transformation from axis-aligned boxes to oriented boxes and corresponding regression targets are:

$$\begin{aligned} t_x^* &= \tfrac{1}{w_r}\big((x^* - x_r)\cos\theta_r + (y^* - y_r)\sin\theta_r\big), \\ t_y^* &= \tfrac{1}{h_r}\big((y^* - y_r)\cos\theta_r - (x^* - x_r)\sin\theta_r\big), \\ t_w^* &= \log\tfrac{w^*}{w_r}, \quad t_h^* = \log\tfrac{h^*}{h_r}, \\ t_\theta^* &= \tfrac{1}{2\pi}\big((\theta^* - \theta_r) \bmod 2\pi\big), \end{aligned} \tag{1}$$

where $(x_r, y_r, w_r, h_r, \theta_r)$ is a stacked vector for representing location, width, height and orientation of an oriented proposal and $(x^*, y^*, w^*, h^*, \theta^*)$ is the ground truth parameters of an oriented bounding box. The corresponding oriented proposals can be obtained by decoding the vector $\boldsymbol{t}$. The Rotated RoIAlign

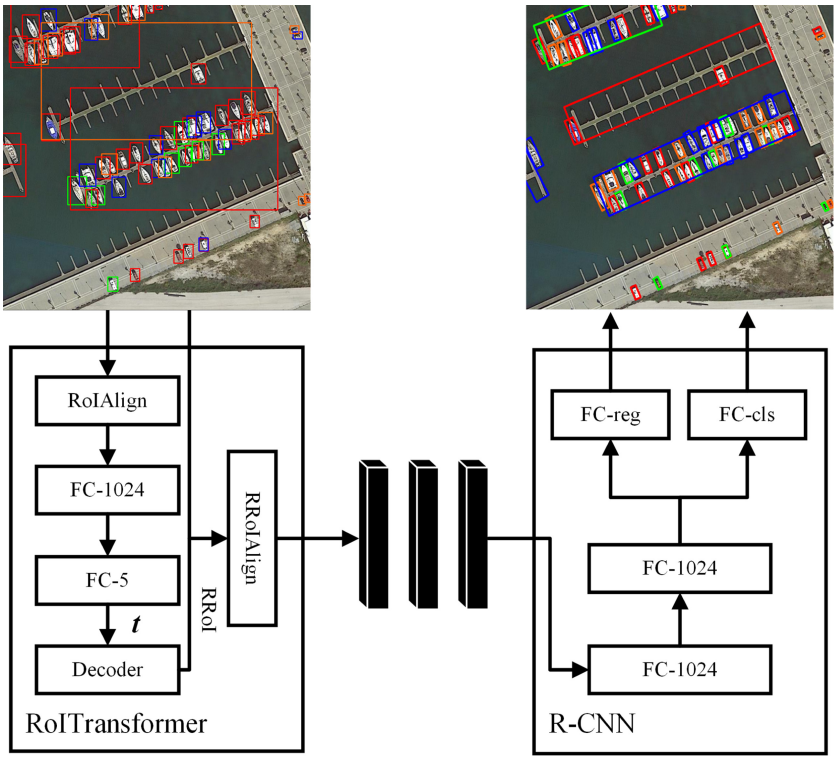


Fig. 3. The detection network.

(RRoIAlign) [18] is used to extract a fixed-length feature vector of each corresponding oriented proposal to maintain the rotation invariance. The final R-CNN stage fine-tunes oriented proposals and outputs D detection results $\mathbf{P}$. The regression targets of R-CNN are the same as Eq. (1). We use the Smooth L1 loss [12] function for the regression loss.

3.2 Segmentation Network

As shown in Fig. 2, the segmentation network is composed of three parts, a bottom module, an attention module and a blender. The bottom module aims at predicting the score maps containing semantic information with lower-level fine-granularity. The attention module predicts the attention maps in the instance-level of each oriented proposal. The blender module is used to merge the scores with attentions to generate the final mask predictions.

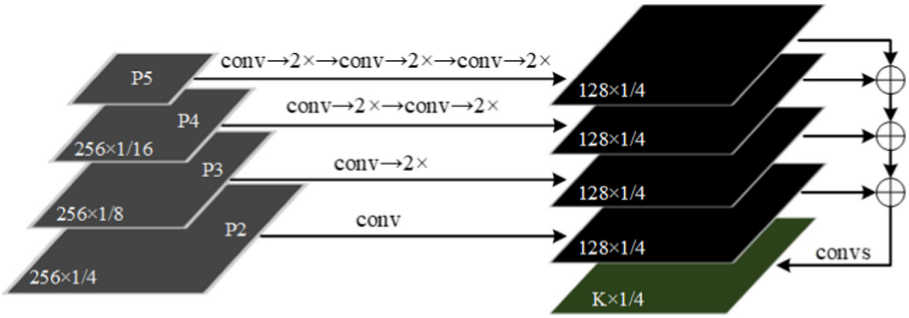


Fig. 4. The bottom module. Each conv is the convolution stage, and 2× is the upsampling stage.

Bottom Module. Figure 4 illustrates our bottom module in detail. We first obtain multi-level features {P2, P3, P4, P5} from the backbone, then perform upsampling stages for each level to yield the feature map at 1/4 scale. Each upsampling stage [24] consists of 3×3 convolution, group norm, ReLU, and 2× bilinear upsampling. The element-wise summation is applied to fused multi-level features. The final four convolution stages and the 1×1 convolution layer are used to generate score maps which are called bases [11], $\mathbf{B}$. $\mathbf{B}$ has a shape of $N \times K \times \frac{H}{s} \times \frac{W}{s}$, where N is the batch size, K is the number of bases, $H \times W$ is the input size and s is the output stride of score map, here s is 4.

Attention Module. After obtaining the oriented proposals from detection network, we use RRoIAlign to extract a fixed-size feature map of each oriented proposal, then predict attention maps $\mathbf{A}$ using an FCN [1,21]. Specifically, FCN has a shape of KM^2 dimensional output for each proposal, which encodes instance-level information into K maps with a resolution of $M \times M$.

Blender. The inputs of the blender module [11] contain bases $\mathbf{B}$, attention maps $\mathbf{A}$ and oriented proposals $\mathbf{P}$ from the detection network. We first use RRoIAlign to extract a fixed-size $R \times R$ feature map $\mathbf{r}_d$ for each proposal $\mathbf{p}_d$ from bases $\mathbf{B}$.

$$\mathbf{r}_d = \text{RRoIAlign}_{R \times R}(\mathbf{B}, \mathbf{p}_d), \quad \forall d \in \{1 \dots D\}. \tag{2}$$

Then we use the bilinear interpolation to resize attention maps $\mathbf{a}_d$ from $M \times M$ to $R \times R$ to ensure the sizes of $\mathbf{a}_d$ and $\mathbf{r}_d$ are the same.

$$\mathbf{a}'_d = \text{interpolate}_{M \times M \to R \times R}(\mathbf{a}_d), \quad \forall d \in \{1 \dots D\}. \tag{3}$$

$\mathbf{a}'_d$ is first normalized with softmax function along the K dimension to make it a set of score maps $\mathbf{s}_d$.

$$\mathbf{s}_d = \text{softmax}(\mathbf{a}'_d), \quad \forall d \in \{1 \dots D\}. \tag{4}$$

Then we apply element-wise product between each $\mathbf{r}_d$, $\mathbf{s}_d$, and sum along the K dimension to get the mask logit $\mathbf{m}_d$:

$$\mathbf{m}_d = \sum_{k=1}^{K} \mathbf{s}_d^k \circ \mathbf{r}_d^k, \quad \forall d \in \{1 \dots D\} \tag{5}$$

where k is the index of the bases.

4 Experiments

4.1 Datasets

iSAID [10] is a large-scale dataset for instance segmentation in aerial images, which contains 2806 aerial images from different sensors and platforms and comprises 655,451 annotated instances of 15 categories. Images with large resolutions (*e.g.* 4000 pixels in width) are commonly present in iSAID, it is necessary to crop the image and detect the objects in the cropped images. There are densely packed oriented objects such as large vehicles, small vehicles, and large elongated objects like harbors, which make segment objects in aerial images challenging.

It is worth to note that the detection network of our system performs oriented boxes regression, however, iSAID dataset does not provide the ground truth parameters of the oriented box $(x^*, y^*, w^*, h^*, \theta^*)$. Here, we use the smallest oriented bounding box of the instance mask as the regression target.

4.2 Implementation Details

The backbone of our detector is ResNet-50 [25] pre-trained on ImageNet [26]. The number of FPN channels is set to 256. Our network is trained with SGD, where the batchsize is 2 and the initial learning rate is set to 0.00125, which is then divided by 10 at $\frac{2}{3}$ and $\frac{8}{9}$ of the entire training. Due to the limited memory, we crop images to 800×800 with the stride of 200 for training and testing. The model is trained and tested at a single scale. By default, we train our model with training set and evaluate it on validation and testing set. Since our detection boxes are oriented, quadrilateral non-maximum suppression with the threshold of 0.3 is used during evaluation. As some configurable parameters of our segmentation network have been comprehensively researched in [11], unless otherwise stated, we take $K = 4$, $M = 28$, $R = 56$ as default.

4.3 Ablation Study

To have a fair comparison, we conduct ablation experiments based on mmdetection [27] framework to evaluate the effect of each component on the validation set of iSAID. The model is not modified except the component being tested.

Table 1. Instance segmentation results using mask mAP on iSAID validation set. Note that ADN means the axis-aligned detection network.

Detection		Segmentation		mAP
ADN	Ours	FCN	Ours	
✓		✓		33.5
	✓	✓		34.0
✓			✓	33.9
	✓		✓	34.4

(a) ADN (b) Ours

Fig. 5. The detection network comparison between ADN and ours. Both are using FCN as segmentation network. (Color figure online)

The Effect of Detection Network. When detecting densely packed oriented objects in aerial images, axis-aligned proposals often contain other instances, which will finally affect the instance segmentation results as shown in Fig. 1. Moreover, the oriented boxes generally contain much less background than the axis-aligned boxes as shown in Fig. 5, which makes the given resolution will be used much more efficiently. To evaluate whether the proposed detection network can handle well the above situation, we conduct ablation experiments as shown in Table 1. No matter what segmentation network is used, when the oriented detection network is used instead of the axis-aligned detection network, the performance will be improved to a certain extent. We also show the comparison in Fig. 5, where in ADN the mask prediction (red part) is not only on the target object, but also on another instance area in the detection result (green box). The background and other objects can be well removed when applying our proposed detection network, which will improve the accuracy of instance segmentation.

Table 2. Class-wise instance segmentation results on iSAID test set. Note that short names are used to define categories: BD-Baseball diamond, GTF-Ground field track, SV-Small vehicle, LV-Large vehicle TC-Tennis court, BC-Basketball court, SC-Storage tank, SBF-Soccer-ball field, RA-Roundabout, SP-Swimming pool, and HC-Helicopter.

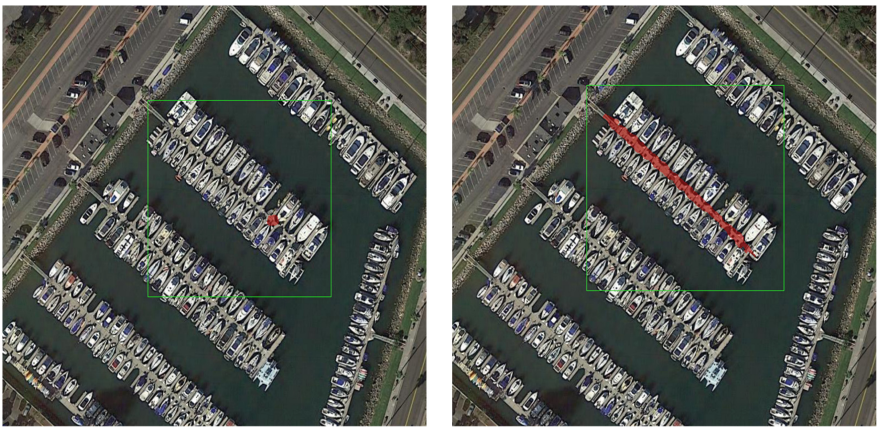
Method	Mask R-CNN [10]	PANet [10]	Ours
Plane	37.7	39.2	40.0
BD	42.5	45.5	51.6
Bridge	13.0	15.1	17.3
GTF	23.6	29.3	27.7
SV	6.9	15.0	13.1
LV	7.4	28.8	**29.6**
Ship	26.6	45.9	44.5
TC	54.9	74.1	74.8
BC	34.6	47.4	48.7
ST	28.0	29.6	34.3
SBF	20.8	33.9	33.7
RA	35.9	36.9	41.2
Harbor	22.5	26.3	**30.4**
SP	25.1	36.1	13.1
HC	5.3	9.5	14.9
mAP	25.7	34.2	35.8

The Effect of Segmentation Network. To evaluate the performance of our segmentation network, we conduct ablation experiments with the same axis-aligned boxes regression method [15] as the detection network. The comparison results are shown in Table 1 and visualized in Fig. 6, from which we can see FCN can hardly provide a convincing mask prediction of large elongated objects (e.g. harbor), especially when they are surrounded by many other category objects (e.g. ship). This is because the features of the harbor and ships are confused in the top-level feature map, which eventually leads to poor prediction results (red part) as shown in Fig. 6(a). However, we use the bottom module to generate score maps, then merge scores with top-levels feature to supplement fine-granularity information about the harbor, accordingly our model can obtain a more accurate result as shown in Fig. 6(b).

4.4 Comparison with State-of-the-Art Methods

We compare our method with other state-of-the-art methods. In order to form a fair comparison, we use a heavier backbone (ResNet-101-FPN) which is the same as original Mask R-CNN [1] and PANet [2]. The comparison experiments are based on mmdetection framework [27]. Compared with other instance

segmentation methods in aerial images, our model can well process objects in arbitrary orientation (e.g. large vehicle) and large elongated objects (e.g. harbor) as shown in Table 2.



(a) FCN (b) Ours

Fig. 6. The segmentation network comparison between FCN and ours. Both are using ADN as detection network. (Color figure online)

5 Conclusion

Instance segmentation in aerial images is a challenging task. In this study, we take a full consideration on the arbitrariness of the orientation. A novel method named Rotated Blend Mask R-CNN is proposed which can well segment instances in aerial images. Compared with methods using axis-aligned boxes, applying oriented bounding boxes can well remove the other instances and background. Besides, we improve the segmentation network by merging top-level coarse instance information with lower-level fine-granularity. Our ablation study proves the effectiveness of each module. The proposed method outperform Mask R-CNN and PANet to a certain extent. In the future, we will explore a more efficient and accurate method for segmenting objects in aerial images.

Acknowledgement. This work was supported by the Youtu Lab of Tencent.

References

1. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. CoRR, vol. abs/1703.06870 (2017)
2. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. CoRR, vol. abs/1803.01534 (2018)

3. Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X.: Mask scoring R-CNN. CoRR, vol. abs/1903.00241 (2019)
4. Lin, T.-Y., et al.: Microsoft COCO: Common objects in context (2014)
5. Everingham, M., Van Gool, L., Williams, C., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. **88**, 303–338 (2010)
6. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. CoRR, vol. abs/1604.01685 (2016)
7. Xia, G.-S., et al.: DOTA: a large-scale dataset for object detection in aerial images. CoRR, vol. abs/1711.10398 (2017)
8. Liu, Z., Wang, H., Weng, L., Yang, Y.: Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds. IEEE Geosci. Remote Sens. Lett. **13**(8), 1074–1078 (2016)
9. Weir, N., et al.: SpaceNet MVOI: a multi-view overhead imagery dataset. CoRR, vol. abs/1903.12239 (2019)
10. Zamir, S.W., et al.: iSAID: a large-scale dataset for instance segmentation in aerial images. CoRR, vol. abs/1905.12886 (2019)
11. Chen, H., Sun, K., Tian, Z., Shen, C., Huang, Y., Yan, Y.: BlendMask: top-down meets bottom-up for instance segmentation (2020)
12. Girshick, R.B., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. CoRR, vol. abs/1311.2524 (2013)
13. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. CoRR, vol. abs/1406.4729 (2014)
14. Girshick, R.B.: Fast R-CNN. CoRR, vol. abs/1504.08083 (2015)
15. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. CoRR, vol. abs/1506.01497 (2015)
16. Dai, J., Li, Y., He, K., Sun, J.: R-FCN: object detection via region-based fully convolutional networks. CoRR, vol. abs/1605.06409 (2016)
17. Cai, Z., Vasconcelos, N.: Cascade R-CNN: delving into high quality object detection. CoRR, vol. abs/1712.00726 (2017)
18. Zhu, Y., Ma, C., Jun, D.: Rotated cascade R-CNN: a shape robust detector with coordinate regression. Pattern Recogn. **96**, 106964 (2019)
19. Zhu, Y., Wu, X., Du, J.: Adaptive period embedding for representing oriented objects in aerial images. CoRR, vol. abs/1906.09447 (2019)
20. Ding, J., Xue, N., Long, Y., Xia, G.-S., Lu, O.: Learning ROI transformer for detecting oriented objects in aerial images. CoRR, vol. abs/1812.00155 (2018)
21. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. CoRR, vol. abs/1411.4038 (2014)
22. Bolya, D., Zhou, C., Xiao, F., Lee, Y.J.: YOLACT: real-time instance segmentation. CoRR, vol. abs/1904.02689 (2019)
23. Bolya, D., Zhou, C., Xiao, F., Lee, Y.: Yolact++: Better real-time instance segmentation (2019)
24. Kirillov, A., Girshick, R., He, K., Dollar, P.: Panoptic feature pyramid networks, pp. 6392–6401 (2019)
25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition, pp. 770–778 (2016)
26. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Li, F.F.: ImageNet: a large-scale hierarchical image database, pp. 248–255 (2009)
27. Chen, K., et al.: MMDetection: open MMLab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)