FAST-Det: Feature Aligned SSD Towards Remote Sensing Detector

Yutong Niu[1](✉), Ao Li[1], Jie Li[2], and Yangwei Wang[2]

[1] School of Computer Science and Technology, Harbin University of Science and Technology,
Harbin, China
yutong_niu@126.com

[2] Shandong Provincial Innovation and Practice Base for Postdoctors, Weihaizhenyu Intelligence
Technology Co., Ltd., Weihai, China

Abstract. Object detection based on large-scale, high-resolution visible light Remote sensing images are widely used in military such as reconnaissance and civilian such as marine resource management. It is also an important task for the application of computer vision in remote sensing images. With the development of deep learning, more and more object detectors use deep network as the backbone, and accurate detection results and indicators can be obtained on conventional images. However, compared with conventional images, remote sensing images have more object numbers and object sizes, and the object distribution is also denser, which makes detection more difficult. At present, there are two types of object detectors: single-stage and two-stage. The single-stage detector directly obtains the detection result based on the feature map and pays more attention to the detection speed, while the two-stage detector generates the region of interest (RoI) by using feature map. More attention is paid to the accuracy of the test results when the test results are obtained through RoIs. This paper proposes a bilateral filtering refining method based on a single-stage detector, which refines the results obtained by a single-stage detector and approaches the performance of a two-stage detector without losing too much detection speed. Experiments conducted on the public large-scale visible light remote sensing dataset DOTA have proved the effectiveness of this method.

Keywords: Deep learning · Object detection · Bilateral filtering

1 Introduction

Now remote sensing technology is in a relatively mature state. Remote sensing images are widely used in military such as reconnaissance and civilian such as marine resource management. At present, the object detection applied to remote sensing images is mainly realized by deep neural network, which is composed of backbone and head. Backbone is usually a convolutional neural network, which uses a convolution operator to extract features from the input image step by step to generate a feature map for object detection;

Published by Springer Nature Switzerland AG 2022. All Rights Reserved
S. Shi et al. (Eds.): 6GN 2021, LNICST 439, pp. 255–263, 2022.
https://doi.org/10.1007/978-3-031-04245-4_22

the head of detector will perform object category and location detection on the feature [1].

Ross Girshick proposed a design idea of a classifier regression for object detection tasks in RCNN [2], and implemented end-to-end multi-task learning in Faster RCNN [3]. Common object detectors still maintain this design idea. Under the multi-task training of classification and regression, feature maps suitable for two tasks can be obtained together. It can be found by using the heat map to visualize the feature map that the features suitable for the classification task do not completely overlap with the features suitable for the regression task, and there is a deviation in the center point of the feature distribution. The reason is that the classification task has the invariance of translation and scale and pays more attention to the extraction of the difference information between the classes; the regression task is more sensitive to the location and scale information of the object and pays more attention to the extraction of the boundary information of the object. This causes the classification and regression feature centers to shift. This feature shift makes the anchor center with a higher classification score not necessarily close to the object geometric center. The offset of the classification regression feature center and the offset of the anchor center and the object geometric center increase the difficulty of regression from the anchor to the prediction bounding box and have a certain impact on the detection results [4, 5].

Compared with conventional images, remote sensing images have more objects, more object sizes, and denser distribution. This makes conventional object detectors unable to obtain accurate detection results on remote sensing images. Compared with the horizontal bounding box, the rotating bounding box is more suitable for object detection tasks with dense object distribution. In this work we present a feature-aligned rotating SSD detector, which can effectively adapt to remote sensing images. Specifically, the FPN structure is used to generate multi-size feature maps for detection. The larger feature map is more suitable for setting small anchors to detect small targets. On the contrary, a large anchor should be set for a smaller feature map to adapt to a large target, because it has a larger receptive field. In the first stage, the object is roughly detected to obtain a prediction bounding box. Compared with the anchor center, prediction bounding box center should be closer to the actual geometric center of the object. Therefore, the bilateral filter is used to calculate the central feature estimation of the prediction bounding box, integrate it into the original feature point position, generate a new feature pyramid, and complete the feature alignment. Using the aligned feature pyramid to perform the second classification regression can get a better detection effect. Through experiments on the DOTA dataset and calculation of indicators, the effectiveness of this model is verified.

2 Related Work

2.1 One-Stage and Two-Stage Object Detector

The object detection network based on deep learning can be divided into two categories according to the detection process. One is a single-stage detector that uses feature maps to detect directly. The other is a two-stage detector that relies on the region of interest. Faster RCNN [3], as a classic two-stage end-to-end object detector, can fully reflect the characteristics of the two-stage object detector. The detection process of this type of detector is carried out in two steps. The first stage is to obtain anchors according to the feature map of the picture in the Region Proposal Network (RPN) and get the region of interest (RoI). The second step is to pool the RoIs to the same size and perform the second detection. The two-stage object detector only classifies foreground and background in the first stage, and the focus is on calculating RoIs. The RoIs has undergone a regression, and its center point is closer to the object geometric center than the anchors, which helps to improve the accuracy of the second classification regression. In the second stage, the RoIs is fine-grained classification and further regression, and more accurate detection results are obtained. The two-stage detector requires RPN for anchor screening and RoIs generation, which requires more detection time. Most of the remote sensing images have a large size, so it is necessary to use a sliding window method for stepwise detection and then merge the results, so using a two-stage detector for remote sensing image detection will consume more time.

The single-stage detector directly performs classification and regression on the characteristics of the picture. The entire detection process only needs one step, and it focuses more on the speed of detection, and is more suitable for the distribution detection method based on sliding windows. SSD [6], as a representative of single-stage object detector, generates anchors of several sizes for each feature point in the feature map calculated by backbone. And directly perform fine-grained classification and bounding box regression on each anchor at one time. Finally, post-processing is performed by non-maximum suppression NMS to suppress redundant bounding boxes. Because there is no RPN sub-network, the detection speed is greatly improved, and it can better adapt to the sliding window method in high resolution image object detection.

2.2 Multi-scale Features Object Detector

Large object sizes and more small objects are the main challenges for object detection in visible light remote sensing images. The object detector used for remote sensing images should have the ability to detect objects of different sizes at the same time. Multi-scale feature maps are used in SSD for detection [6]. Large-size feature maps are more suitable for detecting small objects due to their smaller receptive fields. Therefore, several small-size anchors with aspect ratios are generated in the large feature maps. The small feature map is at the high level of the network and has a larger receptive field, which is suitable for detecting large objects. Therefore, several large-size anchors with aspect ratios are generated in the small feature map. SSD performs NMS on the object bounding boxes generated by these anchors in the post-processing stage to remove redundant bounding boxes. Single-stage and multi-scale features make SSD has fast detection speed and accurate detection results.

However, in order to prevent the low-level features from affecting the performance of the detector, the SSD adds a convolutional layer to the high-level of the backbone to generate a multi-size feature map, and only uses the high-level information. The information in the underlying feature map is particularly important in small objects detection.

In order to make rational use of multiple size feature maps, RetinaNet [7] adds the feature pyramid network FPN [8] on the SSD model. FPN will up-sample the high-level features, expand the size and add it to the bottom-level feature map, which makes each layer of the generated feature pyramid have relatively complete feature information. The detection of large and small objects can be considered at the same time, which improves the recall rate of the detector.

3 The Proposed Method

3.1 Multi-scale and Rotation

In order to make the single-stage object detector SSD more suitable for remote sensing images, we choose ResNet as the backbone of the detector. Because ResNet uses the residual block as the basic structure of the network, the shortcut connection in the residual block can play the role of identity mapping, thereby effectively preventing the degradation caused by the deepening of the network layer. Further improve the feature extraction ability of the input image. And add FPN structure between backbone and head, FPN and ResNet are connected horizontally to calculate feature pyramid. In the feature extraction stage, two paths are formed. ResNet calculates various scale feature maps from low to high, and FPN up samples high-level features from high to low, and adds them to low-level feature maps step by step to perform high-level feature information. Supplement. The feature pyramid calculated in this part maintains the feature information of various sizes from low to high, and the bottom-level information is also supplemented by the high-level information. Small-sized high-level feature maps are suitable for setting large-sized anchors, because they have greater feelings, and are more suitable for detecting large-sized objects. On the contrary, the receptive field of the underlying feature map is smaller and suitable for detecting small objects, and a smaller size anchor should be set.

In the case of dense object distribution, the horizontal bounding box will introduce other object information, which makes the bounding boxes affect each other, which is not conducive to the performance of the detector. In the remote sensing image of airport, harbor and parking lot scenes, there are a large number of objects, and the objects are very dense and the directions are changeable. At the same time, there are also long objects like bridges in remote sensing images. These reasons make it difficult for object detectors that use horizontal bounding boxes to achieve actual detection results.

Further, we add a rotation vector to the traditional 4-dimensional regression vector (center point [x, y] bounding box width w and height h). It also stipulates that the edge with an acute angle to the *x–axis* is defined as the height of the bounding box, the other side is the width of the bounding box.

3.2 Feature Alignment and Bilateral Filter

The object detector based on deep learning uses a multi-task method to train the network. In the training phase, the classification sub-network for object classification and the regression sub-network for object bounding box prediction are trained at the same time. The multi-task training method enables the trained feature extraction network to extract features suitable for two tasks synergistically. However, after using the heat map to visualize the feature map, it can be found that the features suitable for the classification task do not completely overlap with the features suitable for the regression task, and there is a deviation in the center point of the feature distribution. The reason is that the classification task has the invariance of translation and scale, and the category judgment will not change due to different sizes and positions. The feature maps trained by the classification task pay more attention to the extraction of the difference information between the classes. However, the feature map trained on the regression task should have the location feature of the object. The difference of focus makes the classification regression feature center biased. In other words, the classification feature is offset from the geometric center of the object. In actual applications, in order to prevent the disappearance of the gradient of the deep neural network, the idea of migration learning is often used when training the classifier, and the ImageNet classification task is used to pre-train the backbone to enhance its feature extraction ability. This also makes the feature map more biased towards the classification task, which makes it difficult to return to the bounding box.

The single-stage object detector represented by SSD will generate anchors of several sizes at each point in the feature map and use the anchors directly to calculate classification confidence and bounding box prediction. Since the center of the classification feature may not be at the geometric center of the object, the anchor center with high confidence may not be closer to the geometric center of the object. The resulting feature shift phenomenon.

The two-stage detector first adjusts the anchor to the proposal through the RPN, and then determines the region of feature map corresponding to the proposal, which is called the region of interest RoI, and then uses the pooled RoI for fine-grained classification and more accurate bounding box prediction. The movement from the center of the anchor to the center of the RoIs alleviates the feature shift to a certain extent. Inspired by this approach, we propose a method for feature alignment of a single-stage object detector. First, use bilinear interpolation to linearly estimate the feature value of center point. In addition, a nonlinear bilateral filter is used for nonlinear reconstruction, and the reconstructed features are spliced with the original features for secondary detection. The structure of FastDet is shown in Fig. 1. The bilateral filter is composed of a distance filter and a difference filter. The distance filter formula is shown in (1), and the difference filter is shown in (2).

$$\begin{aligned} r_{d_{lt}} &= \frac{\frac{1}{d_{lt}}}{\frac{1}{d_{lt}}+\frac{1}{d_{rt}}+\frac{1}{d_{lb}}+\frac{1}{d_{rb}}} \quad r_{d_{rt}} = \frac{\frac{1}{d_{rt}}}{\frac{1}{d_{lt}}+\frac{1}{d_{rt}}+\frac{1}{d_{lb}}+\frac{1}{d_{rb}}} \\ r_{d_{lb}} &= \frac{\frac{1}{d_{lb}}}{\frac{1}{d_{lt}}+\frac{1}{d_{rt}}+\frac{1}{d_{lb}}+\frac{1}{d_{rb}}} \quad r_{d_{rb}} = \frac{\frac{1}{d_{rb}}}{\frac{1}{d_{lt}}+\frac{1}{d_{rt}}+\frac{1}{d_{lb}}+\frac{1}{d_{rb}}} \end{aligned} \tag{1}$$

where, lt, rt, lb, rb are neighbor pixels of the center point. $d_\bullet$ indicates the distance from the point to lt, rt, lb, rb.

$$\begin{aligned} r_{v_{lt}} &= \frac{\frac{1}{\Delta v_{lt}}}{\frac{1}{\Delta v_{lt}}+\frac{1}{\Delta v_{rt}}+\frac{1}{\Delta v_{lb}}+\frac{1}{\Delta v_{rb}}} \quad r_{v_{rt}} = \frac{\frac{1}{\Delta v_{rt}}}{\frac{1}{\Delta v_{lt}}+\frac{1}{\Delta v_{rt}}+\frac{1}{\Delta v_{lb}}+\frac{1}{\Delta v_{rb}}} \\ r_{v_{lb}} &= \frac{\frac{1}{\Delta v_{lb}}}{\frac{1}{\Delta v_{lt}}+\frac{1}{\Delta v_{rt}}+\frac{1}{\Delta v_{lb}}+\frac{1}{\Delta v_{rb}}} \quad r_{v_{rb}} = \frac{\frac{1}{\Delta v_{rb}}}{\frac{1}{\Delta v_{lt}}+\frac{1}{\Delta v_{rt}}+\frac{1}{\Delta v_{lb}}+\frac{1}{\Delta v_{rb}}} \end{aligned} \tag{2}$$

where, lt, rt, lb, rb are neighbor pixels of the center point. $\Delta v_{\bullet}$ represents the difference between the point interpolation feature value and lt, rt, lb, rb feature value.

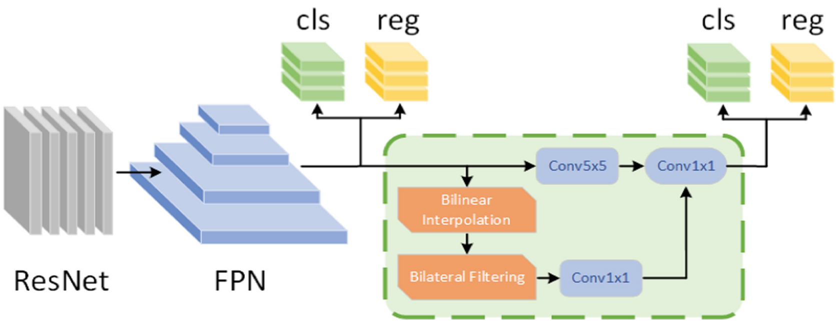

Fig. 1. The structure of FastDet

4 Experiments

4.1 Dataset

The experimental dataset uses Large-scale Aerial Images dataset DOTA [9]. The DOTA dataset contains 2806 aerial images, 15 categories, and a total of 188282 objects. Among them, 1/2 is divided into training set, 1/6 is divided into validation set, and 1/3 is divided into test set (unpublished label). The bounding box labeling method is different from the traditional parallel bounding box labeling method on opposite sides. DOTA uses a quadrilateral of any shape and direction determined by 4 points to label the object. The 15 categories included in DOTA cover common objects of different sizes, as well as objects with different aspect ratios (such as bridges and airplanes). And used the HRSC2016 [10] dataset. This dataset contains ship class and background class. The size ranges from 300 to 1500. The training set contains 436 images, the validation set contains 181 images, and the test set contains 444 images.

4.2 Process and Results

In this paper, we use six baseline methods and our proposed FastDet to conduct experiments on the DOTA dataset. The Baseline method uses R-FCN [11], Faster RCNN with rotating bounding box [3], PIoU [12], ICN [13], RADet [14], O2-DNet [15]. Use the DOTA dataset to test and calculate indicators. The following Table 1 gives a summary of the baseline model and this work presented model experimental results.

The category abbreviations in the table are as follows: BD stands for baseball diamond, GTF stands for ground track field, SV stands for small vehicle, LV stands for large vehicle, TC stands for tennis court, BC stands for basketball court, ST stands for storage tank, SBF stands for soccer ball field, RA It stands for roundabout, SP stands

Table 1. The AP of the baseline model and the model proposed in this work on DOTA

	R-FCN	FR-R	PIoU	ICN	RADet	O^2-DNet	FastDet
Plane	39.57	79.42	80.9	81.4	79.45	89.31	**89.56**
BD	46.13	77.13	69.7	74.3	76.99	82.14	**82.89**
Bridge	3.03	17.7	24.1	47.7	48.05	47.33	**52.79**
GTF	38.46	64.05	60.2	**70.3**	65.83	61.21	63.96
SV	9.1	35.3	38.3	64.9	65.46	71.32	**77.55**
LV	3.66	38.02	64.4	67.8	74.40	74.03	**80.76**
Ship	7.45	37.16	64.8	70.0	68.86	78.62	**86.53**
TC	41.97	89.41	**90.9**	90.8	89.70	90.76	89.56
BC	50.43	69.64	77.2	79.1	78.14	82.23	**85.43**
ST	66.98	59.28	70.4	78.2	74.97	81.36	**83.83**
SBF	40.34	50.3	46.5	53.6	49.92	**60.93**	58.29
RA	51.28	52.91	37.1	62.9	**64.63**	60.17	58.36
Harbor	11.14	47.89	57.1	67.0	66.14	58.21	**66.56**
SP	35.59	47.4	61.9	64.2	**71.58**	66.98	66.61
HC	17.45	46.3	64.0	50.2	**62.16**	61.03	61.87
Avg	30.84	54.13	60.5	68.2	69.09	71.04	**73.64**

for swimming pool, HC stands for helicopter, and Avg stands for the average indicator (AP) of all categories.

In order to get more comprehensive experimental results, we have carried out additional experiments on the HRSC2016 dataset. We choose three baseline models and our proposed FastDet for experiments, R2CNN [16], RRPN [17] and RoI-Transformer [18]. The following Table 2 gives a summary of the baseline model and this work presented model experimental results. The table shows backbone, speed and mAP (using VOC2007 and VOC2012 two calculation methods).

Table 2. AP and speed of the baseline model and FastDet on HRSC2016

Method	Backbone	mAP (VOC2007)	mAP (VOC2012)	Speed
R^2CNN	ResNet101	73.07	79.73	5 fps
RRPN	ResNet101	79.08	85.64	1.5 fps
RoI-Transformer	ResNet101	86.20	–	6 fps
FastDet	ResNet50	**87.60**	89.29	**19 fps**
FastDet	ResNet101	87.52	**89.40**	16 fps

4.3 Result Analysis

We used six baseline models and the model proposed in this paper to conduct experiments on DOTA. According to the experimental results, we can find that the performance of FastDet is significantly better than other baseline networks. The mAP is 138.7% higher than R-FCN. Most types of AP in the DOTA dataset are higher than other models. Especially on densely distributed objects, such as LV and Ship, the AP of FastDet can reach 80.76 and 86.53 respectively.

We also used three baseline models and the model proposed in this paper for supplementary experiments. For the model FastDet proposed in this paper, we used two backbones for experiments. When using the VOC2007 performance calculation method, FastDet with ResNet50 achieved the best results, which mAP is 87.60. When using the VOC2012 performance calculation method, FastDet with ResNet101 achieved the best results, which mAP was 89.40. Therefore, the model proposed in this paper has good performance in a variety of datasets. We also conducted related experiments on the speed of the detector. When using ResNet50 as the backbone, it can detect 19 images per second, which is much higher than the other three models. When using ResNet101 as the backbone, it can detect 16 images per second, which is the fastest model in using this backbone.

5 Conclusion

This paper present a feature-aligned rotating SSD detector FastDet, which uses the FPN structure to generate multi-size feature maps for detecting objects of different sizes. And the refining module is used to refine the detection results of the first stage. The refining module uses a non-linear bilateral filter to estimate the feature value of the center of the prediction bounding box and splice it to the original feature value position to make the feature more complete. Use the refined feature pyramid for the second detection to make the detection more accurate. Through experiments on the FastDet and baseline models on the DOTA and HRSC2016 dataset, it can be proved that our proposed network has higher detection performance.

Acknowledgment. This work was supported in part by the National Natural Science Foundation of China under Grant 62071157, Natural Science Foundation of Heilongjiang Province under Grant YQ2019F011 and Postdoctoral Foundation of Heilongjiang Province under Grant LBH-Q19112.

References

1. Zou, Z., Shi, Z., Guo, Y., et al.: Object detection in 20 years: a survey. arXiv preprint arXiv:1905.05055 (2019)
2. Girshick, R., Donahue, J., Darrell, T., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
3. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. arXiv preprint arXiv:1506.01497 (2015)

4. Liu, L., Ouyang, W., Wang, X., et al.: Deep learning for generic object detection: a survey. Int. J. Comput. Vis. **128**(2), 261–318 (2020)
5. Yang, X., Liu, Q., Yan, J., et al.: R3Det: refined single-stage detector with feature refinement for rotating object. arXiv preprint arXiv:1908.05612 (2019)
6. 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
7. 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, pp. 2980–2988 (2017)
8. Lin, T.Y., Dollár, P., Girshick, R., et al.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
9. Xia, G.S., Bai, X., Ding, J., 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)
10. Liu, Z., Yuan, L., Weng, L., et al.: A high resolution optical satellite image dataset for ship recognition and some new baselines. In: International Conference on Pattern Recognition Applications and Methods, vol. 2, pp. 324–331. SCITEPRESS (2017)
11. Dai, J., Li, Y., He, K., et al.: R-FCN: object detection via region-based fully convolutional networks. arXiv preprint arXiv:1605.06409 (2016)
12. Chen, Z., et al.: PIoU loss: towards accurate oriented object detection in complex environments. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12350, pp. 195–211. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58558-7_12
13. 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
14. Li, Y., Huang, Q., Pei, X., et al.: RADet: refine feature pyramid network and multi-layer attention network for arbitrary-oriented object detection of remote sensing images. Remote Sens. **12**(3), 389 (2020)
15. Wei, H., Zhang, Y., Chang, Z., et al.: Oriented objects as pairs of middle lines. ISPRS J. Photogram. Remote Sens. **169**, 268–279 (2020)
16. Jiang, Y., Zhu, X., Wang, X., et al.: R2CNN: rotational region CNN for orientation robust scene text detection. arXiv preprint arXiv:1706.09579 (2017)
17. Ma, J., Shao, W., Ye, H., et al.: Arbitrary-oriented scene text detection via rotation proposals. IEEE Trans. Multimed. **20**(11), 3111–3122 (2018)
18. Ding, J., Xue, N., Long, Y., et al.: Learning RoI transformer for oriented object detection in aerial images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2849–2858 (2019)