G2Grad-CAMRL: An Object Detection and Interpretation Model Based on Gradient-Weighted Class Activation Mapping and Reinforcement Learning in Remote Sensing Images

Shoulin Yin, Liguo Wang, *Member, IEEE*, Muhammad Shafiq, Lin Teng, Asif Ali Laghari, and Muhammad Faizan Khan

***Abstract*—Remote sensing images (RSIs) contain important information, such as airports, ports, and ships. By extracting RSI features and learning the mapping relationship between image features and text semantic features, the interpretation and description of RSI content can be realized, which has a wide range of application value in military and civil fields, such as national defense security, land monitoring, urban planning, and disaster mitigation. Aiming at the complex background of RSIs and the lack of interpretability of existing target detection models, and the problems in feature extraction between different network structures, different layers, and the accuracy of target classification, we propose an object detection and interpretation model based on gradient-weighted class activation mapping and reinforcement learning. First, ResNet is used as the main backbone network to extract the features of RSIs and generate feature graphs. Then, we add the global average pooling layer to obtain the corresponding feature weight vector of the feature graph. The weighted vectors are superimposed to output class activation maps. The reinforcement learning method is used to optimize the generated region generation network. At the same time, we improve the reward function of reinforcement learning to improve the effectiveness of the region generation network. Finally, network dissecting analysis is used to obtain the interpretable semantic concept in the model. Through experiments, the average accuracy is more than 85%. Experimental results in the public RSI description dataset show that the proposed method has high detection accuracy and good description performance for RSIs in complex environments.**

***Index Terms*—Gradient-weighted class activation mapping, network dissecting analysis (NDA), object detection and interpretation, reinforcement learning, remote sensing images (RSIs).**

I. Introduction

REMOTE sensing image (RSI) interpretation is the core and key link of RSI application. Efficient and accurate interpretation technology is helpful to improve the application level and expand the application field of remote sensing [1], [2]. Currently, the remote sensing survey and update of surveying and mapping, land, forestry, and other industries in China still mainly adopt manual visual interpretation, which is a time-consuming, laborious, costly, and long cycle. It cannot meet the urgent needs of rapid extraction and update of natural resource information in the current rapid economic and social development [3], [4], [5].

In recent years, the pullulating of remote sensing technology makes it no longer difficult to obtain remote sensing data. Under the condition of sufficient data, the object detection methods of natural images cannot be applied to RSIs because of the problems, such as single prediction scale, poor effect of horizontal frame fitting to target, and lack of enhancement of target features [6]. The problems in the field of remote sensing object detection can be concluded as follows.

1) *Scale change problem:* RSI has large scene information. The image reaches a resolution of millions of pixels. Therefore, the target scale is small relative to the image, which leads to the failure to obtain the fine features of the target. In addition, the RSI target scale variation range is wide, which is not conducive to single-scale multiclass target detection. Chen et al. [7] proposed a multiscale object detection framework based on a context feature pyramid, which improved the performance of multiscale object detection by enhancing the connection between scene and object. Zhao et al. [8] proposed a rotation-invariant CNN model for learning rotation invariant, which introduced and learned a new rotation invariant layer to increase the detection effect. Obeso et al. [9] proposed a multiscale object detection algorithm based on an attention mechanism, which introduced the attention mechanism to redis-

Manuscript received 15 October 2022; revised 18 November 2022 and 19 December 2022; accepted 20 January 2023. Date of publication 7 February 2023; date of current version 12 April 2023. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62250410365 and Grant 62071084 and in part by the Leading Talents Project of the State Ethnic Affairs Commission. *(Corresponding author: Muhammad Shafiq.)*

Shoulin Yin and Lin Teng are with the College of Information and Communications Engineering, Harbin Engineering University, Harbin 150000, China (e-mail: yslin@hit.edu.cn; tenglinheu@163.com).

Liguo Wang is with the College of Information and Communications Engineering, Dalian Minzu University, Dalian 116600, China (e-mail: wangliguo@hrbeu.edu.cn).

Muhammad Shafiq is with the Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou 510006, China (e-mail: srsshafiq@gmail.com).

Asif Ali Laghari is with the Software College, Shenyang Normal University, Shenyang 110034, China (e-mail: asif.laghari@smiu.edu.pk).

Muhammad Faizan Khan is with the Department of Information Technology, University of Haripur, Haripur 222146, Pakistan (e-mail: khanmuhammadfaizan@uoh.edu.pk).

Digital Object Identifier 10.1109/JSTARS.2023.3241405

tribute the weights of feature maps in different channels. Although some methods solve the scale problem, time efficiency cannot be guaranteed.

2) *Goal orientation problem:* The objects in RSIs are oriented and densely arranged, and the direction of the objects is irregular. Therefore, the detection model needs to have rotation invariance and a better-quality detection box. Aiming at the problem that the object detection approaches are difficult to distinguish the mixed pixels and the threshold is difficult to select, the adversarial growth algorithm was proposed in [10]. Feng et al. [11] proposed a single-stage target detection algorithm with a dynamic receptive field. Bottom-up short connection pathway and global context up-sampling module were added to the RetinaNet structure to enhance the structural and semantic features of the detection layer. The cascade R-CNN algorithm in [12] continued the two-stage idea based on candidate regions and adopted the cascade detection head structure, which could improve the detection performance step by step and had a good effect on small targets. However, there is still the problem of incomplete feature extraction.
3) *The background is complex and chaotic:* The RSI background is complex and diverse, including a large amount of redundant background information, such as mountains, rivers, and so on. This leads to the blurred boundary between background and target, which is not conducive to the extraction of target features by the model. Avola et al. [13] proposed to enhance features by capturing the correlation between global scenes and local features. Sun et al. [14] proposed nonmaximum suppression constrained by aspect ratio to improve the quality of candidate regions, and used a deformable convolutional neural network to model geometric changes of objects, which effectively improved object detection. Cheng et al. [15] proposed a pixel attention mechanism to suppress image noise and highlight target features, and introduced Intersection over Union (IoU) constant factor into SmoothL1 loss to solve the rotation box boundary problem, so as to make rotation box prediction more accurate. Chen et al. [16] replaced the traditional bounding box with a rotatable border embedded in SSD, so that the algorithm could predict the direction Angle of the target and had rotation invariance. These algorithms are improved based on traditional CNN for RSIs, which improves the performance of RSIs target detection to a certain extent. However, there are still some problems in RSIs, such as target detection angle offset, more missed detection, and low recall rate.
4) *Lack of interpretability:* Deep learning is a “black box” model, which lacks explanatory information about the predicted behavior of the model. As remote sensing technology is related to national security issues, it is essential to execute interpretable analysis on the model to a certain extent to enhance the confidence of the prediction results. For example, Li et al. [17] added a semantic graph module to the pretrained CNN to obtain semantic information of classification and enhance interpretability. Yan et al. [18] proposed the method of gradient attribution, which used the gradient of each pixel in the input model to understand the association between the input and the prediction results. In addition, there are also some visualization methods [19], [20], such as visualizing the regions with large activation values of the convolution kernel and analyzing the information obtained by the model in the image. These interpretable methods generally use human subjective judgment and lack in-depth analysis.

Object detection is to detect objects with different scales and categories in images and give the predicted positions of objects of different categories. In object detection methods, manual selection is usually used in the feature extraction stage, such as scale-invariant feature transformation and orientation gradient histogram [21], [22]. The performance of feature extraction methods largely depends on feature design, which requires a lot of prior knowledge. Therefore, this kind of method has a high design cost, poor feature robustness, and weak generalization ability. Compared with the method of manually designed features, object detection based on deep learning uses CNN to extract image features [23], [24], which has automatic and powerful feature extraction ability, better robustness, and higher detection accuracy. Therefore, the traditional object detection method has been gradually replaced by deep learning-based methods.

Deep learning-based object detection algorithms can be roughly divided into anchor-based algorithms and nonanchor-free algorithms. The difference is whether to use anchor points to extract candidate boxes.

Anchor-based algorithms include two-stage detection models, such as region CNN (R-CNN) series, and one-stage detection models, such as YOLOv2 (You Only Look Once2), SSD (Single Shot MultiBox Detector), etc. [25]. R-CNN first generates candidate boxes for feature extraction and then puts classifiers in these regions to correct and extract targets. Faster R-CNN uses region proposal network (RPN) to deepen the detection task. For a given image, SSD outputs the borders and categories of the target using regression.

Nonanchor-based algorithms discard anchors and obtain box descriptions through other methods, such as YOLOv1 (You Only Look Once Version1), CornerNet [26], ExtrmeNet [27], fully convolutional one stage (FCOS), etc. YOLOv1 performs the regression of target position and category for each pixel of the feature map. CornerNet and ExtremeNet use the key point regression detection box. CornerNet transforms the regression frame positioning problem into a detection and matching problem for the upper left and lower right points. ExtremeNet defines key points as extreme points and groups key points according to the geometric structure. FCOS uses dense prediction to predict detection boxes, and the detector directly takes pixels as training samples, so it does not need anchor points to restrict the selection of features.

Most existing studies recognize and detect targets in RSIs based on deep learning, and achieve high detection accuracy. However, object detection methods cannot generate text descriptions related to RSI content, and there is a semantic gap between low- and high-level semantic features. It cannot realize sensing and understanding of RSIs, and has certain limitations [28]. Wu et al. [29] proposed a novel global context-weaving network (GCWNet) for object detection in RSIs to solve dense instance stacking, large-scale variations, and complex background issues. Wang et al. [30] proposed an end-to-end feature-reflowing pyramid network (FRPNet), which had two advantages that contributed to improving object detection accuracy. Wu et al. [31] proposed a context-driven detection network (CDD-Net) to improve the accuracy of multiclass object detection in RSIs. For capturing the local neighboring objects and features, a local context feature network was proposed to learn the local context of the region of interest. Unlike object detection, image description methods combine computer vision and natural language processing. Image description can extract the target area in remote sensing images. The extracted features include spatial feature, environmental feature and scenarios. It studies the connection between the image features, text semantic features and the mapping relationship.

Currently, most of the research on image description focuses on natural scenes, and there are few studies on image description for remote sensing scenes. Zhou et al. [32] proposed a description generation model based on multiscale and attentional feature enhancement, which realized the description of RSIs. Sun et al. [33] proposed an RSI description model based on deep learning and CNN. Xue and Tong [34] proposed a deep multimodal neural network model, which could be used for the text description of high-resolution RSIs. Lu et al. [35] constructed a public RSI description dataset and used a multimodal method and attention method to generate description of the content of RSIs.

Although the above-mentioned researchers have realized the description of RSIs, it is easy to be affected by the complex background of RSIs, more noise information, and a small proportion of targets, resulting in low accuracy of the generated RSI description results, which cannot meet the requirements of RSI description in complex environments. For example, if the background color is similar to the remote sensing target color, it will be difficult to distinguish the remote sensing target, and clouds, atmospheric particles, and fog will bring great difficulties to the extraction of RSI features.

Our main contributions are as follows. This article presents an object detection and interpretation model based on gradient-weighted class activation mapping and reinforcement learning. The backbone network based on ResNet is used to extract features from RSIs. Then, the global average pooling (GAP)layer is added to obtain the corresponding feature weight vector of the feature graph. The weighted vectors are superimposed to output class activation maps. The reinforcement learning method is used to optimize the generated region generation network. Meanwhile, we improve the reward function of reinforcement learning to improve the effectiveness of the region generation network.

This article is organized as follows. Related works are reviewed in Section II, including deep learning interpretability approaches and Grad-CAM. Section III proposes image object detection and interpretation. Several experiments are conducted in Section IV to show the superiority of the presented method. Finally, Section V concludes this article.

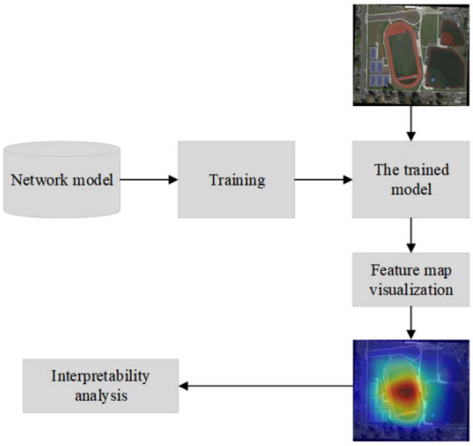

Fig. 1. Traditional visual interpretability method.

II. Related Works

With the rapid development of remote sensing technology, high-resolution RSIs contain increasingly rich information, which greatly promotes the applied research in the field of remote sensing. RSIs contain important information such as airports, ports, and ships. By extracting RSI features and learning the mapping between image features and text semantic features, RSI content can be interpreted and described. It has a wide range of application value in military and civil fields such as national defense security, land monitoring, urban planning, and disaster mitigation [36], [37]. For example, in national defense security, by extracting and capturing important information such as airports and ships in RSIs, text descriptions related to the content of RSIs with smooth semantics can be generated, which can provide military information for military security managers, assist them to make decisions quickly and deploy tasks. In the civil field, the generated RSI text description can accurately provide important information about disaster assessment, farmland utilization, vegetation cover, and urban change, and provide decision support for relevant managers. Therefore, it is of great significance to describe RSIs.

A. Deep Learning Interpretability Approaches

Currently, the interpretability of deep learning is divided into several branches, among which the visualization method is one of the important research directions. Zhang et al. [38] proposed sensitivity analysis to quantify the sensitivity of the model to input variables and visualize regions with high sensitivity, indicating that this region mainly affected model decision-making. Other visualization methods sample the image blocks with the largest convolution kernel activation value [39], and then visualize these activated image blocks to analyze how the networks obtain information. Ke et al. [40] used two visualization techniques (occlusion and guided backpropagation) to find relatively important areas in the image.

As shown in Fig. 1, the interpretable visualization algorithm described above visualizes network feature maps or activation maps without further analysis of these visual features. These

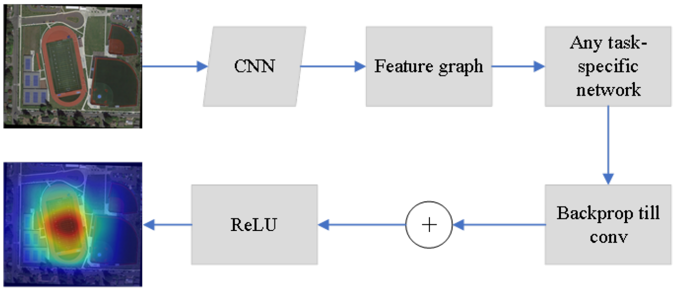

Fig. 2. Structure of Grad-CAM.

methods use human visual observation analysis to obtain the interpretability of network models, which is prone to human subjective judgment errors.

B. Gradient-Weighted Class Activation Mapping (Grad-CAM)

Grad-CAM belongs to the method based on class activation mapping in local interpretation [41]. According to the prediction results of a single image, the heat map highlighting the important region is obtained by combining its feature maps as the interpretation result image. Grad-CAM can also be used for weakly supervised localization problems, that is, only the label information of the image is given, and then the object referred to by the label in the image is located.

The idea of Grad-CAM is to calculate the gradient of the feature map of the last convolutional layer, which is used as the weight to obtain the thermal map for a specific category. Since the thermal map is coarse-grained, the method can also be combined with the visual interpretation method based on backpropagation to get the interpretation map with clear semantics, that is, the high-resolution, pixel-level saliency map. This method is simple and intuitive and can be flexibly applied to models of different tasks, such as image classification, image understanding, and visual question answering, as shown in Fig. 2.

The shallow feature maps of deep neural networks usually encode basic concepts such as color and texture. Deep feature maps encode more advanced concepts of semantic and spatial information. The fully connected layer discards most of the concept of spatial information. Therefore, Grad-CAM selects the feature map output by the last convolutional layer as the original information to provide interpretation. Taking the model performing the classification task as an example, to obtain the thermal map $L^c_{\text{Grad-CAM}}$ about class c, the gradient of the output y^c of the fully connected layer concerning the kth feature map A^k of the convolution layer, namely $\frac{\partial y^c}{\partial A^k}$, is first calculated. Then, GAP is performed to obtain the importance score α^c_k of the feature map for category c, namely

$$\alpha^c_k = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A^k}. \tag{1}$$

Finally, all the feature maps of this convolutional layer are summed by α^c_k weighting and ReLU activation is performed to obtain the saliency map $L^c_{\text{Grad-CAM}}$ about category c, namely

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha^c_k A^k\right). \tag{2}$$

ReLU activation is performed to screen out regions that have a positive impact on category c, that is, these regions can increase the output y^c of the fully connected layer on category c. However, regions with negative influence may be related to other categories, and displaying both positive and negative regions at the same time may lead to relatively chaotic positioning results. The final interpretation result image is obtained by up-sampling and normalization of the thermal map $L^c_{\text{Grad-CAM}}$.

Guided feature inversion [42] is a visual interpretation method based on class activation mapping in local interpretation. In other words, based on the prediction results of a single image, the thermal map of prominent important areas can be obtained by combining its feature maps as the interpretation result image.

First, the original image I_a is fed into the model to obtain the feature map output by each layer. Also, based on the deep feature map, high-level semantic and spatial information are encoded, and the feature map output by the last convolutional layer is selected as the original information to provide interpretation. Then, a weight vector ω is initialized with a constant. An intermediate thermal map m is obtained by weighting the feature map, namely

$$m = \sum_i \omega_i f^{l_1}_i (I_a) \tag{3}$$

where $f^{l_1}_i(I_a)$ represents the ith feature map of the l_1th layer of the model. m is upsampled and normalized, and a perturbed image Φ is generated using m guidance, i.e.,

$$\Phi(I_a, \omega) = I_a \odot m + p \odot (1 - m) \tag{4}$$

where p is a noisy background image. It can be a gray image, a white Gaussian noise image, or the original image after the Gaussian blur image. The last method is used in the document to minimize artifacts from sharp edges. For such unnatural images, it is impossible to judge how much the model is altering its predictions because of artificial traces. The perturbed image Φ will retain the region highlighted by the intermediate thermal map m.

In this way, a generation can be selected to optimize the weight vector ω so that the distance between the original image and the feature map output by the perturbed image in the last convolution layer of the model is as small as possible, namely

$$L_{\text{inv}}(I_a, \omega) = \left|\left| f^{l_0}(\Phi(I_a, \omega)) - f^{l_0}(I_a)\right|\right|^2 + \gamma||\omega||_1. \tag{5}$$

The second term is the L1 constraint, which is to keep the number of importance scores greater than 0 in the weight vector ω as small as possible. Because the model does not need to use all the feature maps to identify an object, and even only needs the corresponding feature map of a part of the object to make a correct prediction, that is, the feature map used for a prediction is sparse. The significance of using the original image is to ensure that the noise of the optimized intermediate thermal map is as small as possible on the one hand, and to reduce the number of parameters to be optimized on the other hand.

At this time, after the first step of optimization, the obtained intermediate heat map does not have class discrimination. It is just a linear superposition of the feature maps, so it highlights all the foreground objects. To make the interpretation result image class-discriminative, an objective function should be added to fine-tune the weight vector ω. The aim is to make the probability of the model predicting the perturbed image into the specified category as high as possible and the probability of its complementary image as low as possible. The complementary image is defined as follows:

$$\Phi_{\text{b}_\text{g}} = (I_a, \omega) = I_a \odot (1 - m) + p \odot m. \tag{6}$$

So, the objective function for the second stage is

$$L_{\text{target}}(I_a, \omega) = -f_c^L(\Phi(I_a, \omega) + \lambda f_c^L \Phi_{\text{b}_\text{g}}(I_a, \omega) + \delta ||\omega||_1 \tag{7}$$

where f_c^L is the prediction probability of model output. Thus, the first term improves the prediction probability of the specified category for the prominent region of the intermediate thermal map, whereas the second term reduces the prediction probability of the complementary region for the specified category.

The thermal map can be obtained by the superposition of random masks. The importance of the region λ retained by mask Q is defined as the prediction probability of the perturbed image obtained by its element-level multiplication with image I. Then, the final interpretation of the importance of the prominent area in the resulting image is the expectation obtained by all masks, namely

$$S_{I,f}(\lambda) = E_Q \left[f(I \odot Q) | Q(\lambda) \right]. \tag{8}$$

After the multiplication of mask and image elements, if the prediction probability of model f is greater, the area retained by this mask is more important.

It expands the above equation according to the expected definition and rewrites it using conditional probability

$$\begin{aligned} S_{I,f}(\lambda) &= \sum_q f(I \odot Q) P[Q = q | Q(Q) = 1] \\ &= \frac{1}{P[Q(\lambda) = 1]} \sum_q f(I \odot Q) P[Q = q, Q(Q) = 1]. \end{aligned} \tag{9}$$

The second term is

$$P(Q = q, Q(\lambda) = 1) = \begin{cases} 0 & if\ q(\lambda) = 0 \\ P[Q = q] & if\ q(\lambda) = 1 \end{cases}. \tag{10}$$

So

$$P(Q = q, Q(\lambda) = 1) = q(\lambda) P[Q = q]. \tag{11}$$

On substituting it into (9), we get

$$S_{I,f}(\lambda) = \frac{1}{P[Q(\lambda)] = 1} \sum_q f(I \odot q) q(\lambda) P[Q = q]. \tag{12}$$

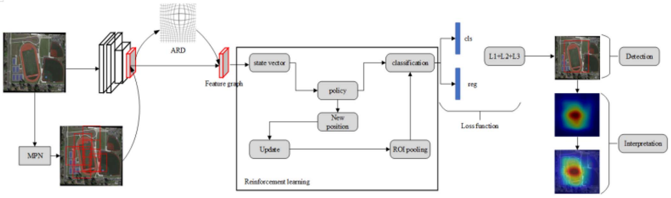

Fig. 3. Proposed G2Grad-CAMRL model.

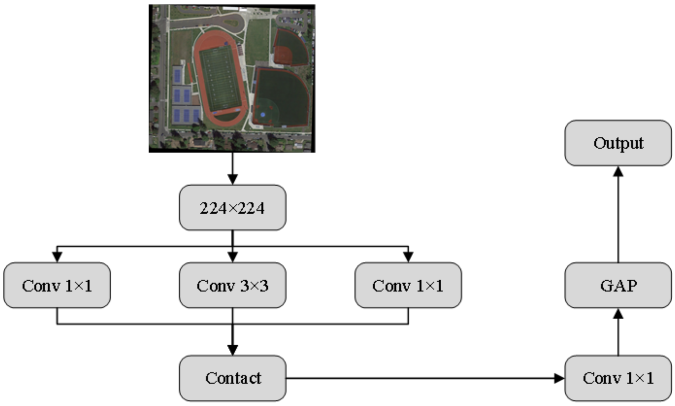

Fig. 4. Mask proposal network.

Since the mask m is distributed in [0-1], $P[Q(\lambda)] = 1 = E[Q(\lambda)]$, i.e.,

$$S_{I,f} = \frac{1}{E[Q(\lambda)]} \operatorname{sum}_q f(I \odot q) \cdot q(\lambda) \cdot P[Q = q]. \tag{13}$$

According to (12), the thermal map can be obtained by weighting the mask obtained by random sampling. The weight is the prediction probability of the disturbed image. When uniformly sampled, $P[Q = q] = 1/N$, i.e.,

$$S_{I,f} \overset{MC}{\approx} \left\{ \frac{1}{E[Q]\, N} \sum_{i=1}^{N} f(I \odot Q_i) \right\} \cdot M_i(\lambda) \Big\}. \tag{14}$$

Because the pixel-level mask may have a great impact on the model, a small part of pixels may be occluded, which may cause a great change in the prediction of the model. In addition, sampling a pixelwise mask computationally takes an exponential amount of space. Therefore, when generating masks to ensure smoothness, smaller masks are first generated, and then they are upsampled back to the image size.

III. Proposed G2Grad-CAMRL

The proposed method is shown in Fig. 3. It includes three main learning stages: mask proposal network (MPN), reinforcement learning, and network dissecting analysis (NDA).

A. *Mask Proposal Network Based on Grad-CAM*

We propose an object MPN combined with Grad-CAM to achieve the purpose of adjusting the proportional relationship between target and background information as shown in Fig. 4.

In this article, GAP is chosen instead of global max pooling (GMP), because the algorithm requires the MPN to obtain the maximum possible feature region to distinguish target categories. GMP can only output the area with the highest identification of the target and completely abandon the feature area with low identification.

We use ResNet to build the MPN combined with Grad-CAM. First, it adjusts the image size to 224 × 224 pixels and inputs it into ResNet [43]. The image is transferred to the convolutional layer in the network, and the output size of this layer is [77 512]. This output is also known as the eigenmap vector. Let $f^k(w,h)$ represent the activation response of kernel unit k at any position (w,h) in the eigenvector graph, where k represents the kth [77] feature map in the vector. Then it inputs $f^k(w,h)$ into the GAP layer and obtains the output

$$F^k = \sum_{p=(w_0,h_0)}^{(w_0+w_l,h_0+h_l)} f^k(p) \tag{15}$$

where $p = (w,h)$. (w_0,h_0) represents the upper-left coordinate of the image, (w_0+w_l, h_0+h_l) represents the coordinate at the lower right corner of the image, w_l is the width of the image, and h_l is the height of the image.

For the image with category c label, Grad-CAM can be calculated by the following formula:

$$S_c = \sum_k \omega_c^k F^k. \tag{16}$$

Substituting (15) into (16), the following equation can be obtained:

$$S_c = \sum_{p=(w_0,h_0)}^{(w_0+w_l,h_0+h_l)} \sum_k \omega_c^k F^k(p). \tag{17}$$

When the image is predicted to be class c, the Grad-CAM value of any coordinate position in the image can be calculated by the following formula:

$$P_c(p) = \sum_k \omega_c^k F^k(p),\ p = (w,h). \tag{18}$$

Combining (17) and (18), it can be seen that Grad-CAM is used to calculate the value of P_c at all pixel positions in the image, which is the basis for ResNet to determine the target category.

I_c is obtained by projecting P_c into the RGB space with the value range [0, 255]. The final thermodynamic map of Grad-CAM is obtained by superimposing I_c with the original image I_o through the following equation:

$$I_h = \alpha I_o\ + \beta I_c + \gamma, \alpha + \beta + \gamma\ =\ 1. \tag{19}$$

According to the following formula, an output value of the MPN is calculated as

$$I_{\text{mask}}(p)\ = \begin{cases} 1 & \text{if } P_c(p) \geq \tau \\ 0 & \text{if } P_c(p) < \tau \end{cases}. \tag{20}$$

Taking S_c as the input of formula (21), the quality score S_{mask} of the MPN is obtained as

$$S_{\text{mask}} = \frac{\exp(S_c)}{\sum_c \exp(S_c)}. \tag{21}$$

According to the following formula, the original image, the target initial positioning mask, and the mask score are used to generate the target initial positioning map I_{out}:

$$I_{\text{out}}(p) = I_o\ (p)\, I_{\text{mask}}(p)\, S_{\text{mask}}(p)\, S_{\text{mask}}. \tag{22}$$

B. Attention Region Deformation in Grad-CAM

To further fully and comprehensively learn the subtle features of the key regions, the deformation sampling method is introduced to generate extended data. Traditional deformation-based data enhancement methods usually distort images randomly [44]. But its effects are not guaranteed. The deformed image in this article can highlight the attention part and suppress the remaining part, so as to help the model continue to learn the differences of subtle features.

First, the deformed image D should be sampled by the input image I. They have the same size. It can be formalized as $D(x,y) = I(f(x,y), g(x,y))$, where x and y represent the position coordinates of the deformed image, that is, the pixel value of the deformed image D at (x,y) is equal to the pixel value of the original image I at a certain position. The horizontal and vertical coordinates of this position are determined by the mapping relations f and g, respectively. The goal of f and g is to adaptively sample the original image according to the size of each pixel value in the Grad-CAM image, that is, the pixel position in the attendance area of the Grad-CAM image is oversampled, and the pixel position in other noncritical areas is reduced or not sampled. According to [14], f and g should be satisfied

$$\int_0^{f(x,y)} \int_0^{g(x,y)} A(x', y')\, dx' dy'\ =\ xy \tag{23}$$

where x and y represent the horizontal and vertical coordinates of the deformed image. $f(x,y)$ and $g(x,y)$ represent the horizontal and vertical coordinates of the original graph to be sampled. x, y, $f(x,y)$ and $g(x,y)$ are the normalized coordinate values. Assuming that the Grad-CAM graph does not reflect the key attention area, that is, the Grad-CAM graph conforms to a uniform distribution with a pixel value equal to 1, then (23) can be satisfied by only setting $f(x,y) = x$ and $g(x,y) = y$, which is equivalent to the original image (no attention area needs to be deformed). However, if the Grad-CAM graph can reflect a key attention area, that is, the pixel value conforms to the nonuniform distribution, then we want to find f and g, it is equivalent to solving for a change that transforms the Grad-CAM graph from an uneven distribution to a uniform distribution. However, in this case, the left side of (23) needs to calculate the integral of the discrete function, which cannot be solved in the higher data category [45], [46], [47]. Therefore, it is very difficult and costly to calculate these two mapping functions accurately.

So, we need to find another approximate solution. When the Grad-CAM graph is not uniform, the goal of the solution formula can be visually understood as the original image pixel $I(x, y)$ is spreading to other pixels with $F(x, y)$ force during sampling. Therefore, f and g can be approximated as

$$\begin{cases} f\left(x, y\right), \frac{\sum_{x',y'} A(x',y')k((x,y),(x',y'))x'}{\sum_{x',y'} A(x',y')k((x,y),(x',y'))} \\ g\left(x, y\right), \frac{\sum_{x',y'} A(x',y')k((x,y),(x',y'))y'}{\sum_{x',y'} A(x',y')k((x,y),(x',y'))} \end{cases} \tag{24}$$

where $k((x, y), (x', y'))$ represents the distance measurement between two points. At this time, the sampling results are related to two factors: 1) the pixel value of each point in the Grad-CAM graph; and 2) the distance between the points to be sampled and each point in the Grad-CAM graph. If the value of a pixel in the Grad-CAM graph is larger, and the distance between the point to be sampled and the point is closer, and the possibility to select the point position in the original graph for sampling is greater. Therefore, this method can finally get a deformation effect similar to the expansion of the attention region, and the existence of distance measurement k also prevents selecting the point corresponding to the maximum position of Grad-CAM in the original image for each sampling. Finally, both the numerator and denominator in (24) can be realized by a convolution operation. In this case, k corresponds to one convolution operation (input and output channels are 1).

If the input image is $I \in R^{C\times H\times W}$, then the mapping functions f and g correspond to a flow field grid $G \in R^{H\times W\times 2}$. G represents the sampling coordinate of output image I at (x, y). $D[x, y, 0]$ represents the index of the width dimension of I. $D[x, y, 1]$ represents the index of the height dimension of I. The output sampling value is the bilinear interpolation result of the four closest corner points of the sampling point. The final sampling result D can be used as an expanded image for training.

C. Reinforcement Learning Strategy

In the training stage, the traditional image description method adopts the backpropagation algorithm to maximize the probability of the next real pixel given the previous real pixel. In the test phase, the probability of the next pixel is predicted based on the pixels previously generated by the model. This method will cause a mismatch between the training phase and the test phase, and lead to the phenomenon of exposure deviation, which causes an easy error and continuous accumulation in the test phase, and reduces the quality of the generated description image. In addition, the cross-entropy loss function optimizes the model in the training stage. In the test phase, discrete and nondifferentiable indicators can assess the quality of the generated images. This method will have the defect of inconsistent optimization direction, which leads to the inability of the network to directly use BLUE and other evaluation indicators for optimization training. When the cross-entropy loss function is minimum, the best evaluation result may not be produced.

To eliminate the defects of exposure bias and inconsistent optimization direction, this method introduces a reinforcement learning strategy [48]. The gradient algorithm in reinforcement learning strategy can train the nondifferentiable discrete variables end-to-end, and directly optimize the model according to BLUE and other indicators to improve the training effect of the model. The reinforcement learning strategy treats ResNet as an agent that interacts with the image and the external environment and defines the learning strategy p to guide the model to predict the next pixel. After generating the image description, the reinforcement learning strategy uses BLUE and other indicators to measure the fit and similarity between the image description and manually annotated reference statements, assigns ResNet an expected reward, and takes minimizing the negative expected reward as the goal to optimize the model, which can be expressed as

$$L\left(\theta\right) \ = \ -E_{w^s\sim p_\theta}\left[r\left(w^s\right)\right] \tag{25}$$

where θ is the model parameter, w^s is the sequence of each pixel, $r(\cdot)$ is the reward function, and $E(\cdot)$ is the expectation function. In practice application, $L(\theta)$ is generally obtained by single sampling with strategy p_θ, and can be expressed as

$$L\left(\theta\right) \approx -r\left(w^s\right), w^s \sim p_\theta. \tag{26}$$

Reinforcement learning adopts the policy gradient algorithm to calculate the gradient of $L(\theta)$, which can be expressed as

$$\nabla_\theta L\left(\theta\right) = \ -E_{w^s\sim p_\theta}\left[r\left(w^s\right)\nabla_\theta \text{log} p_\theta\left(w^s\right)\right]. \tag{27}$$

In practice, to facilitate the solution, Monte Carlo single sampling is used for approximate estimation, which can be expressed as

$$\nabla_\theta L\left(\theta\right) = \ -r\left(w^s\right)\nabla_\theta \text{log} p_\theta\left(w^s\right). \tag{28}$$

Due to the randomness of sampling and the lack of context normalization, the reinforcement learning strategy is used to calculate the gradient resulting in large variance and instability of the training process. To reduce the variance, a benchmark factor b is introduced to constrain and correct the expected reward function, which can be expressed as

$$\nabla_\theta L\left(\theta\right) = \ -E_{w^s\sim p_\theta}\left[r\left(w^s\right) - b)\nabla_\theta \text{log} p_\theta\left(w^s\right)\right]. \tag{29}$$

To maintain an unbiased estimate of the gradient, the benchmark factor b can be any function that does not depend on w^s. When Monte Carlo single sampling is used for approximate estimation, the gradient $\nabla_\theta L(\theta)$ can be expressed as

$$\nabla_\theta L\left(\theta\right) \ = \ -[r\left(w^s\right) - b]\nabla_\theta \text{log} p_\theta\left(w^s\right). \tag{30}$$

Using the chain derivative rule, the final gradient expression is obtained as

$$\nabla_\theta L\left(\theta\right) \ = \sum_{i=1}^{T} \frac{\partial L\left(\theta\right)}{\partial s_t}\frac{\partial s_t}{\partial \theta} \tag{31}$$

where s_t is the input of the Softmax function. When Monte Carlo single sampling is used for approximate estimation, $\frac{\partial L(\theta)}{\partial s_t}$ in (31) can be expressed as

$$\frac{\partial L\left(\theta\right)}{\partial s_t} \approx \left[r\left(w^s\right) - b\right]\left[p_\theta(w_t|h_t) - l'\right] \tag{32}$$

where l' is the one-hot vector representation of pixels. w_t and h_t are the pixel and internal vector representation at time t, respectively.

The reward function is improved according to the features of RSIs to obtain more accurate regional proposals. At each time

Algorithm 1: MPN Based on Reinforcement Learning.

```
Input: feature map to form the initial state quantity $S_0$
  At each iteration step $t$, the agent decides the next action
   based on one policy;
  for Fixate action do
    Visit a new pixel position $Z_t$, calculate fixate reward,
     RoI observation quantity $R_t$;
    Update the location;
    Send $R_t$ to the pooling layer;
    The probability vector of a specific category is inserted
     into the history quantity merged with $S_0$ to form a
     new state $S_t$ and jump to the fixation action;
  end for
  for Done action do
    Break out of the loop;
    End the search;
    Calculate the reward;
    Detection and Classification;
  end for
Return results.
```

step, the agent of MPN of reinforcement learning will calculate whether to terminate the search according to the policy. The strategy is determined by the probability of fixate action and done action. The agent represents the reinforcement learning model designed in this article. Fixate action means that after a large number of interest regions are extracted from features, these regions are screened. If a certain area is selected to calculate the reward, it is to focus on that area. As long as the search is not over, a fixate action is issued to visit the new location. Region of interest (ROI) observations are updated in the domain centered around this new location. To indicate that this area of interest has been selected, it sets all entries in this domain to 1. All ROIs are sent to the pooling layer for class-specific bounding box offset prediction. Nonmaximum suppression [49] is applied to the classified ROI to obtain the most significant information. Since the remaining regions of interest have the final bounding box prediction, they are mapped to some spatial location of the observed history for a particular class. A class-specific probability vector is inserted into the history quantity merged with the base state quantity S_t. With the new state, it takes a new action at $t + 1$ and repeats the process until the action is complete. Then it collects all the selected predictions in the entire trajectory. The RPN pseudocode of reinforcement learning is shown in Algorithm 1.

The agent of reinforcement learning should first balance two RoI selection criteria. 1) High object instance overlap should be generated; 2) The RoI number should be as small as possible to reduce the number of false positives and maintain a manageable processing time. On this basis, two action rewards are set to evaluate the actions issued by the agent: fixate action reward and done action reward.

Considering that RSIs have the characteristics of large image size and small target instances, the original reinforcement learning reward function has simple content and less data volume, which does not perform well on some datasets. Three datasets are explored in the MPN of reinforcement learning. According to the fixation reward and done reward, it is found that the fixation reward obtained by searching an image on the NWPUVHR-10 dataset is relatively dense, and the done reward is generally between -20 and -1. However, the fixation reward obtained by searching an image on DOTA and VisDrone2018 datasets is very sparse, and the done reward ranges from -50 to -20. In the DOTA and VisDrone2018 datasets, the output detection boxes of the instances are few and the target is small, so they are easy to be discarded in the training, which is unable to obtain more fixation rewards and done rewards in the image. It is difficult to converge.

For each object instance, the fixation reward first gives a small negative reward for each fixation action, but the agent also gains a positive reward for increasing IoU with any truth instance of the current image. At each time step t, the difference between the IoU of the instance and the true value and the maximum IoU value (IoU_t^i) of that instance over the entire time step are computed. Trajectory data are collected for all the regions where IoU is calculated within this time step. At the same time, when the IoU threshold is appropriately reduced, the positive reward of fixate action can be increased to encourage the agent to continue searching and obtain the prediction box that may be missed because the target instance is small. It obtains the adjusted fixate reward at time t given as

$$r_t^f = -\beta + \frac{1}{\tau}\sum_i \left(\text{IoU}_t^i - \text{IoU}^{\text{i}}\right) \tag{33}$$

where i indicates the ith object instance. The done action reward is calculated based on the IoU for each instance and truth value. The larger covered area denotes the reward closer to zero, otherwise, it becomes more and more negative. Upon termination, the agent receives a done action reward that reflects the quality of the search trajectory

$$r_t^d = \frac{1}{\tau}\sum_i \left(\text{IoU}^i - \tau\right). \tag{34}$$

The pseudocode of the reward function is shown in Algorithm 2.

D. Loss Function

The weakly supervised network mainly uses the method of weak semantic segmentation to generate the attention weight. It uses the weak semantic mask to guide the learning of the attention weight. The loss function of the weak semantic attention network is the cross-entropy loss, and the specific form is shown in the following equation:

$$L1\left(u_{ij}, u'_{ij}\right) = -\frac{1}{H \times W}\sum_i^H \sum_j^W u_{ij}\text{log}u'_{ij} \tag{35}$$

where H and W represent the length and width of the weak semantic mask. u_{ij} and u'_{ij} represent the weight value of the output point (i, j) of the attention network and the pixel value of the point (i, j) on the weak semantic mask.

Algorithm 2: Reward Function.

```
for Fixate action do
    The maximum IoU between each object instance and the truth value is calculated, denoted as $IoU^i$;
    In each time step $t$, the maximum IoU between ROI and object instance is calculated, denoted as $IoU_t^i$;
end for
    for $\mathrm{IoU}_t^i > \mathrm{IoU}^i \geq IoU$ do
    Cumulative fixate reward;
    $\mathrm{IoU}_t^i = \mathrm{IoU}^i$
    Jump to fixate action;
end for
for Done action do
    Calculate $\mathrm{IoU}^i$;
    Cumulative done reward;
end for
Return results.
```

The regression classification network contains two branches, so it is necessary to calculate the loss of the classification network and the loss of the regression network, respectively. Focal loss [50] is used for classification loss, as shown in the following equation:

$$p\left(t\right) = \begin{cases} p_n, \text{ if } t_n = 1 \\ 1 - p_n, \text{ otherwise} \end{cases} \tag{36}$$

$$L2\left(p_n, t_n\right) = -\frac{1}{N}\sum_{n\,=\,1}^{N} \alpha(1-p_t)^{\gamma}\log\left(p_t\right) \tag{37}$$

where N indicates the total number of prediction boxes, p_n represents the probability distribution of multiple categories, and t_n represents the category label of the target. In focal loss, α and γ are hyperparameters, which are set to 0.2 and 1, respectively.

In addition to the classification loss, smoothL1 loss is also used as the loss function for regression tasks in the classification regression network, as shown in (37)

$$L3\left(v'_{nj}, v_{nj}\right) = \begin{cases} \frac{1}{N}\sum_{n=1}^{N} t'_n \cdot A, \text{ if } \left|v'_{nj} - v_{nj}\right| < 1 \\ \frac{1}{N}\sum_{n=1}^{N} t'_n \cdot B, \text{ otherwise} \end{cases} \tag{38}$$

where $A = \sum_{j\in(x,y,w,h,\theta)} 0.5(v'_{nj} - v_{nj})^2$, $B = \sum_{j\in(x,y,w,h,\theta)} |v'_{nj} - v_{nj}| - 0.5$. N indicates the total number of prediction boxes, t'_n indicates confidence ($t'_n = 1$ indicates the foreground, and $t'_n = 0$ indicates the background), v'_{nj} represents the predicted coordinate vector, and v_{nj} presents the true label coordinate vector.

Therefore, the multitask loss in the model training process in this article is shown in the following equation:

$$L_{\text{final}} = \sigma_1\, L_1 + \sigma_2 L_2 + \sigma_3 L_3 \tag{39}$$

where σ_1, σ_2, and σ_3 are the balance parameters of multitask loss, L_1 is the regression loss, L_2 is attention loss, and L_3 is the classification loss.

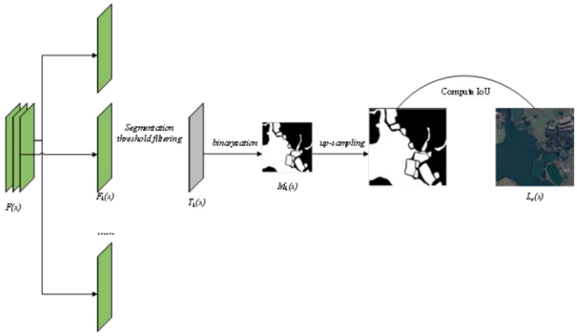

Fig. 5. Network dissecting analysis (NDA).

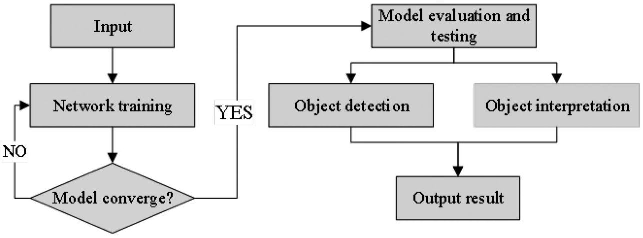

Fig. 6. Process of experiment.

E. Network Dissecting Analysis

Neural networks achieve superior performance at the cost of low interpretability of their black-box representation. However, in fields related to human or social security, such as medical treatment, driving, and remote sensing, deep learning models not only need excellent effects but also need to provide a certain basis for decision-making. In recent years, some interpretable visualization algorithms visualize network feature maps or activation maps and then perform interpretable analysis on model decisions. These methods make use of the subjective analysis of human vision and are prone to errors in judgment. In this article, the method of NDA [51] is improved. The basic principle of NDA is to explore the distribution of activation value of the feature map by using the feature map left by image forward propagation in the convolutional network according to a set of predefined human interpretable semantics and a dataset containing these interpretable semantic annotations. Then, the interpretable semantic information of the convolution kernel in the network is obtained by calculating the similarity between the distribution and the interpretable semantic annotation in the dataset.

First, the human-interpretable semantic concepts of the scene, object, component, material, texture, and color defined in the traditional method are divided in a way that conforms to human understanding. Scene, object, and component are considered high-level semantic concepts, whereas material, texture, and color are considered low-level semantic concepts. Second, the scoring value of the semantic concept is calculated by NDA shown in Fig. 5. Finally, the interpretability is quantified and used to encode the convolution kernel. Taking the second-layer

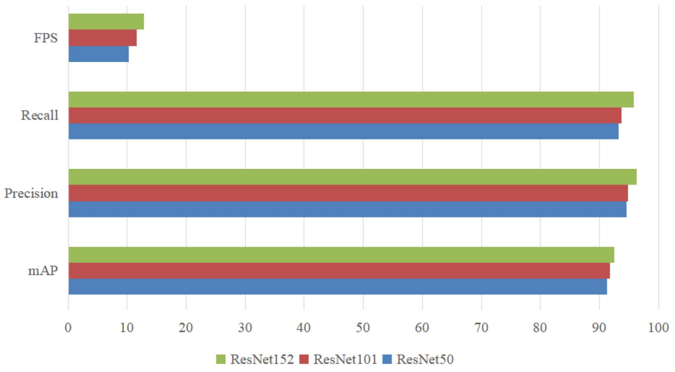

Fig. 7. Bar chart of Table III.

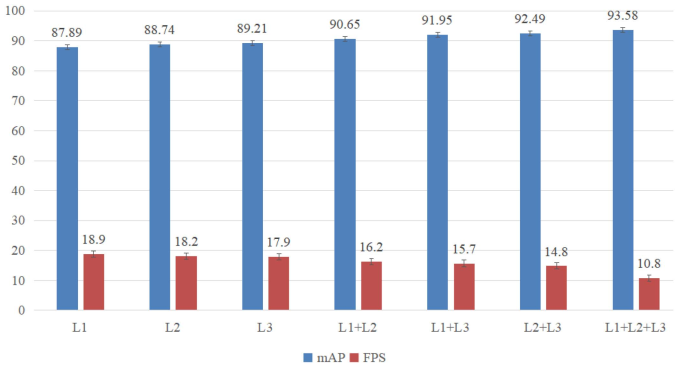

Fig. 8. Data graph of Table IV.

convolution C5_conw2 of the fifth stage in the backbone network ResNet as an example, assuming that the input image is $I(x)$, the feature map $F(x)$ output from 512 convolution kernels in Conv_2 (second-layer convolution) is saved after a forward propagation and used for subsequent interpretability calculation. $F(x)$ contains 512 feature maps, and each feature map corresponds to the semantic distribution of a convolution kernel. For $F_k(x)$ (k is the convolution kernel index), the segmentation threshold g is used to filter the weak semantic information, and the strong semantic information is retained as the semantic feature of the convolution kernel. In this article, the method of calculating threshold using probability distribution in traditional network analysis is improved, as shown in the following equation:

$$g = \frac{1}{H \times W} \sum_{i=1}^{H \times W} p_i \tag{40}$$

where H and W represent the height and width values of $F_k(x)$ and p_i represents the value of the ith pixel. The average value of the activation value is calculated as the threshold g because a value higher than the average value can better represent the semantics of its convolution kernel. The strong semantic feature graph after filtering is $T_k(x)$, and the filtering method is shown in the following equation:

$$T_k^i(x) = \begin{cases} T_k^i(x), & \text{if } T_k^i(x) \geq g \\ 0, & \text{otherwise} \end{cases} \tag{41}$$

where $T_k^i(x)$ represents the pixel value of the ith point on the feature map $T_k(x)$. $T_k(x)$ of each convolution kernel is compared with the marked semantic mask. First, a binarization preprocessing is carried out on $T_k(x)$, and the retained feature activation value is differentiated from the filtered weak semantic information to obtain a binary semantic map $M_k(x)$. Then, it is upsampled to facilitate the calculation of the semantic graph and the semantic mask. Using the IoU calculation method in [52], for mask $L_c(x)$ with different semantic C, the obtained IoU value is the interpretability score of convolution kernel k and semantic c. Finally, the interpretability scores of all convolution kernels in this layer for different semantic concepts are obtained. In this article, the overall average level is used as the scoring threshold f, and the specific calculation method is shown as follows:

$$f = \frac{1}{K} \sum_{k=1}^{K} \sum_{c=1}^{C} \frac{|M_k(x) \cap L_c(x)|}{|M_k(x) \cup L_c(x)|} \tag{42}$$

where K represents the total number of convolution kernels in this layer. The threshold f can be used to obtain the semantic concept whose score of each convolution kernel is greater than the threshold.

IV. Experiments and Analysis

To realize RSI object detection and interpretation, the public RSI description dataset DOTA [53] is used to train and learn the method. It is also compared with other current methods with good image description performance to verify the effectiveness of this method. The experimental process is shown in Fig. 6.

A. Datasets

To verify the effectiveness of the proposed method, a comparative experiment is conducted on the DOTA V1.0 dataset. DOTA dataset is a large public dataset annotated by a rotating box, which is mainly used for RSI object detection tasks. The dataset consists of 2806 RSIs from different sensors and platforms, ranging in size from 800 × 800 to 4000 × 4000 pixels, which contains 188282 target instances of different scales, orientations, and shapes. It mainly includes 15 common categories: Plane (PL), Helicopter (HC), Swimming Pool (SP), Roundabout (RA), Harbor (HA), Baseball Court (BC), Soccer Ball Field (SBF), Tennis Court (TC), Ground Track Field (GTF), Baseball Diamond (BD), Storage Tank (ST), Bridge (BR), Ship (SH), Small Vehicle (SV), and Large Vehicle (LV). In this article, 3/5 of this dataset is selected as the training set, 1/5 as the validation set, and 1/5 as the test set. All images are uniformly cropped into 1024 × 1024 pixels.

B. Evaluative Criteria

Average precision (AP) and mean average precision (mAP) are used to evaluate the detection accuracy of the model. Frames per second (FPS) is used to evaluate the detection speed of the model.

The AP can be calculated as follows:

$$\text{AP} = \int_0^1 p(r)\, dr. \tag{43}$$

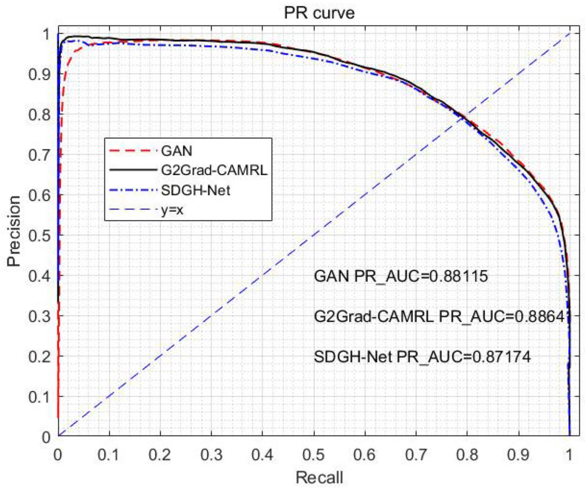

Fig. 9. PR curve.

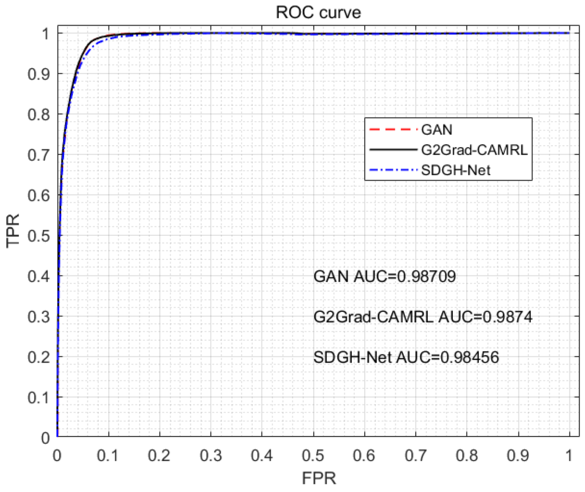

Fig. 10. ROC curve.

It calculates the area enclosed by the curves drawn in the range of precision and recall and the coordinate axes, The value range is [0, 1]. Precision and recall are defined as

$$\begin{cases} \text{Precision} &= \frac{\text{TP}}{\text{TP+FP}} \\ \text{Recall} &= \frac{\text{TP}}{\text{TP+FN}} \end{cases} \tag{44}$$

where TP represents the true positive sample, FP represents a false positive sample, and FN represents the false negative sample.

MAP can be calculated as follows:

$$\text{mAP} = \frac{\sum_{i=1}^{20} \text{AP}_i}{20}. \tag{45}$$

The FPS can be calculated as

$$\text{FPS} = \frac{S_{\text{test}}}{T} \tag{46}$$

where S_{test} is the number of samples in the test set, and T is the time consumed for the testing set.

TABLE I
INTRODUCTION OF EXPERIMENT PARAMETERS

Parameter	Description	Value
ILR	Initial learning rate	1×10^{-5}
Epoch	Training number	50
Adam	optimizer	$\cdots$
AR	attenuation rate	0.8
BS	batch size	1
λ_1	Loss balance weight	0.8
λ_2	Loss balance weight	0.3
λ_3	Loss balance weight	0.3

TABLE II
COMPARISON OF ABLATION EXPERIMENTS

Attention deformation	Grad-CAM	RLS	mAP%
	√	$\cdots$	88.73
√	√	$\cdots$	90.28
	√	√	91.89
√	√	√	94.22

TABLE III
COMPARISON OF ABLATION EXPERIMENTS

Network	mAP%	Precision %	Recall%	FPS
ResNet50	91.28	94.57	93.27	10.25
ResNet101	91.79	94.76	93.72	11.67
ResNet152	92.45	96.33	95.83	12.78

C. Experiment Process

The experiment is carried out on the PyTorch framework, and the specific parameters are shown in Table I.

D. Ablation Experiments

To verify the effectiveness of the G2Grad-CAMRL algorithm, three ablation experiments are designed. ResNet network is used as the benchmark method, and the proposed three modules in this article are used for the comparison experiment. The ablation experiment is performed on the DOTA dataset, and the experimental results are shown in Table II. $\surd$ indicates that the model contains this module. RLS is a reinforcement learning strategy.

Table II shows the effectiveness of each module proposed in this article on the object detection task in the RSI dataset. Since the background occupies a large part in RSIs, Grad-CAM can solve this problem which has a better effect on accuracy improvement. It can be seen from Table II that the G2Grad-CAMRL in this article improves the IoU threshold due to the introduction of reinforcement learning, making the object detection effect better.

To verify the universality of the proposed algorithm on different backbone networks, a set of comparison experiments are designed. The G2Grad-CAMRL is compared on different backbone networks, and the experimental results are shown in Table III. Fig. 7 is the formal bar chart of Table III to give the reader a more objective understanding.

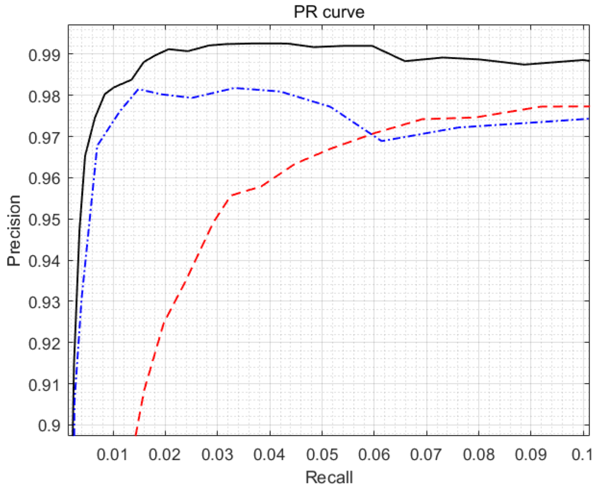

Fig. 11. Enlargement of Fig. 9.

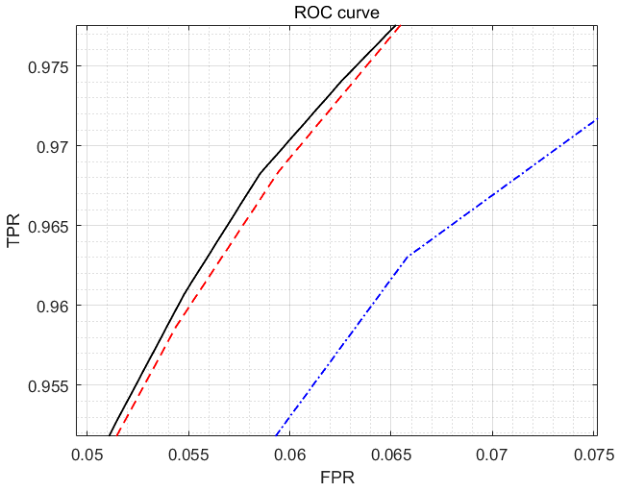

Fig. 12. Enlargement of Fig. 10.

In the comparison experiment, ResNet50, ResNet101, and ResNet152 are used in the backbone network. Table III shows that simply increasing the network depth has a limited effect on improving the object detection effect of RSIs. Because small objects in RSIs have a large proportion, and the features of small targets mostly exist in shallow semantic information, so blindly deepening the network depth is of limited help to dealing with RSI object detection tasks. The mAP of ResNet152 is 92.45%, it has a slight improvement compared to ResNet101 (91.79%) and ResNet50 (91.28%). Meanwhile, FPS values of ResNet50, ResNet101, and ResNet152 are 10.25, 11.67, and 12.78, respectively. Owing to the fewer layers in ResNet50, the FPS is less than ResNet101 and ResNet152. So to save the number of parameters, we use ResNet50 in this article.

Different loss functions can affect the convergence speed of the model and further affect the accuracy. Therefore, we conduct comparative experiments under different loss functions, and the results are shown in Table IV. Its data graph is shown in Fig. 8.

TABLE IV
COMPARISON OF LOSS FUNCTION

Loss function	mAP%	FPS
L1	87.89	18.9
L2	88.74	18.2
L3	89.21	17.9
L1+L2	90.65	16.2
L1+L3	91.95	15.7
L2+L3	92.49	14.8
L1+L2+L3	93.58	10.8

TABLE V
COMPARISON WITH CLASSICAL METHODS (AP/%)

Model	Faster R-CNN	R-FCN	YOLOv2	SSD	G2Grad-CAMRL
PL	82.05	89.63	89.75	88.91	94.58
BD	66.78	72.31	79.63	83.51	92.64
BR	53.21	53.37	55.84	60.78	87.55
GFT	68.55	60.41	77.03	78.45	90.26
SV	52.96	69.93	74.12	80.33	92.67
LV	57.31	65.61	74.25	75.79	93.82
SH	73.52	77.71	84.71	89.17	95.67
TC	76.94	82.16	89.66	92.45	95.78
BC	73.51	78.39	80.13	88.45	94.61
SBF	59.62	62.05	64.71	67.49	77.39
SP	71.97	73.19	76.38	80.27	85.88

TABLE VI
COMPARISON WITH OTHER STATE-OF-THE-ART METHODS (AP/%)

Model	CWDL	GAN	SDGH-Net	G2Grad-CAMRL
PL	90.64	91.24	93.85	94.58
BD	89.21	89.74	91.35	92.64
BR	83.61	85.72	86.93	87.55
GFT	85.55	87.16	89.44	90.26
SV	84.79	88.63	91.05	92.67
LV	88.17	90.23	91.77	93.82
SH	89.74	91.78	93.64	95.67
TC	90.17	91.54	92.85	95.78
BC	90.86	91.74	93.29	94.61
SBF	70.28	71.66	74.15	77.39
SP	81.64	83.01	83.69	85.88

The combined loss function achieves better results, it obtains 93.58% mAP. Even if the effects of these combinations are not too different, the proposed method still has the least FPS. Therefore, the loss function method in this article is competent for the task of object detection.

E. Comparison Experiments With Other Methods

The G2Grad-CAMRL in this article is compared with four classical object detection algorithms, including Faster R-CNN, R-FCN, YOLOv2, and SSD. Faster R-CNN is the benchmark model of the original DOTA dataset. The backbone network used in YOLO is DarkNet19. The backbone networks of other comparison algorithms are ResNet50 as in this article. Then, we select three other state-of-the-art algorithms for comparison including CWDL [54], GAN [55], and SDGH-Net [56]. The experimental results are shown in Tables V and VI, respectively.

Table V shows that the two-stage algorithm Faster region-based convolutional neural network (RCNN) has the worst effect. For the object selected in this article, the highest recognition rate of PL is only 82.05% with Faster RCNN, and the recognition rate of SV is 52.96%. The recognition rate of PL based on

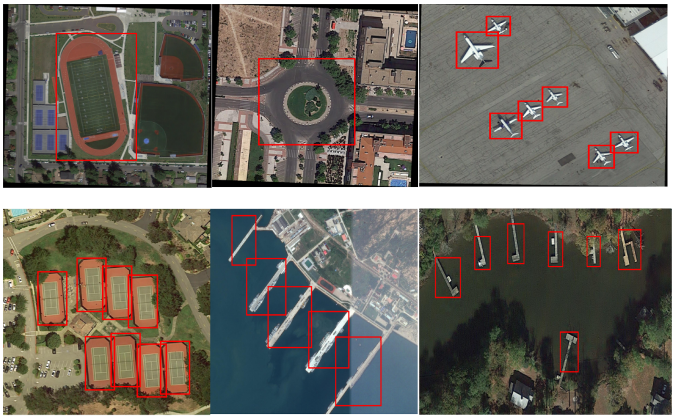

Fig. 13. Part test results with the proposed method.

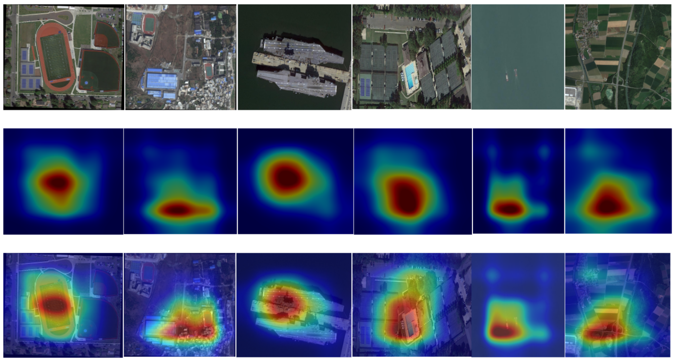

Fig. 14. Partial image interpretation results. First row: original images. Second row: generated intensity map. Third row: results.

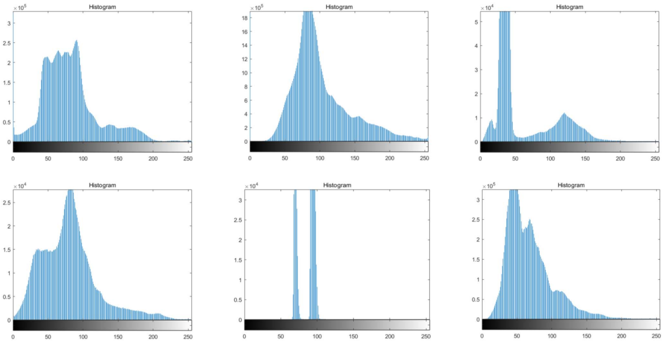

Fig. 15. Histogram of the first row of Fig. 14.

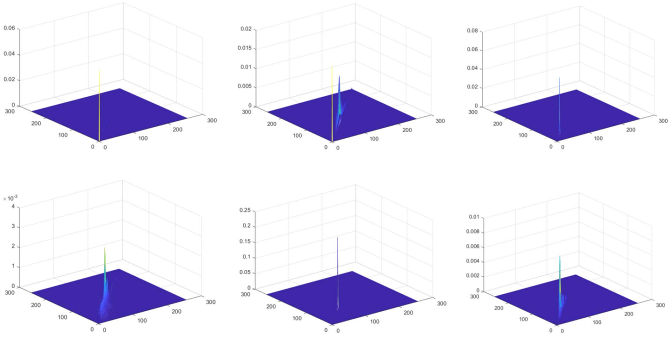

Fig. 16. 2-D histograms of the first row of Fig. 14.

R-FCN is 89.63%, which is 7.58% higher than that based on Faster RCNN. The recognition rate of BR is 53.37%. The highest recognition rate of the first-stage algorithm YOLOv2 and SSD does not exceed 90% because the interference of background features is ignored. In G2Grad-CAMRL, among the three objects with the highest recognition rate, SH, TC, and PL achieve 95.67%, 95.78%, and 94.58%, respectively, which are improved by 6.5%, 3.33%, and 5.67% than that by SSD method, respectively.

From Table VI, all methods have good identification results. For example, the recognition rate of SBF is 70.28% based on CWDL, the recognition rate of SP is 83.01% based on GAN, and the recognition rate of SBF is 74.15% based on SDGH-NET. Based on G2Grad-CamRL, the recognition rate of SBF and SP are 77.39% and 85.88%, respectively. It has a certain improvement over the other three methods.

As can be seen from the comparison in the above tables, the G2Grad-CAMRL remote sensing object detection method is superior to other methods. Good detection results have been achieved on aircraft, small vehicles, large vehicles, ships, etc., indicating that the proposed method has more advantages for the detection of such scenes. Figs. 9 and 10 are the PR and ROC curves for the comparison of GAN, SDGH-Net, and G2Grad-CAMRL.

Fig. 9 is the PR curve trend chart. We only selected three effective methods, including GAN, SDGH-Net, and G2Grad-CAMRL. The area under curve of G2GRAD-CAMRL is 88.64%, which is improved by 1.47% and 0.49% higher than that of SDGH-NET (87.17%) and GAN (88.15%). In terms of the ROC curve, G2GRAD-CAMRL also shows a certain improvement compared with the other two methods. Figs. 11 and 12 are partial enlargements of Figs. 9 and 10, respectively, so that the reader can see the curve trend more clearly. Fig. 13 shows some detection results.

F. Visual Interpretation Effect

To verify the effectiveness of the attention mechanism, Fig. 14 shows the visual interpretation effect of Grad-CAM in the process of generating RSI description text. It can be found that GRAD-CAM, by screening image features, focuses on the highly salient features of the target region, rejects other redundant features and noise information, enhances the perception and understanding of the content of RSIs by the model, and improves the accuracy of description results.

We perform the histogram processing on the first and third rows of Fig. 14 to obtain the results shown below (see Figs. 15 –18). From the point of view of the pixel distribution of the histogram, the pixel distribution of the histogram is denser than that of the original image after processing by the proposed method. This indicates that the sensitive areas of the image can be focused on.

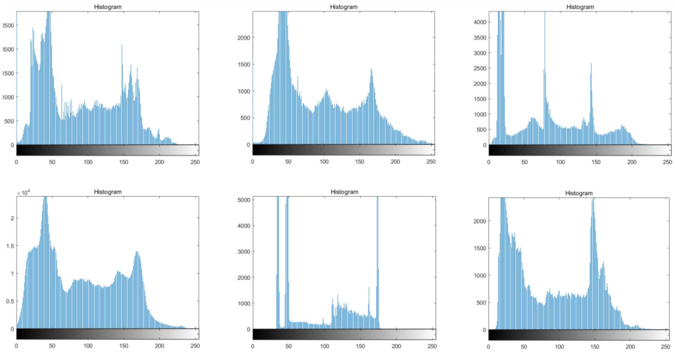

Fig. 17. Histogram of the third row of Fig. 14.

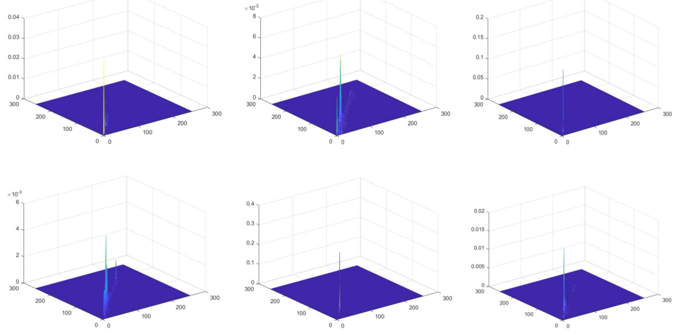

Fig. 18. 2-D histograms of the third row of Fig. 14.

V. Conclusion

To realize the description of RSIs, an RSI description method is proposed by using ResNet to construct the basic network architecture, introducing Grad-CAM, and adopting a reinforcement learning strategy. To verify the effectiveness of the proposed method, the publicly available RSI description dataset is used for training and verification. The experimental results show that the proposed method achieves high accuracy and has good image description performance for RSIs under complex environmental backgrounds, and can realize the interpretation and description of RSIs. In the next step, the model will be improved and optimized to further improve the description performance of RSIs. By specific engineering practice, it will be applied to the aerospace direction.

Conflicts of interest: The authors declare that they have no conflict of interest with respect to the research, authorship and/or publication of this article.

Data availability: The data used to support the findings of this study are available from the corresponding author upon request.

Author contribution: All the authors made contributions to the article in different areas. Shoulin Yin, Liguo Wang, and Muhammad Shafiq conceptualized the study; Liguo Wang and Lin Teng were responsible for investigation; Asif Ali and Lin Teng were responsible simulation; Shoulin Yin and Muhammad Shafiq wrote the original draft; Liguo Wang and Muhammad Shafiq reviewed and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

References

[1] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, “Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3735–3756, 2020.

[2] X. Wang, L. Wang, and Q. Wang, “Local spatial–spectral information-integrated semisupervised two-stream network for hyperspectral anomaly detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5535515.

[3] I. Kotaridis and M. Lazaridou, “Remote sensing image segmentation advances: A meta-analysis,” *ISPRS J. Photogramm. Remote Sens.*, vol. 173, pp. 309–322, 2021.

[4] Y. Bazi et al., “Vision transformers for remote sensing image classification,” *Remote Sens.*, vol. 13, no. 3, 2021, Art. no. 516.

[5] H. Chen and Z. Shi, “A spatial-temporal attention-based method and a new dataset for remote sensing image change detection,” *Remote Sens.*, vol. 12, no. 10, 2020, Art. no. 1662.

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

[7] C. Chen et al., “Object detection in remote sensing images based on a scene-contextual feature pyramid network,” *Remote Sens.*, vol. 11, no. 3, 2019, Art. no. 339.

[8] X. Zhao et al., “Multiscale object detection in high-resolution remote sensing images via rotation invariant deep features driven by channel attention,” *Int. J. Remote Sens.*, vol. 42, no. 15, pp. 5764–5783, 2021.

[9] A. M. Obeso et al., “Visual vs internal attention mechanisms in deep neural networks for image classification and object detection,” *Pattern Recognit.*, vol. 123, 2022, Art. no. 108411.

[10] W. Wang, J. Zhang, W. Zhai, Y. Cao, and D. Tao, “Robust object detection via adversarial novel style exploration,” *IEEE Trans. Image Process.*, vol. 31, pp. 1949–1962, 2022.

[11] H. Feng et al., “SharpGAN: Dynamic scene deblurring method for smart ship based on receptive field block and generative adversarial networks,” *Sensors*, vol. 21, no. 11, 2021, Art. no. 3641.

[12] D. Kumar and X. Zhang., “Improving more instance segmentation and better object detection in remote sensing imagery based on cascade mask R-CNN,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 4672–4675.

[13] D. Avola et al., “MS-faster R-CNN: Multi-stream backbone for improved faster R-CNN object detection and aerial tracking from UAV images,” *Remote Sens.*, vol. 13, no. 9, 2021, Art. no. 1670.

[14] X. Sun et al., “PBNet: Part-based convolutional neural network for complex composite object detection in remote sensing imagery,” *ISPRS J. Photogramm. Remote Sens.*, vol. 173, pp. 50–65, 2021.

[15] G. Cheng, Y. Si, H. Hong, X. Yao, and L. Guo, “Cross-scale feature fusion for object detection in optical remote sensing images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 3, pp. 431–435, Mar. 2020.

[16] J. Chen, L. Wan, J. Zhu, G. Xu, and M. Deng, “Multi-scale spatial and channel-wise attention for improving object detection in remote sensing imagery,” *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 4, pp. 681–685, Apr. 2019.

[17] Y. Li et al., “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] J. Yan et al., “IoU-adaptive deformable R-CNN: Make full use of IoU for multi-class object detection in remote sensing imagery,” *Remote Sens.*, vol. 11, no. 3, 2019, Art. no. 286.

[19] D. Li et al., “A multi-scale fusion convolutional neural network based on attention mechanism for the visualization analysis of EEG signals decoding,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 12, pp. 2615–2626, 2020.

[20] S. Yin and H. Li, “Hot region selection based on selective search and modified fuzzy C-means in remote sensing images,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5862–5871, 2020.

[21] S. Mohtaram et al., “Multi-objective evolutionary optimization & 4E analysis of a bulky combined cycle power plant by CO2/CO/NOx reduction and cost controlling targets,” *Renewable Sustain. Energy Rev.*, vol. 128, 2020, Art. no. 109898.

[22] S. Mohtaram, W. Chen, and J. Lin, “Investigation on the combined Rankine-absorption power and refrigeration cycles using the parametric analysis and genetic algorithm,” *Energy Convers. Manage.*, vol. 150, pp. 754–762, 2017.

[23] Y. Zhong, X. Han, and L. Zhang., "Multi-class geospatial object detection based on a position-sensitive balancing framework for high spatial resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 138, pp. 281–294, 2018.

[24] Z. Deng et al., "Multi-scale object detection in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 3–22, 2018.

[25] Y. Xiao et al., "A review of object detection based on deep learning," *Multimedia Tools Appl.*, vol. 79, no. 33, pp. 23729–23791, 2020.

[26] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6569–6578.

[27] X. Zhou, J. Zhou, and P. Krahenbuhl, "Bottom-up object detection by grouping extreme and center points," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 850–859.

[28] 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.

[29] Y. Wu, K. Zhang, J. Wang, Y. Wang, Q. Wang, and X. Li, "GCWNet: A global context-weaving network for object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5619912.

[30] J. Wang, Y. Wang, Y. Wu, K. Zhang, and Q. Wang, "FRPNet: A feature-reflowing pyramid network for object detection of remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2020, Art. no. 8004405.

[31] Y. Wu, K. Zhang, J. Wang, Y. Wang, Q. Wang, and Q. Li, "CDD-Net: A context-driven detection network for multiclass object detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2020, Art. no. 8004905.

[32] S. Zhou and J. Qiu., "Enhanced SSD with interactive multi-scale attention features for object detection," *Multimedia Tools Appl.*, vol. 80, no. 8, pp. 11539–11556, 2021.

[33] Y. Sun et al., "Aggregating dense and attentional multi-scale feature network for salient object detection," *Digit. Signal Process.*, vol. 130, 2022, Art. no. 103747.

[34] B. Xue and N. Tong, "Real-world ISAR object recognition using deep multimodal relation learning," *IEEE Trans. Cybern.*, vol. 50, no. 10, pp. 4256–4267, Oct. 2019.

[35] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Apr. 2017.

[36] Q. Ming et al., "Dynamic anchor learning for arbitrary-oriented object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 3, pp. 2355–2363.

[37] 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.

[38] X. Zhang et al., "PSNet: Perspective-sensitive convolutional network for object detection," *Neurocomputing*, vol. 468, pp. 384–395, 2022.

[39] M. Akcakaya and A. Nehorai, "MIMO radar sensitivity analysis for target detection," *IEEE Trans. Signal Process.*, vol. 59, no. 7, pp. 3241–3250, Jul. 2011.

[40] W. Ke and D. Huang., "Improving object detection with inverted attention," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2020, pp. 1305–1313.

[41] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[42] M. Du et al., "Towards explanation of DNN-based prediction with guided feature inversion," *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 1358–1367.

[43] F. He, T. Liu, and D. Tao., "Why resNet works? Residuals generalize," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5349–5362, Dec. 2020.

[44] G. Bhattacharya, B. Mandal, and N. B. Puhan, "Multi-deformation aware attention learning for concrete structural defect classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 9, pp. 3707–3713, Sep. 2020.

[45] S. Mohtaram et al., "Evaluating the effect of ammonia-water dilution pressure and its density on thermodynamic performance of combined cycles by the energy-exergy analysis approach," *Mechanika*, vol. 23, no. 2, pp. 209–219, 2017.

[46] S. Mohtaram et al., "Exergy analysis of a multi mixture working fluid absorption refrigeration cycle," *Case Stud. Therm. Eng.*, vol. 15, 2019, Art. no. 100540.

[47] R. Gheisari et al., "Experimental studies on the ultra-precision finishing of cylindrical surfaces using magnetorheological finishing process," *Prod. Manuf. Res.*, vol. 2, no. 1, pp. 550–557, 2014.

[48] L. Brunke et al., "Safe learning in robotics: From learning-based control to safe reinforcement learning," *Annu. Rev. Control, Robot., Auton. Syst.*, vol. 5, pp. 411–444, 2022.

[49] Y. Song et al., "Improved non-maximum suppression for object detection using harmony search algorithm," *Appl. Soft Comput.*, vol. 81, 2019, Art. no. 105478.

[50] S. Karim et al., "Impact of compressed and down-scaled training images on vehicle detection in remote sensing imagery," *Multimedia Tools Appl.*, vol. 78, pp. 32565–32583, 2019.

[51] G. S. Tran et al., "Improving accuracy of lung nodule classification using deep learning with focal loss," *J. Healthcare Eng.*, vol. 2019, 2019, Art. no. 5156416, doi: 10.1155/2019/5156416.

[52] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6541–6549.

[53] G. S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.

[54] Z. Huang et al., "Contrast-weighted dictionary learning based saliency detection for VHR optical remote sensing images," *Pattern Recognit.*, vol. 113, 2021, Art. no. 107757.

[55] X. Li et al., "A deep translation (GAN) based change detection network for optical and SAR remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 179, pp. 14–34, 2021.

[56] Z. Wang et al., "SDGH-Net: Ship detection in optical remote sensing images based on Gaussian heatmap regression," *Remote Sens.*, vol. 13, no. 3, 2021, Art. no. 499.

Shoulin Yin received the M.A. degree in Computer application technology from Shenyang Normal University, Shenyang, China, in 2015. He is currently working toward the Ph.D. degree in information and communication engineering with the College of Information and Communication Engineering, Harbin Engineering University, Harbin, China.

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

Liguo Wang (Member, IEEE) received the M.S. and Ph.D. degrees in signal and information processing from the Harbin Institute of Technology, Harbin, China, in 2002 and 2005, respectively.

From 2006 to 2008, he held a postdoctoral research position with the College of Information and Communications Engineering, Harbin Engineering University, Harbin, where he is currently a Professor. Since 2020, he has been with the College of Information and Communication Engineering, Dalian Minzu University, Dalian, China. He is the author of 2 books in hyperspectral image processing and more than 130 papers in journals and conference proceedings. His main research interests include remote sensing image processing and machine learning.

Muhammad Shafiq received the Ph.D. degree in computer science from the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, in 2018.

He completed his Post-doctorate in 2020 at Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou, China. He is currently a Distinguished Associate Professor with the Cyberspace Institute of Advance Technology, Guangzhou University, Guangzhou, China, and an Adjunct Professor with Shenyang Normal University, Shenyang, China. He is also an Associate Editor at Hindawi, Frontiers, MDPI, IJASE, and CUT University. He has authored more than 70 peer-reviewed articles on topics related to cybersecurity. His current research interests include cybersecurity, network security, IoT security, IoT anomaly and intrusion traffic classification, detection identification, IoT management, network traffic classification, and cloud computing.

Dr. Shafiq is currently an Associate Editor and an Academic Editor at several well-known journals, such as *Wireless Communications and Mobile Computing*. He has been a Lead Guest Editor and a Guest Editor in several Special Issues in well-reported SCI journals. He is the Chair of several conferences. He is an invited keynote speaker and chairperson at several well-known conferences. He is a CP, a PC, and a TPC member of several well-known conferences. He was the recipient of the National Natural Science Foundations of China (NSFC) project on September 29, 2022 for 2023 funded by NSFC (200 000 RMB). In October 2022, he was listed in the top 2% of scientists in the world (according to the recently released list by Stanford University, Stanford, CA, USA. He was also the recipient of the Outstanding Scientific Research Achievements Reward (130 000 RMB) at Guangzhou University in August 2022, and the Outstanding Scientific Research Achievements Reward (30 000 RMB) at Cyberspace Institute of Advanced Technology, Guangzhou University in 2021. He has received multiple awards for Academic Excellence and University Contribution. In 2020, he received the certificate of Appreciation from *Journal Concurrency and Computation: Practice and Experience* by Wiley.

Lin Teng received the M.A. degree in computer application technology from Shenyang Normal University, Shenyang, China, in 2020. She is currently working toward the Ph.D. degree in communication engineering with the College of Information and Communication Engineering, Harbin Engineering University, Harbin, China.

Her research interests include image processing and semantic segmentation.

Asif Ali Laghari received the B.S. and master's degrees in information technology from the Quaid-e-Awam University of Engineering Science and Technology Nawabshah, Nawabshah, Pakistan, in 2007 and 2014, respectively, and the Ph.D. degree in computer science & technology from the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, in 2019.

From 2007 to 2008, he was a Lecturer with the Department of Computer and Information Science, Digital Institute of Information Technology, Pakistan. He is currently a Professor with Shenyang Normal University, Shenyang, China. He authored or coauthored more than 60 technical articles in scientific journals and conference proceedings. His current research interests include machine learning, computer networks, cloud computing, IoT, Fog computing, and multimedia quality of experience management.

Muhammad Faizan Khan received the Ph.D. degree in computer science from Guangzhou University, Guangzhou, China, in 2019.

He is currently an Assistant Professor with the Department of Information Technology, University of Haripur, Haripur, Pakistan.