Computer Science and Information Systems 22(1):327–344 https://doi.org/10.2298/CSIS241110011Y

Multimodal Deep Learning-based Feature Fusion for Object Detection in Remote Sensing Images

Shoulin Yin[1], Qunming Wang[2], Liguo Wang[3], Mirjana Ivanović[4], and Hang Li[5]

[1] College of Information and Communication Engineering, Harbin Engineering University
Harbin, 150001 China
yslin@synu.edu.cn
[2] College of Surveying and Geo-informatics, Tongji University
[3] College of Information and Communications Engineering, Dalian Minzu University
Dalian, 116600, China
wangliguo@hrbeu.edu.cn
Corresponding author:Liguo Wang
[4] Faculty of Sciences, University of Novi Sad, Serbia
mira@dmi.uns.ac.rs
[5] Software College, Shenyang Normal University
Shenyang, 110034 China
lihang@synu.edu.cn

Abstract. Object detection is an important computer vision task, which is developed from image classification task. The difference is that it is no longer only to classify a single type of object in an image, but to complete the classification and positioning of multiple objects that may exist in an image at the same time. Classification refers to assigning category labels to the object, and positioning refers to determining the vertex coordinates of the peripheral rectangular box of the object. Therefore, object detection is more challenging and has broader application prospects, such as automatic driving, face recognition, pedestrian detection, medical detection etc,. Object detection can also be used as the research basis for more complex computer vision task such as image segmentation, image description, object tracking and action recognition. In traditional object detection, the feature utilization rate is low and it is easy to be affected by other environmental factors. Hence, this paper proposes a multimodal deep learning-based feature fusion for object detection in remote sensing images. In the new model, cascade RCNN is the backbone network. Parallel cascade RCNN network is utilized for feature fusion to enhance feature expression ability. In order to solve the problem of different segmentation shapes and sizes, the central part of the network adopts multi-coefficient cascaded hollow convolution to obtain multi-receptive field features without using pooling mode and preserving image information. Meanwhile, an improved self-attention combined receptive field strategy is used to obtain both low-level features with marginal details and high-level features with global semantics. Finally, we conduct experiments on DOTA set including ablation experiments and comparison experiments. The experimental results show that the mean Average Precision (mAP) and other indexes have been greatly improved, and its performance is better than the state-of-the-art detection algorithms. It has a good application prospect in the remote sensing image object detection task.

Keywords: Object detection, remote sensing image, multimodal deep learning, feature fusion.

1. Introduction

In recent years, with the development of remote sensing technology, object detection technology based on remote sensing image has attracted wide attention. Object detection can locate the object of interest on the ground from a distance and identify its category. It has a wide range of applications and prospects in the fields of military defense and civil aviation [1,2]. In particular, the classification and detection of aircraft/airport objects in the application of high-resolution remote sensing images can provide some new solutions for the more efficient and scientific organization and management of civil aviation, military and national defense research and other fields [3].

With the rapid development of deep learning, object detection algorithms based on deep learning are widely used in various fields. At present, the mainstream object detection algorithms are divided into two categories: (1) two-stage object detection algorithm based on candidate regions; (2) single-stage object detection algorithm based on direct regression. Single-stage detection algorithm can achieve a balance in accuracy and running speed, which is a kind of detection framework widely used in object detection at present [4].

Single-stage object detection algorithms mainly include SSD [5] series and YOLO [6] series. Zhong et al. [7] proposed a single-stage rotating object detection method based on anchor frame transformation. Based on YOLOv3, a new feature extraction Network Darknet-53-Dense was designed in this algorithm to improve the feature extraction ability. In addition, an Anchor Transformation Network (ATN) [8] was proposed for the detection head network. The initial horizontal anchor frame was transformed into a rotating anchor frame to improve the accuracy of object detection. Cao et al. [9] proposed an algorithm for object detection based on Dilated convolution pyramid, introducing different sizes of hollow convolution layers into Feature Pyramid Networks (FPN). It built a Hybrid Receptive Field Module (HRFM). By increasing the receptive field to obtain more global feature information, the problem of object occlusion was solved. Murthy et al. [10] proposed an anti-residual object detection algorithm, which was based on YOLOv4. An efficient Lightweight Ghost Convolution model (LGC) was proposed to obtain more feature maps with fewer parameters to improve the network's ability to extract global feature information and improve the overall object detection performance. Zhao et al. [11] proposed a object detection algorithm based on improved YOLOv3, added convolutional layer module to the network structure to classify the object background of samples, and adjusted the size of the anchor frame on the feature map. After the object background probability was output, the samples whose background probability value was lower than the set threshold were filtered out, so as to solve the imbalance of the ratio of positive and negative samples in the original algorithm.

Kumar et al. [12] proposed an improved object detection algorithm (Single Shot multibox Detector (SSD) based on DenseNet [13] and feature fusion. On the basis of SSD network, DenseNetS-31-1 feature extraction network was designed according to DenseNet dense connection to enhance the feature extraction ability of the model. In the multi-scale detection part, the fusion mechanism of multi-scale feature layer was introduced to combine the low-level semantic features and high-level semantic features in the network structure, and then improved the model performance.

The two-phase object detection algorithms mainly use RCNN [14] as a series. Xiao et al. [15] proposed an improved object detection algorithm for Faster RCNN, which

improved the accuracy of the model by introducing two difficult sample mining strategies and alternate training. Jiang et al. [16] proposed an improved Faster R-CNN algorithm for multi-scale object detection. By adopting multi-level feature extraction strategy to extract multi-scale object features and using multi-channel method to generate multi-scale object candidate boxes, the accuracy of the object detection algorithm could be improved to some extent. Detection based on the above single or two stage methods can improve the accuracy of object detection. However, the above improved algorithms have low detection accuracy when the detection object is different, the background is complex and the object is blocked. For example, it can be seen from Figure 1 that the scales of Big Ben, dogs and crowds in the three images from Figure 1(a) to Figure 1(c) are from large to small. The two images shown in Figure 1(d) both contain objects of multiple scales. For MSCOCO data sets commonly used in detection tasks, if the scale of all instances is statistically and sorted according to the pixel ratio of object mask to image, it will be found that 10% of objects in the data set have a scale smaller than 0.0207, and 10% of objects have a scale larger than 0.345, with a large scale span.

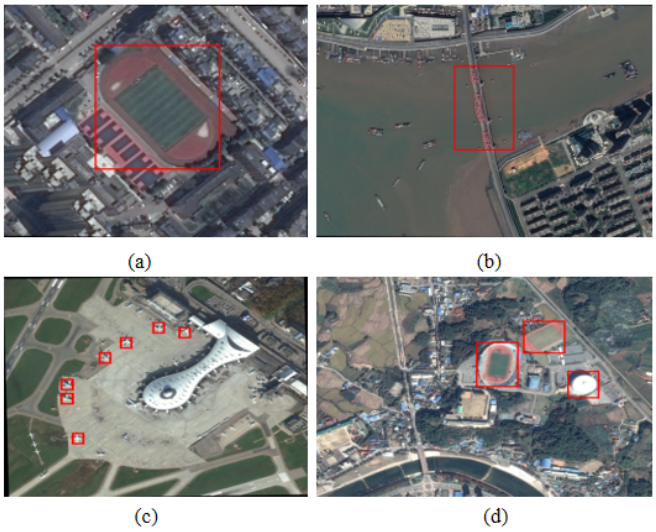

Fig. 1. Remote sensing image objects at different scales. (a) big object; (b) medium object; (c) small object; (d) Multi-scale object

Object detection includes two sub-tasks: object location and classification. The scaling problem lies in the fact that, in the process of deepening convolutional neural networks [17], the ability to express abstract features becomes stronger and stronger, but the shallow spatial information is also relatively lost.

In reference [18], deconvolution layer was added to CNN to fuse the deep and shallow features of CNN network for the detection of buildings in remote sensing images. In reference [19], optimized ResNet model was introduced to solve the significance detection problem of remote sensing images. In reference [20], CNN features with moderate sen-

sitivity field were selected according to the aircraft imaging size in the image, and deep CNN features and shallow CNN features were sampled for superposition fusion. In reference [21], Markov random fields and full convolutional neural networks were introduced to generate high-quality candidate regions. In reference [22], multi-layer CNN features were integrated to describe vehicle objects in remote sensing images, and the hierarchical boost classifier was used to discriminate and achieves good results. Reference [23] used features of different layers of CNN to detect objects of different scales respectively, and improved the detection effect by combining context information. Reference [24] expanded the sample data and combined the object context features to detect aircraft objects in remote sensing images. It can be seen that, for the problem of object scale diversity and small object in remote sensing images, it is a good idea to integrate the corresponding features of different convolutional layers in CNN network, that is, to integrate the detailed information rich in shallow convolutional layer and the semantic information rich in deep convolutional layer in CNN network for feature extraction. However, the use of dimension splicing or pixel-by-pixel addition/multiplication to fuse multi-layer features rarely considers the distribution and scale differences of features of different layers, so feature fusion is still a difficult research task. In addition, the background complexity of remote sensing image has great interference on object detection, so it is necessary to pay more attention to the influence of context information on object detection.

Therefore, according to the above analysis, this paper proposes a multimodal deep learning-based feature fusion for object detection in remote sensing images. In the new model, cascade RCNN is the backbone network. In the multi-scale object detection task, the proposed method makes full use of the features of different scales for fusion, which can greatly improve the robustness of the algorithm.

Our main contributions for this paper are as follows:

1. Parallel cascade RCNN network is utilized for feature fusion to enhance feature expression ability.
2. In order to solve the problem of different segmentation shapes and sizes, the central part of the network adopts multi-coefficient cascaded hollow convolution to obtain multi-receptive field features without using pooling mode and preserving image information.
3. Meanwhile, an improved self-attention combined receptive field strategy is used to obtain both low-level features with marginal details and high-level features with global semantics.

The organizational structure of this paper is assigned as follows. This paper summarizes recent developments in detecting remote sensing objects using deep learning techniques in section 2. The proposed object detection architecture via multi-modal deep learning is presented in Section 3. Section 4 presents experiments including various situations. Section 5 presents the conclusion and future works.

2. Related Works

In order to quantify the scale of the object, usually the area occupied by the object instance (i.e. the number of pixels occupied by the mask) is divided by the area of the image and the result obtained is taken as the relative scale of the object instance (between 0-1), which

is referred to as the scale. Therefore, the relative scales of objects in different images are very different, or the sizes of multiple objects in the same image are very different, which is called the scaling problem. It has always been one of the core challenges that affect the accuracy of object detection.

Table 1 lists the detection results of some object detection algorithms on MSCOCO test set. The "++" symbol indicates that the model uses an image pyramid when inferring. Where, AP refers to the average accuracy when the thresholds of IoU are 0.50:0.05:0.95. AP50 and AP75 are the accuracy when the IoU threshold is 0.50 and 0.75, respectively. APS, APM and APL refer to small, medium and large object AP respectively.

It can be seen from the data in Table 1 that the detection accuracy of small objects of early detectors such as SSD, YOLOv2 and FPN is less than half of that of medium and large objects [25]. In recent two years, the size of detectors has been improved, but there is still an obvious gap between the accuracy of small objects and that of medium and large objects, which seriously affects the improvement of the overall accuracy. Therefore, how to make detectors better cope with objects of different scales (especially small objects) is still an important problem in current object detection research.

Table 1. Different object detection models

Model	Skeleton network	Year	AP	AP50	AP57	APS	APM	APL
Faster RCNN	VGGNet-16	2015	21.9	42.7	–	–	–	–
SSD512	VGGNet-16	2016	28.8	48.5	30.3	10.9	31.8	43.5
Faster RCNN+++	ResNet-101	2016	34.9	55.7	37.4	15.6	38.7	50.9
R-FCN	ResNet-101	2016	29.9	51.9	–	10.8	32.8	45.0
Cascade R-CNN	ResNet-101	2018	42.8	62.1	46.3	23.7	45.5	55.2
DES512	VGGNet-16	2018	32.8	53.2	34.6	13.9	36.0	47.6
TridentNet	ResNet-101	2019	42.7	63.6	46.5	23.9	46.6	56.6
YOLOv3	DarkNet-53	2019	43.9	64.1	49.2	27.0	46.6	53.4
ATSS	ResNet-101	2020	46.3	64.7	50.4	27.7	49.8	58.4
Dynamic R-CNN	ResNet-101	2020	42.0	60.7	45.9	22.7	44.3	54.3
YOLOv4	CPSDarkNet-53	2020	43.5	65.7	47.3	26.7	46.7	53.3

In this paper, we mainly focus on the cascade RCNN to perfect the proposed method in remote sensing images. Cascade RCNN Network consists of feature extraction network (ResNet101), feature pyramid network (FPN), Region Proposal Network (RPN) layer, and cascade detector. Feature extraction network ResNet101 is used to extract image features. The original image is convolved by Conv1, Conv2, Conv3, Conv4 and Conv5 and features of different levels are fused to obtain feature images P2, P3, P4 and P5 of different scales. Then, the feature maps of different scales P2, P3, P4 and P5 are input into RPN to obtain candidate object regions. After the ROI Align [26] operation on the obtained candidate object region, the feature map of Region of Interest (ROI) with uniform size is obtained. Figure 2 shows the Cascade RCNN network structure.

In the detection stage, different from Faster RCNN, Cascade RCNN uses cascade detector for detection, and three detectors set different thresholds respectively for detection. Each detector consists of ROI Align, full connection layer, classification score C and frame regression position coordinate B. During detection, the candidate object region is

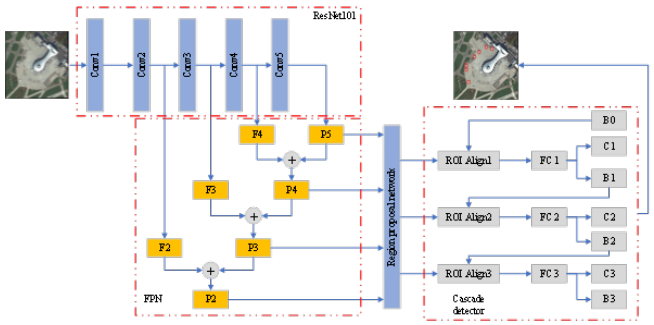

Fig. 2. Cascade RCNN network structure

re-sampled through the frame regression B output by the detector in the previous stage, and the new classification score C and frame regression B are obtained by gradually improving the IoU threshold training, and finally the sample quality and network training effect are improved.

In the process of frame regression, annotated frame P is the predefined anchor, annotated frame G is the object frame, and annotated frame G' is the forecast frame output by the model, whose ultimate purpose is to bring the forecast frame closer to the object frame. When the IoU of candidate box and object box is large, the transformation $d(\cdot)$ from candidate box to prediction box can be regarded as an approximate linear transformation. Define the object box center (G_x, G_y), width and height (G_w, G_h), candidate box center (P_x, P_y), width and height (P_w, P_h), and establish a regression model, as shown in equation (1).

$$\begin{cases} G'_x = & P_x + P_w d_x(P) \\ G'_y = & P_y + P_h d_y(P) \\ G'_w = & P_w exp(d_h(P)) \\ G'_h = & P_h exp(d_h(P)) \\ d_x(P) = & W_*^T \phi(P) \end{cases} \tag{1}$$

Here, $\phi(P)$ is the feature of the candidate frame, and is the parameter to be learned.

The loss function of Cascade RCNN mainly consists of two parts, namely classification error and coordinate regression error, as shown in equation (2).

$$L(p_i, t_i) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*). \tag{2}$$

Where p_i is the probability that anchor prediction is the object. p_i^* is the probability of the real box. $t_i = [t_x, t_y, t_w, t_h]$ is a vector representing the four parameterized coordinates of the prediction box. t_i^* is the coordinate vector of the real box, N_{cls} and N_{reg} both represent the total number of samples, λ is the weight balance factor.

In Formula (2), $L_{cls}(p_i, p_i^*)$ is the error between the predicted class confidence and the object class, and the loss function is the cross entropy loss function. $L_{reg}(t_i, t_i^*)$ is frame regression loss, Smooth is adopted as loss function, as shown in equations (3)-(5).

$$L_{cls}(p_i, p_i^*) = -\log[p_i^* p_i + (1 - p_i^*)(1 - p_i)]. \tag{3}$$

$$L_{reg}(t_i, t_i^*) = Smooth_{L1}(t_i - t_i^*). \tag{4}$$

$$Smooth_{L1}(x) = \begin{cases} 0.5x^2, |x| < 1 \\ |x| - 0.5, others \end{cases} \tag{5}$$

3. Multimodal Deep Learning-based Feature Fusion for Object Detection

Aiming at the problems of low detection rate and object occlusion in Cascade RCNN network, this paper proposes an improved Cascade RCNN network structure diagram, as shown in Figure 3. The improved Cascade RCNN algorithm introduces a Dilated convolution module in ResNet101, which carries out multi-scale feature extraction and enhances the robustness of the model for slices with different sizes. The coordinate attention mechanism is introduced into ResNet101 residual network [27]. The low level features with edge details and high level features with global semantics are obtained by using axial self-attention combined with receptive field strategy to improve the accuracy of object detection.

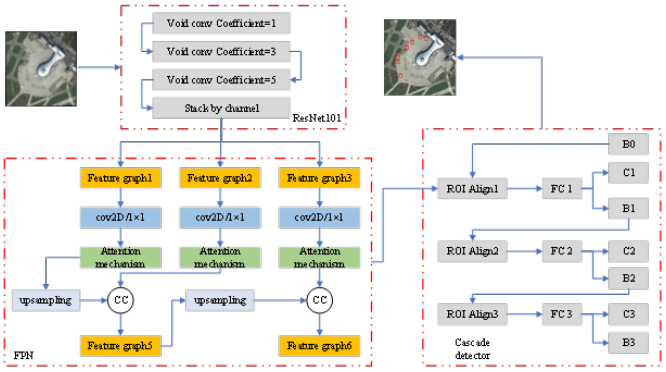

Fig. 3. Proposed object detection structure. CC is Channel-wise Concat.

3.1. Dilated Convolution Module

Considering the complexity of remote sensing images, the correlation of the whole image and the difference of the original resolution distribution of each data sample, it is of great significance to improve the sensing range of the features in the central part of the network and the feature fusion of multiple receptive fields. Pooling can effectively improve the

receptive field of the feature map, but at the same time, important spatial information will be lost due to the decrease of the resolution of the feature map. Therefore, a 3-layer dilated convolution module (3DCM) with skip connection is designed in the central part of FPN to obtain the features of three receptive fields. The structure of this module is shown in Figure 4. C represents the number of output channels of the encoder. It is specifically defined as:

$$d = \varphi(C(\sigma(g_1); \sigma(g_2); \sigma(g_3))). \tag{6}$$

Where g_i represents the output of the $i-th$ layer dilated convolution. σ is the activation function of ELU (exponential linear units). In this paper, the adjustable parameter α of the activation function is set as 1. C represents the stacked feature graphs by channel. $\varphi \in R^{1\times1\times N/2}$ is the convolutional parameter matrix, where N is the feature graph size of input 3DCM module.

Dilated convolution effectively improves the receptive field without introducing new parameter number by injecting holes into the convolution kernel, which is specifically defined as:

$$g[x, y] = \sum_{i}^{M} \sum_{j}^{M} f[x + r \times i, y + r \times j] \cdot h[i, j]. \tag{7}$$

Where x and y represent the coordinates of feature points. i and j are the coordinates of the convolution point. M is the size of the convolution kernel. r is the void coefficient, $r = 1$ in the standard convolution operation. f is the input feature, h is the convolution kernel and g is the output feature.

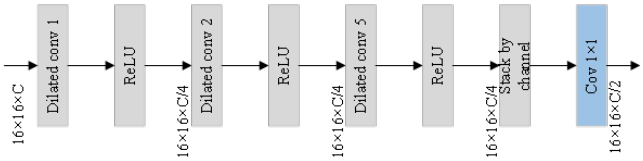

Fig. 4. One kernel

However, if the dilated coefficient is set in the continuous dilated convolution layer without mutual prime, the problem of sampling discontinuity of the feature graph, that is, grid effect, will be generated, and a large amount of feature information will be lost. In order to avoid this effect and take into account the segmentation effect of large and small objects, the dilated coefficient of 3DCM follows the design structure of hybrid dilated convolution (HDC). The dilated coefficients of three layers are set as 1, 2 and 5 respectively. The size of convolution kernel is 3×3 and the step is 1. In this way, the adjacent information of the feature map can be obtained by the 3DCM module and the recognition ability of small object can be improved. In addition, it can also obtain deep receptive field similar to the feature map and improve the recognition ability of large objects.

The receptive field is calculated as:

$$l_k = l_{k-1} + (f_k - 1) \times \prod_{i=1}^{k-1} s_i. \tag{8}$$

Where l_k is the receptive field of each point in the $k-th$ layer. f_k is the size of the $k-th$ convolution kernel. s_i is the convolution step of the $i-th$ layer. According to Formula (8), the actual receptive fields of each layer relative to the output feature map of the encoder are respectively 3, 7 and 17. Since the size of the output feature graph of the encoder is 16×16, the feature points of layer 3 cavity convolution will cover relatively complete information in the input feature graph of 3DCM module. In addition, after stacking channels for feature graphs of different receptive fields, 3DCM uses 1×1 convolution instead of channel addition, which improves the ability of the network to adjust feature weights of different receptive fields adaptively and promotes information fusion.

3.2. Receptive Field Enhancement in Axial Self-attention

In image segmentation, many researchers are studying how to extract low-level features with edge details and global semantic high-level features at the same time. Although self-attention can better achieve the above purposes, this method requires a huge amount of computation, and axial attention solves the above two problems to a certain extent. Therefore, based on this, and combining with the receptive field block (RFB) [28] strategy, this paper designs an axial self-attention receptive field module, as shown in Figure 5. First, the input feature graphs are respectively passed through the receptive field path of 1×3, 1×5, 1×7, 1×1 convolution layer. Secondly, the receptive field is expanded by the cavity convolution layer with cavity rates of 3, 5, 7, 1 respectively, and then the axial self-attention module is entered. Finally, the channel is spliced with the feature map of 1×1 receptive field path and output. Among them, axial self-attention is shown in Figure 6.

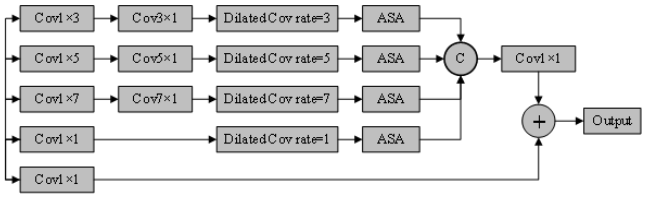

Fig. 5. Axial self-attention field enhancement module

This module is used to build a rich context-dependent model for local features, model the remote dependency relationship, and improve the feature representation of remote sensing image segmentation. Since its purpose is to enhance the features of relatively small objects, this paper selects parallel strategies to compute both horizontal and vertical directions for non-local operations to construct axial attention. At the same time, since the horizontal direction and the vertical direction contribute equally to the output, this paper

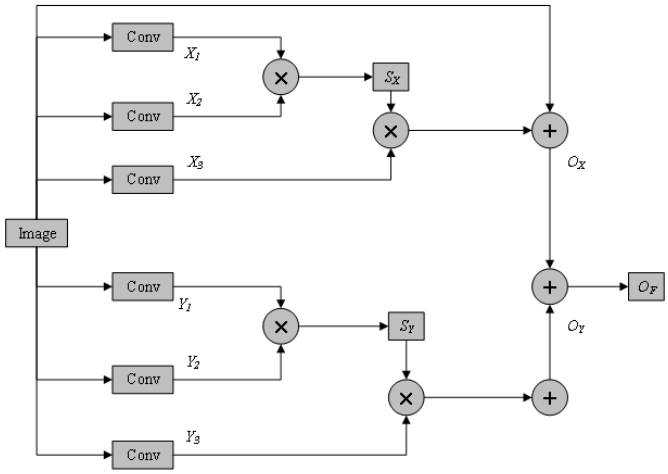

Fig. 6. Axial self-attention module

adopts the method of element by element addition to aggregate the feature maps. The local feature maps are first sent to the convolution layer to generate feature maps, which are reconstructed and transposed into $H \times C \times W$ and $W \times H \times C$ respectively in the horizontal direction (X-axis) to obtain X_1 and X_2. The horizontal space feature maps S_X are calculated by Softmax. i and j represent the position of pixel space. The influence of the $j - th$ position on the pixel of the $i - th$ position can be expressed as:

$$S_{ij} = \frac{e^{(X_{1i}, X_{2j})}}{\sum_{i=1}^{N} e^{(X_{1i}, X_{2j})}}. \tag{9}$$

The more similar the feature representations of the two locations, the stronger the correlation between them. After matrix multiplication of and the feature map reconstructed as $W \times H \times C$, the feature map in the horizontal direction can be obtained by adding the original local feature element by element.

$$O_X = \sum_{i=1}^{N} S_{ij} X_{3i}. \tag{10}$$

Similarly, in the vertical direction (Y-axis), they are reconstructed and transposed as $W \times C \times H$ and $H \times W \times C$, the vertical spatial feature graph S_Y is calculated by Softmax. After matrix multiplication with the reconstructed feature graph $H \times W \times C$, the vertical feature graph O_Y is obtained by adding the original local feature element by element. Finally, it outputs O_F by adding the feature graphs horizontal axis and vertical axis.

4. Experiment Results and Analysis

4.1. Data Set and Experiment Setting

In order to test the performance of the multi-modal deep learning model, 200 high-resolution remote sensing images containing different object categories are collected in the DOTA data set published by Wuhan University [29]. The scale of the original remote sensing image ranges from 800×800 pixels to 4000×4000 pixels. These remote sensing images include roads, trees, houses and other types of complex backgrounds. The spatial resolution range is 0.1-0.3m. Before training, we use DOTA_devkit tools to cut the image into 600×600 pixels, Stride size=100. After data enhancement processing, the images are divided into three sets, of which 4157 are used as the training set, 1064 as the verification set and 1234 as the test set. The model is trained on the training set and tested on the test set. The Non-maximum-suppression (NMS) method with IOU threshold=0.1 is adopted for the final test results to discard repeated detection.

The experimental operating environment is Intel(R) Core (TM) i79700CPU 3.00GHz processor, NIVDIA RTX 30608 GB GDDR6 graphics card 64 GB-DDR4 memory. The environment settings are Cuda10.1 and Cudnn 7.6.4. The network frameworks for deep learning are Pytorch 1.7.1, Python 3.8.8. During training, batch size is set to 4, learning rate to 1.25×10^{-4} and num_epochs to 80. In order to accelerate the model convergence speed, the pre-training weight of ResNet50 on the ImageNet classification task is also introduced as transfer learning.

In this paper, Mean Average Precision (mAP) and Frame Per Second (FPS) are used as evaluation indexes of the model. mAP represents the percentage of the number of correctly recognized single objects in the total number of recognized objects, which is used to measure the overall comprehensive performance of the model. FPS is positively correlated with the speed of model detection.

4.2. Ablation Experiments

In order to verify the effectiveness of multi-scale feature fusion module and spatial and channel attention module in multi-modal deep learning networks, ablation experiments are conducted on DOTA remote sensing data sets, and the experimental results are shown in Table 2.

Table 2. Ablation experiments

Backbone	3DCM	Attention	mAP/%	FPS/s
ResNet101	NO	NO	89.64	1.72
ResNet101	YES	NO	89.98	1.32
ResNet101	NO	YES	90.27	1.37
ResNet101	YES	YES	92.76	1.25

After the introduction of dilated convolution and self-attention multi-scale feature fusion modules, the mAP of the model increases by 3.12% when the backbone network is ResNet101. After adding the self-attention module, the model detection mAP increases

by 0.63% when the backbone is ResNet101. Through the analysis of experimental data, it is found that the introduction of multi-scale feature fusion module and attention module increases the complexity of the whole model, resulting in a 0.47 frame/s FPS reduction, but the model detection speed can still meet the requirements of real-time detection. Some results based on the proposed method are shown in Figure 7.

Fig. 7. Detection results

4.3. Comparison Experiments

To verify the effectiveness of the object detection algorithm in this paper, it is compared with other advanced object detection algorithms containing MSFYOLO, RFEB, ADIR and MSDA. Under the same training and test data sample conditions, $IOU > 0.5$ indicates that the detection is correct, and mAP and FPS are used as evaluation indicators. Tables 3,4,5 show the comparison of detection results between the proposed algorithm and other algorithms on DOTA data sets. To display the intuitionally results, figures 8,9,10 give the objective results.

As shown in Table 3, in the DOTA data set, the detection model presented in this paper has the best detection effect on airport targets, and the detection accuracy reaches 93.87%. In addition, compared with MSFYOLO, RFEB, ADIR and MSDA, the detection model proposed in this paper increases the average accuracy mAP by 12.33%, 9.63%, 5.88%,

Table 3. Comparison experiments on airports

Model	mAP/%	FPS/s
MSFYOLO	81.54	1.97
RFEB	84.24	1.22
ADIR	88.19	2.07
MSDA	90.64	1.73
Proposed	93.87	1.46

Table 4. Comparison experiments on harbors

Model	mAP/%	FPS/s
MSFYOLO	80.81	1.62
RFEB	76.92	1.84
ADIR	82.13	1.45
MSDA	86.95	2.28
Proposed	91.49	1.27

and 3.23%, respectively. Compared with B algorithm, the detection speed of the model is increased by 0.24 frame/s. Due to the introduction of void convolution and self-attention multi-scale feature fusion modules, the model complexity is relatively high, resulting in a relatively slow model detection speed compared with similar models.

From table 4 and table 5, the mAP values of proposed model are 91.49% and 95.58% on harbor and aircraft object respectively. It also illustrates the better detection effect on DOTA.

5. Conclusion

In this paper, an object detection model based on multi-modal deep learning feature fusion in remote sensing images is proposed. A multi-scale feature fusion module is constructed by integrating dilated convolution and self-attention mechanism to enrich the spatial and semantic information of objects and further improve the effectiveness of model detection. Compared with the traditional detection algorithm ResNet101, the accuracy is improved by about 3.12%, which proves the effectiveness of the proposed detection algorithm. Although the detection effect of the method in this paper has been significantly improved on the test set, some problems still exist, and this is also the direction of future research. 1) The detection is time-consuming. Although the detection accuracy of this method has

Table 5. Comparison experiments on aircraft

Model	mAP/%	FPS/s
MSFYOLO	83.83	1.66
RFEB	86.52	1.95
ADIR	88.27	2.84
MSDA	94.84	1.67
Proposed	95.58	1.48

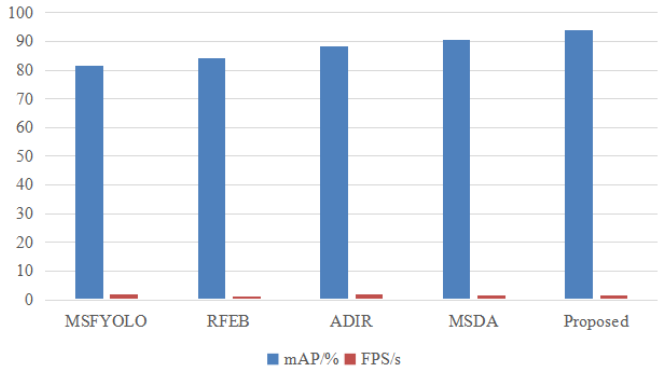

Fig. 8. Visualization result for airports

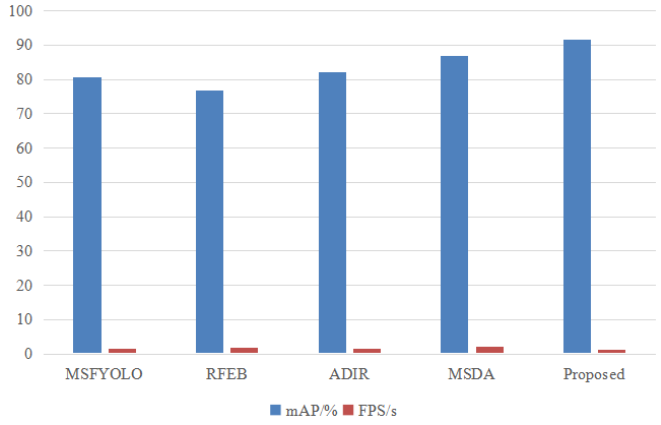

Fig. 9. Visualization result for harbors

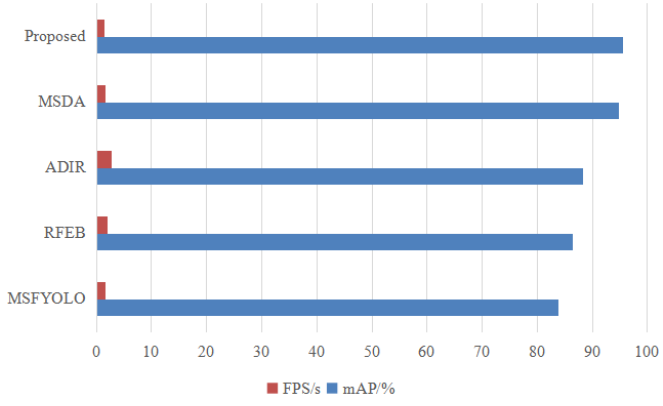

Fig. 10. Visualization result for aircraft

been significantly improved, the calculation amount is relatively large. 2) Context is not used enough. For example, the distribution of objects such as aircraft and oil-tanks has certain linear and clustered characteristics. Better use of these characteristics may further improve the detection accuracy. In the following work, the method of candidate region generation will be studied to improve the quality of candidate region generation and reduce the amount of calculation. At the same time, the application of context information will be deeply mined to explore the distribution correlation between objects and the use of context information in position regression, so as to further improve the accuracy of optical remote sensing image object detection.

Acknowledgments. This work was supported by the Liaoning Province science and technology plan joint project. Project name ”Research on key technologies of object recognition and interpretation in high resolution remote sensing images under complex scenes”.

References

1. Y. Pi, N. D. Nath, H. A. Behzadan, Convolutional neural networks for object detection in aerial imagery for disaster response and recovery, Advanced Engineering Informatics, vol. 43, pp. 101009, 2020.
2. M. R. Marshall et al., ”3-D Object Tracking in Panoramic Video and LiDAR for Radiological Source Object Attribution and Improved Source Detection,” IEEE Transactions on Nuclear Science, vol. 68, no. 2, pp. 189-202, Feb. 2021, doi: 10.1109/TNS.2020.3047646.
3. S. Yin, L. Wang, M. Shafiq, L. Teng, A. A. Laghari and M. F. Khan, ”G2Grad-CAMRL: An Object Detection and Interpretation Model Based on Gradient-Weighted Class Activation Mapping and Reinforcement Learning in Remote Sensing Images,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 16, pp. 3583-3598, 2023, doi: 10.1109/JSTARS.2023.3241405
4. Y. Tao, Z. Zongyang, Z. Jun, C. Xinghua and Z. Fuqiang, ”Low-altitude small-sized object detection using lightweight feature-enhanced convolutional neural network,” Journal of Systems Engineering and Electronics, vol. 32, no. 4, pp. 841-853, Aug. 2021, doi: 10.23919/JSEE.2021.000073.
5. A. Kumar, S. Srivastava, Object detection system based on convolution neural networks using single shot multi-box detector, Procedia Computer Science, vol. 171, pp. 2610-2617, 2020.
6. T. Diwan, G. Anirudh, V. J. Tembhurne, Object detection using YOLO: Challenges, architectural successors, datasets and applications, Multimedia Tools and Applications, vol. 82, no. 6, pp. 9243-9275, 2023.
7. B. Zhong, K. Ao, Single-stage rotation-decoupled detector for oriented object, Remote Sensing, vol. 12, no. 19, pp. 3262, 2020.
8. F. Zhao, J. Li, J. Zhao and J. Feng, ”Weakly Supervised Phrase Localization with Multi-scale Anchored Transformer Network,” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 5696-5705, doi: 10.1109/CVPR.2018.00597.
9. D. Cao, S. Yang, A Method based on Faster RCNN Network for Object Detection, Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science), vol. 15, no. 9, pp. 1239-1244, 2022.
10. B. C. Murthy,F. M. Hashmi, G. A. Keskar, EfficientLiteDet: a real-time pedestrian and vehicle detection algorithm, Machine Vision and Applications, vol. 33, no. 3, pp. 47, 2022.
11. L. Zhao, S. Li, Object detection algorithm based on improved YOLOv3, Electronics, vol. 9, no. 3, pp. 537, 2020.

12. A. Kumar, S. Srivastava, Object detection system based on convolution neural networks using single shot multi-box detector, Procedia Computer Science, vol. 171, pp. 2610-2617, 2020.
13. K. Zhang, Y. Guo, X. Wang, J. Yuan and Q. Ding, "Multiple Feature Reweight DenseNet for Image Classification," IEEE Access, vol. 7, pp. 9872-9880, 2019, doi: 10.1109/ACCESS.2018.2890127.
14. L. Wang, S. Yin, A. Hashem, et al., A novel deep learning-based single shot multibox detector model for object detection in optical remote sensing images,Geoscience Data Journal, 2022. https://doi.org/10.1002/gdj3.162
15. Y. Xiao, X. Wang, P. Zhang, et al., Object detection based on faster R-CNN algorithm with skip pooling and fusion of contextual information, Sensors, vol. 20, no. 19, pp. 5490, 2020.
16. D. Jiang, G. Li, C. Tan, et al., Semantic segmentation for multiscale target based on object recognition using the improved Faster-RCNN model, Future Generation Computer Systems, vol. 123, pp. 94-104, 2021.
17. S. Yin and H. Li, "Hot Region Selection Based on Selective Search and Modified Fuzzy C-Means in Remote Sensing Images," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 5862-5871, 2020, doi: 10.1109/JSTARS.2020.3025582.
18. G. A. Tsihrintzis, P. M. Johansen and A. J. Devaney, "Buried object detection and location estimation from electromagnetic field measurements," IEEE Transactions on Antennas and Propagation, vol. 47, no. 11, pp. 1742-1744, Nov. 1999, doi: 10.1109/8.814957.
19. G. Cheng, Y. Si, H. Hong, X. Yao and L. Guo, "Cross-Scale Feature Fusion for Object Detection in Optical Remote Sensing Images," IEEE Geoscience and Remote Sensing Letters, vol. 18, no. 3, pp. 431-435, March 2021, doi: 10.1109/LGRS.2020.2975541.
20. D. Ortego, J. C. SanMiguel and J. M. Martanez, "Long-Term Stationary Object Detection Based on Spatio-Temporal Change Detection," IEEE Signal Processing Letters, vol. 22, no. 12, pp. 2368-2372, Dec. 2015, doi: 10.1109/LSP.2015.2482598.
21. C. R. del-Blanco, F. Jaureguizar and N. Garcia, "An efficient multiple object detection and tracking framework for automatic counting and video surveillance applications," IEEE Transactions on Consumer Electronics, vol. 58, no. 3, pp. 857-862, August 2012, doi: 10.1109/TCE.2012.6311328.
22. X. Yu, J. Yang, Z. Lin, J. Wang, T. Wang and T. Huang, "Subcategory-Aware Object Detection," IEEE Signal Processing Letters, vol. 22, no. 9, pp. 1472-1476, Sept. 2015, doi: 10.1109/LSP.2014.2299571.
23. L. Pan, W. -S. Chu, J. M. Saragih, F. De la Torre and M. Xie, "Fast and Robust Circular Object Detection With Probabilistic Pairwise Voting," IEEE Signal Processing Letters, vol. 18, no. 11, pp. 639-642, Nov. 2011, doi: 10.1109/LSP.2011.2166956.
24. Q. Hu, S. Paisitkriangkrai, C. Shen, A. van den Hengel and F. Porikli, "Fast Detection of Multiple Objects in Traffic Scenes With a Common Detection Framework," IEEE Transactions on Intelligent Transportation Systems, vol. 17, no. 4, pp. 1002-1014, April 2016, doi: 10.1109/TITS.2015.2496795.
25. Z. -Q. Zhao, P. Zheng, S. -T. Xu and X. Wu, "Object Detection With Deep Learning: A Review," IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 11, pp. 3212-3232, Nov. 2019, doi: 10.1109/TNNLS.2018.2876865.
26. Y. Lin, H. He, Z. Yin and F. Chen, "Rotation-Invariant Object Detection in Remote Sensing Images Based on Radial-Gradient Angle," IEEE Geoscience and Remote Sensing Letters, vol. 12, no. 4, pp. 746-750, April 2015, doi: 10.1109/LGRS.2014.2360887.
27. Z. Cai and N. Vasconcelos, "Cascade R-CNN: High Quality Object Detection and Instance Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 5, pp. 1483-1498, 1 May 2021, doi: 10.1109/TPAMI.2019.2956516.
28. Q. Yao, X. Hu and H. Lei, "Multiscale Convolutional Neural Networks for Geospatial Object Detection in VHR Satellite Images," IEEE Geoscience and Remote Sensing Letters, vol. 18, no. 1, pp. 23-27, Jan. 2021, doi: 10.1109/LGRS.2020.2967819.

29. J. Ding et al., "Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 11, pp. 7778-7796, 1 Nov. 2022.
30. Z. Song, Y. Zhang, Y. Liu, K. Yang and M. Sun, "MSFYOLO: Feature fusion-based detection for small objects,"IEEE Latin America Transactions, vol. 20, no. 5, pp. 823-830, May 2022, doi: 10.1109/TLA.2022.9693567.
31. X. Dong, R. Fu, Y. Gao, Y. Qin, Y. Ye and B. Li, "Remote Sensing Object Detection Based on Receptive Field Expansion Block," IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1-5, 2022, Art no. 8020605, doi: 10.1109/LGRS.2021.3110584.
32. Z. Wu, J. Wen, Y. Xu, J. Yang and D. Zhang, "Multiple Instance Detection Networks With Adaptive Instance Refinement," IEEE Transactions on Multimedia, vol. 25, pp. 267-279, 2023, doi: 10.1109/TMM.2021.3125130.
33. X. Dong et al., "Multiscale Deformable Attention and Multilevel Features Aggregation for Remote Sensing Object Detection," IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1-5, 2022, Art no. 6510405, doi: 10.1109/LGRS.2022.3178479.

Shoulin Yin received the M.A. degree from Shenyang Normal University, Shenyang, China, in 2015. He is currently pursuing the Ph.D. degree with the College of Information and Communication Engineering, Harbin Engineering University, Harbin. His research interests are remote sensing image processing and object detection.

Liguo Wang received his M.A. degree in 2002 and Ph.D. degree in signal and information processing in 2005 from Harbin Institute of Technology, Harbin, China. He held postdoctoral research position from 2006 to 2008 in the College of Information and Communications Engineering, Harbin Engineering University. He is currently a Professor with Information and Communications Engineering, Dalian Minzu University, Dalian, China. His research interests are remote sensing image processing and machine learning. He has published three books, 27 patents, and more than 200 papers in journals and conference proceedings. Email: wangliguo@hrbeu.edu.cn.

Qunming Wang received the Ph.D. degree from The Hong Kong Polytechnic University, Hong Kong, in 2015. He is currently a Professor with the College of Surveying and Geo-Informatics, Tongji University, Shanghai, China. From 2017 to 2018, he was a Lecturer (Assistant Professor) with Lancaster Environment Centre, Lancaster University, Lancaster, U.K., where he is currently a Visiting Professor. Prof. Wang is an Editorial Board Member for Remote Sensing of Environment, and serves as an Associate Editor for Science of Remote Sensing (sister journal of Remote Sensing of Environment) and Photogrammetric Engineering & Remote Sensing.

Mirjana Ivanovic (Member, IEEE) has been a Full Professor with the Faculty of Sciences, University of Novi Sad, Serbia, since 2002. She has also been a member of the University Council for informatics for more than 10 years. She has authored or coauthored 13 textbooks, 13 edited proceedings, 3 monographs, and of more than 440 research articles on multi-agent systems, e-learning and web-based learning, applications of intelligent techniques (CBR, data and web mining), software engineering education, and most

of which are published in international journals and proceedings of high-quality international conferences. She is/was a member of program committees of more than 200 international conferences and general chair and program committee chair of numerous international conferences. Also, she has been an invited speaker at several international conferences and a visiting lecturer in Australia, Thailand, and China. As a leader and researcher, she has participated in numerous international projects. She is currently an Editor-in-Chief of Computer Science and Information Systems Journal.

Hang Li received the B.S., M.S., and Ph.D. degrees in computer applications technology from Dalian Fisheries University, Dalian, China, Shenyang Institute of Technology, Shenyang, China, Northeastern University, Shenyang, China, in 1999, 2002, and 2005, respectively. In 2002, he joined Software College as a Teacher. He is a full Professor and Master Supervisor. He is an Outstanding Young Backbone Teacher of Liaoning General Institutions of Higher Learning (Liaoning Education Department). His research interests include image analysis and processing, big data, and cloud computing. Dr. Li won the Second Prize of National Defense Science and Technology Award of The Commission of Science, Technology and Industry for National Defense (National Defense Science, Technology and Industry Commission 2005GFJ2126-6) and First Prize of Science and Technology Award of China North Industries Group Corporation (2005-BQJ-1-0019-6).

Received: November 10, 2024; Accepted: January 01, 2025.