Discriminative Feature Pyramid Network For Object Detection In Remote Sensing Images

Xiaoqian Zhu[1,2], Xiangrong Zhang[1,2], Tianyang Zhang[1,2], Peng Zhu[1,2], Xu Tang[1,2], Chen Li[3]
[1]*Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education,*
[2]*Xidian University, Xi'an 710071, China,*
[3]*School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China*

southmaiya@gmail.com, xrzhang@mail.xidian.edu.cn, tianyangzhang@stu.xidian.edu.cn,
zhupeng@stu.xidian.edu.cn, tangxu128@gmail.com, cli@xjtu.edu.cn

***Abstract*—Multi-class geospatial object detection in remote sensing images suffer great challenges, such as large scales variability and complex background. Although feature pyramid network (FPN) can alleviate the problem of scale variation to some extent, it causes the loss of spatial and semantic information which is not conducive to object location. To address the above problem, this paper proposes a discriminative feature pyramid network (DFPN) by introducing a global guidance module (GGM) and a feature aggregation module (FAM). Specifically, the global guidance module delivers the high-level semantic information to lower layers, so as to obtain feature maps with stronger semantic information to eliminate the interference caused by complex background. The feature aggregation module enhances the interflow of information between different layers and better captures the discrimination information at each layer. We validate the effectiveness of our method on the NWPU VHR-10 and RSOD datasets, the results outperform baseline by 2.06 and 3.88 points respectively.**

***Index Terms*—Object detection, discriminative feature learning, global guidance module, feature aggregation module**

I. Introduction

Object detection aims to obtain the spatial location of each category object in given images which has a wide range of applications [1] [2]. Thanks to the rapid development of deep convolutional neural networks (CNN), object detection has achieved significant improvements in recent years. Generally, object detection can be divided into two main streams: the two-stage methods and single-stage methods. The two-stage detectors [3]–[5] use the multiple stage regressions to obtain the location of objects. To be specific, Faster R-CNN [4] introduces the region proposal network (RPN) to generate regional proposals and then uses the Region-CNN (RCNN) [6] to obtain the final results. These methods can obtain accurate prediction results but simultaneously leads to increased model complexity and slow inference speed. To accelerate the inference phase, the single-stage detectors [7]–[9] are proposed, they completely eliminate the generation of region proposals and encapsulate all calculations in a single network, leading to faster speed and comparable results. Single Shot MultiBox Detector (SSD) [8] employs feature maps of shallower layers to predict smaller objects whereas deeper features to detect large objects. The performance of the single-stage detection method is seriously hampered by the imbalance between positive and negative samples. RetinaNet [9] solves this problem by employing the focal loss to change the weight of positive and negative samples.

Although many object detection architectures have achieved remarkable performance, they are designed for the natural scene. It is not effective to directly apply them to remote sensing images (RSIs) because of the large margin between RSIs and natural images. Compared to the nature images, the RSIs generally have satisfactory spatial resolution and complex background, which may cause false detection. Besides, the large scale variation, the appearance ambiguous and the complex distribution of objects further increase the difficulty of object detection in RSIs.

To address the above problem, many researchers [10]–[13] are committed to object detection in RSIs. Cheng et al. [10] proposed a new rotation-invariant convolutional neural networks (RICNN) model by introducing a rotation-invariant layer based on the structure of AlexNet to learn the rotation-invariant feature of the objects. Li et al. [11] added multi-angle anchors on RPN, and adopted hybrid constrained Boltzmann machine to fuse local context information. Zhong et al. [12] introduced a position-sensitive balance (PSB) framework to diminish the influence of translation-invariance of the convolutional neural network on object localization. Zhang et al. [13] provided an Encoder-Decoder architecture, called Rotated Feature Network (RFN), to produce rotation-sensitive feature maps for regression and rotation-invariant feature maps for classification. Deng et al. [14] adopted a multi-scale object proposal network (MS-OPN) to generate object candidate boxes which then are used to classification and regression by a proposed accurate object detection network (AODN) to achieve multi-class object detection.

However, these methods do not effectively deal with the complicated background in RSIs. Inspired by [15], we introduce a global guidance module (GGM) to deliver high-level semantic information to low-level features, enriching the semantic information of features to avoid mis-classification of complex background. In order to fully integrate the features

This work was supported by the National Natural Science Foundation of China (Nos. 61772400, 61772399, 61871306), and the 111 Project (No. B07048).

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

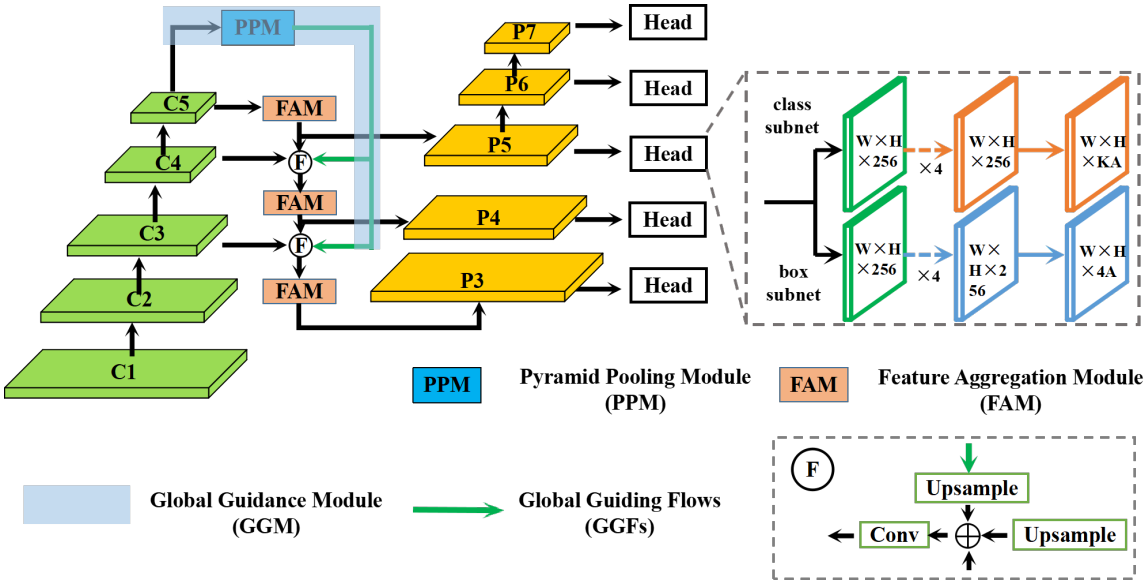


Fig. 1: The overview architecture of our method. We introduce the global guidance module (GGM) and the feature aggregation module (FAM), in which the pyramid pooling module (PPM) and the global guidance flows (GGFs) constitute the GGM.

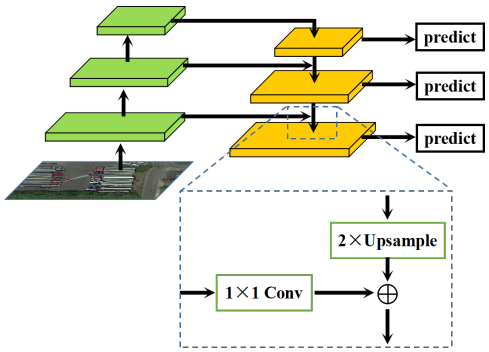


Fig. 2: The construction of feature pyramid network.

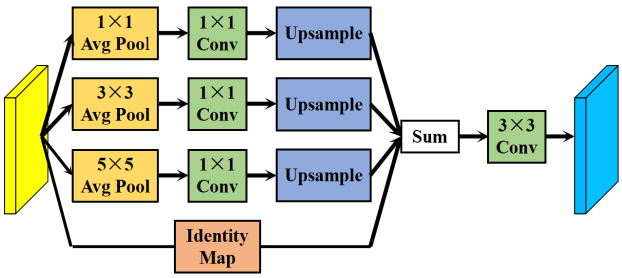


Fig. 3: Detailed illustration of the pyramid pooling module (PPM).

of different levels, we propose a feature aggregation module (FAM) to enhance the interflow of information at different layers to obtain better feature maps, which is more conducive to object location and classification. The major contributions of this paper are summarized as follows.

(1) We introduce the GGM to enhance the semantic information of feature maps.

(2) We propose the FAM to more availably merge the high-level and low-level features.

(3) We conduct comprehensive experiments to verify the effectiveness of the proposed module and achieve remarkable performance.

II. OUR METHOD

The overall architecture of our proposed method is shown in Fig. 1 and it can be regarded as an extension of RetinaNet which better trade-off between detection accuracy and inference speed. As shown in Fig. 1, we first use CNN to extract features from the given image. Then, a discriminative feature pyramid network with proposed GGM and FAM to obtain multi-level feature maps. Finally, the multi-level feature maps are fed into the detection head to generate detection results.

A. Review of Feature Pyramid Network

FPN [16] exploits an additional top-down pathway and the lateral connections to combine higher-level and lower-level features. Fig. 2 shows the construction of FPN, FPN first obtains feature maps of different spatial resolutions from the backbone network through a bottom-up pathway, these feature maps are then upsampled through a top-down pathway to progressively restore the spatial resolution, while the lateral connection is to merge the feature maps of the same spatial size from the top-down pathway and bottom-up pathway by an element-wise sum operation and a 3×3 convolutional layer.

B. Global Guidance Module

In the construction of FPN, the high-level semantic information is gradually diluted in the process of transmission from high-level layers to the shallow ones, leading to the lack of semantic information in shallow feature maps. The above

phenomenon results in mis-classification in the complex background and influences the results of detection. In view of this, we propose the GGM to enhance the semantic information of shallower layers and avoid false predictions in the background regions.

The GGM consists of a pyramid pooling module (PPM) and global guidance flows (GGFs). As demonstrated in Fig. 3, the PPM is composed of four sub-branches, the first sub-branch adopts a global average pooling layer to capture the global guidance semantic information of the input image. In the second and the third branch, we utilize two adaptive average pooling layers with different kernel size to generate feature maps with 3×3 and 5×5 spatial resolution, respectively. This is conducive to gather semantic information of different respective filed. Besides, we apply the identity mapping layer to combine the advanced semantic information from the above three sub-branches with the original feature map.

After we collect global semantic information by the PPM, GGFs is applied to transfer the global semantic information to shallow layers of the pyramid. This operation covers the shortage that the semantic information is progressively weakened from the top-down pathway of FPN. The green arrow in Fig. 1 indicates the GGFs.

C. Feature Aggregation Module

In the architecture of FPN, the low-resolution future map is upsampled and then fused with the adjacent higher resolution one by element-wise sum operation, which leads to the aliasing effect. The FAM is proposed to address the above problem and availably merges the high-level feature maps with the low-level feature maps simultaneously. As depicted in Fig. 4, the FAM contains four parallel branches. In the inference phase, We apply the average pooling layers with different downsampling rates to resize the input feature maps into different resolutions. Then a 3×3 convolutional layer is employed to further integrate futures of each downsample branch. The 3×3 convolution kernel is sufficient to capture changes in features and only brings a small amount of parameters. In order to ensure that the resolution of the input feature map is unchanged, we upsample the feature maps of different branches back to the original resolution. These feature maps are merged by the element-wise sum operation, following a 3×3 convolutional layer. By using different sampling rates (2,4,8), the FAM enables each spatial position of the feature map to observe the local context through different scale-spaces, which are 1, 0.5, 0.25, 0.125 ratios of the original feature map spatial size respectively, thereby further expanding the receptive field of the whole network.

D. Loss Function

The loss function consists of two parts: classification loss and regression loss. We use the focal loss as classification loss to solve the problem of imbalance between positive and negative samples. It can be defined as follows:

$$FL\left(p_t\right) = -\alpha_t \left(1-p_t\right)^{\gamma} \log\left(p_t\right) \tag{1}$$

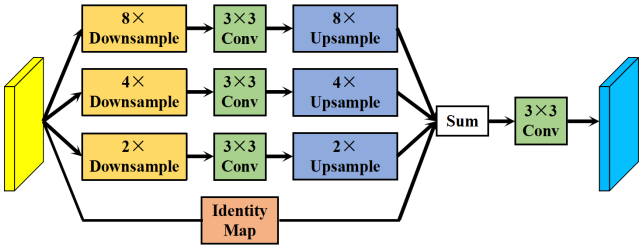


Fig. 4: Detailed illustration of the feature aggregation module (FAM).

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{otherwise} \end{cases} \tag{2}$$

where p and y stands for the classification probability and the ground-truth label of interested object, respectively. The α and γ is the variable factor. In this paper, α and γ are 0.25 and 2 respectively.

As for the regression loss, we use the smooth L_1 loss, which is expressed as follows:

$$L_{reg}\left(t_i, t_i^*\right) = \text{smooth}_{L_1}\left(t_i - t_i^*\right) \tag{3}$$

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{4}$$

where t and t^* are calculated by the following equation:

$$\begin{aligned} t &= \left[\frac{(x-x_a)}{w_a}, \frac{(y-y_a)}{h_a}, \log\left(\frac{w}{w_a}\right), \log\left(\frac{h}{h_a}\right)\right] \\ t^* &= \left[\frac{(x^*-x_a)}{w_a}, \frac{(y^*-y_a)}{h_a}, \log\left(\frac{w^*}{w_a}\right), \log\left(\frac{h^*}{h_a}\right)\right] \end{aligned} \tag{5}$$

where x, y, w, h represent the center coordinates, width, and height of the box. Variables x_a, x, and x^* correspond to the prediction box, anchor box and ground truth box, respectively. (y, w, and h are the same.) Therefore, the loss function of the whole network is:

$$L = \lambda FL + (1-\lambda) L_{reg} \tag{6}$$

where λ is the weighting factor.

III. Experiments

A. Datasets

We evaluate our proposed method on the NWPU VHR-10 dataset and the RSOD dataset. The details of these two datasets are described as follow:

- ***NWPU VHR-10:*** The dataset contains 800 VHR optical RSIs, of which 715 color images are obtained from Google Earth with a spatial resolution range of 0.5 to 2 meters and 85 color infrared images are obtained from Vaihingen data with a spatial resolution of 0.08 meters. The whole dataset consists of 10 categories, including airplanes, ships, tanks, baseball fields, tennis courts, basketball courts, ground track fields, harbors, bridges, and vehicles. There are 650 positive image sets and 150 negative image sets. In this experiment, we only

choose the 650 positive images and random select 80% as the training set and 20% as the testing set.

- ***RSOD:*** The RSOD dataset comes from Google Earth and Tianditu and is divided into four categories: oil tank, aircraft, overpass, and playground. The annotated bounding box contains a total of 4,993 aircraft, 191 playgrounds, 180 overpasses, and 1,586 tanks. In this experiment, 561 images are used for the training set and 375 images are used for the testing set.

B. Implementation Details

In all the experiments, we keep the input image at least 608 in height and no more than 1024 in width. We adopt the ResNet101 as our backbone network. Following RetinaNet, we use the feature maps generated from P_3 to P_7 to predict objects. On these pyramid levels, anchors with an area of 32^2 to 512^2 are set. At each pyramid level, anchors have three different aspect ratios $\{1{:}\ 2, 1{:}\ 1, 2{:}\ 1\}$ and three different sizes $\left\{2^0, 2^{1/3}, 2^{2/3}\right\}$. For the focal loss, we follow the default setting in [9] (e.g. $\gamma = 2, \alpha = 0.25$) and the weighting factor of the loss function λ is set to 0.5. When the IoU of an anchor and any ground truth box is greater than or equal to 0.5, we divide the anchor into the positive class. Otherwise, we divide the anchor into the negative class. In the training phase, we train our model on per GPU with batch size 1 for 50 epoch and use the Adam optimizer with the initial learning rate of 1e-5. In the inference phase, we use a threshold of classification score of 0.05 to filter out the background bounding boxes and only preserve the first 1K bounding boxes. Then, we use the Non-Maximum Suppression (NMS) with an IoU threshold value of 0.5 to obtain the top 100 bounding boxes for each image as the final prediction box of the network. The IoU refers to the ratio of the intersection and union of the areas of the ground truth box and the prediction box.

C. Evaluation Metrics

TABLE I: Ablative study of our proposed modules

	FAM	GGM	mAP	
			NWPU	RSOD
RetinaNet			89.91	92.35
	✓		91.18	94.46
		✓	91.31	93.16
	✓	✓	**91.97**	**96.23**

We evaluate the performance of the model by comparing the values of average precision (AP) and mean average precision (mAP). The AP calculates the mean of the Precision of the Recall interval from 0 to 1. The mAP is the average of AP values for all categories.

The definitions of Precision and Recall are as follows:

$$Precision = \frac{TP}{(TP + FP)} \tag{7}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{8}$$

where *TP* and *FN* indicate the number of positive samples that are correctly divided into positive samples and incorrectly divided into negative samples, respectively. And *FP* indicates the number of negative samples that are incorrectly divided into positive samples.

D. Experiment Result and Analysis

1) Ablation Analysis: To evaluate the effectiveness of our proposed two module, we report the ablation analysis in Table I on both NWPU and RSOD dataset.

- *GGM Only.* First, global semantic information is introduced into the current layer of the pyramid. The feature maps of adjacent higher layers are upsampled and then added to the current layer. Finally, a 3×3 convolutional layer is used to obtain the fused feature map. In this case, the global semantic information is transmitted to the lower layers, which enriches the semantic information of the feature maps in the lower layers of the pyramid. As shown in Table I, for the NWPU VHR-10 dataset, our method increase mAP by 1.4%. For the RSOD dataset, the mAP of our network increased by 1.02%.
- *FAM Only.* The feature map of the higher layer is upsampled and added to the feature map of the adjacent lower layer, and the result of the addition is further integrated through the FAM module. In this way, we better integrate information from the upper and lower layers and strengthen the information interflow between different layers. On the NWPU VHR-10 dataset and the RSOD dataset, the mAP of the model improved by 1.27% and 2.32%, respectively.
- *FAM & GGM.* By combining the FAM and the GGM, the high-level semantic information and the low-level location information are adequately merged so that the feature map of each layer contains more comprehensive information. This allows our network to obtain discriminative features for subsequent regression and classification, reducing misclassification and precisely locating in a complex background. On the NWPU VHR-10 and RSOD datasets, the mAP improved by 2.06% and 3.88%, respectively.

2) Comparisons to the State-of-the-Arts: Table II and Table III demonstrate the overall comparison results on both NWPU and RSOD datasets. For the NWPU dataset, our proposed method achieves the best performance and increases 2.06% compared to the baseline. The visualization results are shown in the Fig. 5. The AP of oil tanks, harbors, and vehicles has been remarkably improved, increasing by 10.97%, 6.4%, and 3.22%, respectively. In Fig. 7(a), RetinaNet mistakenly detects some objects on the water surface as ships. The network notices its background but ignores low-level information such as the outline of the ship. While our network strengthens the fusion of low-level information and high-level information through the application of the FAM, which can reduce the mis-classification of such objects. The low-level information such as contours and textures of oil tanks are less than other categories. The original network did

TABLE II: Performance of our method and other methods on NWPU VHR-10 dataset.

Model	airplane	ship	storage tanks	baseball diamonds	tennis courts	basketball courts	ground track fileds	harbors	bridges	vehicles	mAP
YOLOv1-448 [7]	69.85	42.74	11.04	86.58	42.74	67.63	88.72	55.44	76.46	42.75	58.39
YOLOv3-416 [17]	99.55	81.82	80.30	98.26	80.56	81.82	**99.47**	74.31	89.61	86.98	87.27
SSD-300 [8]	90.63	67.40	71.05	98.59	69.27	79.52	99.24	84.38	87.41	75.20	82.27
SSD-512 [8]	99.49	83.10	**86.00**	95.02	84.25	84.44	93.53	77.29	73.98	**93.00**	87.01
RFBNet-300 [18]	99.48	66.69	75.99	90.91	78.35	88.36	99.24	**87.73**	89.09	86.02	86.19
RFBNet-512 [18]	**99.87**	85.37	81.82	97.66	89.62	**98.73**	98.60	86.94	87.70	89.27	91.56
RetinaNet	99.84	91.20	64.96	99.80	96.96	98.52	98.92	80.12	93.86	74.92	89.91
Ours	99.37	**91.91**	75.93	**99.96**	**98.33**	95.81	98.90	86.52	**94.79**	78.14	**91.97**

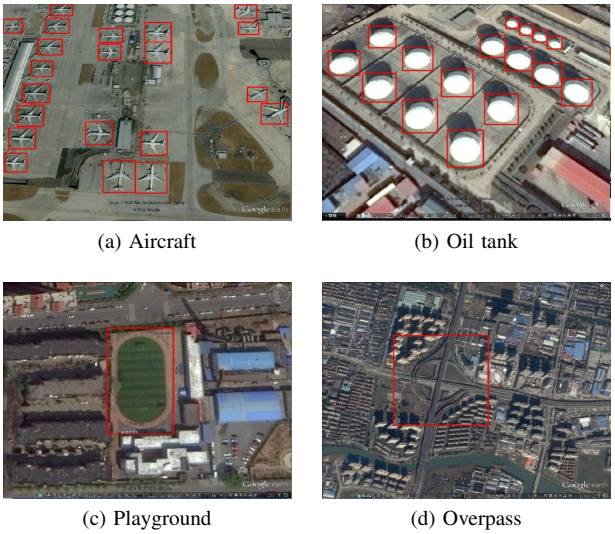
(a) Oil Tanks (b) Harbors (c) Ground Track Fileds and Baseball Diamonds (d) Airplanes (e) Vehicles

(f) Basketball Courts (g) Ships (h) Bridges (i) Tennis Courts (j) Baseball Diamonds

Fig. 5: Qualitative visualization of our method on the NWPU VHR-10 dataset.

TABLE III: Performance of our method and other methods on RSOD dataset

Method	Aircraft	Playground	Oil tank	Overpass	mAP
Faster R-CNN ResNet101 [4]	83.54	97.81	98.11	88.62	92.02
SSD300 VGG16 [8]	71.89	98.58	90.72	90.21	87.85
RFCN ResNet101 [5]	83.69	99.54	98.44	94.42	94.03
YOLOv3 DarkNet53 [17]	88.38	99.65	**98.91**	**96.64**	95.97
NAS-FPN ResNet101 [19]	89.88	97.88	92.50	89.37	92.41
RetinaNet ResNet101	81.34	98.69	97.03	92.32	92.35
Ours	**93.72**	**99.80**	97.38	94.05	**96.23**

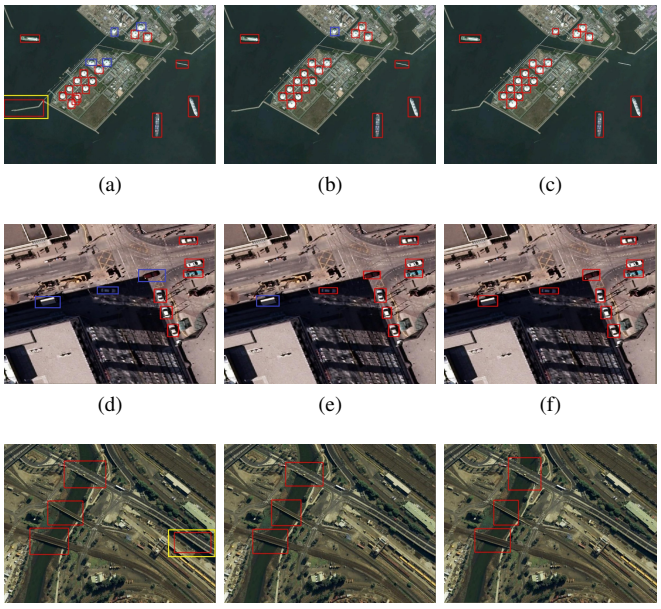


(a) Aircraft (b) Oil tank (c) Playground (d) Overpass

Fig. 6: Qualitative visualization of our method on RSOD dataset.

not perform well in detecting such simple categories when learning other categories with obvious details. In Fig. 7(a), RetinaNet caused a lot of missed detections of tanks due to the lack of feature information, and our network can effectively alleviate this deficiency by introducing higher-level semantic information through the application of the GGM. In Fig. 7(d), the interference of shadow occlusion and dim color makes RetinaNet fail to recognize these vehicles, and our network is capable of learning the general feature of vehicles because of introducing global semantic information. Compared to ground truth, our network does not detect the square vehicle in Fig. 7(e) because such sample are too rare in the training set. In Fig. 7(g), RetinaNet mistakes the highway as a bridge, the FAM enlarges the receptive field of the network to enhance context information, which can reduce such misdetection problems.

As for the RSOD dataset, our method surpasses these famous high-performance detectors such as Faster-RCNN, SSD, NAS-FPN, etc., and reaches the highest mAP of 96.23%, which is 3.88% higher than the RetinaNet. We demonstrate the detection results in Fig. 6. It is apparent from Table III

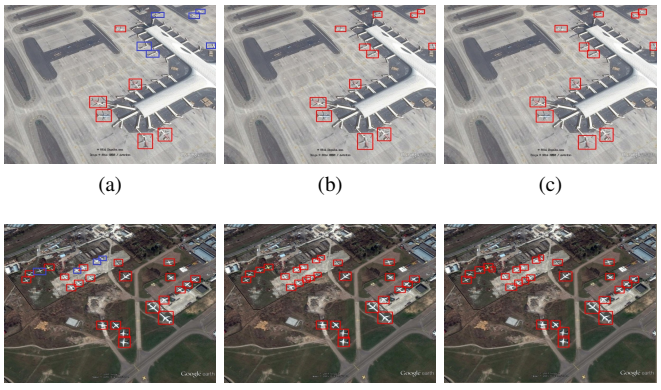


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 7: The comparison results of RetinaNet and our method with proposed DFPN on NWPU VHR-10 dataset. The three columns are detection results of RetinaNet, our method, and the ground-truth, respectively. Yellow boxes indicate false detections, and blue boxes indicate missed detections.



(a) (b) (c)

(d) (e) (f)

Fig. 8: The comparison of aircraft detection results of RetinaNet and our method with proposed DFPN on RSOD dataset. The three columns are detection results of RetinaNet, our method, and the ground-truth, respectively. The blue boxes indicate missed detections.

that our method has made definite improvements in almost all categories. Notably, the improvement of the aircraft is very significant and it increases more than 10%. We consider the background of the aircraft in the RSOD dataset is more complicated, indicating the superior performance of our method for complex backgrounds. In Fig. 8(a), interference from similar objects in the vicinity of the aircraft and the blur caused by the small size make the RetinaNet fail to detect these aircraft. . In Fig. 8(b), there are houses and highways near the aircraft. In such a messy background, RetinaNet dose not recognize these aircraft. By applying the GGM, our network introduces global semantic information to each layer of the pyramid and then uses the FAM to seamlessly combine high-level information with low-level information to obtain discriminative features, so that our network can distinguish the objects from the complex background to reduce missed and false detections caused by background mis-classification.

IV. Conclusion

In this paper, we propose a discriminative feature pyramid network for RSIs object detection by introducing a GGM and a FAM, the GGM captures and processes the global semantic information from high-level, and delivers it to shallower layers to enhance the semantic information of feature maps. The FAM adequately merges high-level semantic information and shallow location information. By applying these two modules, the feature map integrates comprehensive information to obtain discriminative features and effectively mitigate interference of complex backgrounds in remote sensing images. Experiments on the NWPU and RSOD datasets prove the effectiveness of our discriminative feature pyramid network.

References

[1] K. Kuru and W. Khan, "Novel hybrid object-based non-parametric clustering approach for grouping similar objects in specific visual domains," *Applied Soft Computing*, vol. 62, pp. 667–701, 2018.

[2] X. Zhang, X. Wang, X. Tang, H. Zhou, and C. Li, "Description generation for remote sensing images using attribute attention mechanism," *Remote Sensing*, vol. 11, no. 6, p. 612, 2019.

[3] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

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

[5] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in neural information processing systems*, 2016, pp. 379–387.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*, 2016, pp. 21–37.

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

[10] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 12, pp. 7405–7415, 2016.

[11] K. Li, G. Cheng, S. Bu, and X. You, "Rotation-insensitive and context-augmented object detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2337–2348, 2017.

[12] Y. Zhong, X. Han, and L. Zhang, "Multi-class geospatial object detection based on a position-sensitive balancing framework for high spatial resolution remote sensing imagery," *ISPRS journal of photogrammetry and remote sensing*, vol. 138, pp. 281–294, 2018.

[13] Z. Zhang, X. Chen, J. Lie, and K. Zhou, "Rotated feature network for multi-orientation object detection," *arXiv preprint arXiv:1903.09839*, 2019.
[14] Z. Deng, H. Sun, S. Zhou, J. Zhao, L. Lei, and H. Zou, "Multi-scale object detection in remote sensing imagery with convolutional neural networks," *ISPRS journal of photogrammetry and remote sensing*, vol. 145, pp. 3–22, 2018.
[15] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," *arXiv preprint arXiv:1904.09569*, 2019.
[16] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.
[17] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.
[18] S. Liu, D. Huang *et al.*, "Receptive field block net for accurate and fast object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 385–400.
[19] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "Nas-fpn: Learning scalable feature pyramid architecture for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7036–7045.