Object Detection by Integrating Scene-Level Semantic Information and Border Regression Reinforcement

Yu Quan[1], Zhixin Li[1]*, Canlong Zhang[1], Huifang Ma[2]

[1]Guangxi Key Lab of Multi-source Information Mining and Security,
Guangxi Normal University,Guilin 541004,China
[2]College of Computer Science and Engineering,Northwest Normal University,Lanzhou 730070,China
*Corresponding Author. E-mail: lizx@gxnu.edu.cn

***Abstract*—The improvement of object detection performance is mostly focused on the extraction of local information near the region of interest in the image, which results in detection performance in this area being unable to achieve the desired effect. First, a depth-wise separable convolution network(D_SCNet-127 R-CNN) is built on the backbone network. Considering the importance of scene and semantic informations for visual recognition, the feature map is sent into the branch of the semantic segmentation module, region proposal network module, and the region proposal self-attention module to build the network of scene-level and region proposal self-attention module. Second, a deep reinforcement learning was utilized to achieve accurate positioning of border regression, and the calculation speed of the whole model was improved through implementing a light-weight head network. This model can effectively solve the limitation of feature extraction in traditional object detection and obtain more comprehensive detailed features. The experimental verification on MSCOCO17, VOC12, and Cityscapes datasets shows that the proposed method has good validity and scalability.**

***Index Terms*—Object Detection, Separable convolutional neural network, Deep reinforcement learning, Self-attention mechanism**

I. Introduction

In recent years, the development of object detection has been relatively rapid, extending from image processing based on traditional manual feature extraction [1] to object detection based on deep learning [2]. Due to the lack of deep learning, early object detection was a little clumsy in the expression of image features. Therefore, considering its strong learning ability, expression ability, and robustness, the deep convolution neural network was applied for feature extraction. Since then, the development of object detection has entered a new era. Object detection algorithms based on deep learning are mainly divided into two categories: two-stage object detection algorithms (R-CNN [3], Fast R-CNN [4], Faster R-CNN [5], R-FCN [6], and Mask R-CNN [7]) and one-stage object detection algorithms, which are based on integrated convolutional networks (Yolo [8] and SSD [9]). The two algorithms greatly improve the accuracy and speed of object detection and realize the end-to-end object recognition and detection network.

This paper proposes a object detection model based on scene-level region proposal self-attention module. As shown in Figure 1. Firstly, from the perspective of reducing the

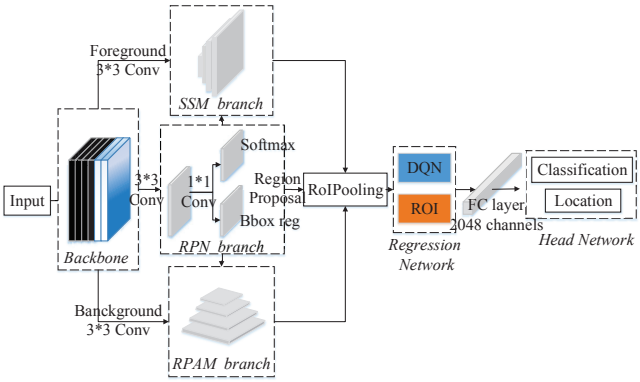

Fig. 1. Scene-level region proposal modle.

number of deep network parameters, a separable convolutional network was introduced into the backbone network structure, which reduced the number of network training parameters by nearly 9 times. In addition, in order to make up for the shortcomings of local feature extraction in object detection, after the convolution operation, an improved feature pyramid network (FPN) [10] processing is used to form a deeply separable convolutional network (D_SCNet-127 R-CNN, where 127 represents the number of network layers in the backbone network). Secondly, in the network part based on candidate region re-identification, the feature map output from FPN is sent to three parallel branches: the semantic segmentation module, the FPN module, and the region suggestion self-attention module [11], Thus, a scene-level regional suggestion self-attention [12] network is constructed. This construction improves the limitations of local feature extraction in traditional target detection and obtains more comprehensive details, thereby avoiding the limitations of local feature extraction.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

II. Proposed Model

A. Deep Separable Convolution Network Module

The deep separable convolution can achieve this well. It not only makes the whole network model more light-weight [13] (the training weight parameter is reduced), but the operation speed is faster (the floating point number operation is reduced) and the performance of the corresponding model is greatly increased by a few percentage points. The so-called depth separable convolution is actually used to solve the traditional volume integral into a deep convolution (condwise convolution) and a 1×1 convolution (pointwise convolution), and performing spatial convolution on each channel of the input. The output channels are then mixed by point-by-point convolution. In short, spatial feature learning [19] is separated from channel feature learning. Moreover, the amount of calculation required for the parameters is greatly reduced, and in the case of a small amount of data, it is a more efficient way to get better model performance. As shown in Figure 2, (a) is a standard convolutional layer filter, and (b) and (c) are a depth convolution and a 1×1 convolution of a depth separable convolution filter, respectively.

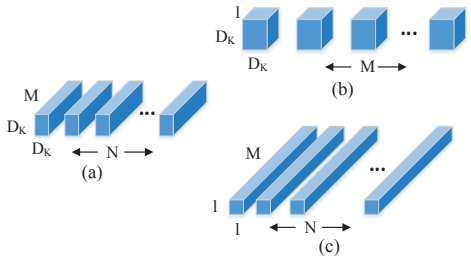

Fig. 2. (a) Standard convolutional layer filter, (b) Deep convolution of depth separable convolution filter, and (c) 1×1 convolution.

In order to compare the calculated volume of the standard convolution and the depth separable convolution (calculating amount, CA), the feature map sizes of the input and output are $D_F{*}D_F{*}M$, $D_F{*}D_F{*}N$, and the convolution kernel is $D_K{*}D_K$, where D_F is the size of the output feature map. Then the calculation amount of the standard convolution filter is $D_K{*}D_K{*}M{*}N{*}D_F{*}D_F$, and the calculation amount of the depth separable convolution filter is the depth convolution calculation amount and the 1×1 convolution calculation amount. Moreover, $D_K{*}D_K{*}M{*}N{*}D_F{*}D_F{+}M{*}N{*}D_F{*}D_F$. The calculation found that the ratio of the depth separable convolution to the traditional convolution calculation is $CA = (1/N) + (1/(D_K * D_K))$.

The computational amount of the corresponding depth separable convolution is reduced to $(1/N) + (1/(D_K * D_K))$ times the conventional convolution.

It consists of 6 stages in the deep separable convolutional network. The first four stages are composed of traditional residual convolution modules, each of which includes Conv-block and batch normalization (BN), a rectified linear unit (ReLU), max pooling, and Identity-block. In the next two stages, the module is composed of Conv-block, Depth-wise conv (Point-wise conv), BN, ReLU, max pooling, and Identity-block. Among the core modules of the proposed D_SCNet-127 R-CNN network structure, it still follows the ResNet network module structure. In addition, batch normalization and activation function layers are added after each convolution layer operation of the convolution module. On the one hand, BN can speed up the network training process, thereby reducing the number of training times, and greatly enhancing the training ability of deep networks. On the other hand, the activation function layer can make the output of some neurons 0, which increases the sparsity of the network and reduces the interdependence of parameters, thereby reducing the occurrence of overfitting problems.

First, we consider the ResNet-50 itself, which has excellent performance and is often used as a target network for detection. Therefore, this paper still retains the first four stages of ResNet-50 (1, 2, 3, 4) in the design of the backbone network. The image to be trained will pass through the backbone network of the deep separable network: first, at the first stage: through the $7 \times 7 \times 64$ convolution operation, BN, ReLU, and Max pooling, this can ensure that the output image is only the original 1/4, so we can have a large enough receptive field. At this stage, this paper still uses the operation of Mask R-CNN to make stage 1 only participate in pre-training. Each of the large layers in stages 2–4 is superposed and superimposed by the same residual module 1×1, 3×3, 1×1 convolution layer. The feature map output from the first stage is sent to the second stage for a deeper feature map extraction operation. The feature map extracted from the second stage is directly sent to the third stage for the feature map extraction operation. The fourth stage receives the feature map output from the third stage and performs feature extraction [18] for deeper and smaller targets.

In the backbone network, stages 5 and 6 are one of the innovations of this article. First, the sixth stage is added to the pre-training model in the backbone network of the deep separable convolutional network to ensure that the resolution of the feature map and the size of the receptive field can be guaranteed. Similarly, in order to obtain a higher feature map resolution and greater receptive field, stages 5 and 6 maintain the feature map size (1/16 of the original image size); this map is enlarged compared with the conventional feature map. Furthermore, each of these stages consists of a main path and a bypass. In the main path, the first three layers still retain the conventional residual convolution module; the fourth layer is a 3×3 convolution, 256-channel depth-wise convolution; then through BN and ReLU; followed by a 1×1 pointwise convolution, BN and ReLU, and Max pooling. In the bypass, in order to ensure that the input feature map can be added to the feature map output on the main road, a $1 \times 1 \times 256$ convolution operation is set, and the step

size is set to 2. It has been found that the use of deep separable convolution networks reduces the computational complexity of parameters by more than nine times compared to conventional convolutional networks. Therefore, increasing the depth separable convolutional network module reduces the pressure on the amount of computation and memory requirements to a certain extent.

The feature map outputs from the second stage to the fifth stage are obtained by the operation of the 1×1 convolution kernel, the 256 channels, the activation function, and the upsampling operation of the $2\times$upsample, obtaining the feature maps P2, P3, P4, P5, and P6. This operation is not performed in the first stage, it is mainly considered that the characteristic map output at this stage belongs to the shallow layer, and the error is large. Next, the feature maps P2, P3, P4, P5, and P6 are uniformly subjected to a 3×3 convolution operation through 256 channels. This operation is mainly used to eliminate the upsampling aliasing effect of the previous stage. Figure 3 shows the structure of the backbone network based on the depth separable convolution.

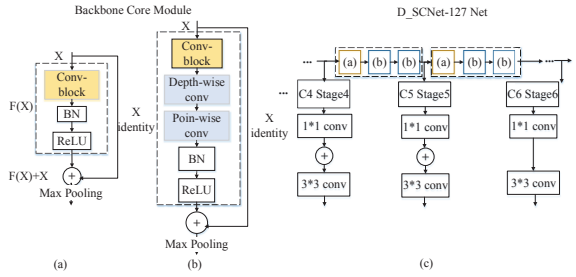

Fig. 3. D_SCNet-127 R-CNN network module.

B. Region Proposal Self-attention Network Module

First, after the training of the deep separable convolutional network model, the recognition process (scene-level and region proposal self-attention module) is divided into three parallel branches [14], namely, the semantic segment module (SSM) branch, the RPN (the structure in the Fast R-CNN) branch, and the region proposal self-attention module (RPAM) branch.

SSM branch: in order to obtain stronger semantic features and improve the performance of object detection, all levels of information from the FPN [15] are combined into a single output to achieve high-density prediction. RPN branch: generate candidate regions through the RPN, and use Softmax to judge foreground information and background information to further obtain accurate candidate boxes. RPAM branch: by introducing a self-attention mechanism, we try to complement foreground information and background information, and this self-attention module combines useful information from the RPN branch. This makes the detection task focus more on the local object to promote the accuracy of background semantics. After RPAM, a small structure called background selection is added to filter out useless background features, which can also be regarded as a small self-attention mechanism. SSM branch: image semantic segmentation is an important branch in the field of artificial intelligence and an important part of understanding images in machine vision technology. Early semantic segmentation methods just focus on the category of segmented objects and classify the semantics of each region (i.e., what the object is in this region). Among these methods, the n-cut(normalized cut) method is the most famous. It mainly considers the relationship weight between pixels comprehensively and divides the image into two parts according to a given threshold. With the improvement of computing power, we began to consider the use of machine learning methods for image semantic segmentation. GrabCut is a method that achieves segmentation by adding human–computer interaction in segmentation. When a full convolutional network (FCN) appears, it means that deep learning begins to enter into the field of image semantic segmentation. The FCN pursues that input is a picture, and output is also a picture, learning the mapping from pixel to pixel; however, this method significantly increases the computational complexity and memory burden, which limits the use of the master Type of dry network.

Therefore, in order to ensure the flexibility of the model itself, it is more important to maintain compatibility with its own model. This paper proposes to use improved FPN and FCN models to achieve semantic segmentation. Figure 4 shows the SSM branches. Traditional semantic segmentation solves the problem of image segmentation at the semantic level. The semantic segmentation branch is actually used to obtain the scene-level information in the feature to compensate for the disadvantages of focusing on the local information of the region of interest in traditional target detection. The image can be classified into pixels using a fully connected network, and the input feature map is fully connected to obtain a fixed-length feature vector. In order to better obtain strong semantic features, improve target detection performance, and achieve high-density prediction, all levels of information from the FPN are combined into a single output. According to the FPN feature generation semantic segmentation output, a simple design is proposed, combining all levels of information of the FPN pyramid into a single output. Each FPN level is upsampled by convolution and bilinear upsampling until the size becomes 1/4, and then these outputs are summed and finally converted to a pixel output.

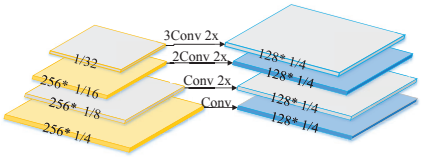

Fig. 4. SSM branch structure module.

Then, using the fully connected layer, the feature map output from the FPN network can be mapped to form a feature vector as a feature of a node in the graph model. We perform the same processing operations on the feature maps of all training images to obtain the scene information of the image. Then, each pair of regions of interest is mapped and transformed to perform cascading operations as edge elements in the graph model. In summary, through this structural reasoning method to iteratively update the node, the last state of the node is used to predict the category of the relevant region of interest and its location information.

The SSM branch receives the feature map from the output of the previous stage and then inputs the feature map to the fully connected layer [16] for processing. Thus, the size of the feature map can be unified, a feature vector can be formed on the input feature map by the operation of the fully connected layer, and the input feature map is convoluted to multi-scale features. The figure is uniformly dimensioned; the $2\times$ upsampling operation is performed on the feature map after the size is made uniform, and the feature maps of the same channel part are merged; then, the feature map is subjected to a convolution operation. The convolution operation is split into two matrices, and the convolution kernel and image are converted into a matrix function. Finally, the eigenvector of the entire feature map formed by the entire feature map is obtained and used as the graph model. Therefore, the global scene information of the image is acquired. Next, mapping transformation is performed on each pair of regions of interest in each feature map, and then a cascading operation is executed, and the relationship map vector between the regions of interest is used as a graph model edge element. Finally, the structure of the acquired graph model is iteratively updated node reasoning, and the last state of the corresponding node is used. The key category and location information related to the region of interest are then measured.

RPN branch: Generate candidate regions through the RPN, and use Softmax to determine foreground information and background information to further obtain accurate candidate frames. In the candidate area network branch, after the feature map is sent to the candidate area network, it will quickly and automatically generate deeper and more accurate category information and location information on the original feature map.

First, a 1×1 convolution operation is performed on the feature map, which can be used to adjust the dimension of the channel and reduce the amount of computation. Second, the feature map is subjected to softmax classification and bounding box regression. Finally, the obtained classification loss results are processed with the regression results to output a more accurate region of interest.

RPAM branch: Attempts to complement foreground and background information by introducing a self-attention mechanism, and this self-attention module applies information of the RPN branch to the RPAM branch. This allows the detection task to focus more on the local target to promote the accuracy of the background semantics. In addition to RPAM, a small structure called background selection is added to filter out unused background features; this structure can also be seen as a small self-attention mechanism. The regional recommendation self-care branch mainly re-identifies the background information through the proposal attention module (PAM) and achieves the goal of complementing the foreground and background information by merging the results of the branching of the candidate regional network. Thus, the accuracy of object detection is improved. The regional recommendation self-attention branch mainly aims to enhance the accuracy of the feature map by identifying the background information to achieve the complementary effect of the foreground and the background information. The self-attention mechanism is a mechanism that increases the fineness of the weak partial region by fusing the acquired feature map's own features with the obtained background information and outputs the classification loss (class_logits), correction loss (bbox_logits), and feature map of the image target.

First, using the knowledge of the self-attention mechanism to construct a self-attention branch of the region suggestion, the important features of the sparse data can be quickly extracted so that the background information features of the feature map can be obtained; then, the feature map input and the branch of the candidate region network are obtained. The extracted feature maps are fused to increase the amount of foreground information; then the background information is obtained, and the two types of information are merged. Thus, the accuracy of target detection is improved. The features acquired by the scene-level-region proposal from the three parallel sub-branches of the attention module are simultaneously sent to the region of interest (ROI) for pooling processing, thereby obtaining the fusion feature. This is shown in Figure 5.

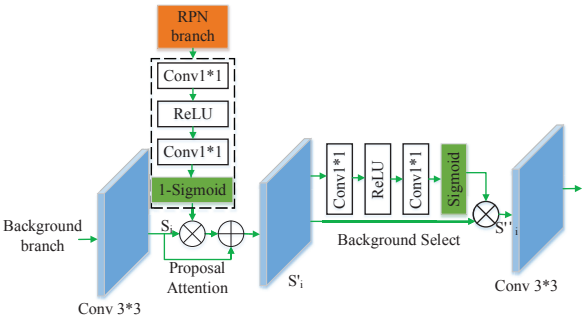

Fig. 5. RPAM structure module.

C. Border Regression Network Module

As an important research direction in the field of machine learning, reinforcement learning [17], [20], [21] defines an agent to interact with the environment continuously and decide the next action according to its observation of the

environment. Caicedo and others put forward an algorithm of target location based on deep reinforcement learning. The algorithm regards the whole picture as an environment and introduces an agent to learn the top-down search strategy for the border box. The agent can perform a series of simple deformation actions for the border box according to the learned strategy and finally locate the target accurately.

At present, the object detection algorithm based on deep learning relies too much on a large number of region proposals to improve the accuracy, and processing a large number of region proposals has become a bottleneck to improving detection speed. The object detection algorithm based on reinforcement learning can search region proposals selectively and thus significantly reduce the number of region proposals to be processed. However, considering that the search is mainly based on a certain proportion of the current region's area, it is challenging to improve accuracy. Therefore, this paper proposes a border regression network model of joint depth reinforcement learning. On the basis of the object detection framework based on deep reinforcement learning, this paper introduces a DQN network fusion ROI regression network. It can infer the location distribution of the whole target in the image according to the information extracted from the current region and achieve accurate positioning by the border regression of the current region. At the same time, it can improve the accuracy and speed of object detection. The whole network model framework in this paper consists of three parts: the backbone network (feature extraction network D_SCNet-127), the object recognition process based on the region proposal (scene-level region proposal self-attention network), and the head network (joint depth reinforcement learning frame regression network model) for post-processing.

As shown in Figure 6, first, the input image is extracted and

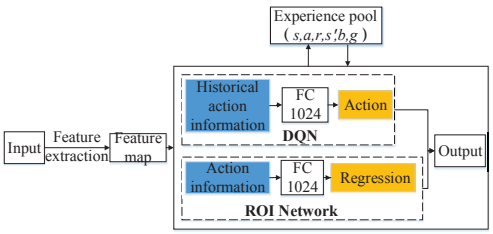

Fig. 6. Framework of regression network based on reinforcement learning.

re-identification by the pre-trained deep separable network model D_SCNet-127 and the scene-level region proposal self-attention module in turn. Second, the extracted feature vector is input into the DQN network, which is responsible for determining the search path. Finally, when the DQN network terminates the search, the ROI regression network is responsible for identifying the candidate according to the feature vector. The area is regressed by the border, and the final detection result is output. The training of the DQN network needs an experience pool to store a large number of experience samples, while the training of the ROI regression network needs a large number of experience samples whose IoUs are greater than a certain threshold. The DQN network focuses on solving the problem of the region exploration strategy, while the ROI regression network mainly improves the accuracy of candidate regions. The training data of the two networks are different from the optimization goal. In order to optimize the ROI regression network and DQN network, this paper constructs a complete reinforcement learning system and defines the specific actions, states, reward functions, and other system components for the target positioning task.

First, action set a is a set of actions that can be taken to achieve the goal. Nine actions are defined to search the region proposal, eight of which are used to transform the region proposal, and the other is used to terminate the search. Instructions for each action vary in proportion to the size of the current bounding box. In addition, the state set s represents the understanding of current environment information. S is a tuple composed of two vectors: a feature vector of the current observation area, and a vector with a fixed size representing the action history of the agent. Connecting two vectors in the RQN network will output a vector with nine dimensions, representing nine actions. The reward function R is a good or bad evaluation of the environment for the selected action in this state, and it is also the optimal strategy to guide the current state learning. The reward function is defined as follows:

$$R_{\mathrm{a}}(a, s \rightarrow s^{'}) = sign(IoU(b^{'}, g) - IoU(b, g) \qquad (1)$$

Here, IoU is the intersection ratio between the target g and the boundary box B:

$$IoU(b, g) = area(b \cap g)/area(b \cup g) \qquad (2)$$

where area () is the area function. For termination actions, the reward function is as follows:

$$R_t(s \rightarrow s^{'}) = \begin{cases} +\eta & if IoU(b, g) \geq \tau \\ -\eta & otherwise \end{cases} \qquad (3)$$

The experience stored in the experience pool is (s, a, R, s', B, g), where s is the current state, a is the X action taken, R is the immediate reward after performing action a in state s, s' is the next state to be converted, B is the coordinates of the current area, and G represents the coordinates of the real area of the target. The data used in ROI regression network training are (s, B, g), and the input data of the network are state s.

III. Experiments

A. Datasets and Evaluation Indicators

Experimental parameter setting: This paper mainly uses the ResNet-50 network as the reference network and evaluates the proposed method from local to global on MSCOCO17, VOC12, and Cityscapes datasets. Here, the MSCOCO dataset selects the standard COCO dataset indicators to evaluate the

experiment, including average precision (AP) and average recall (AR).

There are 80 object categories in the experimental dataset MSCOCO, including 80K images in the training set and 40K images in the test set and the verification set. In this paper, the 40K verification set is divided into 35K and 5K datasets. Then, the 80K training set and 35K verification set are combined to obtain a 115K training set and 5K small verification set. Meanwhile, the Cityscape dataset has 5K high-resolution images (1024×2048 pixels) with accurate pixel annotation: 2975 columns, 500 Val, 1525 test. In addition, there are 20K images with coarse annotation, which we did not use in the experiment.

B. D_SCNet-127 R-CNN Module Experiment

In order to lighten the model, speed up the calculation, and improve the detection performance, a deep separable network is used to replace some residual networks. In order to further compare the advantages of the deep separable convolutional network D_SCNet-127, it is verified from different perspectives on the MSCOCO dataset. Table I shows the mAP, calculation amount, and number of parameters under different network structures.

TABLE I
RESULTS OF IMPACT OF VARIOUS BACKBONE NETWORKS ON FPN ON MSCOCO (%).

Framework Resolution	Model	mAP	Billion Mult-adds	Million Paraments
SSD300	deeLab-VGG	21.1	34.9	33.1
	Inception V2	22.0	3.8	13.7
	D_SCNet-127	20.7	1.0	5.7
Fast R-CNN300	VGG	22.9	64.3	138.5
	Inception V2	15.4	118.2	13.3
	ResNet-50	18.9	19.8	6.7
	D_SCNet-127	22.4	20.6	5.3
Faster R-CNN600	VGG	25.7	149.6	138.5
	Inception V2	21.9	129.6	13.3
	ResNet-50	25.9	24.2	6.7
	D_SCNet-127	31.7	25.8	5.3
Faster R-CNN600	VGG	30.1	215.7	245.9
	Inception V2	28.9	175.6	35.7
	ResNet-50	39.1	45.1	12.4
	D_SCNet-127	52.7	60.9	10.6

First, the one-stage object detection model based on deep learning is compared with the two-stage object detection model based on deep learning, and different backbone networks are loaded in each model for comparative analysis. Table I shows the experimental results of the SDD network structure of the one-stage object detection and the two-stage object monitoring model (the Faster R-CNN network) after 300 iterations. The Faster R-CNN300 network loads the mAP after the D_SCNet-127 network. It has increased by nearly 2 percentage points, and the number of corresponding parameters has also decreased. Therefore, the performance of the two-stage object detection algorithm after loading the depth separable network model is improved compared with the performance of the one-stage object detection. Second, in order to further verify the validity of the model, we performed Faster R-CNN 300 times and 600 different iterations and compared it with Mask R-CNN600. Different backbone network models are loaded under the same network structure, and the D_SCNet-127 network model has much better performance than other models. Under the same two-stage object detection algorithm, it is iterated 600 times at the same time. Mask R-CNN600, after loading the D_SCNet-127 network model, has a 20% higher mAP than the Faster R-CNN600. Therefore, the deep separable convolutional neural network model constructed in this paper has a very friendly performance.

In addition, considering that only an example model of one-stage object detection is given in the comparative test of Table I. In order to make the model more convincing, Table II carries out one-stage object detection model method (SSD, Yolo), two-stage object detection model method (R-CNN, Fast R-CNN, Faster R-CNN, Mask R-CNN and Deeplab) and the proposed method (D_SCNet-127) experimental comparison.

Several pre-trained basic ImageNet models and our own network models were used in the experiment. Among them, the VGG16 network model is called V16. The ResNet50 network model is called R50. The table also shows the training time (Train Speed), the training rate (Test Rate), the test speed (Test Speedup), and the VOC07 and MSCOCO17 datasets. The average accuracies of models are compared.

The training and testing speed in the experiment are based on R-CNN model. First, it is found from Table II that in several two-stage object detection methods, when the basic network V16 is replaced with R50, all aspects of performance are improved. In the case of the D_SCNet-127 method and the Mask R-CNN method, the test speed is increased by 2.3 times in the same R50 network model. Table II intuitively shows that in the two-stage object detection, the proposed method achieves better results than the Mask R-CNN method in terms of detection speed, which also reflects the calculation amount and the number of parameters of the D_SCNet-127 model in Table I. There are many R-CNNs, but a good result can still be obtained in terms of detection speed.

C. Region Proposal Self-attention Network Module Experiment

This section is mainly composed of two parts. One is based on the scene-level semantic segmentation branch experiment, and the other is based on the region proposal self-attention network branch experiment. The previous section mainly uses the Cityscapes dataset. After randomly selecting 32 512×1024 image crops (4 crops per GPU) in the Cityscapes dataset, each scale is constructed by randomly scaling 0.5 to $2.0\times$. We train 65K iterations, starting at a learning rate of 0.01 and reducing it by 10 times in 40K and 55K iterations. This is different from the original Mask R-CNN setup but works well for both instance and semantic segmentation.

Semantic segmentation branch (SSM) experiment: This paper analyzes the effectiveness of using semantic segmentation branches for context information extraction. Table III

TABLE II
RESULTS OF IMPACT OF VARIOUS BACKBONE NETWORKS ON FPN GPU-BASED TRAINING AND TEST RATE ANALYSIS RESULTS FOR MULTIPLE MODELS MSCOCO (%).

Evaluation	R-CNN		Fast R-CNN		Faster R-CNN		Mask R-CNN		D_SCNet-127 R-CNN		SSD		$YOLO_{V3}$	
	V_{16}	R_{50}	V_{16}	R_{50}	V_{16}	R_{50}	V_{16}	R_{50}	V_{16}	R_{50}	V_{16}	R_{50}	V_{16}	R_{50}
Train Time(h)	84	75	9.5	8.0	8.7	7.7	44	15	12.5	11.3	10.0	8.4	10.4	6.4
Train Speedu(×)	1	1.12	8.8	10.5	9.6	10.9	1.9	5.6	8.7	9.6	8.4	10.0	8.0	13.0
Test Rate(s/im)	47	5	0.32	0.25	0.14	0.11	0.2	0.07	0.047	0.03	0.045	0.038	0.047	0.029
Test Speedup(×)	1	9.4	146	188	335	427	522	671	1025	1375	1044	1236	1000	1620
Voc07+COCO17	45.7	49.6	65.1	67.5	70.4	77.6	70.5	77.3	81.6	84.5	76.3	78.1	57.9	73.8

compares the features of the Faster R-CNN, Mask R-CNN, and D_SCNet-127 network models. In the network structure of Mask R-CNN, the object detection experiment is based on the backbone networks of ResNet-50 and ResNet-50-SSM. The network model after loading the SSM branch improves the accuracy of feature extraction by 2.4%, which also shows that the SSM branch is used to fuse the validity of the foreground information in feature extraction. Table III also shows the feature extraction accuracy rate after loading the SSM branch under the D_SCNet-127 network model; the feature extraction accuracy rate is 3.6% higher than before loading. Therefore, it is better to compare the feature segmentation before the feature extraction.

TABLE III
TRAINING RESULTS FOR VARIOUS MODELS BASED ON SSM ON CITYSCAPES(%).

Models	Backbone	mAP
Faster R-CNN	ResNet-50	42.5
Mask R-CNN	ResNet-50	51.2
	ResNet-50-SSM	53.6
D_SCNet-127	ResNet-50	52.3
	ResNet-50-SSM	55.9

Region proposal self-attention module (RPAM) experiment: This paper obtains the global information of the object by introducing a self-attention mechanism in the region proposal module and combining the information of the region of interest in the RPN branch with the semantic information of the foreground in the SSM branch. In order to verify the validity of the RPAM branch in the experiment, the experiments using Faster R-CNN, Mask R-CNN, and D_SCNet-127 network model were compared. As can be seen in Table IV, the introduction of panoptic quality (PQ) as one of the evaluation indicators in the experiment can be interpreted as the product of segmentation quality (SQ) and recognition quality (RQ). We evaluate the accuracy of the RPAM branch in extracting the context information, and IoU is the ratio of the predicted target area to the real area.

Several network models have achieved better results on AP, PQ, and IoU after loading the RPAM branch. The D_SCNet-127 network model has an AP increase of 3.1% and a corresponding PQ increase of 4.1% with the backbone based on ResNet-50 and ResNet-50-RPAM. Therefore, the RPAM branch has a significant effect on the feature extraction.

The scene-level-based region proposal self-attention network is of great significance in the object re-identification process of object detection. The drawbacks that have limited local feature extraction have been addressed, and object detection experiences a qualitative breakthrough in performance.

TABLE IV
TRAINING RESULTS OF VARIOUS MODELS WITH RPAM ON CITYSCAPES DATASET(%).

Models	Backbone	PQ	mAP	IoU
Faster R-CNN	ResNet-50	-	27.8	-
Mask R-CNN-50	ResNet-50	-	31.5	-
D_SCNet-127	ResNet-50	56.1	34.5	74.6
D_SCNet-127	ResNet-50-RPAM	60.2	37.6	77.3

D. Border Regression Network Module Experiment

In this paper, the DQN network is introduced into the ROI regression network through joint deep reinforcement learning to obtain a more accurate target location. In order to test the effectiveness of the network, this paper uses the training set data of VOC07 and VOC12 and uses the test set of VOC07.

Table V shows the experimental results of the accuracy and target localization of several algorithms on the VOC dataset. The experimental results show that the proposed algorithm has obvious advantages in single-category target detection compared to other target-based detection algorithms based on reinforcement learning. Among the airplane categories, the proposed algorithm has better accuracy than the algorithm of Bueno et al. 20.55%. Compared with the algorithm proposed by Caicedo et al., the accuracy of our algorithm is increased by 15.38%. In the dog category, compared with the algorithms of Bueno and Caicedo, the accuracy of our algorithm is increased by 17.46% and 15.07%, respectively. Thus, the proposed algorithm can effectively improve the accuracy of target positioning. Similarly, in the airplane category of target location detection, the proposed algorithm has a maximum improvement of 21.22% in positioning accuracy compared with the algorithm proposed by Bueno and Caicedo. In the dog category, the algorithm is located in the algorithm of Bueno and Caicedo. The accuracy increased by a maximum of 22.40%. This shows that the boundary regression network combined with deep reinforcement learning has great advantages.

In order to further understand the network model of this paper, Figure 6 illustrates the object detection for the

TABLE V
BORDER REGRESSION PERFORMANCE RESULTS OF MULTIPLE BACKBONE NETWORKS BASED ON MSCOCO DATASET(%).

Algorithms	Accuracy		Positioning Accuracy	
	areoplane	dog	areoplane	dog
Bueno [21]	32.23	32.19	32.23	32.18
Caicado [20]	37.40	34.58	37.41	34.58
Our algorithm	**52.78**	**49.65**	**53.45**	**56.98**

aircraft category under simple background conditions. For normal size objects, as shown in Figure 7(b) and (d), the model only needs a few search steps to locate the target aircraft. For large-scale targets, as shown in Figure 7(a), the DQN network can accurately locate the target position by performing only one search action according to the current regional characteristics and can then accurately locate the target area through the regression network. For small-scale targets, as shown in Figure 7(c), due to the small targets, the DQN network will continue to search in the direction of the target area until sufficient information is collected. Then, the search action will be terminated, the current area will be determined as the target area (as shown by the smallest green box), and the target location will be more accurately located by the regression network.

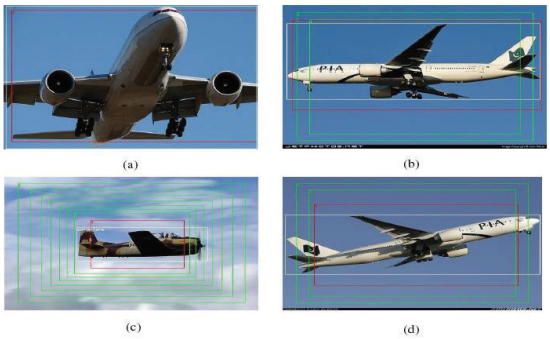

Fig. 7. Object detection and positioning against a simple background. The green box represents the candidate area generated by DQN network each time, the red box represents the final positioning result obtained by combining the regression networks, and the white box represents the real target area.

IV. CONCLUSIONS

The D_SCNet-127 R-CNN model of this paper carries out verification experiments from three parts. A deep separable convolutional network is proposed on the backbone network. On the one hand, the entire network model is lightened, and the operation speed can be improved. In the object re-identification network, the parallel branch of SSM and RPAM is built on the original network of RPN, and a new scene-level area is proposed to re-identify the network. Finally, joint deep reinforcement learning optimizes the object location performance of the regression network. Through the comparison and analysis of several experiments, the model of this paper has achieved very effective results in accuracy.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (Nos. 61966004, 61663004, 61866004, 61762078), the Guangxi Natural Science Foundation (Nos. 2019GXNSFDA245018, 2018GXNSFDA281009, 2017GXNSFAA198365), the Guangxi "Bagui Scholar" Teams for Innovation and Research Project, the Guangxi Talent Highland Project of Big Data Intelligence and Application, Guangxi Collaborative Innovation Center of Multi-source Information Integration and Intelligent Processing.

REFERENCES

[1] Wang X, Han T X, Yan S. An HOG-LBP human detector with partial occlusion handling, in 2009 IEEE 12th international conference on computer vision. IEEE, 2009: 32-39.
[2] Lin K, Yang H F, Hsiao J H, et al. Deep learning of binary hash codes for fast image retrieval, In: Proc. CVPR, 2015: 27-35.
[3] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation, In: Proc. CVPR, 2014: 580-587.
[4] Ross Girshick, 'Fast r-cnn', in Proceedings of the IEEE international conference on computer vision, pp. 1440–1448, 2015.
[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, 'Faster r-cnn: Towards real-time object detection with region proposal networks', In: Proc. NIPS, pp. 91–99, 2015.
[6] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun, 'R-fcn: Object detection via region-based fully convolutional networks', In: Proc. NIPS, pp. 379–387, 2016.
[7] He K, Gkioxari G, Dollár P, et al. Mask r-cnn, in Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.
[8] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, 'You only look once: Unified, real-time object detection', In: Proc. CVPR, pp. 779–788, 2016.
[9] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In European conference on computer vision, pages 21–37. Springer, 2016.
[10] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection, In: Proc. CVPR, 2017: 2117-2125.
[11] Fan H, Ling H. Siamese cascaded region proposal networks for real-time visual tracking, In: Proc. CVPR, 2019: 7952-7961.
[12] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu, 'Dual attention network for scene segmentation', In: Proc. CVPR, pp. 3146–3154, 2019.
[13] Quan Y, Li Z, Zhang F, et al. D_dNet-65 R-CNN: Object Detection Model Fusing Deep Dilated Convolutions and Light-Weight Networks, In: Proc. PRICAI, Springer, Cham, 2019: 16-28.
[14] Liu Y, Wang R, Shan S, et al. Structure inference net: Object detection using scene-level context and instance-level relationships, In: Proc. CVPR, 2018: 6985-6994.
[15] Kirillov A, Girshick R, He K, et al. Panoptic feature pyramid networks, In: Proc. CVPR, 2019: 6399-6408.
[16] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation, In: Proc. CVPR, 2015: 3431-3440.
[17] Mathe S, Pirinen A, Sminchisescu C. Reinforcement learning for visual object detection, In: Proc. CVPR, 2016:2894-2902.
[18] Steder B, Rusu R B, Konolige K, et al. Point feature extraction on 3D range scans taking into account object boundaries, In: Proc. ICRA, 2011: 2601-2608.
[19] Ghiasi G,Lin T Y,Le Q V. Nas-fpn: Learning scalable feature pyramid architecture for object detection, In: Proc. CVPR, 2019: 7036-7045.
[20] Caicedo J C, Lazebnik S. Active object localization with deep reinforcement learning, In: Proc. CVPR, 2015:2488-2496.
[21] Bellver, Miriam, et al. "Hierarchical object detection with deep reinforcement learning." arXiv preprint arXiv:1611.03718 (2016).