SSD Object Detection Algorithm Based on Feature Fusion and Channel Attention

Leilei Fan
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, Shaanxi, China
E-mail: 2547462712@qq.com

Zhiyi Hu
Engineering Design Institute
Army Research Loboratory
Beijing, 100042, China
E-mail: 18992899862@163.com

Jun Yu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, Shaanxi, China
E-mail: yujun@xatu.edu.cn

***Abstract*—Aiming at the problems of low object detection accuracy due to complex background and insufficient semantic information of shallow features in the object detection SSD algorithm, this paper improves the existing SSD algorithm. First, the original vgg16 network is replaced by the ResNet50 network, and the residual network structure as well as the Batch Normalization layer are added, which are used to improve the accuracy of the feature extraction network; Second, a feature fusion module is designed to fuse adjacent feature maps to improve the detection effect by integrating contextual information; Third, the SE attention mechanism is introduced to give channel weights adaptively and enhance the useful feature channels; Finally, the object detection analysis experiments are conducted on the PASCAL VOC2012 dataset. The experimental results show that the improved SSD algorithm in this paper is able to achieve an mean average precision of 72.7% in the data set, which is 2.1% better than the original SSD-VGG16 and greatly improves the object detection effect.**

Keywords-Object Detection; Feature Fusion; Attention Mechanism; SSD Algorithm

I. INTRODUCTION

Object detection has become one of the important research directions in the field of computer vision. Its essence is to find the desired target in a complex background image, give the location information of the target, and judge its class. Object detection technology is widely used in face recognition, medical field, traffic research and other aspects [1].

Object detection based on deep learning can be divided into two groups: two-stage object detection method based on region proposal and one-stage object detection method based on regression [4].In the two-stage series, represented by a series of algorithms of regional convolutional neural network (r-cnn), the task of the first stage is to generate a group of target candidate regions, and then send these candidate regions to the second stage. Then coordinate regression and classification are carried out to gradually realize end-to-end object detection. Although the detection accuracy has been improved, the detection speed is slow due to its large network parameters [6].One-stage series algorithms are represented by single-detector YOLO and single-network multi-scale detector SSD. It discards the stage of extracting candidate regions in the two stage method, and directly obtains the category probability and location of the target, making its network structure simpler. Compared with the two stage object detection method, the detection speed of this method has been improved to a certain extent, but the localization accuracy has decreased,

DOI: 10.2478/ijanmc-2022-0029

and there is still a problem that the model parameters are too large [6].

SSD [20] (full name: Single Shot Detection) is a typical algorithm of one-stage object detection method. In order to detect objects of different scales, it uses shallow feature maps to detect small objects, and uses deep feature maps to detect large objects.

The SSD algorithm encapsulates the target location and target prediction in the forward operation, and can directly generate the category probability and position coordinate value of the object. The final detection result can be obtained by only one step detection. Although the detection speed is faster, the positioning accuracy is decreased. SSD algorithm uses shallow high-resolution feature layer. Due to the lack of feature expression ability of this layer, there may be missed detection and false detection when detecting small-scale targets.

In order to avoid the above problems, this paper improves the standard SSD object detection algorithm. The residual network structure is introduced, and the adjacent feature maps are fused. At the same time, the SE attention mechanism experiment is carried out on the PASCAL VOC2012 dataset.

II. SSD ALGORITHM PRINCIPLE

Liu et al. proposed SSD algorithm [20] in 2015, which combines the advantages of Yolo's fast speed and fast rcnn's accurate positioning. The main innovations of SSD algorithm are as follows: (1) extracting feature maps at different scales; (2) using prior boxes with different aspect ratios. These two important improvements enable SSD algorithm to efficiently detect targets with different scales [20].

The network structure of SSD is shown in Figure 1.Its detection framework [20] consists of two sections: feature extraction network and multi-scale feature detection network. The feature extraction network adopts the VGG16 network structure for the preliminary extraction of image features. Since the SSD network only needs to extract features without classification in this part, the fully connection layer in VGG16 is replaced by the convolution layer. The second part is used to detect the predicted feature maps of different scales generated by the feature extraction network. The spatial resolution of the shallow feature map is higher than that of the deep feature map, which can more accurately identify the detailed information such as the edge, contour and texture of the image [18].The deep feature map has a large receptive field and strong information representation ability, but it lacks detailed information compared with the shallow feature map detect smaller sized targets and the deep feature map detect larger targets.

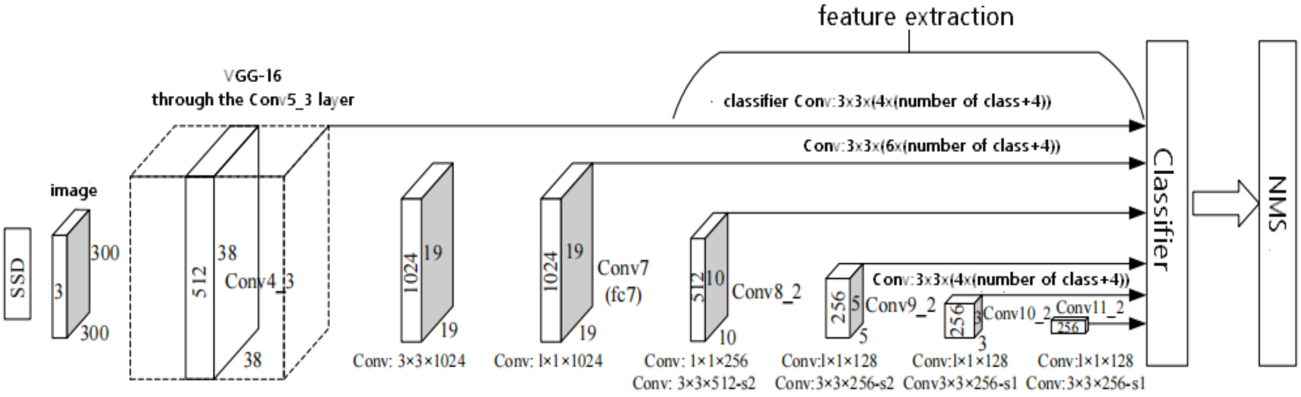

Figure 1. Schematic diagram of SSD network structure

A. Feature extraction network

SSD adopts multi-scale feature detection in object detection. The size of the input image can be 300×300、512×512 [18], etc. The original SSD basic network uses VGG16 as the backbone network. The multi-scale pyramid feature map is generated by adding several convolution layers with gradually reduced size to the basic network: First, the FC7 layer in VGG16 is replaced by the convolution layer Conv7, and all dropout layers and FC8 layers are removed; Secondly, the feature layers Conv8, Conv9, Conv10 and Conv11 are added. The SSD network uses the Conv4_3 layer in vgg16 as the first prediction feature map, and the feature maps obtained from the Conv7, Conv8_2, Conv9_2, Conv10_2, and Conv11_2 layers as the subsequent prediction feature maps. So, a total of 6 prediction feature maps with different scales are obtained. Taking the input image size of 300×300 as an example, the size of the six predicted feature maps obtained by the feature extraction network are 38×38, 19×19, 10×10, 5×5, 3×3, 1×1.

B. Multi-scale feature detection network

For the six predicted feature maps, each cell of each predicted feature map will extract k = 4 to k = 6 default boxes according to different aspect ratios, and finally obtain 8732 prior boxes. If the size of a feature map is H×W, it means that H×W×k default boxes will be generated on this feature map. The k values on the six predicted feature maps are: 4, 6, 6, 6, 4, 4.The size and number of default boxes generated by each feature map are shown in Table Ⅰ.

TABLE I. DEFAULT DIMENSION AND QUANTITY OF FEATURE MAP

Feature Map	Width and height of the feature map	Default boxes size	Number of default boxes
Feature Map1	38 x 38	21{1/2,1,2}; $\sqrt{21\,\mathrm{x}\,45}$ {1}	38 x 38 x 4
Feature Map2	19 x 19	45{1/3,1/2,1,2,3}; $\sqrt{45\,\mathrm{x}\,99}$ {1}	19 x 19 x 6
Feature Map3	10 x 10	99{1/3,1/2,1,2,3}; $\sqrt{99\,\mathrm{x}\,153}$ {1}	10 x 10 x 6
Feature Map4	5 x 5	153{1/3,1/2,1,2,3}; $\sqrt{153\,\mathrm{x}\,207}$ {1}	5 x 5 x 6
Feature Map5	3 x 3	207{1/2,1,2}; $\sqrt{207\,\mathrm{x}\,261}$ {1}	3 x 3 x 4
Feature Map6	1 x 1	261{1/2,1,2}; $\sqrt{261\,\mathrm{x}\,315}$ {1}	1 x 1 x 4

After obtaining all default boxes, we will calculate the scores of c categories and 4 coordinate offsets (boundary box regression parameters) for each default box: For the M×N, P-channel predicted feature map, two 3×3 convolution kernels of P channels are used to generate probability scores and coordinate offsets of corresponding default boxes. Let s be the number of all default boxes on these six prediction feature maps, then the number of convolution kernels required for calculating the category is s × c, and the number of convolution kernels required to calculate the coordinate offset is s ×4.

The loss function of multi-scale feature detection network is divided into two parts: category loss and position offset loss. The calculation formula of the target loss function is shown in (1). Among them, N is the number of matched positive samples, α is 1, $L_{conf}(x,c)$ is the predicted category loss, $L_{\mathrm{loc}}(x,l,g)$ is the position offset loss.

$$L(x,c,l,g)=\frac{1}{N}(L_{conf}(x,c)+\alpha L_{loc}(x,l,g)) \quad (1)$$

For the predicted category loss, the calculation formula is shown in (2).

$$L_{\text{conf}}(x,c) = -\sum_{i\in Pos}^{N} x_{ij}^{p}\log(\hat{\mathrm{c}}_i^p) - \sum_{i\in Neg}\log(\hat{\mathrm{c}}_i^0)$$
$$\text{where } \hat{c}_i^p = \frac{\exp(c_i^p)}{\sum_p \exp(\hat{c}_i^p)} \tag{2}$$

Among them, $-\sum_{i\in Pos}^{N} x_{ij}^{p}\log(\hat{\mathrm{c}}_i^p)$ is the category loss of positive samples and $-\sum_{i\in Neg}\log(\hat{\mathrm{c}}_i^0)$ is the category loss of negative samples;

$\hat{c}_i^p$ is the category probability p of predicting the GT box corresponding to the i-th default box.

x_{ij}^p ={0,1} is the j-th GT box (category is p) matched by the i-th default box.

$\hat{c}_i^0$ is the probability that the i-th default box is predicted to be the background.

The equation of positional offset loss $L_{\text{loc}}(x,l,g)$ is shown in (3). Among them, l_i^m is the predicted regression parameters (center coordinate x, center coordinate y, width w, height h) corresponding to the i-th positive sample; g_j^m is the coordinate of the j-th GT box matched to the positive sample i; d_i^m is the coordinate of the i-th default box; $\hat{\mathrm{g}}_{\mathrm{j}}^{\mathrm{m}}$ is the regression parameter calculated from the coordinates of the i-th default box and the coordinates of the j-th GT box matched to it.

$$L_{loc}(x,l,g) = \sum_{i\in Pos}^{N}\sum_{m\in\{cx,cy,w,h\}} x_{ij}^{k}\, smooth_{l1}(l_i^m - \hat{g}_i^m)$$
where
$$\hat{g}_j^{cy} = (g_j^{cy} - d_i^{cy})/d_i^h,\ \ \hat{g}_j^{cy} = (g_j^{cy} - d_i^{cy})/d_i^h$$
$$\hat{g}_j^{w} = \log(\frac{g_j^w}{d_i^w}),\ \ \hat{g}_j^{h} = \log(\frac{g_j^h}{d_i^h}) \tag{3}$$
$$smooth_{l1}(x) = \begin{cases} 0.5x^2, |x|<1 \\ |x|-0.5,\ \ other \end{cases}$$

III. IMPROVED SSD NETWORK STRUCTURE

The improvements in this paper based on the SSD algorithm include:

1) In this paper, ResNet50 network is used as a feature extraction network. First, the step of the first Residual structure of the Conv4 layer in the ResNet50 network is changed from 2 to 1; second, the structures of Conv5, Avg Pool, FC and Softmax behind the Conv4 layer are deleted, so the first predicted feature map is obtained; finally, five downsampling operations are added after the first predicted feature map. Among them, the Batch Normalization layer is added after each convolution operation; in addition, the bias parameter terms in the convolution operation and the weight decay of the parameters of the Batch Normalization layer are removed. The network structure is shown in Figure 2.

2) A total of 6 predicted feature maps are obtained from the operation of 1). To enhance the semantic information of the feature maps, the shallow feature maps are fused with the deep feature maps to obtain the new prediction feature maps. The new prediction feature maps are input to the network for subsequent operations.

3) Introducing the SE attention mechanism. After the fused predicted feature maps are input to the SE attention module, the SE attention mechanism obtains the importance of each channel in the feature maps. Based on this importance level, the SE attention mechanism assigns a weight value to each feature channel so that the feature channels that are useful for the current task are enhanced and the feature channels that are not useful for the current task are suppressed.

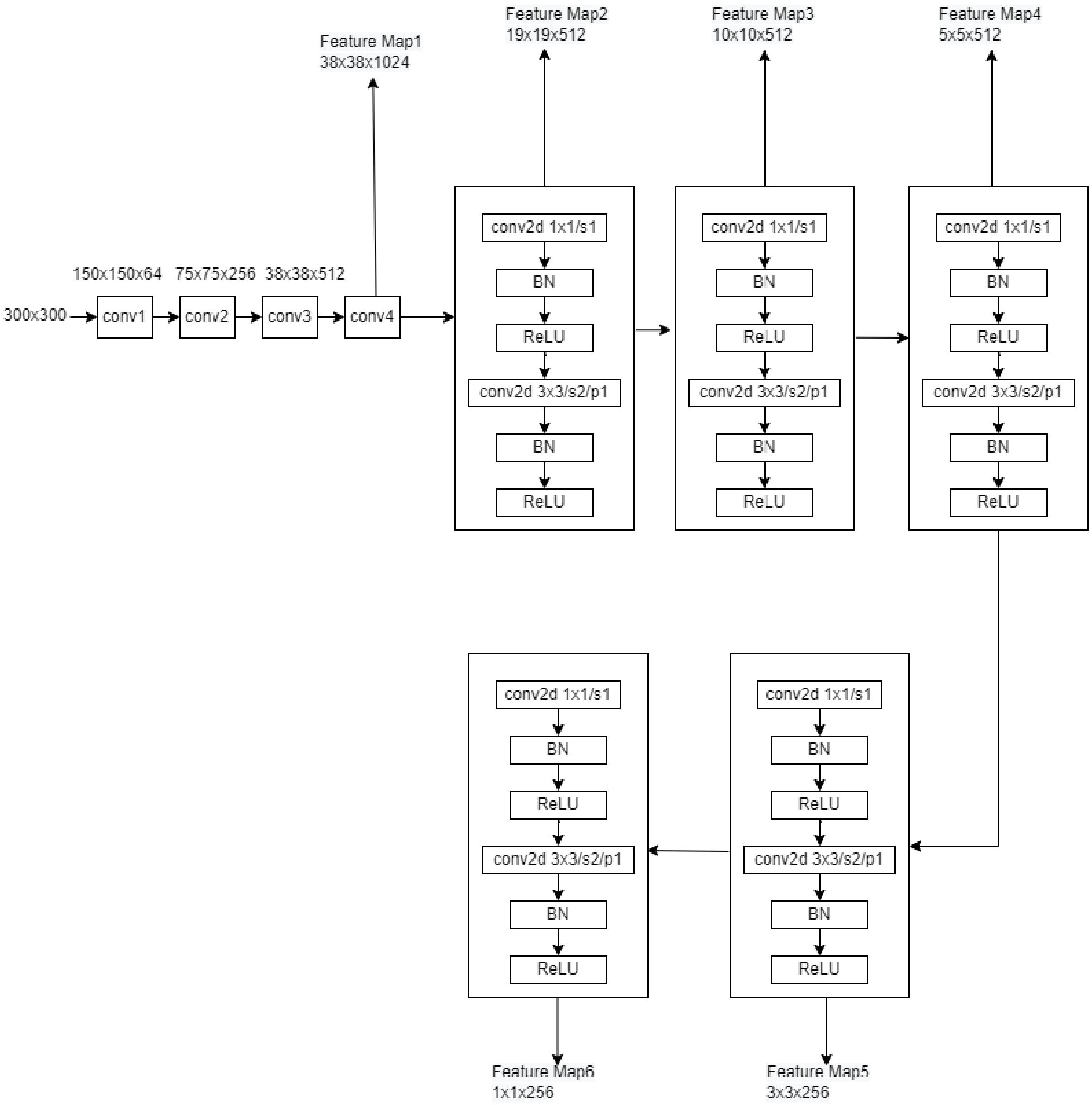

Figure 2. Schematic diagram of the base network with improved SSD

A. The Improvement of Backbone Network

In this paper, the backbone in the SSD network is replaced by the VGG16 network with the ResNet50 network. In the object detection algorithm, the selection of the classification network has a significant impact on the performance of the algorithm. Removing the fully connected layer and loss layer from the classification network will get the basic network part of the object detection algorithm.VGG16 consists of 13 convolutional layers, 3 fully connected layers and 5 pooling layers. The outstanding feature of vgg16 is its simple structure, which is easier to build by stacking several convolutional and pooling layers. However, the disadvantage of VGG16 is that the number of layers is shallow and the feature extraction is not sufficient[8].The ResNet50 network consists of a series of Residual structures and uses Batch Normalization to accelerate the training; and the ResNet50 network is much smaller than VGG16, and its speed and accuracy are superior to VGG16. Therefore, in this paper, ResNet50 is chosen to replace the VGG16 network in the SSD network.

B. Feature Fusion

The shallow feature maps are large in size and contain sufficient detail information, but too little semantic information. The deep feature maps have large receptive fields and contain sufficient semantic information, but detail information is gradually lost as the receptive field increases [3]. Therefore, fusing features of different scales is an important means to improve the performance of object detection. In this paper, bilinear interpolation is used to achieve upsampling, which increases the image resolution and retains more feature information; the concatenate method is used to stitch the shallow feature map with the deep feature map in the depth direction. The feature fusion method is shown in Figure 3.

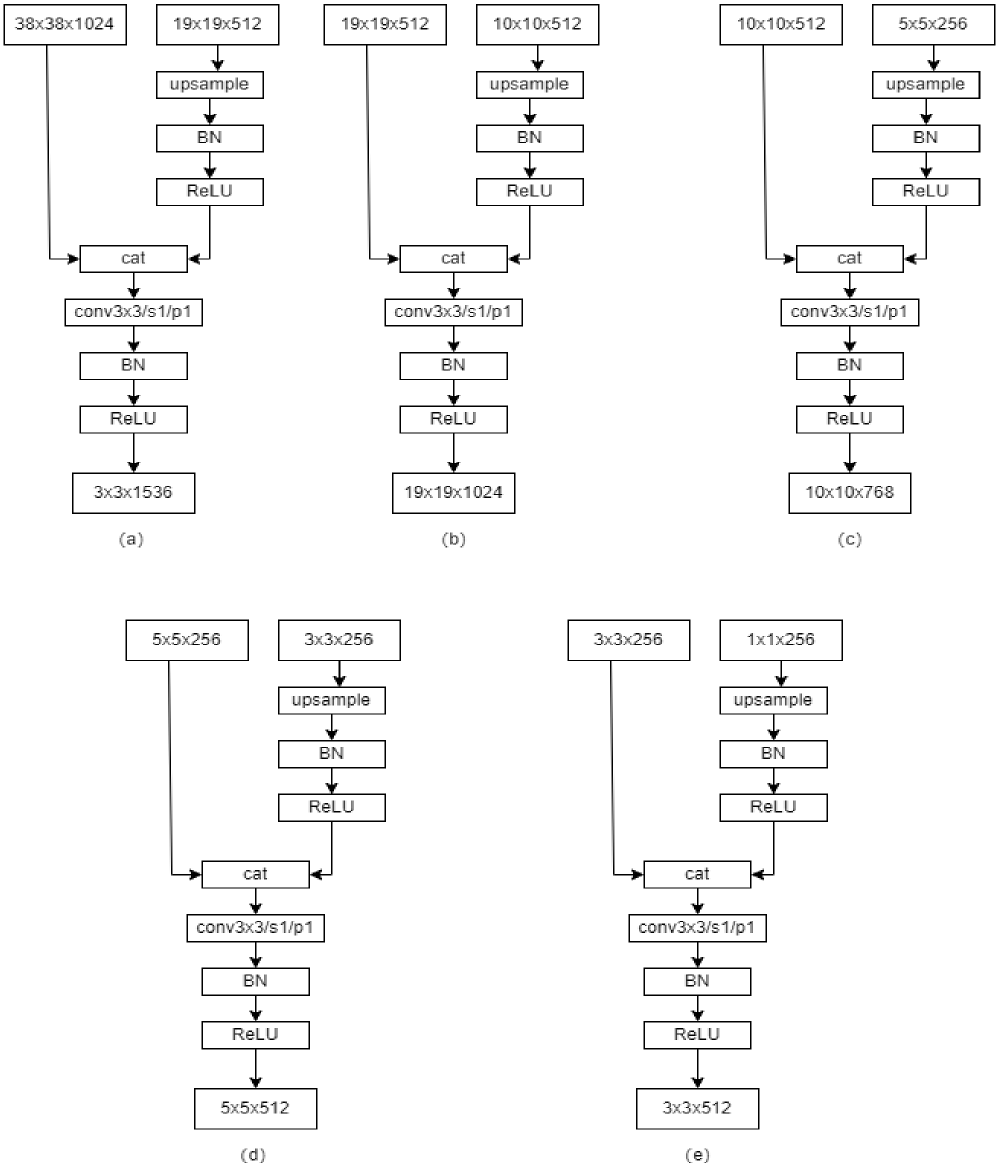

Figure 3. Schematic diagram of the fusion of shallow features and deep features

Taking the fusion of Conv4 feature layer and Conv7 feature layer as an example, as shown in Figure 3(a), the specific fusion is as follows:

1) First, the Conv7 feature layer is first upsampled by bilinear interpolation, Batch Normalization and linear activation (ReLU), so that the width and height of the feature map of this layer are the same as the width and height of the feature map of the Conv4 layer. Then, the feature map of size 38×38×512 is obtained.

2) Second, the feature map obtained from the Conv7 layer is concatenated with the feature map from the Conv4 layer to obtain a feature map of size 38×38×1536.

3) Finally, the feature map of size 38×38×1536 is subjected to a 3×3 convolution, batch normalization and linear activation (ReLU) to obtain the final first predicted feature map.

In Figure 3(b), the feature maps of the Conv7 layer are fused with the feature maps of the Conv8_2 layer, and the obtained feature maps are used to replace the feature maps of the Conv7 layer.

In Figure 3(c), the feature maps of the Conv8_2 layer are fused with the feature maps of the Conv9_2 layer, and the feature maps of the Conv8_2 layer are replaced with the resulting feature maps.

In Figure 3(d), the feature maps of the Conv9_2 layer are fused with the feature maps of the Conv10_2 layer, and the feature maps of the Conv9_2 layer are replaced with the obtained feature maps.

In Figure 3(e), the feature maps of the Conv10_2 layer are fused with the feature maps of the Conv11_2 layer, and the feature maps of the Conv10_2 layer are replaced with the obtained feature maps.

The operation flow of their fusion is similar to Figure 3(a), and the final size of the new predicted feature layers are obtained as:

38×38×1536,19×19×1024,10×10×768,5×5×512,3×3×512,1×1×256.

C. SE attention mechanism

SENet emerged to solve the loss problem caused by the different importance of different channels of the feature map in the convolutional pooling process. In the traditional convolutional pooling process, each channel of the feature map is equally important by default. In real-world problems, the importance of different channels varies. SENet can adaptively recalibrate the channel feature responses by explicitly modeling the interdependencies between channels. The role of SENet is to obtain the weights of each channel of the incoming feature map, allowing different channels to have different effects on the task results with different weights. So, the use of SENet allows the network to focus on the channels that need to play the most role in the detection task.

The process of SEblock is divided into two steps: Squeeze and Excitation.

(1) Squeeze: Obtaining the global compressed feature volume of the current feature map by performing Global Average Pooling on the feature map layer.

(2) Excitation: The weights of each channel in the feature map are obtained by a two-layer fully connected bottleneck structure, and the weighted feature map is used as the input of the next layer of the network.

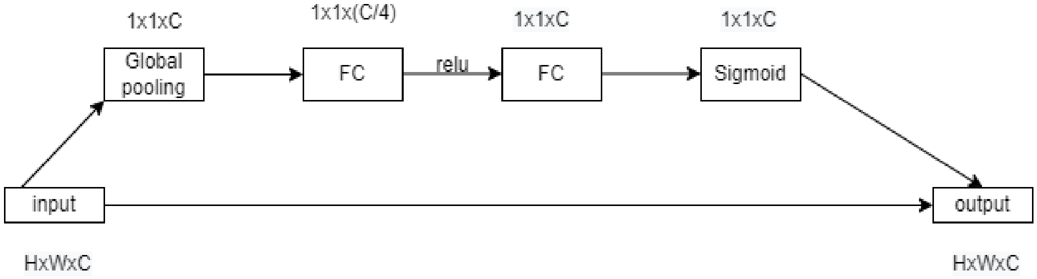

Figure 4. Schematic diagram of SE module

The four steps of the SE attention mechanism are illustrated in Figure 4, as follows:

(1) Performing global average pooling of the input feature maps.

(2) Performing two full connections, the first with a smaller number of fully connected neurons and the second with the same number of fully connected neurons as the number of channels of the input feature map.

(3) After completing two full connections, the weights (between 0 and 1) of each channel in the input feature map are obtained by performing another Sigmoid to fix the value between 0 and 1.

(4) After obtaining this weight value, multiply this weight value by the input feature map.

The feature map contains a large number of channels, and the judgment of the network varies from channel to channel. Some channels contain rich information and some channels hardly play a role. The purpose of introducing the SE attention mechanism is to make full use of the feature channels that contain important information. The SE attention mechanism is added to the back of the six predicted feature maps obtained after feature fusion, so that it adaptively assigns weights to each channel in the feature maps, strengthening the useful feature channels while suppressing the useless ones. The SSD model with the added SE attention mechanism is shown in Figure 5.

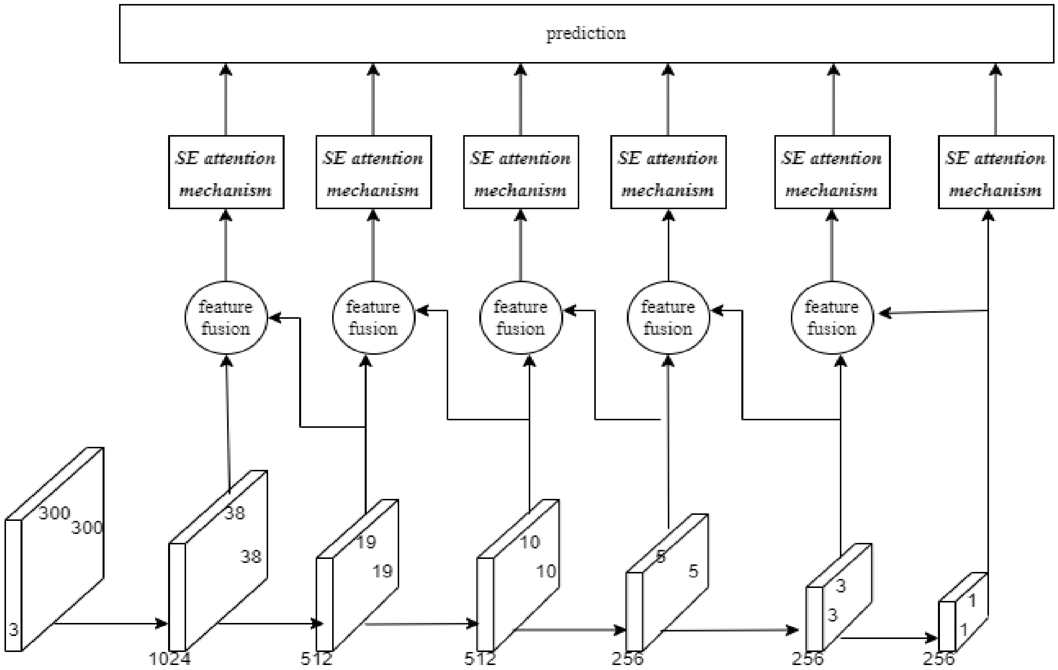

Figure 5. Schematic diagram of the network model with improved SSD

IV. Experimental procedure and analysis of results

A. Preparation of dataset and development environment

This experiment uses the PASCAL VOC2012 dataset, which contains 11530 images and 27450 targets labeled. There are 20 categories of targets to be recognized in this dataset, which are divided into four main categories: Vehicles, Household, Animals, and People; Vehicles class include Aero plane, Bicycle, Boat, Bus, Car, Motorbike, and Train; Household class contains: Bottle, Chair, Dining table, Potted plant, Sofa, TV/Monitor; Animals class contains: Bird, Cat, Cow, Dog, Horse, Sheep. The images in the dataset are annotated with corresponding XML files for the location and class of the target [4].

The environment used for the experiments is shown in Table Ⅱ.

TABLE II. ENVIRONMENT CONFIGURATION TABLE

Hardware	**Processor** **Video Cards**	**Intel(R)Core(TM) i7-6500U** **GeForce_RTX_2080_Ti**
Software	Operating System	windows10
	Deep Learning Framework	pytorch-gpu
	Compiler Language	python
	Compilers	pycharm

B. Evaluation Indicators

To comprehensively evaluate the accuracy of the SSD algorithm in detecting targets, this paper chooses to use the Mean Average Precision (mAP) as the evaluation criterion. MAP represents the average of all Average Precision (AP), and each Average Precision(AP) is measured using the intersection over Union (IOU). Samples with IOU above the threshold are positive samples, and samples below the threshold are negative samples. The calculation of AP requires Precision, Recall, as shown in (4):

$$\begin{aligned} &\text{Presion} = \text{TP}/(\text{TP}+\text{FP}) \\ &\text{Recall} = \text{TP}/(\text{TP}+\text{FN}) \\ &AP = \int_0^1 p(r)dr \end{aligned} \tag{4}$$

Among them,TP is the positive sample with positive prediction, FP is the negative sample with positive prediction, FN is the positive sample with negative prediction, and p(r) is the precise-recall curve.

C. Setting of training parameters

TABLE III. TRAINING PARAMETERS SETTING TABLE

Parameter	**Value**
learning rate	0.0005
momentum	0.9
weight_decay	0.0005
batch size	16
epoch	50
step_size	5

D. Experimental results and analysis

In the experiments, transfer learning strategy is used to reduce the training difficulty and improve the detection results. The improved SSD algorithm in this paper is trained on the ResNet50 pre-training model, and its loss function remains the same as that of the original SSD-VGG16 algorithm. After the training, the improved algorithm in this paper produces the loss curve shown in Figure 6. From this figure, it can be seen that the algorithm in this paper starts to converge when the iteration reaches the 25th epoch. The final training loss value is 2.72.

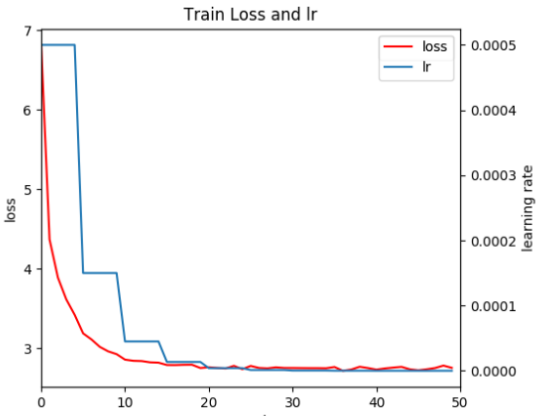

Figure 6. Loss curve of improved SSD algorithm

The distribution of mAP values is shown in Figure 7. It can be seen that the mAP value reaches a maximum of 72.67%, which is an increase of 2.1% over the original SSD-VGG16 network.

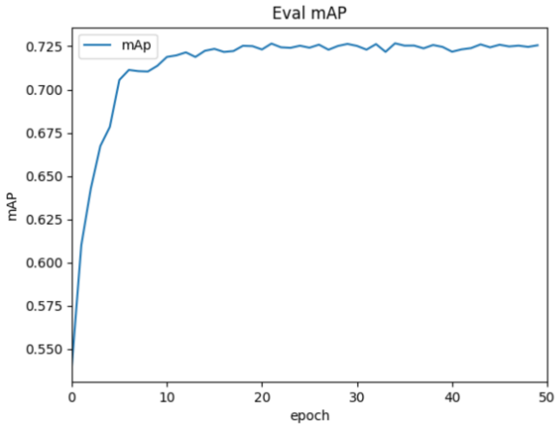

Figure 7. mAP graph of improved SSD algorithm

In this paper, the SSD-VGG16 algorithm, the SSD-ResNet50 algorithm and the improved algorithm of this paper are respectively trained and validated on the PASCAL VOC2012 dataset. And we calculates the mAP of each algorithm to detect the target. From Table Ⅳ, we can see that the mAP of the improved algorithm in this paper improves from 70.6% to 72.7% compared with the original SSD-VGG16 algorithm, thus the detection effect of this paper's algorithm is better than the original SSD algorithm.

TABLE IV. COMPARISON OF DETECTION RESULTS OF DIFFERENT ALGORITHMS ON THE PASCAL VOC2012 DATASET

Algorithm	mAP
SSD-VGG16	70.6%
SSD-ResNet50	72.1%
The improved algorithm	72.7%

V. CONCLUSION

In order to improve the detection accuracy of SSD algorithm for object detection, this paper improves the network structure of SSD object detection algorithm, and introduces multi-layer feature fusion module and channel attention mechanism. Multi-layer feature fusion at different scales based on the ResNet50 feature network increases the semantic information of the original feature map. At the same time, the SE attention mechanism is added to enhance the focus on feature channels for each layer of the network. The comparison experiments on the PASCAL VOC2012 dataset show that the improved SSD algorithm in this paper has a 2.1% higher mean average precision and better detection capability compared to other commonly used algorithms.

REFERENCES

[1] Li Weiqiang, Wang Dong, Ning Zhengtong, Lu Mingliang, Qin Pengfei. A review of fruit target detection algorithms under computer vision [J]. Computers and Modernization, 2022(06):87-95.

[2] Lv Lu, Cheng Hu, Zhu Hongtai, Dai Nianshu. A review of target detection research and application based on deep learning [J]. Electronics and Mounting, 2022, 22(01):72-80.DOI:10.16257/j.cnki.1681-1070.2022.0114.

[3] Deng Quan, Lin Xingxing. Improved SSD-based algorithm for marine life detection [J]. Computer Technology and Development, 2022, 32(04):51-56.

[4] Xie F, Zhu D.B.. A review of deep learning target detection methods [J]. Computer Systems Applications, 2022, 31(02):1-12. doi:10.15888/j.cnki.csa.008303.

[5] Peng Hongxing, Li Jing, Xu Huiming, Chen Hu, Xing Zheng, He Huijun, Xiong Juntao. Litchi detection based on multiple feature enhancement and feature fusion SSD [J]. Journal of Agricultural Engineering, 2022, 38(04):169-177.

[6] Jia Kexin, Ma Zhenghua,Zhu Rong,Li Yonggang. Attention mechanism to improve lightweight SSD model for sea surface small target detection [J]. Chinese Journal of Graphics, 2022, 27(04):1161-1175.

[7] Zheng Qiumei, Xu Linkang, Wang Fenghua, Lin Chao. Pyramid scene resolution network based on improved self-attention mechanism[J/OL]. Computer Engineering:1-9[2022-07-08]. DOI:10.19678/j.issn.1000-3428.0063652.

[8] Huang Yichuan, Li Lianghai, Ma Jijun, Cui Huimin. Research on aerial photography target detection algorithm based on improved SSD algorithm [J]. Telemetry and Remote Control, 2022, 43(03):79-85.

[9] Gao Na, Wu Qing, Zhang Man-ho. Multi-scale feature enhancement algorithm for SSD target detection [J]. Journal of Hebei University of Technology, 2022, 51(02):23-30.

[10] Guo Jianzhong, Yu Tengfei, Cui Yuding, Zhou Xinglin. Research on improved SSD-based vehicle small target detection algorithm [J]. Computer Technology and Development, 2022, 32(03):1-7.

[11] Yin, Meng-Yuan. Research on SSD target detection algorithm based on attention mechanism[C]. Anhui University of Technology, 2021.

[12] Sun Peng. Research on small target detection based on improved SSD model[C]. Nanjing University of Posts and Telecommunications, 2021.

[13] Yao Guanghua, Wu Xuncheng, Zhang Xuexiang, and Squire Jun. A small target detection method incorporating contextual information features [J]. Computer and Digital Engineering, 2022, 50(05):1018-1022.

[14] Yang Haojie, Wang Lu, Yang Shengwei. A feature fusion-based road target detection method [J]. Journal of Changsha University, 2022, 36(02):1-6.

[15] Li Kequan, Chen Yan, Liu Jia Chen, Mou Xiangwei. A review of deep learning-based target detection algorithms [J/OL]. Computer Engineering:1-17[2022-07-09]. DOI:10.19678/j.issn.1000-3428.0062725.

[16] Liu X. Research on improved SSD-based target detection algorithm [D]. Xiangtan University, 2021. DOI:10.27426/d.cnki.gxtdu.2021.002054.

[17] Wang Yanni, Yu Lixian. Attention and multi-scale effective fusion of SSD target detection algorithm [J]. Computer Science and Exploration, 2022, 16(02):438-447.

[18] He Jingyuan, Xie, Shenglong, Tian Yuan, Tian Qinqin. Multi-scale feature fusion for target detection algorithm [J]. Henan Science, 2021, 39(07):1045-1051.

[19] Wu T-C, Wang X-T, Cai Y-J, Jing Y-B, Chen C-Y. Lightweight SSD target detection method based on feature fusion [J]. Liquid Crystal and Display, 2021, 36(10):1437-1444.

[20] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector[C]// European Conference on Computer Vision. Springer, Cham, 2016: 21-37.