# SSD Target Detection Algorithm Based on Multi-Scale Fusion and Attention

Chenyang Jin
School of Software, Yunnan University, Kunming, China
3216380542@qq.com

Lei Li
School of Software, Yunnan University, Kunming, China
1453560627@qq.com

Mengting Li
School of Software, Yunnan University, Kunming, China
jkc265616@mail.ynu.edu.cn

Yijian Pei
School of Information Science & Engineering, Yunnan University, Kunming, China
yndxpyj@163.com

## ABSTRACT

Aiming at the problems of weak effective information in feature maps and high miss-detection rate of difficult targets when traditional SSD target detection algorithms perform target detection, we propose an improved SSD target detection algorithm. First, add a CBAM module after each feature layer of the SSD. CBAM is a hybrid module that combines spatial attention and channel attention. This module strengthens the network's ability to discriminate targets and backgrounds, improves the expression of effective feature weights, and suppresses interference from irrelevant information; then, adopt the idea of FPN to construct a feature fusion module, which effectively integrates feature layers of different scales, thereby improving the network's ability to detect difficult targets. Verifying the method proposed in this paper on the PASCAL VOC data set fully proves that the improved network performance has been greatly improved.

## CCS CONCEPTS

• **Computing methodologies**; • **Artificial intelligence**; • **Computer vision**;

## KEYWORDS

Target detection, SSD, CBAM, Multi-scale feature

**ACM Reference Format:**
Chenyang Jin, Lei Li, Mengting Li, and Yijian Pei. 2021. SSD Target Detection Algorithm Based on Multi-Scale Fusion and Attention. In *The 5th International Conference on Computer Science and Application Engineering (CSAE 2021), October 19–21, 2021, Sanya, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3487075.3487087

## 1 INTRODUCTION

Target detection is an important branch in the field of computer vision. Its main purpose is to analyze and process a given video or picture, point out the category of each target, and draw a bounding box near the target to mark the location of the target. Traditional target detection algorithms are based on manually extracted features, the accuracy of the algorithm is low and the generalization performance is weak. In recent years, deep learning technology has developed rapidly. The target detection algorithm based on deep learning uses convolutional neural networks to extract features. It has made breakthroughs in detection accuracy and is widely used in video surveillance, smart logistics, and medical mage analysis, unmanned driving and other fields [1]. The current mainstream target detection methods are mainly divided into two-stage detection methods and single-stage detection methods. Fast-RCNN [2], Faster-RCNN [3], etc. are all classic two-stage detection methods. You Only Look Once (YOLO) [4–6],Single Shot MultiBox Detector (SSD) [7], etc. are all typical single-stage detection methods. The two-stage detection method mainly divides the extraction of the candidate frame and the target detection into two steps. First, the candidate is screened, and then the candidate is classified and regressed. The single-stage target detection is to integrate it into a network, and directly output the detection result. In actual application scenarios, the speed of target detection is slow, and the detection effect is often affected by the deformation, occlusion, and environmental changes of the target object. Therefore, designing a target detection algorithm with good real-time performance, strong robustness and high detection accuracy has very important research significance.

Among them, the SSD algorithm draws on both the idea of YOLO grid and the anchor mechanism of Faster R-CNN, so that the SSD can make rapid predictions and obtain the position of the target accurately. However, because the SSD network cannot fully capture contextual information and multi-scale information, although high-level features have rich semantic information, their resolution is low, and their ability to perceive details is poor. The shallow feature layers cannot make full use of contextual semantic information, so there are more false detections and missed detections for difficult targets such as small targets.

Aiming at the detection shortcomings of the SSD algorithm, this paper uses VGG-16 [8] as the basic backbone network and introduces the CBAM attention mechanism [9] and feature fusion module [10] to reduce the missed detection rate and false detection rate of the network. The attention mechanism can improve the network's ability to learn key information, and feature fusion can enrich the feature layer information used for target detection in the SSD. The combination of the two can effectively improve the accuracy of target detection. In the second section, we introduced the


*CSAE 2021, October 19–21, 2021, Sanya, China*

ACM ISBN 978-1-4503-8985-3/21/10…$15.00
https://doi.org/10.1145/3487075.3487087

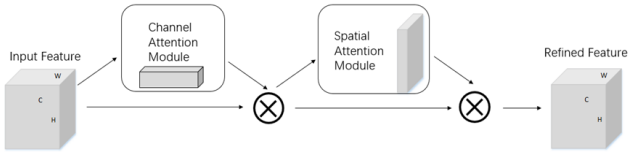


**Figure 1: CBAM Attention Module Structure.**

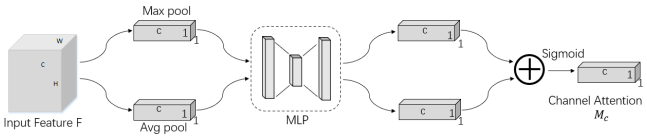


**Figure 2: Channel Attention Module Structure.**

CBAM module and the multi-scale feature fusion module in detail, and then added them to the SSD network. The third section introduces the experimental conditions, gives the experimental results, and analyzes the experimental results. Finally, we summarized the work of this article and future research directions.

## 2 METHOD

### 2.1 CBAM Attention Mechanism

The CBAM attention mechanism includes a channel attention module and a spatial attention module, which can strengthen the network's ability to discriminate targets and backgrounds, and effectively improve network performance [11]. The structure of the CBAM attention module is shown in Figure 1

### 2.2 Channel Attention Module

The channel attention module automatically obtains the importance of each feature channel through network learning, and finally assigns different weight coefficients to each channel, thereby strengthening important features and suppressing unimportant features. The structure of the channel attention module is shown in Figure 2

The working principle of the channel attention module: Assuming that the input feature map of the channel attention module is $F \in R^{H\times W\times C}$, input the average pooling layer and maximum pooling layer respectively to obtain two feature maps $F_{avg} \in R^{1\times 1\times C}$ and $F_{max} \in R^{1\times 1\times C}$, and then use the parameter-sharing MLP to upgrade and reduce the dimensions of the two feature maps. The MLP contains three layers, and the number of nodes in the hidden layer is C/r (r is the feature dimensionality reduction parameter), and the number of nodes in the output layer is C. The advantage of this processing is that it has more nonlinearity, can better fit the complex correlation between channels, and greatly reduces the amount of parameters and calculations. The two feature maps obtained are added element-wise, and finally the Sigmoid activation function is used to output $M_C(F) \in R^{1\times 1\times C}$, which is the weight parameter, and $M_C(F)$ perform element-wise multiplication with the input feature map F to obtain the feature layer $F' \in R^{H\times W\times C}$. The calculation method of the channel attention weight parameter

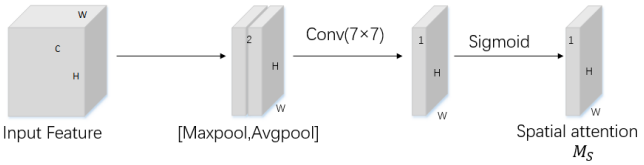


**Figure 3: Spatial Attention Module Structure.**

can be obtained by the following formula:

$$M_C(F) = \sigma\left(W_1\left(W_0\left(F_{avg}\right)\right) + W_1\left(W_0\left(F_{max}\right)\right)\right) \tag{1}$$

Where $\sigma$ represents the Sigmoid function, $W_0$ is the hidden layer weight, and $W_1$ is the weight coefficient of the MLP output layer.

### 2.3 Spatial Attention Module

Spatial attention aims to improve the feature expression of key areas. In essence, the spatial information in the original image is transformed into another space through the spatial conversion module and the key information is preserved, thereby enhancing the specific target area of interest while weakening irrelevant the background area [12]. The structure of the spatial attention module is shown in the figure below (Figure 3):

The working principle of the spatial attention module: Since the channel attention module and the spatial attention module are cascaded, the feature layer $F'$ output by the channel attention module is used as the input of the spatial attention module, and passes through an average pooling layer and a maximum pooling layer obtains the feature layers of one-dimensional channels $F'_{avg} \in R^{H\times W\times 1}$ and $F'_{max} \in R^{H\times W\times 1}$, and splice them into a feature map with 2 channels, and then use a 7×7 convolution kernel to reduce its dimensionality. At this time, the feature map is compressed to H×W×1, and then through the Sigmoid activation function, the weight coefficient of the spatial attention module is obtained as $M_S(F') \in R^{H\times W\times 1}$, which is multiplied by the input feature layer F′ to obtain the output feature layer $F''$ of the spatial attention module [13]. After passing through the entire CBAM module, the target information of the input feature layer can be emphasized. The weight parameter of the spatial attention module can be obtained by the following formula, $f^{7\times 7}$ represents a convolution kernel with a size of 7×7.

$$M_S\left(F'\right) = \sigma\left(f^{7\times 7}\left(\left[F'_{avg}; F'_{max}\right]\right)\right) \tag{2}$$

The whole process of the CBAM attention mechanism can be summarized as: the input feature layer F first passes through the channel attention module to obtain the weight coefficient $M_C(F)$, and multiplies it with the input feature layer F to obtain the F′ feature layer, and then uses F′ as the input of the spatial attention module, the weight coefficient $M_S(F')$ of the spatial attention module is obtained, and $M_S(F')$ is multiplied by F′ to obtain the output feature layer $F''$ of the entire CBAM module. The whole process can be deduced by the following formula:

$$F' = M_C(F) \otimes F \tag{3}$$

$$F'' = M_S\left(F'\right) \otimes F' \tag{4}$$

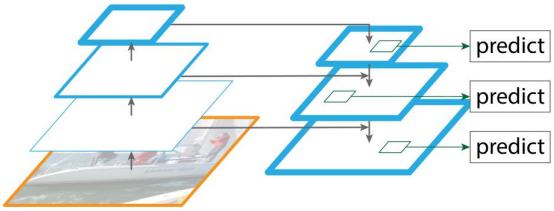


**Figure 4: FPN Feature Pyramid Structure.**

## 2.4 Add Feature Fusion Module

The design idea of SSD is to make predictions on different feature layers at the same time. Its disadvantage is that the obtained features are not robust, because many features are directly obtained from the shallower feature layer, and these are some weak features [14]. The shallow features of SSD are used for small target detection, but the receptive field of the shallow feature map is small, and the representation ability of semantic information is weak. In order to improve the ability of the SSD network to detect small targets, we use the FPN feature pyramid module to achieve feature fusion, and its structure is shown in Figure 4

The whole process of feature fusion is as follows: First, we perform deep convolution on the input image to generate feature layers of different scales, and then perform up-sampling operations on the high-level feature layers to make them have corresponding sizes. We accumulate the low-level feature layer and the high-level feature layer, so that each feature layer merges the information of multiple feature layers, and then outputs them in different feature layers, which effectively improves the performance of the network [15].

Feature fusion between different scales requires up-sampling of high-level feature layers. The upsampling method used in this article is the nearest neighbor interpolation algorithm in the linear difference method. The basic principle is to assign the value of a known pixel to the position of the nearest neighbor pixel. As shown in Figure 5, the pixel value of area A is determined by the point (i, j), and the pixel value of area B is determined by the point (i+1, j). By analogy, we can get all the pixel values $f(i+u, j+v)$ in the area.

## 2.5 The Improved SSD Network

In order to solve the problem of the SSD algorithm for the detection of difficult targets, we incorporated the attention mechanism and feature fusion module into the original network. Firstly, insert the CBAM attention mechanism module into the different feature layers extracted from the original SSD to enhance the ability of the feature map to express key information. Secondly, the FPN module is used to integrate the deep network and the shallow network to improve the semantic information representation ability of the shallow network. The improved network structure is shown in Figure 6, and the feature layers in the figure have been processed by the CBAM attention module.

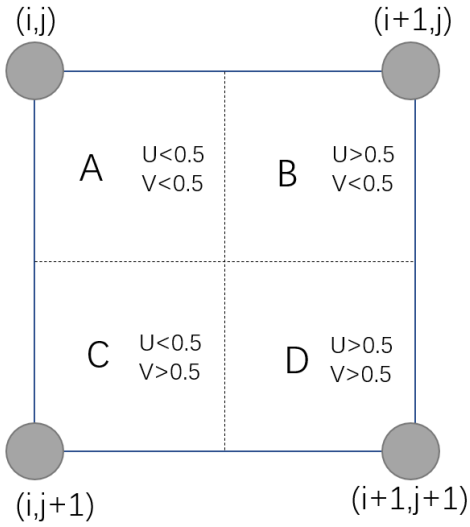


**Figure 5: Nearest Neighbor Interpolation Algorithm.**

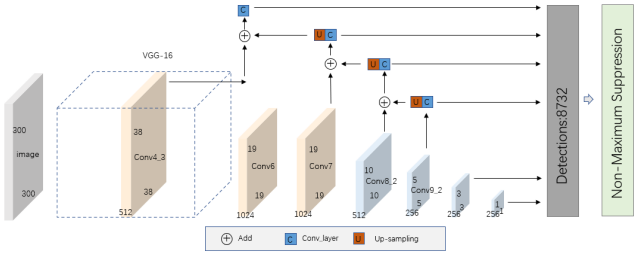


**Figure 6: Improved SSD Network Structure.**

# 3 MODEL TRAINING AND EXPERIMENTAL RESULTS ANALYSIS

## 3.1 Setup Before Experiment

In order to evaluate the improved network performance, we use the PASCAL VOC data set for verification. The data set contains a total of 20 object categories, such as birds, people, cars, sheep, cows, potted plants, sofas, chairs, and so on. In the training phase, the PASCAL VOC2007 train data set and the PASCAL VOC2012 train data set are used as the training set and the validation set, which contains a total of 16551 pictures. In the test phase, the PASCAL VOC2007 test data set is used as the network evaluation data set, which contains a total of 4952 pictures. The graphics card used in the experiment is RTX3090, the size of the graphics card is 24G, and Pytorch is used for neural network construction.

We use a series of conventional data enhancement methods, such as horizontal flip, color distortion, random cropping, etc. Through the data enhancement strategy, the training data set can be enriched during the training process, so as to achieve the purpose of improving the network performance. The improved algorithm still uses classification and regression to predict the target. The Softmax function is used for classification confidence prediction, and the

**Table 1: Experimental Results**

| Method | mAP/% | aero | bike | bird | boat | bottle | bus | car | cat | chair | cow |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Fast | 69.8 | 76.0 | 77.2 | 68.5 | 60.2 | 40.1 | 83.5 | 77.8 | 85.9 | 45.2 | 81.2 |
| Faster | 72.1 | 76.5 | 78.9 | 70.4 | 66.2 | 51.2 | 84.3 | 83.2 | 86.2 | 49.7 | 82.3 |
| SSD | 75.6 | 77.8 | 84.3 | 72.3 | 68.5 | 48.4 | 84.7 | 85.3 | 87.9 | 58.4 | 82.9 |
| DSSD | 79.1 | 79.3 | 83.2 | 80.6 | 68.8 | 63.7 | **87.5** | 86.8 | 84.3 | 62.1 | 84.5 |
| ION | 79.6 | 80.2 | **85.3** | 78.5 | **74.4** | 61.3 | 86.5 | **87.5** | 89.7 | 63.2 | **86.9** |
| ours | **80.8** | **80.7** | 84.6 | **81.5** | 72.2 | **65.3** | 87.2 | 85.8 | **90.9** | **64.2** | 85.6 |
| Method | mAP/% | table | dog | horse | mbike | person | plant | sheep | sofa | train | tv |
| Fast | 69.8 | 67.5 | 81.2 | 81.8 | 78.3 | 65.2 | 35.6 | 66.5 | 75.5 | 76.6 | 72.5 |
| Faster | 72.1 | 64.8 | 82.4 | 83.2 | 80.4 | 68.4 | 38.9 | 68.6 | 72.9 | 78.6 | 75.4 |
| SSD | 75.6 | 75.2 | 84.7 | 85.9 | 83.5 | 75.8 | 48.5 | 72.6 | 74.8 | 82.8 | 78.6 |
| DSSD | 79.1 | **81.3** | 86.8 | 85.8 | 87.6 | 79.6 | 53.2 | 76.8 | 78.4 | 86.5 | 84.8 |
| ION | 79.6 | 79.2 | 84.9 | 86.7 | 85.3 | 81.6 | 52.8 | 76.6 | **81.5** | 86.9 | 82.7 |
| ours | **80.8** | 80.5 | **88.6** | **89.2** | **88.4** | **83.3** | **54.2** | **78.5** | 79.6 | **88.4** | **86.2** |

regression function is used to output the target position when positioning. The target loss function is composed of confidence error and position error.

The Adam optimizer is used in the network training stage, and the two-step training method is adopted to prevent the weights obtained through the initialization of the pre-training model from being destroyed. First freeze backbone network trained to detect network, batch_size set to 32, the initial learning rate of 0.01, a total of 5000 iterations, and each iteration, the learning rate is reduced by 5%. Then release the frozen backbone network parameters and train the entire network from scratch, batch_size still set to 32, the initial learning rate of 0.01, a total of 5000 iterations, and each iteration, the learning rate is also reduced by 5%. A total of 10000 iterations throughout the training phase.

## 3.2 Experimental Results

We use AP and mAP as the performance evaluation index of the algorithm. And compared with target detection algorithms such as SSD300, Fast RCNN, Faster RCNN, DSSD [16], ION300 [17], etc. The detailed results of the detection accuracy of each category are shown in Table 1

Among the listed 6 methods, the highest single-category target AP value is displayed in bold, DSSD occupies 2 items, ION occupies 5 items, and the method proposed in this paper occupies 13 items. The mAP of the improved SSD is 5.2% higher than the original SSD, 1.7% higher than DSSD, and 1.2% higher than ION. According to the results, we can easily see that the AP of the method proposed in this paper is the maximum in most categories, indicating that the improved method proposed is effective.

In order to compare the difference between the algorithm in this paper and other mainstream target detection algorithms more comprehensively, the detection speed and detection accuracy of various algorithms on the PASCAL VOC2007test data set are further compared. The results are shown in Table 2. As can be seen from the table, our algorithm has improved accuracy and detection speed compared with Faster RCNN, SSD512, DSSD321, and RSSD300 [18]. Especially compared with other methods based on SSD algorithm improvement, our method has more advantages. Compared with the original SSD300, although the FPS has dropped by 10.4, the detection accuracy has increased by 5.2%.

## 4 CONCLUSIONS

We have improved the traditional SSD target detection algorithm. First, we added a CBAM hybrid attention mechanism between the feature layers. CBAM includes a channel attention module and a spatial attention module, which can effectively improve the weight expression of target features. Then a multi-scale feature fusion module is added, which can enrich the semantic information and detailed information of the feature layer. It can be concluded from the experimental results that our proposed method has achieved good results, effectively improved the accuracy of target detection.

**Table 2: Comparison of Detection Speed and Detection Accuracy**

| Method | Network | FPS | Anchor | Size | mAP/% |
|---|---|---|---|---|---|
| Faster | VGG-16 | 7.8 | 6000 | ~600×1000 | 72.1 |
| SSD300 | VGG-16 | 48.6 | 8732 | 300×300 | 75.6 |
| SSD512 | VGG-16 | 22.4 | 24564 | 512×512 | 78.2 |
| RSSD300 | VGG-16 | 34.8 | 8732 | 300×300 | 78.5 |
| DSSD321 | ResNet-101 | 13.5 | 17080 | 321×321 | 79.1 |
| ours | VGG-16 | 38.2 | 8732 | 300×300 | 80.8 |

However, due to the addition of new network modules and the increase of model parameters, the improved network detection speed has decreased. In the future work, we will continue to optimize the network to make the detection accuracy higher and the real-time performance better.

## REFERENCES


[1] ZHENG Y P, LI G Y, LI Y (2019). Survey of application of deep learning in image recognition[J]. Computer Engineering andApplications, 55(12): 20-36. (in Chinese with English abstract)

[2] GIRSHICK R B (2015). Fast R-CNN[J]. 2015 IEEE International Conference on Computer Vision (ICCV), 1440-1448.

[3] REN S, HE K, GIRSHICK R B, *et al.* (2015). Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39: 1137-1149.

[4] REDMON J, DIVVALA S, GIRSHICK R, *et al.* (2016). You only look once: unified, real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Las Vegas, NV, USA, June 27-30, 2016. Piscataway: IEEE, 779-788.

[5] REDMONJ, FARHADI A (2017). YOLO 9000: better, faster, stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, Ju-ly 21-26,2017. Piscataway: IEEE, 6517-6525.

[6] REDMON J, FARHADI A (2018). Yolov3: an incremental improvem-ent[EB/OL]. (2018-04-08) [2021-05-13]. https://arxiv.org/abs/1804.02767

[7] LIU W, ANGUELOV D, ERHAN D, *et al.* (2016). SSD: single shot multibox detector[C]//Proceedings of the European Conference on Computer Vision, Berlin,Germany, October 11-14, 2016. Berlin, Heidelberg: Springer, 21-37.

[8] SIMONYAN K, ZISSERMAN A (2014). Very deep convolutional networks for large scale image recognition[J]. Computer Science, 1409-1556.

[9] WOO S,PARK J,LEE J Y,et al. (2018). CBAM: convolutional block attention module [M//Computer Vision-ECCV 2018.Cham: Springer International Publishing, 3-19.

[10] LIN T-Y, DOLLÁR P, GIRSHICK R, *et al.* (2017). Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, July 21-26, 2017. Piscataway: IEEE, 936-944.

[11] CHU X,YANG W,OUYANG W, *et al.* (2017). Multi-context atten-tion for human pose estimation[C]//2017 IEEE conference on computer vision and pattern recognition.

[12] GOVARDHAN P, PATI U C (2014). NIR image based pedestrian detection in night vision with cascade classification and validati-on[C]//Proceedings of the IEEE International Conference on Advanced Communications, Control and Computing Tech-nologies, Ramanathapuram, India, May 8-10, 2014. Piscataway: IEEE, 1435-1438.

[13] JADERBERG M, SIMONYAN K, ZISSERMAN A, *et al.* (2016). Spatial transformer networks[EB/OL].(2016-02-04)[2021-05-13]. https://arxiv.org/abs/1506.02025v3

[14] DENG J,DONG W, SOCHER R, *et al.* (2009). Image Net: a large-scale hierarchical image database[C]//2009 IEEE Confer-ence on Computer Vision and Pattern Recognition. June 20-25,2009, Miami, FL, USA. IEEE, 248-255.

[15] JI Z, KONG Q, WANG H, *et al.* (2019). Small and Dense Commodity Object Detection with Multi-Scale Receptive Field Attention[C]//Proceedings of the 27th ACM International Conference on Multimedia. October, 2019. 1349-1357.

[16] FU C-Y, LIU W, RANGA A, *et al.* (2017). DSSD : deconvoluti-onal single shot detector[EB/OL].(2017-01-23)[2021-05-13]. https://arxiv.org/abs/1701.06659.

[17] BELL S, ZITNICK C L, BALA K, *et al.* (2016). Inside-Outside Net: detecting objects in context with skip pooling and recurrent neural networks[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, June 27-30, 2016. Piscataway: IEEE, 2016: 2874-2883.

[18] JEONG J, PARK H, KWAK N (2017). Enhancement of SSD by concatenating feature maps for object detection [EB/OL]. (2017-05-26)[2021-05-13]. https://arxiv.org/abs/1705.09587v1.