Multi-Vision Network for Accurate and Real-Time Small Object Detection in Optical Remote Sensing Images

Wenxuan Han, Alifu Kuerban, Yuchun Yang, Zitong Huang, Binghui Liu, and Jie Gao

***Abstract*— Accurate and real-time detection of airplanes, cars, and ships in remote sensing images is an important but challenging task that plays an important role in both military and civilian life. The most challenging issues posed by this task are the intensive and tiny size of objects and the complexity of application scenarios. In this letter, we propose a multi-vision small object detector that can rapidly and accurately detect airplanes, cars, and ships in remote sensing images. We make the following three contributions: a multiscale residual block (MRB) is proposed, whereby dilated convolution is employed in a cascade residual block to capture multiscale context information, thus improving the feature representation ability of convolutional neural networks; a multiscale receptive field enhancement module (MRFEM) is proposed that combines features obtained using dilated convolution at different dilation rates to further enhance the multiscale feature representation of the remote sensing targets; and a multi-vision network (MVNet) is presented that uses multiple low-level feature maps with multi-branch convolution to detect small objects. Experimental results show that the proposed method can achieve a significant mean average precision (mAP) of 94.70% in remote sensing images and can run at 24 FPS on a single NVIDIA 1080Ti GPU.**

***Index Terms*— Multiscale receptive field, multiscale residual learning, remote sensing image, small object detection.**

I. Introduction

THE development of high spatial resolution remote sensing technology resulted in an abundance of information in remote sensing images, enabling the detection and identification of geospatial objects. Object detection in remote sensing images has been extensively researched, and the developed detection methods can be used for military object identification, traffic management, resource exploration, and environmental monitoring [1]–[6].

Object detection in traditional images has been extremely successful in recent years because of the rapid development of deep learning methods, such as region-based convolutional neural networks (R-CNNs) [7]–[9], you only look once (YOLO) methods [10], [11], and single-shot detectors (SSDs) [12]. The development of the convolutional neural network (CNN) and its successful application to object detection in traditional images have recently resulted in the development of several CNN-based methods [13]–[16] that implement object detection tasks for remote sensing images. Wang *et al.* [13] proposed an end-to-end multiscale visual attention network (MS-VAN) to extract multiscale feature maps to improve the detection performance for small objects. Qu *et al.* [14] proposed an SSD-based detector called DFSSD to detect vehicles in remote sensing images. This framework utilized the structure of FPN network to fuse the low-level feature map with high resolution and the high-level feature map with rich semantic information. A hyperscale object detection framework for multiple spatial resolution (MSR) remote sensing imagery has been proposed to alleviate the extreme scale-variation problem by learning hyperscale feature representation [15]. Qin *et al.* [16] proposed a specially optimized one-stage network (SOON) that effectively extracts, understands, and analyzes a combination of feature and semantic information for small objects in remote sensing images. However, considerable room for improvement remains in terms of the accuracy and real-time performance of these models.

As previously mentioned, the accurate detection of small objects in remote sensing images in real time is a challenging task. In this letter, we propose a novel and efficient SSD to meet these challenges. The contributions of this letter are summarized below.

1) We propose a multiscale residual block (MRB) that employs dilated convolution in a cascade residual block to capture multiscale context information, thereby improving the feature representation ability of CNNs.
2) We propose a multiscale receptive field enhancement module (MRFEM) to combine features obtained by dilated convolution at different dilation rates to further enhance the feature representation of remote sensing targets.
3) We present a multi-vision network (MVNet) that uses multiple low-level feature maps with multi-branch convolution to detect small objects.

II. Proposed Method

Our goal is to leverage the SSD [12] to build a novel detection network to detect airplanes, cars, and ships in remote sensing images. Fig. 1 is an overview of the proposed MVNet framework. The proposed MVNet has two main components: an MRB and an MRFEM.

Manuscript received April 21, 2020; revised August 14, 2020, October 21, 2020, and November 27, 2020; accepted December 7, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61562084. *(Corresponding author: Alifu Kuerban.)*

Wenxuan Han, Alifu Kuerban, Zitong Huang, Binghui Liu, and Jie Gao are with the Key Laboratory of Software Engineering Technology, College of Software, Xinjiang University, Urumqi 830046, China (e-mail: 1943099816@qq.com; ghalipk@xju.edu.cn).

Yuchun Yang is with the College of Resources and Environment, Northwest A&F University, Xianyang 712100, China.

Data is available on-line at https://share.weiyun.com/prCpb8EB.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/LGRS.2020.3044422.

Digital Object Identifier 10.1109/LGRS.2020.3044422

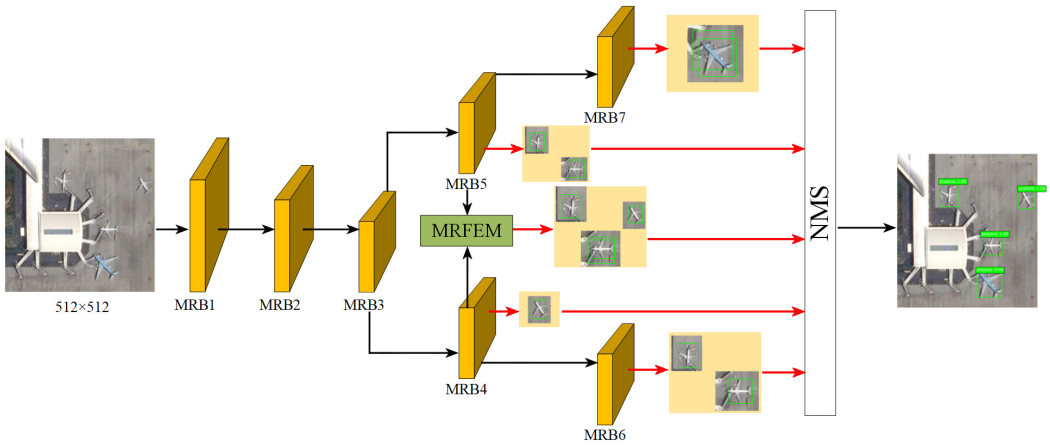

Fig. 1. Overview of the proposed MVNet. It contains two main components: MRB and MRFEM.

A. Multiscale Residual Block

In recent years, CNNs have played an important role in computer vision tasks, including classification, object detection, and image segmentation. Most CNNs use a relatively small kernel, usually 3×3, because the use of a large kernel is accompanied by a large number of parameters and high computational costs. However, a small kernel cannot cover a large image area. To meet the requirement of few parameters while covering a large image area, CNNs use a cascaded structure with a small kernel and down-sampling layers to gradually reduce the size of the input and to increase the receptive field of the network. However, increasing the number of convolutional layers results in the vanishing gradient problem. The most typical example of such a structure is the VGG network [17], which consists of 13 convolution layers. ResNet [18] introduces a residual learning framework to resolve the vanishing gradient problem. However, ResNet cannot capture sufficient context information.

We propose an MRB to capture multiscale context information. MRB uses dilated convolution [19], which can effectively improve the accuracy of object detection [14], [20], at different dilation rates to enlarge the receptive field of the kernel over that of ResNet without additional costs. Fig. 2 shows the second multiscale residual block (MRB2) used in our MVNet as an example. First, a 1×1 Conv is applied to change channels of the input feature maps, followed using three 3×3 dilated convolutions with different dilation rates: 1, 3, and 5. Then, a 1×1 Conv is used to regain the initial number of feature maps. A shortcut connection is used to facilitate identity mapping, which can address the problems of vanishing gradient and poor results of deep networks. MRB has different receptive fields: a small dilation rate corresponds to a small receptive field for details that can provide information about small objects and/or parts of the objects, whereas a large dilation rate can provide more reliable details about large objects and/or context information.

B. Multiscale Receptive Field Enhancement Module

We design a MRFEM to enhance the multiscale feature representation of remote sensing targets. MRFEM introduces rich

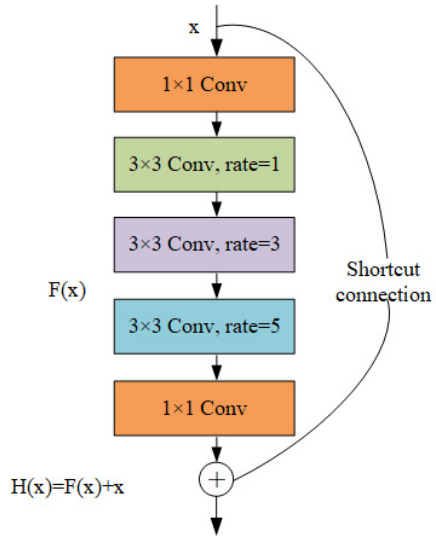

Fig. 2. Idea of multiscale residual learning.

context information by smartly combining features obtained by dilated convolution at different dilation rates, which increases the accuracy of multiscale object detection.

As shown in Fig. 3, we fuse two feature maps with the same resolution that are obtained from the dilated convolution with different dilation rates using add. Finally, a 1×1 Conv is used to realize the final spatial array of the MRFEM, which has strong multiscale feature representation.

C. Multi-Vision (MVNet) Detector

Fig. 1 shows that the proposed MVNet detector is built upon a one-stage SSD framework. We use the proposed MRB and MRFEM to design a novel backbone network and appropriate detection layers. The details of model construction are given below.

1) New Backbone Network: The backbone network of the original SSD is VGG16 [17], which is widely used in image classification and feature extraction. However, VGG16 is a deep neural network and is prone to gradient vanishing. In addition, the excessively deep convolutional layers of

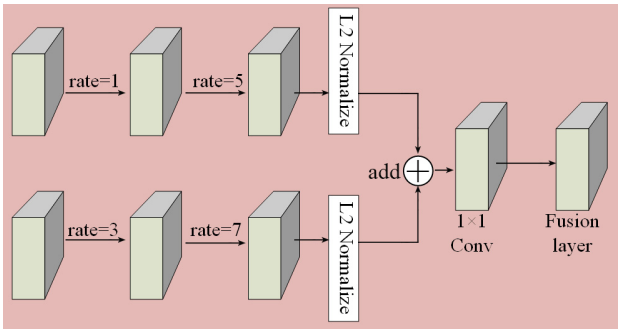

Fig. 3. MRFEM.

TABLE I
BACKBONE NETWORK DETAILS

Name	MRB	Ouput
MRB1	Conv, 1×1, 128 Conv3_1, 3×3, 128, rate=1 Conv3_2, 3×3, 128, rate=3 Conv3_3, 3×3, 256, rate=5 Conv, 1×1, 128	256×256
MRB2	Conv, 1×1, 128 Conv3_1, 3×3, 128, rate=1 Conv3_2, 3×3, 128, rate=3 Conv3_3, 3×3, 256, rate=5 Conv, 1×1, 128	128×128
MRB3	Conv, 1×1, 128 Conv3_1, 3×3, 128, rate=1 Conv3_2, 3×3, 128, rate=3 Conv3_3, 3×3, 256, rate=5 Conv, 1×1, 128	64×64
MRB4	Conv, 1×1, 64 Conv4_1, 3×3, 128, rate=1 Conv4_2, 3×3, 128, rate=5 Conv, 1×1, 64	32×32
MRB5	Conv, 1×1, 64 Conv5_1, 3×3, 128, rate=3 Conv5_2, 3×3, 128, rate=7 Conv, 1×1, 64	32×32
MRB6	Conv, 1×1, 32 Conv6_1, 3×3, 128, rate=1 Conv6_2, 3×3, 64, rate=3 Conv, 1×1, 32	16×16
MRB7	Conv, 1×1, 32 Conv7_1, 3×3, 128, rate=3 Conv7_2, 3×3, 64, rate=5 Conv, 1×1, 32	16×16

VGG16 are not useful for small object detection. We design a novel backbone based on MRB, which offers the following advantages: 1) residual learning is introduced to prevent gradient vanishing; 2) dilated convolution is used to capture the multiscale context, thereby enhancing feature representation; and 3) a multi-branch convolution structure is used to detect small objects on multiple low-level feature maps. Table I presents the architectures of the backbone network.

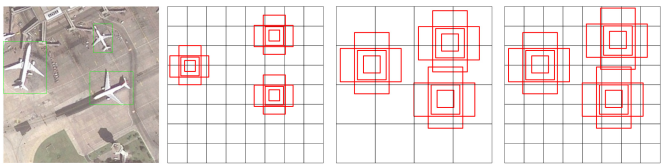

Fig. 4. Strategy of our proposed MVNet to generate anchor boxes: we not only generate anchor boxes of different sizes on feature maps of different scales [see (b) and (c)] but also generate anchor boxes of different sizes on feature maps of the same scales [see (b) and (d)]. (a) Image with ground truth boxes. (b) 8 × 8 feature map with small anchor boxes. (c) 4 × 4 feature map with large anchor boxes. (d) 8 × 8 feature map with large anchor boxes.

TABLE II
ANCHOR BOXES PARAMETERS OF EACH DETECTION LAYER

Detection layer	Size	Scales	Aspect rations
MRB4	32×32	0.02	1:1, 1:2, 2:1
MRB5	32×32	0.06	1:1, 1:2, 2:1
MRFEM	32×32	0.12	1:1, 1:2, 2:1
MRB6	16×16	0.24	1:1, 2:1, 1:2, 3:1, 1:3
MRB7	16×16	0.48	1:1, 2:1, 1:2, 4:1, 1:4

2) Detection Layers: We select MRB4, MRB5, MRFEM, MRB6, and MRB7 as the detection layers. During training, SSD generates anchor boxes [shown by the red box in Fig. 4(b) and (c)] with different scales and aspect ratios for feature maps of different scales to match the ground truth box [shown by the green box in Fig. 4(a)]. The anchor is assigned a positive label when the intersection over union (IoU) between the anchor and the ground truth box exceeds a threshold and is assigned a negative label otherwise. However, it may not be a good choice to generate anchor boxes of different sizes only for feature maps at different scales. Thus, we improve the detection accuracy by also generating anchor boxes of different sizes on feature maps at the same scales, as shown in Fig. 4(d). Thus, we have presented the essential features of our proposed multi-vision detector. Multi-branch convolution achieves the purpose of obtaining multiple feature maps at the same resolution to detect small targets on multiple shallow feature maps with many small target features [4]. In addition, we adopt the k-means clustering method to analyze the most suitable anchor scales and aspect ratios of small objects in remote sensing images. The parameters of the anchor boxes of each detection layer are shown in Table II.

III. EXPERIMENTS

In this section, we evaluate the performance of different models for our remote sensing image data set. All the experiments were conducted using an NVIDIA GTX1080TI video card, 11 GB memory, and a CPU E5-2450. We trained and tested the deep learning framework, Keras. We first evaluated existing architectures [4], [9], [11], [12], [14], [20] on our data set. In addition, we used ablation experiments to evaluate and to analyze the proposed model.

A. Data Preparation

Recall that the objective of this study is to detect airplanes, cars, and ships in remote sensing images. Public data sets,

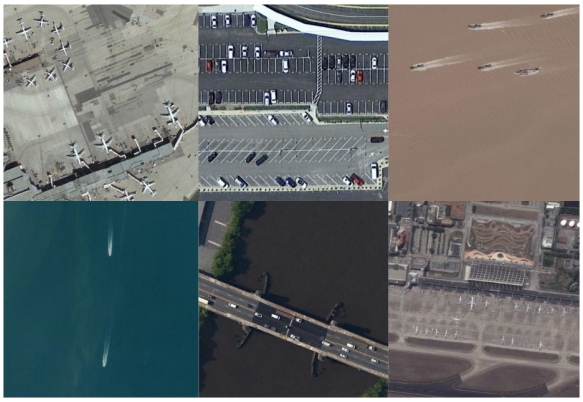

Fig. 5. llustrative examples of our data set.

TABLE III
STATISTICAL DETAILS OF DATA SET

Class	Image quantity	Instances quantity
Airplane	1500	17540
Car	500	9790
Ship	1200	4725

such as DOTA [21] and NWPU VHR-10 [22], are instrumental in promoting research on Earth observation and remote sensing. NWPU VHR-10 is a ten-class geospatial data set of 650 positive images for object detection, and DOTA is a 15-class data set of 2806 aerial and satellite images for multiobject detection. However, these two data sets contain few images of airplanes, cars, and ships: in particular, the NWPU VHR-10 data set contains a total of only 650 images.

To better evaluate the performance of airplane, car, and ship detection in optical remote sensing images, we collected a total of 1300 high-resolution optical remote sensing images, mainly from Google Earth and World View-2, including 300 images of cars, 500 images of ships, and 500 images of airports: the spatial resolution of the Google Earth images ranges from 0.5 to 1.5 m, and the spatial resolution of the World View-2 images is 0.5 m. We augmented our collected images by random cutting, random rotation, and random blurring. After being randomly rotated by 90°, 180°, or 270°, flipped along the Y-axis with a probability of 0.25, and randomly blurred, the 1300 images were expanded to 3200 images, and each image is of the size about 500 × 500 pixels. In the experiment, the data set was divided into a training set and a test set (comprising 20% of all the data). Fig. 5 shows some example images. Table III shows the statistical details of the data set.

B. Model Comparison

We train the data set using popular existing frameworks and compare the results with those obtained using the proposed method, and the results are shown in Table IV. All the models are trained for 100k iterations. We adopt adaptive moment estimation (Adam) as the optimization function to train our model, where the learning rate starts at 0.001 and decreases to 0.0001 after 70k iterations.

TABLE IV
EVALUATION OF DETECTION RESULTS OBTAINED USING EXISTING METHODS

Method	Airplane	Car	Ship	mAP(%)	FPS
SSD [12]	83.06	75.14	59.37	72.53	36
Faster R-CNN [9]	85.05	74.23	71.37	76.88	16
YOLOv3 [11]	91.17	74.45	78.00	81.20	59
SAPNet [4]	96.37	92.73	91.83	93.64	9
DFSSD [14]	95.16	87.21	92.45	91.60	14
MSCA [20]	97.24	90.18	93.12	93.51	19
MVNet (ours)	96.58	92.05	95.47	94.70	24

TABLE V
MODEL SIZE AND COMPUTATIONAL COST OF THE PROPOSED METHOD

Model	Input size	Model size	Computational cost (ops)	mAP(%)
SSD [12]	512×512	276MB	17.51B	72.53
MVNet (ours)	512×512	29.7MB	38.74B	94.70

TABLE VI
RESULTS OF ABLATION EXPERIMENTS

Method	mAP (%)
A: SSD [11]	72.53
B: A + Multi-vision structure	75.41
C: B + New anchor parameters	90.57
D: C + MRB	93.84
MVNet: D + MRFEM	94.70

Table IV shows that our method achieves a 94.70% mean average precision (mAP), which is 22.17%, 17.82%, and 13.50% higher than that of SSD [12], Faster R-CNN [9], and YOLOv3 [11], respectively. Our model outperforms other state-of-the-art algorithms [4], [14], [20] for remote sensing. Moreover, our method processes an image in 42 ms (24 FPS) using a single NVIDIA GTX 1080Ti GPU. The original SSD generates a total of 8732 anchor boxes compared to a total of 15,360 anchor boxes generated in our experiment. An excessive number of anchor boxes decreases the processing speed for our method when compared to the original SSD. Examples of the results produced by the best performance model are shown in Fig. 6. Table V shows the model sizes and computational cost of the proposed method and original SSD. It was observed that the model size of the proposed method was 29.7 MB, which is smaller than the original SSD. However, the proposed method requires 38.74 billion operations to perform inference, which is higher than the original SSD.

We conduct ablation experiments on our data sets to evaluate the effectiveness of the proposed method. The ablation experimental results are shown in Table VI.

1) Multi-Vision Structure: Table VI shows that the original SSD achieves a mAP of 72.53%. We modify the SSD backbone network for training using the structural parameters in Table I. In addition, we choose conv4_2, conv5_1, conv5_2, conv6_2, and conv7_2 as the detection layer. It is noted that 1 × 1 convolution, dilated convolution, and the shortcut

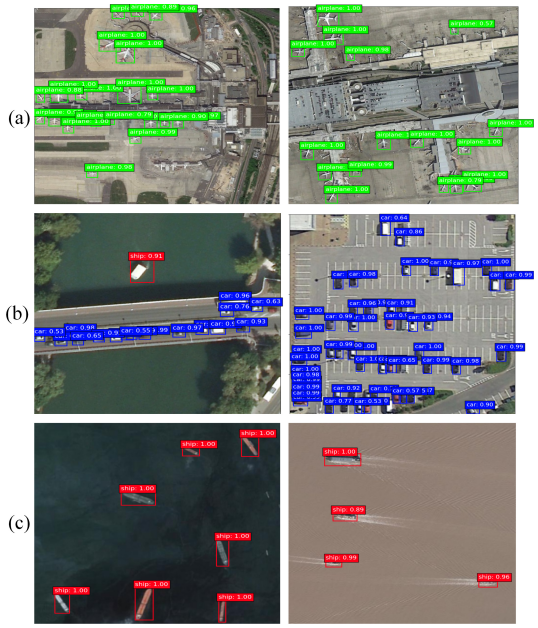

Fig. 6. Three classes of detection instances. (a) Airplane. (b) Car and ship. (c) Ship.

connection are not used. Table V shows that the use of multiple shallow detection layers considerably increases the mAP by 2.88% (from 72.53% to 75.41%).

2) New Anchor Parameters: The anchor setting of the original SSD is not suitable for small remote sensing targets. We use Experiment B to set the same anchor parameters as Table II for the five detection layers. The results show that the mAP increases by 15.16% (from 75.41% to 90.57%).

3) Multiscale Residual Block: MRB is used to extract multiscale context and prevent gradient vanishing. Table VI shows that replacing the backbone network of Experiment C with the backbone network shown in Table I increases the mAP by 3.27% to 93.84%, indicating the effectiveness of MRB.

4) Multiscale Receptive Field Enhancement Module: The MRFEM is designed to increase the accuracy of multiscale object detection. Table VI shows that the use of MRFEM clearly increases the mAP by 0.86%.

IV. Conclusion

In this letter, we developed a multi-vision small object detector to detect airplanes, cars, and ships in remote sensing images with high speed and accuracy. We first proposed an MRB to capture multiscale context information, thereby improving the feature representation ability of CNNs. We also introduced an MRFEM, which combines different features obtained by dilated convolution at different dilation rates to enhance the feature representation of remote sensing targets. We used the MRB and MRFEM to develop an MVNet that uses multiple low-level feature maps to detect small objects. Experimental results show that the proposed method can achieve the state-of-the-art detection in real time.

References

[1] Y. Hu, X. Li, N. Zhou, L. Yang, L. Peng, and S. Xiao, "A sample update-based convolutional neural network framework for object detection in large-area remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 947–951, Jun. 2019.

[2] Y. Yu, T. Gu, H. Guan, D. Li, and S. Jin, "Vehicle detection from high-resolution remote sensing imagery using convolutional capsule networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 12, pp. 1894–1898, Dec. 2019.

[3] X. Zhang, G. Wang, P. Zhu, T. Zhang, C. Li, and L. Jiao, "GRS-det: An anchor-free rotation ship detector based on Gaussian-mask in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, early access, Sep. 4, 2020, doi: 10.1109/TGRS.2020.3018106.

[4] S. Zhang, G. He, and H. B. Chen, "Scale adaptive proposal network for object detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 864–868, Jun. 2019.

[5] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.

[6] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016, doi: 10.1109/TGRS.2016.2601622.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 580–587.

[8] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1440–1448.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2015, pp. 91–99.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[11] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

[12] W. Liu, D. Anguelov, and D. Erhan, *SSD: Single Shot MultiBox Detector*. Cham, Switzerland: Springer, 2016, pp. 21–37.

[13] C. Wang, X. Bai, S. Wang, J. Zhou, and P. Ren, "Multiscale visual attention networks for object detection in VHR remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 310–314, Feb. 2019.

[14] J. Qu, C. Su, Z. Zhang, and A. Razi, "Dilated convolution and feature fusion SSD network for small object detection in remote sensing images," *IEEE Access*, vol. 8, pp. 82832–82843, 2020.

[15] Z. Zheng *et al.*, "HyNet: Hyper-scale object detection network framework for multiple spatial resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 1–14, Aug. 2020.

[16] H. Qin, Y. Li, J. Lei, W. Xie, and Z. Wang, "A specially optimized one-stage network for object detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, early access, Mar. 24, 2020, doi: 10.1109/LGRS.2020.2975086.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[19] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[20] J. Chen, L. Wan, J. Zhu, G. Xu, and M. Deng, "Multi-scale spatial and channel-wise attention for improving object detection in remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 4, pp. 681–685, Apr. 2020.

[21] G.-S. Xia *et al.*, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.

[22] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS J. Photogramm. Remote Sens.*, vol. 98, pp. 119–132, Dec. 2014.