Computing and Informatics, Vol. 42, 2023, 1404–1427, doi: 10.31577/cai_2023_6_1404

REMOTE SENSING TARGET DETECTION INSPIRED BY SCENE INFORMATION AND INTER-OBJECT RELATIONS

Yi Ding

College of Design and Engineering
National University of Singapore
4 Engineering Drive 3, Block E4, #05-42, Singapore 117583
e-mail: e0973777@u.nus.edu

Xiangru Lyu

School of Computer Science and Electronic Engineering
University of Surrey
Guildford GU2 7XH, UK
e-mail: xl01241@surrey.ac.uk

Liuyang Yan, Lan Rong, Keyang Cheng[*]

School of Computer Science and Communication Engineering
Jiangsu University
Zhenjiang, 212013, China
e-mail: liuyangyan@stmail.ujs.edu.cn, 1748675160@qq.com,
kycheng@ujs.edu.cn

Abstract. Remote sensing target detection has been widely used in industries. In various application scenarios, complicated contexts may inhibit target identification and reduce detection accuracy, especially in multi-target detection tasks. In this paper, a new remote sensing target detection method based on structural reasoning is proposed to improve target detection performance by integrating inter-object relationships and scene information. Based on inter-object information, a relation structure graph is designed to reduce errors and missed targets. To establish

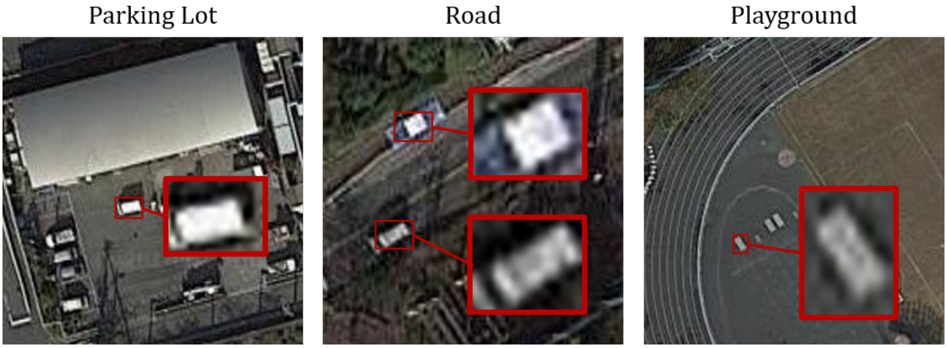
[*] Corresponding author

contextual constraints, semantic is used as a prior information for Bayesian criterion based on scene information. Experiments conducted on HRRSD dataset show that the average accuracy of the proposed method is 10.7 % higher than the state-of-the-art algorithms. The experimental results confirm that the proposed algorithm can achieve significant improvements and adapt to complex scenes in remote sensing by mining contextual information at both feature and semantic levels.

Keywords: Remote sensing, inter-object relation, scene information

1 INTRODUCTION

Based on the technological progress of data acquisition, remote sensing image processing has been adopted for widespread use across industries, including high-resolution remote sensing imagery [1], tropical cyclone detection [2], disaster control [3, 4, 5], road detection [6, 7], and municipal construction [8]. However, various scenes and complex context information in remote sensing images often pose great challenges to target detection, especially in multi-target detection. In complex environments, targets with small sizes or vague shapes perhaps lose their distinctiveness from the complicated background. In addition, various factors contribute to the variability of remote sensing images, such as time, climate, shadow, occlusion, and illumination [9]. Figure 1 shows some representative detection problems in remote sensing images. The objects, including a bus stop, a van, and several playground landmarks, are similar in shape and color. It is not easy to distinguish without the help of context information. All these problems cause a reduction in the accuracy and confidence of detection. Therefore, it has become a challenge to detect targets with features and scales in remote sensing images of various scenes and different contexts.



Parking Lot

Road

Playground

Figure 1. Representative detection problems in remote sensing images

In order to solve the aforementioned problems and to model the scenes and backgrounds, namely context constraints, researchers have made considerable efforts to advance the detection performance in complex contexts of remote sensing images. Among the traditional target detection methods, some context knowledge-based algorithms are proposed in context-aware detectors, which is an offshoot of knowledge-based object detection [10, 11, 12, 13]. The most frequently used context knowledge lies in how targets in images interact with adjacent regions, i.e., object-background relationships. For instance, the appearance of the shadow is supposed to be an essential clue for the existence of nearby buildings [12]. Based on prior knowledge, the researchers have tried to encode relationships between manufactured architectures and their shadows into handcrafted rules for building detection. In traditional remote sensing target detection studies, the knowledge-based algorithms have shown the great potential of transforming implicit knowledge into explicit detection rules. However, two problems have gradually emerged as knowledge-based methods develop. On one hand, to encode a sophisticated and fuzzy understanding of seemingly endless targets is hard. On the other hand, it is usually challenging to manually determine the optimal extent of detection rules: false positive cases will increase when controls need to be strict enough, whereas targets will be missed under insufficiently strict rules [9].

Compared to traditional methods, deep learning has offered a rich diversity of remote sensing target detection methodologies and shown significant advantages in recent years [14, 15, 16, 17, 18]. Learned features have quickly overtaken manual ones. With the help of a multi-layer structure, the convolutional neural network (CNN) can extract more sophisticated and effective features on both low and high levels than handcrafted features in traditional target detection. Many different methods based on CNN are proposed and applied to remote sensing object detection, such as dilated CNN [17] and Region CNN [18]. However, despite various deep-learning-based algorithms, only some consider contextual information. Since conventional knowledge-based algorithms have proven the effectiveness of context constraints in remote image target detection, ignoring the information may reduce the accuracy [19]. Furthermore, many algorithms are developed from early neural networks for natural object detection tasks. However, different from object detection in natural images, targets in remote sensing images are always accompanied by detailed scenes and blurred backgrounds such as various terrains, landscapes, and climates which indicates higher importance of context information.

Motivated by context constraints in knowledge-based algorithms, we propose a new deep learning network model for remote sensing target detection. Compared to limited handcrafted rules in traditional methods, the proposed algorithm using learnable context constraints can adapt to complicated situations in authentic remote-sensing images. More precisely, the model is based on two genres of contextual information, i.e., scene information and inter-object relationship. Thus, the model can use the relationships of the target with the background and with other targets to determine the final target. The main contribution of this work can be summarized as follows:

An object context-constrained module is proposed. This module is used to model inter-object relationships. It contains a Region Proposal Network (RPN) and an Object Relationship Structure Graph (ORSG). After receiving the information on region proposals from the RPN, the structure graph can aggregate the information and model the relationships between targets, including relative positions and class flag pairs. The module can learn the adaptive relationships between specified classes and arbitrary targets by building context constraints.

A scene context-constrained module is engaged. Scene-target relations are modeled in this module. In our proposed model, the scene context-constrained module uses a multi-layer perception (MLP) to predict scene classes of remote sensing images. Based on the predicted scene classes, a Bayes criterion is used to determine the appearance probability of detected targets.

This paper is structured as follows. The related research is summarized and briefly reviewed in Section 2. In Section 3, we elaborate on our algorithm in detail. Section 4 introduces the experimental design with the employed datasets. Finally, we give the conclusion of our work in Section 5.

2 RELATED WORK

2.1 Object Detection

Generally, traditional methods in remote sensing target detection can be classified as three groups, template matching-based, object-based image analysis-based (OBIA) and knowledge-based algorithm [9]. The template matching-based methods first generate templates for different target classes by hand-crafting or learning algorithms. The template is used to match a source image at any possible position and measure the similarity of each match [20]. However, template-based methods need more robustness for targets varying in size or direction; thus, many pieces of early research are concentrated on this issue, such as deformable template matching [21, 22]. In the knowledge-based method, prior knowledge is translated into detection rules. The authors in [23] proposed a geometric model detecting buildings with a shape like “E”, “F” or “T”, which is the handcrafted features extracted by the model. In another work [12], buildings can be assumed to have a rectangular shape in remote sensing images, and this rule can be considered a generic model for detecting buildings. Concerning OBIA methods, also called target-based methods, the input image is first segmented into homogeneous pixel groups by selected criteria such as scale, shape, or compactness, and then these groups are classified into different target classes. However, segmentation algorithms and the images’ complexity can greatly influence the performance of the OBIA method [24].

Apart from the traditional methods, mainstream target detection algorithms are mostly based on deep learning networks, which can be grouped into two genres: two-stage and one-stage. In the architecture of two-stage networks, a sub-network called region proposal network (RPN) is independent of the feature extraction part, including Region-based CNN (R-CNN) family models [25, 26] and Feature Pyramid Networks (FPN) [27]. In one-staged-based algorithms, an end-to-end model is designed to determine the bounding box of predicted targets and calculate their class probability simultaneously, such as single-shot multi-box detector (SSD) [28], you-only-look-once (YOLO) family algorithms [29, 30], etc. Law and Deng proposed a new approach based on a single convolution neural network, where the object detection is considered as a task of predicting a pair of keypoints (top-left corner and bottom-right corner) [31]. Inspired by the keypoint-based approach, CenterNet is proposed by Duan et al., which integrates the information from the center part of the region proposal to enhance the predictions [32]. MultiTask-CenterNet (MCN) [33] is based on the CenterNet and accomplishes multi-tasks like object detection, like semantic segmentation and depth estimation. Besides the convolution-based neural network, FLDS used a multistage residual hybrid attention module to learn robust and powerful features for target detection [34]. In comparison, one-stage methods have advantages in location and classification accuracy, whereas two-stage methods are superior in computational speed.

2.2 Contextual Information

Naturally, context information is beneficial to target detection in a complicated environment. For instance, the probability of a bridge appearing on the river is much higher than on the sports field. In early studies, many knowledge-based researchers have attempted to find a theoretical explanation of how targets interact with pixels in their neighboring regions [9, 23]. Much early empirical research proves that it is helpful to utilize contextual information to enhance object detection performance [13]. Taking building recognition as a representative example, many pieces of research suppose the shadow as a shred of evidence for the existence of manufactured buildings [11, 12]. Bückner et al. proposes a semantic net for building detection, which contains top-down operators for image segmentation and bottom-up operators for target labeling [10].

In recent years, many deep-learning-based research have proposed some novel neural networks that integrate contextual information into object detection, mainly based on convolutional neural networks. Li et al. propose a two-stage convolutional neural network to integrate global and local contextual information, containing a context network that detects objects [35]. Similarly, Li et al. use a fusion network to combine multi-angle and multi-scale characteristics of targets in remote sensing images [36]. However, scene-level features are ignored in the proposed network. Zhang et al. designed a novel CNN called CAD-Net with a spatial-and-scale-aware attention module that can capture relations between global scenes and local objects [37]. Based on CAD-Net, Gong et al. propose a novel context-aware CNN which includes multi-layer feature maps of context-regions-of-interests (context ROIs) [38]. Sun et al. dedicated a context refinement module that can aggregate multi-scale feature maps to extract context information on different levels [39].

Cheng et al. proposed a one-stage network containing two sub-networks to handle object and scene contexts [15].

However, there are several shortcomings of these methods:

1. These methods usually focus only on the internal features of the targets and ignore the external semantic features around the targets, especially the relationships between the targets.

2. These methods still only target visual features when utilizing scene information, lacking the guidance of high-level semantic knowledge.

3. These methods do not utilize both target relationship information and scene context information, but simply utilize one-sided information.

Therefore, a remote sensing image target detection method based on target relationship information and scene semantic information is designed.

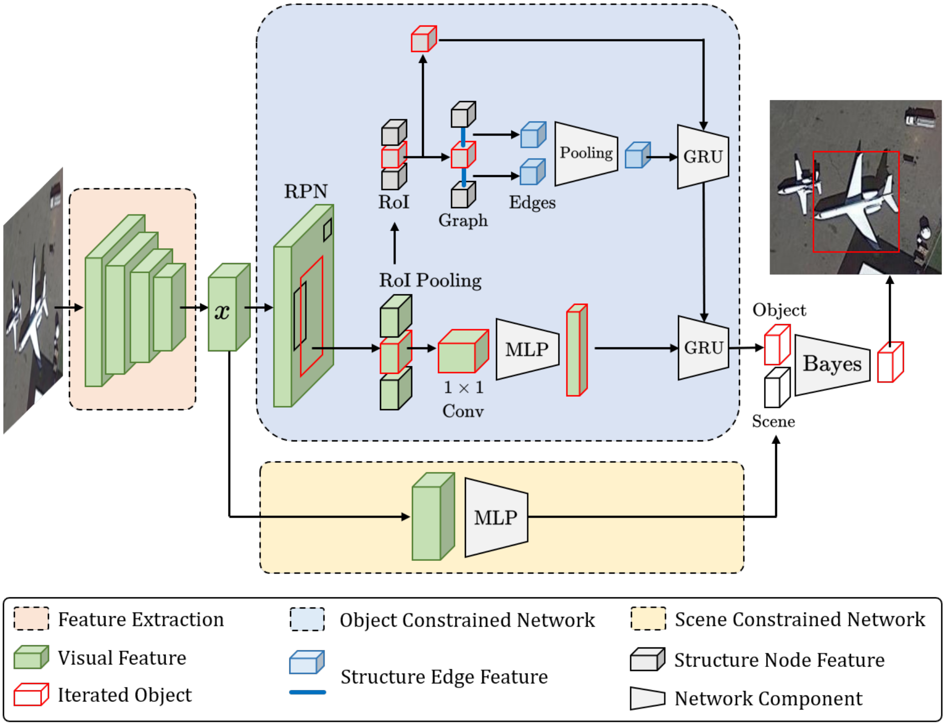


Figure 2. Overall network structure

3 PROPOSED APPROACH

Figure 2 depicts the framework of our proposed detection network. The network contains three main parts, including feature extraction, object-constrained network, and scene constrained network. The CNN-based encoder extracts the features of the input image and feeds the features into two sub-network. Based on the target information proposed by the RPN, the object-constrained network uses the gated recurrent unit (GRU) to build inter-object context information. The scene context-constrained sub-network uses a multi-layer perceptron (MLP) to classify scenes and then uses Bayesian criterion to describe scene-object relationships. The detection network benefits from two sub-networks that can integrate the local object context and the global scene context to infer inter-object and scene-object relationships. The framework will be detailed in the following sections.

3.1 Inter-Object Context Constraints

In this section, we delineate the object-constrained module, which describes the relationships between objects and builds the inter-object context constraints on the detection output.

3.1.1 Inter-Object Relationship Structure Graph

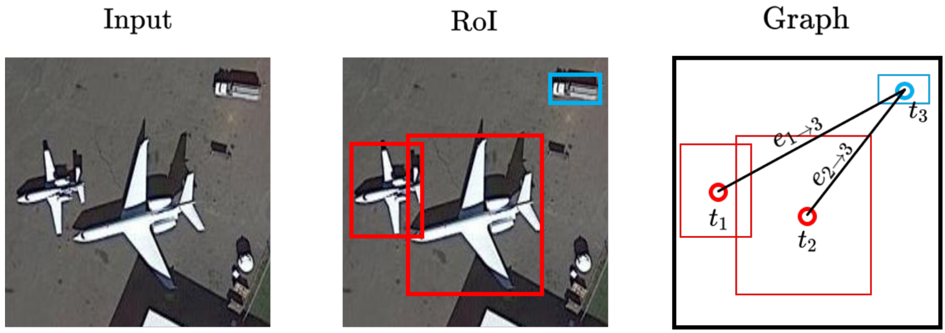


Figure 3. Inter-object relationship structure graph in remote sensing images

A wealth of context information on the inter-object level makes a difference to object detection. For instance, the airplane might not appear on a basketball playground but will likely coexist with aircraft towing tractors. In the proposed network, we use a graphical model called an inter-object relationship structure graph to describe the relations among targets.

Figure 3 shows a typical circumstance in a remote sensing image, which depicts constructing an inter-object relationship. It is natural to model the inter-object relationships as a directed graph $G = (T, E)$, where nodes $t \in T$ denote the possible targets proposed by an arbitrary detection network since the graphical model is independent of the target detection network. In this work, we build the structure graph based on the regions of interest (RoIs) proposed by an RPN network [26]. The RPN will propose k RoIs, where the number k is determined by the RPN network. Each RoI corresponds to a node v_i in the structure graph. The RoI pooling layer will extract the visual features of corresponding nodes, denoted as f_i^v.

In addition to visual features, position information is also used in the network. The edge $e \in E$ denotes the relations between the pair of target nodes. Specifically, $e_{j\to i}$ represents the edge from node v_j to v_i, which indicates the influence of the node v_j on v_i. The structure graph integrates the information of visual features and the relative position of RoIs to model the relation. To describe the information of the relative position, we use a vector $R^p_{j\to i}$.

$$R_{j\to i} = \left[w_i, h_i, s_i, w_j, h_j, s_j, d_{1x}, d_{1y}, d_{2x}, d_{2y}, \log\left(\frac{w_i}{w_j}\right), \log\left(\frac{h_i}{h_j}\right)\right],$$

where

$$d_{1x} = \frac{(x_i - x_j)}{w_j}, \quad d_{1y} = \frac{(y_i - y_j)}{h_j},$$

$$d_{2x} = \frac{(x_i - x_j)^2}{w_j^2}, \quad d_{2y} = \frac{(y_i - y_j)^2}{h_j^2}.$$

(x_i, y_i) is the center coordinate of RoI. w_i, h_i are the width and height of the region. s_i represents the area of the region. For node v_i, the position and visual features are integrated into the edge vector e. Therefore, the edge $e_{j\to i}$ is calculated by

$$e_{j\to i} = \mathrm{Relu}\left(W_p R^p_{j\to i}\right) \cdot \tanh\left(W_v \left[f_i^v, f_j^v\right]\right),$$

where W_p and W_v are learnable weight matrices of the visual feature and position, respectively. $[\cdot]$ denotes the concatenation of vectors.

3.1.2 Gated Recurrent Unit

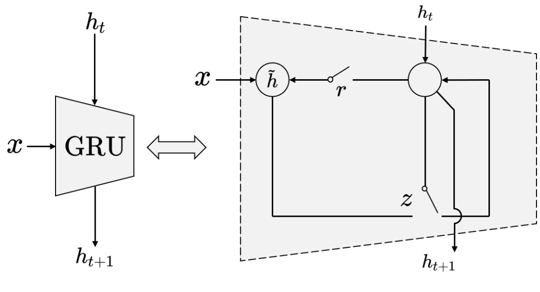


Figure 4. Gated recurrent unit

For nodes in the graph, we hope the network can manipulate the information from a particular node and also memorize the manipulation result of other nodes. Therefore, the network can learn the interactions between different notes. The mechanism here is similar to a neural memory network, such as Recurrent Neural Network (RNN). The Gated Recurrent Unit (GRU) [40], a lightweight practical RNN module, is used in this paper. The GRU is used to aggregate the information passed from the structure graph. Figure 4 depicts the details of the GRU.

The *reset* gate r controls whether the GRU ignores the previous hidden state h_t. Similarly, the *zero* gate z controls whether a new hidden state $\tilde{h}$ is used to update the next hidden state h_{t+1}. r and z are both jointly determined by hidden state h_t and state input x. They are computed by

$$r = \sigma \left(W_r \left[x, h_t \right] \right),$$

$$z = \sigma \left(W_z \left[x, h_t \right] \right),$$

where $[\cdot]$ represents the vector connection. W_r and W_z are the learnable weight matrix for the *zero* and *reset* gates, respectively. σ is the logistic sigmoid function, which is the activation function of the gates. r decides the use of a new hidden state $\tilde{h}$ which is computed by

$$\tilde{h} = \tanh \left(Wx + U \left(r \odot h_t \right) \right),$$

where W and U are learnable weights for the hidden state $\tilde{h}$. $\odot$ represents the element-wise multiplication. The weights of the hidden state imitate the selective memory. Thus, the hidden state can determine whether upcoming information from new nodes is relevant and whether to ignore useless information.

Finally, the output unit h_{t+1} is computed by h_t and $\tilde{h}$,

$$h_{t+1} = zh_t + (1 - z)\tilde{h}.$$

In summary, the GRU has a lightweight practical memory cell to remember long-term information in sequence. When the network sequentially processes the proposed target, the GRU can aggregate the context information from different target nodes in the structure graph. The next section will elaborate how the node information from the graph is fed into the GRU.

3.1.3 Object Constraints

Figure 5 shows the complete inference process, which takes the inference on the third target t_3 as an instance. After the structure graph is built, the edge information is integrated with the visual feature as the final message passed to the GRU, which is denoted as $m_{i \to j}$, calculated by

$$m_{i \to j} = e_{i \to j} \cdot f_i^v.$$

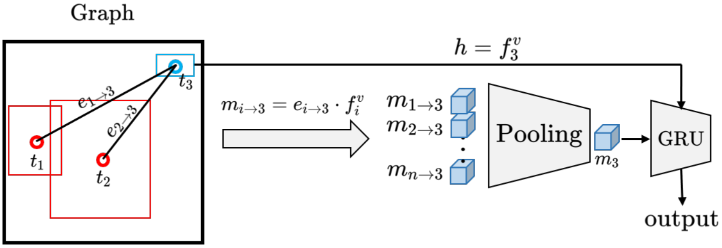


Figure 5. The process of structure inference

We use a pooling layer to consolidate all the relation information. Empirically, it can be found that the maximum pooling can extract the most important message in the structure graph, while using average pooling can be interfered with the ROI of a large number of unrelated regions. Therefore, the final message passed to the GRU is m_i, calculated by

$$m_i = \max_{j\in T} \text{pooling}\left(e_{i\to j} \cdot f_i^v\right)$$

At the beginning, the structure randomly decides a node as the starting node, such as t_3 in Figure 5. The GRU takes the visual feature f_3 of the iterated RoI as the initial hidden state and the message m_3 as the input. In the following iterations, the GRU will take the message from a new inter-object relation as new input vectors and compute the next hidden states. After iterating all the remaining nodes, the final integrated node embeddings will be used to predict the bounding boxes and object classes.

3.2 Context Constraint

Naturally, the presence of a target will indicate a scene. For example, vehicles are often seen on roads rather than on the playground. Therefore, utilizing scene semantic information will significantly improve the accuracy and credibility of the target detection network. But unlike natural images, the scenes of remote sensing images are often too complex to play an active role in target detection. So it is challenging to build an effective semantic information model for remote sensing image target detection tasks. This paper presents a scenario context constraint model based on the Bayes criterion. The network is built by the Bayes' theorem,

$$p(t|s) = \frac{p(s|t)p(t)}{p(s)},$$

where $p(t)$ and $p(s)$ represent the probability of the target category and the probability of the scene category, respectively. $p(t|s)$ represents the probability that the scene of the target i appears in the scene j. $p(t|s)$ is obtained by a probability matrix that computes the correlation relationship between the target and the scene, as shown in Figure 6.

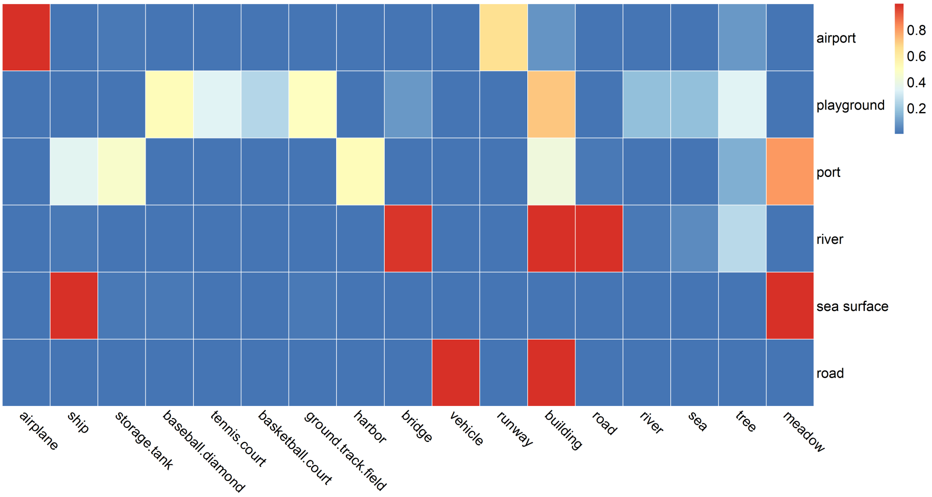


Figure 6. The probability matrix between the target and the scene

The probability i of target category $p_t(i)$ is obtained at layer Softmax in the following way:

$$p_t(i) = \frac{\exp(f_i)}{\sum_{k=1}^{M} \exp(f_k)}.$$

f_i is the feature map of the target, and M is the total number of target categories. The target probability matrix $p_t = \{p_t(1), p_t(2), \dots, p_t(M)\}$ is constructed by probability $p_t(i)$. Scenes are classified by CaffeNet [41], an image classifier with excellent performance. In layer Softmax, the probability distribution of scenario categories can be constructed by

$$p_s(j) = \frac{\exp(f_j)}{\sum_{k=1}^{N} \exp(f_k)},$$

where f_j represents the feature graph of the j^{th} scene, N denotes the total number of scene categories, and the scene probability matrix p_s is composed of probability $p_s(j)$.

3.3 Loss Function

The framework proposed in this paper is shown in Figure 2. The model mainly consists of target context constrained network and scene context constrained network. The target context constrained network has two output layers including discrete probability distribution and bounding box regression migration. The loss function of the target context constrained network is defined as

$$l_t = l_{\text{cls}}\left(p_t, p_t^*\right) + \lambda l_{\text{bbox}}\left(b, b^*\right).$$

The loss contains a classification loss and a bounding box loss, where the classification loss is calculated by

$$l_{\text{cls}}(p_t, p_t^*) = -\log\left[p_t p_t^* + (1-p_t)(1-p_t^*)\right].$$

The loss for the predicted bounding boxes is

$$l_{\text{bbox}} = \text{smoothL}_1\left(b - b^*\right),$$

where

$$\text{smoothL}_1(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1, \\ |x| - 0.5, & \text{otherwise.} \end{cases}$$

b and b^* represent the predicted boundary box and the real boundary box, respectively, p_t^* represents the real probability of the target. The smoothL$_1$ loss is used as the loss function of the prediction frame, which is used to make the loss l_{bbox} more robust to outliers and prevent the gradient explosion. The scene context constraint network has an output layer that gives the probability of classification. Assume that the true class probability of the scenario is $p_s^*(i)$. The loss function of the scene context constrained network is defined as

$$l_{\text{s}} = -\sum_{i=1}^{N} p_s^*(i) \log p_s(i).$$

The target context constraint network and the scene context constraint network are trained separately. When the classification of the target scene is inaccurate, the information provided for the Bayesian classifier is not conducive to the convergence. Thus, we train the Scene Constrained Network first. After the Scene classifier is optimized, we jointly train it with the Object Constrained Network.

4 EXPERIMENT RESULTS AND ANALYSIS

4.1 Datasets

NWPU VHR-10 [42] and HRRSD [43] datasets are used to verify the performance of remote sensing image target detection based on target relationship and scene information. These two datasets are widely used in the field of remote sensing in Google Earth and Baidu Map. The NWPU VHR-10 has 800 images and a total of 10 target categories. HRRSD has 21761 images in 13 target categories, including boats, bridges, track fields, oil storage tanks, basketball courts, tennis courts, airplanes, baseball fields, ports, cars, intersections, T-junctions, and parking lots. In this dataset, the image background is complex and diverse, which can make a reasonable verification of our proposed method.

4.2 Experimental Environment and Parameters

The experimental environment is 24 GB NVIDIA TATIAN GPUs, the Python framework is PyTorch, and the operating system is Ubuntu 14.04. VGG-16 is used as the backbone network in the experiment. First, we train the target context constraint network, including the target detection network, and then train the scene context constraint network, including scene classification. Both target and scenario context constraint networks are optimized end-to-end and brought together by Bayesian criteria. During training, the learning rate of the network is 0.001, and the weight is 0.0005. Average accuracy (AP) and average AP (mAP) are used as evaluation indexes with the formulas:

$$\mathrm{AP} = \sum_{k=1}^{N} P(k)\Delta r(k),$$

$$\mathrm{mAP} = \frac{\sum_{k=1}^{N} \mathrm{AP}_i}{K},$$

where $P(k)$ is the precision-recall curve, N is the total amount of data, and k is the index of each sample point. AP is a comprehensive evaluation standard for accuracy and recall, ranging from 0 to 1. mAP is the average of all kinds of APs.

4.3 Experiment Results and Analysis

Ablation experiments are performed on NWPU VHR-10 and HRRSD datasets. Five mainstream methods (SSD [28], FRCNN [26], YOLOv4 [29], DCIFF [44], and SCCM-BR [15]) are introduced for comparative experiments to verify the effectiveness of the proposed method.

4.3.1 Ablation Experiments

Figure 7 and Figure 8 show the ablation experiments under the data sets NWPU VHR-10 and HRRSD, respectively. Legend a indicates the situation where the target relationship diagram is not included in the scene context constraint network. Legend b represents that only the target relationship diagrams are used. Legend c means that only the scene context constraint network is used. Finally, the legend d indicates our complete neural network. The original FRCNN (Faster R-CNN, FRCNN) is first used for detection, and it can be seen that the overall target detection result was not ideal. When adding the target-relational structure diagram and training the end-to-end target context constraint network, the detection result is greatly

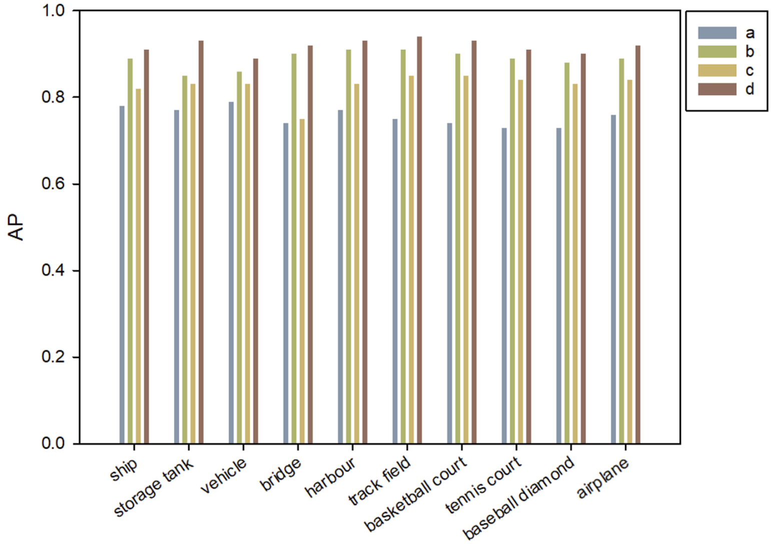


Figure 7. Ablation experiments with NWPU VHR-10 dataset

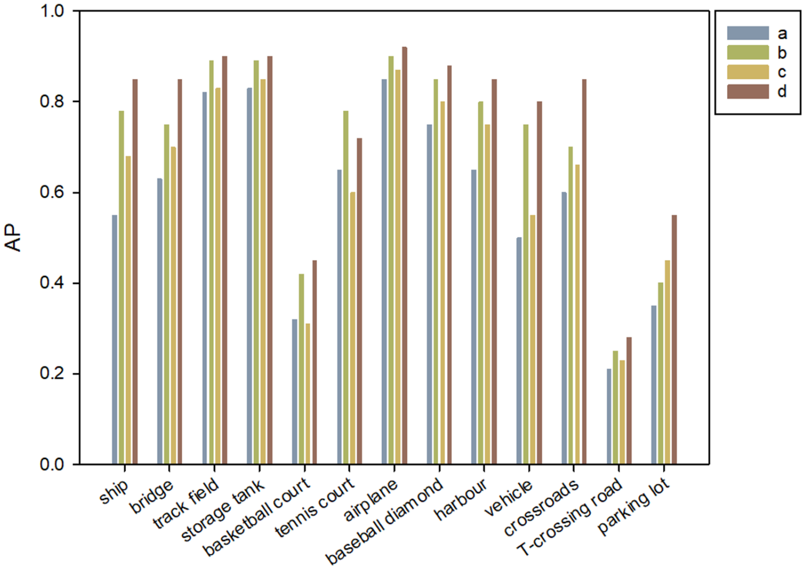


Figure 8. Ablation experiments with HRRSD dataset

improved, especially for the track-and-field and basketball court, where the target context is closely related. The results demonstrate the effectiveness of the target-relational structure diagram. When the scene context constraint network is added, the detection effect is improved to some extent, but the effect is not as apparent as the target context constraint network. However, the detection effect is improved for the target in specific scenes, such as ports, bridges, etc. The results show that both the target context constraint network and the scene context constraint network have made extraordinary contributions.

Scene	Airport	Basketball field	Parking lot	Playground
Accuracy	92.34 %	95.91 %	85.61 %	97.01 %
Scene	Harbour	River	Road	Sports field
Accuracy	92.93 %	93.76 %	97.94 %	95.82 %
Scene	Storage tank	Urban street	Sea	
Accuracy	95.44 %	87.10 %	90.43 %	

Table 1. Scene classification accuracy

In this paper, 21 761 images in HRRSD dataset are divided into 11 categories of scenarios, with approximately 50 % of the images used for training and tuning the model and the other 50 % for testing the model. The result of scene classification is shown in Table 1. It can be seen that the accuracy of most scene classifications can reach 90 %. It is worth noting that parking lots are similar to city streets, so their accuracy is reduced.

4.3.2 Comparative Experiment

In this experiment, five state-of-the-art methods (SSD, FRCNN, YOLOv4, DCIFF, and SCCM-BR) are selected to compare with the method in this paper. Table 2 and Table 3 respectively show the detection results of different detection methods in NWPU VHR-10 data set and HRRSD data set, including AP value and average mAP value of detection accuracy of different categories. Table 2 shows the detection results of these methods on the NWPU VHR-10 dataset. As seen from Table 2, our method obtains the highest average detection accuracy mAP among all methods and the highest AP value among all targets. For objects with explicit scenes, such as ships, ports, bridges, and vehicles, the AP value increases significantly because the scene context constraint network can constrain corresponding targets according to different scenarios. For baseball court, tennis courts, basketball courts, and track and field with similar scenes, the role of the scene context constraint network is relatively weakened, but the role of the target context constraint network is enhanced. Therefore, these targets still achieve good detection results, proving that the target context constraint network can enhance the target based on the objects around the target.

	SSD	FRCNN	YOLOv4	DCIFF	SCCM	Ours
Ship	0.881	0.780	0.884	0.895	0.915	0.921
Storage tank	0.869	0.827	0.873	0.872	0.901	0.911
Vehicle	0.863	0.845	0.872	0.865	0.890	0.901
Bridge	0.855	0.832	0.864	0.855	0.887	0.899
Harbour	0.860	0.755	0.871	0.877	0.894	0.902
Track field	0.890	0.706	0.881	0.891	0.901	0.913
Basketball field	0.901	0.714	0.899	0.907	0.913	0.921
Airplane	0.906	0.896	0.890	0.910	0.919	0.923
Baseball diamond	0.908	0.860	0.894	0.901	0.910	0.919
Tennis court	0.883	0.799	0.830	0.735	0.905	0.908
mAP	0.882	0.801	0.876	0.871	0.904	0.912

Table 2. Detection results on the NWPU VHR-10 dataset

	SSD	FRCNN	YOLOv4	DCIFF	SCCM	Ours
Ship	0.535	0.623	0.614	0.745	0.706	0.841
Bridge	0.408	0.525	0.524	0.635	0.607	0.789
Track field	0.529	0.650	0.853	0.802	0.870	0.905
Storage tank	0.589	0.651	0.652	0.741	0.763	0.856
Basketball field	0.451	0.534	0.504	0.623	0.742	0.912
Tennis court	0.457	0.531	0.654	0.741	0.801	0.901
Airplane	0.698	0.811	0.931	0.809	0.875	0.954
Baseball diamond	0.543	0.687	0.791	0.702	0.698	0.805
Harbour	0.656	0.753	0.759	0.801	0.816	0.904
Vehicle	0.327	0.457	0.514	0.614	0.678	0.781
Crossroads	0.468	0.668	0.708	0.696	0.751	0.893
T-junction road	0.587	0.678	0.774	0.697	0.789	0.847
Parking lot	0.589	0.665	0.689	0.784	0.759	0.863
mAP	0.525	0.634	0.690	0.722	0.758	0.865

Table 3. Detection results on the HRRSD dataset

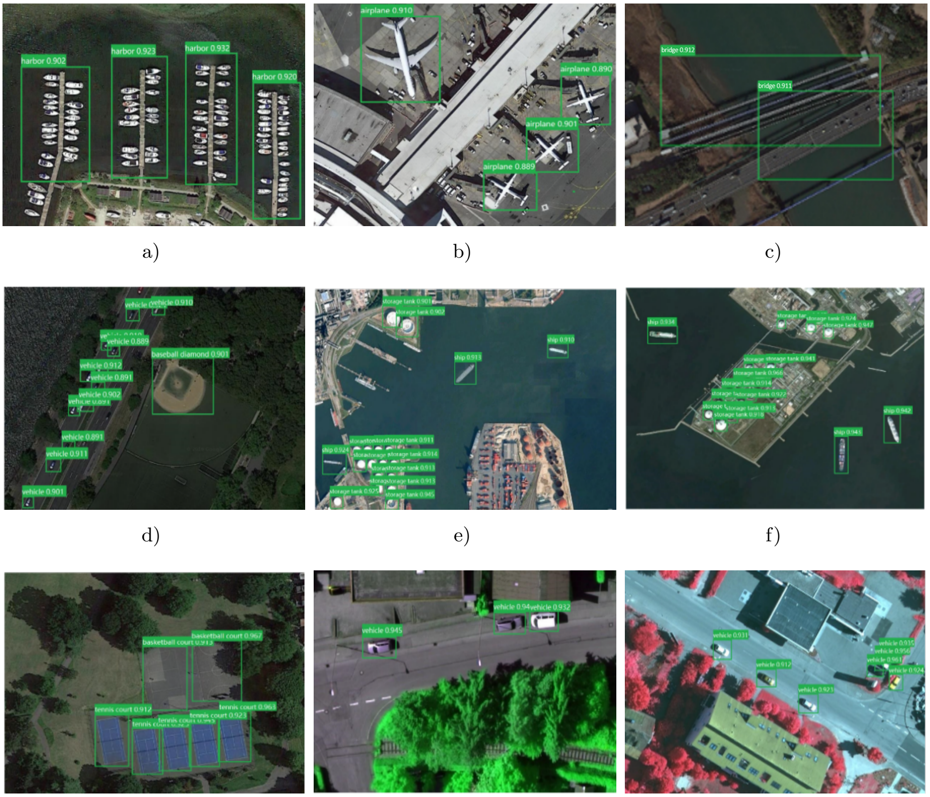
Table 3 displays the detection results of these methods on the HRRSD dataset. As shown in Table 3, SSD adopts multiple feature mapping and pixel resampling stages, so its performance is better than some traditional methods. However, SSD extracts each pixel's features, reducing the performance of small target detection. For example, the detection result of automobiles is only 0.327. The detection accuracy of vehicles is improved by 0.454, and the average detection accuracy mAP is improved by 0.34, which indicates the effectiveness of our method, especially for the detection of small targets. FRCNN and YOLOv4 are currently two of the most advanced target detection algorithms. However, due to the lack of correlation, these two algorithms are not effective in detecting remote sensing images and cannot identify multiple clustered targets in complex backgrounds. Remote sensing images have complex target types and backgrounds, which require targeted processing. Compared with these methods, our method improved by 0.231 and 0.175 mAP, respectively, indicating the effectiveness of our target relational structure diagram and scenario context constraint network. DCIFF integrates multi-scale and object background information. It achieves an average detection accuracy of 0.634 mAP on remote sensing images containing multiple targets. However, this method does not use background information and has particular error detection. Our method integrates background information and increases the average detection accuracy mAP by 0.143. The result effectively proves the effectiveness of our scene context constraint network. SCCM-BR network uses LSTM to extract features and adds scene information to constrain detection results. In the complex scene of post storage tanks, ships, and ports, AP values reach 0.763, 0.706, and 0.816, respectively, but the detection effect of similar targets such as baseball courts, tennis courts, basketball courts, and track and field is not good. By comparison, the target relational structure graph is added in our method to aggregate information from the target context and improve the detection effect of similar targets. The mAP value increases by 0.107, which powerfully demonstrates the effectiveness of the target relational structure graph.

4.3.3 Visualization Results

Some typical detection results of our method on the HRRSD dataset are presented in Figure 9. Figure 9 a) shows the ports detected. Figure 9 b) shows the detected aircrafts of different sizes and directions. Figure 9 c) shows the detected bridges. Figure 9 d) shows objects of different shapes, sizes, and colors detected against a complex background, including vehicles and baseball fields. Figures 9 e) and 9 f) show a dense array of oil storage tanks and vessels of different sizes and directions detected against a complex background. Figure 9 g) shows the detected tennis and basketball courts. Figures 9 h) and 9 i) show the detection results of vehicles under different backgrounds. The experimental results demonstrate that our method can obtain accurate and stable target detection and recognition results in different categories and complex scenes.

5 CONCLUSION

Remote sensing targets are always accompanied by complex scenes and blurred backgrounds, which bring difficulties to the task of target detection in remote sensing images. In this paper, a novel remote sensing target detection model using structure-based inference has been proposed, including a target context constraint network and a scene context constraint network. The target context constraint network is used to detect targets, and its key component module is the target relationship structure graph, which is used to self-adaptively aggregate information from different targets. The scene context constraint network, on the other hand, builds a scene semantic information model to help improve the accuracy of target detection. Experiments were conducted on NWPU VHR-10 and HRRSD datasets. The results indicate that



a) b) c)

d) e) f)

g) h) i)

Figure 9. Results of our proposed network

the proposed method effectively improves the detection accuracy of remote sensing images in complex scenes and can also be adapted to the complex scene information of remote sensing images.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 623-72215) and Special fund project of Jiangsu Science and Technology Plan (No. BE-2022781).

REFERENCES

[1] Zhu, Q.—Guo, X.—Deng, W.—Shi, S.—Guan, Q.—Zhong, Y.—Zhang, L.—Li, D.: Land-Use/Land-Cover Change Detection Based on a Siamese

Global Learning Framework for High Spatial Resolution Remote Sensing Imagery. ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 184, 2022, pp. 63–78, doi: 10.1016/j.isprsjprs.2021.12.005.

[2] Pang, S.—Xie, P.—Xu, D.—Meng, F.—Tao, X.—Li, B.—Li, Y.—Song, T.: NDFTC: A New Detection Framework of Tropical Cyclones from Meteorological Satellite Images with Deep Transfer Learning. Remote Sensing, Vol. 13, 2021, No. 9, Art. No. 1860, doi: 10.3390/rs13091860.

[3] Dong, L.—Shan, J.: A Comprehensive Review of Earthquake-Induced Building Damage Detection with Remote Sensing Techniques. ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 84, 2013, pp. 85–99, doi: 10.1016/j.isprsjprs.2013.06.011.

[4] Jha, M. N.—Levy, J.—Gao, Y.: Advances in Remote Sensing for Oil Spill Disaster Management: State-of-the-Art Sensors Technology for Oil Spill Surveillance. Sensors, Vol. 8, 2008, No. 1, pp. 236–255, doi: 10.3390/s8010236.

[5] Hoque, M. A. A.—Phinn, S.—Roelfsema, C.—Childs, I.: Tropical Cyclone Disaster Management Using Remote Sensing and Spatial Analysis: A Review. International Journal of Disaster Risk Reduction, Vol. 22, 2017, pp. 345–354, doi: 10.1016/j.ijdrr.2017.02.008.

[6] Shi, Q.—Liu, X.—Li, X.: Road Detection from Remote Sensing Images by Generative Adversarial Networks. IEEE Access, Vol. 6, 2017, pp. 25486–25494, doi: 10.1109/ACCESS.2017.2773142.

[7] Wang, W.—Yang, N.—Zhang, Y.—Wang, F.—Cao, T.—Eklund, P.: A Review of Road Extraction from Remote Sensing Images. Journal of Traffic and Transportation Engineering (English Edition), Vol. 3, 2016, No. 3, pp. 271–282, doi: 10.1016/j.jtte.2016.05.005.

[8] Ryznar, R. M.—Wagner, T. W.: Using Remotely Sensed Imagery to Detect Urban Change: Viewing Detroit from Space. Journal of the American Planning Association, Vol. 67, 2001, No. 3, pp. 327–336, doi: 10.1080/01944360108976239.

[9] Cheng, G.—Han, J.: A Survey on Object Detection in Optical Remote Sensing Images. ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 117, 2016, pp. 11–28, doi: 10.1016/j.isprsjprs.2016.03.014.

[10] Bückner, J.—Pahl, M.—Stahlhut, O.—Liedtke, C. E.: A Knowledge-Based System for Context Dependent Evaluation of Remote Sensing Data. In: Van Gool, L. (Ed.): Pattern Recognition (DAGM 2002). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 2449, 2002, pp. 58–65, doi: 10.1007/3-540-45783-6_8.

[11] Akçay, H. G.—Aksoy, S.: Building Detection Using Directional Spatial Constraints. 2010 IEEE International Geoscience and Remote Sensing Symposium, 2010, pp. 1932–1935, doi: 10.1109/IGARSS.2010.5652842.

[12] Liow, Y. T.—Pavlidis, T.: Use of Shadows for Extracting Buildings in Aerial Images. Computer Vision, Graphics, and Image Processing, Vol. 49, 1990, No. 2, pp. 242–277, doi: 10.1016/0734-189X(90)90139-M.

[13] Ok, A. O.—Senaras, C.—Yuksel, B.: Automated Detection of Arbitrarily Shaped Buildings in Complex Environments from Monocular VHR Optical Satellite Imagery. IEEE Transactions on Geoscience and Remote Sensing, Vol. 51, 2012,

No. 3, pp. 1701–1717, doi: 10.1109/TGRS.2012.2207123.

[14] Dong, R.—Xu, D.—Zhao, J.—Jiao, L.—An, J.: Sig-NMS-Based Faster R-CNN Combining Transfer Learning for Small Target Detection in VHR Optical Remote Sensing Imagery. IEEE Transactions on Geoscience and Remote Sensing, Vol. 57, 2019, No. 11, pp. 8534–8545, doi: 10.1109/TGRS.2019.2921396.

[15] Cheng, B.—Li, Z.—Xu, B.—Dang, C.—Deng, J.: Target Detection in Remote Sensing Image Based on Object-and-Scene Context Constrained CNN. IEEE Geoscience and Remote Sensing Letters, Vol. 19, 2021, pp. 1–5, doi: 10.1109/LGRS.2021.3087597.

[16] Zhang, W.—Wang, S.—Thachan, S.—Chen, J.—Qian, Y.: Deconv R-CNN for Small Object Detection on Remote Sensing Images. IGARSS 2018, 2018 IEEE International Geoscience and Remote Sensing Symposium, 2018, pp. 2483–2486, doi: 10.1109/IGARSS.2018.8517436.

[17] Qu, J.—Su, C.—Zhang, Z.—Razi, A.: Dilated Convolution and Feature Fusion SSD Network for Small Object Detection in Remote Sensing Images. IEEE Access, Vol. 8, 2020, pp. 82832–82843, doi: 10.1109/ACCESS.2020.2991439.

[18] Deng, Z.—Lei, L.—Sun, H.—Zou, H.—Zhou, S.—Zhao, J.: An Enhanced Deep Convolutional Neural Network for Densely Packed Objects Detection in Remote Sensing Images. 2017 International Workshop on Remote Sensing with Intelligent Processing (RSIP), IEEE, 2017, pp. 1–4, doi: 10.1109/RSIP.2017.7958800.

[19] Bell, S.—Zitnick, C. L.—Bala, K.—Girshick, R.: Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2874–2883, doi: 10.1109/CVPR.2016.314.

[20] Chaudhuri, D.—Kushwaha, N. K.—Samal, A.: Semi-Automated Road Detection from High Resolution Satellite Images by Directional Morphological Enhancement and Segmentation Techniques. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Vol. 5, 2012, No. 5, pp. 1538–1544, doi: 10.1109/JSTARS.2012.2199085.

[21] Fischler, M. A.—Elschlager, R. A.: The Representation and Matching of Pictorial Structures. IEEE Transactions on Computers, Vol. C-22, 1973, No. 1, pp. 67–92, doi: 10.1109/T-C.1973.223602.

[22] Ahmadi, S.—Valadan Zoej, M. J.—Ebadi, H.—Abrishami Moghaddam, H.—Mohammadzadeh, A.: Automatic Urban Building Boundary Extraction from High Resolution Aerial Images Using an Innovative Model of Active Contours. International Journal of Applied Earth Observation and Geoinformation, Vol. 12, 2010, No. 3, pp. 150–157, doi: 10.1016/j.jag.2010.02.001.

[23] Fua, P.—Hanson, A. J.: Using Generic Geometric Models for Intelligent Shape Extraction. Proceedings of the Sixth National Conference on Artificial Intelligence (AAAI '87) – Volume 2, 1987, pp. 706–709.

[24] Blaschke, T.: Object Based Image Analysis for Remote Sensing. ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 65, 2010, No. 1, pp. 2–16, doi: 10.1016/j.isprsjprs.2009.06.004.

[25] Girshick, R.—Donahue, J.—Darrell, T.—Malik, J.: Rich Feature Hierar-

chies for Accurate Object Detection and Semantic Segmentation. 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 580–587, doi: 10.1109/CVPR.2014.81.

[26] Ren, S.—He, K.—Girshick, R.—Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (Eds.): Advances in Neural Information Processing Systems 28 (NIPS 2015) – Volume 1. Curran Associates, Inc., 2015, pp. 91–99, `https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf`.

[27] Lin, T. Y.—Dollár, P.—Girshick, R.—He, K.—Hariharan, B.—Belongie, S.: Feature Pyramid Networks for Object Detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 936–944, doi: 10.1109/CVPR.2017.106.

[28] Liu, W.—Anguelov, D.—Erhan, D.—Szegedy, C.—Reed, S.—Fu, C. Y.—Berg, A. C.: SSD: Single Shot Multibox Detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.): Computer Vision – ECCV 2016. Springer, Cham, Lecture Notes in Computer Science, Vol. 9905, 2016, pp. 21–37, doi: 10.1007/978-3-319-46448-0_2.

[29] Bochkovskiy, A.—Wang, C. Y.—Liao, H. Y. M.: YOLOv4: Optimal Speed and Accuracy of Object Detection. CoRR, 2020, doi: 10.48550/arXiv.2004.10934.

[30] Redmon, J.—Farhadi, A.: YOLOv3: An Incremental Improvement. CoRR, 2018, doi: 10.48550/arXiv.1804.02767.

[31] Law, H.—Deng, J.: CornerNet: Detecting Objects as Paired Keypoints. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.): Computer Vision – ECCV 2018. Springer, Cham, Lecture Notes in Computer Science, Vol. 11218, 2018, pp. 765–781, doi: 10.1007/978-3-030-01264-9_45.

[32] Duan, K.—Bai, S.—Xie, L.—Qi, H.—Huang, Q.—Tian, Q.: CenterNet: Keypoint Triplets for Object Detection. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6568–6577, doi: 10.1109/ICCV.2019.00667.

[33] Heuer, F.—Mantowsky, S.—Bukhari, S. S.—Schneider, G.: MultiTask-CenterNet (MCN): Efficient and Diverse Multitask Learning Using an Anchor Free Approach. 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2021, pp. 997–1005, doi: 10.1109/ICCVW54120.2021.00116.

[34] Qiao, S.—Pang, S.—Luo, G.—Pan, S.—Chen, T.—Lv, Z.: FLDS: An Intelligent Feature Learning Detection System for Visualizing Medical Images Supporting Fetal Four-Chamber Views. IEEE Journal of Biomedical and Health Informatics, Vol. 26, 2022, No. 10, pp. 4814–4825, doi: 10.1109/JBHI.2021.3091579.

[35] Li, J.—Wei, Y.—Liang, X.—Dong, J.—Xu, T.—Feng, J.—Yan, S.: Attentive Contexts for Object Detection. IEEE Transactions on Multimedia, Vol. 19, 2017, No. 5, pp. 944–954, doi: 10.1109/TMM.2016.2642789.

[36] Li, K.—Cheng, G.—Bu, S.—You, X.: Rotation-Insensitive and Context-Augmented Object Detection in Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing, Vol. 56, 2018, No. 4, pp. 2337–2348, doi: 10.1109/TGRS.2017.2778300.

[37] Zhang, G.—Lu, S.—Zhang, W.: CAD-Net: A Context-Aware Detection

Network for Objects in Remote Sensing Imagery. IEEE Transactions on Geoscience and Remote Sensing, Vol. 57, 2019, No. 12, pp. 10015–10024, doi: 10.1109/TGRS.2019.2930982.

[38] Gong, Y.—Xiao, Z.—Tan, X.—Sui, H.—Xu, C.—Duan, H.—Li, D.: Context-Aware Convolutional Neural Network for Object Detection in VHR Remote Sensing Imagery. IEEE Transactions on Geoscience and Remote Sensing, Vol. 58, 2020, No. 1, pp. 34–44, doi: 10.1109/TGRS.2019.2930246.

[39] Sun, X.—Wang, P.—Wang, C.—Liu, Y.—Fu, K.: PBNet: Part-Based Convolutional Neural Network for Complex Composite Object Detection in Remote Sensing Imagery. ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 173, 2021, pp. 50–65, doi: 10.1016/j.isprsjprs.2020.12.015.

[40] Cho, K.—van Merrienboer, B.—Bahdanau, D.—Bengio, Y.: On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. CoRR, 2014, doi: 10.48550/arXiv.1409.1259.

[41] Krizhevsky, A.—Sutskever, I.—Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira, F., Burges, C. J., Bottou, L., Weinberger, K. Q. (Eds.): Advances in Neural Information Processing Systems 25 (NIPS 2012). Curran Associates, Inc., 2012, pp. 1097–1105, `https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf`.

[42] Cheng, G.—Han, J.—Zhou, P.—Guo, L.: Multi-Class Geospatial Object Detection and Geographic Image Classification Based on Collection of Part Detectors. ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 98, 2014, pp. 119–132, doi: 10.1016/j.isprsjprs.2014.10.002.

[43] Zhang, Y.—Yuan, Y.—Feng, Y.—Lu, X.: Hierarchical and Robust Convolutional Neural Network for Very High-Resolution Remote Sensing Object Detection. IEEE Transactions on Geoscience and Remote Sensing, Vol. 57, 2019, No. 8, pp. 5535–5548, doi: 10.1109/TGRS.2019.2900302.

[44] Cheng, B.—Li, Z.—Xu, B.—Yao, X.—Ding, Z.—Qin, T.: Structured Object-Level Relational Reasoning CNN-Based Target Detection Algorithm in a Remote Sensing Image. Remote Sensing, Vol. 13, 2021, No. 2, Art. No. 281, doi: 10.3390/rs13020281.

Yi Ding graduated with his B.Sc. degree from Nanjing University and completed his M.Sc. degree at the National University of Singapore. His research interests lie in deep learning and computer vision.

Xiangru Lyu is currently an M.Sc. student studying the programme of communication networks and software at the School of Computer Science and Electric Engineering at the University of Surrey, U.K. His research interests are in the areas of deep learning, software engineering and communication networks.

Liuyang Yan received her B.Sc. degree in Internet of Things engineering from the Jiangsu University, Zhenjiang, China, in 2020, where she is currently pursuing her M.Sc. degree in computer technology. Her research interests include computer vision and image processing.

Lan Rong is currently pursuing her M.Sc. degree at the School of Computer Science and Communication Engineering at Jiangsu University, China. Her research interests include machine learning and computer vision.

Keyang Cheng received his Ph.D. degree in computer application technology from the School of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2015. He was a Postdoctoral Researcher with the University of Warwick, Coventry, U.K., in 2016. He is currently serving as Professor with the Department of Computer Science and Telecommunications Engineering, Jiangsu University, Zhenjiang, China. He has coauthored more than 50 journal and conference papers. His current research interests include artificial intelligence and Internet of Things.