Few shot object detection in remote sensing images

Xingyu Zhang[a,b], Haopeng Zhang*[a,b], and Zhiguo Jiang[a,b]

[a]Department of Aerospace Information Engineering, School of Astronautics, Beihang University, Beijing 102206, China
[b]Beijing Key Laboratory of Digital Media, Beijing 102206, China

ABSTRACT

Object detection in remote sensing images has far-reaching significance. However, compared with object detection tasks in the field of natural images, there are still some challenge problems that need to be improved in remote sensing, due to the similarity and unbalanced scale between objects in remote sensing images. Currently, object detection algorithms based on deep learning have reached an incredible grade with the concerted efforts of researchers, and have been used in various aspects including remote sensing. However, these methods always need a lot of labeled data, while collecting a dataset is labor intensive and time consuming. Focusing on ship detection in remote sensing images, we introduce a few shot learning algorithm with attention mechanism aiming to detect ships of unknown classes with only a few labeled remote sensing images. Experiments on HRSC-2016 dataset validate the effectiveness of our method.

Keywords: few shot learning, object detection, remote sensing, deep learning.

1. INTRODUCTION

In the field of computer vision, the completion of target detection tasks has reached a high level with the concerted efforts of researchers, such as Faster-RCNN,[1] YOLO series,[2] etc. However, these algorithms need a lot of data to train. Therefore, large-scale labeled datasets play an important role in the development of object detection tasks. Obviously, the current deep learning methods are unreasonable in two main aspects. On the one hand, the dataset production process is time-consuming and labor-intensive. On the other hand, large-scale-data-driven learning methods do not conform to human cognitive processes. Thus few shot detection has become a hot research topic. Similarly, the application of few shot object detection in remote sensing images is also very important and urgent.

Few shot ship detection of ship in remote sensing images is a challenging task. It is more difficult to label the remote sensing images due to large viewing field and clutter background. Since remote sensing data is more difficult to obtain than natural images, it is more difficult to obtain ship data as well. Therefore, the detection problem of new types of ships without sufficient samples needs to be solved urgently. In addition, since the size of ships in remote sensing images is also relatively small, the performances of existing algorithms degrade and need to be improved.

To address these issues mentioned before, we propose a few shot ship detection algorithm in this paper. Our few shot ship detection framework can detect new types of ships without retraining and finetune. The structure of our framework is shown in Figure 1. We use the Multi-branches parallel network structure[3] and YOLO[2] detection algorithm, and add an attention mechanism to get better performance. Multi-branches parallel network is a weight shared feature extractor and its backbone is improved darknet. In order to solve the problem of unbalanced target scale, our feature extractor extracts three feature maps at different scales.[4] Particularly, every feature map integrates the information of multi-layer feature maps for extracting more sufficient features. The input of the multi-branches parallel network structure network are small images which entered is a close-up of the target and a large remote sensing image containing the target. They are entered in parallel. By averaging the corresponding feature maps on each scale of small images, we get the average representation of features of the

Further author information: (Send correspondence to Haopeng Zhang)
Haopeng Zhang: E-mail: zhanghaopeng@buaa.edu.cn, Telephone: +86 10 6171 6978

target on three scales. We use average representation of features of the target from small images as a convolution kernel that slide on query feature map in depth-wise cross correlation way. By this way, we obtain an attention feature map of the query image which provides better feature representation of target. In a word, our attention mechanism is mainly achieved by obtaining matching relationship between object pairs (support set and query set).

The contributions of this paper are as follows. Firstly, we propose a few shot ship detection framework with an attention mechanism, which can detect new types of ships in remote sensing images without retraining and finetune and needs few labeled data. Secondly, we use a one-stage detection algorithm to increase the speed of the algorithm while ensuring the detection accuracy. Therefore, it has high practicability. Thirdly, our method can achieve good experimental performance in the task of few shot ship detection in remote sensing images.

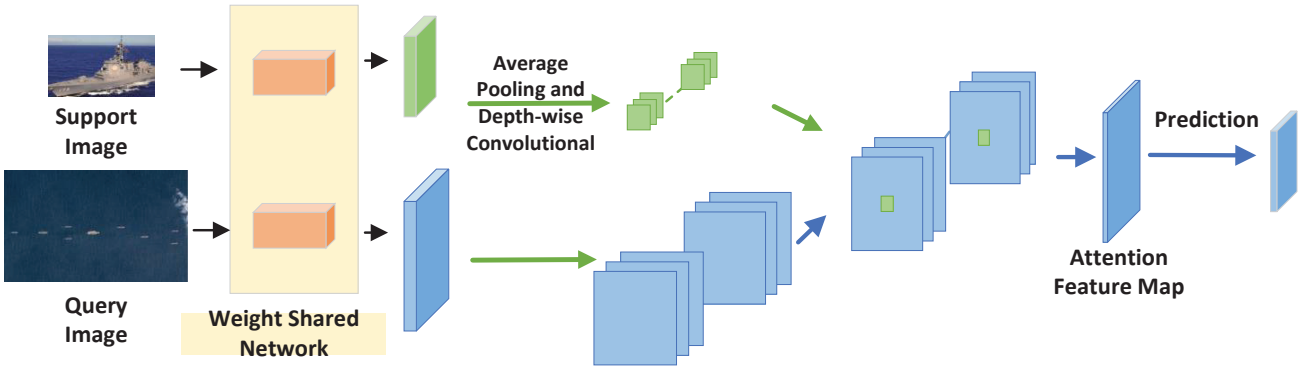

Figure 1. Overview of framework architecture. The support image and the query image are processed by a weight shared network for extract feature. The attention feature map is obtained after depth-wise cross correlation, which has enhanced feature of objects. Then, the enhanced feature map is input to the detect head.

2. METHOD

In this section, we first define our task of few shot ship detection in remote sensing images and then introduce the details of our framework.

2.1 Problem definition

Given a small image containing a close-up of the target, we regard it as the support image and call the set of support images as the support set. We regard the large image containing the target with label as the query image. Similarly, we call the set of query images as the query set. Our task is to find all the targets belonging to the support image in the query image with bounding boxes and give the target category information. For every train step, if we input K different categories and N pictures for each category, we define the task as N-way K-shot detection.[5]

2.2 Feature Extraction

For N-way K-shot task, we construct a weight-shared structure with (K+1) branches, one for query image and the others for support images. It should be noticed that in order to facilitate the presentation, we only drew one support map branch in Figure 1. Following weight-shared network, we average the corresponding feature maps on each scale of K support images. Since the YOLO[2] algorithm has both excellent accuracy and rapidity in the detection task, we use the backbone of YOLO v5 (improved darknet) as feature extractor to extract features of support images and query images. In order to effective detection of targets of different scales, the feature extractor extracts three features of different scales. Each feature map integrates the information of multi-layer feature maps for extracting more sufficient features. The structure of feature extractor is shown in Figure 2.

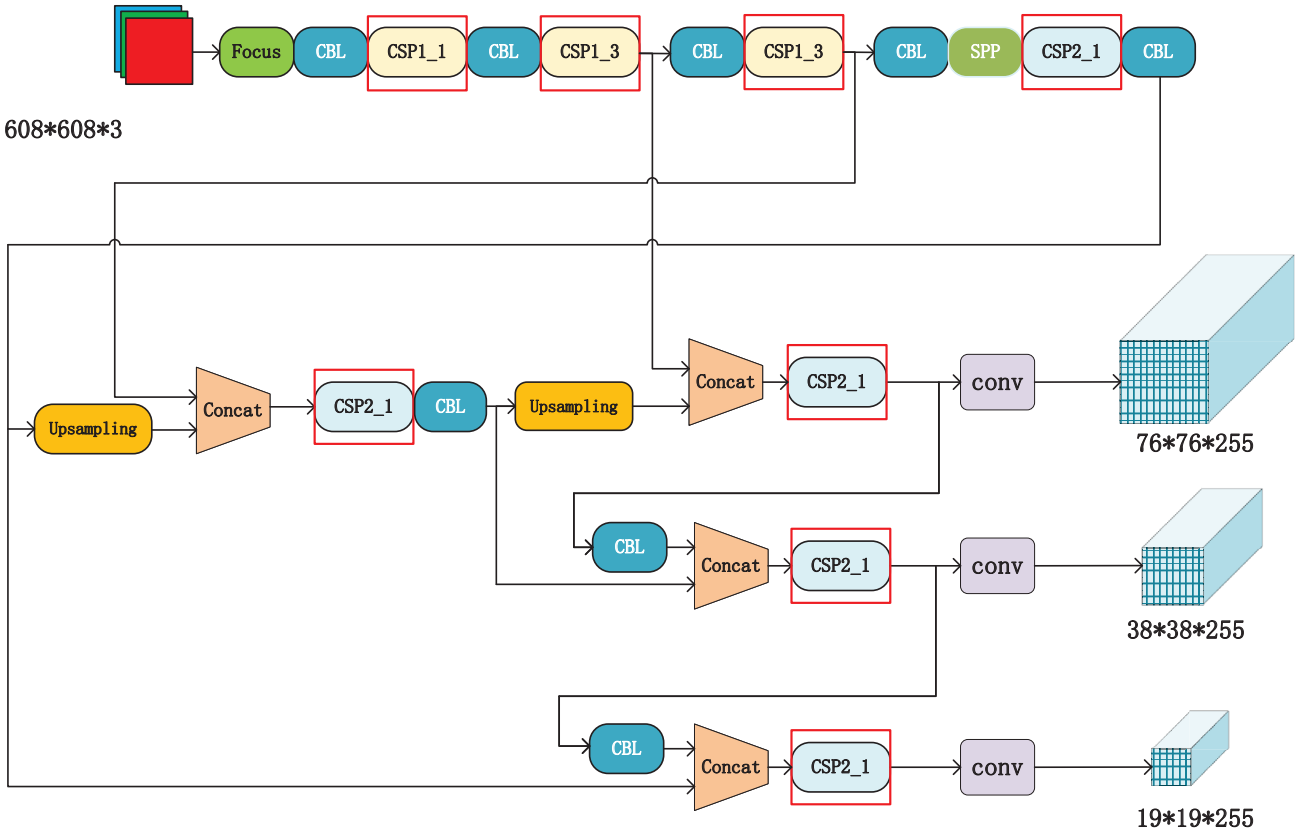

Figure 2. The structure of our feature extractor. Three feature maps with different scales are obtain, and each feature fuses the information of multi-layer feature maps.

2.3 Attention feature map

We get three different mean feature maps of different scales of target after feature extraction. In addition, we also have three different feature maps of different scales of query image. We denote the centralized representation of feature maps of the support set as $S_k \epsilon t^{b_k \times b_k \times c_k}$ and feature map of the query set as $Y_k \epsilon t^{h_k \times w_k \times c_k}$, where $k = 1, 2, 3$. We compute the similarity of query set and support set as

$$G_{h,w,k,c} = \sum_{i,j} S_{i,j,c,k} \cdot Y_{h+i-1,w+j-1,c,k}, i, j \epsilon 1, \cdots, b$$

where $G_{h,w,c,k}$ is attention feature map. We use S_k as a convolution kernel that slide on query feature map Y_K in depth-wise cross correlation way. After that, $G_{h,w,c,k}$ will be sent to the prediction model.

2.4 Prediction

The loss function of the target detection task is composed of two parts, i.e. classification loss and bounding box regression loss. In this paper, we use CIoU loss[6] and DIoU NMS,[7] respectively.

2.4.1 CIoU Loss

On the basis of DIoU, the scale information of the aspect ratio of the bounding box is considered. Its calculation formula is

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha\upsilon$$

where ρ represents the Euclidean distance, b, b^{gt} represent the center point of the prediction box and the truth box, and c is the diagonal distance of the smallest outer rectangle. α is the parameter to trade-off. α is computed as $\alpha = \frac{\upsilon}{(1-IoU)+\upsilon}$ and υ is a parameter used to measure the consistency of the aspect ratio. It is computed as $\frac{4}{\pi}(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w}{h})^2$.

2.4.2 DIoU NMS

In Non-Maximum Suppression (NMS), the IoU index is often used to suppress redundant detection boxes, where the overlapping area is the only factor. DIoU-NMS regards DIoU as the criterion of NMS. Because not only the overlap area should be considered in the suppression criterion, but also the center point distance between the two boxes should be considered. DIoU considers the overlap area and the center distance of the two frames at the same time. For the prediction box M with the highest score, the s_i update formula of DIoU-NMS can be formally defined as

$$s_i = \begin{cases} s_i, & IoU - R_{DIoU}(M, B_i) < \varepsilon, \\ 0, & IoU - R_{DIoU}(M, B_i) \geq \varepsilon). \end{cases}$$

3. EXPERIMENT

3.1 Dataset

We use the ship data in HRSC-2016 dataset[8] to perform our experiments. HRSC-2016 dataset has a total of 1055 pictures, which contain 19 different ship categories. We have chosen 617 pictures of them as training sets, and the remaining 438 pictures are testing sets.

3.2 Training details

Our model is trained on one GeForce GTX 1080Ti in end-to-end way. The learning rate is set as 0.001. We train the model 200 epochs. We unify the size of the input image to 640×640. We use a classic evaluation metrics,[9] including precision, recall, and AP_{50}. The calculation formulas for precision and recall rate are

$$precision = \frac{tp}{tp + fp}$$

$$recall = \frac{tp}{tp + fn}$$

where tp represents the number of positive samples that are correctly identified, fp represents the number of negative samples that are detected as positive samples, and fn represents the number of positive samples that are not correctly identified.

3.3 Comparative experiments

Table 1. The performance of our method in different shot number.

Shots	Dataset	mAP_{50}	Precision	Recall
3	HRSC2016	3.58	0.11	0.23
5	HRSC2016	3.7	0.13	0.27
10	HRSC2016	3.61	0.2	0.26

We first did experiments in 3-Shot, 5-Shot and 10-Shot to get better setting of our method. Experimental results are shown in Table 1. As can be seen from Table 1, the performance of 5-Shot is the best. Thus we performed comparative experiments between our method and other object detection methods with the setting of 5-Shot. For fairness, all compared methods were trained in the same way and all of them adopted the setting of 1-way 5-shot. The baseline method is getting bounding box through YOLO and randomly assigning the

Table 2. Results of comparative experiments.

Method	Dataset	mAP_{50}	Precision	Recall
Faster-Rcnn	HRSC2016	2.1	0.07	0.13
FSOD[10]	HRSC2016	3.5	0.11	0.22
YOLO	HRSC2016	1.9	0.07	0.11
baseline	HRSC2016	1	0.07	0.11
ours	HRSC2016	3.7	0.13	0.27

category information. The results in Table 2 show that our algorithm has reached the better performance on the HRSC-2016 dataset than other methods.

It can be seen from the results of Table 2 that our method can extract feature of targets in a more accurate way. Comparing the indexes of YOLO and our method, we can see that our method is better than YOLO in all aspects, indicating that our model has better learning ability. Because our method is a one-stage object detection method, it has advantage in reasoning speed compared with FSOD which is a two-stage method. Besides, our method also have higher detection accuracy than FSOD. Figure 3 shows the visualization results of our algorithm.

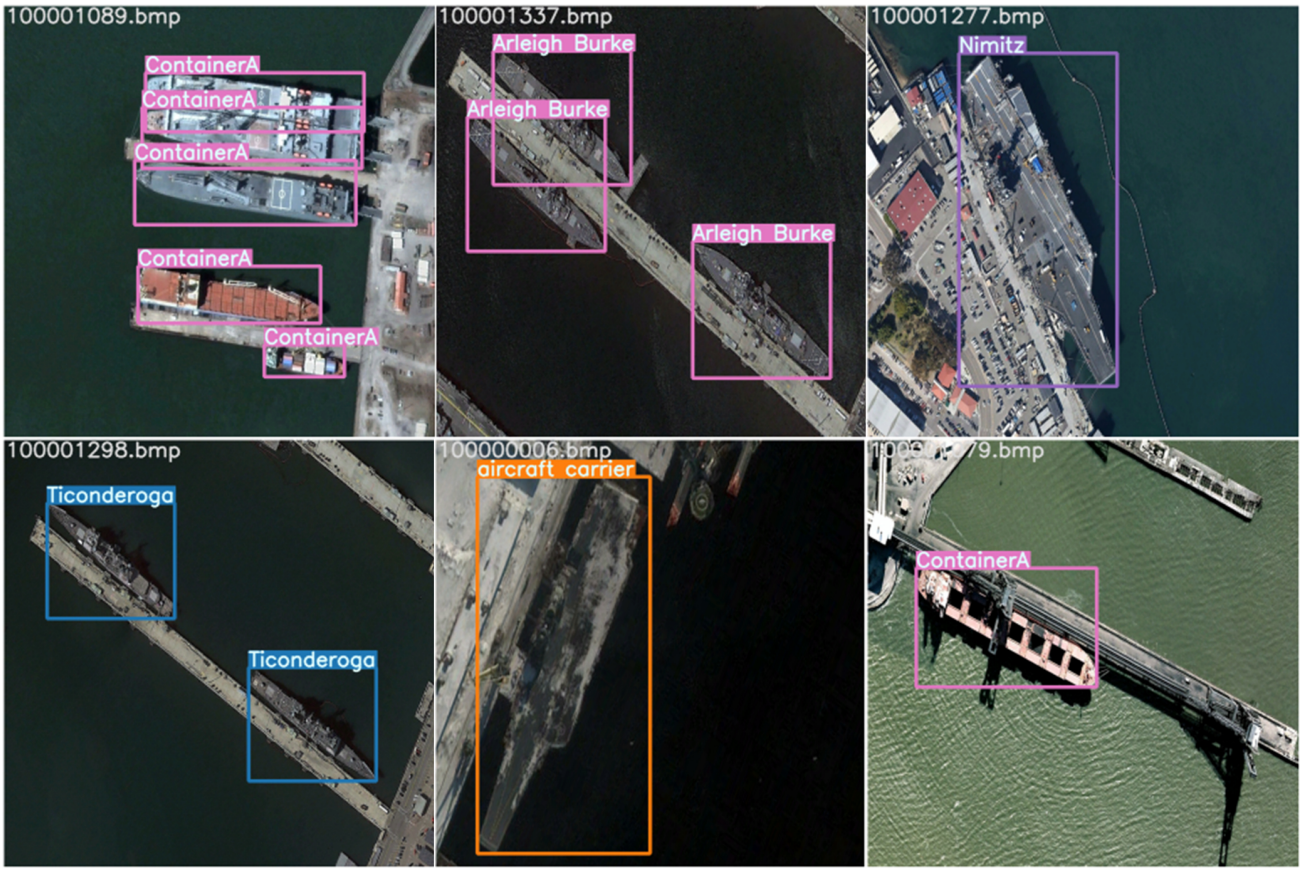

Figure 3. Visualization results of our algorithm.

4. CONCLUSION

In this paper, we propose a few shot learning algorithm for ship detection in remote sensing image with an attention mechanism, in order to detect ships of unknown classes with only a small labeled dataset. The algorithm contains a one stage pipeline, a feature extractor, and an attention module. It uses multi-branch structure and support set to study the feature of targets quickly with the backbone of the newest YOLO algorithm series. Compared with other methods, experiments on HRSC-2016 dataset show that our method can get improved detection performance. In future work, we will continue to improve the performance of our algorithm by introducing meta-learning.

ACKNOWLEDGMENTS

This work was supported in part by the National Key Research and Development Program of China (Grant No. 2019YFC1510905), and the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] Ren, S., He, K., Girshick, R., and Sun, J., "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems* **28**, 91–99 (2015).

[2] Redmon, J. and Farhadi, A., "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767* (2018).

[3] Chopra, S., Hadsell, R., and LeCun, Y., "Learning a similarity metric discriminatively, with application to face verification," in [*2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*], **1**, 539–546, IEEE (2005).

[4] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S., "Feature pyramid networks for object detection," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 2117–2125 (2017).

[5] Schwartz, E., Karlinsky, L., Shtok, J., Harary, S., Marder, M., Pankanti, S., Feris, R., Kumar, A., Giries, R., and Bronstein, A. M., "Repmet: Representative-based metric learning for classification and one-shot object detection," *arXiv preprint arXiv:1806.04728* **4323** (2018).

[6] Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., and Ren, D., "Distance-iou loss: Faster and better learning for bounding box regression," in [*Proceedings of the AAAI Conference on Artificial Intelligence*], **34**(07), 12993–13000 (2020).

[7] Bodla, N., Singh, B., Chellappa, R., and Davis, L. S., "Soft-nms–improving object detection with one line of code," in [*Proceedings of the IEEE international conference on computer vision*], 5561–5569 (2017).

[8] Liu, Z., Yuan, L., Weng, L., and Yang, Y., "A high resolution optical satellite image dataset for ship recognition and some new baselines," in [*International conference on pattern recognition applications and methods*], **2**, 324–331, SCITEPRESS (2017).

[9] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P., "Focal loss for dense object detection," in [*Proceedings of the IEEE international conference on computer vision*], 2980–2988 (2017).

[10] Fan, Q., Zhuo, W., Tang, C.-K., and Tai, Y.-W., "Few-shot object detection with attention-rpn and multi-relation detector," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 4013–4022 (2020).