A Contextual Bidirectional Enhancement Method for Remote Sensing Image Object Detection

Jun Zhang, Changming Xie, Xia Xu, Zhenwei Shi, *Member, IEEE*, and Bin Pan, *Member, IEEE*

***Abstract*—In remote sensing images, the backgrounds of objects include crucial contextual information that may contribute to distinguishing objects. However, there are at least two issues that should be addressed: not all the backgrounds are beneficial, and object information may be suppressed by backgrounds. To address these problems, in this article, we propose the contextual bidirectional enhancement (CBD-E) method to simultaneously remove unexpected background information and enhance objects' features. CBD-E integrates the features of different background regions sequentially in two directions. On the one hand, a gate function is used to filter out unexpected information in the background and thus improve the recall of detection. On the other hand, a spatial-group-based visual attention mechanism is adopted to enhance the features of objects to reduce the false alarm. The gate function provides an approach to selecting meaningful information in the background, while the spatial-group- based visual attention mechanism enhances the information control ability of the gate function. In the experiments, we have validated the effectiveness of both the gate function and the visual attention mechanism and further demonstrated that the proposed contextual fusion strategy performs well on two published data sets.**

***Index Terms*—Bidirectional fusion, context, remote sensing object detection, visual attention.**

I. Introduction

OBJECT detection in remote sensing images has attracted more and more attention and has achieved remarkable results in recent years [1]–[4]. Compared with images from various angles captured by ground-level sensors, remote sensing images of the overhead view typically contain richer and more distinguishable co-occurring characteristics between objects and the background [5], [6]. Usually, the objects and the background in remote sensing images have certain contextual relationship. For example, ships often appear on the sea and airplanes tent to locate on airports. Therefore, the introduction of context can significantly improve the performance of remote sensing object detection algorithms [7]–[9].

However, it is inevitable that some unexpected things appear around the detected object. Unexpected information that exists in the background may mislead the model so that true objects are identified as negative samples. For example, planes may be identified in a square, and parked cars may be detected in a wasteland. In contextual information fusion, to establish a powerful joint representation of an object, it is necessary to filter out unexpected information in the background. In this article, we construct a contextual-based remote sensing object detection network called contextual bidirectional enhancement (CBD-E). CBD-E is motivated by the idea that context generally contribute to the detection results, but under certain circumstances the background may mislead the detectors. Therefore, we propose two simultaneous strategies to address this issue: filtering out the unexpected background and enhancing the objects.

On the one hand, we integrate the gated bidirectional fusion (GBD) structure [10] into our CBD-E model to suppress the unexpected background. In GBD-Net [10], multiple sets of information of contextual regions are fused in a certain order, namely, first in a positive order and then in a negative order. The gate function implemented by convolution simulates the gate to control the information flow, and filters out unexpected information during the fusion process. By introducing the context, the risk of missed detection may be reduced. It is worth noting that the background is retained and gradually becomes dominant in CBD-E, which is quite different from existing multiscale enhancement or visual attention based approaches.

On the other hand, to further enhance the objects in the image, we improve the bidirectional fusion structure via a visual attention based approach, spatial groupwise enhancement (SGE) [11]. Unlike objects in natural scene images, objects in remote sensing are observed in the overhead view, in which case their structures tend to be stable and occlusions are seldom. In CBD-E, to force the detector to focus more on the object itself, we introduced the idea of visual attention. Additionally, subfeatures generated by neural networks can usually be distributed in multiple groups to represent various semantic entities [12]. SGE strengthens the subfeatures separately according to grouping. It is worth noting that this is more advantageous for gate functions to control the flow of information. By focusing on object itself, CBD-E reduces the risk of false alarm and improves precision.

Manuscript received May 14, 2020; revised July 27, 2020; accepted August 2, 2020. Date of publication August 11, 2020; date of current version August 20, 2020. This work was supported in part by the National Key R&D Program of China under the Grant 2017YFC1405605, in part by the Natural Science Foundation of Hebei under the Grant F2019202062, in part by the China Postdoctoral Science Foundation under the Grant 2020M670631, and in part by the Science and Technology Program of Tianjin under Grant 18ZXZNGX00100 and Grant 18YFCZZC00060. *(Corresponding author: Xia Xu.)*

Jun Zhang and Changming Xie are with the School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China, and also with the Hebei Province Key Laboratory of Big Data Calculation, Tianjin 300401, China (e-mail: zhangjun@scse.hebut.edu.cn; xiechangming@hotmail.com).

Xia Xu is with the College of Computer Science, Nankai University, Tianjin 300350, China (e-mail: xuxia@nankai.edu.cn).

Zhenwei Shi is with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China (e-mail: shizhenwei@buaa.edu.cn).

Bin Pan is with the Statistics and Data Science, Nankai University, Tianjin 300071, China (e-mail: panbin@nankai.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2020.3015049

Overall, the proposed method is inspired by the effectiveness of contextual correlation, where removing unexpected background and enhancing the objects are conducted simultaneously. The two contributions of CBD-E can be summarized as follows.

1) We develop a GBD structure to suppress the unexpected background in the context area around the object.
2) We conduct SGE to improve the object saliency and highlight the features of the object.

II. Related Works

A. Generic Method

Object detection methods based on deep learning are widely used in remote sensing scenes. These detectors are broadly divided into two types: two-stage and single-stage. Two-stage methods have better results in detection accuracy. They first generate some candidate boxes, and then further determine the category and adjust the position, such as Faster RCNN [13]. One-stage methods have a faster speed when inferring. They do not need to generate proposals, such as SSD [14] and YOLO [15]. The researchers designed the feature pyramid networks (FPN) [16] with FPN based on the two-stage algorithm. In recent years, weakly supervised detectors have been developed, which only require scene-level annotations for training. WSDL [17] exploits both the separate scene category information and mutual cues between scene pairs to sufficiently train deep networks. Our proposed CBD-E uses FPN as a baseline. We will introduce it in detail in Section III.

B. Context Method

In recent years, researchers have tried various methods to use context to enhance the performance of remote sensing detectors. Liu *et al.* proposed detection methods [18]–[21] to extract context by segmenting remote sensing images before detecting objects. Based on the prior knowledge of an object and its context, Sun *et al.* established a robust context model [22]–[24] to obtain the degree of correlation between an object and its context. The rapid development of deep neural networks, especially convolutional neural networks (CNNs), enables the context to be fully utilized. Zhao *et al.* extracted features and designed an object detection model [25], [26] using a CNN for fusion context based on the conditional random field. Zhang *et al.* designed a deep learning model [27] with a contextual feature extraction structure. In the researchers' model, the context structure is built on the deepest of several feature extraction branches, and contextual information is introduced to each layer of the detection branches. Yang *et al.* designed two-stage object detectors that consisted of a region proposal stage and a refinement stage [28]–[30]. Before the refinement stage, the authors introduced the features of the region around the candidate box as context to be integrated into the local features. Gong *et al.* proposed CA-CNN [31] that obtains scene information by mapping contextual regions of interest (RoIs) mined from the foreground proposals to multilevel feature maps. Ke *et al.* designed LCFFN [32] with multiple branches. The model extracts context from RoIs that expand by a fixed ratio and combines them with local features. The essence of the aforementioned methods is to obtain object features that contain the richest possible contextual information to allow the subsequent structure to perform more accurate classification and regression.

C. Gate Function

Opening a gate means allowing someone or something to pass, otherwise it means blocking them. The gate function is a structure that simulates the gate through a mathematical operation. It is usually used to transmit or block some information. In LSTM [33], researchers have designed various gate functions to realize the update of long-term memory and short-term memory. In GRF [34], the researchers use the characteristics of the gate function to control the features to select the appropriate branch. In highway networks [35], the gate function is used to open a channel to the deep layer, which solves the problem of gradient disappearance. In our proposed CBD-E, the gate function is designed to filter out unexpected information in the background.

D. Visual Attention

When observing things, humans always have different levels of attention in focus and background. Inspired by the human visual system, the researchers designed the attention structure to simulate the focus of human observation. This mechanism is widely used in object detection in remote sensing scenes. Chen *et al.* proposed a multiscale spatial and channel-wise attention mechanism for enhancing objects in different backgrounds [36]. In MA-FPN [37], the researchers extract attention information from shallow features and optimizes deep features to make the network track object regions more accurately. Xue *et al.* developed a supervised multidimensional attention network to detect small object in a cluttered background [38]. In our proposed CBD-E, the visual attention is designed to enhance the objects.

III. Method

The overall framework of CBD-E is shown in Fig. 1. First, we generate three context regions with different sizes based on the candidate box. The feature map of each region is obtained through RoIAlign [39], and they have the same size. Then, we use the bidirectional fusion structure to interact the features of the original region and the background region in order of size. In the process of transmission, we filter out the unexpected information in the background through the gate function. And the visual attention mechanism is used to enhance the features of objects suppressed by context. Finally, the fused feature maps are fed into the parameter sharing detection network to generate multiple prediction boxes, and the optimal box is selected by nonmaximum suppression (NMS). More details are discussed in the following sections.

A. Background

In the current detection methods, CNNs are widely used as an effective way of extracting features [40]–[43]. In CBD-E, FPN is used to generate abstract feature maps that describe remote sensing image content. In the current two-stage detection

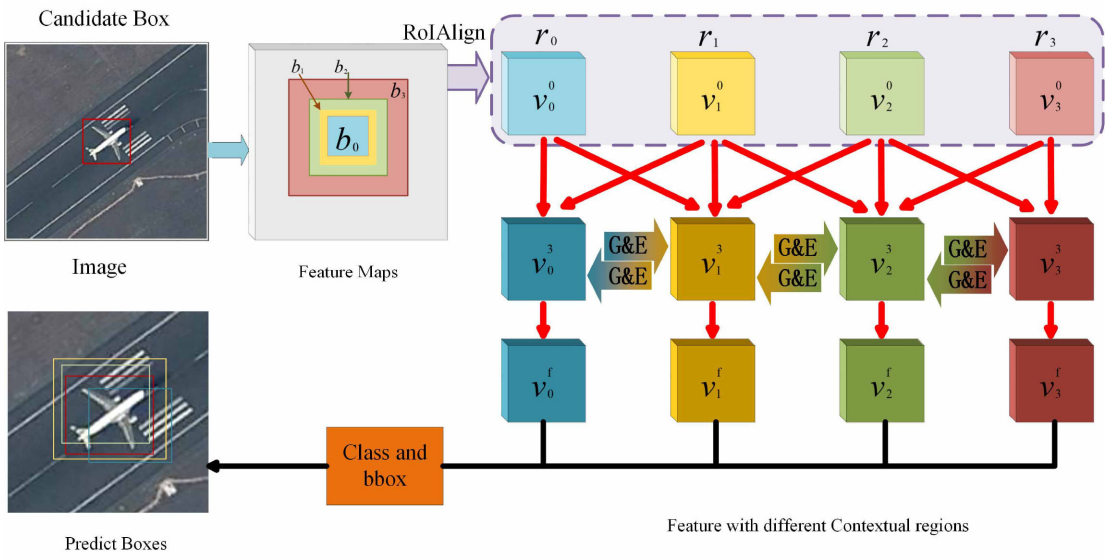


Fig. 1. Overview of our framework. Parameters on black arrows are shared across branches, while parameters on red arrows are not shared. G and E represents gate function and visual enhancement, respectively.

algorithms, a rough set of object proposals is usually generated first as candidate boxes and then is refined by adjustment of their coordinates and prediction of their categories [44], [45]. In this article, we generate the candidate boxes on the feature map through region proposal network (RPN) [13] and use r to represent the region of a candidate box. Mathematically, we define the coordinate expression b of r as follows:

$$b = [x, y, w, h] \tag{1}$$

where x and y represent the central point coordinates of r, and w and h represent the length and width of r, respectively.

Based on the candidate box b_0, RoIAlign is usually used to obtain a set of fixed-size regional feature maps v_0 by twisting and scaling. Then, the original refinement procedure of the existing two-stage detectors commonly entails performing the following operations to refine region r_0:

$$\begin{cases} s = f_{\text{cls}}(v_0) \\ b_f = f_{\text{reg}}(b_0, v_0) \end{cases} \tag{2}$$

where f_{reg} and f_{cls} are regression and classification operations, respectively. In a two-stage detector, f_{reg} is generally a linear regression operation used to obtain the coordinates b_f of the predicted box, and f_{cls} is usually a softmax operation used to obtain the confidence score s of the refined predicted box. If the context is not considered, b_f is usually the final predicted box of the object.

B. Context Fusion Method Based on a GBD Structure

In a remote sensing scene, the specific background where the object is located will often provide the information crucial to distinguishing the object. In this article, we introduce a gated bidirectional structure into remote sensing object detection to fuse information from different context regions. In the following sections, we will show how to select different regions and the fusion process.

1) RoIAlign of Features for Different Context Regions: The feature maps of several different support regions of an object are selected as the contextual information of the object. In CBD-E, we obtain the context region r_p of r_0 through b_0. The following actions are usually performed:

$$b^p = [x^0, y^0, (1+p)w^0, (1+p)h^0] \tag{3}$$

where b^p is the coordinate of r_p. In CBD-E, we adjust the value of p to obtain three sets of coordinates b_1, b_2, and b_3, which correspond to the three context regions r_1, r_2, and r_3, as shown in Fig. 1. From the corresponding regions, features v_1, v_2, and v_3 are obtained through RoIAlign and warped into $14 \times 14 \times 256$ to obtain the same size.

The context scale value p determines the amount of padded context. A large p value means that a larger background region can be obtained and that richer contextual information can be introduced. Different from the multiscale enhancement algorithm that always pays attention to the features of the object itself, CBD-E pays more attention to the contextual information when the p is larger. As p increases, the features of the object itself will be compressed more, and contextual information will dominate. A smaller p value means that more features of the object itself can be retained but that less contextual information will be introduced. Since each region has the same central point, there is an inevitable overlap between the context regions. If the difference between p values is smaller, the model will also contain more redundant information. When the value of p is very large, the context region of a large object may exceed the range of the feature map. During the experiment, 1.7 is a reliable upper limit of p. The context close to the object is usually more closely related to the object. Therefore, as the value of p decreases, the interval of p decreases. Next, we set $p = 0.4$, 0.8, and 1.7 by default, corresponding to r_1, r_2, and r_3, respectively.

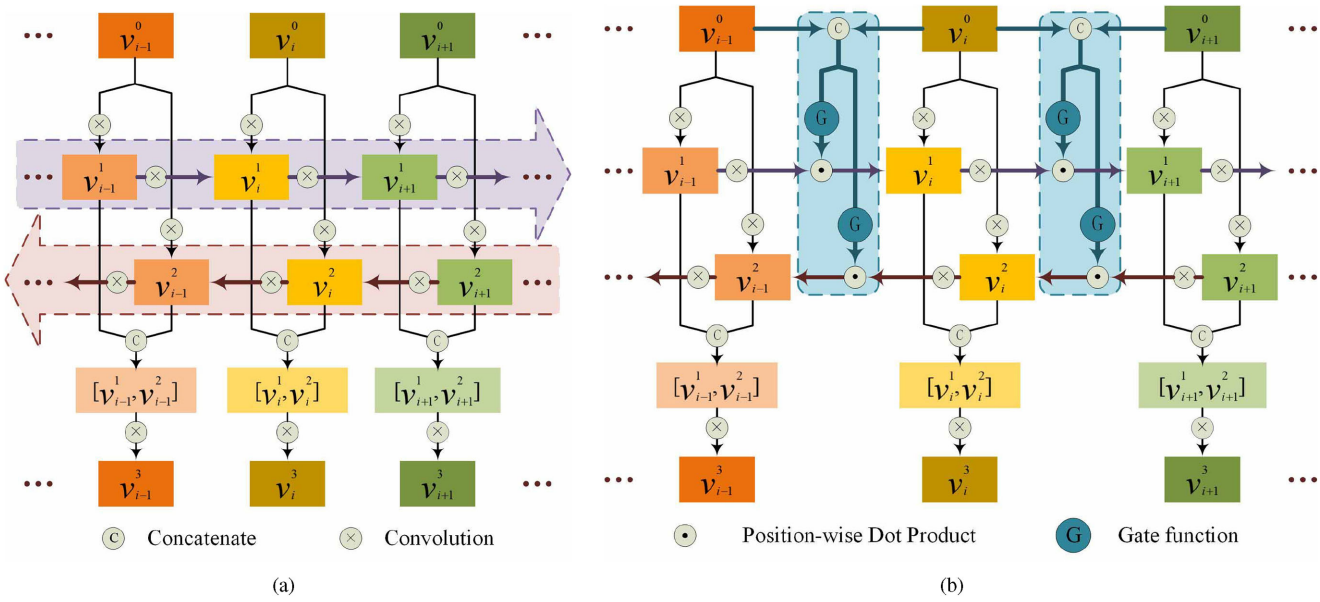


Fig. 2. (a) Illustration of bidirectional structures that fuse features of different regions. (b) Illustration of a bidirectional fusion structure with gate functions.

2) Bidirectional Fusion Structure: Fig. 2(a) shows the architecture of a bidirectional fusion structure. It takes features v_0, v_1, v_2, and v_3 as inputs and outputs features v_0^3, v_1^3, v_2^3, and v_3^3 for a single candidate box. It builds two-directional connections between multiple regions of different sizes. One directional connection starts from features with the smallest region size and ends at features with the largest region size. The other does the opposite. The connection starting from the smallest region gradually introduces the object into the larger background. This means that the category of the object is gradually verified in a larger background. The change of the region size from large to small means that the category of the object is gradually inferred from the continuously shrinking background. For example, a baseball diamond is located in the playground, and the playground is located in the school. The gradient change of region size enables information of different regions to be gently fused together, and the resulting fusion structure will be more stable.

For a single candidate box b_0, v_i is represented as v_i^0, which is the input of the ith branch. The forward propagation for the proposed bidirectional structure can be summarized as follows:

$$v_i^1 = \sigma(v_i^0 \otimes w_i^1 + b_i^{0,1}) + v_{i-1,i}^1 \tag{4}$$

$$v_{i-1,i}^1 = \sigma(v_{i-1}^0 \otimes w_{i-1,i}^1 + b_{i-1}^1) \tag{5}$$

$$v_i^2 = \sigma(v_i^0 \otimes w_i^2 + b_i^{0,2}) + v_{i,i+1}^2 \tag{6}$$

$$v_{i,i+1}^2 = \sigma(v_{i+1}^0 \otimes w_{i,i+1}^2 + b_{i+1}^2) \tag{7}$$

$$v_i^3 = \sigma(\text{cat}(h_i^1, v_i^2) \otimes w_i^3 + b_i^3) \tag{8}$$

where $\otimes$ represents the convolution operation. Variables b_*^* and w_*^* represent the biases and filters of convolutional layers, respectively. Operator $\sigma(\cdot)$ represents an elementwise RELU used as a nonlinear function. Function cat() concatenates CNN feature maps along the channel direction. If i =0, v_0^1 is only generated by v_0^0: $v_0^1 = \sigma(v_0^0 \otimes w_0^1 + b_0^{0,1})$. If i =3, v_3^2 is only generated by v_3^0: $v_3^2 = \sigma(v_3^0 \otimes w_3^2 + b_3^{0,2})$.

To enhance stability, we add an identity mapping, as is widely done in practice [46], [47]. The operation is as follows:

$$v_i^f = v_i^0 + \alpha v_i^3. \tag{9}$$

The result v_i^f will be sent to the subsequent detection network for classification and regression. Constant α is a crucial parameter; its impact is discussed in the following sections. In subsequent sections, α defaults to 0.1.

3) Gate Functions for Filtering out Unexpected Information: The background region of an object includes unexpected information that is detrimental to improved detection performance. Instead of retaining and passing all the information of the background to the next region during the fusion process, we introduce gate functions to filter out unexpected information and adapt message passing for individual candidate boxes.

Our gate function is generated by two adjacent branches. In the approach of GBD-Net [10], the gate function generated by v_{i-1}^0 controls the transfer of information from $(i-1)$th branch to ith branch. This means that only $(i-1)$th branch controls what messages are passed. Unlike GBD-Net, we believe that ith branch that receives information should also participate in rate control. Therefore, we introduce v_i^0 in the generation of the gate function, as shown in Fig. 3. And the others are also generated by two data sources. We analyzed the contribution of this change in Section IV-F.

Fig. 2(b) shows the application of gate functions. A gate function is implemented by convolution with a kernel size of 3 $\times$ 3 and the sigmoid nonlinear function. The generation process and control information passing of the gate structure are as follows:

$$G_i^1 = \text{sigm}(\text{cat}(v_{i-1}^0, v_i^0) \otimes w_{i-1}^g + b_{i-1}^g)) \tag{10}$$

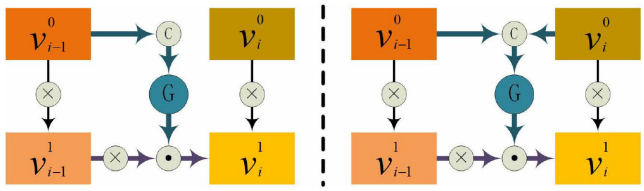


Fig. 3. Illustration of the gate functions generated by different data sources. The right function is used in CBD-E.

$$G_i^2 = \text{sigm}(\text{cat}(v_i^0, v_{i+1}^0) \otimes w_{i+1}^g + b_{i+1}^g)) \tag{11}$$

$$v_i^1 = \sigma(v_i^0 \otimes w_i^1 + b_i^1) + G_i^1 \bullet v_{i-1,i}^1 \tag{12}$$

$$v_i^2 = \sigma(v_i^0 \otimes w_i^2 + b_i^2) + G_i^2 \bullet v_{i,i+1}^2 \tag{13}$$

where G is the gate function used to control message passing, $\bullet$ denotes elementwise product, w_*^g and b_*^g are weights and offsets of the convolution, respectively, and $\text{sigm}(x) = 1/[1 + \exp(-x)]$. The value of G represents the passing rate of information, so it should be within the range of (0, 1). The output of sigm() is in the range of (0, 1). And the function is derivable and its curve is symmetrical around 0.5. So sigm() is suitable as a gate. In element-level operations, multiplying by 0 means that the information is erased. When $G = 0$, it means that the pass rate is 0, and the unexpected information is erased. Conversely, the context that is helpful for detecting objects is allowed to pass by the gate.

C. Enhance Objects Features Based on Visual Attention Mechanism

In this article, we introduce a visual attention mechanism to prevent the introduction of contextual information leading to suppressing the object information. SGE adjusts the importance of subfeatures of corresponding positions by generating an attention factor for each spatial position in each semantic group. Unlike SENet [47], which focuses on partial channel information, it enables each group to autonomously enhance its learning expression. In our method, SGE can enhance the fused information v_i^1 and v_i^2 grouping in the context fusion process, as shown in Fig. 4.

In SGE, feature maps are first divided into E groups according to dimension. Then, each group in a space has a vector representation $\mathcal{V}$, $\mathcal{V} = \{v_1 \ldots v_m\}, v_j \in \mathbb{R}^{\frac{C}{E}}, m = h \times w$. We further assume that this group gradually captures a specific semantic response (such as the serve area of a tennis court) during network learning. In this group space, it is ideal to obtain the feature maps that have a strong reaction only on the serve area. And the rest of it is barely activated and becomes a zero vector. However, due to the inevitable noise and the existence of similar patterns, it is usually difficult for CNNs to obtain a well-distributed feature response. Therefore, we use the overall information of the whole group space to further enhance the learning of the semantic feature of crucial regions. Specifically, we use the global statistical feature obtained through the spatial averaging function $Fgp(\cdot)$ to approximate the semantic vector this group learns to represent. For a single feature group $\mathcal{V}$, the operations are as follows:

$$e = F_{gp}(v) = \frac{1}{m}\sum_{j=1}^{m} v_j. \tag{14}$$

Next, using the global feature, we can enhance the features of this group. The enhancement operation realized by the dot product enhances or weakens the features of different groups to different degrees. Therefore, the model pays more attention to some particular grouping. For each position, we have

$$c_j = e \cdot v_j. \tag{15}$$

To prevent the bias in the magnitude of coefficients between various samples, we normalize c over the space

$$\hat{c}_j = \frac{c_j - \mu_c}{\eta_c + \epsilon}, \mu_c = \frac{1}{m}\sum_{k}^{m} c_k, \eta_c^2 = \frac{1}{m}\sum_{k}^{m}(c_k - \mu_c)^2 \tag{16}$$

where ϵ is a constant used to enhance stability, with the value of 1e-5. To ensure that the normalization of the insertion network can represent the identity transformation, we introduce two parameters γ and β. Compared with other attention methods [48]–[50], it is more lightweight with fewer parameters. For each coefficient $\hat{c}$

$$a_j = \gamma\hat{c}_j + \beta. \tag{17}$$

Parameters γ and β are trainable; γ is initialized to 0, and β is initialized to 1. Finally, to obtain the enhanced feature vector $\hat{v}_j$, the original v_j is scaled by the generated crucial coefficients a_j via a sigmoid function

$$\hat{v}_j = v_j \cdot \text{sigm}(a_j). \tag{18}$$

Ultimately, all feature groups in $\hat{\mathcal{V}} = \{\hat{v}_1 \ldots \hat{v}_m\}$, $\hat{v}_j \in \mathbb{R}^{\frac{C}{E}}$, $m = h \times w$ will be enhanced. Fig. 5 illustrates the GBD structure with SGE.

D. Loss Function

In the training step, we use a multitask loss function, like Faster RCNN. The loss of RPN includes two parts: classification loss L_{rc} and regression loss L_{rl}. In the second stage, the loss includes classification L_{sc} and regression L_{sl}. L_{rc} is log loss over two classes. And L_{sc} is a multiclass cross-entropy loss. The loss function for an image is defined as

$$\begin{aligned} L = {} & \lambda_0 \frac{1}{N_{rc}} \sum_{f=1} L_{rc}(z_f^{c1}, t_f^{c1}) + \lambda_1 \frac{1}{N_{rl}} \sum_{q=1} t_f^{c1} L_{rl}(\mathbf{z}_q^{l1}, \mathbf{t}_q^{l1}) \\ & + \lambda_2 \frac{1}{N_{sc}} \sum_{f=1} L_{sc}(z_f^{c2}, t_f^{c2}) + \lambda_3 \frac{1}{N_{sl}} \sum_{f=1} t_f^{c2} L_{sl}(\mathbf{z}_f^{l2}, \mathbf{t}_f^{l2}). \end{aligned} \tag{19}$$

Here, t_f^{c1} is 1 if the anchor point is positive, otherwise it is 0. And t_f^{c2} is the classification ground-truth label. f is the index of an anchor in a minibatch, and q is the index of an anchor of the image. λ_* are weights, and N_{**} are used to normalize the corresponding loss part. z_*^{c*} are the classification prediction values at different stages. $L_{*l} = \sum_{i=1}^{4} \mathbb{Z}(z_{f,i}^{l*} - t_{f,i}^{l*})$, and $\mathbb{Z}$ is Smooth L1 loss. The ground-truth bounding box offsets is

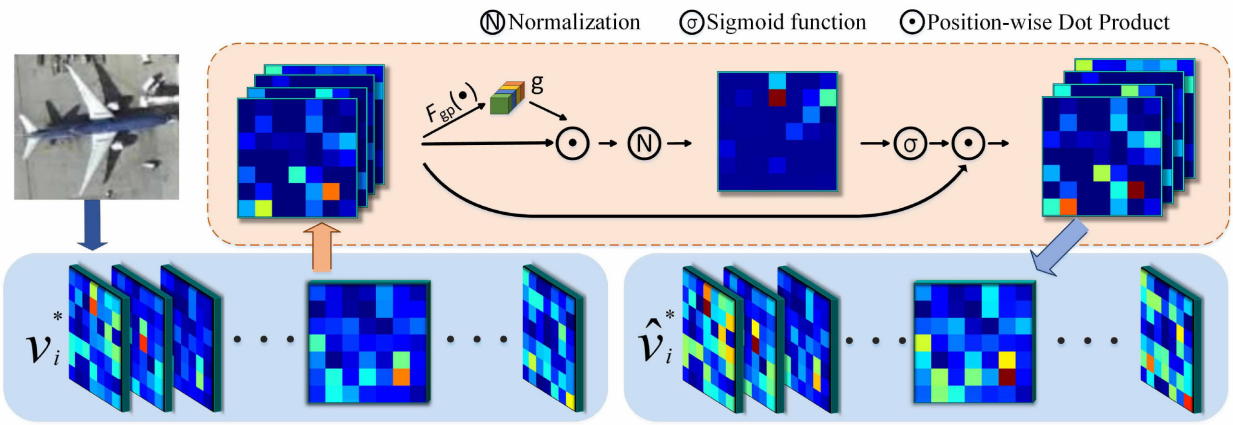


Fig. 4. Illustration of the SGE module enhancement process. The subfeatures of each group are processed in parallel, and the correlation between the global statistical features of the whole group and the local positional features of the group is used as the attention guidance to enhance the features of the group.

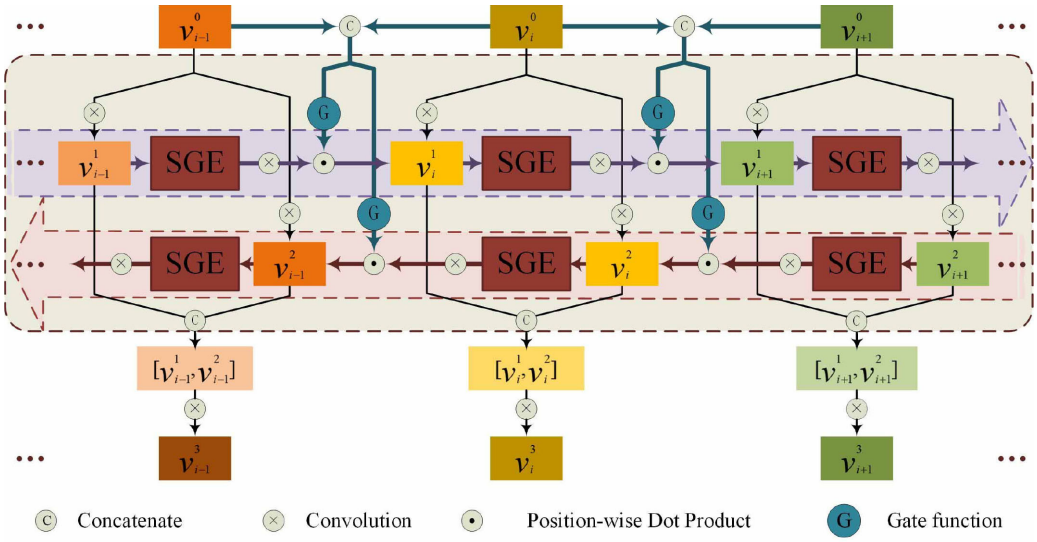


Fig. 5. Illustration of our GBD structure with SGE.

denoted by $\mathbf{z}_f^{l*} = [z_{f,1}^{l*}, z_{f,2}^{l*}, z_{f,3}^{l*}, z_{f,4}^{l*}]$. And the predicted offset is denoted by $\mathbf{t}_f^{l*} = [t_{f,1}^{l*}, t_{f,2}^{l*}, t_{f,3}^{l*}, t_{f,4}^{l*}]$.

E. Discussion

CBD-E fuses accurate and concise contextual information without suppressing features of the object of interest. In context fusion, we filter out unexpected information by the gate function and enhance the feature of the object by SGE. Finally, multiple groups of feature maps rich in contextual information will predict the object through the subsequent detection network. Parameters are shared during detection. Fig. 1 shows the detection process of CBD-E, where the contextual information of r_0 is provided by feature maps of r_1, r_2, and r_3. In conjunction with the contextual information, the following refinement operations are usually performed in the second stage:

$$\begin{cases} (s_0, s_1, s_2, s_3) = f_{cls}(v_0^f, v_1^f, v_2^f, v_3^f) \\ (b_{f0}, b_{f1}, b_{f2}, b_{f3}) = f_{\text{reg}}(b_0, v_0^f, v_1^f, v_2^f, v_3^f) \end{cases} \tag{20}$$

where v_1, v_2, and v_3 are the regional feature maps of b_1, b_2, and b_3, respectively, $(b_{f0}, b_{f1}, b_{f2}, b_{f3})$ are the four predicted boxes obtained by regression of b_0, and (s_0, s_1, s_2, s_3) are confidence scores of the four predicted boxes. The prediction for an object can only be one box instead of four. The final prediction box will be selected from the four predictions by proper NMS. The detailed process of CBD-E is shown in Algorithm 1.

IV. Experiments

In this section, we examine the performance of CBD-E through several experiments. We first introduce the data set and the evaluation methods used in the experiment. Then, we describe the implementation details of CBD-E. Next, we compare

Algorithm 1: Contextual Bidirectional Enhancement.

Input: Candidate box $b_0 = [x^0, y^0, w^0, h^0]$, contextual regions number I, a set of context scale values $P = \{p_0 = 0, \ldots, p_I\}$, feature maps extracted by CNN V.
Output: Final prediction box b_f.

```
for i = 0 to I do
    Obtain contextual box b_i by scaling the b_0 by p_i, based on (3).
    Obtain the contextual feature v_i^0 of b_i on V by RoIAlign.
    if i == 0 then
        v_i^1 is generated by v_i^0 through convolution.
    else
        Generate G_i^1 by v_i^0 and v_{i-1}^0.
        v_{i-1}^1 is visually enhanced by SGE and filtered by the gate G_i^1.
        Obtain v_i^1 by fusing v_i^0 and the feature v_{i-1}^1 of the previous branch.
    end if
end for
for i = I to 0 do
    if i == I then
        v_i^2 is generated by v_i^0 through convolution.
    else
        Generate G_i^2 by v_i^0 and v_{i+1}^0.
        v_{i+1}^2 is visually enhanced by SGE and filtered by the gate G_i^2.
        Obtain v_i^2 by fusing v_i^0 and the feature v_{i+1}^2 of the previous branch.
    end if
    Obtain v_i^f by fusing v_i^1 and v_i^2.
    Generate the predict box b_fi by v_i^f.
end for
Select the optimal prediction box b_f in {b_f0, ..., b_fI} through NMS.
```

our approach with baseline networks and other state-of-the-art context fusion algorithms. In addition, we compare CBD-E with some of the most advanced remote sensing object detection algorithms. Finally, we analyze the influence of key parameters in CBD-E on detection performance.

A. Data Sets

Two public data sets, namely, NWPU [51], RSOD [52], and DIOR [53], are used for comparison. Both of them are available online.[1][2][3]

1) NWPU: The NWPU data set is a challenging ten-class object detection data set published in conjunction with object detection tasks annotated by the Northwestern Polytechnical University. These ten classes of objects are ship, vehicle, airplane, bridge, storage tank, tennis court, harbor, basketball court, ground track field, and baseball diamond. The NWPU data set has two parts, the positive set and the negative set, which contain a total of 800 images, ranging in size from 533 × 597 to 1028 × 1728. All 3210 objects in this data set exist in the 650 images in the positive set. The 150 images in the negative set contain no objects. Of these images, 715 from Google Maps are high-resolution remote sensing images with spatial resolutions ranging from 0.5 to 2 m, and the remaining 85 images are color-infrared images with the spatial resolution of 0.08 m. The positive set and the negative set are randomly divided into three parts according to the ratio of 6:2:2. Corresponding data are combined as training set, verification set, and testing set. We directly use the original images as input to the network, instead of splitting them or resizing.

2) RSOD: RSOD is an open remote sensing object detection data set that was labeled by a Wuhan University team in 2017. The data set contains 976 images and four categories: aircraft, playground, overpass, and oil tank. These images contain a total of 6950 objects, including 1586 oil tanks in 165 images, 191 playgrounds in 189 images, 4993 aircraft in 446 images, and 180 overpasses in 176 images. The data set is randomly divided into a training set, a validation set and a testing set according to 6:2:2 ratios.

3) DIOR: The dataset is a large open source dataset recently developed and contains 23 463 images. It covers a wide range of categories. These images contain 20 kinds of objects, such as vehicle, train station, harbor, etc. The size of the image is unified to 800 × 800. The ground coverage is wide, covering more than 80 countries. Moreover, these images are carefully collected under different seasons, weathers and imaging conditions. And it has high interclass similarity and intraclass diversity. We use officially divided training set, validation set, and test set.

B. Evaluation Metrics

We use average precision (AP) to quantitatively evaluate the performance of an object detection system. This standard is widely used in many object detection studies [31], [32]. Precision measures the fraction of detections that are true positives, and recall measures the fraction of positives that are correctly identified. Assuming that TP, FP, and FN denote the numbers of true positives, false positives, and false negatives, respectively, precision and recall can be formulated as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{21}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{22}$$

If the intersection over union (IOU) between the predicted bounding box and the ground truth exceeds 0.5, the respective instance is considered to be TP. Otherwise, the prediction is treated as FP. The AP metric involves computing the average value of precision over the interval from recall = 0 to recall = 1, i.e., the area under the precision–recall curve (PRC) obtained by plotting precision and recall. AP is particularly suitable

[1]Online. [Available]: http://jiong.tea.ac.cn/people/JunweiHan/NWPUVHR10dataset.html
[2]Online. [Available]: https://github.com/RSIA-LIESMARS-WHU/RSOD-Dataset-
[3]Online. [Available]: http://www.escience.cn/people/gongcheng/DIOR.html

TABLE I
COMPUTED MAP (%) COMPARISON OF EIGHT DIFFERENT METHODS ON THE NWPU DATA SET

Class	Faster R-CNN	SSD300	RetinaNet	FPN	CA-CNN	LCFFN	GBD	CBD-E (ours)
Ship	**94.39**	83.74	90.27	92.01	88.06	87.47	85.55	92.43
Vehicle	82.02	38.43	86.89	91.19	88.47	86.39	**91.71**	89.08
Airplane	99.35	90.61	95.78	99.99	98.71	96.81	98.74	**99.99**
Bridge	74.42	98.18	97.06	86.70	86.84	**98.91**	94.98	97.08
Harbor	87.94	88.21	83.86	92.63	92.99	91.62	89.58	**93.17**
Ground track field	97.37	**99.99**	99.76	97.53	99.59	99.37	99.70	99.55
Tennis court	92.03	87.61	90.80	**96.33**	92.38	93.12	95.55	95.21
Storage tank	66.65	77.37	86.30	85.73	94.23	94.48	**95.02**	94.65
Basketball court	85.25	69.28	90.63	90.53	**98.83**	90.36	91.24	90.22
Baseball diamond	93.37	97.44	93.06	94.70	98.04	98.14	97.46	**98.39**
mAP	87.28	83.09	89.62	92.73	93.81	93.67	93.95	**94.98**

for evaluation algorithms that predict both object location and category because the criterion reflects the stability of the model. Higher AP values indicate better model performance and vice versa. For multiclass detection tasks, mAP is usually used to evaluate the model's average performance across all classes.

C. Implementation Details

The implementation of CBD-E is based on FPN. We use ResNet-101 [46] as the backbone network to extract features. We migrated the ResNet-101 pretrained on ImageNet [54] to our model and further fine-tuned it on remote sensing data sets [55], [56]. For the NWPU and RSOD data sets, predictions are made on P2, P3, P4, P5, and P6 of feature maps extracted by the FPN. The maximum number of positive samples selected by RPN is no more than 6000, and the IOU between the labels is more than 0.7 for positive samples. We set the batch size of the above methods to 1. In the second stage, where the number of positive samples does not exceed 128, the minibatch size is set to 256. For the NWPU data set, we trained for 30 000 steps, and our initial learning rate was set to 1e-3 and changed to 1e-4 and 1e-5 at 10 000 and 15 000 steps, respectively. For the RSOD data set, we trained for 45 000 steps, and our initial learning rate was set to 1e-3 and changed to 1e-4 and 1e-5 at 15 000 and 30 000 steps, respectively. For both data sets, we use the data augmentation of rotation and flip. To prove the outstanding performance of our method among algorithms of the same kind, we reproduced the context fusion methods CA-CNN and LCFFN of other authors. For fairness of this comparative experiment, we reproduced the above methods on the same baseline network FPN and used the same training parameters and data augmentation.

In this article, we compare CBD-E with other advanced object detection methods in the field of remote sensing, such as Faster R-CNN, SSD300, and RetinaNet [57]. Additionally, we use the same data set partition for all of the above methods.

D. Results and Analysis

1) Experiments on NWPU: The predictions of CBD-E on this data set are shown in Fig. 6(a). The experimental results are shown in Table I, where the best results are marked in bold. Compared with the baseline, CBD-E achieved an overall performance improvement of 2.25%, and obtained the excellent detection result of 94.98% in comparison with seven methods. According to Table I, there are several classes, including bridge, ground track field, baseball diamond, and storage tank, for which CBD-E improves over the baseline by 10.38%, 2.02%, 3.69%, and 8.92%, respectively. As shown in the PRCs in Fig. 7, GBD significantly increases the recall of multiple classes compared to the baseline. And CBD-E obtained a more perfect curve. On this data set, CA-CNN and LCFFN attain an mAP of 93.81% and 93.67%, which are excellent results. These similar algorithms and CBD-E have excellent performance. In Table I, the experimental results of Faster R-CNN, SSD300, and RetinaNet all originate from [58]. Through comparative experiments, the effect obtained by CBD-E is comparable to that of state-of-the-art remote sensing target detection algorithms. During inference, CBD-E can run at 2.0 FPS on the GPU.

In CBD-E, the improvement in detection accuracy of storage tanks and bridges is very large. Although a storage tank has a distinct outline, it has almost no texture, as shown in Fig. 8. Storage tanks are usually found only in factories. Compared with the baseline, CBD-E has a more accurate response to the factory, as shown in Fig. 8. We speculate that the introduction of CBD-E as context has made a tremendous contribution to the improvement of storage tank detection performance. It is worth mentioning that CA-CNN and LCFFN are also significantly better than Faster R-CNN, SSD300, RetinaNet, and FPN in the detection of storage tanks. We attribute the improved performance of detecting these objects to the context. Bridges are usually built over rivers. We visualized the feature maps during the detection process and found that riparian line that are not detected objects still have a high degree of activation response, as shown in Fig. 8. We speculate that the introduction of rivers as context is an important factor to improve the accuracy of bridge detection.

2) Experiments on RSOD: Due to the differences between the data sets, we analyzed the sensitivity of the parameters on RSOD. Unlike NWPU, $p = 0.2, 0.6$, and 1.4 is more suitable. The others are consistent with NPWPU. The predictions of CBD-E are shown in Fig. 6(a). The experimental results are shown in Table II, where the best results are marked. Compared to the baseline, the overall performance of the CBD-E has been greatly improved, and the improvement of some categories is significant. CBD-E has a mAP of 94.23%, which is outstanding among many methods. GBD reached 93.61%, which still has a considerable increase compared to the baseline. Table II shows

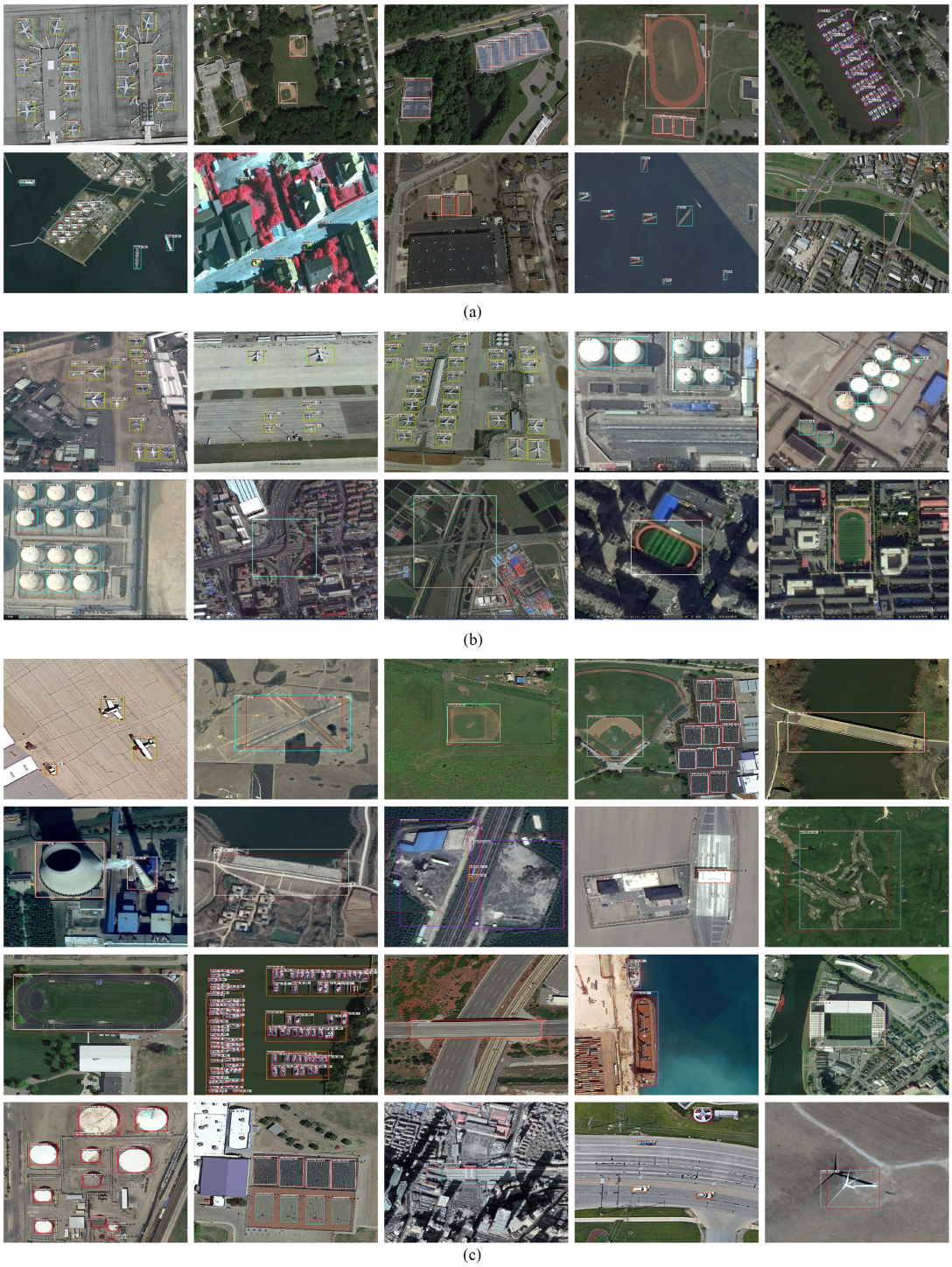


Fig. 6. Select detection results on the test sets of NWPU, RSOD, and DIOR. The red boxes represent the ground truth. (a) Predictions of CBD-E on the NWPU data set. (b) Predictions of CBD-E on the RSOD data set. (c) Predictions of CBD-E on the DIOR data set.

that for oil tanks, CBD-E attains a significant improvement by 4.48% compared with the baseline. The storage tanks in the NWPU data set and the oil tanks in RSOD are instances of the same object. The oil tank images in the RSOD data set are slightly clearer, but the texture is still not rich enough, as shown in Fig. 6. CA-CNN and LCFFN are also significantly better than Faster R-CNN, SSD300, RetinaNet, and FPN in the detection of oil tanks. The introduction of contextual information also enables better detection results for oil tanks in the RSOD data set. On RSOD, the inference speed of CBD-E can reach 3.3 FPS.

In Table II, the experiment results of Faster R-CNN, SSD300, and RetinaNet all also originate from [58]. As shown in Table II, CBD-E is observed to still have the excellent mAP among the context fusion algorithms. Compared with other object detection methods in remote sensing scene, CBD-E still has outstanding performance.

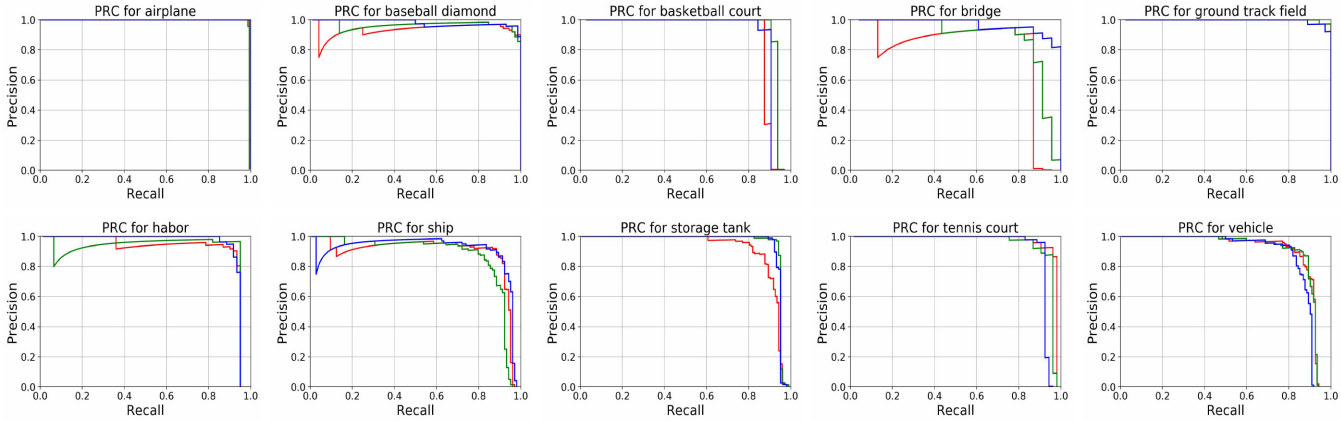


Fig. 7. PRCs of the baseline (red), GBD (green), and CBD-E (blue) for NWPU.

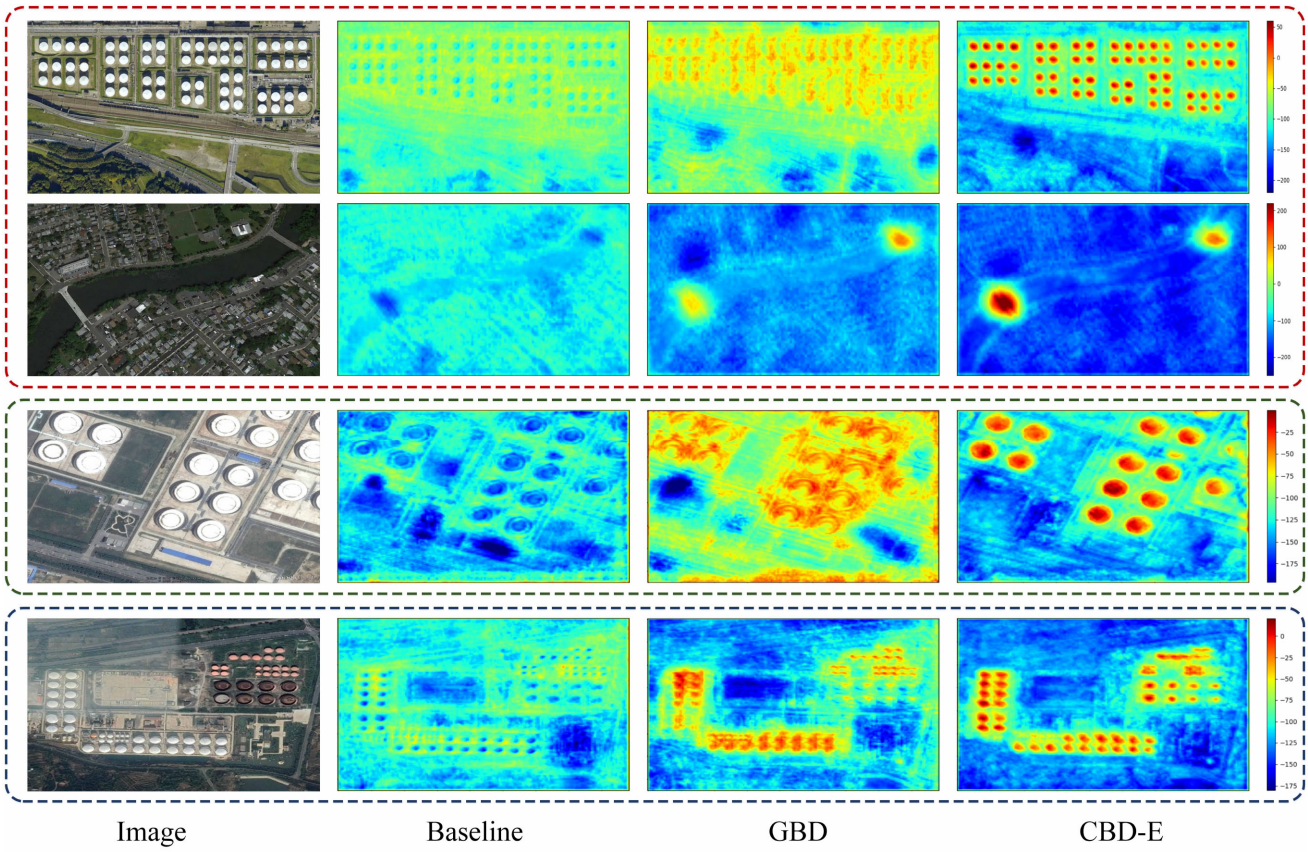


Fig. 8. Activation response during detection. The objects in the red box are from NWPU, oil tanks in the green box are from ROSD, and storage tanks in the blue box are from DIOR.

3) Experiments on DIOR: Due to the differences between the data sets, we analyzed the sensitivity of the parameters on DIOR. When the object is relatively large, we find that the expanded RoI often exceeds the range of the image. So, we set $p = 0.2, 0.6$, and 1.0. The predictions of CBD-E on this data set are shown in Fig. 6(a). The experimental results are shown in Table III, where the best results are marked in bold. Compared with the baseline, CBD-E achieved an overall performance improvement of 2.4%, and obtained the excellent detection result of 67.76% in comparison with seven methods. According to Table III, there are several classes, including storage tank, baseball field, and basketball court, for which CBD-E improves over the baseline by 10.0%, 8.2%, and 6.5%, respectively. By visualizing the features of the P2 layer in FPN, we found that the feature extraction network of CBD-E has a stronger activation effect on the storage tank than the background, as shown in Fig. 8.

TABLE II
COMPUTED MAP (%) COMPARISON OF EIGHT DIFFERENT METHODS ON THE RSOD DATA SET

Class	Faster R-CNN	SSD300	RetinaNet	FPN	CA-CNN(g)	LCFFN	GBD	CBD-E (ours)
Aircraft	83.54	71.89	80.57	95.32	95.44	95.80	95.44	**95.81**
Playground	97.81	98.58	96.97	99.50	**99.87**	99.59	99.56	99.75
Oil tank	**98.11**	90.72	96.69	87.99	91.38	89.22	91.12	92.47
Overpass	88.62	90.21	**90.25**	88.27	86.21	87.69	88.31	88.88
mAP	92.02	87.85	91.19	92.77	93.23	93.08	93.61	**94.23**

TABLE III
COMPUTED MAP (%) COMPARISON OF EIGHT DIFFERENT METHODS ON THE DIOR DATA SET

Class	Faster R-CNN	SSD300	RetinaNet	FPN	CA-CNN(g)	LCFFN	GBD	CBD-E (ours)
Airplane	53.6	59.5	53.3	54.5	54.3	54.6	54.3	54.2
Airport	49.3	72.7	77.0	76.7	76.8	77.1	76.5	77.0
Baseball field	78.8	72.4	69.3	63.3	64.4	75.8	71.2	71.5
Basketball court	66.2	75.7	85.0	80.6	80.5	87.4	85.4	87.1
Bridge	28.0	29.7	44.1	44.8	46.2	40.4	42.2	44.6
Chimney	70.9	65.8	73.2	72.5	72.1	76.7	71.6	75.4
Dam	62.3	56.6	62.4	60.0	59.5	63.8	62.1	63.5
Expressway service area	69.0	63.5	78.6	75.6	76.1	71.5	76.3	76.2
Expressway toll station	55.2	53.1	62.8	62.1	64.5	62.7	63.2	65.3
Golf course	68.0	65.3	78.6	76.0	77.3	76.4	79.6	79.3
Ground track field	56.9	68.6	76.6	77.3	79.0	78.1	77.3	79.5
Harbor	50.2	49.4	49.9	46.4	44.2	43.8	45.1	47.5
Overpass	50.1	48.1	59.6	57.3	57.2	57.0	59.3	59.3
Ship	27.7	59.2	71.1	71.8	67.5	74.5	72.5	69.1
Stadium	73.0	61.0	68.4	68.2	70.2	68.5	69.8	69.7
Storage tank	39.8	46.6	45.8	54.3	60.1	64.1	61.2	64.3
Tennis court	75.2	76.3	81.3	81.1	83.2	84.0	84.1	84.5
Train station	38.6	55.1	55.2	59.5	64.1	56.3	59.5	59.4
Vehicle	23.6	27.4	44.4	43.6	41.0	42.9	42.4	44.7
Wind mill	45.4	65.7	85.5	81.0	82.9	86.8	85.3	83.1
mAP	54.1	58.6	66.1	65.3	66.1	67.1	66.9	67.8

TABLE IV
DETECTION MAP (%) FOR FEATURES WITH DIFFERENT REGIONS' COMBINATION ON THE NWPU DATA SET

Contextual region size (p)	Single region					Multiple regions					
	-0.2	0	0.4	0.8	1.7	0+0.4	0+0.8	0+0.4+0.8	0+0.4+1.7	0+0.4+0.8+1.7	-0.2+0.4+0.8+1.7
mAP (ours GBD)	89.56	92.73	92.75	92.01	90.53	93.22	93.38	93.61	93.53	93.95	93.67
mAP (GBD-Net)	*	*	*	*	*	93.19	93.37	93.44	93.48	93.63	93.28

And the factory where the storage tank is located also has strong activation.

In Table III, the experiment results of Faster R-CNN, SSD300 and RetinaNet all also originate from [53]. Compared with other object detection methods in remote sensing scene, CBD-E still has outstanding performance. During inference, CBD-E can run at 3.6 FPS on DIOR.

E. Parameter Analysis

To investigate the influence of different sizes of context regions, we perform a series of experiments on the NWPU data set. Among them, contextual features of different region sizes are added one-by-one. The experimental results obtained with various parameters are shown in Table IV. $p = 0$ means that no background area is introduced, and this region is the original candidate box. If there is only one region feature, we observe that the larger the background region is, the worse the performance is. Additionally, the larger the gradient of the background area is, the worse the performance is. We speculate that the central area of an object tends to contribute more to instance classification. Based on this conjecture, we explored the detection performance when p takes a negative value, for example, $p = -0.2$. Unfortunately, we do not get better performance, even if it only serves as a branch.

We research the influence of parameter α mentioned in Section III using the NWPU data set. The experimental results for various values of α are shown in Table V. When studying the influence of α, we fix the value of E at 64. We find that no performance improvement is attained if α is larger or smaller than

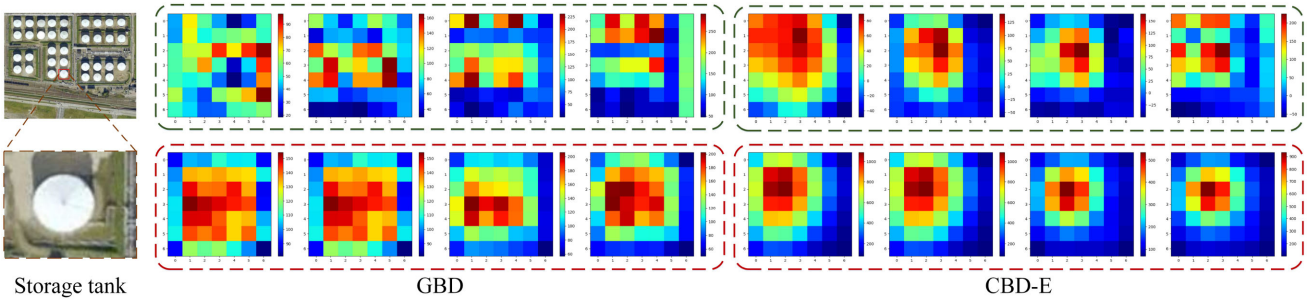


Fig. 9. Activation response of the NWPU's storage tank in the second stage. Unprocessed features are enclosed in green boxes. The features in the red box are optimized by the algorithm.

TABLE V
DETECTION MAP (%) FOR FEATURES WITH DIFFERENT VALUES OF α AND E ON THE NWPU DATA SET

α	0.05	0.1	0.2	0.5	1	*
mAP	93.23	93.95	93.74	93.21	92.43	92.03
E	4	8	16	32	64	128
mAP	93.05	93.66	93.92	94.52	94.98	93.05

TABLE VI
INFERENCE SPEED (FPS) ON DIFFERENT DATA SETS

Method	NWPU	RSOD	DIOR	FLOPs	Params
Baseline	2.6	4.1	4.6	259.8M	120.8M
GBD	2.2	3.5	3.8	291.5M	152.5M
CBD-E	2.0	3.3	3.6	291.6M	152.5M

0.1. We tried to use the fusion result v^3 directly for prediction but obtained worse performance, as shown in the last column in Table V. Using the NWPU data set, we also explore the influence of parameter g of SGE, as shown in Table V. When studying the influence of g, we fix the value of α at 0.1. We find that the performance of the network increases first and then decreases with the increase in E. According to the experimental results, we select the value of E to be 64.

F. Ablation Study

We performed ablation studies to verify the effectiveness of various parts of CBD-E. We explore the activation state of the object by obtaining heat maps synthesized by its feature maps during the detection process, as shown in Fig. 9. In the four groups of original features obtained by RoIAlign, the relative activation state between the object and its context is not optimal. The activation intensity of the object itself cannot be much higher than its background. And the activation amount of the object itself is negative correlated with the activation amount of its context. After CBD-E combing the context and enhancing the object, the four groups of features of the storage tank all have considerable changes, as shown in Fig. 9. The four groups of features of the storage tank all have more effective context information and its own activation intensity is higher. Regions that introduce less background also have rich contextual information. And the features of the object in context-rich regions have been properly improved. It is worth noting that the object itself has a fairly strong activation relative to its context. In order to further verify that this high activation phenomenon is caused by visual attention, we removed the SGE part in the CBD-E model (GBD). The contrast in activation intensity between objects and context is no longer so strong, as shown in Fig. 9.

Our method also has a considerable impact on the front-end network. By visualizing the features of the P2 layer of FPN, we found that the CBD-E feature extraction network has a stronger activation of the object relative to the background. And the activation degree and scope of the object's context are also more object-oriented, as shown in Fig. 8. The strong response contrast obviously means that the model pays more attention to the detected object. However, after removing the SGE part (GBD), the response to the object is significantly reduced, and the contrast strength between the object and the context is reduced.

On the NWPU, we did an ablation experiment on our proposed GBD structure, by comparing the original view GBD-Net. As shown in Table IV, our proposed GBD has a slight improvement in a variety of regional combinations. And the GBD is more prominent when the gap between adjacent branches is larger. We speculate that when the regional difference is large, the acceptance branch should be more involved in the control of the gate function.

G. Computational Cost

On the three data sets, we conducted an ablation study on the inference speed and computational cost of CBD-E. All models are tested on a single Nvidia GeForce GTX 1080Ti GPU with 11-GB memory. As shown in Table VI, the inference speed of CBD-E can reach 2-3 FPS. On the NWPU, we tested the computational cost and parameter amount. Compared to the baseline, the FLOPs of CBD-E increased by 10%. The parameters of CBD-E increased by 20% from the baseline, and the visual attention only increased by 1536. As the number of parameters increases, CBD-E inevitably increases the computational cost compared to the baseline. And the computational cost of visual attention is slight.

V. Conclusion

In this article, we propose the CBD-E method for contextual information fusion in remote sensing scene object detection, where gate functions and visual attention are involved in the deep network. The gate functions effectively remove the unexpected information in the context fusion process, and SGE highlights the feature of the object itself that is suppressed by the background. The former's role is to improve recall, and the latter's role is to reduce false alarm. The joint learning of the two strategies enhances the expression ability of the algorithm. By analyzing the performance of multiple methods on open remote sensing data sets, we conclude that CBD-E has excellent performance in context-dependent algorithms, and has achieved comparable effect to state-of-the-art methods.

Acknowledgment

The authors would like to thank C. Gong from Northwestern Polytechnical University and L. Yang from Wuhan University, for providing the NWPU and RSOD data sets in their study, respectively.

References

[1] 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.

[2] Z. Shao, L. Wang, Z. Wang, W. Du, and W. Wu, "Saliency-aware convolution neural network for ship detection in surveillance video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 3, pp. 781–794, Mar. 2020.

[3] J. Pang, C. Li, J. Shi, Z. Xu, and H. Feng, "$\mathcal{R}^2$ -CNN: Fast tiny object detection in large-scale remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5512–5524, Aug. 2019.

[4] Y. Zhang, Y. Yuan, Y. Feng, and X. Lu, "Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5535–5548, Aug. 2019.

[5] W. Huang, Q. Wang, and X. Li, "Feature sparsity in convolutional neural networks for scene classification of remote sensing image," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 3017–3020.

[6] Y. Wang, T. Wang, G. Zhang, Q. Cheng, and J. Wu, "Small target tracking in satellite videos using background compensation," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.2978512.

[7] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," 2019, *arXiv:1905.05055*.

[8] H. G. Akçay and S. Aksoy, "Building detection using directional spatial constraints," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2010, pp. 1932–1935.

[9] A. O. Ok, C. Senaras, and B. Yuksel, "Automated detection of arbitrarily shaped buildings in complex environments from monocular VHR optical satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 3, pp. 1701–1717, Mar. 2013.

[10] X. Zeng *et al.*, "Crafting GBD-Net for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 9, pp. 2109–2123, Sep. 2018.

[11] X. Li, X. Hu, and J. Yang, "Spatial group-wise enhance: Improving semantic feature learning in convolutional networks," 2019, *arXiv:1905.09646*.

[12] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Adv. Neural Inform. Process. Syst. 30*. Curran Associates, Inc., 2017, pp. 3856–3866.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[14] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[15] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788. [Online]. Available: http://arxiv.org/abs/1506.02640

[16] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 936–944.

[17] Y. Li, Y. Zhang, X. Huang, and A. L. Yuille, "Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 146, pp. 182–196, 2018.

[18] G. Liu, Y. Zhang, X. Zheng, X. Sun, K. Fu, and H. Wang, "A new method on inshore ship detection in high-resolution satellite images using shape and context information," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 3, pp. 617–621, Mar. 2014.

[19] S. Bo and Y. Jing, "Region-based airplane detection in remotely sensed imagery," in *Proc. 3rd Int. Congr. Image Signal Process.*, Oct. 2010, vol. 4, pp. 1923–1926.

[20] Y. Li, X. Sun, H. Wang, H. Sun, and X. Li, "Automatic target detection in high-resolution remote sensing images using a contour-based spatial model," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 5, pp. 886–890, Sep. 2012.

[21] P. Zhong and R. Wang, "Object detection based on combination of conditional random field and Markov random field," in *Proc. 18th Int. Conf. Pattern Recognit.*, Aug. 2006, vol. 3, pp. 160–163.

[22] X. Sun, K. Fu, H. Long, Y. Hu, L. Cai, and H. Wang, "Contextual models for automatic building extraction in high resolution remote sensing image using object-based boosting method," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2008, vol. 2, pp. II–437–II–440.

[23] A. O. Ok, "Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts," *ISPRS J. Photogramm. Remote Sens.*, vol. 86, pp. 21–40, 2013.

[24] O. Firat, G. Can, and F. T. Y. Vural, "Representation learning for contextual object and region detection in remote sensing," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 3708–3713.

[25] W. Zhao, S. Du, Q. Wang, and W. J. Emery, "Contextually guided very-high-resolution imagery classification with semantic segments," *ISPRS J. Photogramm. Remote Sens.*, vol. 132, pp. 48–60, 2017.

[26] X. Yao, J. Han, L. Guo, S. Bu, and Z. Liu, "A coarse-to-fine model for airport detection from remote sensing images using target-oriented visual saliency and CRF," *Neurocomputing*, vol. 164, pp. 162–172, 2015.

[27] G. Zhang, S. Lu, and W. Zhang, "CAD-Net: A context-aware detection network for objects in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 1–10, Dec. 2019.

[28] X. Yang, H. Sun, X. Sun, M. Yan, Z. Guo, and K. Fu, "Position detection and direction prediction for arbitrary-oriented ships via multitask rotation region convolutional neural network," *IEEE Access*, vol. 6, pp. 50 839–50 849, 2018.

[29] M. Kang, K. Ji, X. Leng, and Z. Lin, "Contextual region-based convolutional neural network with multilayer fusion for SAR ship detection," *Remote Sens.*, vol. 9, no. 8, 2017, Art. no. 860.

[30] Y. Ren, C. Zhu, and S. Xiao, "Small object detection in optical remote sensing images via modified Faster R-CNN," *Appl. Sci.*, vol. 8, no. 5, 2018, Art. no. 813.

[31] Y. Gong *et al.*, "Context-aware convolutional neural network for object detection in VHR remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 34–44, Jan. 2020.

[32] 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.

[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[34] Z. Shen *et al.*, "Improving object detection from scratch via gated feature reuse," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2019.

[35] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2015.

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

[37] X. Ying *et al.*, "Multi-attention object detection model in remote sensing images based on multi-scale," *IEEE Access*, vol. 7, pp. 94 508–94 519, 2019.

[38] X. Yang *et al.*, "SCRDet: Towards more robust detection for small, cluttered and rotated objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 8231–8240.

[39] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

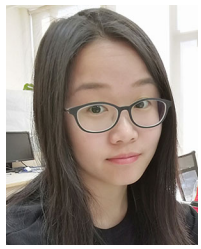
[40] P. Ding, Y. Zhang, W.-J. Deng, P. Jia, and A. Kuijper, "A light and faster regional convolutional neural network for object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 141, pp. 208–218, 2018.
[41] Z. Deng, H. Sun, S. Zhou, J. Zhao, L. Lei, and H. Zou, "Multi-scale object detection in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 3–22, 2018.
[42] B. Pan, X. Xu, Z. Shi, N. Zhang, H. Luo, and X. Lan, "DSSNet: A simple dilated semantic segmentation network for hyperspectral imagery classification," *IEEE Geosci. Remote Sens. Lett.*, to be published, doi: 10.1109/LGRS.2019.2960528.
[43] J. Zhang, J. Liu, B. Pan, and Z. Shi, "Domain adaptation based on correlation subspace dynamic distribution alignment for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.2985072.
[44] Z. Dong, M. Wang, Y. Wang, Y. Zhu, and Z. Zhang, "Object detection in high resolution remote sensing imagery based on convolutional neural networks with suitable object scale features," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2104–2114, Mar. 2020.
[45] Y. Zhang, Y. Yuan, Y. Feng, and X. Lu, "Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5535–5548, Aug. 2019.
[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[47] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[48] J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and J. Li, "Visual attention-driven hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 8065–8080, Oct. 2019.
[49] J. M. Haut, R. Fernandez-Beltran, M. E. Paoletti, J. Plaza, and A. Plaza, "Remote sensing image superresolution using deep residual channel attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9277–9289, Nov. 2019.
[50] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019.
[51] 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, 2014.
[52] 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.
[53] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020. [Online]. Available: http://dx.doi.org/10.1016/j.isprsjprs.2019.11.023
[54] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[55] Y. Wang *et al.*, "Supervised high-level feature learning with label consistencies for object recognition," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2019.2955557.
[56] X. Feng, J. Han, X. Yao, and G. Cheng, "Progressive contextual instance refinement for weakly supervised object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.2985989.
[57] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2980–2988.
[58] X. Ying *et al.*, "Multi-attention object detection model in remote sensing images based on multi-scale," *IEEE Access*, vol. 7, pp. 94 508–94 519, 2019.

Jun Zhang received the B.S. and Ph.D. degrees from the Hebei University of Technology (HEBUT), Tianjin, China, in 1999 and 2011, respectively.

He is currently an Associate Professor with the School of Artificial Intelligence, HEBUT. His research interests include machine learning and intelligent computing.

Changming Xie received the B.S. degree in electronic information science and technology from Hebei Normal University, China, in 2018. He is currently working toward the M.S. degree with the School of Artificial Intelligence, Hebei University of Technology, Tianjin, China.

His research interests include machine learning and intelligent computing.



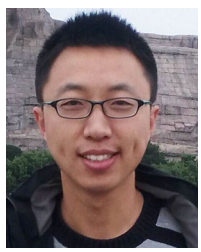
Xia Xu received the B.S. and M.S. degrees from the School of Electrical Engineering, Yanshan University, Qinhuangdao, China, in 2012 and 2015, respectively, and the Ph.D. degree from the School of Astronautics, Beihang University, Beijing, China, in 2019.

She is currently working as an Assistant Professor with the College of Computer Science, Nankai University, Tianjin, China. Her research interests include hyperspectral unmixing, multiobjective optimization, and remote sensing image processing.

Zhenwei Shi (Member, IEEE) received the Ph.D. degree in mathematics from the Dalian University of Technology, Dalian, China, in 2005.

From 2005 to 2007, he was a Postdoctoral Researcher with the Department of Automation, Tsinghua University, Beijing, China. From 2013 to 2014, he was a Visiting Scholar with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA. He is currently a Professor and the Dean of the Image Processing Center, School of Astronautics, Beihang University, Beijing, China. He has authored or coauthored more than 100 scientific papers in related journals and proceedings, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON NEURAL NETWORKS, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON IMAGE PROCESSING, and IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION. His research interests include remote sensing image processing and analysis, computer vision, pattern recognition, and machine learning.

Dr. Shi was the recipient of the Best Reviewer Awards for his service to IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, in 2017. He has been an Associate Editor for the Infrared Physics and Technology since 2016.



Bin Pan (Member, IEEE) received the B.S. and Ph.D. degrees from the School of Astronautics, Beihang University, Beijing, China, in 2013 and 2019, respectively.

Since 2019, he has been an Associate Professor with the School of Statistics and Data Science, Nankai University, Tianjin, China. His research interests include machine learning, remote sensing image processing, and multiobjective optimization.