Context-Aware Convolutional Neural Network for Object Detection in VHR Remote Sensing Imagery

Yiping Gong, Zhifeng Xiao, Xiaowei Tan, Haigang Sui, Chuan Xu, Haiwang Duan, and Deren Li

***Abstract*— Object detection in very-high-resolution (VHR) remote sensing imagery remains a challenge. Environmental factors, such as illumination intensity and weather, reduce image quality, resulting in poor feature representation and limited detection accuracy. To enrich the feature representation and mine the underlying context information among objects, this article proposes a context-aware convolutional neural network (CA-CNN) model for object detection that includes proposal generation, context feature extraction, feature fusion, and classification. During feature extraction, we propose integrating a context-regions-of-interests (Context-RoIs) mining layer into the CNN model and extracting context features by mapping Context-RoIs mined from the foreground proposals to multilevel feature maps. Finally, the context features extracted from multilevel layers are fused into a single layer, and the proposals represented by the fused features are classified by a softmax classifier. In this article, through numerous experiments, we thoroughly explore the influence of key factors, such as Context-RoIs, different feature scales, and different spatial context window sizes. Because of the end-to-end network design approach, our proposed model simultaneously maintains high efficiency and effectiveness. We conducted all model testing on the public NWPU VHR-10 data set. The experimental results demonstrate that our proposed CA-CNN model achieves significantly improved model performance and better detection results compared with the state-of-the-art methods.**

***Index Terms*— Contextual information mining, convolutional neural network (CNN), object detection.**

I. Introduction

OBJECT detection is a fundamental and meaningful task in the fields of computer vision and image processing. With the development of remote aerospace imaging technology and the increasing volume of data available, there is an urgent need for a fast and accurate object detection algorithm. Although extensive research [1]–[9] has been conducted, accurately locating objects in a complex environment remains a challenging problem. Based on the feature extraction methods they use, object detection methods can be divided into two general approaches: handcrafted feature-based methods and deep learning-based methods. The handcrafted feature-based methods generate hundreds of candidate boxes that may contain objects of interest by using the selective search [10] and edge boxes [11] algorithms. Then, the hand-designed features, such as histograms of oriented gradients (HOG) [12], local binary pattern (LBP) [13], and scale-invariant feature transform (SIFT) [14] features, are extracted from each box based on prior knowledge. Finally, these features are classified by a classifier, such as a support vector machine (SVM) [15]. The recent success of deep learning approaches, especially the convolutional neural network (CNN), has attracted considerable interest in deep learning-based object detection methods, and many models have achieved the state-of-the-art performances [1], [16]–[21]. In contrast to artificially designed methods, deep learning methods use proposal generation methods, such as MultiBox [22], DeepBox [23], and region proposal networks (RPNs) [21] to generate fewer but higher quality candidates. Moreover, the features extracted by CNNs outperform hand-designed features.

Different from natural scene images, remote sensing images are easily disturbed by environmental interferences, such as illumination intensity and weather, resulting in low image quality and poor feature representation. To address this problem, many researchers [24]–[27] have studied the contributions of contextual information to feature representation and object detection and proven that considering contextual information helps to reduce object uncertainty and increase detection accuracy. A spatial region with a window size that is too small cannot fully capture the relationships between an object and its surroundings, while a larger window may introduce excessive noise, which can reduce the object detection accuracy. Although many works have made use of spatial context information, determining an appropriate boundary for the spatial context window is difficult because no rigorous theory exists [28], [29]. Inspired by the works in [30] and [31], we propose integrating a context-regions-of-interests (Context-RoIs) mining layer into the CA-CNN model. This layer can automatically generate Context-RoIs with a size not smaller than those of regions-of-interests (RoIs) from the potential context proposals surrounding the RoIs; then, it fuses the Context-RoIs features with the RoIs features prior to classification.

Our contributions are as follows.

1) *Context Feature Extraction:* Information from an object's surroundings is important for accurate object detection; for example, a bridge with a boat passing by is

Manuscript received August 3, 2018; revised December 1, 2018, April 8, 2019, and June 17, 2019; accepted July 16, 2019. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB0502600, in part by the Youth Foundation of China High Resolution Earth Observation under Grant GFZX04061502, and in part by the National Natural Science Foundation of China under Grant 41601443. *(Corresponding author: Zhifeng Xiao.)*

Y. Gong, Z. Xiao, X. Tan, H. Sui, C. Xu, and D. Li are with the State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: gongyp15@163.com; xzf@whu.edu.cn; cugtxw@163.com; haigangsui@263.com; xc992002@foxmail.com; drli@whu.edu.cn).

H. Duan is with SZ DJI Technology Co., Ltd., Shenzhen 518048, China (e-mail: geoduanhaiwang@163.com).

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.2019.2930246

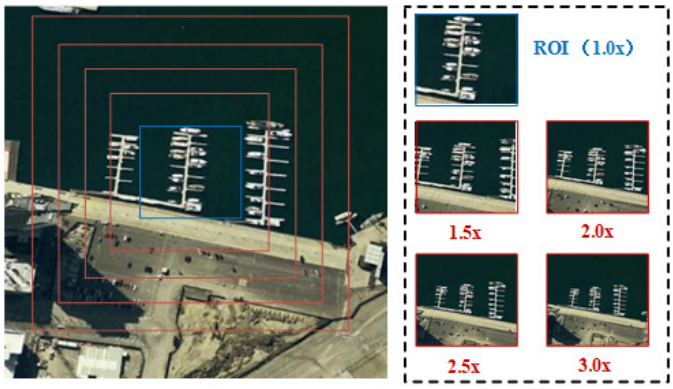

Fig. 1. Context windows with different sizes.

easy to distinguish from a road surrounded by buildings or farmland. To take advantage of the information among objects, we mine an adaptive Context-RoI for each RoI to extract contextual features; then, we fuse the contextual features with the RoI features. In addition, we present a series of experiments conducted with context windows of different sizes to estimate their effects on object detection. As shown in Fig. 1, the context window sizes from 1.5× to 3.0× are the relative sizes of the RoIs.

2) *CA-CNN:* We propose an end-to-end model that simultaneously extracts the features from RoIs and Context-RoIs and fuses them into a single feature. The experimental results on the NWPU VHR-10 data set show a great improvement in object detection performance.

II. Related Works

Object detection based on contextual information is a focus of object detection research and has received considerable attention. The poor image quality caused by interference factors, such as sensors and the environment, makes it difficult to locate objects accurately. Considerable efforts based on contextual information have been made to solve these problems. According to Biederman's division method [32], which is widely used in the field of computer vision to classify contextual information, context can be divided into three categories: semantic context, spatial context, and scale context. We describe these three categories in detail in the following.

A. Semantic Context

Semantic context describes the likelihood that an object will appear in a specific type of scene while not appearing in other types of scenes. The semantic context can also be expressed as a symbiotic relationship with other objects; for this purpose, semantic coding [29], [33] or a co-occurrence matrix [24], [34] can be used to determine the correlations between the objects. However, this method requires the accurate identification of other objects; consequently, unreliable detection information will severely affect the final detection results.

B. Scale Context

On one hand, the scale context describes the relative scales of different objects in the same scene, emphasizing their size relationships; on the other hand, it also characterizes the coarse and fine features of the same object appearing in feature maps at different levels. Fine features represent high-level semantic information that is relevant to the overall nature of the object, whereas coarse features represent low-level statistical features, such as edges and shapes. Convolutional and pooling operators cause the receptive fields to become increasingly large, which is detrimental to the detection of small objects, and the lost detail information cannot be repaired through upsampling or deconvolution. Thus, many scholars have made great efforts to preserve the useful low-level features. Yu and Koltun [35] developed a new convolutional network module that uses dilated convolutions to systematically aggregate multiscale contextual information without losing resolution. The presented context module avoids pooling operations and instead uses dilated convolutions to extend the receptive fields, thereby improving the accuracy of the state-of-the-art semantic segmentation systems. HyperNet [36] performs maximum pooling operations on lower level convolution layers and deconvolution operations on higher level convolution layers to sample all feature maps at the same size; then, it normalizes the sampled feature maps through local response normalization. The aggregation of the normalized feature maps is a hyper-feature that replaces the last convolution layer for proposal generation and object detection. Inside-outside net (ION) [31] applied RoI-pooling operations to feature maps at different scales and then concatenated the $L2$-normalized RoI feature maps. On the PASCAL VOC 2012 data set, ION achieved an improvement in object detection from the state-of-the-art mean average precision (mAP) of 73.9% to an mAP of 76.4%. This improvement provides strong evidence that context and multiscale representations improve small object detection. Multi-scale CNN (MS-CNN) [37] extracts and classifies candidate boxes from multiple output layers, combining various scale-specific detectors into a strong multiscale detector. The successes of all these techniques demonstrate that multiscale feature fusion can compensate for the semantic gap between the low- and high-level features to improve the accuracy of object detection, especially for small objects.

C. Spatial Context

Spatial context implicitly describes the symbiotic relationships between an object of interest and its surrounding environment by emphasizing positional relationships. The spatial context can be subdivided into four aspects: component context, neighborhood context, target context, and scene context. The component context refers to the spatial relationships among the components of the object itself; for example, the wings of an aircraft always appear on both sides of the fuselage. The neighborhood context refers to the spatial positional relationships between an object and the pixels within a certain neighborhood; corresponding features are extracted from pixel-level statistics in these local regions. The algorithms most commonly used for this purpose are HOG [12]

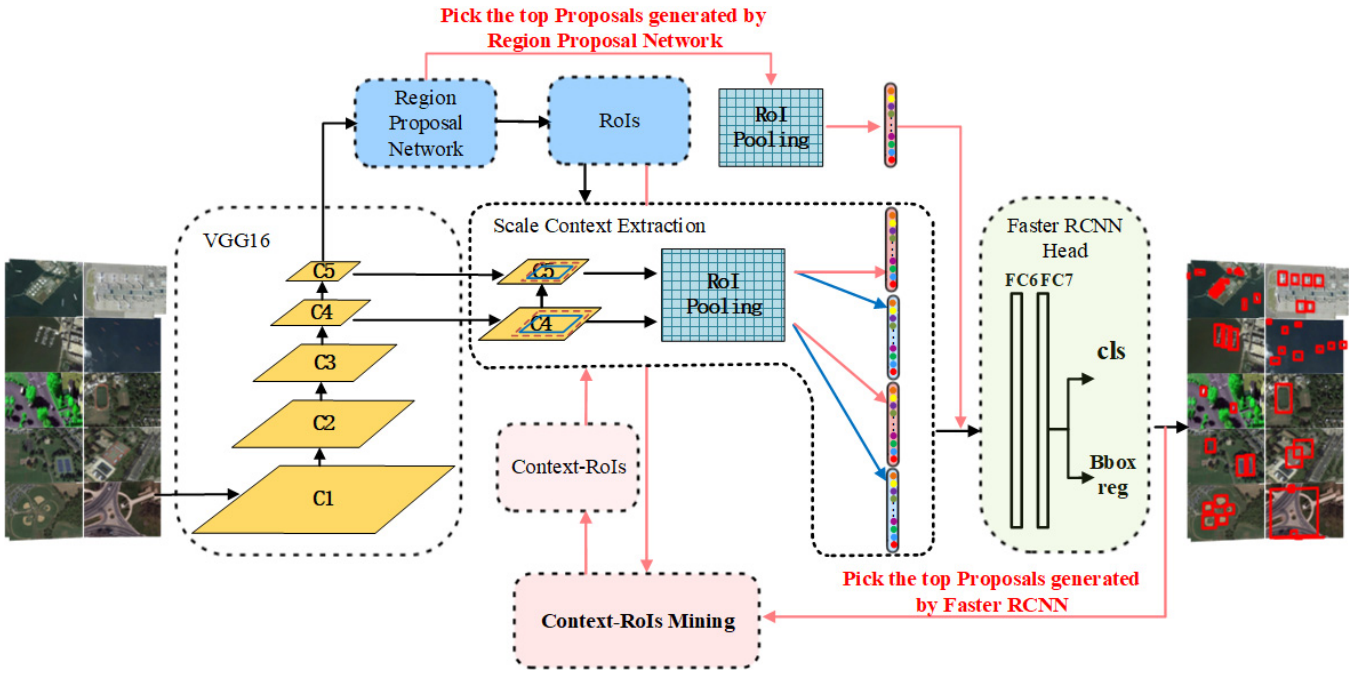

Fig. 2. Framework of the CA-CNN model.

and SIFT [38]. The target context refers to spatial location relationships among multiple objects, such as orientation and distance. For example, tennis and basketball courts often coexist. The scene context refers to the spatial context of an object and its surrounding environment. For example, aircrafts always park at airports, while ships dock in harbors or float on the sea. Many researchers have attempted to model such spatial context relationships. ION [31] makes use of pixel information from four directions around an object to extract spatial context features for object detection. MS-CNN [37], a stacked network that consists of branches of image recognition neural networks (IRNNs), extracts context features by introducing a larger window in each branch: the sizes of these windows are mined adaptively.

Although the consideration of many types of contextual information has been proposed in pursuit of more accurate object detection, the lack of a systematic theory still constrains the application of contextual information in other detection tasks. Many problems remain to be explored for practical applications, such as which features are appropriate to fuse and how to fuse them effectively. We present a detailed study of these contextual issues in Section III.

III. Methodology

In our experiments, we used the Faster RCNN from VGG16 [21] as the backbone network in our model. Fig. 2 shows an overview of our proposed CA-CNN model, which comprises three stages: proposal generation, context feature extraction, and feature fusion and classification. For context feature extraction, we add a context feature extraction layer to the CNN following the last convolution layer to generate more effective features. Finally, we use a softmax classifier for classification.

A. Proposal Generation

The purpose of region proposal generation, which is the first step in the object detection process, is to locate candidate regions that may contain objects of interest. Traditionally, to obtain as many region proposals as possible, an exhaustive search method known as the sliding window method is commonly used for proposal generation; however, this approach results in high redundancy and expensive computation. More recently, the selective search [10] and edge boxes [11] algorithms were proposed for proposal generation. These algorithms reduce the number of boxes generated and improve the detection efficiency. However, the number of region proposals generated using these methods is still too large because objects may appear anywhere. Moreover, the repeated convolution operations on each region on the input image severely affect the detection efficiency. To solve these problems and accelerate the detection process, we integrated an RPN into our CNN model to more efficiently generate higher quality region proposals by applying the sliding window method only to the final convolutional feature maps. Consequently, the convolution operation is performed only once for each image because the RPN shares the final convolution result with the primary network. The sliding windows applied to the final convolutional feature maps are called anchors. To capture objects of different sizes and shapes, we define a series of anchors with different sizes and aspect ratios. The output of the RPN consists of RoIs, whose coordinates are mapped from the region proposals in the feature maps to the corresponding regions of the original image.

B. Context-RoI Mining

To mine the underlying context information surrounding the RoIs, we propose a Context-RoIs mining layer that generates a

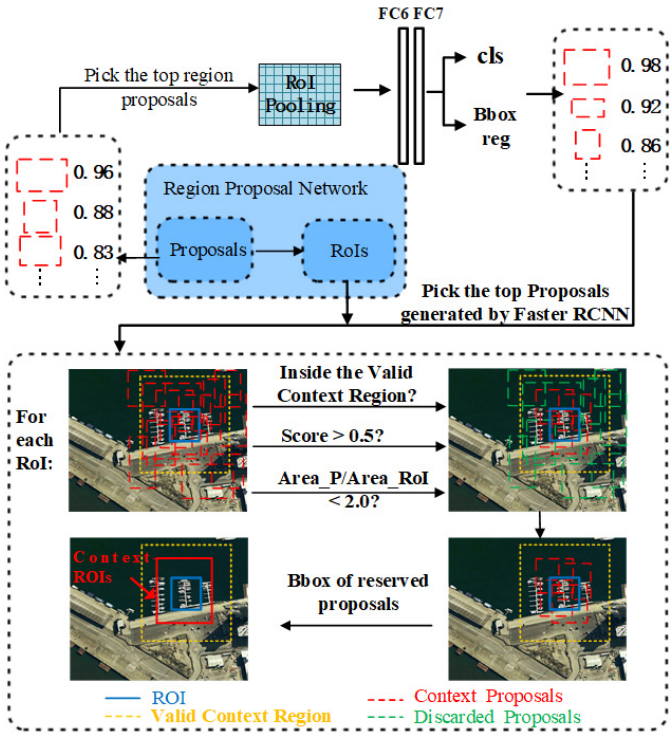

Fig. 3. Mining Context-RoIs. Blue box: RoI generated by the RPN. Red boxes: proposals used to generate Context-RoIs. Yellow box: valid context region.

Context-RoI for each RoI. An RPN generates a large number of proposals and preserves only the top 256 proposals as RoIs, which are sized to exactly cover the real objects. However, before the training process completes, the proposals generated by RPN may cover parts of the objects or the background around the objects. Therefore, we use these proposals to generate Context-RoIs to extract context information. The process of mining Context-RoIs is shown in Fig. 3, where the blue box indicates the RoI, the red boxes indicate the proposals used to generate Context-RoIs, and the yellow box indicates the valid context region. The mining process includes two components as follows.

1) *Potential Proposal Preparation:* First, all the proposals generated by the RPN are arranged in reverse order according to their probabilities, which are predicted by the softmax classifier. Then, we pick the top 10 000 proposals and send them to the core network to evaluate the probabilities and rectified coordinates of ten classes. Only the proposals that definitely belong to one of the ten categories are preserved. In other words, we keep only the proposals whose probability is above 0.5 in a certain category. During the early training stage, because the probabilities of ten categories may not exceed 0.5, we use the proposals generated by the RPN instead. Finally, the proposals are sorted in reverse order by their probabilities, and we select the top 6000 proposals to perform context proposal mining.
2) *Context Proposal Mining:* For each RoI generated from the RPN, we first define a valid context region with a size three times that of the RoI. Then, we define all the proposals as $\phi = \{p_0, p_1, \ldots, p_n\}$, where p_i represents the ith proposal modified by the RPN and n is the total number of proposals. In our experiment, n is 6000. Then, the proposals are used to generate Context-RoIs when they satisfy the following constraints.
   a) The bounding box of the proposal must be inside the valid context region. The closer the proposal is to the central target, the more effective the information it contains, and vice versa. Therefore, we set the valid region size to 3.0× relative to the RoI to avoid too much noise interference.
   b) The proposal confidence (probability) must be larger than 0.5. The proposal used to generate Context-RoIs should be a foreground box itself. The category of the foreground box surrounding the RoI may be the same as that of RoI or may differ. Regardless of the category of the foreground box, the relationship between the foreground box and the RoI is what we want to mine.
   c) The proposal area must be no more than twice that of the RoI. We assume that the context candidate proposals are similar in size to the RoI. Context window that is too large will introduce considerable noise, leading to a degradation in the RoI.
3) *Context-RoIs:* For each RoI, we simply adopt the minimum bounding box of the context proposals and the RoI as the Context-RoI.

C. Context Feature Extraction and Classification

To compensate for the semantic gap between the high- and low-level features, we map the RoIs and Context-RoIs to feature maps at different levels to extract scale context features. As shown in Fig. 4, after RoI generation and Context-RoI mining, the RoIs and Context-RoIs are first mapped to the conv4 and Conv5 layers for scale context feature extraction; then, the scale context features extracted from different layers are fused into one comprehensive feature. The fused single feature is the final feature representation for each proposal and will be fed into a softmax classifier to obtain the probability of each proposal. In addition, to reduce the bounding box redundancy, a nonmaximum suppression (NMS) algorithm is applied with a threshold of 0.3.

IV. Experimental Results

This section first introduces the data set used in our experiments and the implementation details of our proposed CA-CNN model. Then, we comprehensively analyze the key factors that influence the detection results, such as different feature scales and different sizes of spatial context windows. Next, we explore the influences of the context features on all ten classes of the experimental data set individually. Finally, we compare the results of our proposed method with those of other state-of-the-art methods.

A. Data Set Description

A large number of publicly available data sets exist for natural image processing. However, few such data sets contain

TABLE I
TRAINING AND VALIDATION DATA SETS

category	object size(pixel)	object number	enhancement method	object size after enhancement	object number after enhancement
ship	20∼90	134	R\M\D	10∼97	2144
airplane	28∼112	322	R\D	17∼112	2576
storage tank	24∼70	73	S\M\D	12∼70	1168
vehicle	19∼82	293	R\D	10∼82	2344
harbor	30∼155	108	R\M\D	16∼155	1728
tennis court	36∼79	80	R\M\D	18∼79	1280
baseball diamond	38∼134	64	R\M\S	20∼134	2560
bridge	40∼256	58	R\M\S	20∼256	2320
basketball court	42∼156	34	R\M\D\S	21∼156	2720
ground track field	131∼305	27	R\M\D\S	70∼305	2160

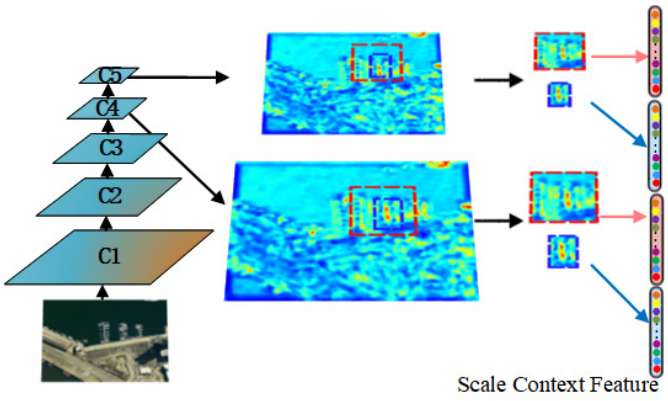

Fig. 4. Scale context feature extraction process.

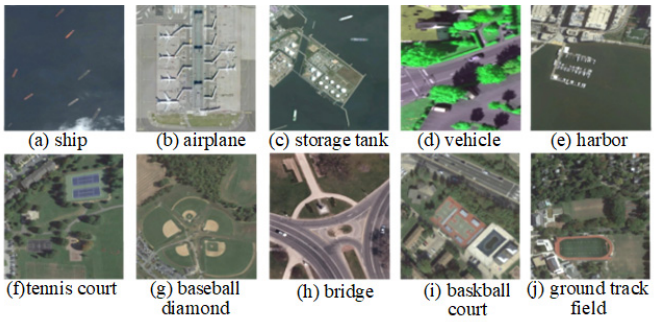

Fig. 5. NWPU VHR-10 data set. (a) Ship. (b) Airplane. (c) Storage tank. (d) Vehicle. (e) Harbor. (f) Tennis court. (g) Baseball diamond. (h) Bridge. (i) Basketball court. (j) Ground track field.

remote sensing imagery. The NWPU VHR-10 [39] data set is a challenging ten-class object detection data set commonly used in object detection tasks that were annotated by Northwestern Polytechnical University. This data set contains a total of 800 images, of which 650 contain objects and the remaining 150 contain no objects. As shown in Fig. 5, the ten object classes are ship, airplane, storage tank, vehicle, harbor, tennis court, baseball diamond, bridge, basketball court, and ground track field. Of these images, 715 are high-resolution remote sensing images collected from Google Maps with spatial resolutions ranging from 0.5 to 2 m, and the remaining 85 images are pansharpened color-infrared images with a spatial resolution of 0.08 m.

We randomly selected half the data set as the training data set and the other half as the test data set. Then, we randomly selected 20% of the training data set as a validation data set.

TABLE II
DETAILS OF THE DATA ENHANCEMENT

Enhancemnet method	Parameters
Rotation (R)	angle = [90,180,270]
Mirror(M)	horizontal
Downsampling(D)	ratio = [0.5 ∼ 0.8]
Shifting(S)	direction = [left, right, top, bottom], ratio = 0.2

TABLE III
TEST DATA SET

Category	Object size(pixels)	Object number
ship	17∼88	107
airplane	28∼110	399
storage tank	39∼86	185
vehicle	19∼82	303
harbor	25∼194	103
tennis court	36∼85	128
baseball diamond	41∼126	74
bridge	53∼203	55
basketball court	40∼207	31
ground track field	108∼277	28

Table I shows the pixel sizes and the numbers of objects in the training and validation data sets. Because the NWPU VHR-10 data set alone was not sufficient for training, we enhanced the training data set by applying rotation, mirroring, downsampling, and shifting transformations, as shown in Table II. For the downsampling method, we varied the sampling ratio from 0.5 to 0.8 for every remote sensing image. For the storage tank class, the rotation enhancement method was ineffective because of the rotational symmetry of the round storage tanks; therefore, we used only the other three enhancement methods on these images. In addition, because the number of objects in each category was unbalanced, we adopted different enhancement methods for different categories to balance the object quantities. The total numbers of objects and the object sizes in the training and validation data sets after enhancement are shown in Tables I and III lists the pixel sizes and the numbers of objects in the test data set.

B. Implementation Details

We based the construction of our CA-CNN model based on the successful VGG16 (Faster RCNN) model pretrained on the ImageNet data set [40]. The basic VGG16 model is composed of five convolution layers, two connection layers, and one classification layer. The first two convolution layers contain two individual convolution operations, and the last three convolution layers contain three individual convolution

TABLE IV
TRAINING PARAMETERS

parametes	values
Max Iteration	60000
Stepsize	30000
Learning Rate	0.001
Anchor Size	[1,2,4,8,16]
Ratios	[0.5,1,2]
Batch Size	256
Momentum	0.9
Weight Decay	0.0005
Droupout	0.5
Weights Initilization	Xaiver
Max Length of Input Image	2000 pixel
Max Length of Resized Image	2000 pixel

TABLE V
PERFORMANCES OF OUR PROPOSED CA-CNN MODEL AND THE VGG16 MODEL IN TERMS OF AP VALUES

Category	VGG16(Faster RCNN)	CA-CNN
ship	0.8932	**0.9055**
airplane	0.9878	**0.9991**
storage tank	**0.9876**	0.9001
vehicle	0.8899	**0.8900**
harbor	0.8040	**0.8897**
tennis court	**0.9037**	0.9016
baseball diamond	0.9091	**0.9965**
bridge	0.5809	**0.7962**
basketball court	0.9032	**0.9091**
ground track field	0.9057	**0.9091**
mAP	0.8765	**0.9097**

operations. Except for the last convolution layer, which is followed by an RoI-pooling layer, the convolution layers are each followed by a max-pooling layer. The activation function used in all the hidden layers is ReLU [41]. In our experiment, we extracted features from the output of the final convolution result of each convolution layer, for example, Conv5_3 and Conv4_3. The training parameters and network parameters are shown in Table IV. To capture objects of various sizes, we defined the anchor sizes of 1, 2, 4, 8, and 16 that corresponded to input image regions with the sizes of 16, 32, 64, 128, and 256 pixels, respectively. In addition, to avoid image compression, we set the max length of the input images and the resized images from 600 to 2000 pixels, which is larger than the pixel size of our data set. We train on one GPU for 60 000 steps and the weights of the whole network are initialized by Xaiver. The initial learning rate is 0.001, which is decreased by 10 at the 30 000 steps. We train on one GPU for 60 000 steps and the weights of the whole network are initialized by Xaiver.

C. Evaluation Protocols

We adopted the curve (PRC) and the average precision (AP) to evaluate the object detection performance.

1) Precision–Recall Curve: PRC is a widely used measure applied in many studies on object detection [1], [39], [42]–[44]. Assuming that TP, FP, and FN denote the numbers of true positives, false positives, and false negatives, respectively, the precision and recall metrics can be formulated as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{2}$$

A detection result will be predicted as a true positive when it has an intersection over union (IOU) with the ground truth of no less than 0.5; otherwise, it is considered a false positive.

2) Average Precision: In the field of computer vision and object detection, AP is one of the common criteria. AP is especially suitable for evaluating algorithms that simultaneously predict the position and class of a target because this criterion reflects the stability of the model. A higher AP value indicates a better model performance and vice versa. For multiclass detection tasks, mAP is often used to assess the mean performance of a model across all classes.

D. Results and Analysis

To evaluate the contributions of the context features to successful object detection, we first quantitatively compare the performance of our proposed CA-CNN model with that of the underlying VGG16 model. Then, we thoroughly analyze the factors that can strongly influence the detection results, including different object scales (feature maps at different levels) and different sizes and combinations of spatial context windows, and present an overall comparison. Finally, we compare our proposed CA-CNN model with other state-of-the-art methods.

1) Results: As introduced in Section III, we extract context features from the Conv5_3 and conv4_3 layers based on each RoI and its Context-RoIs. Table V compares the results of our proposed CA-CNN model with those of the basic VGG16 model in terms of AP. As shown in Table V, our proposed CA-CNN model achieves a detection result of 90.97%, while VGG16 achieves 87.65%. With the exception of a decrease in the storage tank, vehicle, and tennis court classes, our proposed method leads to better detection results, especially for the bridge class, which shows the largest improvement (21.53%). However, although our method achieves the best mAP, the AP values of our method for the storage tank and tennis court class are lower than those of VGG16. It is known that the features of one single object are much different from those of two or more. Context-RoIs are usually larger than the objects and cover part or whole of the objects nearby as well as the center object, making the features extracted from Context-RoIs contradicting those from RoIs. Therefore, for some storage tanks that are tightly arranged, features extracted from RoIs may bring about the interference between the objects of the same classes; however, the appearance of related objects or scenes will help to improve the detection accuracy. For bridges and harbors, the appearance of ships will improve the feature representation and detection accuracy. For ships, scenes, such as bridges or harbors, also help to improve the detection accuracy of ships.

Fig. 6 shows a visualization of the detection results of our proposed CA-CNN model and VGG16 on the ten classes. The first, third, fifth, and seventh lines show the detection results of our proposed model, while the second, fourth, sixth, and eighth lines show the results of VGG16. This results comparison yields two conclusions as follows.

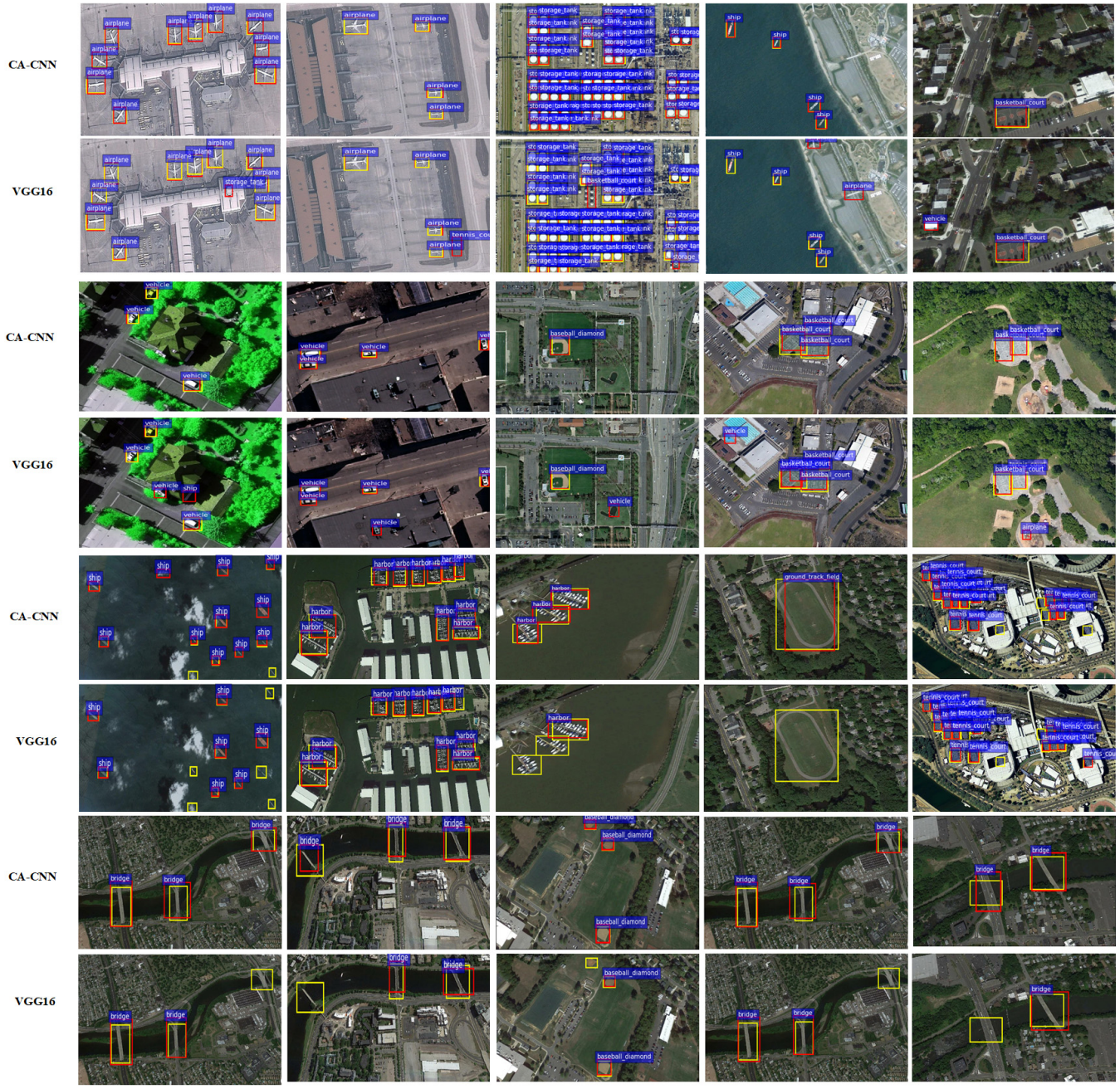

Fig. 6. Performances of our proposed CA-CNN model and VGG16. The first, third, and fifth lines show the detection results of our proposed model, while the second, fourth, and sixth lines show the results of VGG16.

1) Considering context features can reduce false positives; for example, the "tennis court" and "storage tank" predictions for the airport scene, the "basketball court" prediction for the factory, the "ship" prediction alongside the house, the "vehicle" prediction for the field, and the "airplane" prediction for the residential area generated by the VGG16 model are all eliminated in our detection results.
2) Considering context features can improve the recall. For the ship, harbor, ground track field, and bridge classes, our proposed method detects more objects than does the VGG16 model.

2) Context Feature Evaluation: We thoroughly explore the influences of the scale of features and the size of the spatial window on the object detection performance from the perspective of the mAP, mean recall (mR), and mean precision (mP), as shown in Table VI. For this comparison, we defined context window sizes ranging from 1.5× to 3.0× to explore the effects of different sizes on the detection and to explore the performance of features extracted from feature maps at different levels. In the part related to the multiscale, it can be seen that the feature levels can greatly affect the detection results. Compared with the high-level features extracted from Conv5, the lower-level features extracted from conv4 yield

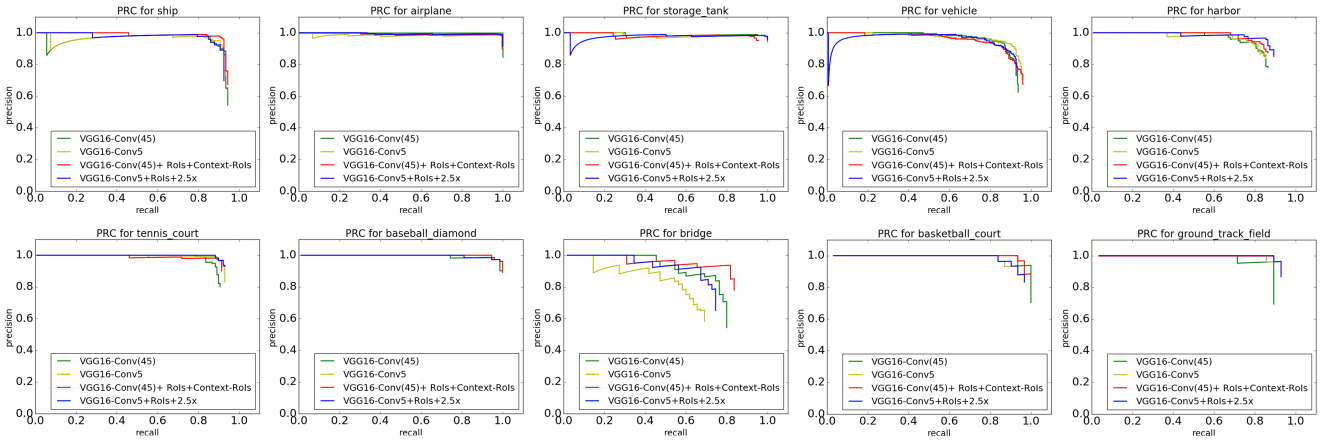

Fig. 7. PRCs of the proposed CA-CNN [VGG16-Conv(45)+RoIs+Context-RoIs] model and other approaches for the ship, airplane, storage tank, vehicle, harbor, tennis court, baseball diamond, bridge, basketball court, and ground track field object classes.

TABLE VI

OVERALL PERFORMANCE OF THE MODELS EXPLORED IN THIS ARTICLE

Method	mAP	mR	mP
VGG16-Conv5	0.8765	0.9249	**0.8172**
VGG16-Conv4	0.9021	0.9427	0.6739
VGG16-Conv(45)	**0.9033**	0.9345	0.7321
VGG16-Conv(345)	0.8899	**0.9458**	0.5580
RoIs(1.0X)	0.8765	0.9249	**0.8172**
RoIs+(1.5X)	0.8838	0.9298	0.8177
RoIs+(2.0X)	0.8883	**0.9438**	0.8287
RoIs+(2.5X)	0.8974	0.9398	0.8369
RoIs+(3.0X)	0.8905	0.9397	**0.8404**
RoIs+Context-RoIs	**0.8985**	0.9397	0.8413
VGG16-Conv(45)+RoIs+Context-RoIs	**0.9097**	**0.9397**	**0.8595**

better mAP and mP values while reducing the mP because of the tradeoff between precision and recall. Through the multilevel feature fusion, conv(45) and conv(345) both yield mAP and mR values above those of Conv5, but their mP is lower. Therefore, we can conclude that multiscale features are beneficial for improving the AP and recall. In the second part, related to the size of the spatial window, we can see that among all spatial window sizes, the automatically mined Context-RoIs method achieves the highest mAP and relatively high mR and mP scores. In the last part of the table, we extract context features from both conv4 and conv5, demonstrating that our proposed method achieves the best performance among all the investigated methods

Fig. 7 shows the PRCs for all the classes. We selected several typical models for comparison, including VGG16-Conv5 (backbone), VGG16-Conv(45), VGG16-Conv5+2.5×, and CA-CNN [VGG16-Conv(45)+RoIs+Context-RoIs]. For all categories, our proposed method achieves a detection recall of over 80%. For the bridge class, the models that use context information achieve obvious improvements compared with the basic model, VGG16-Conv5, which uses features extracted from Conv5_3. However, for the storage tank class, considering the context results in a decrease in the recall and mAP compared with the basic VGG16-Conv5 model. This result may occur because the proximity of other storage tanks interferes with feature representation and judgment for any single storage tank; therefore, for objects of this type, the features extracted from RoIs may be sufficient.

TABLE VII

COMPARISON OF OUR PROPOSED CA-CNN MODEL WITH RICA, HYPERNET, RICNN, COPD, AND EXEMPLAR-SVMS

Method	Average run time per image (seconds)
COPD	1.16
RICNN	8.47
VGG16	0.11
RICA	2.89
CA-CNN	2.82

Fig. 8 shows the mined Context-RoIs and Conv5_3 feature maps. The left column shows the ground truth (yellow), mined Context-RoIs (red dotted box), and the detected boxes (red), and the middle and right columns show the Conv5_3 feature maps of the CA-CNN method and VGG16, respectively. As shown in Fig. 8, Context-RoIs are larger than the ground truth and contain both the background information and the central object. The feature maps indicate that except for the tennis court and storage tank categories, the Context-RoIs mining method makes the objects visually more significant and easier to detect compared with VGG16. However, for tightly arranged objects such as storage tanks, the Context-RoIs mining method may introduce interference features from the neighboring storage tanks, resulting in a lower confidence fused feature and, thus, in missed detection and a reduction in precision.

3) Comparisons With Other Methods: To quantitatively evaluate the performance of our proposed method, we compared it with several state-of-the-art methods. The results are shown in Table VII. We implemented the HyperNet on our data set, while the detection results for Exemplar-SVMs, rotation-invariant CNN (RICNN), the collection of part detectors (COPD) method, and the rotation-insensitive and context-augmented (RICA) methods were reported in [18], [30], [39], and [45]. Clearly, our proposed CA-CNN model achieves the best performance among all the methods considered in Table VII, with an improvement of 3.85% compared to RICA, nearly 2.27% compared to HyperNet, 18.34% compared to RICNN, 10.29% compared to COPD, and 44.38% compared to the Exemplar-SVMs model. Although our proposed method outperforms these other methods, it achieves a lower AP for the storage tanks and tennis court classes

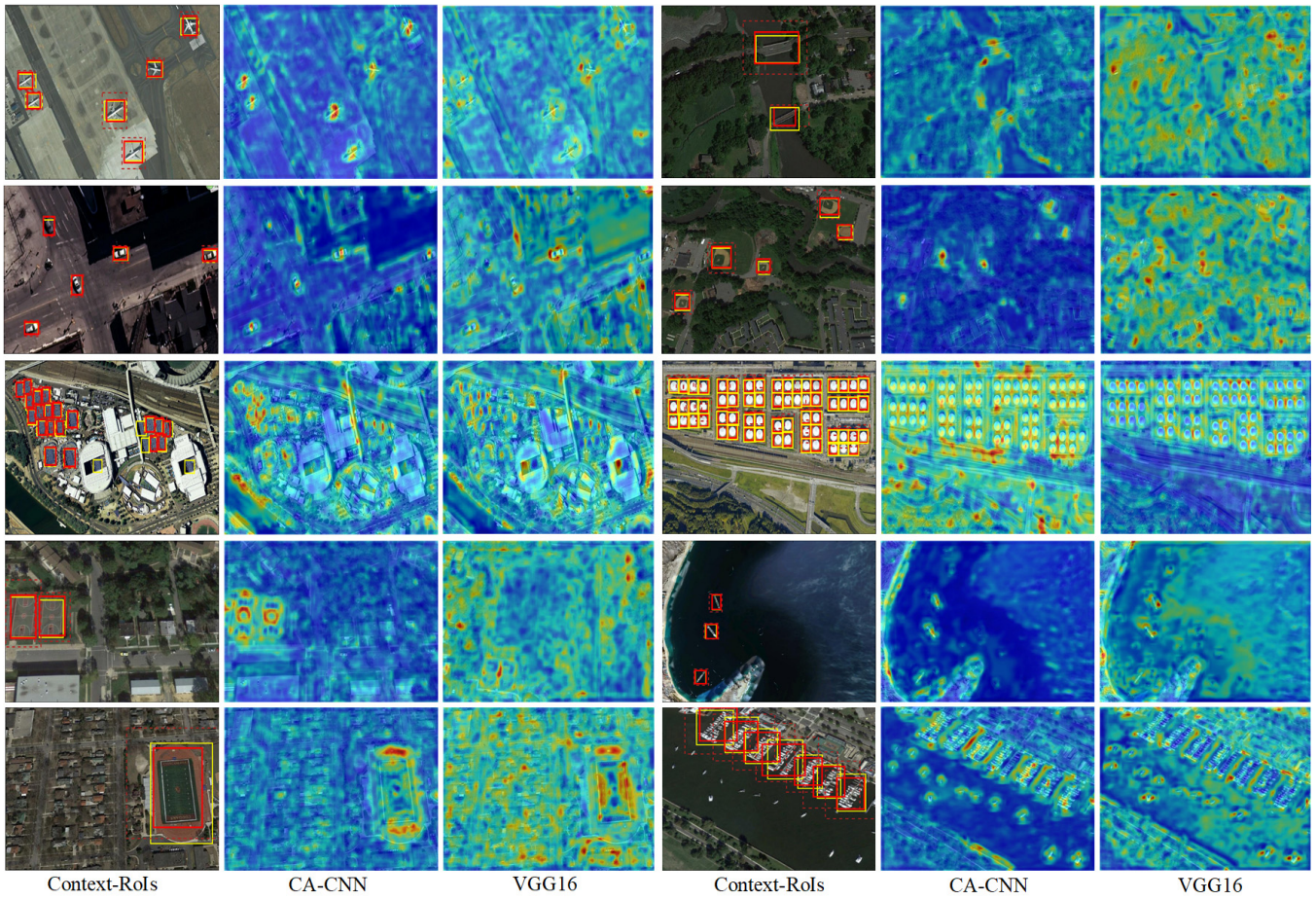

Fig. 8. Context-RoIs mined by CA-CNN and Conv5_3 feature maps of CA-CNN and VGG16. (Left column) Ground truth (yellow), mined Context-RoIs (red dotted box), and detected boxes (red). (Middle and right column) Conv5_3 feature maps of the CA-CNN method and VGG16, respectively.

TABLE VIII
COMPUTATION TIME COMPARISON OF THE FIVE DIFFERENT METHODS

Category	mAP	ship	airplane	storage tank	vehicle	harbor	tennis court	baseball diamond	bridge	basketball court	ground track field
Exemplar-SVMs	0.4659	0.3704	0.8411	0.7087	0.4600	0.3307	0.3145	0.8091	0.2414	0.4378	0.2457
COPD	0.8068	0.8173	0.8911	0.9732	0.8330	0.7339	0.7327	0.8938	0.6286	0.7341	0.8299
RICNN	0.7263	0.7734	0.8835	0.8527	0.7110	0.6860	0.4083	0.8812	0.6151	0.5845	0.8673
HyperNet	0.8870	0.8976	0.9947	**0.9869**	0.8865	0.8037	**0.9067**	0.9091	0.6890	0.9030	0.8927
RICA	0.8712	**0.9080**	0.9970	0.9061	0.8714	0.8029	0.9029	0.9291	0.6853	0.8013	0.9081
CA-CNN	**0.9097**	0.9055	**0.9991**	0.9001	**0.8900**	**0.8897**	0.9016	**0.9965**	**0.7962**	**0.9091**	**0.9091**

than does HyperNet. Table VIII shows the computation times of the five different methods. Compared with other methods, the Context-RoI mining process consumes more computation time.

V. CONCLUSION

In this article, we propose an end-to-end CA-CNN that simultaneously extracts the features from RoIs and Context-RoIs and fuse them for object detection. The experiments demonstrate the effectiveness of our CA-CNN. Our contributions are summarized in the following.

1) *Context-RoIs Mining Layer:* We propose a Context-RoIs mining layer that is integrated into the CA-CNN and proved to be effective for most objects. The experiments demonstrated that the Context-RoIs mining method achieves the mAP of 89.85% over all the test images, which is 2.2% higher than that of VGG16 (Faster RCNN). Compared with methods that use manually designed spatial windows (a range of context window sizes from 1.5× to 3.0×), Context-RoIs mining method also achieves the best performance.
2) *End-to-End Structure* The process of feature extraction and feature fusion is all integrated into CA-CNN. Moreover, instead of extracting object features at only one scale, it extracts the RoI and Context-RoI features from both multilevel feature maps. The experiments demonstrated that the use of multilevel features further improved the mAP from 89.85% to 90.97%, substantially improving the object detection performance.

REFERENCES

[1] 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.

[2] J. Leitloff, S. Hinz, and U. Stilla, "Vehicle detection in very high resolution satellite images of city areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 7, pp. 2795–2806, Jul. 2010.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.

[4] 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.

[5] S. Tuermer, F. Kurz, P. Reinartz, and U. Stilla, "Airborne vehicle detection in dense urban areas using HoG features and disparity maps," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 6, pp. 2327–2337, Dec. 2013.

[6] H. Grabner, T. T. Nguyen, B. Gruber, and H. Bischof, "On-line boosting-based car detection from aerial images," *J. Photogramm. Remote Sens.*, vol. 63, no. 3, pp. 382–396, 2008.

[7] Ö. Aytekin, U. Zöngür, and U. Halici, "Texture-based airport runway detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 3, pp. 471–475, May 2013.

[8] G. Tang, Z. Xiao, Q. Liu, and H. Liu, "A novel airport detection method via line segment classification and texture classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 12, pp. 2408–2412, Dec. 2015.

[9] 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.

[10] 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.

[11] C. L. Zitnick and P. Dollár, *Edge Boxes: Locating Object Proposals from Edges*. Springer, 2014.

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 886–893.

[13] C. Chen, B. Zhang, H. Su, W. Li, and L. Wang, "Land-use scene classification using multi-scale completed local binary patterns," *Signal Image Video Process.*, vol. 10, no. 4, pp. 745–752, 2016.

[14] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Sep. 1999, pp. 1150–1157.

[15] A. Ukil, "Support vector machine," *Comput. Sci.*, vol. 1, no. 4, pp. 1–28, 2002.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," 2013, *arXiv:1311.2524*. [Online]. Available: https://arxiv.org/abs/1311.2524

[17] 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.

[18] 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.

[19] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 2553–2561.

[20] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2018–2025.

[21] 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.*, 2015, pp. 91–99.

[22] C. Szegedy, S. E. Reed, D. Erhan, and D. Anguelov, "Scalable, high-quality object detection," *CoRR*, vol. abs/1412.1441, 2014. [Online]. Available: http://arxiv.org/abs/1412.1441

[23] W. Kuo, B. Hariharan, and J. Malik, "DeepBox: Learning objectness with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2479–2487.

[24] A. Torralba, "Contextual priming for object detection," *Int. J. Comput. Vis.*, vol. 53, no. 2, pp. 169–191, 2003.

[25] B. Li, T. F. Wu, and S.-C. Zhu, "Integrating context and occlusion for car detection by hierarchical And-Or model," in *Proc. ECCV*, vol. 8694, 2014, pp. 652–667.

[26] T.-H. Vu, A. Osokin, and I. Laptev, "Context-aware CNNs for person head detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2893–2901.

[27] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Auto-context convolutional neural network (auto-net) for brain extraction in magnetic resonance imaging," *IEEE Trans. Med. Imag.*, vol. 36, no. 11, pp. 2319–2330, Nov. 2017.

[28] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1271–1278.

[29] L. Wolf and S. Bileschi, "A critical view of context," *Int. J. Comput. Vis.*, vol. 69, no. 2, pp. 251–261, 2006.

[30] K. Li, G. Cheng, S. Bu, and X. You, "Rotation-insensitive and context-augmented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2337–2348, Apr. 2018.

[31] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2874–2883.

[32] I. Biederman, R. J. Mezzanotte, and J. C. Rabinowitz, "Scene perception: Detecting and judging objects undergoing relational violations," *Cogn. Psychol.*, vol. 14, no. 2, pp. 143–177, 1982.

[33] T. G. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," *J. Artif. Intell. Res.*, vol. 2, pp. 263–286, Jan. 1995.

[34] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[35] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: https://arxiv.org/abs/1511.07122

[36] T. Kong, A. Yao, Y. Chen, and F. Sun, "HyperNet: Towards accurate region proposal generation and joint object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 845–853.

[37] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 354–370.

[38] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[39] 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.

[40] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[41] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, Jun. 2011, pp. 315–323.

[42] G. Cheng *et al.*, "Object detection in remote sensing imagery using a discriminatively trained mixture model," *ISPRS J. Photogramm. Remote Sens.*, vol. 85, no. 9, pp. 32–43, 2013.

[43] 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.

[44] J. Han *et al.*, "Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding," *ISPRS J. Photogram. Remote Sens.*, vol. 89, pp. 37–48, Mar. 2014.

[45] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of exemplar-SVMs for object detection and beyond," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 89–96.

Yiping Gong received the B.S. degree from Lanzhou University, Lanzhou, China, in 2015, and the M.S degree from Wuhan University, Wuhan, China, in 2018, where she is currently pursuing the Ph.D. degree with the State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing.

Her research interests include object detection from remote sensing images and instance segmentation in 3D space.

Zhifeng Xiao received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2008.

From 2014 to 2015, he was a Visiting Scholar with the Computational Biomedicine Imaging and Modeling Center, Rutgers University, New Brunswick, NJ, USA. He is currently an Associate Professor with the State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing, Wuhan University. His work consists of object detection in remote sensing images, large-scale content-based remote sensing image retrieval, and scene analysis on remote sensing images. His research interests include remote sensing image processing, computer vision, and machine learning.

Xiaowei Tan received the B.S. degree from the China University of Geosciences, Wuhan, China, in 2017. She is currently pursuing the master's degree in photogrammetry and remote sensing with the State Key Laboratory of Surveying, Mapping and Remote Sensing Information Engineering, Wuhan University, Wuhan.

Her research interests include the application of semantic segmentation in remote sensing.

Haigang Sui received the B.S. degree and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1996 and 2002, respectively.

He is currently a Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University. His research interests include change detection of remote sensing, target recognition, and disaster analysis.

Chuan Xu received the B.S. degree from Huazhong Agricultural University, Wuhan, China, in 2006, and the M.S. and Ph.D. degrees from Wuhan University, Wuhan, in 2009 and 2013, respectively, where she is currently pursuing the Ph.D. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing.

Her research interests include algorithm development and application for synthetic aperture radar image segmentation, classification, image registration, and target recognition.

Haiwang Duan received the M.S. degree from the State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2017.

He is currently with an intelligence company, SZ DJI Technology Co., Ltd., Shenzhen, China. His research interests include object classification and detection.

Deren Li received the Ph.D. degree in photogrammetry from the University of Stuttgart, Stuttgart, Germany, in 1986.

He was the President of the former Wuhan Technical University of Surveying and Mapping from 1996 to 2000. He is currently a Professor and the Chair of the Academic Committee of the State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China. He has published eight books and more than 400 papers. His research interests include photogrammetry and remote sensing, global navigation satellite systems, and geographic information systems (GISs), and their innovation integrations and applications in China.

Dr. Li was elected as an Academician of the Chinese Academy of Sciences in 1991, the Chinese Academy of Engineering, and the Euro–Asia Academy of Sciences in 1995. He was the founding President of the Asia GIS Association from 2003 to 2006. He was the President of the Chinese Society of Geodesy, Photogrammetry and Cartography and the International Society for Photogrammetry and Remote Sensing Commissions III and VI. He is also the Vice President of the China Society of Image and Graphics, a Chief Scientist of the Optics Valley of China, and the Co-Chair of the Committee on Earth Observation and Satellites and the Integrated Global Observing Strategy Partnership. In the 1980s, his research findings on a posterior variance estimation-based iteration weighted method for bundler location was recognized internationally and named the Deren Li Method. His research on the separability theory of model errors scientifically solved a hundred-year baffling problem in geodetic science and earned him the 1988 Best Paper Award of the German Society for Photogrammetry and Remote Sensing and the Hansa Luftbild Award. He was a recipient of more than ten national- and provincial-level awards and prizes, such as the Sci-tech Progress Award, the National Excellent Textbook Award, and the Excellent Educational Achievement Awards.