applied sciences

MDPI

Article

ADSSD: Improved Single-Shot Detector with Attention Mechanism and Dilated Convolution

Jian Ni, Rui Wang * and Jing Tang

School of Information and Electrical Engineering, Hebei University of Engineering, Handan 056038, China
* Correspondence: wangrui19981101@163.com

Abstract: The detection of small objects is easily affected by background information, and a lack of context information makes detection difficult. Therefore, small object detection has become an extremely challenging task. Based on the above problems, we proposed a Single-Shot MultiBox Detector with an attention mechanism and dilated convolution (ADSSD). In the attention module, we strengthened the connection between information in space and channels while using cross-layer connections to accelerate training. In the multi-branch dilated convolution module, we combined three expansion convolutions with different dilated ratios to obtain multi-scale context information and used hierarchical feature fusion to reduce the gridding effect. The results show that on PASCAL VOC2007 and VOC2012 datasets, our 300 × 300 input ADSSD model reaches 78.4% mAP and 76.1% mAP. The results outperform those of SSD and other advanced detectors; the effect of some small object detection is significantly improved. Moreover, the performance of the ADSSD in object detection affected by factors such as dense occlusion is better than that of the traditional SSD.

Keywords: SSD; small object detection; attention mechanism; multi-branch dilated convolution

Citation: Ni, J.; Wang, R.; Tang, J. ADSSD: Improved Single-Shot Detector with Attention Mechanism and Dilated Convolution. *Appl. Sci.* **2023**, *13*, 4038. https://doi.org/10.3390/app13064038

Academic Editor: Jan Egger

Received: 3 March 2023
Revised: 17 March 2023
Accepted: 21 March 2023
Published: 22 March 2023

1. Introduction

In the field of computer vision, many tasks are necessary to determine the class and position of objects; that is why object detection has a high status in this field. The main task is to determine the proposal region, extract features from the region, and then use the extracted features for accurate localization and classification. It is the basis for other image processing tasks. Nowadays, it has been applied in various fields, for instance, UAV reconnaissance in the military field [1], pest detection in the agriculture field [2], brain tumor detection in the medical field [3], and so on. But these technologies are still defective and have room for improvement. The improvement of object-detection technology will also lead to the advancement of these technologies. However, given the emphasis and difficulty of object detection, poor performance of small object detection will also reduce the overall detection effect. Therefore, research on small object detection is of great significance.

As deep learning makes more and more progress in various studies, Convolutional Neural Networks (CNNs) are becoming widely used in object-detection algorithms [1,2]. Unlike traditional detection algorithms, CNNs have an excellent performance in accuracy and speed. In deep learning-based object-detection algorithms, commonly used anchor frame object-detection methods can be divided into two categories. One is two-stage object detection: First, distinguish between background and prospect to get region of interest (RoI); second, use CNNs to extract features in the RoI and regression prediction again. Region-Based Convolutional Neural Networks (R-CNNs) are commonly used [4]. Due to the low detection efficiency of R-CNNs, Redmon J et al. proposed another detection algorithm: one-stage objection detection, called "You Only Look Once" (YOLO) [5]. Although the speed of detection has improved, the performance of small object detection is still unsatisfactory. To improve the performance of small object detection, Liu et al. proposed a Single-Shot MultiBox Detector (SSD) [6].

Unlike YOLO, the SSD performs better because it outputs and predicts at different scales. The SSD takes VGG-16 as the backbone network and uses pyramid structures to obtain multi-scale feature information for location and classification. The deeper network layers are used to detect larger objects, and smaller objects are detected by shallow network layers, so that the SSD will perform better in the detection of small objects. However, background information still affects the extraction of small object feature information. We use the attention mechanism to strengthen the connection between the channel and the spatial feature information; reduce the weight of the background information; achieve the purpose of suppressing background information and focusing on key information; and then improve the extraction ability of feature information. At the same time, since only shallow features are downsampled, resulting in a lack of context information to assist small object detection, we use convolution modules with different expansion ratios to capture local and global context information. In addition, we use hierarchical feature fusion to solve the problem of local information loss and information obtained from a distance being not relevant due to dilated convolution.

To address these issues, this paper proposed an algorithm ADSSD based on an improved SSD.

(1) To suppress background information and improve the ability to extract small object feature information, we propose a module that is improved based on a CBAM, called a Residual Convolutional Block Attention Module (RCBAM).
(2) We constructed a multi-branch dilated convolution to extract contextual information, introduced hierarchical feature fusion (HFF) for de-gridding, and used cross-layer connections to improve learning efficiency and prevent degradation.
(3) We have improved the detection speed while improving the detection accuracy, which is not available in other algorithms.

In Section 2, we cover the work involved in the module. In Section 3, we introduce the network structure of the ADSSD and the structure of the module used in the improvement. In Section 4, YOLO, the SSD, and other detectors are compared with the ADSSD, and the experimental results are analyzed. In Section 5, the conclusion and prospect of the future are provided.

2. Related Work

2.1. Deep Learning-Based Object-Detection Algorithm

The introduction of deep learning in object detection can make object detection have a superior performance in detection efficiency and detection accuracy. Therefore, many deep-learning methods [7,8] have been applied to the field of object detection, and excellent results have been achieved. Ross Girshick et al. first attempted to combine deep learning with object detection and proposed the first two-stage detection R-CNN [4]. Later, to improve the speed of detection, Fast R-CNN [9] and Faster R-CNN [10] were proposed one after another. However, for work that requires real-time performance, its detection speed is not enough. Hence, "You Only Look Once" (YOLO) [5] was proposed by Redmon J et al. Despite the improved accuracy and speed, it is still unsatisfactory in small object detection. Therefore, Liu et al. proposed SSD [6]. It can output and predict feature layers at different scales, which is more suitable for small object detection.

2.2. Visual Attention Mechanism

Although the existing algorithms perform well in object detection, they are still insufficient for small object detection. They are easily affected by background information, resulting in false detections, and missed detections. Jecong-Seon Lim et al. [11] proved that the introduction of an attention mechanism can effectively suppress unnecessary shallow feature information, thereby improving the ability of small object detection. Therefore, attention mechanisms are used in various fields, such as remote-sensing image detection [12], object segmentation [13], and so on.

In 2018, CVPR Jie Hu et al. proposed a channel attention mechanism Squeeze-and-Excitation Network (SENet) [14]. Xiaocong et al. [12] proposed a new module called DAM based on SENet, which is applied in the field of remote sensing. However, they all focused on the connection information between channels to suppress unwanted features and reduce noise, but ignored the connection of information in space. Moreover, Max Jaderberg proposed a Spatial Transformer Network (STN) [15]; it learns the deformation of the input to compete the preprocessing operation suitable for the task. But it only pays attention to the spatial information connections and ignores the connection information between channels. Jing et al. [16] used the improved MS-CAB attention module to fuse feature information on spaces and channels. But the parallel approach could not better establish the feature connection between space and channel. Sanghyun Woo et al. proposed a mixed attention mechanism CBAM [17]. This attention mechanism is a kind of mixed attention mechanism module, which can achieve better results than the previous attention mechanisms. No one is currently trying to integrate a CBAM into an SSD.

2.3. Dilated Convolution

A small receptive field and low resolution are also important factors affecting the accuracy of small object detection. Before the CVPR in 2017, Fisher Yu et al. proposed a new way of convolution called dilated convolution [18]. This approach can expand the receptive field while maintaining resolution without degradation or increase in computation. Jiang et al. [19] enhanced the utilization of shallow features by increasing the receptive field by expanding convolution but did not increase the amount of computation while expanding the receptive field. Li Y et al. proposed a trident network [20], connecting the size of the receptive field and the detection of objects of different scales. It was confirmed that increasing receptive fields is helpful in developing the task of categorizing. Moreover, it can enhance the characteristics of small objects to enrich semantic information, thereby reducing the interference of redundant information and improving the ability to recognize small objects. The multi-branch dilated convolution based on a residual network used by Qunjie et al. [21] proved that it is helpful for improving the detection effect of small objects, but it does not consider the gridding effect of dilated convolution. In response to this problem, we chose to carry out layered feature fusion.

3. Method

In this section, we will introduce the structure of the SSD and the details of the ADSSD.

3.1. Single-Shot MultiBox Detector (SSD)

To maintain the speed of detection and have higher accuracy in small object detection, we chose to use an SSD as the basis of the network. As shown in Figure 1, we can see that the SSD used VGG-16 as the backbone network and inherited YOLO's idea of regression, and a similar prior box was proposed based on the anchor frame mechanism of Faster R-CNN [10]. To have higher accuracy on both small objects and large objects, the SSD selects six feature maps with different scales as output layers, and default boxes that distribute different scales and proportions are on these feature maps. Then the SSD generates position information and category probabilities for all anchor boxes. Finally, through non-maximum suppression (NMS), the SSD excludes bounding boxes that do not meet the criteria.

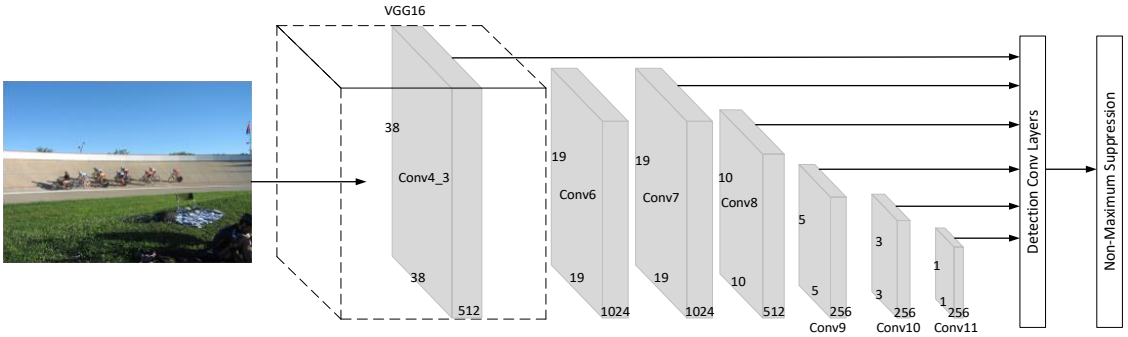

Figure 1. Network of Single-Shot MultiBox Detector (SSD).

Smaller objects are detected using shallow layers with few convolutions. However, the interference of background information and the lack of context information led to low small objects detection. To solve these issues, we proposed an improved SSD called an ADSSD.

Next, we will describe the ADSSD in detail.

3.2. ADSSD

In Figure 2, the ADSSD is proposed based on the SSD. We used the Residual Convolutional Block Attention Module (RCBAM) to reduce the influence of background factors. Then we added the multi-branch dilated convolution module (MDCM) to obtain information on multi-scale receptive fields and get more contextual information. In the following section, we will introduce the above modules in detail. These modules can be easily added to the original inspection network.

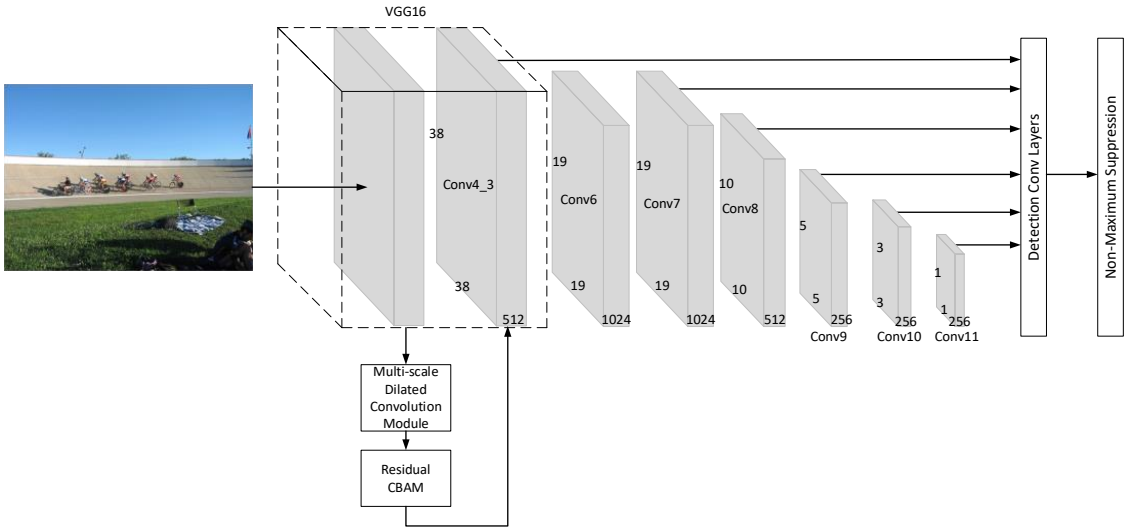

Figure 2. The single-shot object detection with attention mechanism and dilated convolution (ADSSD).

3.2.1. Residual Convolutional Block Attention Module

The attention mechanism can be widely understood as focusing only on local information to suppress redundant information on detection. There are currently three categories of attention mechanisms. They are spatial, channel, and mixed attention mechanisms.

The CBAM combines the advantages of spatial and channel attention mechanisms. In Figure 3, we can see the general structure of the CBAM. It serializes spatial attention and channel attention while emphasizing meaningful features in both spatial and channel dimensions.

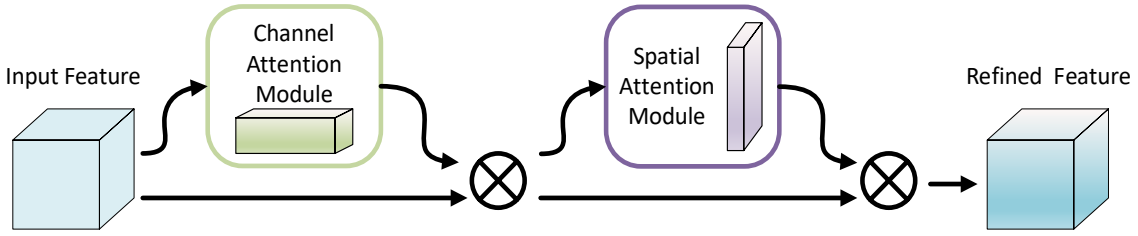

Figure 3. An overview of Convolutional Block Attention Module (CBAM). The module has two sequential sub-modules: channel and spatial attention modules.

Taking $\boldsymbol{F} \in R^{C\times H\times W}$ as input, CBAM yields in order $\boldsymbol{M_C} \in R^{C\times 1\times 1}$ and $\boldsymbol{M_s} \in R^{1\times H\times W}$ as illustrated in Figure 3. The entire attention process can be described as:

$$F' = M_C(F) \otimes F \tag{1}$$

$$F'' = M_s(F') \otimes F' \tag{2}$$

where $\otimes$ denotes element-wise multiplication.

In Figure 4, the channel attention module establishes connections between channels by squeezing the spatial dimension of the feature map.

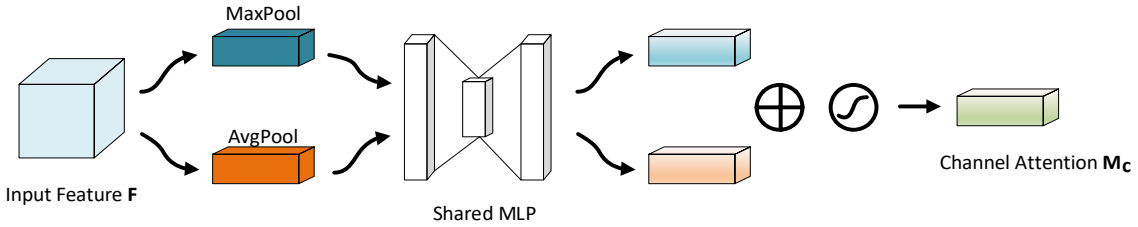

Figure 4. The channel attention module of CBAM. Channel attention map $\boldsymbol{M_C}$ is generated by feeding the pooled feature maps into a shared network.

In the following formula, F^c_{avg} and F^c_{max} represent the feature maps after average-pooling and max-pooling. The channel attention is computed as:

$$\begin{aligned} M_C(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma\left(W_1\left(W_0\left(F^c_{avg}\right)\right) + W_1(W_0(F^c_{max}))\right) \end{aligned} \tag{3}$$

where σ means the sigmoid function, $W_0 \in R^{C/r\times C}$, and $W_1 \in R^{C\times C/r}$.

In addition, as shown in Figure 5, we obtain spatial attention by applying pooling operations along the channel axis.

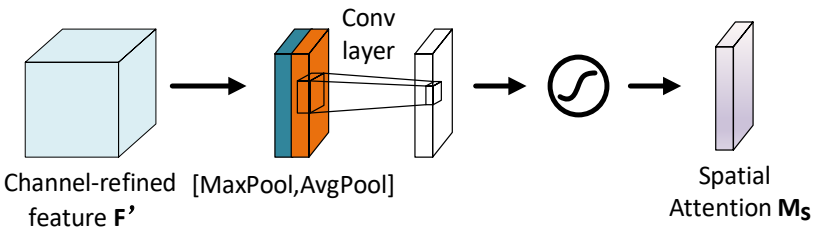

Figure 5. The spatial attention module of CBAM. Spatial attention map $\boldsymbol{M_s}$ is generated by pooling operations on the feature map.

The spatial attention module uses $F^s_{avg} \in R^{1\times H\times W}$ and $F^s_{max} \in R^{1\times H\times W}$ to represent the feature maps after pooling operations along the channel axis. The spatial attention is computed as:

$$\begin{aligned} M_s(F) &= \sigma\left(f^{7\times 7}([AvgPool(F); MaxPool(F)])\right) \\ &= \sigma\left(f^{7\times 7}\left(\left[F^s_{avg}; F^s_{\max}\right]\right)\right) \end{aligned} \tag{4}$$

where $f^{7\times 7}$ represents a convolution operation with the filter size of 7×7.

By reading many papers, we found that no one has yet introduced a CBAM on an SSD. Therefore, to improve the performance, we introduced the improved CBAM. Inspired by Kaiming He et al. [20], we chose to build a residual CBAM. This module makes a skip connection between the input and output layer. This improvement can effectively alleviate the problem of network degradation and speed up training.

Unlike others, we inserted the module into the backbone network. This method can improve detection accuracy without slowing down the detection speed. The specific structure of the residual CBAM is shown in Figure 6.

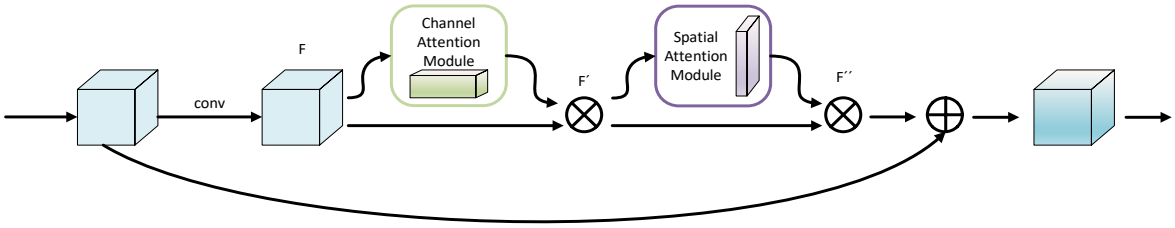

Figure 6. Residual CBAM.

In the next section, we will introduce another module we built called a multi-branch dilated convolution module (MDCM).

3.2.2. Multi-Branch Dilated Convolution

In view of the problem of the detection performance regarding small objects being degraded due to the low resolution of the receptive field of deep networks, dilated convolution can be used to solve it. Inspired by trident networks [22] and the module that Qunjie et al. [21] proposed, we proposed a multi-branch dilated convolution.

As shown in Figure 7, we introduced the MDCM to aggregate context information. We first used the 1×1 convolution to reduce the amount of computation. Then we used 3×3 convolution with different dilation rates to obtain context information from the feature map.

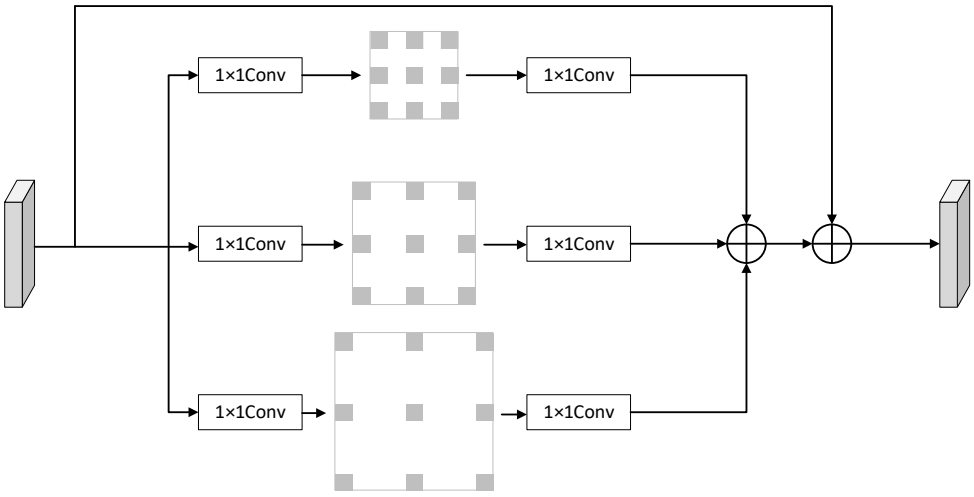

Figure 7. Multi-branch dilated convolution module (MDCM).

Faced with a gridding problem caused by the dilated convolution, we were inspired by Sachin Mehta et al. [23] to introduce hierarchical feature fusion (HFF). We can see the improved MDCM in Figure 8. This improved method does not increase the complexity of the module and leads to slower detection speed. Finally, we used skip connections to speed up learning. The dilation rate in our model is defined as {3,5,7}. To ensure detection accuracy and speed, we embedded the module into the backbone network.

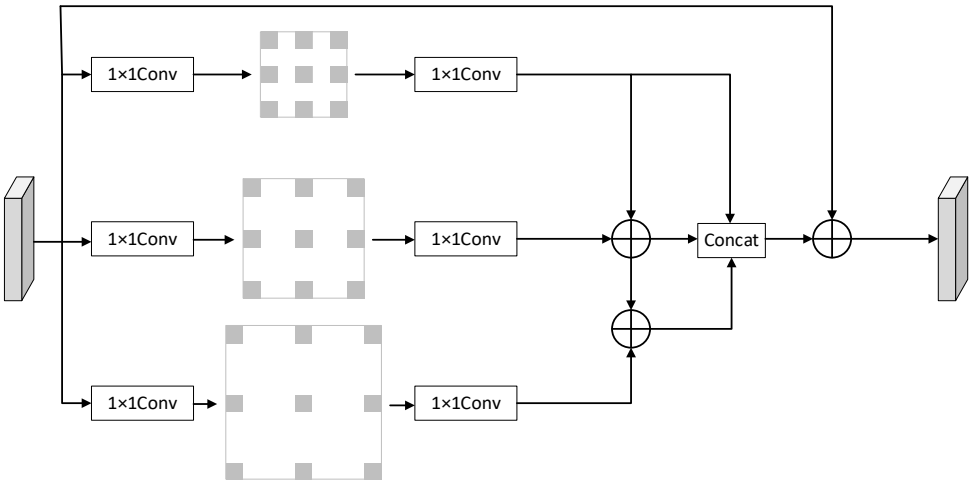

Figure 8. Improved MDCM.

In this study, we used dilated convolution with a convolution kernel size of 3×3. Our formula for calculating the dilated convolution kernel size is as follows:

$$k' = k + (k-1)(r-1) \tag{5}$$

where k is mean kernel size, and r is the dilated rate.

We use R_n to represent the receptive field of each pixel in the nth layer; k' is kernel size of dilated convolution; and s_i is the stride. The R_n is calculated as:

$$R_n = R_{n-1} + \left(k' - 1\right) \times \prod_{i=1}^{n-1} s_i \tag{6}$$

Table 1 shows the receptive field size of the dilated convolution calculated by the above formula.

Table 1. Receptive field size with different dilated rates.

Layers (n)	1	2	3
Kernel (n)	3×3	3×3	3×3
Stride (s_i)	1	1	1
Dilated rate (r)	3	5	7
Padding	3	5	7
Kernel (k')	7×7	11×11	15×15
RF (R_n)	9×9	19×19	23×23

4. Experiments and Dataset

In this section, we first introduce the datasets used in the experiment and evaluation criteria for the data. Then we present the parameters used in the experiment.

4.1. Dataset

We use PASCAL VOC [24] as the training and testing dataset. There are two versions of the PASCAL VOC dataset, including VOC 2007 and VOC 2012. Each dataset has trainval

set, val set, and test set. The dataset can be downloaded at Pascal VOC Dataset Mirror (pjreddie.com, accessed on 1 March 2022). In the experiment, we used 300 × 300 images as input images. In addition, we used VOC 07 + 12 trainval as our training dataset and VOC 07 as the test dataset. The ratio of our training and testing sets was about 8:2. At the same time, we also uploaded the result file generated using the VOC2012 test dataset to the competition website (PASCAL VOC Challenge performance evaluation server (ox.ac.uk, accessed on 28 February 2023)) for evaluation.

The dataset has 20 categories, including airplane, bicycle, and so on. We can see the specific information in Table 2.

Table 2. The categories in the PASCAL VOC 2007 and the number of images. The PASCAL VOC 2012 has the same categories, but different numbers of images.

Category	Trainval 12	Trainval 07	Test 07
Airplane	670	238	204
Bicycle	552	243	239
Bird	765	330	282
Boat	508	181	172
Bottle	706	244	212
Bus	421	186	174
Car	1161	713	721
Cat	1080	337	322
Chair	1119	445	417
Cow	303	141	127
Dining table	538	200	190
Dog	1286	421	418
Horse	482	287	274
Motorbike	526	245	222
Person	4087	2008	2007
Potted plant	527	245	224
Sheep	325	96	97
Sofa	507	229	223
Train	544	261	259
Tv monitor	575	256	259
Total	11,540	5011	4952

4.2. Evaluation Criteria

We used mean average precision (mAP) to evaluate the average accuracy of all categories and frames per second (FPS) to evaluate detection speed. The value of the AP is the area below the PR curve, where it is plotted based on the relationship between precision rate and recall rate. The AP is expressed as:

$$AP = \int_0^1 P(r)\mathrm{d}r \tag{7}$$

where $P(r)$ means the PR curve.

The precision and recall are required for drawing the PR curve. They were formulated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

where TP, FP, and FN refer to the number of correctly predicted positive, incorrectly predicted positive, and incorrectly predicted negative samples.

We calculate it in the same way by dividing the number of images by the total detection time.

Based on the PASCAL VOC [24] dataset settings, we used AP, mAP, and FPS for evaluation.

4.3. Parameters Setting

The software environment is: Windows10, Pytorch1.7.1, Python3.8.3, and CUDA11.1. The batchsize of the model is 32, the initial learning rate is 0.001, and the weight decay is 0.0005. This experiment environment configuration is shown in Table 3. We used about two days for training, a total of 120,000 rounds of training. We chose to have a learning rate decline between 80,000 and 100,000.

Table 3. Experiment environment configuration.

Name	Model	Number
CPU	Intel (R) Xeon (R) Gold 6130	1
Graphics	GeForce RTX 2080Ti 11G	1

To avoid the problem of overfitting due to insufficient data volume, this algorithm selects the training data of VOC 07 + 12 as the training set and tests on the test dataset of VOC 2007 and VOC 2012.

4.4. Results on VOC 2007

Table 4 shows the detection results compared with other detectors. All of these are trained on a dataset including VOC 2007 trainval and VOC 2012 trainval and evaluated on VOC 2007 test set. The ADSSD used the VGG-16 as the backbone network; the size of input image was 300 × 300. The detection accuracy reached 78.4%, and the detection speed was 54.1 FPS.

Table 4. Performance in different object-detection algorithms.

Method	Backbone	Input Size	GPU	FPS	mAP(%)
R-CNN [4]	AlexNet	1000 × 1000	-	0.07	50.2
SPP-NET [25]	AlexNet	224 × 224	-	0.5	63.1
Fast R-CNN [9]	VGG16	1000 × 600	Titan X	0.5	70.0
Faster R-CNN [10]	VGG16	1000 × 600	Titan X	7.0	73.1
HyperNet [26]	VGG16	1000 × 600	Titan X	0.9	76.3
OHEM [27]	VGG16	1000 × 600	Titan X	7.0	74.6
ION [28]	VGG16	1000 × 600	Titan X	1.3	76.5
YOLO [5]	GoogleNet	448 × 448	Titan X	45.0	69.0
SSD [6]	VGG16	300 × 300	2080Ti	46	77.1
DOSD [29]	DenseNet [30]	300 × 300	Titan X	17.4	77.7
ADSSD (ours)	VGG16	300 × 300	2080Ti	54.1	78.4

In the table above, we can see that ADSSD is superior to two-stage detection in both detection speed and accuracy, such as R-CNN, SPP-NET, etc. At the same time, the ADSSD also outperforms other one-stage detection models.

To compare the performance of models on each cluster, we also select a partial one-stage detection model for comparison in Table 5.

In this table, we can see that the performance of the ADSSD is better than that of the two-stage detectors, including Fast R-CNN, Faster R-CNN, and so on. Compared with Faster R-CNN, SSD, YOLO, and YOLOv3, the improved algorithm proposed in this paper improves the overall mAP by RCBAM and MDCM.

Table 5. The test detection result on PASCAL VOC2007.

Method	mAP	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow
Fast [9]	70.0	77.0	78.1	69.3	59.4	38.3	81.6	78.6	86.7	42.8	78.8
Faster [10]	73.1	76.5	79.0	70.9	65.5	52.1	83.1	84.7	86.4	52.0	81.9
YOLO [5]	69.0	75.0	77.8	60.7	57.7	54.4	77.5	84.4	73.8	53.2	66.0
YOLOv3 [31]	74.8	80.1	85.0	72.5	65.3	**64.8**	86.0	86.2	85.9	55.6	72.1
SSD300 [6]	77.1	80.7	83.1	76.4	**71.1**	50.7	84.2	85.8	86.5	**61.7**	81.2
Ours	**78.4**	**83.1**	**85.2**	**77.4**	69.4	53.7	**86.1**	**86.7**	**88.6**	61.3	**83.2**
Method	**mAP**	**table**	**dog**	**horse**	**mbike**	**person**	**plant**	**sheep**	**sofa**	**train**	**tv**
Fast	70.0	68.9	84.7	82.0	76.6	69.9	31.8	70.1	74.8	80.4	70.4
Faster	73.1	65.7	84.8	84.6	77.5	76.7	38.3	73.6	73.9	83.0	72.6
YOLO	69.0	68.3	70.4	79.4	77.2	79.8	44.3	63.7	69.2	77.6	68.6
YOLOv3	74.8	69.6	83.3	84.1	82.3	**81.4**	44.2	70.8	70.0	83.3	73.6
SSD300	77.1	78.0	85.0	86.0	84.3	78.3	49.5	76.7	79.6	86.9	76.5
Ours	**78.4**	**78.7**	**85.0**	**87.0**	**85.8**	79.7	**52.7**	**77.7**	**80.9**	**88.0**	**78.4**

4.5. Results on VOC2012

To verify that the ADSSD had improved on other datasets, we chose VOC2012 to evaluate the trained model. We put the results in Table 6. We can see an overall 0.5% improvement in VOC2012 and the best performance in several categories.

Table 6. The test detection result on PASCAL VOC2012.

Method	mAP	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow
Fast [9]	68.4	82.3	78.4	70.8	52.3	38.7	77.8	71.6	89.3	44.2	73.0
Faster [10]	70.3	84.9	79.8	74.3	52.3	49.8	77.5	75.9	88.5	45.6	77.1
YOLO [5]	57.9	77.0	67.2	57.7	38.3	22.7	68.3	55.9	81.4	36.2	60.8
YOLOv2 [32]	73.4	86.3	82.0	**74.8**	59.2	**51.8**	79.8	76.5	90.6	52.1	78.2
SSD300 [6]	75.6	88.2	82.9	74.0	**61.1**	47.3	**82.9**	78.9	91.6	57.8	80.2
Ours	**76.1**	**88.3**	**83.5**	74.2	60.9	48.0	82.7	**79.0**	**91.7**	**58.3**	**80.3**
Method	**mAP**	**table**	**dog**	**horse**	**mbike**	**person**	**plant**	**sheep**	**sofa**	**train**	**tv**
Fast	68.4	55.0	87.5	80.5	80.8	72.0	35.1	68.3	65.7	80.4	64.2
Faster	70.3	55.3	86.9	81.7	80.9	79.6	40.1	72.6	60.9	81.2	61.5
YOLO	57.9	48.5	77.2	72.3	71.3	63.5	28.9	52.2	54.8	73.9	50.8
YOLOv2	73.4	58.5	89.3	82.5	83.4	81.3	49.1	77.2	62.4	83.8	68.7
SSD300	75.6	63.9	89.3	85.5	85.9	82.2	49.7	78.8	73.1	86.6	71.6
Ours	**76.1**	**65.9**	**89.5**	**85.5**	**86.5**	**82.6**	**51.0**	**79.1**	**73.7**	**87.4**	**73.4**

4.6. Ablation Studies

We conducted ablation experiments to compare the effects of adding different modules. As shown in Table 7, we used the ADSSD compared with SSD, SSD + RCBAM, and SSD + RCBAM + MDCM. They were trained on the same platform.

Table 7. Improved SSD model compared with SSD, SSD + RCBAM, and SSD + RCBAM + MDCM.

RCBAM	MDCM	Improve-MDCM	mAP	FPS
			77.1	46
✓			77.6	58.5
✓	✓		78.1	57.3
✓		✓	78.4	54.1

Reducing the impact of background information is crucial. Small object detection is often affected by background factors, and the use of attention mechanisms can establish

connections on spaces and channels, reduce the influence of background factors, and improve performance. We used the improved CBAM to improve performance by 0.5%.

Obtaining contextual information is useful. We used dilated convolution to obtain multi-scale context information while reducing the impact of scale transformation on the model. The addition of a dilated convolution module resulted in a 0.5% performance improvement.

Reducing the gridding effect is crucial. When the dilated rate is not 1, there will be many gaps between the used data, and the discontinuities between the original data used will cause a lack of information. To reduce the gridding effect, we made hierarchical feature fusion of dilated convolution to strengthen the connection between the data. This method further improved the detection effect by 0.3%.

4.7. Visualization

To represent our method more intuitively, we selected some images for detection, and comparison results are illustrated in Figure 9. We showed only visualizations with classification scores above 0.5. We can see that the ADSSD was able to detect more objects.

(**a**) (**b**)

Figure 9. Test of images after improvements: (**a**) on the left is use of the SSD; (**b**) on the right is use of the ADSSD.

4.8. Intensive Object Detection

To compare the algorithm on a standard dataset, we also selected a subset of images affected by dense occlusion and small targets for testing. The results are shown in Figure 10. We can see that the ADSSD is better to detect occluded and smaller targets. This is because attention mechanisms have been added to the ADSSD to distinguish between background and target and determine the location of the target through the contextual information obtained by the dilated convolution module.

Through comparison, we can see that the ADSSD performed better when detecting occluded and smaller objects.

(**a**) (**b**)

Figure 10. Testing dense target images: (**a**) on the left is use of the SSD; (**b**) on the right is use of the ADSSD.

5. Conclusions

In this paper, aiming at the problem that object detection is easily affected by background information and lack of context information, we proposed an improved SSD called an ADSSD that integrates an attention mechanism and dilated convolution. In addition, we introduced hierarchical feature fusion to solve the gridding effect. We verified the ADSSD at VOC2007. According to the results, the method proposed by the ADSSD enhanced the ability to extract object information, thereby improving the accuracy of detection by 78.4%. In Section 4.6, the performance of the ADSSD was significantly better than that of the SSD in dense occlusion or small object detection.

Our improvements mitigate the impact of background information and obtain richer context information. To further improve the ADSSD, the introduction of deformable convolution to enhance the adaptability of the network to objects with scale transformations would be a good choice. It will be tested in our future work.

Author Contributions: J.N. and R.W. proposed the idea of the paper. R.W. and J.T. helped manage the annotation group and helped clean the raw annotations. R.W. conducted all experiments and wrote the manuscript. R.W., J.N. and J.T. revised and improved the text. R.W. and J.N. are the people in charge of this project. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Handan Science and Technology Research and Development Program (19422091008-35) and the Hebei Science and Technology Program (21350101D).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: A part of the dataset is available at: Pascal VOC Dataset Mirror (pjreddie.com).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-Captured Scenarios. *arXiv* **2021**, arXiv:2108.11539.
2. Sun, H.; Xu, H.; Liu, B.; He, D.; He, J.; Zhang, H.; Geng, N. MEAN-SSD: A Novel Real-Time Detector for Apple Leaf Diseases Using Improved Light-Weight Convolutional Neural Networks. *Comput. Electron. Agric.* **2021**, *189*, 106379. [CrossRef]
3. Ranjbarzadeh, R.; Bagherian Kasgari, A.; Jafarzadeh Ghoushchi, S.; Anari, S.; Naseri, M.; Bendechache, M. Brain Tumor Segmentation Based on Deep Learning and an Attention Mechanism Using MRI Multi-Modalities Brain Images. *Sci. Rep.* **2021**, *11*, 10930. [CrossRef] [PubMed]
4. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *arXiv* **2014**, arXiv:1311.2524.
5. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
6. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *Computer Vision—ECCV 2016*; Springer: Cham, Switzerland, 2016; pp. 21–37.
7. Tong, K.; Wu, Y.; Zhou, F. Recent Advances in Small Object Detection Based on Deep Learning: A Review. *Image Vis. Comput.* **2020**, *97*, 103910. [CrossRef]
8. Shokofeh, A.; Nazanin, T.S.; Negin, M.; Shadi, D.; Amirali, R. Review of Deep Learning Approaches for Thyroid Cancer Diagnosis. *Math. Probl. Eng.* **2022**, *2022*, 5052435.
9. Girshick, R. Fast R-CNN. *arXiv* **2015**, arXiv:1504.08083.
10. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2015; Volume 28.
11. Lim, J.-S.; Astrid, M.; Yoon, H.-J.; Lee, S.-I. Small Object Detection Using Context and Attention. In Proceedings of the 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Jeju Island, Republic of Korea, 13–16 April 2021; pp. 181–186.
12. Lu, X.; Ji, J.; Xing, Z.; Miao, Q. Attention and Feature Fusion SSD for Remote Sensing Object Detection. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 5501309. [CrossRef]
13. Zhou, T.; Wang, S.; Zhou, Y.; Yao, Y.; Li, J.; Shao, L. Motion-Attentive Transition for Zero-Shot Video Object Segmentation. *arXiv* **2020**, arXiv:2003.04253. [CrossRef]
14. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. *arXiv* **2018**, arXiv:1709.01507.
15. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial Transformer Networks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2015; Volume 28.
16. Lian, J.; Yin, Y.; Li, L.; Wang, Z.; Zhou, Y. Small Object Detection in Traffic Scenes Based on Attention Feature Fusion. *Sensors* **2021**, *21*, 3031. [CrossRef] [PubMed]
17. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. *arXiv* **2018**, arXiv:1807.06521.
18. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv* **2016**, arXiv:1511.07122.
19. Jiang, J.; Zhai, D.H. Single-stage object detection algorithm based on atrous convolution and feature enhancement. *Comput. Eng.* **2021**, *47*, 232–238, 248.
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
21. Yin, Q.; Yang, W.; Ran, M.; Wang, S. FD-SSD: An Improved SSD Object Detection Algorithm Based on Feature Fusion and Dilated Convolution. *Signal Process. Image Commun.* **2021**, *98*, 116402. [CrossRef]
22. Li, Y.; Chen, Y.; Wang, N.; Zhang, Z. Scale-Aware Trident Networks for Object Detection. *arXiv* **2019**, arXiv:1901.01892.
23. Mehta, S.; Rastegari, M.; Caspi, A.; Shapiro, L.; Hajishirzi, H. ESPNet: Efficient Spatial Pyramid of Dilated Convolutions for Semantic Segmentation. *arXiv* **2018**, arXiv:1803.06815.
24. Everingham, M.; Gool, L.V.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2009**, *88*, 303–308. [CrossRef]
25. Purkait, P.; Zhao, C.; Zach, C. SPP-Net: Deep Absolute Pose Regression with Synthetic Views. *arXiv* **2017**, arXiv:1712.03452.
26. Kong, T.; Yao, A.; Chen, Y.; Sun, F. HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 845–853.
27. Shrivastava, A.; Gupta, A.; Girshick, R. Training Region-Based Object Detectors with Online Hard Example Mining. *arXiv* **2016**, arXiv:1604.03540.
28. Bell, S.; Zitnick, C.L.; Bala, K.; Girshick, R. Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2874–2883.
29. Shen, Z.; Liu, Z.; Li, J.; Jiang, Y.-G.; Chen, Y.; Xue, X. DSOD: Learning Deeply Supervised Object Detectors from Scratch. *arXiv* **2017**, arXiv:1708.01241.
30. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. *arXiv* **2018**, arXiv:1608.06993.

31. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
32. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. *arXiv* **2016**, arXiv:1612.08242.