Chapter 8
Small Object Detection of Remote Sensing Images Based on Residual Branch of Feature Fusion

Xiaoling Feng

Abstract In recent years, the detection of remote sensing images has been developed widely, and small objects have been paid more and more attention. The existing small object detection methods fuse the multi-scale features of different layers directly when using the feature pyramid network. However, due to the decrease of channels in feature fusion, the top-level feature of pyramid will lose information of the object, which is disadvantageous to detect small object ion. In order to fuse multi-scale features more effectively, we propose an object detection method based on the residual branch of feature fusion (RBFF), which is specially used to detect small objects. Our approach improves the network structure of the feature pyramid. We also recalculated the weights to reduce the semantic gap in feature fusion. In addition, we also introduce sub-pixel convolution to reconstruct the low-frequency information of the feature map accurately, to obtain the feature map with more information. The experimental results show that our method has a good effect.

8.1 Introduction

With the advance of deep learning, object detection can be divided into two groups: two-stage detectors and one-stage detectors. Two-stage detectors such as [1, 2] first generate some RoIs in the first stage and make an object classification and RoI-wise bounding box regression next. One-stage detectors, e.g., YOLO [3] and SSD [4], do not generate the RoIs and directly detect objects. Owing to extreme imbalance of foreground–background class, the performance of two-stage detectors is usually better than one-stage detectors. Anchor-free detectors are used to address this problem, such as [2, 5, 6]. It alternatively transforms object detection into a points detection problem to avoid complex computations of anchors and run faster.

To recognize and locate objects in remote sensing images more effectively, the research of remote sensors detection is urgent. In recent years, the research on object

X. Feng (✉)
Tiangong University, Tianjin 300387, China
e-mail: 1930081292@tiangong.edu.cn

R. Kountchev et al. (eds.), *New Approaches for Multidimensional Signal Processing*,
Smart Innovation, Systems and Technologies 270,
https://doi.org/10.1007/978-981-16-8558-3_8

detection is mostly based on Convolution Neural Network (CNN). For example, Region-based Convolutional Neural Networks [7] (R-CNN), known as a pioneering method, first generated region proposals using selective search and then refined them by extracting regional features from a convolution network. A region proposal network and an end-to-end trainable detector have been proposed to improve performance, which is named Faster R-CNN [8]. The Feature Pyramid Networks [9] (FPN) constructed a feature pyramid and predicted different objects at different pyramid feature maps by the scales of the region proposal. RetinaNet [10] chose a feature pyramid network likely FPN as its backbone and introduced a new focal loss to alleviate the imbalance between easy and hard examples. In aerial images, however, since the objects are mostly very small, these methods do not have good results in detecting them. This presents us with great challenges.

In recent years, many methods based on feature pyramid have been proposed. This is because FPN can combine low-level high-resolution information with higher-level strong semantic information, and simultaneously predict at different levels using lower-level features and higher-level features. As a result, targets in remote sensing images are not too small to be ignored by the detectors. Mou et al. [11] proposed a method to establish a feature pyramid network at all scales with strong semantic feature maps, which use a top-down pathway and horizontal connection. The feature map of different layers was responsible for detecting objects of different sizes. A dense feature pyramid network (DFPN) has been proposed by Yang et al. [12] to achieve automatic detection of ships: each feature map was closely linked and combined by concatenation.

With the improvement of the above methods, the ability of FPN network to recognize small objects has been improved, but some problems still exist. FPN proposes different features at each layer of the image pyramids, and then makes corresponding predictions. The shallow networks in the feature pyramid are more concerned with details and location information, while the upper layers focus more on semantics, which helps locate objects. First, feature maps of higher levels contributed to enhance the semantic information of lower levels. Second, the topmost convolution layer losses some information due to a few feature channels and is not compatible with other feature levels since it only has single-scale context information. So, the feature map on the top layer is very important to detect. To improve this shortcoming, we propose a method to enrich the top-level feature information. We use a five-layer feature pyramid network ($C_1 - C_5$), and our method uses residual branch to get a new convolution layer C_6. Residual branch is used to indoctrinate the original branches with different spatial background information. Generation of a new convolution layer C_6 is used to alleviate the loss of information due to reduced channel convergence.

In addition to the above method, we also introduce super-resolution (SR) technology to enrich some detailed information of feature maps. Image super-resolution refers to make recovery in images or image sequences from low-resolution (LR) to high-resolution (HR). In general, the higher the resolution of an image the more detail and information it contains. However, the resolution is not the same as the pixel size. For example, an image that is multiplied by five by an interpolation does not tell you how much detail it contains. Image super-resolution is concerned with recovering the

Fig. 8.1 The figure is an example of the SR technique, **a** is the ground truth, **b** is the low-resolution image, and **c** is the recovered high-resolution image

missing details in the image, that is high-frequency information. Figure 8.1 shows an example of SR technology, where a is the clear image, b is an image that needs to be restored to high resolution, and c is the result of the restoration. As you can see from the image, the restored image with SR contains more details and information. We use sub-pixel convolution to enrich the detail in the case of high-level details so that C_5 has more information. We hope this method can reduce the information loss and improve the performance of generated feature pyramids.

In order to realize the above method, we first improve the network structure of the traditional feature pyramid and propose a module to add a convolution layer before multi-scale feature fusion. The module also recalculates the fusion weight to fuse the extracted multi-scale feature layers more effectively. Finally, we introduce sub-pixel convolution to improve the semantic richness of the feature map to reduce the loss of detail.

8.2 Methods

Previous methods cannot solve the problem of incompatibility between high-level feature map and other level feature map. We propose a new RBFF network consisting of residual branches and sub-pixel convolution which is to detect small objects in aerial images. Figure 8.2 shows the framework of our method. The module we designed performs several operations on the tensor in order to fuse feature maps more efficiently. In addition, we use the sub-pixel convolution to enrich the high-frequency information of the feature map. Our method is described in detail below.

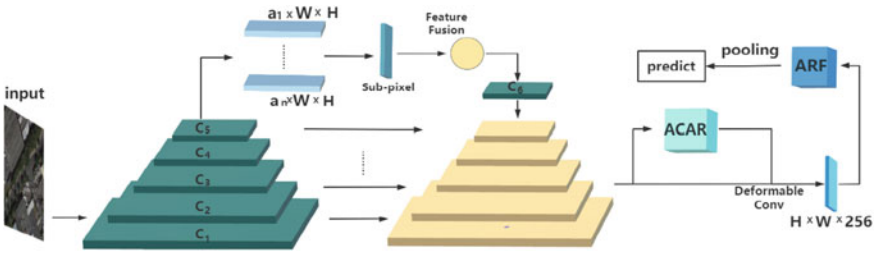

Fig. 8.2 The figure shows the RBFF network architecture

Our method adds a residual branch to generate a new feature map C_6 and recalculate weights. These features are then fused with recalculated weights. The ACAR module consists of the anchor classification branch and the anchor regression branch. Then we sent the anchor box and input feature maps into the deformable convolution [6] to extract aligned features. Finally, the active rotating filter [13] (ARF) is used to extract invariant directional features and produce the final detection results.

8.2.1 Sub-pixel Convolution

Most remote sensing images are very large. For example, the size of images in the DOTA dataset is about 4000 × 4000, and small objects like vehicles have very little information in the image. In addition, when the image is extracted by the feature pyramid network, there is less detail left, making it impossible to fully identify small objects in the image. The appearance of image super-resolution technology solves this problem.

In general, both I^{LR} and I^{HR} can have C color channels, thus they are represented as real-valued tensors of size $r\mathrm{H} \times \mathrm{rW} \times \mathrm{C}$ and $r\mathrm{H} \times \mathrm{rW} \times \mathrm{C}$, respectively. There is a way to realize image super resolution is convolution that uses fractional stride of $\frac{1}{r}$ in the LR space. But this way will increase the computational cost because that process happens in the HR space. So, we use a convolution with stride of $\frac{1}{r}$ in LR space filters W_a of size k_a with weight spacing $\frac{1}{r}$, which do not active all W_a convolution. And we do not need to activate weights and do not need to calculate the weights which are between pixels. The activated pattern has activated at most $\lceil \frac{k_a}{r} \rceil^2$ weights. These patterns are activated periodically throughout the convolution, relying on the different sub-pixel positions: mod (a, r), mod (b, r) where a, b is the coordinates of output pixel in HR space. In this paper, we use a more effective way called sub-pixel convolution to achieve the above process when mod $(k_a, r) = 0$:

$$U^{SR} = t^K\left(U^{LR}\right) = VB(S_K \times t^{K-1}\left(U^{LR}\right) + c_K) \tag{8.1}$$

where VB is a periodic shuffle operator that ranges the elements of the $H \times W \times C \cdot r^2$ tensor again into a tensor of the size $rH \times rW \times C$. This operation can mathematically be described as follows:

$$PS(T)x, y, c = T_{\lfloor x/r \rfloor, \lfloor y/r \rfloor}, c \cdot r \cdot mod(y, r) + c \cdot mod(x, r) \tag{8.2}$$

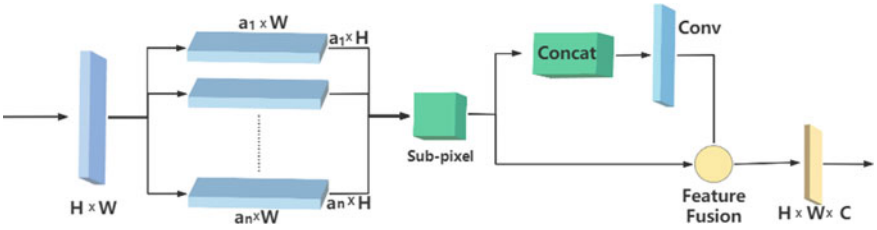

Fig. 8.3 The diagram shows the detailed structure of the residual branch that we propose. First of all, the topmost feature map has to go through three scales of adaptive pooling. Then the feature is amplified by sub-pixel convolution and then horizontally concatenated

8.2.2 Residual Branches

In the feature pyramid network, the top-down feature fusion process in the pyramids loses information at the top level due to fewer channels. To this end, we use a ratio-invariant adaptive pooling on the topmost layer of the feature pyramid to produce feature pyramid with different scales ($a_1 \times S$, $a_2 \times S$, ., $a_n \times S$) of multiple contextual features. To avoid the aliasing effects caused by interpolation, we have set three different scales to fit these contextual functions rather than simply summarizing them. Next sub-pixel convolution is used to scale up to the scale of S for subsequent fusion. Each context feature then independently passes through a 1×1 convolution layer, to reduce the channel dimension to 256 of the feature maps. Finally, in order to construct a feature pyramid, we use a 3×3 convolution layer at each feature map, as shown in Fig. 8.3.

8.3 Methods

8.3.1 Data Set

Our experiments were running primarily on the DOTA [14] dataset, which contains 2,806 aerial images of approximately 4000×4000 in size and 188,282 instances. And the dataset has 15 categories: plane (PL), ship (SH), storage tank (ST), baseball diamond (BD), tennis court (TC), basketball court (BC), ground track field (GTF), harbor (HA), bridge (BR), large vehicle (LV), small vehicle (SV), helicopter (HC), roundabout (RA), soccer ball field (SBF), and swimming pool (SP). It is marked as a quadrilateral with an arbitrary shape and orientation determined by four points rather than a traditional horizontal box. Specifically, first mark an initial point (x_1, y_1) and then mark 2, 3, and 4 in clockwise order. The initial point is usually selected at the head of the object. If it is an object such as a port with no obvious visual shape, choose the upper-left corner as the first point, as shown in Fig. 8.4.

Fig. 8.4 The figure shows how the dataset labels are defined

Function of Loss. The loss function of our method consists of two parts. The loss function is defined as follows:

$$L = \frac{1}{N_R}\left(\sum_i L_c\left(c_i^R, l_i^*\right) + \sum_i 1_{l_i^* \geq 1} L_r\left(x_i^R, g_i^*\right)\right) + \frac{\lambda}{N_M}\left(\sum_i L_c\left(c_i^F, l_i^*\right) + \sum_i 1_{l_i^* \geq 1} l_r\left(x_i^F, g_i^*\right)\right), \tag{8.3}$$

where λ is a loss balance parameter, **1** is an indicator function, N_R and N_M are the numbers of positive samples in the ACAR and ARF, respectively, i is the index of a sample in a minibatch. c_i^R and x_i^R are the predicted category and refined locations of the anchor i in ACAR. c_i^F and x_i^F are the predicted object category and locations of the bounding box in ARF. l_i^* and g_i^* are the ground-truth category and locations of the anchor i. The Focal loss [10] and smooth $L1$ loss are adopted as the classification loss L_C and the regression loss L_R, respectively. The hyperparameters of Focal loss Lc are set to $\alpha = 0.25$ and $\gamma = 2.0$. We use the same training procedure as in Detectron [15].

8.3.2 Ablation Study

Residual Branches. In our approach, the network is enhanced by changing its structure and adding a new branch. To compare with another method, we use ResNet-50 as the backbone of the two methods. S^2A-Net [16] was chosen for comparison with our method. The result of using and not using residual branch are shown in Table 8.1. We use S^2A-Net to represent the S^2A-Net method and RBFF to show our method. Our method provides better detection results for small objects on the DOTA validation dataset.

Sub-pixel convolution. To test the impact of adding sub-pixel convolution on improving the accuracy of small target detection, we work on two tests with our

Table 8.1 Experimental results with different networks

Network	PL	BR	SV	LV	SH	TC	BC	ST
S^2A-Net	89.64	47.01	66.87	83.26	88.41	90.69	63.09	87.39
RBFF	**89.74**	**47.42**	**67.91**	**83.34**	**88.72**	**90.72**	**65.26**	**88.21**

Table 8.2 Comparison of the results of the experiment

Network	PL	BR	SV	LV	SH	TC	BC	ST
S^2A-Net	89.64	47.01	66.87	83.26	88.41	90.69	63.09	87.39
RBFF	**89.89**	**47.42**	**69.85**	**83.49**	**88.82**	**90.69**	**65.62**	**88.29**

network, one using sub-pixel convolution and the other not. Here we use sub-pixel to denote the network using sub-pixel convolution and S^2A-Net to denote the method we did not use. The result of adding sub-pixel convolution or not is shown in Table 8.2. The table shows that the use of sub-pixel convolution has a positive impact on the detection of small objects in general.

8.3.3 Comparison of Experimental Results

The RBFF method was compared with other popular methods in the DOTA dataset. The results of the experiment are shown in Table 8.3. In contrast to many previous works [13, 17] was designed to detect large scale targets, our experimental results presented in the table show detection results for nine types of objects which is aimed at evaluating the small objects. The mAP in the last row of the table is also the

Table 8.3 Comparison with other methods on DOTA dataset. FFA-3(M) implies the use of the multi-stage detector of FFA-3 for experiments

Method	Back	PL	GTF	SV	LV	SH	TC	ST	SBF	HA	mAP
RetinaNet [10]	R101	88.82	65.72	67.11	55.82	72.77	90.55	76.30	54.19	63.71	70.05
FFA-3 [18]	R101	88.80	57.90	63.60	75.90	79.60	**90.80**	82.90	54.30	66.90	71.49
FFA-3(M) [18]	R101	89.60	58.90	67.20	76.50	81.40	90.01	83.40	55.70	73.20	75.11
R^3Det [19]	R101	89.54	62.52	70.84	74.29	77.54	**90.80**	83.54	61.97	65.44	75.12
S^2A-Net [16]	R101	89.64	74.13	66.87	83.26	88.41	90.69	87.39	73.53	73.58	80.83
RBFF	R50	**90.05**	67.30	67.83	**83.33**	**88.62**	90.61	87.64	70.07	73.34	79.87
RBFF	R101	89.91	**75.82**	**70.49**	82.99	88.50	90.73	**87.92**	**74.65**	**75.25**	**81.81**

average of the detection of these 9 types of objects. From the result, it is clear that our method outperforms some previous detection methods. With the default input size, e.g., 1024 × 1024, RBFF can run at 399 ms per image on the RTX2080. A single-scale test can run at 66 ms per image. Finally, some visualization of detection results can be seen in Figs. 8.5 and 8.6.

Fig. 8.5 The figure shows visualization results of our method. In the figure, the four pictures on the left are detection results of the S^2A-Net, and the four pictures on the right are the detection results of our method. Significantly more objects are identified in the red boxes in the four pictures on the right than on the left

Fig. 8.6 This figure shows part of detection results obtained by our method

8.4 Conclusion

In this paper, a novel method for remote sensing detection has been proposed based on the feature pyramid network. Our method uses the residual branch to improve the network structure and reduce the feature loss that occurs during feature fusion. The features are then scaled by sub-pixel convolution. Our method uses the focal loss to better rebalance the variant scales of the bounding box. Multi-scale testing can significantly improve detection performance. Our RBFF was trained using ResNet-50-FPN and ResNet-101-FPN, both achieved good performance on DOTA dataset. I hope that our approach will be useful in the field of remote sensing object detection or data statistics.

References

1. Girshick, R.: Fast R-CNN. In: ICCV, pp. 1440–1448 (2015)
2. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points (2019). arXiv:1904.07850
3. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: CVPR, pp. 779–788 (2016)
4. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: SSD: single shot multibox detector. In: ECCV, pp. 21–37 (2016)
5. Yang, Z., Liu, S., Hu, H., Wang, L., Lin, S.: Reppoints: point set representation for object detection. In: ICCV, pp. 9656–9665 (2019)
6. Law, H., Deng, J.: Cornernet: detecting objects as paired keypoints. In: ECCV (2018)
7. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, Columbus, OH, United states, pp. 580–587 (2014)
8. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017)
9. Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature Pyramid Networks for Object Detection, vol. 2017-January, Honolulu, HI, United States, pp. 936–944 (2017)
10. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. IEEE Trans. Pattern Anal. Mach. Intell. **42**(2), 318–327 (2020)
11. Mou, L., Zhu, X.X.: Vehicle instance segmentation from aerial image and video using a multi-task learning residual fully convolutional network. IEEE Trans. Geosci. Remote Sens. **56**(11), 6699–6711 (2018)
12. Yang, X., Sun, H., Fu, K., Yang, J., Sun, X., Yan, M., Guo, Z.: Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks. Remote Sens. **10**(1) (2018)
13. Zhou, Y., Ye, Q., Qiu, Q., Jiao, J.: Oriented response networks. In: CVPR, pp. 4961–4970 (2017)
14. Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L.: DOTA: a large-scale dataset for object detection in aerial images. In: CVPR, pp. 3974–3983 (2018)
15. Girshick, R., Radosavovic, I., Gkioxari, G., Dollár, P., He, K.: Detectron (2018). https://github.com/facebookresearch/detectron
16. Han, J., Ding, J., Li, J., Xia, G.-S.: Align deep features for oriented object detection. IEEE Trans. Geosci. Remote Sens. (2021)
17. Chen, K., Ouyang, W., Loy, C.C., Lin, D., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J.: Hybrid Task Cascade for Instance Segmentation, pp. 4969–4978 (2019)

18. Fu, K., Chang, Z., Zhang, Y., Xu, G., Zhang, K., Sun, X.: Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. ISPRS J. Photogramm. Remote. Sens. **161**, 294–308 (2020)
19. Yang, X., Liu, Q., Yan, J., Li, A.: R3det: refined single-stage detector with feature refinement for rotating object. CoRR, vol. abs/1908.05612 (2019)