Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 2701217, 19 pages
https://doi.org/10.1155/2022/2701217

Research Article

Parallel CNN Network Learning-Based Video Object Recognition for UAV Ground Detection

Huanyu Liu,[1] Jiaqing Qiao,[1] Lu Li,[2] Lei Wang,[3] Hongyu Chu,[3] and Qingyu Wang[2]

[1]*School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150001, China*
[2]*Defence Industry Secrecy Examination and Certification Center, Bejing 100001, China*
[3]*Key Laboratory of Space Physics, China Academy of Launch Vehicle Technology, Beijing 10001, China*

Correspondence should be addressed to Jiaqing Qiao; qiaojiaqing@hit.edu.cn

Received 20 July 2021; Accepted 21 March 2022; Published 14 April 2022

Academic Editor: M. Hassaballah

Video object recognition for UAV ground detection is widely used in target search, daily patrol, environmental reconnaissance, and other fields. So, we propose the novel parallel deep learning network with the ability of the global and local joint feature extraction for the UAV video target detection. This paper focuses on solving the problems of feature extraction and target background discrimination required by target discovery to realize target discovery. Break through the key problems of real-time target recognition, such as multiscale targets, high background complexity, many small targets, dense target arrangement, and multidirection, and put forward an optimized network scheme, aiming at the problem of multiscale of image target and aiming at the problem of large change of target scale in image. In the network, the corresponding targets with different sizes and different aspect ratios are matched to make the different targets match the closest, and then, the position of the detection box is fine-tuned by regression. For the special problem of image viewing angle and for the rotation invariance of the airborne down looking image of the target, the usual solution is through data enhancement; that is, through the rotation transformation of the training data, the neural network can learn the rotation invariance of the target. Aiming at the problem of multi-directional image target and aiming at the problems of large target aspect ratio, large target tilt angle, and changeable direction in the target, we propose to use the tilt detection frame instead of the ordinary rectangular detection frame. Aiming at the problem of dense arrangement of image targets and aiming at a large number of densely arranged targets in the image, a feature refining module is proposed, which can effectively improve the detection performance of the detector for densely arranged targets. The experimental results shows that the proposed algorithm achieves more than 10% on the target detection accuracy with focal length change of 1-10 times. The detection accuracy meets the requirements of practical application.

1. Introduction

UAVs have been widely used in ground-to-air photography due to their small size, fast movement speed, wide coverage, etc. In recent years, remote sensing data has been widely used in various fields. Remote sensing images are playing an important role in many aspects, such as geographic information resource exploration [1], ground important information observation, geographic mapping, meteorology, civil and military communication [2], military information detection [3], and sensitive information capture and battlefield situation awareness [4]. In the military field, the automatic recognition of sensitive targets is a very important research direction in military reconnaissance and military early warning. It is also particularly important to integrate automatic target recognition technology into a system with high practicability and good robustness. At present, the realization of information acquisition in domestic remote sensing images is in the stage of transformation from traditional manual judgment to intelligent automatic recognition. Many units at home and abroad are gradually carrying out the work of platform and systematization of remote sensing image target information extraction technology. Many universities and research institutions, such as the University of Littleton, the School of Computer Science of Carnegie Mellon University, and the Department of Geographic

Information of the University of Maryland, have conducted in-depth research on this. Many universities and institutions in China, such as Wuhan University, Central China University of Science and Technology, and the Institute of Applied Sciences of the Chinese Academy of Sciences, have carried out relevant research. At present, the research on the processing and recognition system of remote sensing images at home and abroad, on the one hand, is the customized target recognition system for specific interested targets, such as the ship target recognition system mentioned by Kodors et al. [5], an airport recognition system proposed by Liu et al. [6], and the building group recognition system proposed by Li et al. [7]. Most of these studies are highly targeted, aiming at the specific goal in a specific scene, it has achieved good processing effect, and the generalization of the system is relatively weak. On the other hand, it is a universal remote sensing image automatic processing system with good universality, such as Sahara [8], a semiautomatic image scene understanding system based on multisource remote sensing images developed by Druyts et al. A high-resolution remote sensing image processing system Scorpius [9], developed by Guindon in American research institutions, provides classification, recognition, and tracking functions for some specific military targets. There is also an automatic recognition system for specific target bridges and ports in remote sensing images based on traditional image processing ideas proposed by Xueqiang and Runsheng [10]. In general, domestic related research started late and developed rapidly. The current platform or equipment-specific remote sensing image target recognition system targets fewer categories of research objects, and the universality of the system is relatively weak [11]. There is still a certain distance from the practical application of customization with strong practicability. Aiming at a variety of remote sensing target automatic recognition systems in practical application scenarios, the universality of the system is weak.

From the perspective of imaging, remote sensing image has rich imaging details and single imaging angle, but its imaging scale and illumination change greatly. Affected by weather conditions, there is more and messy ground object information and a large amount of background information in imaging. The basis of remote sensing target recognition is the accurate description of the visual features of the target in the studied remote sensing image and the construction and expression of the prior knowledge of the image target. In recent years, the research on target recognition algorithm in remote sensing images is mainly designed for targets closely related to human activities or military activities, such as roads [12], building clusters [13], aircraft [14, 15], large bridges [16], highways [17], oil tanks [18], and ports [19]. Based on the various methods of target recognition in traditional remote sensing image processing in recent years, it can be found that they can be roughly divided into two cases. They are feature-based method and model-based method.

The basis of model-based remote sensing image recognition method is the construction of remote sensing image target model. The construction of the model often depends on the accumulation of prior knowledge of target and background. The model-based method focuses on the special structural elements, prominent features, or combinations in the target, such as the long and straight runway structure in airport detection, the large parallel line structure in bridge and highway detection, and the dense short lines in building clusters. The abstract modeling of the research goal is completed by constructing special functions and vectors. For example, the remote sensing image road recognition method based on geographic information model proposed was by Barzohar and Cooper [20]. Qi et al. designed the port model and port prior knowledge base in combination with the prior knowledge such as port geometric features and coastline and improved the port recognition algorithm [21]. In addition, some model-based remote sensing images directly detect research targets with the help of energy function, such as Markov random field [20], conditional random field [22], and geometric random field model [23].

Feature-based remote sensing image target recognition technology is a widely studied target recognition technology. Its research basis is the construction and description of target features in the image. Image features include many aspects. One is the recognition based on image gray, texture, color, and other statistical information. Image statistical information is an important part of image features. This part of the research started earlier and more. In 2002, Yijun et al. proposed an automatic detection method of building clusters in aerial images based on gray statistical information and unsupervised clustering in remote sensing images [24], researchers from the German Aerospace Center proposed an automatic road extraction method based on the statistical information of target color in remote sensing images [25], and Min et al. proposed an automatic road network extraction method for high-resolution remote sensing images based on Gaussian Markov random field texture features [26]. The extraction of roads and bridges is realized by using texture statistical information and neural network. Jinzong et al. proposed a fast airport target recognition method using spatial frequency, mean value, variance, energy, and other information after regional screening of images in literature [27]. The second is a target recognition method based on corner features, geometric features, line features, edge features, regional features [28], and some artificially constructed local invariant features [29], such as a ship detection method [30] proposed by Jin et al. in 2014, which combines Harris corner detector and image local significance calculation to accurately construct and describe the characteristics of ships. X. Wang et al. designed a feature construction method based on the combination of image line feature detection and sift operator [31] in document [32], supplemented by tree classifier as decision strategy to realize the detection of airport targets. Yuan et al. proposed a hierarchical classifier. For the method of aircraft target detection in remote sensing image, Haar feature and AdaBoost classifier are used as the preliminary detector, and hog feature and SVM support vector machine are used as the top detector to realize the detection of aircraft target in remote sensing image with resolution of 1 m [33]. In these detection methods, many scholars use one or more

characteristics to express and describe the image and target and use one or more methods to express each characteristic, such as using gray level cooccurrence matrix and LBP local binary mode to describe texture features and then introducing SVM, support vector machine and decision tree, AdaBoost cascade classifier, and other machine learning ideas to assist decision-making and achieve good recognition results.

In 2012, Krizhevsky et al. [34] proposed that alexnet won the championship of the image classification challenge in that year by far surpassing the second place in the Imagenet image classification challenge. The error rate of its top 5 classification decreased by 10% compared with the classification champion in 2011. This excellent performance makes the deep neural network return to the public vision again and once again leads to an upsurge of deep neural network research. Neural network has a long history. Psychologists McCulloch and Pitts proposed MCP neuron model as early as 1943 [35]. Its model has many basic concepts in modern neural networks, such as input parameters, weights, and activation functions. In 1998, Lynet [36] proposed by Lécun et al. was regarded as the pioneering work of convolutional neural network (CNN). The network contains the basic components of modern convolutional neural network structure such as convolution layer, pooling layer, and full connection layer, successfully applied to handwritten digit recognition. Deep neural network has developed rapidly in recent years. In addition to the development and innovation of the network structure, the rapid development of GPU, the great enhancement of hardware computing power, and the explosive growth of network data in the Internet era have made the deep neural network develop rapidly. In 2017, the last Imagenet challenge ended, and the accuracy of the champion of object classification has reached 97.3%. The excellent performance of deep neural network makes it widely developed in other fields. The performance of deep neural network in classification task proves its excellent ability of feature extraction and expression, so it has also attracted extensive research in the field of target detection.

For the target detection task, the network needs to find the position of the object in the input image and give its category at the same time. The early object detection based on deep learning mostly adopts the window drawing method to extract ROI (region of interest). This method is essentially an exhaustive image classification method, which has a large amount of calculation, consumes a lot of computing resources, and has low efficiency. In 2013, Uijlings et al. proposed an image selective search mechanism [37], which uses four kinds of information such as image color, texture, size, and spatial overlap and uses a similar clustering method to divide the image into several regions to generate a waiting area, which greatly reduces the amount of classification calculation. In 2014, Girshick et al. integrated the selective search method into the neural network and proposed the r-cnn network [38], which uses selective search to extract the proposal in the image, which greatly improves the speed and accuracy of target detection. r-cnn network has also become a classic work of deep learning application in the field of target detection. In the same year, some scholars proposed a target detection network based on spatial pyramid pooling [39] spp net. The author applied the idea of pyramid commonly used in traditional image processing to CNN. Thus, the multiscale feature detection in convolutional neural network is realized. In 2015, Girshick proposed an upgraded version of r-cnn fast r-cnn [40], which has greatly improved in speed and accuracy compared with r-cnn. In the same year, Ren et al. further improved the network and proposed the fast r-cnn network [41]. In the network, a very classic RPN network was designed to extract the proposal. The ROI region extraction, feature extraction and expression, candidate region classification, and location refinement were unified into a deep network. Compared with r-cnn, the training time was accelerated by 250 times, and the target detection speed reaches the speed of 5 fps, which achieves the double improvement of speed and accuracy. In addition to the r-cnn series, many excellent deep detection networks form a situation in which a hundred flowers bloom. Redmon et al. proposed an end-to-end detection network Yolo [42] in 2016, which can predict the location reliability and probability of all categories of targets at one time and realize real-time target detection. Dai et al. proposed an r-fcn network in the article [43] published by nips in 2016, which is excellent in speed and accuracy. The map on VOC 2007 and 2012 data sets reached 83.6% and 82%, respectively, and each test image took only 170 ms. Lin et al. proposed a characteristic pyramid type target recognition network FPN [44] in cvpr2017, which greatly improved the problem of low accuracy of small target detection. At the beginning of 2018, the author team of Yolo proposed an improved version of yolo-v3 [45], which not only improved its small target detection accuracy but also improved its speed. The excellent performance of deep neural network in various image recognition competitions proves its good generalization and universality and can extract and describe the characteristics of targets well. In 2017, Neil Lawrence, an Amazon machine learning expert, used Gaussian process. The interpretability of deep learning is a current research hotspot [46]. With the development of deep learning, using depth structure to improve the effect of kernel mapping has become a research trend in recent years. According to the "black box" characteristics of deep learning, analyzing the internal mechanism of the black box and making the deep learning network interpretable is one of the mainstream directions of research [47] [48]. The research on the interpretability of deep learning shows a rapid growth trend, and the growth rate is faster and faster, including both theory and application, reflecting the theoretical value and application value of the interpretability of deep learning, and explaining the importance of the interpretability of deep learning from the side [49]. A neural network structure is formalized into a formula, which provides an idea for the interpretability of deep learning. The attention area of the features extracted by the network is intuitively expressed by visualization [50], which is very useful for the interpretation and traceability of the network reasoning results.

The rest paper is organized as follows. In the Section 2, we present a novel framework for the object recognition, and in Section 3, we have the experiments to compare the proposed algorithms with other methods.

2. Proposed Parallel CNN Network Learning Framework

2.1. Framework. As shown in the remote sensing image, the size of the objects of interest is too small, and the area of interest is too dense, as shown in the remote sensing image. The similarity between similar targets is high, so it is difficult to accurately distinguish individual categories. Therefore, the detection and individual recognition of dense small targets is a great challenge in the field of remote sensing image processing. The traditional target detection and recognition methods rely on the feature expression designed manually, which strongly depends on the professional knowledge and the characteristics of the data itself, and it is difficult to learn an effective classifier from the massive data to fully mine the association between the data. The feature representation and learning ability of deep learning with more abstract and semantic significance can provide an effective framework for target extraction in images. Therefore, the main algorithm of the detection module is the detection algorithm based on deep learning target.

The target detection algorithm based on deep learning has fast detection speed and high detection accuracy. Since the input image is usually smaller than 10000 pixels in the size of the learning network, it is usually required to input the remote sensing image with a fixed pixel size, but it is not required to scale the input image to 10000 pixels in the size of the learning network first. Due to the scale reduction operation of the input and a series of downsampling operations in the convolution neural network, the features extracted from the originally small target after processing by the convolution neural network are not significant, and the target contains only a small number of pixels. This method has caused great problems for the final detection result, and the accuracy of target detection is obviously low. To solve these problems, this project adopts the method of local area circular detection to improve the detection accuracy of deep neural network for small targets. The specific scheme is shown in Figure 1.

The input data of the whole network framework is the local area image with the size of $M * n$ from the upper left corner of the airport remote sensing image. The network only processes a single local remote sensing image at a time, and the total target detection of the complete image is obtained through multiple cyclic processing. Specifically, the whole image is divided into n subimages with $m * N$ scale. Each subimage obtains the target positioning frame, target rough classification, and classification confidence of the local area of the image through the detection process of deep neural network. After N cycles, the processing of the whole image can be completed. The positioning frames of all the acquired targets are mapped back to the original image, and NMS (nonmaximum suppression) is used to merge multiple positioning frames of the same target to realize the complete target detection of the whole image.

2.2. Methods. At present, the mainstream deep learning target detection and recognition networks mainly include fast RCNN, Yolo, and SSD.

2.2.1. Fast RCNN Network. In fast RCNN network, RPN network is used instead of selective search method to obtain candidate areas. The network structure of fast RCNN is shown in Figure 2. By sliding the window on the feature map and then building a neural network for object classification+box position regression, the position of the sliding window provides the general position information of the object, and the box regression provides the more accurate position of the box. Because of the slow speed of the RC NN, the accuracy of the RC NN is very good.

2.2.2. SSD Network. SSD network adopts the basic network structure of vgg16, uses the first five layers, and then uses astrous algorithm to convert FC6 and fc7 layers into two convolution layers. In addition, three convolution layers and an average pool layer are added. Feature maps at different levels are used to predict the offset of default box and the score of different categories. Finally, the final detection results are obtained by NMS. The SSD network structure is shown in Figure 3. SSD network outputs a series of discrete bounding boxes, and these bounding boxes are obtained on different levels of feature maps with different aspect ratios. Therefore, SSD network can not only ensure the accuracy of target recognition but also do not reduce the recognition speed. This paper uses SSD network architecture as the architecture of feature extractor.

The specific structure of the target detection network used in this project is shown in Figure 4. Firstly, in the network backbone structure, it continues the basic network structure of vgg16 network. The first five layers still use the five convolution layers of vgg16 network, abandon the sixth and seventh full connection layers of vgg16 network, and construct two new convolution layers by hole convolution.

2.2.3. Type Spectrum Level Feature Learning Based on Hybrid Attention Mechanism. N1 network has strong high-level semantic information extraction ability, but the type spectrum level fine information extraction ability is weak. We designed the N2 network of the second channel. The network has a main channel and two subchannels. The main channel is used to extract the shallow spatial information and deep abstract semantic information of the input remote sensing image. In order to enhance the interaction between shallow features and deep features and use the deep semantic information of the network to guide the learning of the shallow spatial information of the network, two subchannels are designed. The upper subchannel uses the deep information to guide the shallow information, and the lower subchannel uses the shallow information to guide the deep information and integrates the semantic information of N1 network. It can realize the complementarity of deep semantic information and shallow spatial information. With a shallow network layer design, we can learn fine type spectral level feature information from less type spectral level sample data.

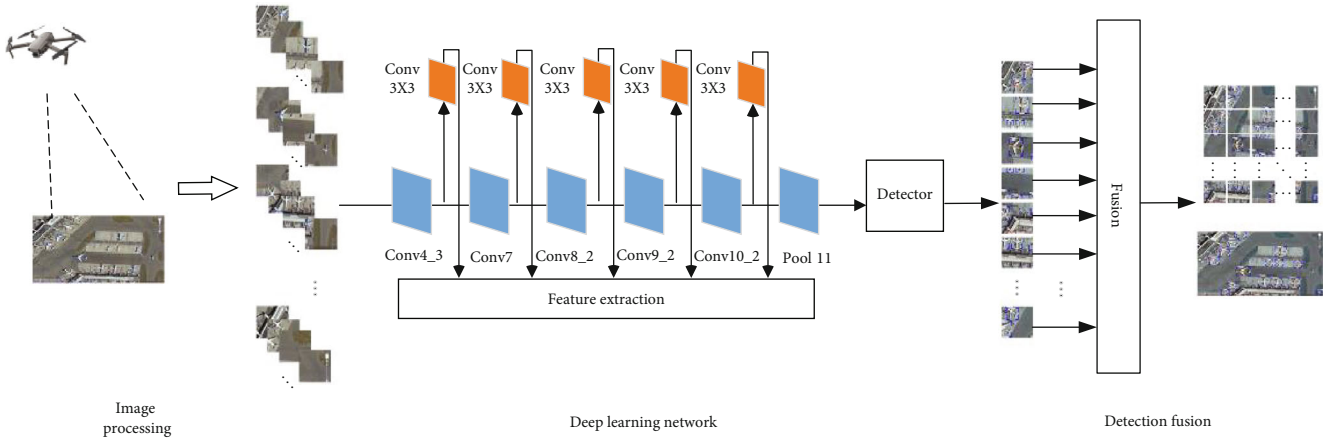


Figure 1: The proposed framework.

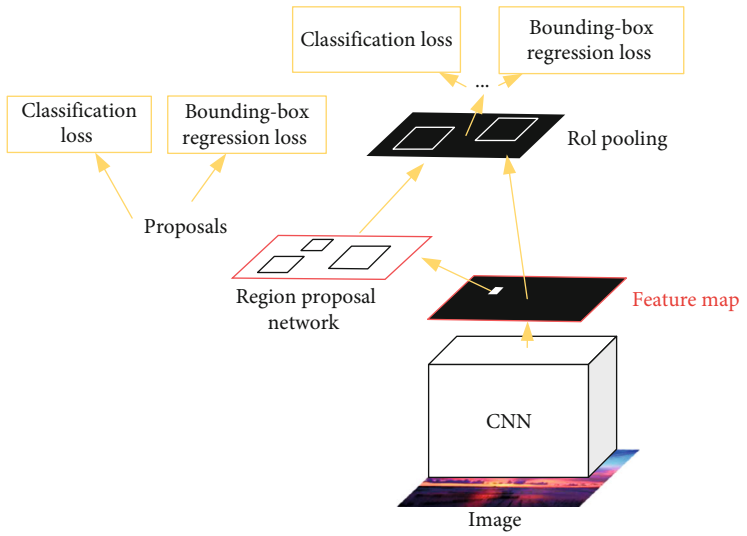
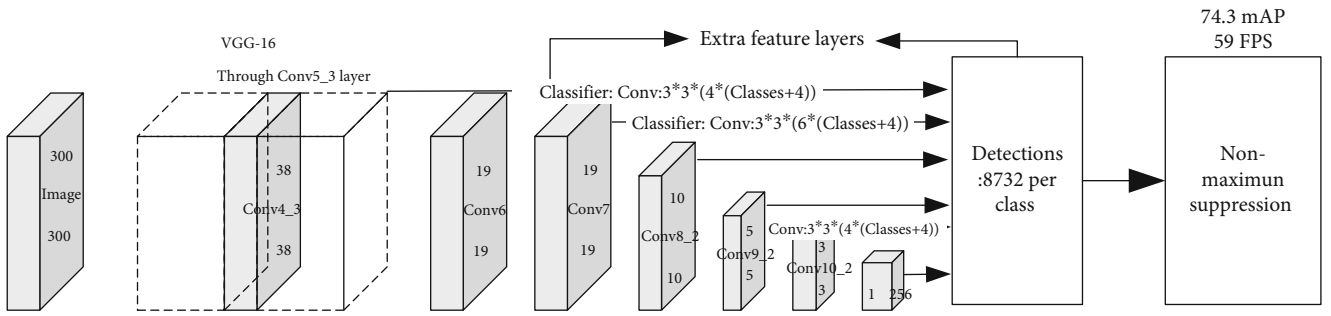


Figure 2: Faster RCNN.



Conv: 3*3*1024 Conv: 1*1*1024 Conv: 1*1*256 Conv: 1*1*128 Conv:1*128 Conv: 1*1*128
Conv: 3*3*512-S2 Conv: 3*3*256-S2 Conv: 3*3*256-S1 Conv: 3*3*256-S1

Figure 3: SSD.

In N2 network, the attention residual module rar based on cycle is designed, as shown in Figure 5. F_i represents the feature map extracted by the current layer, and F_j represents the feature map extracted by other layers. F_j transmits the information to F_i in the form of residual and finally outputs the new feature F_i' of the current channel through the operation shown in the following formula.

$$F_i' = \left(1 + A\left(\mathrm{Cat}\left(F_i, F_j\right)\right)\right) * \left[\Phi\left(\mathrm{cat}\left(F_i, F_j\right)\right) + F_j\right]. \quad (1)$$

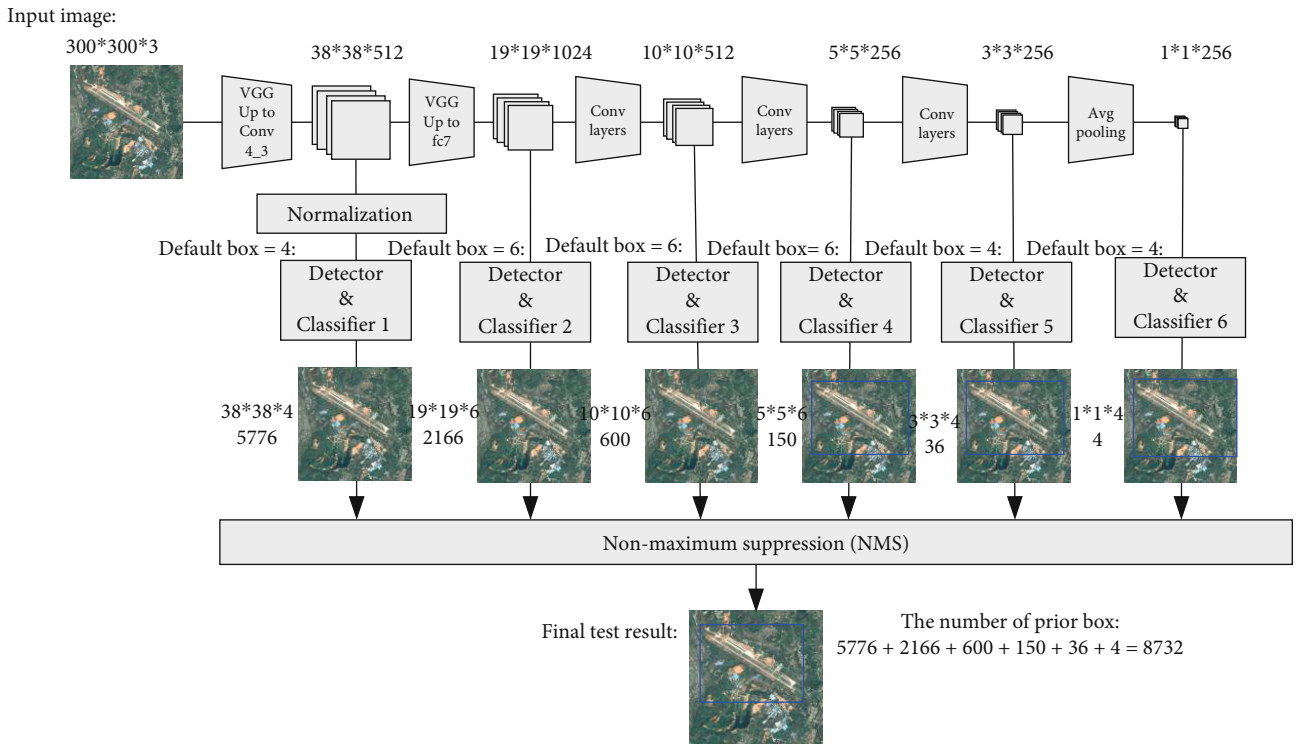


Figure 4: Multiscale detection in the network.

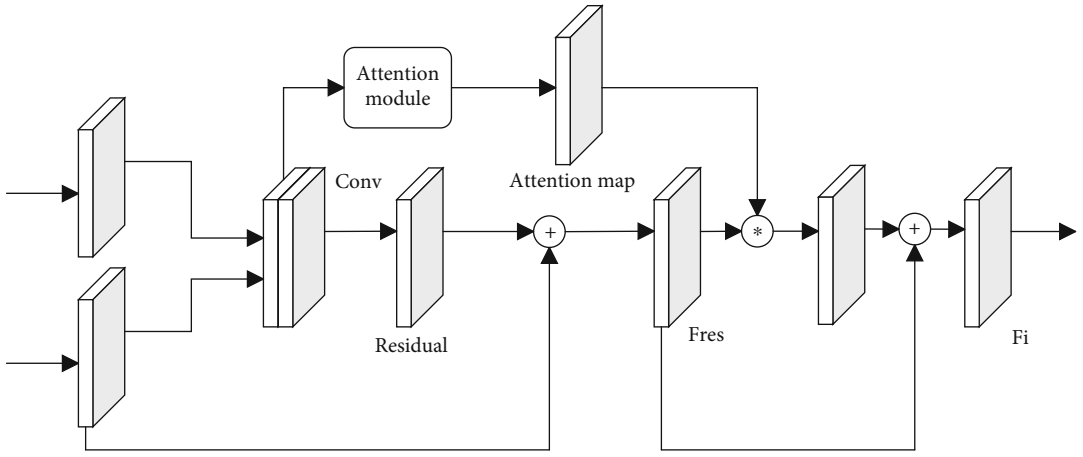


Figure 5: Block diagram of cyclic residual attention module.

The residual structure can better absorb the semantic information of the external layer, retain its own information, and make the network training more efficient. This design uses the small amount of spectral level fine sample data and a large amount of semantic sample data constructed by us to train our network at the same time and uses the information of semantic network to guide the learning of the characteristics of each layer of spectral level network, so as to train our network semisupervised with the data set constructed by us and better learn the spectral level characteristics of remote sensing images and realize the purpose of automatic fine annotation of remote sensing image spectral level.

Then, the width and height of each priority box can be calculated by

$$w_k^a = s_k\sqrt{a_r}, \tag{2}$$

$$h_k^a = s_k/\sqrt{a_r}, \tag{3}$$

where S_k is a parameter of each layer, and its calculation formula is shown in

$$S_k = s_{\min} + \frac{s_{\max} - s_{\min}}{m-1}(k-1), k \in [1, m]. \tag{4}$$

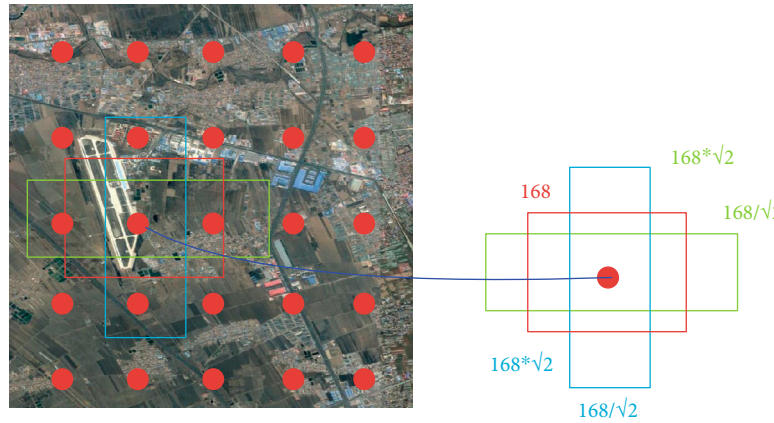


Figure 6: Generation of candidate frames in the network.

For a ratio of 1, that is, an aspect ratio of 1, two candidate boxes with an aspect ratio of 1 are generated around each anchor point, and use $s_k' = \sqrt{s_k s_{k+1}}$ extra to generate a box with an aspect ratio of 1. In this way, for each anchor point, you can get 6 different boxes, see Figure 6.

$$L_{\text{conf}}(x, c) = -\sum_{i \in Pos}^{N} x_{ij}^{p} \log\left(\hat{c}_i^{p}\right) - \sum_{i \in Neg} \log\left(\hat{c}_i^{0}\right), \tag{5}$$

$$\hat{c}_i^{p} = \frac{\exp\left(c_i^{p}\right)}{\sum_p \exp\left(c_i^{p}\right)}. \tag{6}$$

So, its loss function is shown as

$$\begin{cases} L_{loc}(x, l, g) = \sum_{i \in Pos}^{N} \sum_{m \in \{cx, cy, w, h\}} x_{ij}^{k} \text{smooth}_{L1}\left(l_i^m - \hat{g}_j^m\right), \\ \hat{g}_j^{cx} = \left(g_j^{cx} - d_i^{cx}\right)/d_i^{w} \quad \hat{g}_j^{cy} = \left(g_j^{cy} - d_i^{cy}\right)/d_i^{h}, \\ \hat{g}_j^{w} = \log\left(\frac{g_j^{w}}{d_i^{w}}\right) \quad \hat{g}_j^{w} = \log\left(\frac{g_j^{w}}{d_i^{w}}\right). \end{cases} \tag{7}$$

The total loss function in the network is the weighted sum of the above two loss functions as shown as

$$L(x, c, l, g) = \frac{1}{N}\left(L_{\text{conf}}(x, c) + \alpha L_{\text{loc}}(x, l, g)\right), \tag{8}$$

where N is the number of positive samples.

2.3. Network Optimization. In order to improve the ability of multimodal image feature extraction, mosaic data enhancement method is introduced in data enhancement. The data enhancement method must consider the characteristics of data set and task, so that the enhanced image is different from the original image without damaging the information contained in the sample. The schematic diagram of mosaic data enhancement method is shown in the figure. Mosaic data enhancement method samples four images from the data set at a time and places the four images in the upper left, upper right, lower left, and lower right corners of the new image, respectively, so as to generate a new image with an area four times that of the original image. Then, the synthesized image is subjected to conventional data enhancement methods such as perspective transformation, flipping, color gamut change, and so on. Finally, the center of the synthesized image is cropped to restore the size of a single image. For all the detection frames in the four pictures, the useless detection frames are removed according to the number of the remaining parts, length width ratio, and area after cutting. Finally, the remaining detection frames are translated to the correct position to complete the data enhancement process.

Mosaic data enhancement method makes use of the flexibility of target translation in target detection task to combine four images into one. Through mosaic method, a large number of new samples can be generated in a limited data set while maintaining the distribution law of the original samples, so as to maintain the fidelity and improve the richness of the samples at the same time. The target detection network can be divided into backbone network part, feature fusion part, and prediction part. In the backbone network part, the newly proposed cspnet is introduced. Cspnet is a new network design concept. The design purpose of cspnet is to enable the network to obtain richer gradient fusion information on the premise of reducing the amount of calculation. By separating the gradient flow, the gradient flow can propagate on different network paths. By transforming concat and transition operations, the gradient flow after propagation will have great correlation differences. In addition, cspnet can greatly reduce the amount of calculation and improve the reasoning speed and accuracy. Cspnet first divides the feature map into two parts according to the channel and only does the original convolution network operation for one part. Then, connect the results of some features through the convolution network with the previous feature map, and finally get the final result through convolution fusion. This cross phase partially connected network can further improve the performance of the backbone network. In the part of feature fusion, the panet structure which can balance the accuracy and speed is selected. Target detection network usually needs to detect targets with different spatial scales in the same detection scene at the same time,

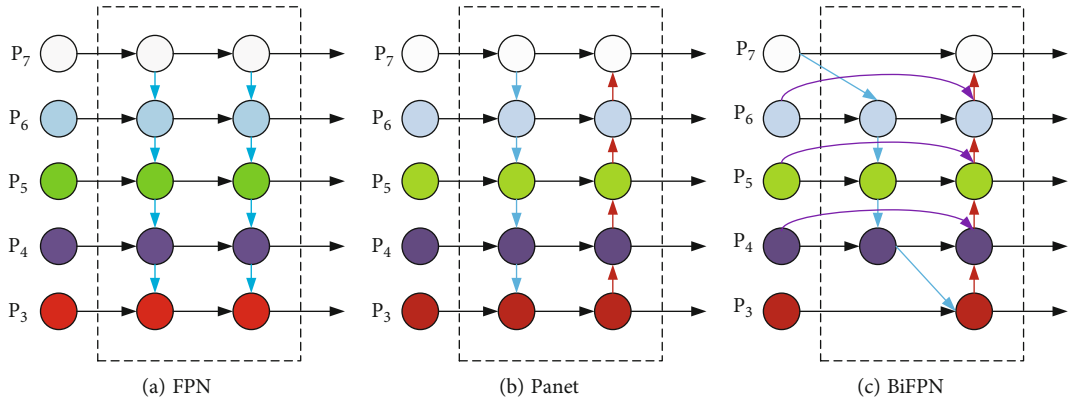


Figure 7: Comparison of different feature fusion methods.

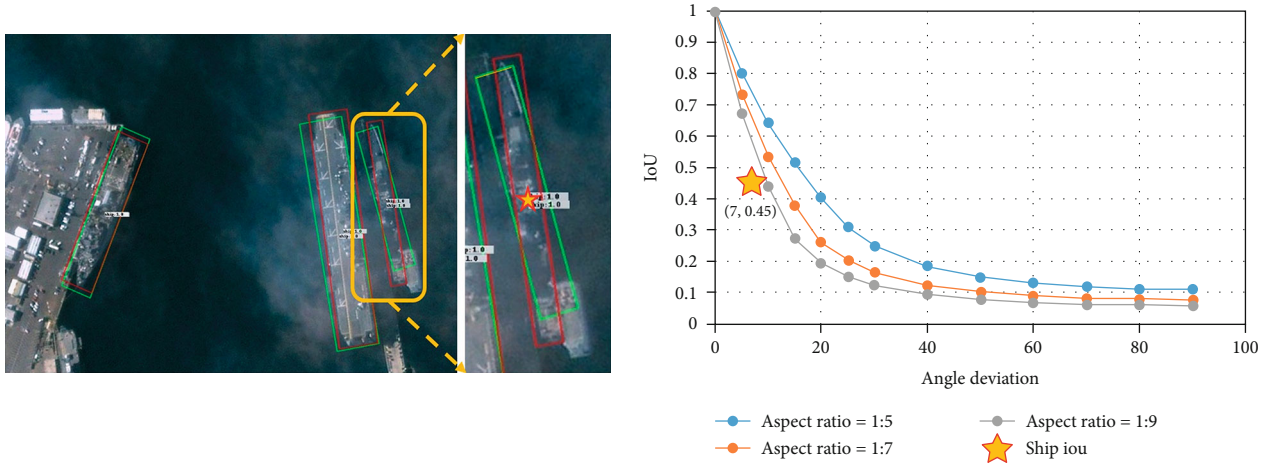


Figure 8: Ship target IOU is sensitive to the angle change of detection frame.

so it is necessary to design a network structure that can deal with multiple scales at the same time. Generally speaking, the shallower the CNN model is, the larger the scale of the feature map is, and more local features can be captured. The deep feature map has a smaller scale and can capture more semantic features. Because the deep feature map lacks local feature information, it is not conducive to the detection of small targets. Using feature fusion can solve this problem to a certain extent. Finally, in the prediction part, the sigmoid function is used to map the location information output by the network to 0~1 instead of the traditional exponential function. At the same time, it is easy to predict the difficulty of the half axis when the divergence function is unstable. The sigmoid function has saturation characteristics and converges to 1 on the positive semiaxis, which makes the network training and prediction more stable. However, the saturation characteristic of sigmoid function will make it difficult to regress the position close to the lattice point. Therefore, the scaling coefficient is used to solve this problem, as shown in the formula.

$$b_x = 2\sigma(t_x) + c_x, \tag{9}$$

$$b_y = 2\sigma(t_y) + c_y. \tag{10}$$

Efficientdet summarizes the previous methods on multiscale feature fusion and proposes bifpn, which can fuse multiscale features with a small number of parameters. This paper holds that the traditional FPN only has a bottom-up feature fusion path, which is not conducive to the information fusion of large-scale feature map. Panet adds a top-down path to improve the detection accuracy, as shown in Figure 7. In the past, the most important feature fusion method is adding or connecting. The amount of calculation is small, but the fusion effect is not good. The connection mode has a large amount of calculation due to the increase of the number of channels. Bifpn proposes a new feature fusion method—weighted addition—that is, the two feature maps are multiplied by their respective weights and then added. This method balances the speed and accuracy, and the effect is good.

Aiming at the problems of large target aspect ratio, large target tilt angle, and changeable direction in remote sensing targets, we propose to use tilt detection frame instead of ordinary rectangular detection frame. When detecting the target with large inclination angle and large aspect ratio, the ordinary rectangular detection frame has to cover the target with a rectangular frame much larger than the target, while most areas in the actual detection frame are background, which is particularly unfavorable for remote sensing target detection. As

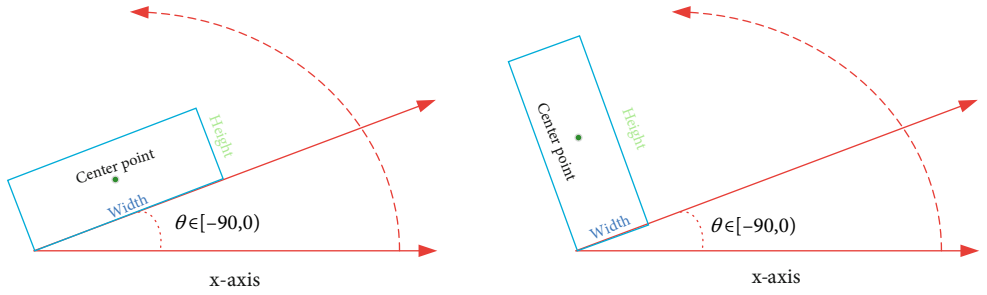


Figure 9: Tilt detection box.

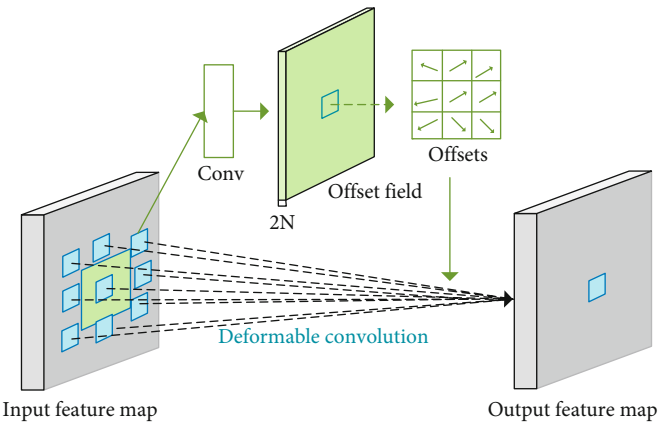


Figure 10: Composition of deformable convolution.

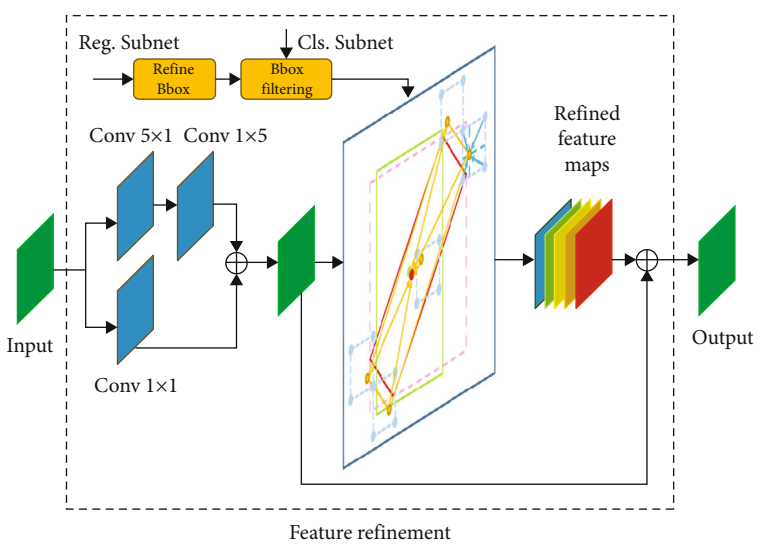


Figure 11: Architecture of feature refining module.

shown in Figure 8, the influence of ship target detection frame IOU on rotation angle deviation is shown.

The detection frame with large aspect ratio is very sensitive to the direction of the target. Even if the rotation angle of the detection frame changes slightly, it will lead to a serious decline of IOU, while the aspect ratio of ships and other types of targets generally reaches more than 5. R3det uses tilt detection frame to solve this problem. As shown in Figure 9, it is a tilt detection frame, which has one more angle parameter than the ordinary rectangular detection frame, and the value is between. The author uses IOU smooth L1 loss to regress the parameters, and the formula is as follows, which has achieved good results.

$$L = \frac{1}{N}\sum_{n=1}^{N} t'_n \sum_{j\in\{x,y,w,h,\theta\}} \frac{L_{\text{reg}}\left(v'_{nj}, v_{nj}\right)}{\left|L_{\text{reg}}\left(v'_{nj}, v_{nj}\right)\right|}\left|-\log\left(IoU\right)\right|, \quad (11)$$

Figure 12: Some examples from data set.

where $L_{\text{reg}}(v'_{nj}, v_{nj})$ is smooth L1 loss and t'_n is the indicator vector of foreground and background. When the label is foreground, it is 1. $L_{\text{reg}}(v'_{nj}, v_{nj})/|L_{\text{reg}}(v'_{nj}, v_{nj})|$ is the direction of gradient descent, and $|-\log(IoU)|$ is the mode of gradient descent; that is, a larger gradient is applied to the detection frame with larger deviation. Using IOU as the loss model ensures the consistency with the evaluation index.

At the same time, r3det also subdivides anchor into different rotation angles, which increases the detection accuracy of targets with various tilt angles. Finally, with resnet-50 as the backbone, r3det reached 70.16% of map on dota data set.

Aiming at the problem of small target in image and aiming at the problem that it is difficult to detect small target in image, the conventional solutions mainly include increasing the resolution of input image and reusing network shallow features for small target detection. Then, these two methods will greatly increase the amount of calculation of the network. Through experiments, it is found that the IOU of small targets to anchor is usually very low. Therefore, the anchor threshold for small targets should be appropriately lowered so that small targets can match more anchors. In addition, using data enhancement can also improve the ability of the network to detect small targets and appropriately oversampling the pictures containing small targets. In yolov3, there is a weighting term for the target, and the author deliberately balances the weight between the large and small targets. We can increase this weight to further improve the importance of small targets, so as to improve the detection accuracy of small targets.

Aiming at the rotation invariance of airborne down looking image of remote sensing target, the usual solution is to enhance the data; that is, make the neural network learn the rotation invariance of the target by rotating the training

Figure 13: Imaging effect and actual shape of ten types of combat vehicle targets in SAR image.

data. Because the ordinary convolution operation has no rotation invariance, a deformable convolution method is proposed to replace the ordinary convolution in recent years, and the ability to learn spatial geometric deformation is introduced into the convolution neural network for the first time. The original convolution operation will sample

Table 1: Classification results of depth mapping targets.

Targets	Training	Testing	SVM	Proposed
2S1&BRDM_2	1000	1000	89%	92.9%
BRDM_2&ZSU234	1000	1000	89.7%	90.4%
2S1&BRDM_2	400	400	91.7%	92%
D7&T62	400	400	100%	100%
T62&2S1	400	400	85%	85.5%

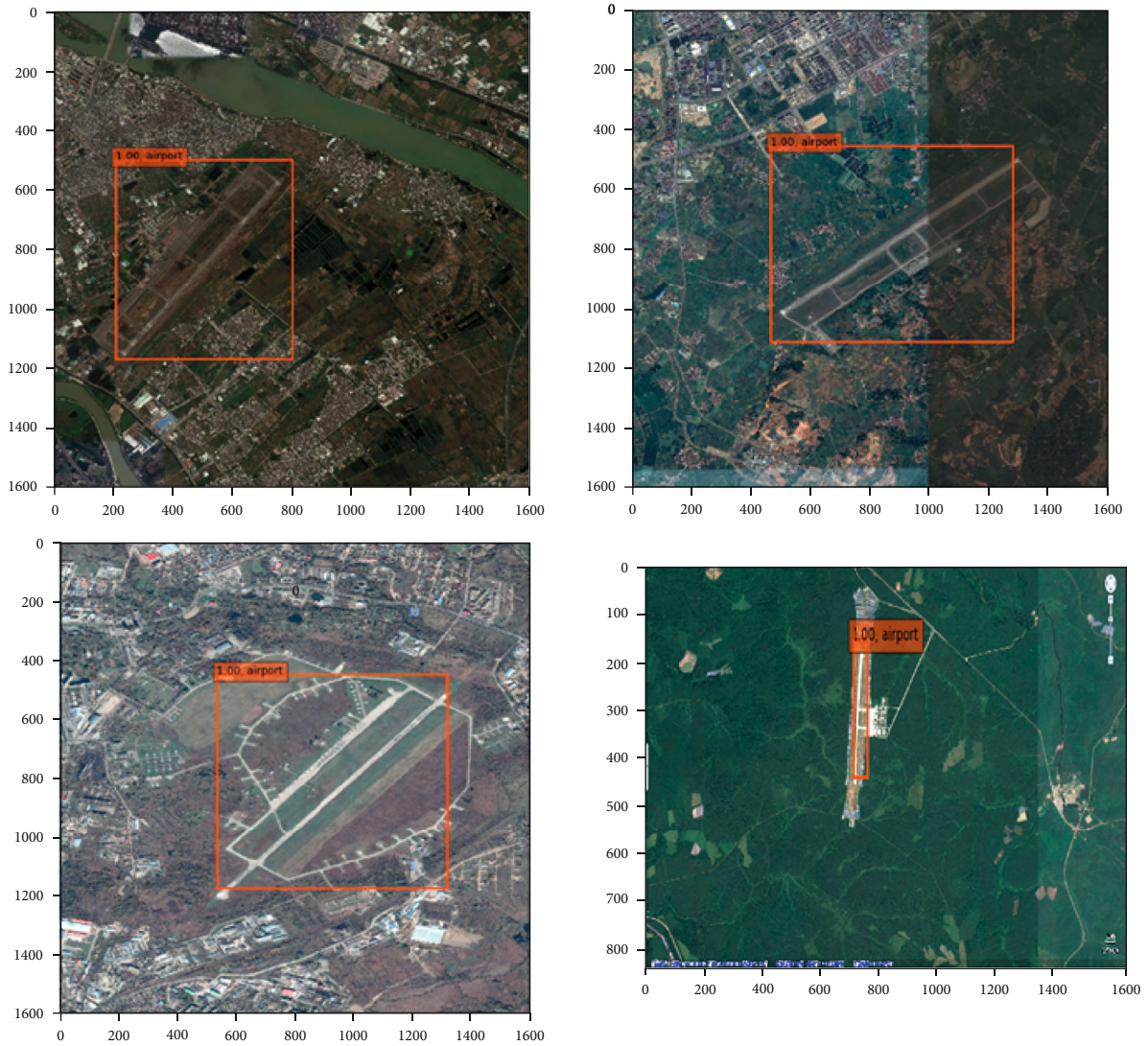


Figure 14: Airport identification results.

based on the regular grid position at each position of the input image, and then convolute the sampled image value and output it as the position. Deformable convolution is to learn the offset sampling position through a group of convolutions, as shown in Figure 10.

Deformable convolution first obtains a convolution result with twice the number of channels as the original result through convolution (due to the need to predict the deviation in X direction and Y direction). This convolution result is to predict the deviation between the convolution input source position and the original convolution position. The convolution result is taken as the pixel deviation and added with the pixel position of the original convolution input to obtain the input pixel position of the deformed convolution. Since the pixel position is a floating-point number at this time, the pixels adjacent to each position need to be bilinear interpolated to obtain the input pixel value of deformable convolution. Finally, the result of deformable convolution can be obtained by convolution. The advantage of deformable convolution is that it can model the rotation

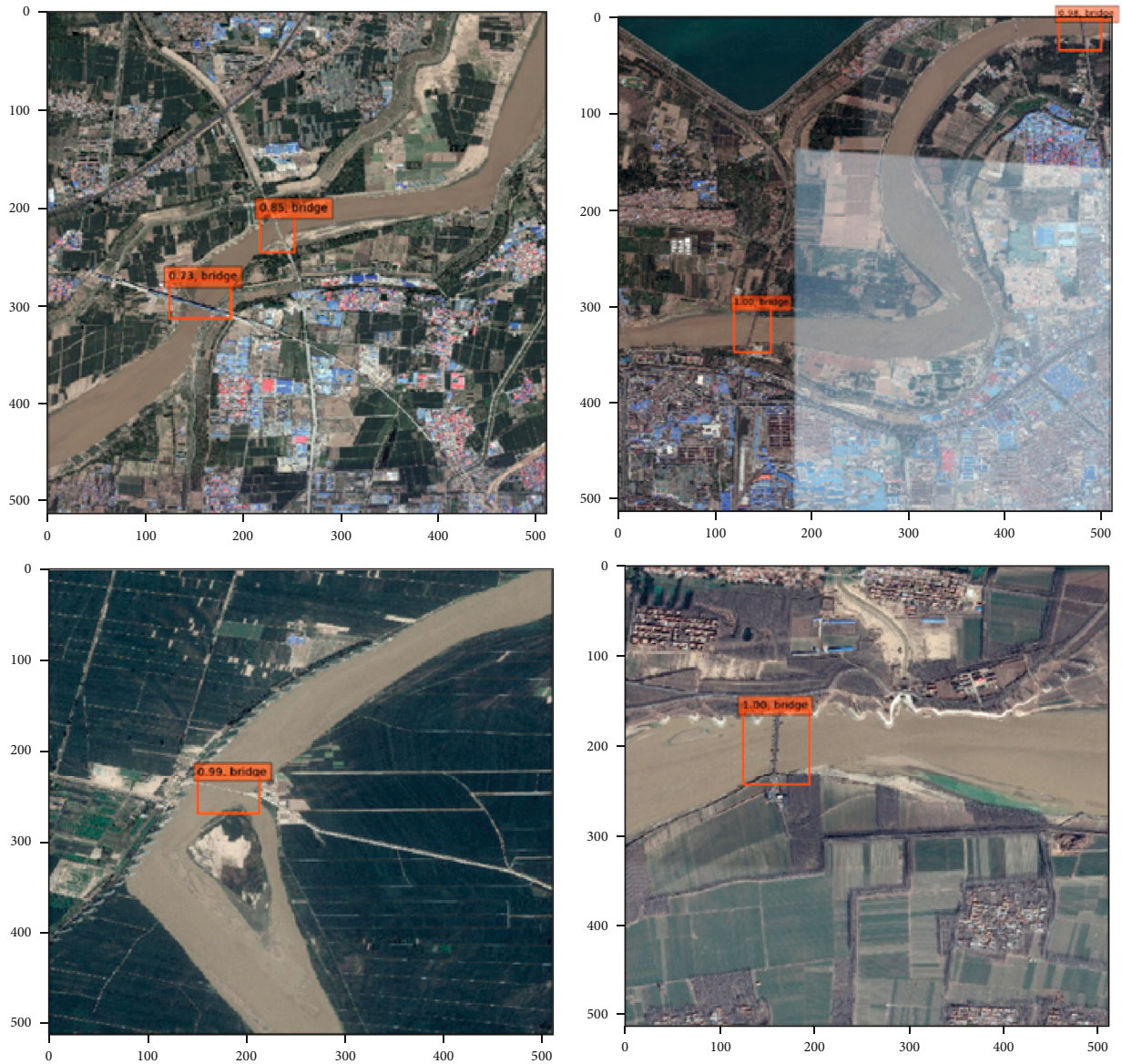


Figure 15: Bridge identification results.

offset of features, while the disadvantage is that it increases the amount of calculation of the model. Therefore, deformable convolution is usually added only at the position close to the output of the network.

The deformable convolution network uses resnet-101 as the backbone. The original convolution is replaced by 3×3 deformable convolution at the last 1, 2, 3, and 6 layers of RESNET, respectively. The map is increased by 3.2% at most on the coco data set. With the use of more deformable convolution, the performance of the model is gradually improved, but the speed of the model is also significantly reduced.

For the problem of dense arrangement of image targets and for a large number of densely arranged targets in remote sensing images, r3det proposes a feature refining module, which can effectively improve the detection performance of the detector for densely arranged targets. The main idea of the feature refining module is to collect the feature vectors at the four corners corresponding to each feature pixel, so as to enrich the feature information representing the target at the center. The architecture diagram of feature refining module is shown in Figure 11.

In the feature refining module, firstly, the input feature map is convoluted with 1×1 after two asymmetric convolutions of 5×1 and 1×5. Select the detection box whose confidence probability is higher than the threshold from the currently predicted detection box. For each pixel of the feature map, find the coordinates of the five points (center point and four corners) of the detection frame predicted by the pixel, calculate the feature vector there with the interpolation algorithm, and sum it as the feature vector of the new feature map. After traversing the entire feature map, the original feature map and the new feature map are added as the final output.

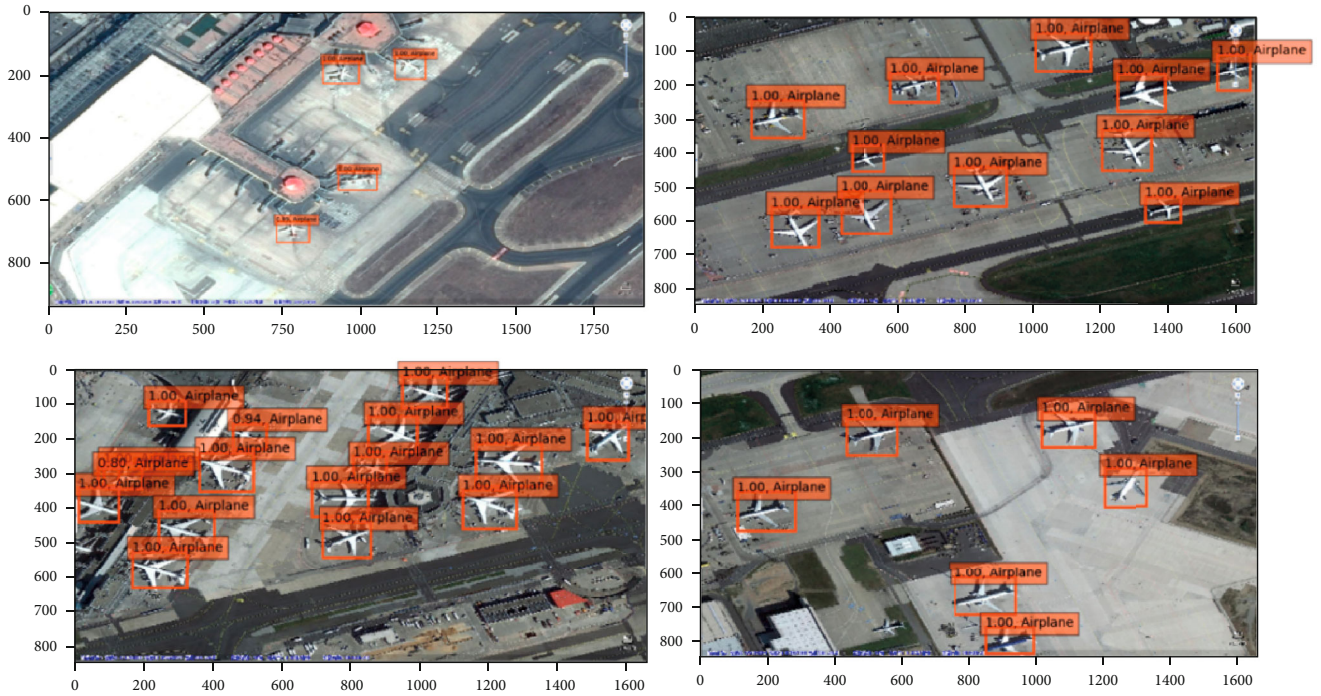


Figure 16: Aircraft identification results.

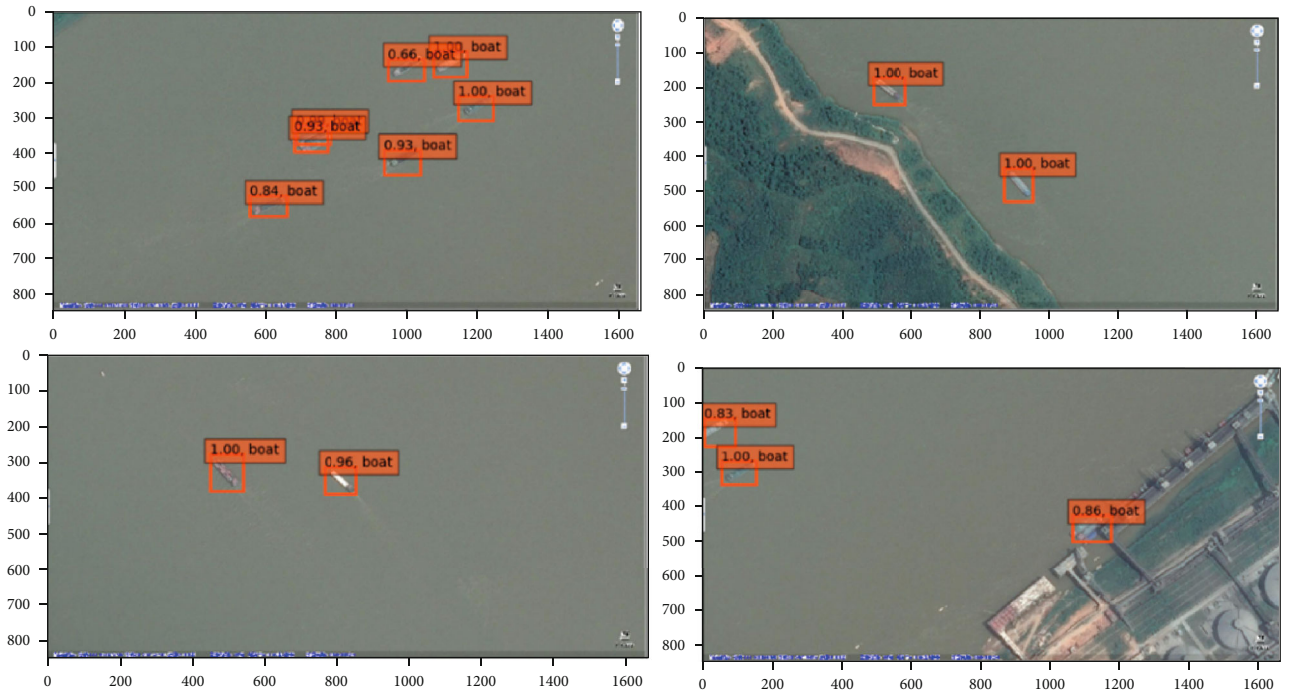


Figure 17: Ship identification results.

The feature refining module can be used alone or in cascade. When multiple modules are used in cascade, the threshold of the latter module is higher than that of the previous module, which can continuously screen out the detection frames with low confidence probability and integrate and refine the characteristics of the detection frames with high confidence probability. When there are a large number of dense side-by-side targets in the detection image, a large number of prediction frames will appear in the same target attachment. Through the feature refining module, most of the detection frames with low confidence probability can be removed and the rest can be refined. Therefore, it can not only improve the detection ability of densely arranged targets but also improve the detection speed.

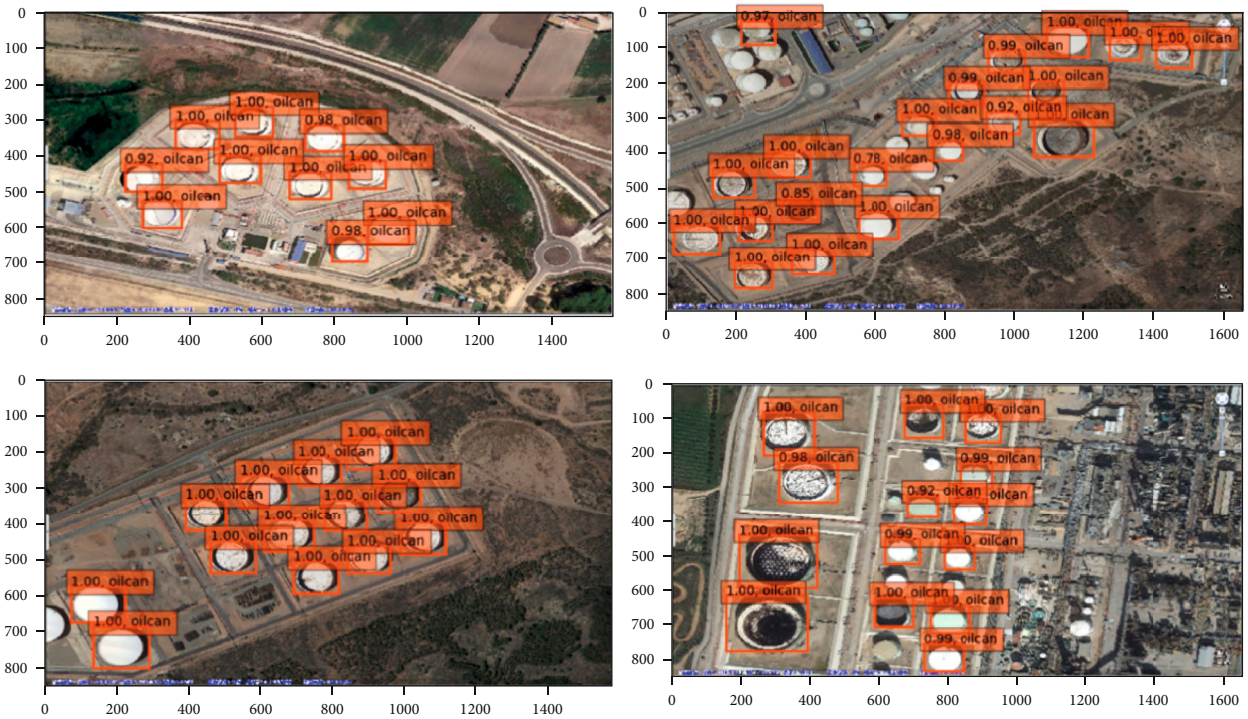


Figure 18: Oil tank identification results.

Figure 19: Target distribution recognition results.

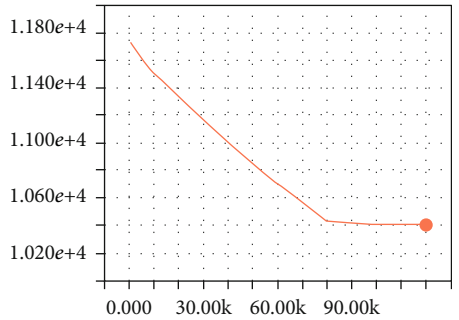


Figure 20: Compression stage L2_ loss_ 1 change curve.

3. Experimental Verification of Algorithm Performance

3.1. Experimental Data. A total of 760 source images were collected. The resolution is 0.5 m. A total of 1121 images including targets such as < port > were collected from port source data, with a resolution of 0.5 m. The collected < oil tank > target images are 900 high-resolution images with a resolution of 0.5 m. The collected < ship > target images are 533 high-resolution images with a resolution of 0.5 m. For the target type data of < Airport >, 500 pieces of airport data with a resolution of more than 6 m were collected. For

Table 2: Experimental results based on channel pruning.

Sheared layer	Pruning rate	Pruning time	Proportion parameters	Model accuracy
NULL	0	—	—	77.64%
VGG16	0.2	16 h	67.34%	77.43%
VGG16	0.3	20 h	53.58%	77.05%
VGG16	0.4	36 h	41.63%	75.47%
VGG16	0.5	40 h	31.56%	73.29%
All layers	0.2	18 h	66.50%	76.98%
All layers	0.3	22 h	52.32%	76.42%
All layers	0.4	38 h	39.96%	74.89%
All layers	0.5	40 h	29.47%	72.48%

Table 3: Experimental results of transplantation based on parameter quantification model.

Data set type	Reasoning time	Model accuracy
VOC Ref	1200 ms	75.98%
VOC Qua	220 ms	75.87%
Ref	1000 ms	86.77%
Qua	200 ms	86.24%

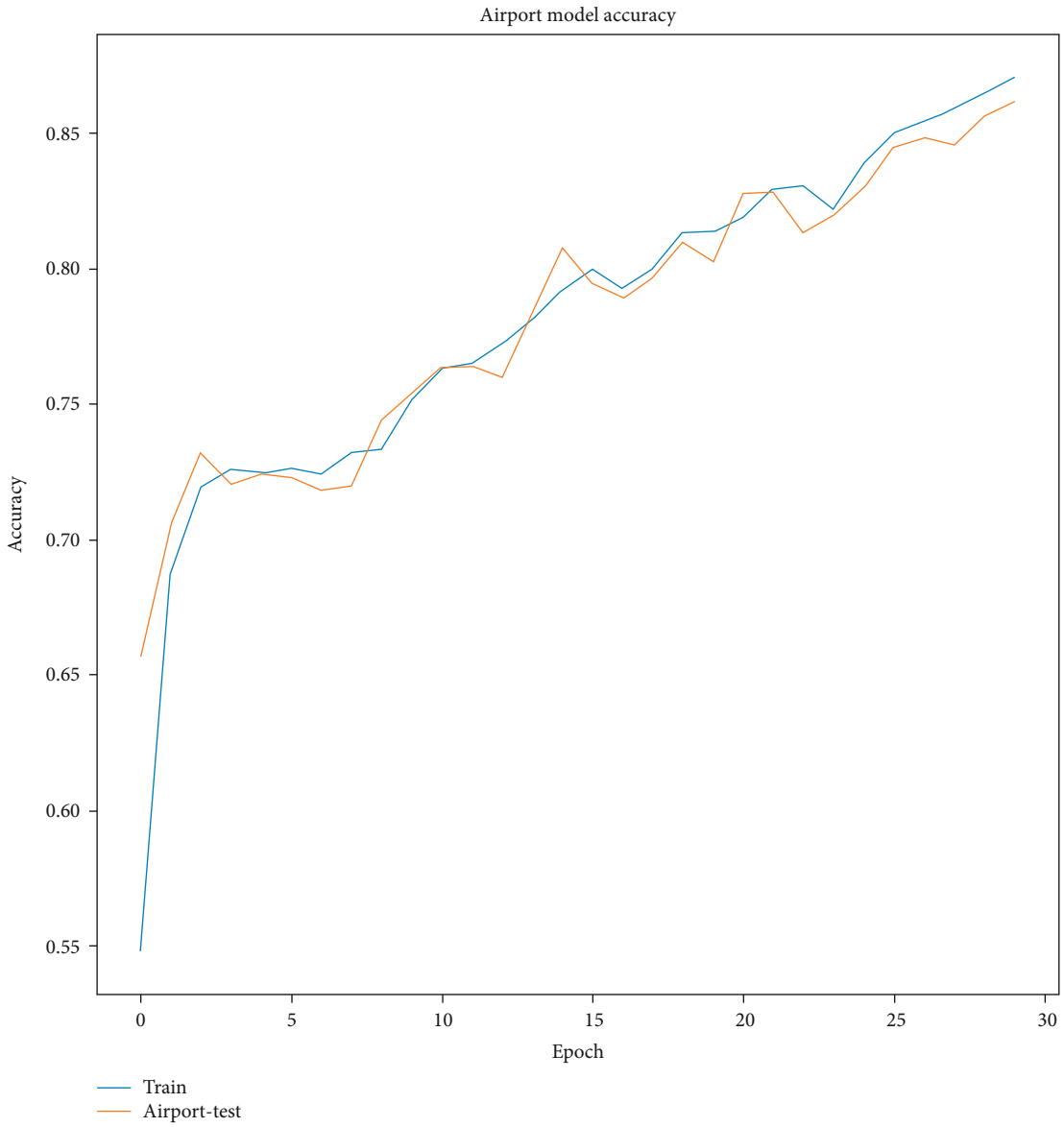


Figure 21: Accuracy of airport target recognition.

the target of < bridge >, a total of 558 source data were collected, with a resolution of 6 m or more. Collect 1080p visible video data of airborne down view for more than 1 hour. Some examples from the data set are shown in Figure 12. The SAR image target is used as the verification object, and ten types of combat vehicle targets at different angles are selected as the data set. Firstly, the original images in MSTAR data set are preprocessed and cut to the same size, and then, 1000 images are selected as the training set and 1000 images as the test set. In the follow-up experiment, in order to test the influence of the number of samples in the data set on the feature extraction results, a comparative experiment was carried out with 400 training sets and 400 test sets as the data objects under the same other conditions. Imaging effect and actual shape of ten types of combat vehicle targets in SAR image are shown in Figure 13.

3.2. Experimental Results. In order to verify the feature extraction performance of deep mapping, the feature extraction effects of the two modes are verified by constructing different data sets. Finally, the target classification results after feature extraction of deep mapping are compared with other common methods. Firstly, the feature extraction effect is verified through the classification task. Ordinary SVM only supports the two classification method. Therefore, in the experiment of verifying the performance of depth kernel mapping, this part also adopts the task form of two classification, carries out two-to-two combination for different types of chariot targets, and then constructs the data set according to the combination structure. Verify the effect of depth kernel mapping structure feature extraction in classification under two different data set sizes. The results are shown in Table 1.

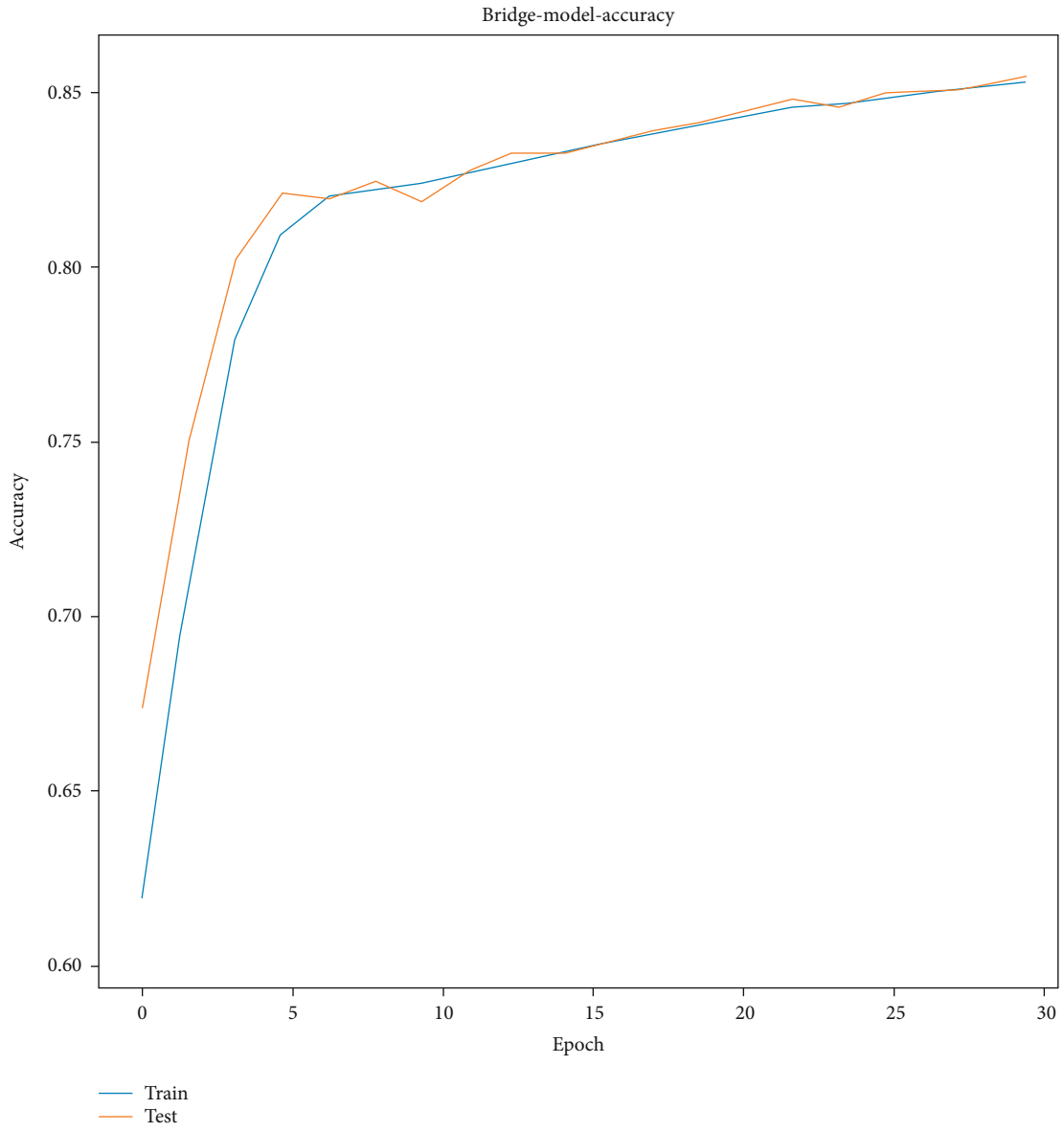


Figure 22: Accuracy of bridge target recognition.

It is used for the recognition ability of airborne/spaceborne visual targets, including image targets such as vehicles, aircraft, bridges, airports, ports, oil tanks, ships, and target cloth, with an overall recognition rate of more than 85%.

The results of airport, bridge, aircraft, ship, oil tank, and target distribution are separately shown in Figures 14–19.

In the experiment, the original data set shall be used for retraining after each layer of pruning, until the training is stopped when the accuracy of the model after pruning reaches the accuracy of the model without pruning, or the accuracy does not decline after retraining. Figure 20 is a schematic diagram of the variation curve of L2 loss function in channel pruning.

Backbone only (trunk pruning, i.e., the pruned layer in Table 2 is the part of vgg16 feature extraction) and all layers are determined. After the two strategies of layers, the pruning rate of VGg model channel is preliminarily set to 0.2, 0.3, 0.4, and 0.5. The specific experimental results are shown in Table 2.

In the selection of quantitative model required by the experiment, SSD target detection model based on public data set voc207 and self-built airborne down looking data set is selected. After the upper computer quantifies the model parameters, the unquantified model and the model based on parameter quantization are transplanted to the selected NVIDIA TX1 platform for testing. The experimental results are shown in Table 3.

According to the experimental results, it can be found that after the model trained by the upper computer is transplanted to the embedded platform, due to the limitation of computing resources of the embedded platform, the model reasoning time is calculated in seconds. However, the reasoning time of the quantified model is about 10 times faster than that of the nonquantified model, but the accuracy of the

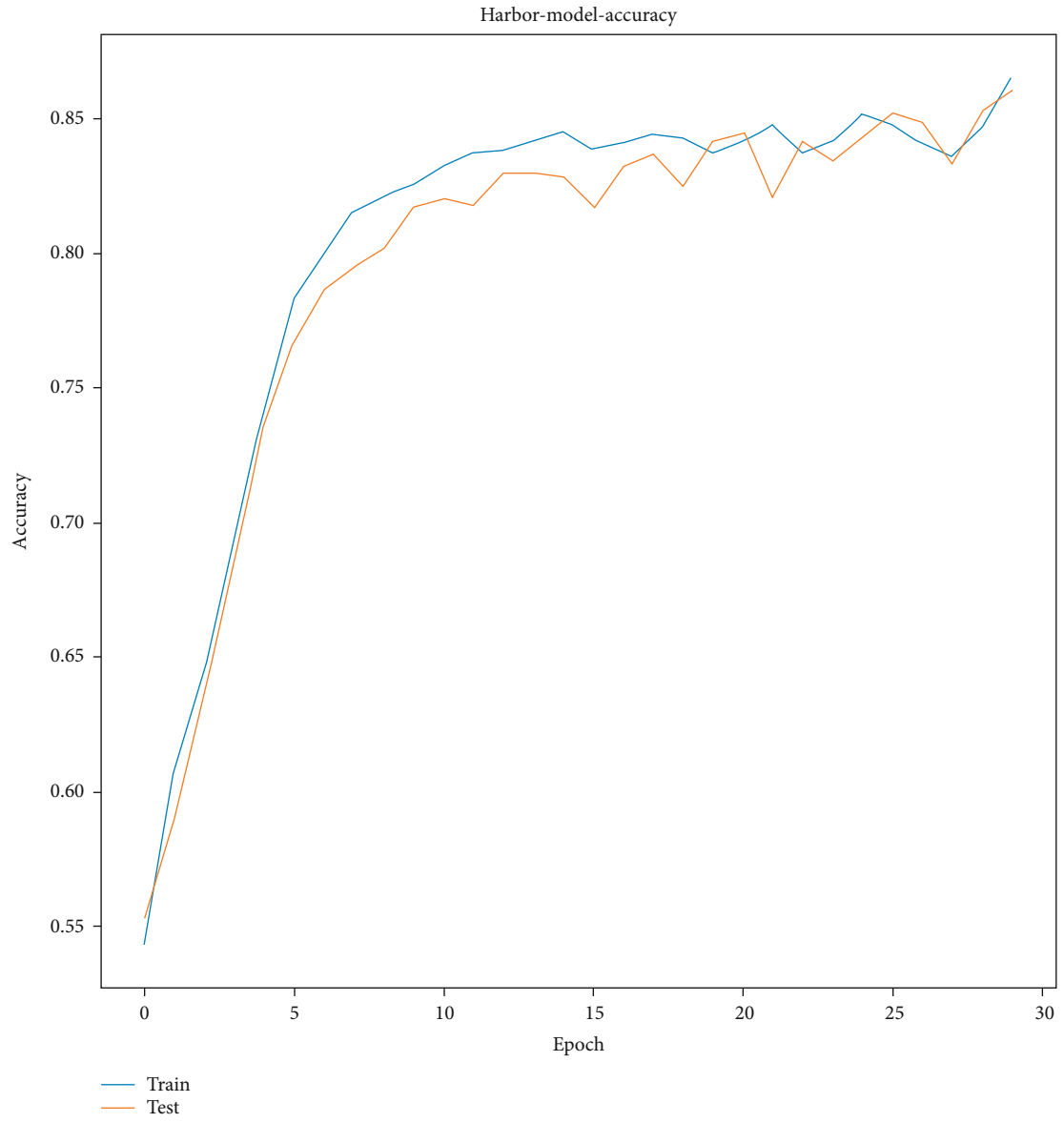


Figure 23: Accuracy of port target recognition.

model has hardly decreased. It can be seen that the compression method based on parameter quantization is effective to accelerate the reasoning time of the model.

For airports, bridges, ports, and other targets with low resolution, the accuracy calculation method shown in formula is adopted as the evaluation index according to the needs of the Cooperative Institute. The performance of the recognition accuracy of low-resolution targets in the test task in the remote sensing target recognition system with different resolution on the training set and the test set is shown in Figures 21–23.

For airport targets, the test set is 100 test images containing airport targets. As can be seen from Figure 21, after 30 epochs, the recognition accuracy of the system for the airport in the test image reaches 86%. For bridge targets, the test set is 120 test images including airport targets. As can be seen from Figure 22, after 30 epochs, the accuracy of bridge recognition reaches 86%. For bridge targets, the test set is 240 test images including airport targets. As can be seen from Figure 23, after 30 epochs, the accuracy of bridge recognition reaches 87%.

The test results show that after 30 iterations, the recognition accuracy of the system for airport, bridge, and port targets in the training set and test set are 86%, 86%, and 87%, respectively (see Table 4 for the recognition time, which meets the requirements of recognition accuracy not less than 85% and recognition time not more than 3 s in the system technical indicators). The recognition accuracy and recognition time are shown in Table 4.

For aircraft, oil tanks, ships, and other targets with high resolution, the test set is adopted as the index according to the needs of the Cooperative Institute. The test results are shown in Table 5. The recognition accuracy of oil tank, aircraft, and ship is 86%, 87% and 85%, respectively, which

Table 4: Target recognition accuracy of remote sensing image under low resolution.

Target type	Sample numbers	Speed (FPS)	Accuracy
Airport	800	15	86%
Bridge	750	16	86%
Port	600	15	87%

Table 5: Target recognition accuracy of remote sensing image under high resolution.

Target type	Sample numbers	Speed (FPS)	Target type
Oil tank	408	15	86%
Aircraft	327	16	87%
Warship	403	15	85%

meets the requirement that the recognition rate in the system technical index is not less than 85%. The identification time is less than 3000 ms, which meets the requirement that the identification time is no more than 3 s in the index requirements.

4. Conclusion

Aiming at the problem that the UAV video target detection accuracy is not high and cannot meet the needs of practical application, a new architecture based on deep learning is proposed. The architecture adopts the method of combining deep learning with traditional template matching and puts forward an optimization idea for target recognition. It fully extracts the local and global information of the image, with an average detection rate of 86.1%. The optimization idea of multithreading improves the speed of the algorithm, and the average detection time is 56.6 ms. The actual results show that the proposed algorithm effectively improves the recognition accuracy and speed.

Data Availability

The data sets used to support the findings of this study have not been made available because they are personal data collected by authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] J. T. Al-Bakri and Y. Y. Al-Jahmany, "Application of GIS and remote sensing to groundwater exploration in Al-Wala Basin in Jordan," *Journal of Water Resource & Protection*, vol. 5, no. 10, pp. 962–971, 2013.

[2] J. N. Sweet, "The spectral similarity scale and its application to the classification of hyperspectral remote sensing data," in *IEEE Workshop on Advances in Techniques for Analysis of Remotely Sensed Data, 2003*, pp. 92–99, Greenbelt, MD, USA, 2004.

[3] A. Chang, Y. Eo, S. Kim, Y. Kim, and Y. Kim, "Canopy-cover thematic-map generation for military map products using remote sensing data in inaccessible areas," *Landscape & Ecological Engineering*, vol. 7, no. 2, pp. 263–274, 2011.

[4] R. D. Hudson Jr. and J. W. Hudson, "The military applications of remote sensing by infrared," *Proceedings of the IEEE*, vol. 63, no. 1, pp. 104–128, 1975.

[5] S. Kodors, A. Rausis, A. Ratkevics, J. Zvirgzds, A. Teilans, and I. Ansone, "Real estate monitoring system based on remote sensing and image recognition technologies," *Procedia Computer Science*, vol. 104, pp. 460–467, 2017.

[6] D. Liu, L. He, and L. Carin, "Airport detection in large aerial optical imagery," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, QC, Canada, 2004.

[7] Y. Li, M. Li, F. Li, X. Sun, and W. Liu, "Real-time interactive object extraction system for high resolution remote sensing images based on parallel computing architecture," in *2010 18th International Conference on Geoinformatics*, pp. 1–6, Beijing, China, 2010.

[8] P. Druyts, W. Mees, and D. Borghys, *Semi-automatic Help for Aerial Region Analysis*, 2007.

[9] B. Guindon, "Computer-based aerial image understanding: a review and assessment of its application to planimetric information extraction from very high resolution satellite images," *Canadian Journal of Remote Sensing*, vol. 23, no. 1, pp. 38–47, 1997.

[10] X. Xueqiang and W. Runsheng, "An automatic recognition system for specific targets in remote sensing images," *Remote sensing technology and application*, vol. 15, pp. 179–183, 2000.

[11] J. Zhu, S. Qiang, and C. Fenge, "Research status and development trend of remote sensing big data," *Chinese Journal of image and graphics*, vol. 21, pp. 1425–1439, 2016.

[12] G. Fu, H. Zhao, C. Li, and L. Shi, "Road detection from optical remote sensing imagery using circular projection matching and tracking strategy," *Journal of the Indian Society of Remote Sensing*, vol. 41, no. 4, pp. 819–831, 2013.

[13] S. D. Mayunga, D. J. Coleman, and Y. Zhang, "Semi-automatic building extraction in dense urban settlement areas from high-resolution satellite images," *Empire Survey Review*, vol. 42, no. 315, pp. 50–61, 2010.

[14] H. X. da Da, *Research on Key Technologies of Typical Target Recognition in Large-Scale Visible Light Remote Sensing Images*, Harbin Institute of technology, 2013.

[15] R. Hulik, M. Spanel, P. Smrz, and Z. Materna, "Continuous plane detection in point-cloud data based on 3D Hough transform," *Journal of Visual Communication & Image Representation*, vol. 25, no. 1, pp. 86–97, 2014.

[16] W. Wang, J. Sun, and R. Hu, "Knowledge-based bridge detection from SAR images. Systems engineering and electronic technology (English version)," vol. 20, pp. 929–936, 2009.

[17] X. Li, S. Zhang, X. Pan, P. Dale, and R. Cropp, "Straight road edge detection from high-resolution remote sensing images based on the ridgelet transform with the revised parallel-beam Radon transform," *International Journal of Remote Sensing*, vol. 31, no. 19, pp. 5041–5059, 2010.

[18] X. Xu, X. Li, and C. Liu, "Building damage detection based on single high-resolution remote sensing imagery," in *International Conference on Automatic Control and Artificial Intelligence (ACAI 2012)*, pp. 618–621, 2012.

[19] L. Chun, Y. Junjun, and Y. Jian, *Small Port Detection Based on Polarimetric SAR Image Combining Shoreline Feature Points*, Journal of Tsinghua University: Natural Science Edition, 2015.

[20] M. Barzohar and D. B. Cooper, "Automatic finding of main roads in aerial images by using geometric-stochastic models and estimation," *Pattern Analysis & Machine Intelligence IEEE Transactions on*, vol. 18, no. 7, pp. 707–721, 1996.

[21] C. Qi and Y. T. Army, "Model based remote sensing image port detection," *Signal processing*, vol. 26, pp. 941–945, 2010.

[22] P. Zhong and R. Wang, "A multiple conditional random fields ensemble model for urban area detection in remote sensing optical images," *IEEE Transactions on Geoscience & Remote Sensing*, vol. 45, no. 12, pp. 3978–3988, 2007.

[23] Y. Zhang, *Research on Target Detection Method Based on Conditional Random Field*, Xi'an University of Electronic Science and Technology, 2014.

[24] Y. Yijun, Z. Rongchun, and W. Wenbing, "Automatic detection of artificial buildings in aerial images Computer engineering," vol. 28, pp. 20-21, 2002.

[25] B. Sirmacek and C. Unsalan, "Road detection from remotely sensed images using color features," in *Proceedings of 5th International Conference on Recent Advances in Space Technologies - RAST2011*, pp. 112–115, Istanbul, Turkey, 2011.

[26] M. Wang, J. Luo, and Z. Chenghu, "Road network extraction from high resolution remote sensing images by combining Gaussian Markov random field texture model and support vector machine," *Journal of remote sensing*, vol. 9, pp. 271–276, 2005.

[27] L. Jinzong, M. Lisheng, and L. Dongdong, "Fast recognition of airport targets in large-scale high-resolution remote sensing images," *Photoelectron laser*, pp. 1083–1088, 2010.

[28] A. R. Zamir and M. Shah, "Image geo-localization based on multiple nearest neighbor feature matching using generalized graphs," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 36, no. 8, pp. 1546–1558, 2014.

[29] L. Jingneng, *Image Local Invariant Feature Extraction Technology and Its Application*, Shanghai Jiaotong University, 2012.

[30] B. Jin, Y. Cong, W. Zhou, and G. Wang, "A new method for detection of ship docked in harbor in high resolution remote sensing image," in *2014 IEEE International Conference on Progress in Informatics and Computing*, pp. 341–344, Shanghai, China, 2014.

[31] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[32] X. Wang, B. Wang, and L. Zhang, "Airport Detection in Remote Sensing Images Based on Visual Attention," in *Neural Information Processing*, pp. 475–484, Springer-Verlag Berlin Heidelberg, 2011.

[33] Y. Yuan, J. Zhiguo, and Z. Haopeng, "Aircraft target detection in remote sensing images based on hierarchical classifier," *Aerospace return and remote sensing*, vol. 35, pp. 88–94, 2014.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[35] W. S. P. W. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[36] Y. Lécun, L. Bottou, and Y. Bengio, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[37] J. R. Uijlings, K. E. Sande, and T. Gevers, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[38] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, Columbus, OH, USA, 2014.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[40] R. Girshick, "Fast R-CNN," *Computer Science*, 2015.

[41] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[42] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, Las Vegas, NV, USA, 2016.

[43] J. Dai, Y. Li, and K. He, *R-FCN: Object Detection via Region-Based Fully Convolutional Networks*, 2016.

[44] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944, Honolulu, HI, USA, 2016.

[45] J. Redmon and A. Farhadi, *YOLOv3: An Incremental Improvement*, 2018.

[46] C. B. Azodi, J. Tang, and S. H. Shiu, "Opening the Black Box: Interpretable Machine Learning for Geneticists," *Trends in Genetics*, vol. 36, no. 6, pp. 442–455, 2020.

[47] R. Moraffah, M. Karami, R. Guo, A. Raglin, and H. Liu, "Causal interpretability for machine learning-problems, methods and evaluation[J]," *ACM SIGKDD Explorations Newsletter*, vol. 22, no. 1, pp. 18–33, 2020.

[48] H. Kaur, H. Nori, and S. Jenkins, "Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, pp. 1–14, 2020.

[49] K. Beckh, S. Müller, and M. Jakobs, "Explainable machine learning with prior knowledge: an overview," 2021, arXiv preprint arXiv:2105.10172.

[50] A. Chatzimparmpas, R. M. Martins, I. Jusufi, and A. Kerren, "A survey of surveys on the use of visualization for interpreting machine learning models," *Information Visualization*, vol. 19, no. 3, pp. 207–233, 2020.