# MULTI-SCALE CONTEXT-AWARE R-CNN FOR FEW-SHOT OBJECT DETECTION IN REMOTE SENSING IMAGES

*Haozheng Su*[1], *Yanan You*[1*], *Gang Meng*[2]

[1]School of Artificial Intelligence, Beijing University of Posts and telecommunications
[2]Beijing Institute of Remote Sensing Information, Beijing, China

IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-2792-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/IGARSS46834.2022.9883807

## ABSTRACT

In the field of remote sensing image object detection, the popular CNN-based methods need a large-scale and diverse dataset that is costly, and have limited generalization abilities for new categories. The few-shot object detection can be driven using only a few annotated samples. Existing few-shot detection methods are mainly designed for natural images, which ignore multi-scale objects and complex environments in remote sensing images. To tackle these challenges, we propose a two-stage multi-scale method based on context mechanism. Guided by the context-aware module, the multi-scale contextual information around the object is effectively extract and adaptively is combined into the ROI features to enhance the classification ability of the detector, which can reduce the classification confusion. Comparative experiments on public remote sensing image dataset RSOD show the effectiveness of our method.

***Index Terms***— few-shot object detection, multi-scale, context-aware, remote sensing images

## 1. INTRODUCTION

Object detection is a challenging fundamental part in the field of remote sensing image analysis and processing. In recent years, researchers have proposed a lot of approaches based on the popular CNN-based frameworks and achieved excellent performance [1] [2]. But these methods need to be driven by a large-scale and diverse dataset. For the task of detecting new categories, collecting a large number of annotated samples is expensive. On the other hand, training a detector with only a few annotated samples tend to cause the overfitting problem, which leads to a sharp decline in generalization ability. Therefore, a method that can learn robust detection ability from a few annotated samples of new categories is very necessary for object detection in remote sensing images.

Recently, few-shot object detection (FSD) has drawn immense research attention in the field of computer vision, which aims at obtaining satisfactory model performance through only a few annotated samples. Specifically, by providing a dataset that consists of the base classes with a lot of annotated samples as the source domain and the novel classes with only a few annotated samples as the target domain, the detector can detect the categories of both the source domain and the target domain.

Existing FSD methods are divided into methods based on transfer learning, such as two-stage fine-tuning approach (TFA) [3], and methods based on meta learning, such as Meta R-CNN [4]. Yang et al. [5] proposed the one-stage SSD-style method context-transformer, which can exploit contexts from images to improve the classification ability. However, the above methods are designed for common object detection in natural images. The scale of objects in remote sensing images changes more greatly than those in natural images. And the environment in remote sensing images is more complex than natural images, which leads to higher requirements for object location. These conditions make FSD more challenging.

In this paper, we proposed multi-scale context-aware R-CNN for FSD in remote sensing images, which is based on the context mechanism. In order to obtain better location ability, the two-stage detection algorithm Faster RCNN [6] is used as the basic framework, and the feature pyramid network (FPN) is introduced to solve the multi-scale problem. For object classification, inspired by the context mechanism of the context-transformer [5], the context-aware module is added to our framework. Normally, the size of a remote sensing image is larger than that of a natural image, and the environment is more complex, unlike context-transformer that extracts contextual information from the whole image, we focus on the contextual information of the surrounding background and combine it into ROI features to assist model transfer, which can effectively reduce the difficulty of object classification. Some experiments on public remote sensing images dataset RSOD [7] demonstrate the effectiveness and accuracy of our method.

## 2. PROPOSED METHOD

### 2.1. Overview of the Proposed Method

The framework of our proposed method is shown in Fig. 1. The model training is divided into two stages. The first stage uses large-scale data of the source domain to train a basic de-

*Corresponding to youyanan@bupt.edu.cn.

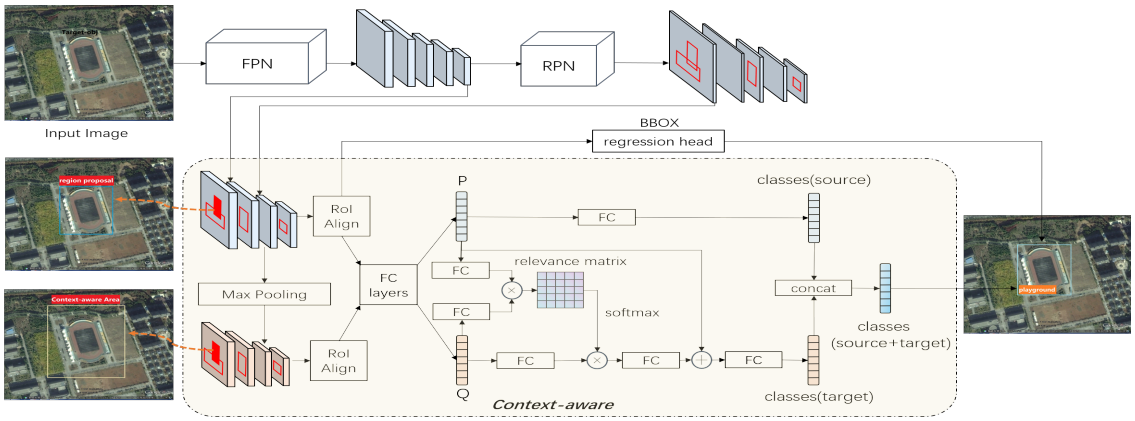


**Fig. 1**. The framework of our proposed method. We regard base classes and novel classes as the source domain and the target domain respectively. Context-aware module explores the contextual information to guide the classification in the target domain.

tector. In the second stage, the context-aware module is used to assist in model fine-tuning. For the input image, the FPN feature extractor is to embed images into the feature maps pyramid, and the RPN network obtains proposals from feature maps. Then the proposals are used for object classification and BBOX regression.

In the object classification part, max-pooling is used to build context-aware areas of the feature maps pyramid. The proposals on the context-aware areas are generated by the proposals on the feature maps pyramid, which are expanded to keep the size of the region on the feature maps unchanged according to the max-pooling step. ROI features vectors of the proposals and their corresponding contextual information are obtained by ROI align and $FC$ layers transformation. Then the relevance matrix between them is constructed, and the contextual information is combined into the ROI features under the guidance of the matrix for the object classification in the target domain. In order to reduce the difficulty of training, the classifier for the categories of source domain is retained and then fine-tuned because it has been trained by a large amount of data. In the end, the classification of the target domain and the classification of source target are output together. On the contrary, unlike object classification that is category-specific, RPN localization and BBOX regression are usually category-irrelevant [5]. Therefore, the detector can locate new categories by fine-tuning with only a few annotated samples in the target domain effectively.

### 2.2. Context-aware Module

For few-shot object detection, a transferred detector often performs well in location but fails in classification., We extract context-aware areas through context-aware module, and combine the contextual information of the background around the object to reduce classification confusion.

**Context-aware Area Construction**. Usually, people focus on the contextual information around the object rather than every detail in the image. Inspired by this, max-pooling is used to build context-aware areas for feature maps pyramid,

$$A_k = MaxPooling(F_k), \quad k = 1, 2, 3, 4, \tag{1}$$

where $k$ is the different scales of the feature maps pyramid, $F_k$ is the feature maps in different scales, $A_k$ is the context-aware areas in different scales. We keep the region size of the RPN extracted proposals on the feature maps unchanged, which is equivalent to expanding the actual region of the proposals. When expending the proposals, more background areas are included, which contain the contextual information between the objects and their surrounding experiments.

**Relevance Matrix**. We use the widely used dot product kernel to find relevance between each ROI features vector $P$ and corresponding contextual information vector $Q$ in the embedded space. As a result, a relevance matrix $R$ between ROI features and contextual information is obtaining,

$$R = FC(P) \times FC(Q)^T, \tag{2}$$

full connection layer $FC$ can increase the learning flexibility of kernel computing. These operations allow to automatically find important contextual information which can reduce the classification confusion caused by the lack of annotation samples.

**Contextual Information Combination**. After finding the relevance between ROI features and contextual information, the contextual information is combined into ROI features as a kind of relational attention. Firstly, we use softmax in each row of the relevance matrix to obtain the importance of each contextual information. Then it is used to obtain a weighted contextual vector $V$,

$$V = softmax(R) \times FC(Q), \tag{3}$$

full connection layer $FC$ is used to increase learning flexibility. Finally, our weighted contextual vector $V$ is restructured

through a full connection layer, and combined into the ROI features $P$ to obtain the feature vector $P^{'}$ for object classification in the target domain,

$$P^{'} = P + FC(V), \tag{4}$$

Then this vector is fed into a fully connected to obtain object scores in target domain. By combining contextual information, we hope to improve the performance of classifier to reduce classification confusion.

**Classification Acquisition**. In order to reduce the difficulty of training and retain the object classification ability of detector in the source domain, the context-aware module is applied to object classification of the target domain, while for object classification of the source domain that is trained with a large amount of data, we use the BOX classifier at the last layer of the pretraining model and fine-tune it to obtain object scores of the source domain. Finally, object scores of the source domain and object scores of the target domain are concatenated as the final output of the classifier.

### 2.3. Training Strategy

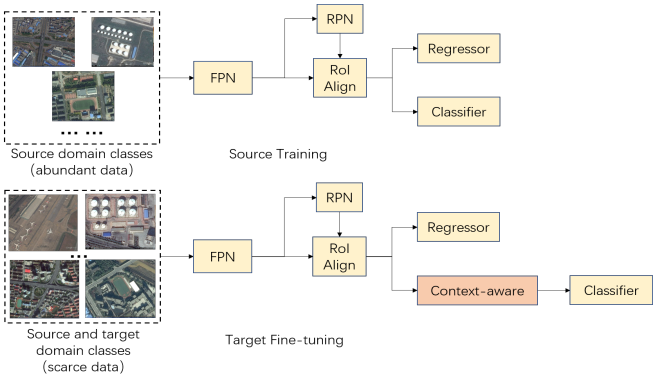


**Fig. 2**. Illustration of our two-stage training strategy.

We adopt a two-stage training strategy that is similar to the common frameworks [3] [4] [5]. As shown in Fig. 2, our training stage is divided into two parts: source training and target fine-tuning. In the first stage source training, only base classes with a large number of annotated samples are used to train a two-stage multi-scale basic detector, and the context-aware module is not used in this stage. In the second stage target fine-tuning, the context-aware module is used to assist in model fine-tuning, and base classes and novel classes are used to train the model. Because each novel class has only k annotated samples, we randomly select k samples for each base class to balance the number of samples.

## 3. EXPERIMENTS

We construct our model and run our experiments based on the deep learning framework Pytorch. 2 Tesla T4 GPUs are used for model training and testing.

### 3.1. Dataset and Setting

We evaluated our proposed method on a public remote sensing image dataset RSOD [7]. As shown in Fig.3, RSOD is a 4-class geospatial dataset for object detection, in which the 4 classes are aircraft, oil tank, overpass and playground. We randomly select 70% of each class in the images as the train set, and the rest of the images as the test set which contains all categories. Among the 4 classes, 1 class is selected to be the novel class as the target domain, while the other 3 classes are selected to be the base ones as the source domain. The train set of the base class is used as the input when training the source domain detector in the first stage. In the second stage, k instances are selected for each base class and novel class as the train data, where k equals 1, 2, 3, 5, and 10, and these images are randomly selected from the train set.


**Fig. 3**. The samples of 4 classes in RSOD.

To verify the effectiveness of our proposed method, our method is compared with 6 baselines. The first baseline is to train the Faster R-CNN with k instances in each base class and novel class together. We define it as FRCN+few. The second baseline is the same like as the first one except that there are abundant annotations in base classes. We define it as FRCN+joint. The third baseline is to train the Faster R-CNN using the two-stage train strategy that is similar to usours. We define it as FRCN+ft-full. The rest 3 baselines include transfer learning-based approach TFA [3], which freezes the feature extraction of the network and only fine-tunes the box predictor of the Fast R-CNN detector, meta learning-based approach Meta R-CNN [4], and context-based approach context-transformer [5]. The experimental setup and training strategy of these 3 baselines are the same as our method.

Average precision (AP) is used as evaluation metrics. We fellow the PASCAL VOC2007 development kit to evaluate the performance of our method. each experiment three times and average the results to obtain relatively stable experimental results.

**Table 1**. Few-shot object detection results (AP of novel class). Split1, 2, 3 and 4 represent the situation where the novel class is aircraft, oiltank, overpass, and playground. Red and blue indicate the best and the second best.

| | split1 | | | | | split2 | | | | | split3 | | | | | split4 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method/shot | 1 | 2 | 3 | 5 | 10 | 1 | 2 | 3 | 5 | 10 | 1 | 2 | 3 | 5 | 10 | 1 | 2 | 3 | 5 | 10 |
| FRCN+few | 0 | 1.21 | 6.06 | 9.09 | 13.7 | 1.27 | 6.06 | 9.09 | 9.9 | 42 | 0 | 3.62 | 7.83 | 22.3 | 40.5 | 10.7 | 35.4 | 42.6 | 61.1 | 79.2 |
| FRCN+joint | 0.19 | 4.55 | 9.09 | 11.4 | 16.5 | 0.7 | 4.55 | 23 | 45.3 | 50.4 | 1.82 | 4.66 | 11.8 | 24.6 | 46.3 | 9.62 | 28.8 | 43.7 | 70.4 | 81.6 |
| FRCN+ft-full | 5.64 | 8.72 | 14.1 | 20 | 23.8 | 6.06 | 10.9 | 32.5 | 54.2 | 62.3 | 4.57 | 9.84 | 16.9 | 31.6 | 57.6 | 28.1 | 53.4 | 71.5 | 76.6 | 85.3 |
| TFA | 6.06 | 7.93 | 13.4 | 18.8 | 22.3 | 9.09 | 9.91 | 27.9 | 44.6 | 50 | 2.43 | 8.2 | 12.8 | 28.3 | 52.8 | 23.1 | 47.8 | 63.8 | 70.1 | 83.7 |
| Meta R-CNN | 6.84 | 12 | 17 | 21.5 | 27 | 7.45 | 13.7 | 33.8 | 55.5 | 62.1 | 6.38 | 10.6 | 18.3 | 34.6 | 61.9 | 32.3 | 59.7 | 78.3 | 81.7 | 90.3 |
| Context-Transformer | 10.5 | 12.8 | 19.1 | 24.4 | 28 | 3.83 | 12.8 | 34.7 | 57.7 | 61.5 | 10.4 | 13.6 | 18.1 | 40.5 | 65.6 | 34 | 64.5 | 80.4 | 86.3 | 94.8 |
| Ours | 12.9 | 16.2 | 22.5 | 27.7 | 32.1 | 9.09 | 15.5 | 37.4 | 58.4 | 63.2 | 11.5 | 15.2 | 20.8 | 42.7 | 66.9 | 37.6 | 68.1 | 84.8 | 90.3 | 96.6 |

### 3.2. Dataset and Setting

The experiment results on the novel class are in Table 1, and we have the following observations. First, the performance of FRCN+few is the worst in most cases, which indicates that Faster R-CNN has poor ability to detect the novel class without a large number of annotated samples. Second, the performance of FRCN+ft-full is better than the one of FRCN+joint in most cases. All these show that our two-stage training strategy performs better than the training model with all classes together. Compared with the transfer learning-based method TFA and the meta learning-based method Meta R-CNN, Context-Transformer and our method perform better in most cases, which shows the outstanding performance of context mechanism for FSD in remote sensing. Finally, compared with the one-stage SSD-style context-transformer, our two-stage multi-scale context-aware method has significantly better performance in most cases, which proves the effectiveness of our method introducing the context mechanism into the two-stage multi-scale detection framework. For the basic detector, a few annotated samples may not be enough to effectively distinguish between the different classes such as overpasses and playgrounds, but contextual information in different backgrounds between them can greatly reduce the problem of classification confusion.

## 4. CONCLUSION

In this paper, we propose a few-shot object detection method multi-scale context-aware R-CNN, to deal with the problem of detecting the categories of target domain with only a few annotated samples in remote sensing images. Based on the context mechanism, our proposed method combines the contextual information around the object in multi-scale into the ROI features, effectively reducing the classification difficulty of model transfer with only a few annotated samples in the two-stage object detection framework, which improves the detection accuracy. The comparative experimental results on the public remote sensing image dataset RSOD show the effectiveness of our method. Few-shot object detection in remote sensing images is a very challenging problem, and we will further explore this field to achieve better performance.

## 5. ACKNOWLEDGEMENT

This work is supported by National Natural Science Foundation of China (62101060), and Beijing Natural Science Foundation, China (Grant No. 4214058).

## 6. REFERENCES

[1] Ding J, Xue N, Long Y, Xia G.-S, and Lu Q, "Learning roi transformer for oriented object detection in aerial images," in *CVPR*, 2019, pp. 2849–2858.

[2] Yanan Y, Bohao R, Gang M, Zezhong L, Fang L, and Zhixin L, "Opd-net: Prow detection based on feature enhancement and improved regression model in optical remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 6121–6137, 2021.

[3] Wang X, Huang T. E, Darrell T, Gonzalez J. E, and Yu F, "Frustratingly simple few-shot object detection," in *ICML*, 2020, pp. 9919–9928.

[4] Yan X, Chen Z, Xu A, Wang X, Liang X, and Lin L, "Meta r-cnn: Towards general solver for instance-level low-shot learning," in *ICCV*, 2019, pp. 9577–9586.

[5] Yang Z, Wang Y, Chen X, Liu J, and Qiao Y, "Context-transformer: Tackling object confusion for few-shot detection," in *AAAI*, 2020, pp. 12653–12660.

[6] Ren S, He K, Girshick R, and Sun J, "Faster r-cnn: Towards realtime object detection with region proposal networks," in *NeurIPS*, 2015, pp. 91–99.

[7] Xiao Z, Liu Q, Tang G, and Zhai X, "Elliptic fourier transformation-based histograms of oriented gradients for rotation- nally invariant object detection in remote-sensing images," *Remote Sensing*, vol. 36, pp. 618–644, 2015, 2.