ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746853

SINGLE-SHOT BALANCED DETECTOR FOR GEOSPATIAL OBJECT DETECTION

Yanfeng Liu, Qiang Li, Yuan Yuan, Qi Wang[*]

School of Computer Science and School of Artificial Intelligence, Optics and Electronics (iOPEN),
Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China

ABSTRACT

Geospatial object detection is an essential task in remote sensing community. One-stage methods based on deep learning have faster running speed but cannot reach higher detection accuracy than two-stage methods. In this paper, to achieve excellent speed/accuracy trade-off for geospatial object detection, a single-shot balanced detector is presented. First, a balanced feature pyramid network (BFPN) is designed, which can balance semantic information and spatial information between high-level and shallow-level features adaptively. Second, we propose a task-interactive head (TIH). It can reduce the task misalignment between classification and regression. Extensive experiments show that the improved detector obtains significant detection accuracy with considerable speed on two benchmark datasets.

Index Terms— Geospatial object detection, one-stage detector, multi-scale balance learning, task-interactive head

1. INTRODUCTION

Geospatial object detection in remote sensing images (RSIs) is a fundamental task for earth observation. It has received more and more attention because of its wide applications, such as disaster monitoring [1], land cover classification [2], building extraction [3], etc. Currently, geospatial object detection is still a challenging problem mainly due to the imbalanced multi-scale objects and time-consuming computation costs.

In recent years, the development of deep learning has greatly promoted the research of object detection, and many excellent algorithms have been proposed [4–10]. These methods are divided into two categories: two-stage and one-stage. The typical two-stage methods [4, 5, 7] achieve excellent detection performance on various public natural scene datasets. Unfortunately, they involve complex computation in two stages and cannot achieve real-time detection for RSIs. Recently, CBD-E [11] and CSFF [12] are designed for geospatial object detection specifically, but they still suffer from complex calculations (as shown in Table 1). One-stage approaches execute the bounding boxes regression directly and obtain more considerable running speeds than two-stage methods. However, these algorithms cannot perform well in RSIs (e.g., [6, 10, 13, 14] as shown in Table 1). We find that there are two challenging problems in RSIs for one-stage detectors: 1) the imbalance of multi-scale features is further aggravated due to the complicated background of RSIs; 2) the diversity of geospatial objects magnifies the feature imbalance between classification and localization sub-tasks.

To handle with the problems mentioned above, a single-shot balanced detector (S2BDet) is proposed in this paper. First, we propose a balanced feature pyramid network to enhance the feature representation ability for multi-scale objects. Compared with the original FPN (e.g., [5,7]), we design a multi-scale balanced module (MSBM), which can capture more effective spatial information on shallow-level features and richer channel information on high-level features. This strategy enables the fused features to select effective spatial/channel information of various scales simultaneously. Second, to alleviate the task misalignment between the classification and localization, a task-interactive head is designed. It compensates for misalignment between two sub-tasks by feature interaction without additional loss function to control the learning process.

The main contributions are summarized as follows:

- We design a balanced feature pyramid network (BFPN), which contains two MSBM blocks to capture more effective features at different scales in RSIs.
- We propose a task-interactive head (TIH). It achieves task feature alignment for one-stage detectors through feature interaction of two sub-networks.
- The proposed S2BDet achieves significant improvement on two remote sensing datasets, which has the most excellent performance in keeping speed/accuracy trade-off compared with several state-of-the-arts.

2. RELATED WORK

2.1. Feature Pyramid Network for Object Detection

Feature Pyramid Network (FPN) [5] is a classical method for multi-scale object detection by using upsampling and element-wise summation to coalesce feature maps with different scales. RetinaNet [6] and YOLOv3 [14] also adopt this strategy. Although FPN improves the performance of various

[*] Qi Wang is the corresponding author. This work was supported by the National Natural Science Foundation of China under Grant U21B2041, U1864204, 61632018, and 61825603.

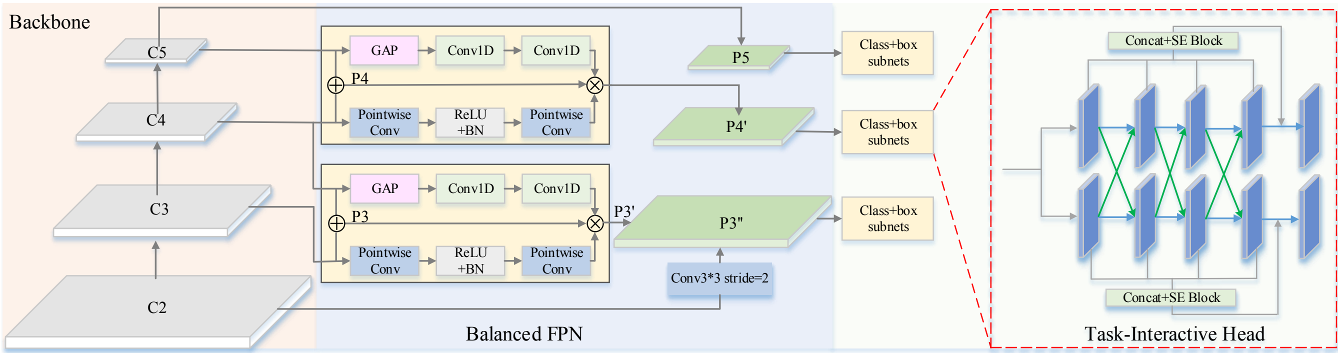

Fig. 1. Illustration of our proposed S2BDet. P_6 and P_7 have been omitted due to space constraints. Like RetinaNet, P_6 is obtained from C_5 by 3×3 convolution of stride = 2, and P_7 is obtained from P_6 by 3×3 convolution of stride = 2.

detectors, its simple feature fusion method limits the improvement of multi-scale detection performance. PAFPN [7] redesigns FPN and achieves better prediction results than FPN [5]. Recently, Dai *et al.* [15] propose an attention fusion strategy for FPN. Its effectiveness has been verified in multiple tasks [16]. It provides a new idea to solve the multi-scale problem of object detection by attention mechanisms.

2.2. Task Misalignment of One-Stage Detectors

Recent one-stage detectors [6, 17] predict two separate outputs by two different sub-networks to classify and regress bounding boxes. This approach can make the network independently to learn the two sub-tasks, reducing the learning difficulty via separate loss functions. However, Feng *et al.* [18] mention that the spatial features learned by them are different because of the divergence of loss functions and learning mechanisms for classification and localization. To overcome this misalignment, a sample assignment scheme and a task-aligned loss function are designed in [18]. Different from that, we propose a task-interactive head to alleviate this inconsistency without additional loss to control the learning process.

3. METHODOLOGY

The proposed S2BDet is summarized in Fig. 1. It applies two MSBM blocks to generate a multi-scale convolutional feature pyramid (BFPN). Then, the TIH consists of two sub-networks, one for classifying anchors and the other for regressing from anchors to ground-truth bounding boxes.

3.1. Balanced Feature Pyramid Network

As for the original FPN, RetinaNet [6] uses feature pyramid levels P_3 to P_7, where P_3 to P_5 are computed from C_3 through C_5 of backbone, i.e.,

$$P_i = \text{conv3} \times 3(\text{Bilinear2x}(P_{i+1}) \oplus \text{conv1} \times 1(C_i)), \quad (1)$$

where $\oplus$ indicates element-wise summation. Note that P_5 is obtained by only C_5 as illustrated in [5]. However, this direct addition strategy does not take into account the imbalance between semantic features at different scales. As discussed in [15, 16], the high-level features have rich channel information while the low-level features have more detailed spatial information. Inspired by that, we propose the MSBM blocks to capture more spatial information and richer channel information during multi-scale feature fusion.

As illustrated in Fig. 1, the MSBM block adopts global average pooling (GAP) and two 1D convolutional layers with non-linear function to explore channel attention from high-level features of backbone. For P_i, the channel attention CA_i can be defined as

$$CA_i = \text{conv1D}(\text{ReLU}(\text{conv1D}(\text{GAP}(C_{i+1})))), \quad (2)$$

where $\text{conv1D}(\cdot)$ denotes the operation of 1D convolution as defined in [19], and $\text{ReLU}(\cdot)$ indicates the ReLU function. Considering that the low-level feature maps have more detailed spatial information, MSBM block deploys two pointwise convolutional layers to capture spatial attention SA_i, which can be defined as

$$SA_i = \text{PointConv}(\text{ReLU}(\text{PointConv}(C_i))), \quad (3)$$

where $\text{PointConv}(\cdot)$ indicates the function of pointwise convolution [20]. With estimated channel fusion weights CA_3, CA_4 and spatial fusion weights SA_3, SA_4, the refined integrated features can be generated as

$$P_i^{'} = CA_i \otimes P_i \otimes SA_i \quad (i = 3, 4), \quad (4)$$

where $\otimes$ denotes element-wise multiplication.

Furthermore, we discover that RetinaNet does not take full advantage of the shallow features of C_2. In BFPN, we utilize a simple 3×3 convolutional layer with stride = 2 to integrate C_2 into the pyramid feature $P_3^{'}$ by element-wise multiplication, i.e.,

$$P_3^{''} = C_{stride=2}(C_2) \otimes P_3^{'}, \quad (5)$$

where $C_{stride=2}$ indicates the operation of 3×3 convolution with stride = 2. To be consistent with RetinaNet, we gather $P_3^{''}$, $P_4^{'}$, P_5, P_6 and P_7 as final detection features.

Table 1. Comparison with representative detectors on DIOR dataset [21]. Best results are marked in bold.

Method	AL	AT	BF	BC	BG	CM	DM	EA	ES	GC	GF	HB	OP	SP	SD	ST	TC	TS	VH	WM	mAP	time(ms)
Two-stage methods																						
FRCNN [4]	53.6	49.3	**78.8**	66.2	28.0	70.9	62.3	69.0	55.2	68.0	56.9	50.2	50.1	27.7	73.0	39.8	75.2	38.6	23.6	45.4	54.1	-
FPN [5]	54.0	74.5	63.3	80.7	44.8	72.5	60.0	75.6	62.3	76.0	76.8	46.4	57.2	71.8	68.3	53.8	81.1	59.5	43.1	81.2	65.1	112.5
PANet [7]	63.0	69.6	71.9	81.3	45.9	72.3	52.5	62.2	63.2	69.3	79.3	47.4	58.2	72.0	**73.9**	**70.5**	87.1	53.1	**54.1**	85.8	66.6	169.5
CBD-E [11]	54.2	77.0	71.5	87.1	44.6	75.4	63.5	76.2	65.3	79.3	79.5	47.5	59.3	69.1	69.7	64.3	84.5	59.4	44.7	83.1	67.8	277.8
CSFF [12]	57.2	79.6	70.1	87.4	**46.1**	76.6	62.7	82.6	73.2	78.2	**81.6**	50.7	59.5	73.3	63.4	58.5	85.9	**61.9**	42.9	86.9	68.0	101.0
One-stage methods																						
YOLOv3 [14]	**72.2**	29.9	74.0	78.6	31.2	69.7	26.9	48.6	54.4	31.1	61.1	44.9	49.7	**87.4**	70.6	68.7	87.3	29.4	48.3	78.7	57.1	36
YOLOv4 [13]	71.3	51.2	66.5	86.9	33.2	72.7	45.4	55.9	47.4	65.6	60.6	**56.0**	51.8	82.5	63.8	62.0	80.3	52.5	42.4	73.1	61.1	**25**
RetinaNet [6]	53.7	77.3	69.0	81.3	44.1	72.3	62.5	76.2	66.0	77.7	74.2	50.7	59.6	71.2	69.3	44.8	81.3	54.2	45.1	83.4	65.7	83.3
RetinaNet* [6]	54.6	83.6	73.5	87.9	40.1	74.4	66.1	83.2	61.8	80.0	80.0	43.2	59.6	71.2	66.6	48.1	**87.5**	58.0	43.0	86.0	67.4	82.9
O^2-DNet [10]	61.2	80.1	73.7	81.4	45.2	75.8	64.8	81.2	**76.5**	79.5	79.7	47.2	59.3	72.6	70.5	53.7	82.6	55.9	49.1	77.8	68.4	131.2
Ours	59.8	**84.9**	73.5	**88.3**	45.2	**77.6**	**69.1**	**83.7**	71.4	**81.7**	80.6	50.1	**61.8**	71.3	68.4	50.0	87.3	54.0	43.1	**88.1**	**69.5**	87.8

RetinaNet* denotes our implementation, higher than the official performance. Airplane (AL), airport (AT), baseball field (BF), basketball court (BC), bridge (BG), chimney (CM), dam (DM), expressway service area (EA), expressway toll station (ES), golf course (GC), ground track field (GF), harbor (HB), overpass (OP), ship (SP), stadium(SD), storage tank (ST), tennis court (TC), train station (TS), vehicle (VH), and wind mill (WM).

3.2. Task-Interactive Head

Object detection contains two sub-tasks, i.e., regression and classification. Deep learning methods deploy two sub-networks to deal with two sub-tasks respectively. As illustrated in [18], there is a degree of misalignment when two separate branches are used to make predictions. Based on this discovery, we propose a task-interactive head to alleviate the task misalignment between classification and localization.

On the one hand, we adopt an efficient information interaction strategy to balance the feature layers between the two sub-networks. As shown in Fig. 1, suppose $F_i(\cdot)$, $L_i(\cdot)$ (i = 1,2,3,4) as the convolutional layers of the two sub-networks, and the output features of them can be represented as

$$Fout_{i+1} = F_{i+1}(Fout_i \oplus Lout_i), \tag{6}$$

$$Lout_{i+1} = L_{i+1}(Lout_i \oplus Fout_i) \tag{7}$$

where $Fout_i$ and $Lout_i$ denote the output features of classification and localization sub-networks, respectively. By this operation, the features of two sub-networks can be mutual supervised by each other.

On the other hand, we introduce Squeeze and Excitation (SE) blocks [22] to perform channel control as displayed in Fig. 1. It further highlights the performance of TIH through information interaction between different channels, which can be defined as

$$Fout_4^{'} = Fout_4 \otimes \text{SE}(\text{Concat}(Fout_{1,2,3,4})) \tag{8}$$

$$Lout_4^{'} = Lout_4 \otimes \text{SE}(\text{Concat}(Lout_{1,2,3,4})) \tag{9}$$

where SE$(\cdot)$ denotes the function of SE block, and Concat$(\cdot)$ indicates the operation of channel concatenation. $Fout_4^{'}$ and $Lout_4^{'}$ are utilized for minimizing the classification loss function and regression loss function, respectively.

3.3. Loss Function

The multi-task loss is used to balance classification and localization tasks [4] in object detection. With reference to RetinaNet, the loss function of S2BDet is defined as

$$L_{total} = \lambda_1 L_{cls} + \lambda_2 L_{reg}, \tag{10}$$

where L_{cls}, L_{reg} indicate classification and regression loss. For consistency, we set λ_1 = λ_2 = 1.

The focal loss proposed by RetinaNet [6] is adopted as the classification loss in our detector, i.e.,

$$L_{cls} = -\alpha_t (1 - p_t)^{\gamma} log(p_t), \tag{11}$$

where α_t and γ are hyperparameters to moderate the weights between easy and hard examples. In our experiments, we set α = 0.25 and γ = 2 as the same as RetinaNet. p_t is defined as

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise}, \end{cases} \tag{12}$$

where $y = 1$ specifies the ground-truth and p is the estimated probability for the category. Besides, we introduce Smooth L1 loss as L_{reg} referring to Faster RCNN [4].

4. EXPERIMENTAL RESULTS

4.1. Experiments Setup

Datasets: The datasets used for evaluation are DIOR [21] and HRRSD [23]. DIOR is the largest dataset for horizontal geospatial target detection. It includes 23463 images with 192472 instances of 20 classes, which is divided into 11725 images as training subset and 11738 images as testing subset. HRRSD contains 21761 images with 55740 objects of 13 categories. It is segmented into 10818 images for training and 10943 images for testing.

Evaluation Metrics: We introduce the average precision of each class (AP), the mean average precision of all classes (mAP) and running time per image for evaluation metrics.

Training Settings: Our experimental environment is PyTorch framework in Ubuntu 18.04 operating system, and all experiments are performed on 4 NVIDIA GTX 1080Ti GPUs. We adopt stochastic gradient descent algorithm (SGD) for optimizing parameters, and total epochs is 20. The initial learning rate of SGD is 0.02. The weight decay and momentum of SGD are 0.0001 and 0.9, respectively. In all experiments, we employ 0.5 horizontal flips as data augmentation.

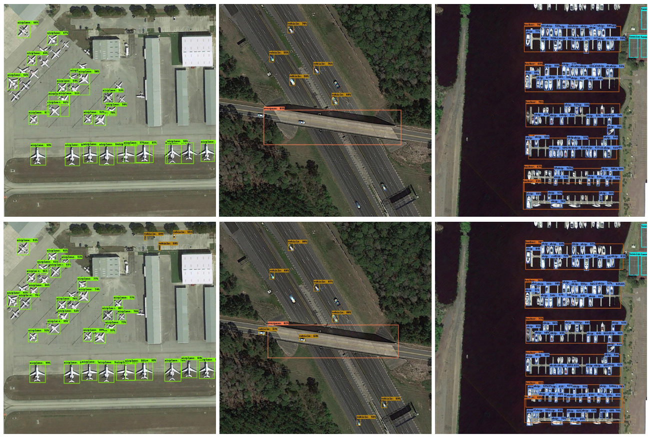

Fig. 2. Typical comparative detection results on DIOR [21]. The top line is the detection results of the baseline while the bottom line is the detection results of S2BDet.

4.2. Ablation Study

We set up ablation experiments on DIOR test set as shown in Table 2. RetinaNet [6] is the baseline which reaches 67.44 mAP on DIOR. To detect multi-scale objects in RSIs effectively, we propose the BFPN. Our method only with BFPN obtains 69.05% mAP (1.61% ↑). It illustrates that the strategy of spatial localities and channels of multi-scale features via BFPN indeed contributes to the network for interpreting RSIs. Furthermore, our method achieves 68.69% mAP when removing the integration of C_2 in BFPN, which shows the effectiveness of combining the shallow feature C_2 into BFPN. Our method shows 68.62% mAP when adding TIH to replace the original detection head. It reflects that the employment of TIH reduces the feature misalignment between two sub-tasks, and thus a considerable detection improvement is gained. When integrating both BFPN and TIH into baseline, we observe that the detection accuracy is further improved, reaching 69.49% mAP (2.05% ↑). Meanwhile, our method has corresponding improvements on AP$_{75}$, AP$_s$, AP$_m$ and AP$_l$. Fig. 2 demonstrates the typical comparative detection results on DIOR. It is clear that S2BDet is able to detect more geospatial objects than baseline, especially small vehicles, airplanes and ships.

4.3. Comparison with the State-of-the-Arts

We compare our method with several state-of-the-arts in Tables 1, 3 on DIOR [21] and HRRSD [23]. As for DIOR, our approach achieves the best results in eight of the total 20 categories and reaches the running speed of 87.8ms per image (only 4.9ms ↓ than baseline). The mAP of S2BDet is superior to two-stage algorithms (e.g., CBD-E [11], CSFF [12]) and one-stage method (O^2-Det [10]) which are designed specifically for remote sensing imagery. Among them, CBD-E and CSFF is time-consuming because of the complex computation of two stages (277.8ms and 101.0ms per image respectively). In addition, O^2-Det does not pay attention to the feature misalignment of the detection head, and its performance of mAP and running time are not as good as our method. Although YOLOv3,v4 [13, 14] achieve fastest running speeds on DIOR (36ms and 25ms per image respectively), their detection accuracies are too low. Compared with the two-stage methods [4, 5, 7, 11, 12], S2BDet has advantages not only in running speed but also in detection accuracy. In conclusion, S2BDet has the most excellent performance in keeping speed/accuracy trade-off compared with other approaches.

Table 2. Ablation experiments on DIOR test set [21]. Best results are marked in bold. "⋇" denotes BFPN without integration of C_2.

BFPN	TIH	mAP	AP$_{75}$	AP$_s$	AP$_m$	AP$_l$
×	×	67.44	49.68	8.71	35.83	65.82
⋇	×	68.69	50.67	9.33	37.30	66.79
✓	×	69.05	50.86	9.93	37.15	66.91
×	✓	68.62	49.16	9.13	35.67	65.19
✓	✓	**69.49**	**51.35**	**9.94**	**37.34**	**67.14**

Table 3. Comparison with representative detectors on HRRSD dataset [23]. Best results are marked in bold.

Method	YOLOv3	FCOS	FRCNN	RetinaNet	Ours
mAP	71.80	82.26	83.10	83.53	**85.20**

With the respect of HRRSD [23], we compare our method with several representative detectors such as YOLOv3 [14], FCOS [17], Faster RCNN [4] and RetinaNet [6]. Among competitors, our method is 1.67% mAP higher than the baseline as shown in Table 3, which illustrates the effectiveness of our proposed modules.

5. CONCLUSION

In this paper, an improved detector S2BDet with two novel upgrades based on RetinaNet is proposed for remote sensing imagery. We conduct ablation and comparison experiments to testify the function of the proposed balanced feature pyramid network and task-interactive head. Our detector outperforms many state-of-the-art algorithms on two large geospatial object detection datasets. Compared with them, S2BDet breaks the performance disadvantage of one-stage algorithms and reaches state-of-the-art results. Besides, the presented detector maintains a considerable speed advantage because the proposed modules have fewer parameters. In the future, we will explore a task alignment learning strategy based on one-stage detectors to further improve the prediction performance for geospatial object detection.

6. REFERENCES

[1] Qi Wang, Zhenghang Yuan, Qian Du, and Xuelong Li, "GETNET: A General End-to-End 2-D CNN Framework for Hyperspectral Image Change Detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 3–13, 2019.

[2] Yanfeng Liu, Qiang Li, Yuan Yuan, Qian Du, and Qi Wang, "ABNet: Adaptive Balanced Network for Multi-scale Object Detection in Remote Sensing Imagery," *IEEE Trans. Geosci. Remote Sens.*, 2021.

[3] Qi Wang, Jianzhe Lin, and Yuan Yuan, "Salient Band Selection for Hyperspectral Image Classification via Manifold Ranking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1279–1289, 2016.

[4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.

[5] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, et al., "Feature Pyramid Networks for Object Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2117–2125.

[6] Tsung-Yi Lin, Priya Goyal, Ross Girshick, et al., "Focal Loss for Dense Object Detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2999–3007.

[7] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia, "Path Aggregation Network for Instance Segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 8759–8768.

[8] Zhiguo Li, Yuan Yuan, and Dandan Ma, "Selection Based on Statistical Characteristics for Object Detection," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2021, pp. 1485–1489.

[9] Tianyuan Wang, Can Ma, Haoshan Su, and Weiping Wang, "CSPN: Multi-Scale Cascade Spatial Pyramid Network for Object Detection," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2021, pp. 1490–1494.

[10] Haoran Wei, Yue Zhang, Zhonghan Chang, Hao Li, Hongqi Wang, and Xian Sun, "Oriented objects as pairs of middle lines," *ISPRS J. Photogramm. Remote Sens.*, vol. 169, pp. 268–279, 2020.

[11] Jun Zhang, Changming Xie, Xia Xu, Zhenwei Shi, and Bin Pan, "A Contextual Bidirectional Enhancement Method for Remote Sensing Image Object Detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4518–4530, 2020.

[12] Gong Cheng, Yongjie Si, Hailong Hong, Xiwen Yao, and Lei Guo, "Cross-scale feature fusion for object detection in optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 3, pp. 431–435, 2021.

[13] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[14] Joseph Redmon and Ali Farhadi, "Yolov3: an incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[15] Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard, "Attentional Feature Fusion," in *IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2021, pp. 3559–3568.

[16] Yimian Dai, Yiquan Wu, Fei Zhou, and Kobus Barnard, "Asymmetric Contextual Modulation for Infrared Small Target Detection," in *IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2021, pp. 949–958.

[17] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He, "FCOS: Fully Convolutional One-Stage Object Detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 9626–9635.

[18] Chengjian Feng, Yujie Zhong, and Yu Gao, "TOOD: Task-aligned One-stage Object Detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2021.

[19] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 11531–11539.

[20] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[21] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, 2020.

[22] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-Excitation Networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7132–7141.

[23] Yuanlin Zhang, Yuan Yuan, Yachuang Feng, and Xiaoqiang Lu, "Hierarchical and Robust Convolutional Neural Network for Very High-Resolution Remote Sensing Object Detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5535–5548, 2019.