Multiscale Visual Attention Networks for Object Detection in VHR Remote Sensing Images

Chen Wang, Xiao Bai, Shuai Wang, Jun Zhou, and Peng Ren

***Abstract*— Object detection plays an active role in remote sensing applications. Recently, deep convolutional neural network models have been applied to automatically extract features, generate region proposals, and predict corresponding object class. However, these models face new challenges in VHR remote sensing images due to the orientation and scale variations and the cluttered background. In this letter, we propose an end-to-end multiscale visual attention networks (MS-VANs) method. We use skip-connected encoder–decoder model to extract multiscale features from a full-size image. For feature maps in each scale, we learn a visual attention network, which is followed by a classification branch and a regression branch, so as to highlight the features from object region and suppress the cluttered background. We train the MS-VANs model by a hybrid loss function which is a weighted sum of attention loss, classification loss, and regression loss. Experiments on a combined data set consisting of Dataset for Object Detection in Aerial Images and NWPU VHR-10 show that the proposed method outperforms several state-of-the-art approaches.**

***Index Terms*— Multiscale feature, object detection, VHR remote sensing image, visual attention.**

I. Introduction

With the development of remote sensing technologies, very high-resolution remote sensing images with the ground sampling distance (GSD) of 2 m or less become easily available [1], facilitating the applications of disaster control, urban monitoring, and traffic planning, and so on. As one of the fundamental tasks in these applications, object detection has attracted great attention from researchers. The challenge in this task comes from the varying orientations and scales of objects in VHR remote sensing images because they are taken from either airplanes or satellites. Furthermore, while high-resolution images bring in fine details of ground objects, they also produce complex and cluttered background.

Manuscript received January 31, 2018; revised May 1, 2018 and July 31, 2018; accepted September 19, 2018. Date of publication October 29, 2018; date of current version January 21, 2019. This work was supported by the National Natural Science Foundation of China under Project 61772057, in part by Beijing Natural Science Foundation under Project 4162037, and in part by the State Key Laboratories of Software Development Environment. *(Corresponding author: Xiao Bai .)*

C. Wang, X. Bai, and S. Wang are with the School of Computer Science and Engineering, Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100191, China (e-mail: baixiao.buaa@gmail.com).

J. Zhou is with the School of Information and Communication Technology, Griffith University, Nathan, QLD 4111, Australia.

P. Ren is with the College of Information and Control Engineering, China University of Petroleum—East China, Qingdao 266580, China.

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2018.2872355

Many approaches treat object detection as a classification task with two main steps: region proposal and feature extraction. This allows a set of candidate image regions with their corresponding features to be used as the input to a classifier which then generates probabilistic values as the output [2]. Region proposal methods [3], [4] generate a limited number of category-independent potential regions which may contain objects. Faster R-convolutional neural network (R-CNN) [5] embeds the region proposal extraction into the CNN network so that a network model completes the detection task end-to-end without the need to generate the region proposals manually. This makes the detection speed of the algorithm greatly improved.

Rotation and scale invariant feature extraction is another key step for object detection in remote sensing images. Bai *et al.* [6]–[8] proposed a series of structure feature description methods for object detection. Cheng *et al.* [9] proposed a rotation-invariant framework based on collection of part detectors. Han *et al.* [10] applied deep Boltzmann machine to infer structure information from low-level Scale Invariant Feature Transform (SIFT) descriptors and encoded middle-level features to effectively describe geospatial objects. Although these rotation invariant descriptors improve object detection in remote sensing images, their limitation is that they still use hand-crafted low-level features such as SIFT, HOG (Histogram of Oriented Gradient), or texture features, rather than automatically learned image representation.

Deep CNN methods have achieved much better accuracy compared with traditional methods in VHR object detection. Long *et al.* [11] focused on accurate localization of detected objects in VHR remote sensing images, combining three steps: region proposal, classification, and accurate object localization. The Rotation Invariant CNN (RICNN) [12] model has achieved state-of-the-art performance by learning a new rotation-invariant layer on the basis of the existing CNN architectures. Li and Wang [13] proposed an object detection method in a coarse-to-fine manner. Deep CNN models for car detection [14], ship detection [15], and aircraft detection [16] concentrate on the feature extraction for specific geospatial objects. Unfortunately, these deep CNN models use only the feature maps from the highest layer of the network to predict object and have not effectively handled the problems of scale variation and cluttered background.

In this letter, we propose a multiscale visual attention networks (MS-VANs) method to solve the aforementioned problems. Our method is based on the observation that: 1) detailed local features are required to represent small objects, which, however, cannot be generated from deep layers

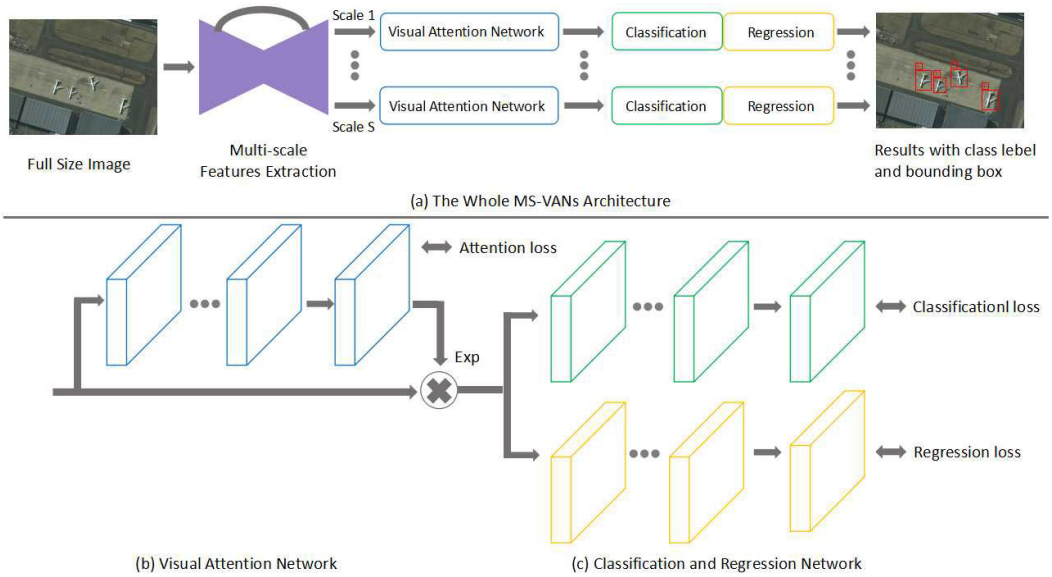

Fig. 1. Structure of the proposed MS-VANs. (a) Whole MS-VANs architecture: MS-VANs takes a full-size image as input and outputs object detection results with possible class label and coordinates of corresponding bounding box. (b) Visual attention network: visual attention network learns an attention mask which highlights the features from object regions. (c) Classification and regression networks: these are two separated branches predicting the probability value and bounding-box coordinates at each pixel.

as they have large receptive field and low-spatial resolution and 2) human vision system focuses the attention on some parts of the visual space instead of processing the whole scene at once [17], highlights the features from discriminative regions, and suppresses the background features to encourage high classification performance.

The proposed MS-VANs method takes a full-size VHR remote sensing image as input and outputs possible class of objects with their corresponding bounding boxes. Our method uses the skip-connected encoder–decoder model to extract multiscale feature maps, with skip-layers linking the feature maps between subsampling and upsampling layers. Then, visual attention model is applied to learn an attention mask, highlighting the features from object regions in each feature map. The output of visual attention networks is used for classification and bounding box regression. A hybrid loss function, consisting of attention loss, classification loss, and regression loss, is applied to train the network. The main contributions of this letter are summarized as follows.

1) We proposed an MS-VANs model trained end to end by a hybrid loss function that is a weighted sum of attention loss, classification loss, and regression loss. Different from CNN architectures which need region proposal operation, our proposed MS-VANs model directly predicts class probability value and bounding box coordinates at each pixel of multiscale feature maps. This effectively reduces the computational complexity and obtains higher accuracy.
2) We applied the skip-connected encoder–decoder network to object detection in VHR remote sensing images. Objects are detected from feature maps in different scales with specifically designed anchors, instead of from only the feature maps of the highest layer. It ensures the accuracy of all scales of object.
3) We applied a visual attention network between the feature extraction network and the classification and regression network, which highlights the features from the object regions and in turn decreases the negative effects of cluttered backgrounds. It is the first time that encoder–decoder model and visual attention network are integrated to solve object detection problems in remote sensing images.
4) Our MS-VANs model outperforms the state-of-the-art methods on combined Dataset for Object Detection in Aerial Images (DOTA) and NWPU VHR-10 remote sensing multiclass object detection data set.

II. Multiscale Visual Attention Networks for Object Detection

In this section, we describe our proposed MS-VANs method that aims to classify individual objects and localize each object using a bounding box. The structure of the MS-VANs model is illustrated in Fig. 1. The whole network is trained end to end on large VHR remote sensing object detection data set. In the network structure, the encoder–decoder model extracts multiscale feature maps, and the visual attention network learns a mask and highlights the features from object regions. Therefore, the whole network is trained by a hybrid loss function with three loss terms.

A. Multiscale Feature Extraction

The proposed MS-VANs model takes the full-size image as input and does not need any region proposal operation. We then propose a skip-connected encoder–decoder model to extract multiscale features from the original images.

We propose to extract multiscale feature maps and use features of each scale to detect the object. The intuition is that lower network layers have smaller receptive fields, better matched to small objects. Conversely, higher layers with high-level semantic information are best suited for large objects. Therefore, we adopt a subsampling operation to

progressively decrease the resolution of feature maps and enlarge the receptive field. After every two convolutional layers, subsampling layers are used to downsample the feature maps. Our network contains five levels of subsampling, so the features are subsampled by a factor of 32 at most. In order to get the original size of the input image, feature maps are upsampled progressively for which the upsampling parameters are learned automatically. There are two convolutional layers and one upsampling layer in each upsampling level.

To fuse features of the same scale and retain high-frequency information, we stack feature maps obtained before subsampling to the feature maps obtained after upsampling. The stacked feature maps at each upsampling level have the same resolution. All convolutional layers use kernel of 3×3 in size and are followed with batch normalization and LeakyReLU operation. Instead of using max pooling or average pooling, we use 3×3 convolutional filters with strides 2 to increase or decrease the resolution of feature maps, which avoids discrete gradient.

B. Visual Attention Networks

The multiscale features extracted using the above-mentioned encoder–decoder model describes different object scales in different layers. Since we want to highlight the features from geospatial objects and decrease the influence of cluttered background, learning an attention mask to guide the network to "see" is a viable choice. In this letter, we propose a visual attention model as illustrated in Fig. 1. This visual attention model takes feature maps at different scales as input. Five sequential convolution layers (kernel size $=$ 3, stride $=$ 1) are applied to extract features and followed by a convolution layer (kernel size $=$ 1) to reduce the number of feature maps to one. The ground truth attention mask is obtained by filling the ground-truth bounding box as 1 and other positions as 0. The attention masks are first fed to an exponential operation and then fused with the input feature maps with a dot product.

The complex and cluttered background consists of many redundant and even distractive texture data for accurate object detection. Our visual attention network learns to keep and highlight the useful information of geospatial objects and, in turn, ensures high detection accuracy.

C. Loss Funcion

The visual attention networks are followed by two branches. The classification branch is used to predict the probability value at each pixel. The regression subnetwork estimates the top-left and bottom-right coordinates of object bounding box. We simultaneously predict multiple anchors [5] which are centered at each position and associated with a scale and aspect ratio. We observe that the scale of most geospatial objects ranges from 33 to 350 pixels. If the size of an object is less than 30 pixels, it is not used as training samples.

We set five anchors from areas of 512^2–32^2, with the scale step of 0.5. In addition, the aspect ratios for anchors are set to 1:1, 1:2, and 2:1, respectively. Thus, there are $A = 15$ anchors at each position. For a feature map of a size $W \times H$, there are $W \times H \times A$ anchors in total. Intersection-over-Union (IoU) is defined as the percentage of pixels in the intersection of two regions. Anchors with the highest IoU and with any ground-truth bounding box larger than 0.6 is assigned as positive samples, which will contribute to the computation of the regression loss. If the highest IoU between an anchor and all bounding boxes is less than 0.3, the anchor is assigned to negative samples. Unsigned anchors are not used during the training.

We propose a hybrid loss function for the MS-VANs model

$$L = \sum_{i \in S} \left(\frac{1}{N_i^{\text{cls}}} \sum_{j \in A_i} L_{\text{cls}}\left(p_j, p_j^*\right) + \lambda_1 \frac{1}{N_i^{\text{reg}}} \sum_{j \in A_i} I(\text{positive}) L_{\text{reg}}\left(t_j, t_j^*\right) + \lambda_2 L_a\left(m_i, m_i^*\right) \right) \tag{1}$$

where i is the index of output feature map from different encoder–decoder layers and A_i is a set of anchors defined in the ith feature map. p_j is a vector representing the classification probabilities and p_j^* is a one-hot vector representing the ground-truth class label of an object. t_j is a vector representing the four parameterized coordinates of the predicted bounding box and t_j^* is the ground truth associated with a positive anchor. m_i is the attention mask learned by visual attention network and m_i^* is the ground truth mask. $L_{\text{cls}}(p_j, p_j^*)$ is the cross-entropy loss function for multiclass classification. N_i^{cls} is the number of anchors which contribute to the classification loss. I(positive) means that the regression loss is activated only for positive anchors. $L_{\text{reg}}(t_j, t_j^*)$ is L1 loss and $L_a(m_i, m_i^*)$ is pixelwise sigmoid cross entropy. These three losses are weighted by two balancing parameters λ_1 and λ_2. By tuning λ_1 and λ_2 between 0.8 and 1.3 with a step of 0.05, the best values of these two balancing parameters are identified as $\lambda_1 = 0.95$ and $\lambda_2 = 1.1$.

D. Data Augmentation

To enhance the rotation invariance, we propose a random rotation strategy that generates object images with different orientations. Based on the training set, we rotate the images for a set of angles ranging from 5° to 355° with a step of 5°. As the CNN model extracts local features of objects, our random rotation strategy ensures that multiple orientations of an object can contribute to the training process.

III. Experiments

In this section, we evaluate the performance of our model on the DOTA [1] and NWPU VHR-10 [9] data sets. We not only compare the object detection performance in terms of average precision (AP) and running time with other state-of-the-art methods but also show visualized results of our proposed method.

A. Data Set Description

DOTA [1] is a challenging 15-class remote sensing object detection data set, which contains 2806 aerial images captured by different sensors. The images range from about

TABLE I

COMPARISON OF SIX METHODS IN TERMS OF MAP VALUE (%). THE SHORT NAMES FOR CATEGORIES ARE DEFINED AS: P-PLANE, S-SHIP, ST-STORAGE TANK, BD-BASEBALL DIAMOND, TC-TENNIS COURT, BC-BASKETBALL COURT, GTF-GROUND TRACK FIELD, H-HARBOR, B-BRIDGE, SV-SMALL VEHICLE, LV-LARGE VEHICLE, SBF-SOCCER BALL FIELD, RA-ROUNDABOUT, SP-SWIMMING POOL, AND HC-HELICOPTER. RUN TIME MEANS THE AVERAGE RUNNING TIME FOR DETECTING OBJECTS IN AN IMAGE OF $512 * 512$ PIXELS

	P	S	ST	BD	TC	BC	GTF	H	B	SV	LV	SBF	RA	SP	HC	MAP	Run time(s)
RICNN [12]	62.1	41.3	27.1	26.0	51.7	43.1	28.5	31.5	20.1	30.4	28.4	25.3	28.1	26.5	13.4	32.2	3.24
Long [11]	69.2	46.3	29.3	28.6	55.7	45.1	30.3	34.1	19.5	33.5	29.6	27.4	31.9	33.5	14.2	35.2	1.32
Li [13]	71.7	48.2	31.5	30.6	59.2	49.1	33.5	35.1	20.5	36.9	31.1	30.7	34.8	35.1	15.3	37.6	1.02
Faster R-CNN [5]	78.4	47.2	58.6	76.8	89.7	74.3	67.3	60.8	32.1	51.9	51.5	55.3	48.7	55.2	40.8	59.2	0.15
Yolo v3 [18]	79.0	49.8	59.2	77.1	89.9	74.8	68.1	61.5	33.9	52.8	52.2	55.5	49.0	55.9	41.7	60.0	**0.075**
Ours	**79.2**	**52.6**	**60.1**	**77.4**	**90.3**	**75.1**	**68.5**	**61.7**	**34.1**	**53.0**	**52.4**	**55.8**	**49.1**	**56.2**	**42.3**	**60.5**	0.11

800×800 to 4000×4000 pixels in size and contain objects exhibiting a wide variety of scales, orientations, and shapes. 188 282 instances are annotated by experts. The height of the horizontal bounding box ranges from 10 pixels to more than 300 pixels. The GSD ranges from 0.11 to 0.55 m. NWPU VHR-10 is a 10-class remote sensing object detection data set. This data set contains 715 images from Google Earth with GSD ranging from 0.5 to 2 m and 85 images from Vaihingen data set with GSD value of 0.08 m. The height of the horizontal bounding box ranges from 33 to 418 pixels. 2934 instances are annotated. Both data sets contain object classes of Plane, Ship, Storage tank, Baseball diamond, Tennis court, Basketball court, Ground track field, Harbor, Bridge, and Small vehicle. The DOTA data set has five extra categories, including Soccer ball field, Helicopter, Swimming pool, Roundabout, and Large vehicle. The DOTA and NWPU VHR-10 data sets contain 650 and 1869 positive images, respectively, with each image containing at least one target to be detected with the ground-truth information. In the experiments, we combine these 2519 images into a large data set and divide them into three groups: 1619 images for training, 300 images for validation, and the rest 600 images for testing.

B. Experimental Results and Analysis

In the first experiment, we evaluate the performance of our MS-VANs model on the combined DOTA and NWPU VHR-10 data set and compare it with five state-of-the-art models. These methods are RICNN [12], Long *et al.* [11], Li *et al.* [13], Faster R-CNN [5], and Yolo v3 [18]. RICNN introduces a new rotation invariant layer on the basis of the existing CNN architectures. The method of Long *et al.* [11] focuses on accurate localization of detected objects in VHR remote sensing images, consisting of three steps: region proposal, classification, and accurate object localization. Li *et al.* [13] detected an object in a coarse-to-fine manner. Faster R-CNN [5] uses region proposal networks to estimate region proposals so the network can be trained end to end. Yolo v3 [18] is the latest version of Yolo network, which does not need region proposal operation and detects object directly at each pixel of feature maps. For a fair comparison, we used all hyperparameters with the highest performance in the original papers for the above-mentioned state-of-the-art models.

Our MS-VANs model was implemented using PyTorch and was optimized with RMSprop. The model was trained on a single Nvidia Titan X GPU with four images per mini-batch. Our model was trained for 100k iterations. The learning rate was initially set as 0.01 and then was dropped by 10 after 60k iterations. For experiments on all data sets, we set $A = 15$, $\lambda_1 = 0.95$, $\lambda_2 = 1.1$. On the basis of the hybrid loss, our MS-VANs model simultaneously predicted object class at each pixel of multiscale feature maps, instead of using region proposal method to extract a region of interest (ROI) and predict the object class for each ROI. With all bounding box coordinates as the output, we used nonmaximum suppression operation to select the most accurate locations.

As illustrated in Table I, our MS-VANs model achieves the best results in all 15 classes in terms of average precision AP. The MS-VANs model learns attention masks for multiple scales of feature maps, respectively, where features from object regions are highlighted and redundant features from the cluttered background are suppressed. This decreases the number of false positives and increases the average accuracy. The augmentation operation considers different orientations of each object and encourages rotation-invariant features extracted from our MS-VANs model. The average running time of our approach to process an image of 512×512 pixels is by average 0.11 s. Even though the proposed MS-VANs predicts object class and bounding box at every pixel, the forward propagation runs only once, which can also decrease the time cost. When both mean average precision (MAP) and running time are taken into consideration, our method shows clear advantages over the baseline approaches.

We also carry out a quantitative analysis to show the effect of changing GSD on the classification accuracy. To obtain images with higher GSD value, we blurred all images in the combined data set using a Gaussian kernel (standard deviation $= 3$) and then downsampled them by a scale factor of 4. Maintaining all the network architecture and training set, we trained and tested our model with these low-resolution images. The MAP value of all 15 classes decreases by 3% compared with results using the original data set. Along with the increasing of GSD, objects with size less than 120×120 pixels in original images can be barely detected, which leads to lower MAP values. On the other hand, for a higher value of GSD, most objects are represented by only a few pixels resulting in a significant interclass similarity, which leads to a lower average precision value. For example, the AP value of "Small vehicle" and "Large vehicle" decreases by 1.3% and 0.9%, respectively. Since the images contain more redundant context information with lower GSD, they are beneficial to extract representative and discriminative features of objects, especially for objects with high interclass similarity. While high-resolution images bring in fine details of ground objects,

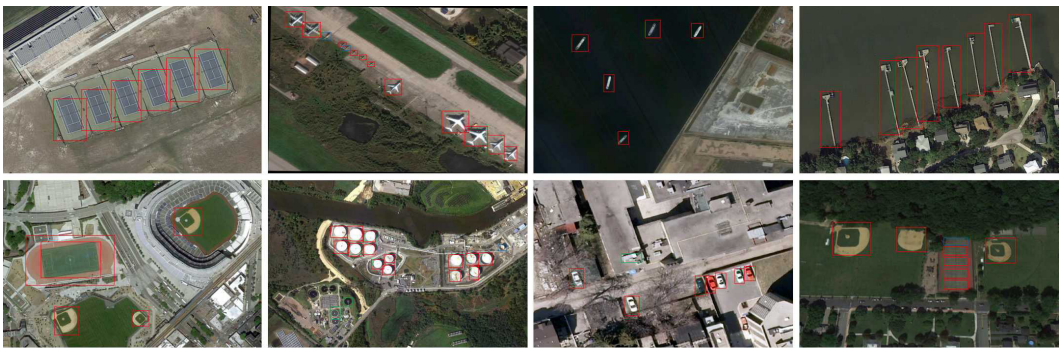

Fig. 2. Detection results from the proposed approach. Red rectangles: true positives. Green rectangles: false positives. Blue rectangles: false negatives.

they may also produce complex and cluttered background. The multiscale features and visual attention network proposed in this letter can solve this problem well.

Furthermore, we also evaluate the performance of our proposed network when it is trained on DOTA data set and tested on NWPU VHR-10 data set. 10 categories that are covered in both DOTA and NWPU VHR-10 data sets are used for the experiment. The MAP value is 58.6% which is 1.9% lower than the result in Table I. Although the platform, sensors, and GSD are different between these two data sets, our proposed model has certain robustness.

Fig. 2 visualizes the results of the proposed MS-VANs. The true positives, false positives, and false negatives are indicated by red, green, blue rectangles, respectively. A rectangle rooftop is predicted as a vehicle in the third image of the second row. Two black circular architectures are detected as storage tank in the second image of the second row. This is because from the top view, these false positives look similar as ground-truth samples. As multiscale features are adopted to detect objects, both large-scale objects (e.g., ground track field) and small-scale objects (e.g., ship) are accurately detected. In addition, it can be seen in the first image from the second row, different scales of baseball diamond are detected, which also demonstrates the effectiveness of our MS-VANs model.

IV. Conclusion

In this letter, we have introduced an MS-VANs model to detect objects in VHR remote sensing images. Instead of using region proposal methods to extract ROI and predict the object class for each ROI, we propose a visual attention-based network and simultaneously predict object class at each pixel of the feature maps. Encoder–decoder model is applied to extract multiscale features, and each scale of feature maps is used to detect objects, which ensure the accuracy of all scales of the object. Visual attention network is learned to highlight the features from the object region and decrease the influence of cluttered backgrounds. Furthermore, data augmentation is adopted to enhance the orientation invariance capability of the MS-VANs model. Experiments on combined DOTA and NWPU VHR-10 data set show that our MS-VANs model has outperformed several state-of-the-art approaches.

References

[1] G.-S. Xia *et al.*, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE CVPR*, Jun. 2018.

[2] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016.

[3] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2189–2202, Nov. 2012.

[4] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Apr. 2013.

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

[6] X. Bai, H. Zhang, and J. Zhou, "VHR object detection based on structural feature extraction and query expansion," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 10, pp. 6508–6520, Oct. 2014.

[7] H. Zhang, X. Bai, J. Zhou, J. Cheng, and H. Zhao, "Object detection via structural feature selection and shape model," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4984–4995, Dec. 2013.

[8] C. Yan, X. Bai, P. Ren, L. Bai, W. Tang, and J. Zhou, "Band weighting via maximizing interclass distance for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 7, pp. 922–925, Jul. 2016.

[9] 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.

[10] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.

[11] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2486–2498, May 2017.

[12] 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.

[13] X. Li and S. Wang, "Object detection using convolutional neural networks in a coarse-to-fine manner," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 1, pp. 2037–2041, Nov. 2017.

[14] M. ElMikaty and T. Stathaki, "Detection of cars in high-resolution aerial images of complex urban environments," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5913–5924, Oct. 2017.

[15] H. Lin, Z. Shi, and Z. Zou, "Fully convolutional network with task partitioning for inshore ship detection in optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1665–1669, Oct. 2017.

[16] F. Zhang, B. Du, L. Zhang, and M. Xu, "Weakly supervised learning based on coupled convolutional neural networks for aircraft detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 9, pp. 5553–5563, Sep. 2016.

[17] T. V. Nguyen *et al.*, "Image re-attentionizing," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1910–1919, Dec. 2013.

[18] J. Redmon and A. Farhadi. (2018). "YOLOv3: An incremental improvement." [Online]. Available: https://arxiv.org/abs/1804.02767