Region-Enhanced Convolutional Neural Network for Object Detection in Remote Sensing Images

Jianjun Lei, *Senior Member, IEEE*, Xiaowei Luo, Leyuan Fang, *Senior Member, IEEE*, Mengyuan Wang, and Yanfeng Gu, *Senior Member, IEEE*

***Abstract*— The convolutional neural networks (CNNs) have recently demonstrated to be a powerful tool for object detection. However, with the complex scenes in remote sensing images, feature extraction of the object in the CNN will be serIously affected by background information. To address this issue, in this article, a region-enhanced CNN (RECNN) is proposed for the object detection of remote sensing images. The RECNN introduces the saliency constraint and multilayer fusion strategy into the CNN model, which can effectively enhance the object regions for better detection. Specifically, the saliency map is extracted and utilized to guide the training of the proposed model to strengthen saliency regions in feature maps. In addition, since different layers can reflect the object regions in varied resolutions, a multilayer fusion strategy is introduced to connect different convolutional layers and explore the context, where the feature maps of object regions are further enhanced. Experimental results on a publicly available ten-class object detection data set demonstrate the superiority of the RECNN over several competitive object detection methods.**

***Index Terms*— Multilayer fusion strategy, object detection, region-enhanced convolutional neural network (RECNN), remote sensing images, saliency constraint.**

I. Introduction

AS A development of remote sensing imaging technology, high-resolution remote sensing images are able to be collected and used to provide rich information for image analysis. Object detection is a typical and important issue in the field of image analysis [1]. With the abundant resolution information, object detection has also become a hot research topic for analyzing the high-resolution remote sensing images, which can be applied on many civil [2] and military applications [3], such as town planning [4], [5] and monitoring [6], [7].

The objective of object detection is to recognize the categories of objects and determine their exact locations in images. To achieve the object detection, two steps are usually adopted [8]–[10], i.e., feature extraction and classifier design. Objects in remote sensing images are characterized by the discriminative features at the feature extraction step and then are classified from the background at the classification step based on the extracted features. Many hand-crafted features were earlier used in object detection [11]–[13]. Each hand-crafted feature is specifically designed to effectively characterize one property of object, such as key point, gradient, color contrast, or rotation. However, such a single feature is hard to effectively represent objects in different scenes of remote sensing images. Since multiple features can characterize different properties of objects, multiple features were combined to enhance the effectiveness of feature extraction [14]–[16]. After feature extraction, several classifiers can be adopted to estimate whether objects are included in image patches, e.g., support vector machine (SVM) [17], manifold regularization and entropy regularization [18], sparse representation-based classifier (SRC) [19], and template matching [20]. Although the fusion strategy of features can deliver better performance, the hand-crafted features still cannot detect the objects in a very complex situation of the remote sensing images, such as diversified objects, a large variation of illuminations, occlusions, or complicated scenes.

Recently, deep learning model [21] has made a great breakthrough and been successfully used in many image processing and computer vision applications [22], [23]. The great success of the deep model is due to the fact that it can automatically extract a hierarchy of effective features from the input images. Also, the deep model has been extended to object detection in remote sensing images [24]. The typical deep learning models utilized in object detection are the restricted Boltzmann machines (RBMs) and convolutional neural networks (CNNs). Compared with RBM, the CNN can effectively capture the spatial information in the fixed image patches generated by scanning over the whole image. Furthermore, fully convolutional networks (FCNs) [25] and faster region-based CNNs, i.e., Faster R-CNN [26], were proposed to exploit the spatial information in global remote sensing images during feature extraction. Deep learning models for object detection utilize deep layer structures to extract features, which provide promising performances on object detection. However, the extracted features from the deep learning models are still easily disturbed by cluttered background information due to the variety of scenes and rich detail information in high-resolution remote sensing images.

Manuscript received April 23, 2019; revised October 18, 2019 and December 16, 2019; accepted January 17, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61722112, Grant 61922029, and Grant 61520106002. *(Corresponding author: Leyuan Fang.)*

Jianjun Lei, Xiaowei Luo, and Mengyuan Wang are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: jjlei@tju.edu.cn).

Leyuan Fang is with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China (e-mail: fangleyuan@gmail.com).

Yanfeng Gu is with the School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mail: guyf@hit.edu.cn).

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

Digital Object Identifier 10.1109/TGRS.2020.2968802

In this article, to reduce the interference of background information and enhance the object region, a region-enhanced CNN (RECNN) is proposed for object detection in remote sensing images, where the saliency constraint and multilayer fusion strategy are utilized jointly in an end-to-end CNN model. In RECNN, the saliency map and pixel-level loss function are first explored to guide the training of network in the field of object detection for remote sensing images, where the saliency regions are effectively enhanced with the beneficial saliency information. Moreover, based on the structure of RECNN, a robust multilayer fusion strategy is developed to explore the context of different resolutions in network and further strengthen the response of object-like regions in feature maps.

The remainder of this article is organized as follows. Section II gives a brief introduction of the related work. Section III introduces the proposed method in detail. The experimental results are presented in Section IV. Finally, the conclusion is drawn in Section V.

II. Related Works

In this section, we review the related works including traditional hand-crafted features for object detection and deep learning based methods for object detection.

A. Hand-Crafted Features for Object Detection

During the past decades, many works [11]–[20] attempt to detect objects with hand-crafted features, such as the histogram of oriented gradient (HOG) [11]–[13], local binary pattern (LBP) [18], Hough transform [19], or the rotation-invariant feature, i.e., radial-gradient angle (RGA) [20]. Although hand-crafted features can effectively characterize the corresponding property of object, it is difficult for one single hand-crafted feature to represent objects in complex scenes of remote sensing images. To enhance the effectiveness of extracted features, multiple features are combined for object detection. In [14], local contrast of intensity, orientation and global color contrast were combined for feature extraction and image segmentation. In [15], several moments and the dense SIFT were used jointly to construct a robust feature. The cascaded combination of line segment and statistical region merging features were introduced in [16] for the efficient object location and recognition. However, it is challenging for hand-crafted features to detect objects in complex and changeable situations, thus limiting its application in remote sensing images.

B. Deep Learning Models for Object Detection

For object detection in remote sensing image, the deep Boltzmann machine (DBM) [24] was formed by stacking three RBM models to extract high-level features based on the low-level and middle-level features, which are used for detection with SVM. Similarly, the deep belief network (DBN) [27] was constructed by multiple stacked RBM models for powerful feature extraction, where the extracted feature is fed to softmax for prediction. Recently, many frameworks [28]–[30] based on CNN have been proposed for object detection in remote sensing images. To detect objects of different orientations, Cheng *et al.* [28] proposed a rotation-invariant CNN, i.e., RICNN, to acquire the rotation-invariant features by a learned rotation-invariant layer, and classified the categories of object by SVM based on the rotation-invariant features. Zhao *et al.* [29] utilized a variant of CNN named vanilla network for aircraft landmarks regression, which predicts the locations of each landmark and judges the distance between the landmarks of each detected result and template to fulfill aircraft recognition task. The network in [31] takes an image of arbitrary size as input and produces the same size detection map. In [32], an entire image was input into the proposed network, and the categories scores and coordinates of bounding boxes of object were predicted.

To exploit the global image information and break the limitation of the sliding-window, recent works [33], [34] based on the FCN or the single shot detectors [35]–[37] were proposed for object detection in remote sensing images. The FCN [25] took an input image of arbitrary size and produced the same size classification map by upsampling with multiple upsampling factors, where the global image information is exploited. In [35]–[37], the feature maps of single shot detectors in the feed-forward process were directly used to detect objects, where satisfied performances on speed and accuracy are achieved. Inspired by these networks introduced in natural images, Zhang *et al.* [33] proposed a coupled CNN, which includes an FCN-based network called CRPNet and a CNN-based network called LOCNet, to extract the candidate regions and recognize airplanes. Based on the single shot detection networks, Zou and Shi [34] introduced a paradigm called random access memories (RAMs) for object detection in remote sensing images under object priors, which adaptively updates the detection model to maximize its posterior determined by both training and observation. Different from these object detection methods, the proposed method designs an end-to-end RECNN, which can alleviate the interference of a cluttered background in high-resolution remote sensing images.

III. Proposed Method

A. Architecture of RECNN

Due to the complex scene in remote sensing images, it is usually difficult for traditional deep models to distinguish the objects from the background, limiting the detection performance. To address this issue, as shown in Fig. 1, the proposed RECNN method utilizes two region-enhanced branches, i.e., saliency information and multilayer fusion, to enhance the object regions in the feature maps and thus can better detect the object. Specifically, the saliency usually corresponds to the main object regions [38]. Therefore, a saliency reconstruction branch (RB) is designed to reconstruct the saliency map and incorporates the binary saliency map into the loss function to enhance the saliency area in the feature maps. In addition, since different layers can reflect the object regions in varied resolutions, the contextual information of different convolutional layers is explored by the fusion module (FM) of the detection branch, which makes the objects of interest more distinguishable. As shown in Fig. 2, the two branches are

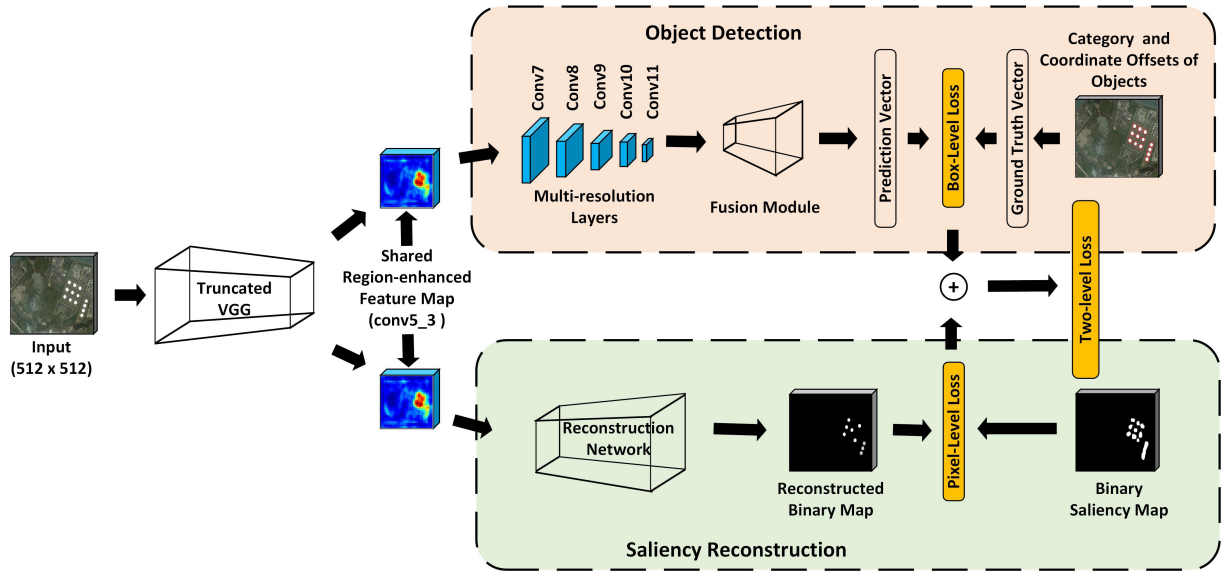


Fig. 1. Architecture of the proposed RECNN object detection network. Truncated VGG is taken as the base network of RECNN for CNN feature extraction. In view of multiscale object detection and binary saliency maps reconstruction, some extra convolutional layers and the reconstruction network are both added after the last layer of VGG (conv5_3). The FM is further connected to the convolutional layers for context exploitation.

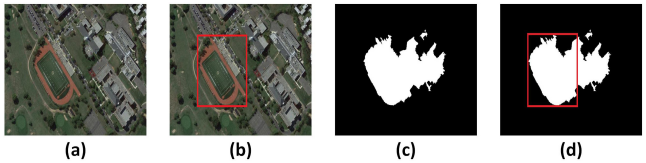


Fig. 2. Characteristics of two-level constraints. (a) Original image. (b) Detected region with box-level constraint that can fit the coordinates and predict the class scores of the bounding box of object. (c) Detected region with pixel-level constraint that can more precisely recognize each pixel of object. (d) Detected region with two-level constraints.

separately supervised under the constraints of pixel-level and box-level, where the object regions are effectively distinguishable, and thus the detection accuracy can be improved.

B. Saliency Map Reconstruction and Pixel-Level Constraint

The cluttered background information in high-resolution remote sensing images severely interferes with the object recognition of CNN-based networks. Therefore, the object regions in feature maps should be enhanced so that the features of objects can be more easily identified in complex scenarios. Considering the saliency map can promote the network to focus on the saliency regions, which have the high probabilities to cover the object regions and weaken the interference of background, the saliency information is utilized to guide the training of RECNN to enhance the object regions in the feature maps. To avoid the performance degradation caused by the inaccurate saliency information, the saliency information is applied with the pixel-level loss function in the RECNN, where the valid saliency information is adaptively extracted by the proposed network during training.

To guide the network to focus more on the object area, the binary saliency map is introduced into the training process. The truncated VGG, i.e., VGG-16 [39] without the last three full connected layers, is used to extract low-level features for the detection branch in RECNN. Inspired by the conclusion that an appropriate decoder network should consist of a hierarchy of decoders corresponding to their encoders [40], the RB and the truncated VGG are used to construct an hourglass encoder–decoder network for better saliency map reconstruction. The reconstruction network, a symmetrical network to the truncated VGG, is connected to the end of the truncated VGG, and a hierarchy of reconstructed results corresponding to the pooling results in VGG can be obtained, where layers are upsampled by a special upsampling layer [40]. During the saliency reconstruction, the saliency regions in the feature maps of VGG are enhanced. Meanwhile, with the proposed structure, where the two branches share the same truncated VGG, these saliency-enhanced feature maps can be propagated to the detection branch, where the responses of the saliency regions are also strengthened for object detection.

To incorporate the saliency information into loss function, the saliency map (obtained by the method in [41]) is transformed to a binary saliency map with a threshold τ (some examples are shown in Fig. 3), which can be used for binary semantic segmentation. The value of each pixel $p_i \in \{0, 1\}$ in binary saliency map indicates its class, i.e., saliency or background. By referring to the salient segmentation methods [42], [43], the output map produced by RB is transformed to a corresponding two-channel map through the last convolutional layer of the branch. One channel represents the probability that each pixel belongs to a significant region while the other channel is used to reflect the probability that each pixel belongs to a background region. In other words, by using a two channel map, the confidences of background class and target class are obtained. At training stage, a training image and its corresponding binary saliency map are separately taken as the input and label. The softmax layer is adopted for pixel classification, which is described as

$$c_i(k) = \frac{e^{p_i(k)}}{e^{p_i(1)} + e^{p_i(2)}} \tag{1}$$

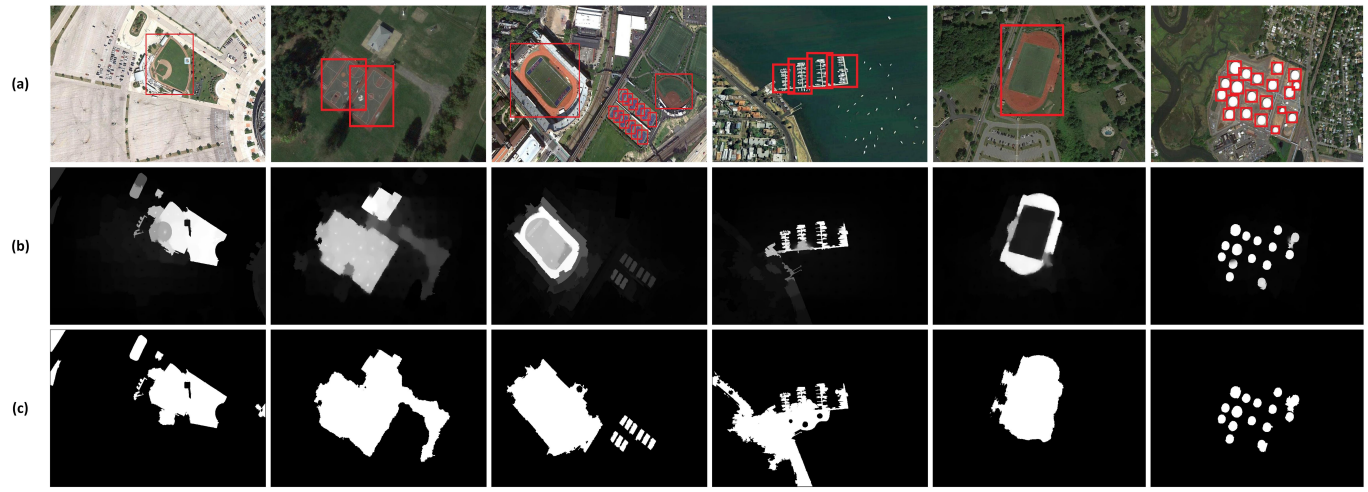


Fig. 3. Some examples of training set. (a) Input images. (b) Detected saliency maps. (c) Transformed binary saliency maps.

where $p_i(k)$ denotes the kth channel value of the ith pixel in the output map produced by the reconstruction network, and $c_i(k)$ denotes the pixel value in classification map C (whose size is $W \times H$) generated by a softmax layer. To construct the pixel-level constraint term for training, the pixel-level softmax classification loss function is adopted, which aims to minimize the misclassification error between C and its groundtruth classification map $\bar{C}$. The pixel-level loss function is given by

$$L_p(C, \bar{C}) = -\frac{1}{W \times H} \sum_{i=1}^{W \times H} \sum_{k=1}^{2} \bar{c}_i(k) \cdot \log[c_i(k)]. \quad (2)$$

C. Multilayer Fusion and Box-Level Constraint

Inspired by the efficiency of the computational detection networks [35], [36], whose multiscale feature maps in feed-forward process are directly used to detect multiscale objects, five extra convolutional layers, whose sizes decrease progressively with the stride of 2, are added after the conv5_3 of the truncated VGG to achieve a balance between detection accuracy and speed. These extra convolutional layers in the detection branch of the RECNN have contextual information of different resolutions, where the shallower convolutional layers can preserve rich spatial detailed structures to support the accurate localization of objects, while the deeper layers capture high-level abstraction semantic information to strengthen the feature representation of objects. Exploiting the contextual information in the detection branch can make each prediction layer in RECNN to achieve better performance on object recognition and location, which further strengthens the responses of object regions and improves detection accuracy. The multilayer fusion strategy and an FM are designed in the RECNN, as shown in Fig. 4, where deconvolution modules and prediction modules [36] separately fuse the two different resolutions and produce prediction results for object detection.

Similar to Single Shot multibox Detector (SSD) [35], the default box mechanism is adopted in RECNN for object detection. We can obtain six prediction feature maps of different scales after prediction modules. Then, six sets of prediction vectors are generated through the prediction layers. After that, these prediction vectors are concatenated and integrated to predict the category and coordinate offsets of objects. The box-level loss is the linear combination of softmax loss and smooth $L1$ loss for category classification and bounding boxes regression, which is computed by

$$L_b(L, V, \bar{L}, \bar{V}) = \frac{1}{N} \begin{bmatrix} L_{\text{soft max}}(L, \bar{L}) \\ +\beta L_{\text{smooth}L1}(V, \bar{V}) \end{bmatrix} \quad (3)$$

where N is the number of the default boxes used for prediction and β is the weight term. The softmax loss over the predicted category scores of object boxes L and the corresponding groundtruth category labels $\bar{L}$ is formulated as

$$L_{\text{soft max}}(L, \bar{L}) = \sum_{m=1}^{M} I(\bar{L}) \cdot \log \left[\frac{e^{V(m)}}{\sum_{j=1}^{M} e^{V(j)}} \right] \quad (4)$$

where M is the number of categories, $I(\bar{L}) = \{0, 1\}$ is an indicator corresponding to groundtruth category labels.

The smooth $L1$ loss over the predicted coordinate offsets of object boxes V and the groundtruth coordinate offsets $\bar{V}$ is calculated by

$$L_{\text{smooth}L1}(V, \bar{V}) = \begin{cases} 0.5(V - \bar{V})^2, & |V - \bar{V}| < 1 \\ |V - \bar{V}| - 0.5, & |V - \bar{V}| \geq 1. \end{cases} \quad (5)$$

D. Network Training and Object Detection

As mentioned above, the pixel-level and the box-level constraints are simultaneously applied on the RECNN during training, which enables the proposed RECNN to focus on the object regions in feature maps. At the training stage, an entire training optical remote sensing image is taken as an input of the RECNN. The binary saliency map is set as the label of pixel-level loss in (2), and the label of box-level loss in (3) is set to the categories and coordinate offsets of objects. By combining the pixel-level and box-level loss functions,

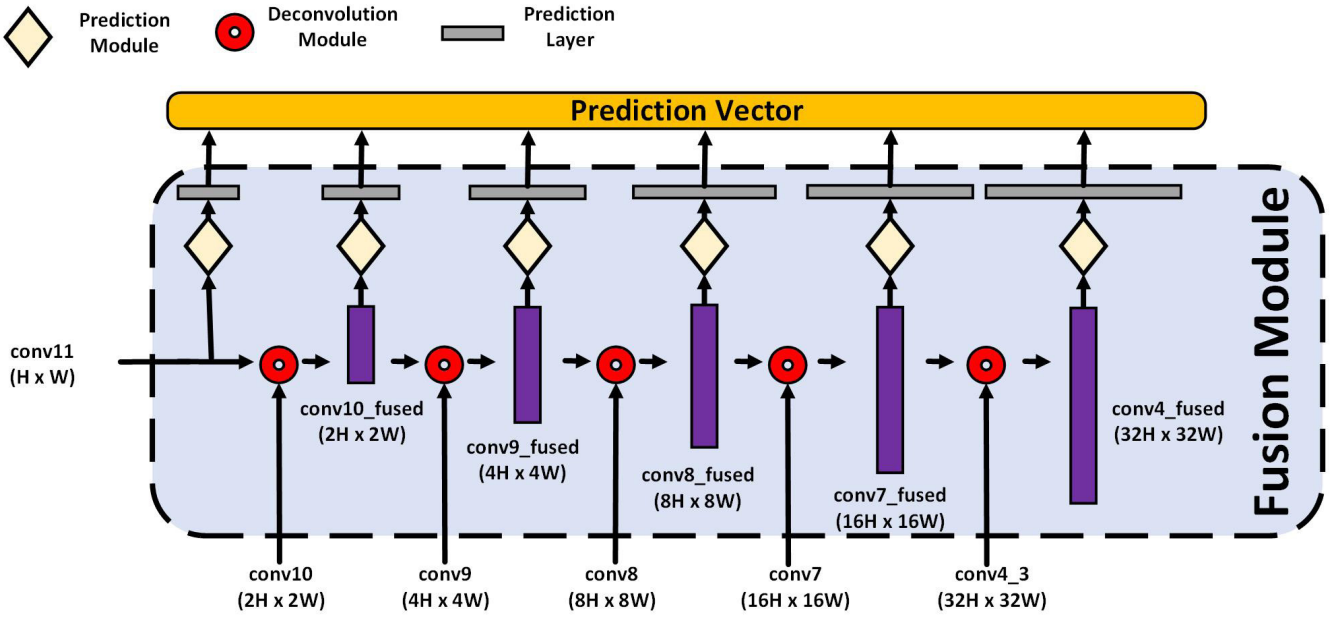


Fig. 4. Architecture of multilayer FM. Each deconvolution module fuses the two input layers of different resolutions and produces a fused feature map with high resolution. Each prediction module is used to obtain a prediction result, which has the same size with the input layer.

the following proposed two-level loss for RECNN training is obtained:

$$L(C, L, V, \bar{C}, \bar{L}, \bar{V}) = \alpha L_p(C, \bar{C}) + L_b(L, V, \bar{L}, \bar{V}) \quad (6)$$

where α is the weight term.

The goal of object detection is to recognize the categories and locate the positions of objects. After training, the detection branch can capture the saliency information in the image, and the RB has no direct effect on detection results. To reduce the parameters and accelerate computation speed, the RB is removed from RECNN at object detection stage, which makes the proposed network be more suitable for practical application.

IV. Experiments

A. Experiments Settings

NWPU VHR-10 data set [28] is adopted in our experiments to analyze the performance of the proposed network. The challenging data set is widely used for the evaluation of detection methods [11], [34]. It is used for ten-class object detection in very high resolution optical remote sensing imagery, and a total of 715 very high-resolution optical color remote sensing images are collected in data set, which are separately acquired from Google Earth with the spatial resolution ranging from 0.5 to 2 m, as well as 85 pansharpened color infrared images obtained from Vaihingen data with a spatial resolution of 0.08 m. Following the training-testing split criterion and experimental settings in [28], the 150 images which contain no objects are removed and the rest of the data set is divided into 20% for training, 20% for validation, and 60% for testing. The training samples are separately rotated with the angle of $\varphi = \{10°, 20°, \ldots, 350°\}$, which extends the quantity of training samples by 36 times. In addition, the training samples are further augmented by randomly modifying the brightness, contrast, hue, and saturation.

We adopt the average precision (AP) as the metric to quantitatively evaluate the performances of object detection methods. For the calculation of the AP metric, the intersection-over-union ratio (IoU) and precision-recall curve are used as detailed below.

The IoU calculates the intersection area and union area between the bounding boxes of groundtruth and detection results. A detection result is considered as corrected if the IoU between the bounding boxes of detection result and groundtruth exceeds the IoU threshold. The IoU is formulated as

$$\alpha_{\text{IoU}} = \frac{\text{area}(B_{\text{det}} \cap B_{\text{gt}})}{\text{area}(B_{\text{det}} \cup B_{\text{gt}})} \quad (7)$$

where $\text{area}(B_{\text{det}} \cap B_{\text{gt}})$ denotes the intersection area of a detection result and groundtruth bounding boxes, and $\text{area}(B_{\text{det}} \cup B_{\text{gt}})$ denotes their union area. In our experiments, the IoU threshold is set to 0.5.

The precision-recall curve reflects the tradeoff between precision and recall. The precision metric measures the fraction of detection results which accurately cover objects in images. The recall metric measures the fraction of detected objects correctly. For the prediction results, a result is considered as a true-positive if the IoU between it and groundtruth exceeds an IoU threshold, otherwise a false-positive. In addition, if several detection results are considered as true-positives for the same groundtruth, only one is marked, and others are considered false-positives. For the objects in images, the missed objects are considered as false-negatives, and the rests are true-negatives. Then, the precision and recall metrics can be formulated as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (8)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (9)$$

where true-positives, false-positives, and false-negatives are, respectively, denoted by TP, FP, and FN.

The AP computes the mean value of precision within the range from recall $= 0$ to recall $= 1$, i.e., the area under the precision-recall curve. Higher AP value means better performance and vice versa. The mean AP (mAP) shows the mean value of the AP of all classes.

B. Implementation Details

To prevent overfitting and gain better generalization, the VGG-16 pretrained on ImageNet data set [44] is taken as the pretrained model of RECNN, and the model of the RECNN is then trained on experiment data set. The objective function of RECNN is optimized by stochastic gradient descent (SGD) [45] with the batch size of eight, a momentum of 0.9 and weight_decay of 0.0005. The training phase of the RECNN is divided into three stages. The RECNN without an FM is trained at the first stage. The initial learning rate is 0.0001, and then it is decayed by 0.1 every 10 000 iterations. At second training phase, the trained model in the first stage is taken as the pretrained model, and the FM is only trained by freezing all of the weights in the pretrained model for better exploiting the context. In this phase, the initial learning rate is set to 0.0001, and then it is decreased similar to the first stage. Finally, based on the model trained in the second phase, we adjust all the parameters in the proposed network, where the whole RECNN is trained with the learning rate of 0.0001 and decrease by 0.1 every 3000 iterations. There are 120.3 million trainable parameters for the proposed network. The detection branch has 86 million trainable parameters, whereas the RB contains 14.7 million. Meanwhile, the rest 19.6 million parameters are contained in the base network. For the generation of a binary saliency map, the threshold value τ is empirically set as the mean value of each saliency map. The weight term β is set to 1 in our work by following the study in [35]. The experiments are run on a PC with an NVIDIA Titan X GPU, and 64 GB of memory. The operating system is Ubuntu14.04, and the implementation environment is under Caffe [46] with CUDA kernels. It takes about 19 h to complete the three stages of network training. As for testing, it takes about 0.26 s to test an image on average.

C. Parameter Optimization

The weight term α of pixel-level loss is an important parameter at the training stage of RECNN, which determines the effect of saliency information on the performance of the network. To set the optimal value of α, the relationship between mAP and α is investigated on the validation set. The mAPs of five RECNN models under five values of α, i.e. $\alpha = \{0.001, 0.01, 0.1, 1, 10\}$, are recorded in Fig. 5. The α is fixed at three phases in the training stage of each RECNN model.

The histogram of mAPs under different values of α in Fig. 5 reflects the trend on the performance of RECNN. As can be observed in the figure, the mAP of RECNN increases first and then decreases when $\alpha > 0.01$. When $\alpha = 0.001$, the loss of pixel-level plays a small role at training phase,

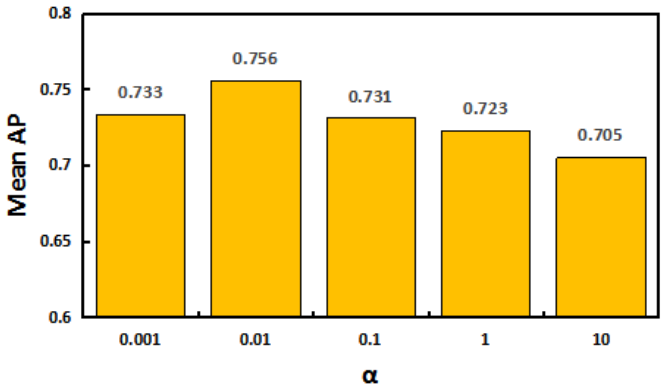


Fig. 5. Histogram of mAP obtained with the validation set under different values of the weight term α.

and saliency information makes a slight impact on RECNN, which constrains the utilization of the saliency information in the proposed network. As the increasing of α, the bigger influence is made by the loss of pixel-level constraint during objective optimization, and the saliency information is utilized to capture the regional features of salient area and weaken the influence of the background, where the score of mAP increases and reaches the highest score of 0.756 when $\alpha = 0.01$. As α continues to increase, mAP starts to decrease, where the performance of the network is limited by the inexact information in auxiliary saliency map. When $\alpha = 0.01$, the best result is obtained in the histogram, and thus the weight term α of pixel-level loss is set to 0.01 in subsequent experiments.

D. Performance Evaluation and Comparison

In this experiment, we evaluate the RECNN on the testing set. To test the performance of RECNN on multiclass object detection, we select six state-of-the-art detection methods for comparison, i.e., Collection Of Part Detectors (COPDs) [11], RICNN [28], RAMs [34], SSD [35], YOLO [37], and Faster R-CNN [26]. The COPD is a multidetector system with the hand-crafted feature, which utilizes HOG for feature representation and 45 seed-based part linear SVM classifier for detection with a sliding window approach. The RICNN learns a RICNN model, which focuses on the problem of object rotation variations in remote sensing images. The RAM accesses model distribution from training data and randomly adjusts the model at the detection stage for obtaining better adaptability. The SSD is a CNN-based end-to-end detection network with multiresolution feature maps. The YOLO is a realtime framework for detection, which can be run at a variety of image sizes to provide a smooth tradeoff between speed and accuracy. The faster R-CNN has a combined region proposal with object detection, thus completing end-to-end training for target detection. The detection results of the seven methods are shown in Table I. In order to achieve the fair comparison, the IoU threshold is uniformly set to 0.5 in the implementation of all network structures in Table I.

As can be seen in Table I, the AP scores for ten classes objects intuitively indicate the performance of each detection

Fig. 6. Some visual detection results of RECNN in testing set. The true-positives, false-positives, and false-negatives are indicated by green, red, and blue bounding boxes, respectively (categories: 1-airplane, 2-ship, 3-storage tank, 4-baseball diamond, 5-tennis court, 6-basketball court, 7-ground track field, 8-harbor, 9-bridge, and 10-vehicle).

TABLE I

PERFORMANCE COMPARISONS OF RECNN AND STATE-OF-THE-ART METHODS IN TERMS OF AP VALUES. THE BOLD NUMBER DENOTES THE HIGHEST VALUES IN EACH ROW

	COPD [11]	RICNN [28]	RAM [34]	SSD [35]	YOLO [37]	Faster R-CNN [26]	RECNN
Airplane	0.623	0.884	**0.941**	0.774	0.906	0.884	0.877
Ship	0.689	0.773	0.855	0.558	**0.907**	0.680	0.640
Storage Tank	0.637	0.853	0.859	0.581	0.709	0.777	**0.862**
Baseball Diamond	0.833	0.881	**0.896**	0.893	0.807	0.888	0.889
Tennis Court	0.321	0.408	0.671	0.827	**0.909**	0.901	0.897
Basketball Court	0.363	0.585	0.639	0.851	0.684	0.860	**0.873**
Ground Track Field	0.853	0.867	0.489	**0.997**	0.715	0.903	0.993
Harbor	0.553	**0.686**	0.629	0.335	0.322	0.515	0.563
Bridge	0.148	0.615	0.587	0.608	0.629	0.655	**0.726**
Vehicle	0.440	0.711	**0.816**	0.577	0.695	0.753	0.596
Mean AP	0.546	0.726	0.738	0.701	0.728	0.782	**0.792**

algorithm. The RECNN achieves improvement on the mAP score, which is boosted to 0.792 from 0.782, which indicates the effectiveness of the proposed algorithm. In terms of the object categories of the bridge, the gains on the AP score of RECNN is 7.1% compared with the second-best method. Compared with those detection methods in remote sensing images [11], [28], [34], the RECNN outperforms the second-best algorithm by 22.6%, 23.4%, 12.6%, and 11.1% on the object categories of tennis court, basketball court, ground track field, and bridge. In addition, the proposed network obtains comparable performance on the categories of airplane, storage tank, baseball diamond, and basketball court. However, the AP scores for ship, harbor, and vehicle are not very satisfactory. The reason may be that objects of these classes are usually small in size and tend to be densely packed, which makes it difficult to determine their boundaries. Although the AP scores for these classes are not very satisfactory, the RECNN still obtains AP gains of 8.2%, 22.8%, and 1.9% compared with SSD, which demonstrates the effectiveness of RECNN. To further demonstrate the effectiveness of the proposed algorithm, some visual detection results are shown in Fig. 6. It can be observed that, even with the large variations in the orientations, sizes of objects, and the cluttered background, most of the objects parked densely or sparsely in VHR optical remote sensing images are accurately recognized and located.

E. Ablation Study

In this experiment, we quantitatively and qualitatively evaluate the performance of each part of RECNN. The networks are trained on the training set and evaluated on the validation set for ablation study. In order to analyze the effectiveness of saliency information and fusion strategy, we implement experiments to compare the performance of four different networks with different parts of RECNN. To obtain the basic performance, Network-A is designed with truncated VGG and extraconvolutional layers, and it is trained as the baseline, where the RB and FM are both removed. To analyze the effect of FM, the reconstruction network is abandoned but the fusion part remains in Network-B. Network-C is implemented

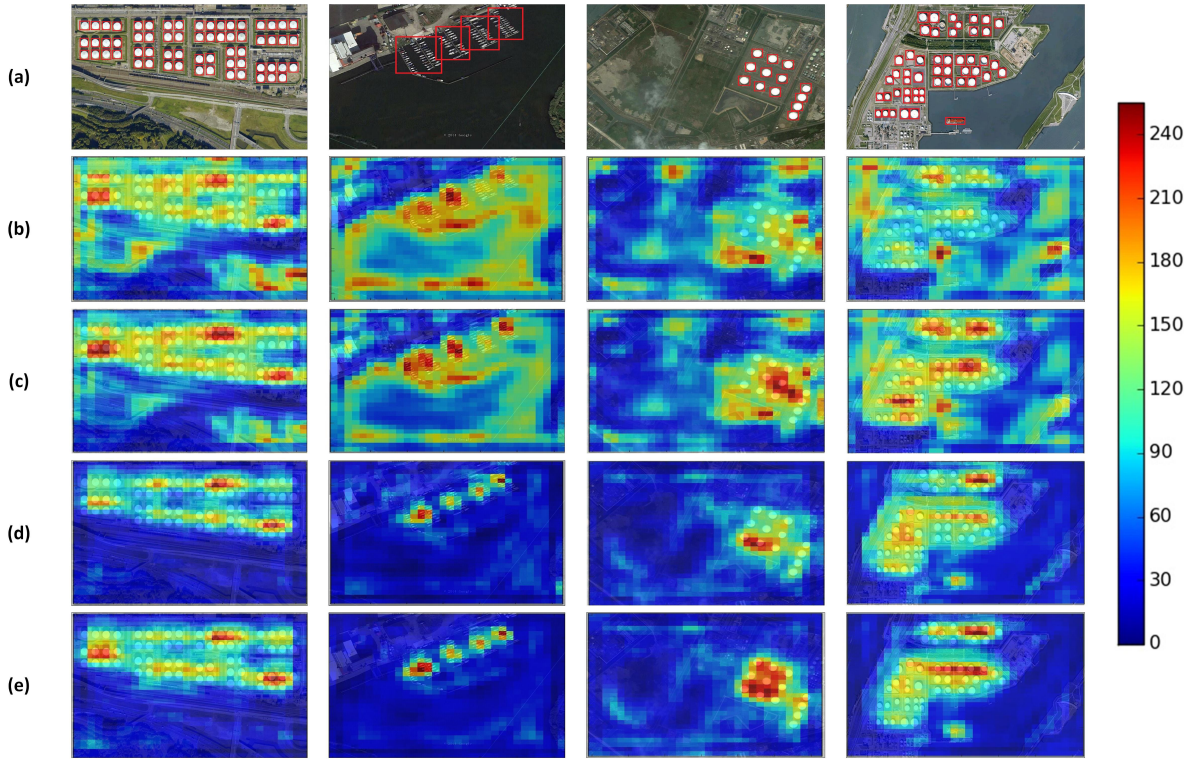


Fig. 7. Heat maps of the shared feature maps of each conv5_3 in four networks. (a) Input images. (b) Network-A. (c) Network-B. (d) Network-C. (e) RECNN.

TABLE II
COMPARISIONS OF THE DIFFERENT ARCHITECTURES OF RECNN

	Network-A (without RB and FM)	Network-B (without RB)	Network-C (without FM)	RECNN (whole network)
Airplane	0.771	0.782	0.761	0.851
Ship	0.546	0.557	0.554	0.586
Storage Tank	0.508	0.739	0.707	0.837
Baseball Court	0.893	0.898	0.888	0.890
Tennis Court	0.766	0.846	0.796	0.859
Basketball Court	0.757	0.775	0.828	0.784
Ground Track Field	0.906	0.906	0.908	0.995
Harbor	0.314	0.451	0.506	0.528
Bridge	0.628	0.681	0.698	0.721
Vehicle	0.593	0.638	0.690	0.611
Mean AP	0.668	0.727	0.733	0.756

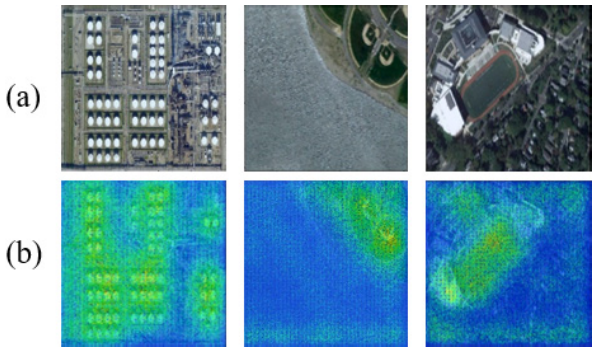


Fig. 8. Some samples of saliency maps in testing mode. (a) Input images. (b) Saliency maps.

with truncated VGG, extra convolutional layers, and the RB, which is used to analyze the influence of saliency information and the feasibility of the proposed two-branch architecture. The RECNN is also adopted for comparison to confirm the effectiveness of the whole network. The weight α of the Network-C and the RECNN are both set to 0.01.

Table II shows the AP sores for ten classes and the mAP of each network. Without auxiliary saliency information and FM, Network-A obtains baseline mAP of 0.668. For Network-B and Network-C, the FM and saliency information separately can achieve the mAP gains of 5.9% and 6.5% compared with the baseline. The effectiveness of the fusion strategy and auxiliary saliency information are confirmed by the gains of the AP score. With saliency map and FM, RECNN obtains the best result in most categories and at least 2.3% gain on mAP compared with other networks, which demonstrate the feasibility of the two-branch structure. Some heat maps of the conv5_3 layers of four networks from validation are shown in Fig. 7. A comparison of the heat maps in (b)–(e) shows that the responses of the object-like regions in shared feature maps have been enhanced by multilayer fusion, which is beneficial to strengthening the recognition ability of the network. Compared with (b) and (c), the responses of the background area in (d) and (e) are weakened, but the features of object regions are enhanced, which means the regional features and location accuracy have been improved. The contextual information is fully exploited by the FM, where the regional features in the

proposed network are thus enhanced. With the proposed two-branch structure, the reconstructed saliency information is well utilized, and the network is thus promoted to focus on the object area in the shared feature maps.

In order to further show the effect of saliency information guided training in RB, a verification experiment is performed in testing mode. Some samples of output saliency maps obtained by the RB in testing mode are shown in Fig. 8. From these results, it can be observed that the saliency information guided training process can promote the network to focus on the saliency regions and make objects of interest more distinguishable.

V. Conclusion

In this article, we presented an end-to-end RECNN for object detection in VHR optical remote sensing images. The proposed RECNN adaptively captures auxiliary saliency information by RB and pixel-level loss function, which is the first time to capture saliency information via binary semantic segmentation for object detection in remote sensing images. In addition, an FM is designed in RECNN, which exploits contextual information among multiresolution layers, and the ability of feature representation of object regions in feature maps is strengthened. The quantitative and qualitative comparison results on a publicly available NWPU VHR-10 data set have demonstrated the effectiveness of the proposed network.

In the future, we will focus on exploring the method to utilize the information in RB and expect to improve the performance of the network with reasonable architecture.

References

[1] 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.

[2] L. Cao *et al.*, "Weakly supervised vehicle detection in satellite images via multi-instance discriminative learning," *Pattern Recognit.*, vol. 64, pp. 417–424, Apr. 2017.

[3] Y. Liu, G. Gao, and Y. Gu, "Tensor matched subspace detector for hyperspectral target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 4, pp. 1967–1974, Apr. 2017.

[4] K. Stankov and D.-C. He, "Detection of buildings in multispectral very high spatial resolution images using the percentage occupancy hit-or-miss transform," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 10, pp. 4069–4080, Oct. 2014.

[5] Q. Zhang, X. Huang, and G. Zhang, "A morphological building detection framework for high-resolution optical imagery over urban areas," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 9, pp. 1388–1392, Sep. 2016.

[6] W. Zhang, X. Lu, and X. Li, "A coarse-to-fine semi-supervised change detection for multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3587–3599, Jun. 2018.

[7] Z. Zou and Z. Shi, "Ship detection in spaceborne optical image with SVD networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 5832–5845, Oct. 2016.

[8] Z. Xiao, Y. Gong, Y. Long, D. Li, X. Wang, and H. Liu, "Airport detection based on a multiscale fusion feature for optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 9, pp. 1469–1473, Sep. 2017.

[9] S. Das, T. T. Mirnalinee, and K. Varghese, "Use of salient features for the design of a multistage framework to extract roads from high-resolution multispectral satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3906–3931, Oct. 2011.

[10] J. Xie, N. He, L. Fang, and A. Plaza, "Scale-free convolutional neural network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6916–6928, Sep. 2019.

[11] 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.

[12] S. Qiu, G. Wen, Z. Deng, Y. Fan, and B. Hui, "Automatic and fast PCM generation for occluded object detection in high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1730–1734, Oct. 2017.

[13] S. Qiu, G. Wen, and Y. Fan, "Occluded object detection in high-resolution remote sensing images using partial configuration object model," *IEEE J. Sel. Top. Appl. Earth Observat. Remote Sens.*, vol. 10, no. 5, pp. 1909–1925, May 2017.

[14] D. Zhang, J. Han, G. Cheng, Z. Liu, S. Bu, and L. Guo, "Weakly supervised learning for target detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 4, pp. 701–705, Apr. 2015.

[15] L. Zhang and Y. Zhang, "Airport detection and aircraft recognition based on two-layer saliency model in high spatial resolution remote-sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 4, pp. 1511–1524, Apr. 2017.

[16] B. Li, X. Cui, and J. Bai, "A cascade structure of aircraft detection in high resolution remote sensing images," in *Proc. IEEE Geosci. Remote Sens. Symp.*, Jul. 2016, pp. 653–656.

[17] Q. Luo and Z. Shi, "Airplane detection in remote sensing images based on object proposal," in *Proc. IEEE Geosci. Remote Sens. Symp.*, Jul. 2016, pp. 1388–1391.

[18] Z. Dan, N. Sang, Y. Chen, and X. Chen, "Remote sensing object recognition based on transfer learning," in *Proc. Int. Conf. Fuzzy Syst. Knowl. Discrete*, 2014, pp. 930–934.

[19] N. Yokoya and A. Iwasaki, "Object detection based on sparse representation and Hough voting for optical remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2053–2062, May 2015.

[20] Y. Lin, H. He, Z. Yin, and F. Chen, "Rotation-invariant object detection in remote sensing images based on radial-gradient angle," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 4, pp. 746–750, Apr. 2015.

[21] G. E. Hinton, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

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

[23] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2016, pp. 1440–1448.

[24] 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.

[25] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[26] 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 Inform. Process. Syst.*, 2015, pp. 91–99.

[27] W. Diao, X. Sun, X. Zheng, F. Dou, H. Wang, and K. Fu, "Efficient saliency-based object detection in remote sensing images using deep belief networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 2, pp. 137–141, Feb. 2016.

[28] 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.

[29] A. Zhao *et al.*, "Aircraft recognition based on landmark detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1413–1417, Aug. 2017.

[30] Z. Fang, W. Li, J. Zou, and Q. Du, "Using CNN-based high-level features for remote sensing scene classification," in *Proc. IEEE Geosci. Remote Sens. Symp.*, Jul. 2016, pp. 2610–2613.

[31] 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.

[32] Z. Deng *et al.*, "An enhanced deep convolutional neural network for densely packed objects detection in remote sensing images," in *Proc. Remote Sens. Intell. Process.*, 2017, pp. 1–4.

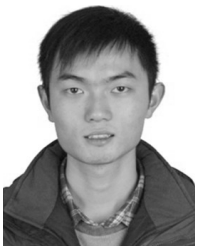
[33] F. Zhang, B. Du, L. Zhang, and M. Xu, "Weakly supervised learning based on coupled convolutional neural networks for aircraft detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 9, pp. 5553–5563, Sep. 2016.
[34] Z. Zou and Z. Shi, "Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1100–1111, Mar. 2018.
[35] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[36] C. Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," 2017, *arXiv:1701.06659*. [Online]. Available: https://arxiv.org/abs/1701.06659
[37] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," 2016, *arXiv:1612.08242*. [Online]. Available: https://arxiv.org/abs/1612.08242
[38] J. Lei *et al.*, "A universal framework for salient object detection," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1783–1795, Sep. 2016.
[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556
[40] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[41] Y. Yuan, C. Li, J. Kim, W. Cai, and D. D. Feng, "Reversion correction and regularized random walk ranking for saliency detection," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1311–1322, Mar. 2018.
[42] M. Kampffmeyer, N. Dong, X. Liang, Y. Zhang, and E. P. Xing, "ConnNet: A long-range relation-aware pixel-connectivity network for salient segmentation," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2518–2529, May 2019.
[43] J. Cheng, Y. Tsai, W. Hung, S. Wang, and M. Yang, "Fast and accurate online video object segmentation via tracking parts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7415–7424.
[44] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[45] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.
[46] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," 2014, *arXiv:1408.5093*. [Online]. Available: https://arxiv.org/abs/1408.5093

Jianjun Lei (Senior Member, IEEE) received the Ph.D. degree in signal and information processing from the Beijing University of Posts and Telecommunications, Beijing, China, in 2007.

He was a Visiting Researcher with the Department of Electrical Engineering, University of Washington, Seattle, WA, USA, from August 2012 to August 2013. He is currently a Professor with Tianjin University, Tianjin, China. He is also on the Editorial Boards of *Neurocomputing and China Communications*. His research interests include 3-D video processing, virtual reality, and artificial intelligence.

Xiaowei Luo received the B.S. degree in information and communication engineering from the Tianjin University, Tianjin, China, in 2016, where he is currently pursuing the M.S. degree.

His research interests include image processing, remote sensing image classification, and object detection in remote sensing image.



Leyuan Fang (Senior Member, IEEE) received the Ph.D. degree from the College of Electrical and Information Engineering, Hunan University, Changsha, China, in 2015.

From September 2011 to September 2012, he was a Visiting Ph.D. Student with the Department of Ophthalmology, Duke University, Durham, NC, USA, supported by the China Scholarship Council, where he was a Post-Doctoral Researcher with the Department of Biomedical Engineering, from August 2016 to September 2017. He is currently a Professor with the College of Electrical and Information Engineering, Hunan University. His research interests include sparse representation and multiresolution analysis in remote sensing and medical image processing.

Dr. Fang was a recipient of the 2nd-Grade National Award at the Nature and Science Progress of China in 2019.

Mengyuan Wang received the B.S. degree in information and communication engineering from the Tianjin University, Tianjin, China, in 2019, where she is currently pursuing the M.S. degree with the Tianjin International Engineering Institute.

Her research interests include computer vision, image processing, and object detection in remote sensing image.

Yanfeng Gu (Senior Member, IEEE) received the Ph.D. degree in information and communication engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2005.

He was a Lecturer with the School of Electronics and Information Engineering, HIT, where he was appointed as an Associate Professor in 2006, and enrolled in the First Outstanding Young Teacher Training Program. From 2011 to 2012, he was a Visiting Scholar with the Department of Electrical Engineering and Computer Science, University of California at Berkeley, Berkeley, CA, USA. He is currently a Professor with the Department of Information Engineering, HIT. He has authored over 60 peer-reviewed articles and 4 book chapters. He is also the Inventor or a Coinventor of seven patents. His research interests include hyperspectral remote sensing image processing, multimodal hyperspectral imaging, and high-resolution remote sensing image processing.

Dr. Gu serves as an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and *Neurocomputing*. He is also a Peer Reviewer of several international journals, such as the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and the *Remote Sensing of Environment*.