EAI Endorsed Transactions
on Scalable Information Systems

Research Article **EAI.EU**

Multi-scale information fusion based on convolution kernel pyramid and dilated convolution for Wushu moving object detection

Yuhang Li[1,*]

[1]Department of Arts and Sports, Henan Technical College of Construction, Zhengzhou, 450064, China

Abstract

In complex background, the accuracy of moving object detection can be affected by some factors such as illumination change, short occlusion and background movement. This paper proposes a new multi-scale information fusion based on convolution kernel pyramid and dilated convolution for Wushu moving object detection. The proposed model uses a variety of ways to fuse the feature information. First, the multi-layer feature map information with different sizes is fused by the per-pixel addition method. Then the feature map of different stages is splicing in the channel dimension to form the information fusion feature layer with rich semantic information and detail information as the prediction layer of the model. In this model, convolution kernel pyramid structure is introduced into the anchor frame mechanism to solve the multi-scale problem of detecting objects. The number of parameters increased by large convolution kernel is reduced by using dilated convolution to reduce the number of anchor frames reasonably. Experimental results show that the proposed fusion algorithm has certain anti-interference ability and high precision for moving object detection in complex environment compared the state-of-the-art methods.

Keywords: moving object detection, multi-scale information fusion, dilated convolution, convolution kernel pyramid.

Received on 10 September 2021, accepted on 17 September 2021, published on 21 September 2021

doi: 10.4108/eai.21-9-2021.170965

1. Introduction

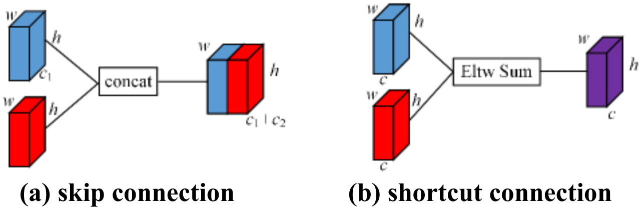
Wushu is a traditional Chinese sport that focuses on both internal and external training, with fighting and defense as its main content, routine and combat as its sport form. Moving object detection refers to the detection of video sequence images and the segmentation of moving object from the complex background. The detection, monitoring and tracking of the object and other subsequent processing are widely used in real-time monitoring and other fields. The background difference method and frame difference method are the two main methods in moving object detection [1-3]. The background difference method is used to analyze the background model and every frame image to be detected. The moving object detection image is obtained by difference calculation. In this method, a stable background model is firstly established, and then the final object image is obtained by subtracting the background model and the image to be detected by comparing the background model with the original image. In addition, the

[*]Corresponding author. Email: xdwangxd@163.com

inter-frame difference method detects the contour of the moving object through the difference operation of adjacent frame images. It has good adaptive ability, simple calculation principle and good real-time performance [4,5]. However, the gap phenomenon exists in the process of object contour extraction, which results in partial feature points being ignored and greatly reduces the detection accuracy [6]. In the two methods, the illumination changes will cause great interference, hinder the acquisition of background model, and affect the accuracy of object detection [7].

With the improvement of computer computing capability, especially the maturity of scientific computing technology based on general purpose Computing unit (GP-GPU), image processing and computer vision have achieved rapid development and made great progress [8,9]. Due to the order of magnitude improvement in computing power, computer vision technology based on deep neural network technology has also made new development, and has made great progress in image classification, segmentation and other tasks [10]. As an important task of computer vision, object detection is also optimized by using deep learning technology.

SqueezeDet [11] serves as a fast object detection framework based on deep learning, it continuously learns from data, automatically extracts the required features, and has natural multi-scale and translational invariance, which makes great progress in the field of object detection. However, in the real scene, the image is seriously affected by lighting, color and other factors. The change of features leads to the failure of model judgment. How to extract invariable features is particularly critical [12]. Due to the pooling operation in convolutional neural network, the resolution becomes smaller and information is lost. At the same time, the change of object distance in the real scene will lead to a large scale change, often resulting in a decrease in detection accuracy. In order to effectively overcome the above problems, this paper proposes a new moving object detection algorithm based on convolution kernel pyramid and dilated convolution. The new algorithm firstly uses Skip Connection and Shortcut Connection to connect feature maps with different resolutions, so as to obtain the lost information and form new feature maps with rich semantics. As shown in figure 1(a), it denotes Skip Connection, that is, two feature maps of the same size are spliced together according to channel dimensions to form a new feature map. Figure 1(b) is Shortcut connection, that is, two feature graphs of the same size and channel dimension are added together in the way of "Eltw sum" to obtain a new feature graph (i.e. the addition of corresponding elements). Then, convolution kernel pyramid structure is introduced into the anchor frame mechanism to solve the problem of mismatch between anchor frame and feature region, so as to detect multi-scale objects more accurately. The introduction of dilated convolution increases the receptive field of convolution kernel without increasing the number of parameters, and determines the number of anchor frames according to the generated predictive tensor, which reduces the time complexity.



(a) skip connection **(b) shortcut connection**

Figure 1. Two connection ways

2. Related works

Moving object detection is a popular research field in computer vision. The current research is mainly based on deep learning methods. The object detection network based on deep learning consists of two basic parts: feature extraction module and object detection module [13,14]. When convolutional neural network (CNN) is used to extract image features, the deep feature graph has rich object semantic information and it is sensitive to category information. However, it lacks detailed information and is often used in classification tasks. However, shallow feature maps have rich details and are sensitive to position, translation and rotation, but lack semantic information. object detection includes classification and object location. The former classifies candidate regions and requires object semantic information [15]. The latter locates candidate areas, requiring details such as location. In order to improve the performance of object detection, feature information of different depths is often fused to facilitate object classification and location.

According to whether feature information fusion is carried out, the object detection network is divided into two types: without feature information fusion and with feature information fusion. In the object detection network without feature information fusion, one type of prediction is based on single-layer feature map, such as two-stage methods (Fast RCNN [16], Faster R-CNN [17]), single-stage methods YOLO (You Only Look Once) [18] and YOLOv2 [19]. The other is to predict on multiple feature graphs, such as SSD (single-shot multi-box detector) [20] and MS-CNN (multi-scale CNN) [21]. In the object detection network with feature information fusion, one kind of prediction is based on a single fused feature graph. For example, HyperNet and Inside-outside not (ION) fuse features of different levels by splicing [22]. The other is prediction over multiple fused feature graphs, such as deconvolutional single-shot dectector (DSSD) [23], which fuses information by multiplying it pixel by pixel. YOLOv3 [24], FPN (Feature Pyramid

Network) [25] and Mask R-CNN [26] fuse information by adding each pixel.

In order to take advantage of different depth feature maps, this paper proposes to fuse feature information at two stages. Firstly, multiple convolutional layers are added after the feature extraction network, and the feature information is fused layer by layer from deep to shallow by pixel-by-pixel addition to form feature maps with rich semantic information and detailed information. Secondly, in order to further enhance the fusion of feature information, the method of channel splicing is used to splicing the feature images with different stages in the fusion feature images obtained in the previous step, forming the feature images with richer semantic and detailed information.

For the problem of multi-scale object detection, the proposed solutions mainly include the following three categories:

The first approach uses an image pyramid network (which extracts features on images of different sizes) such as the Scale Normalization for Image Pyramids (SNIP) algorithm [27] and the face detection algorithm HR (Hybnid Resolution) and gets good results. Its disadvantage is that the algorithms have high time complexity. In order to reduce the time complexity, sparse image pyramid can be adopted, that is, only three different input image sizes can be adopted.

The second approach is to solve the multi-scale problem of the object by using the anchor frame mechanism on the single-layer feature map. For example, Faster RCNN uses RPN (Region Proposal Network) network to extract candidate regions on the deepest feature map. In order to detect objects of different scales, the RPN network predicts nine anchor frames of different sizes and ratios at each anchor point in the feature graph. In addition, RFCN (Region-based Fully convolutional Network) [28] and YOLOv2 also adopt anchor frame mechanism for prediction.

The two-step detection algorithm represented by Faster R-CNN mainly uses candidate frame location, and then classifies candidate frame by classification network and further coordinate regression, which can obtain detection results with higher accuracy. The algorithm process is divided into two steps: 1) Firstly, a Region Proposal Network (RPN) is used to extract regions of interest (RoIs) from an image; 2) Using multi-task classification and regression network to conduct sub-classification and location regression for the regions of interest extracted in step 1 [3].

YOLO1 and SSD are the main one-stage detection algorithms. In order to speed up the processing with a certain amount of accuracy, it pre-sets a set of anchors directly predicting the category and location of the object in the image. The basic idea of one-stage detection method is to extract regions of interest and classify multiple categories, which is similar to RPN network in nature. In order to improve its detection accuracy and generalization ability, SSDS performs detection on multi-scale feature maps, fuses and filters multi-scale detection objects, and refines the final detection results.

Compared with the two-stage detection algorithm, the one-stage detection algorithm is an end-to-end network, where gradients are transmitted well, so the network is relatively easy to train. In addition, the one-stage detection algorithm has no candidate extraction process and it is fast, which is suitable for many real time scenes (such as unmanned driving and video object detection). SqueezeDet is a detection algorithm similar to YOLO. Firstly, a pre-trained model in ImageNet is used to extract high-dimensional feature images from an image. Compared with YOLO, SquezeDet uses convolution layer instead of full connection layer, which greatly reduces the parameters of the entire network and further improves the generalization capability and speed of the network. However, due to the feature graph of single scale, SqueezeDet performs poorly in unmanned driving, video detection and other rapidly changing scenes.

The third way is to make predictions on a feature pyramid. The SSD object detection network predicts objects with different scales on different feature maps, forming the prototype of feature map pyramid. Both DSSD and FPN are predicted on the pyramid of feature graph to deal with multi-scale problem of object.

The above three ideas are effective means to solve the multi-scale problem of detection objects. In this paper, the anchor frame mechanism is improved to solve the multi-scale problem of the object. In the anchor frame mechanism of RPN, each anchor point on the feature graph for prediction corresponds to 9 anchor frames of different sizes and ratios. During prediction, 1×1 convolution kernel is used to predict the position and confidence of multiple anchor frames of different sizes (i.e. the probability that the object contained in the anchor frame belongs to a certain category). Therefore, for anchor frames of different sizes corresponding to anchor points, the same feature area on the feature graph is used in prediction, resulting in the mismatch between the feature area used in prediction of RPN network and the corresponding anchor frame area. Therefore, this paper proposes to introduce convolution kernel pyramid structure into the anchor frame mechanism to detect objects with different sizes, so that the size of convolution kernels corresponding to anchor frames of different sizes is different, while the size of convolution kernels corresponding to anchor frames of the same size but with different ratios is the same, so as to alleviate the problem of mismatch. In addition, large convolutional kernels will increase the number of parameters, in order to reduce the time complexity, the model adopts the dilated convolutional mechanism to design convolutional kernels with different sizes of receptive fields [29]. Under the action of convolution kernels with different sizes, prediction tensors with different resolutions are generated on feature graphs with rich semantic and detailed information (i.e. feature graphs). The model determines the number of anchor frames according to the generated prediction tensor, making small objects correspond to small anchor frames and the number is large, and large objects correspond to large anchor frames and the number is small, thus reducing the number of anchor frames reasonably.

Context information is very important in object detection. For example, if a person is wearing cat ears, the detection algorithm is very likely to misdetect if only cat ears are seen. If context information can be connected, false checks can be avoided. Dilated convolution can effectively collect multi-scale and context information. Figure 2 shows dilated convolution of 3×3. Red is the center of the convolution kernel and blue is the surrounding point. Compared with traditional convolution, dilated convolution has an extra hyperparameter rate. In figure 2, (a), (b) and (c) are dilated convolution with rate=1,2,3 respectively. Figure 1(a) is exactly the same as traditional convolution, which can be regarded as a special case of dilated convolution. As you can see from figure 2, you can get information around the object by setting different rates. Inspired by the existing method, this paper combines dilated convolution, skip connection and shortcut connection to combine with context and multi-scale information and enhance the feature expression ability.

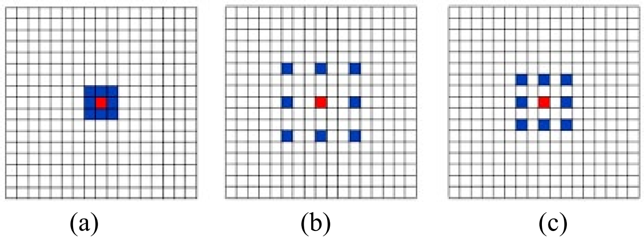


Figure 2. Dilated convolution

3. Proposed CKP-DC for moving object detection

The object detection model based on deep learning takes the feature extraction module of classification network as the basic network. Better classification networks include VGGNet and ResNet, etc. Considering performance and speed comprehensively, VGG16 network is selected as the basic network of CKP-DC in this paper. The convolution kernel pyramid and dilated convolution is abbreviated as CKP-DC.

The proposed object detection model is shown in figure 3. The blue feature map represents the original prediction layer of SSD. The arrow represents the resolution of the feature map by the operation of bilinear interpolation. "+" denotes the fusion with the previous layer feature map by adding each pixel. The green feature graph and arrow indicate the feature graph in different stages. F is the size of the convolution kernel. D represents the dilated convolution coefficient.

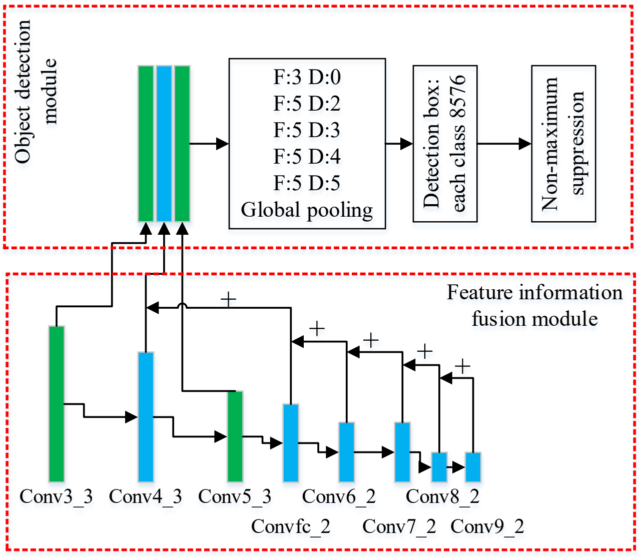


Figure 3. Proposed moving object detection framework

3.1. Multi-scale information fusion

The fusing process of the CKP-DC model starts at the deepest level of the original SSD prediction layer. First, the bilinear interpolation is used to increase the resolution of the feature image. Then, it uses the method of adding pixel by pixel to fuse the previous layer feature map. In this way, the layers are sampled and fused up to the shallowest layer of the original SSD prediction layer to form a feature map containing both detail information and semantic information. This fusion process can be seen from the blue feature diagram of the feature information fusion module in figure 3 and the fusion part shown by the line. In addition, feature maps of different stages are spliced to further enhance the semantic and detail information of the predicted feature maps. This fusion process can be seen in the green feature map of the feature information fusion module in figure 3 and the fusion part shown by the line. Then the feature images fused in different ways are spliced into the final prediction feature images by channel splicing. Because it preserves detail information and semantic information better, using such features is not only beneficial to the detection of large objects, but also can enhance the ability of the model to detect small objects.

Detection is performed on the last feature graph, and the width and height of the feature graph are 1/16 of the original image, that is, the object of 16×16 in the original image is mapped to the feature graph with a size of only 1×1, and the information loss is very serious. Therefore, the small moving object detection is very difficult. For the convolutional neural, due to the existence of down-sampling, many feature images with different resolutions are generated in the middle layer, which have natural multi-scale information. Generally, the larger feature image in the front has local details of the object, while the smaller feature image in the back has richer semantics. Therefore, this paper considers using Skip Connection to combine feature maps with different resolutions and detect them on larger feature maps. This has two advantages:

(1) The newly obtained layer contains both rich semantics and local details of objects, and can make good use of multi-scale information in convolutional neural network;

(2) After the feature graph expanding, the number of selected anchor increases and the sampling becomes more intensive, so the location of the object can be better obtained.

Generally, the contextual information around the object plays a very important role in object detection. By setting different rates, dilated convolution can obtain object information of different ranges.

Therefore, we should first combine dilated convolution of different rates in parallel to extract information around the object as shown in figure 4 and figure 5. Similar residual networks fuse input and output results enabling dilated convolution to better extract context information.

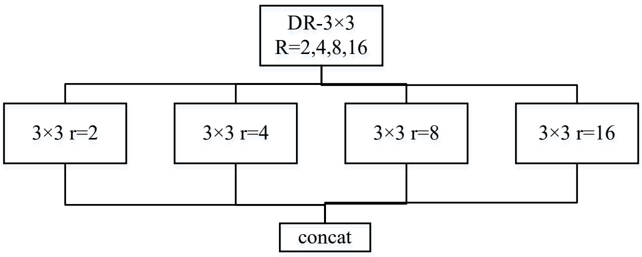


Figure 4. DR module

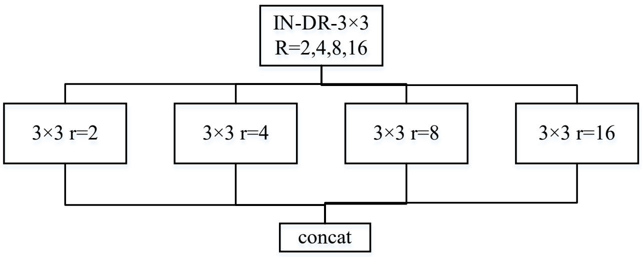


Figure 5. IN-DR module

Dilated convolution can rapidly increase the receptive field in series, and the appropriately large receptive field is helpful for object detection. Therefore, this paper considers series of dilated convolution. In order to make parameter utilization higher. In this paper, the number of channels through each layer is halved in the series process, and the feature graph generated in the middle is also taken as the output of the module, so as to realize feature reuse, as shown in figure 6. When the input and output are 512 channel feature graphs, the series and parallel modules can save about 1/3 parameters.

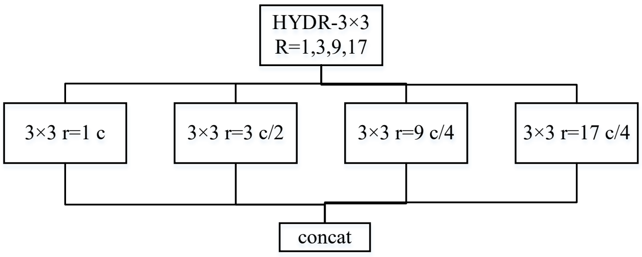


Figure 6. HYDR module

In this paper, dilated convolution and Skip Connection are combined. The network structure of the algorithm is shown in figure 7. The pooling of convolutional neural network is often accompanied by the loss of information. In this paper, shortcut connection is also used to connect smaller feature graphs with larger feature graphs in the form of "Eltw sum" to mitigate the loss of information in the pooling process [30]. In order to keep the pre-training model unchanged, 1×1 convolution in the shortcut connection module is initialized to 0 in this paper.

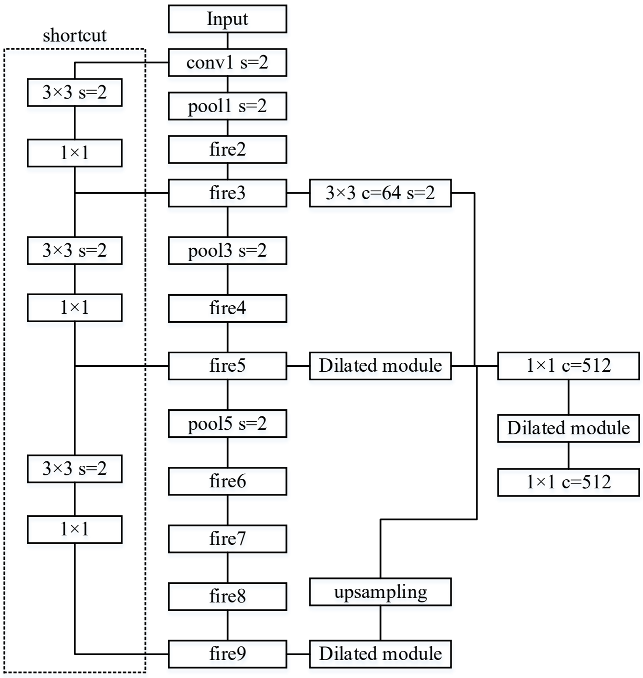


Figure 7. Multi-scale information fusion

3.2. Moving object detection

In order to detect objects of different scales on feature maps with rich semantic information and detailed information, convolution kernel pyramid structure is adopted. Convolution operation is carried out with convolution check feature graphs of different sizes to generate prediction tensors of different sizes, confidence degree and location information based on the classification of the prediction object. The convolution operation of the feature graph with different sizes corresponds to the convolution operation of the receptive field with different sizes on the original graph, thus facilitating the detection of objects with different sizes. As large convolution kernels bring a large number of parameters, the computation of the model is greatly increased. Therefore, dilated convolution is adopted in the convolution kernel pyramid to increase the convolution kernel receptive field without increasing the number of parameters. The pyramid module of convolution kernel based on dilated convolution is shown in figure 8. The prediction tensor is generated through the action of different convolution kernels. Two convolution operations are performed on each set of output tensors, and then the categories and positions of boundary boxes are predicted respectively, as shown in the light red and light blue rectangular boxes in the figure.

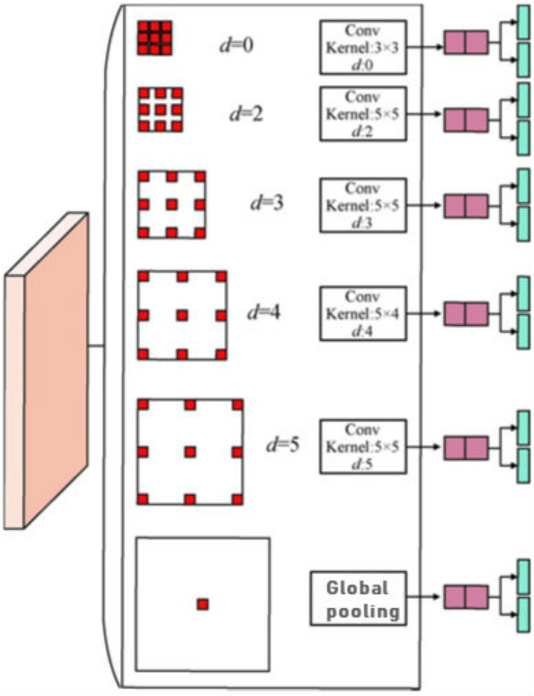


Figure 8. Filter pyramid based on dilated convolution

The final feature map for prediction has a resolution of 38×38 pixels. Multiple sets of convolution kernels of different sizes are designed to cover receptive fields of different sizes on the original image as evenly as possible, so as to better predict objects of different scales. The minimum convolution kernel size is $k_{\min}$ and its value is 3. The maximum convolution kernel size is $k_{\max}$ and the value is 38. n tensors of different sizes are predicted by using n convolution kernels of different sizes (n=6). In order to keep the convolution kernel size evenly distributed between 3 and 38, the size of the m-th convolution kernel is:

$$k_{\min} + \frac{k_{\max} - k_{\min}}{n-1} \times (m-1) \tag{1}$$

The resolution of each group of output prediction tensors should also meet the requirement of uniform distribution so as to make the number of anchor frames with different sizes more effective and reasonable. Therefore, the size of the predicted tensor of the m-th output is:

$$\left\lceil 38/2^{m-1} \right\rceil \tag{2}$$

In the formula, $\lceil \quad \rceil$ means rounding up. According to the requirements of the convolution kernel size distribution and the resolution of the output prediction tensor, and through experimental verification, a set of convolution kernel size definition mechanism is designed, as shown in table 1. r and d represent the actual convolution kernel size and cavity coefficient. s stands for step size. p represents the filling condition. e is equivalent to the size of ordinary convolution kernel in the case of dilated convolution. o represents the resolution of the prediction tensor. e satisfies the following:

$$e = r + (r+1) \times (d-1) \tag{3}$$

o satisfies the following:

$$o = \left\lfloor \frac{38 - e + 2p + s}{s} \right\rfloor \tag{4}$$

In the formula, $\lfloor \quad \rfloor$ means rounding down. The experimental results show that the design scheme can meet the requirement that the convolution kernel uniformly covers different receptive fields on the original image. Compared with the structure that using dilated convolution in multiple series, the dilated convolution structure designed in this paper uses multiple different dilated convolutions in parallel, and there is only one dilated convolution structure in the backbone network. Therefore, the Gridding effect of dilated convolution is not obvious. In addition, the dilated convolution coefficients in this paper are small, each dilated convolution coefficient is different and its greatest common divisor is less than 1, which conforms to the main features of HDC (Hybrid Dilated Convolution) module [31], further reducing the Gridding effect.

Table 1. Designed dilated convolution filter mechanism

Convolution kernel	The actual convolution kernel size	Dilated coefficient	Stride	Filling situation	The size of the ordinary convolution kernel	The resolution of the prediction tensor
1	3	0	1	1	3	38
2	5	2	2	4	11	18
3	5	3	2	0	17	11
4	5	4	3	0	23	6
5	5	5	3	0	29	3
6	-	-	-	-	38	1

Anchor box mechanism

The number of anchor frames in the SSD model is generated based on the resolution of the feature graph for which the prediction is made. If the resolution of the feature map is N×N, the input image is divided into N×N grids. Four or six anchor frames of varying ratios are generated for each grid. If this scheme is also adopted in this paper, the same number of anchor frames of different proportions will be generated for each convolution kernel on the original image, and 46208 (38×38 (2×4+4×6)) anchor frames will be generated with the input of the model of 300×300 pixels. It can be seen from YOLOv2 that the average detection accuracy will decrease if the number of anchor frames is too much [32]. Therefore, this paper proposes a scheme to reduce the number of anchor frames according to the resolution of the generated predictive tensor.

Convolution kernel pyramid and anchor frame mechanism are combined to deal with multi-scale problem of object detection. Convolution check of the same size should have multiple anchor frames, which have the same size and different ratios. Convolution of different sizes should have different sizes of anchor frames. Specifically, under the action of convolution kernels of different sizes, the same feature graph generates prediction tensors with different resolutions, so that the grid number divided on the original image and the resolution of prediction tensors remain the same. In this way, the small object corresponds to the small frame, and the number is larger. A large object corresponds to a small number of anchor frames. This can not only solve the problem of detecting objects of different sizes, but also reduce the number of anchor frames reasonably. Using this mechanism, under the 300×300 pixels, it has 8576 anchor frames (38×38×4+18×18×6 +10×10×6+6×6×6+3×3×6+1×1×4), which is similar to the 8732 anchor frames in SSD, it is reasonable.

This paper determines the size of anchor frame according to the size of convolution kernel and experiments. When the convolution kernel size is k, assuming that the ratio of an anchor frame is r, then the width W and height H of the anchor frame are respectively:

$$W = k \times \alpha \times \sqrt{r} \tag{5}$$

$$H = k \times \alpha / \sqrt{r} \tag{6}$$

Where, α is the hyperparameter according to the actual situation. When the ratio is 1, one more case is added, i.e,

$$W = H = \alpha\sqrt{k \times (k+7)} \tag{7}$$

Therefore, the size of the anchor frame is different for convolution kernels of different sizes. Convolution kernels of the same size have the same size but different ratios.

Modal training

The CKP-DC model initializes model parameters using SSDS as pre-training models. Similar to SSD, the data is enhanced to improve the detection accuracy and robustness of the model. In the matching process of anchor frame and truth label, each truth label is matched to any anchor frame whose IOU (intersection over Union) is greater than 0.5. When a truth label has no matching object, matching it with the largest anchor box of its IOU. For the anchor frame with no matching, the front one is selected as the negative sample according to the predicted confidence, so that the ratio of negative sample to positive sample is 3:1. The loss function definition of the model is the same as that of SSD model, which consists of the sum of smooth L1 positioning loss and Softmax classification loss.

4. Experiments and analysis

The proposed method is tested and evaluated on PASCAL VOC [33], remote sensing data set UCASAOD [34] and real martial arts moving data. The code is implemented on the Caffe deep learning framework, utilizing some artifacts of the SSD and DSSD Caffe open source libraries. All experiments are performed on an HP workstation equipped with a Titan X GPU. SSD is used as the pre-training model of the proposed method, and the model is fine-tuned on PASCALVOC, UCAS-AOD remote sensing dataset and real martial arts moving data. mAP (Mean Average Precision) is used to evaluate the performance of the new method. The proposed method is compared with other advanced deep learning object detection methods in terms of mAP and detection speed. Precision and Recall are defined as:

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

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

Where TP indicates that the positive class is judged to be positive class. FN indicates that positive class is judged to be negative class. FP refers to the negative class is judged to be positive class. TN means the negative class is judged to be negative. AP is defined as the average of the maximum accuracy at 11 different recall rate levels [0, 0.1, 0.2,...,1].

$$AP = \frac{1}{11} \sum_{r\in\{0,\cdots,1\}} P_{\max}(r) \tag{10}$$

Where, $P_{\max}(r)$ refers to the maximum accuracy when the recall rate is r. AP is the evaluation index of a single category. mAP refers to the average AP of multiple objects.

4.1. Experiments on PASCAL VOC 2007

DSSD uses ResNet-101 network as the basic network to extract features. In order to facilitate comparison, the DSSD model with VGG-16 version is established based on VGG-6 network in the experiment, and the model is trained according to the training strategy of DSSD in the original text. In order to make equitable comparison with other advanced algorithms, the CKP-DC model is trained on the PASCAL VOC2007 and PASCAL VOC2012 joint training set. The results are evaluated on PASCAL VOC2007 test set. During training, the weight of the original SSD model is fixed first, and only the additional network parts are trained. In the first 7×10^4 iterations, the learning rate is 0.001. In the next 3×10^4 iterations, the learning rate is 0.0001. The entire network is then fine-tuned to train at a learning rate of 0.001 in the first 2×10^4 iterations, and then at a learning rate of 0.0001 in the second 2×10^4 iterations.

Table 2 shows the comparison of the model parameters with and without dilated convolution. Both basic networks are the same, and the number of basic network parameters is not considered. It can be seen that the parameter number is significantly reduced using dilated convolution.

Table 2. Number of model parameters comparison

State	Parameter number/10^6
With dilated convolution	28.67
Without dilated convolution	123.68

Table 3 shows the results of the proposed detection algorithm and the current advanced algorithms on PASCAL VOC2007 set. All the methods in the table are based on VGG-16, SSD300, DSSD300 and CKP-DC results are obtained from the experiments in this paper, and other results are from the original literature. The input image size is 300×300pixel. Under the VGG-16 network, the CKP-DC achieves the accuracy of 79.3% mAP. Compared with the two-stage method, CKP-DC is 9.3% higher than Fast RCNN, 6.1% higher than Faster RCNN with anchor frame mechanism, 3.7% higher than ION with channel dimension splice and fusion feature information, and 1.1% higher than MRCNN. Compared with single-stage method, CKP-DC is 1.8% higher than SSD model and 0.9% higher than DSSD model on the same input image and base network. In addition, CKP-DC is better than some improved SSDS, DSSD. CKP-DC is 0.7% and 0.4% higher than MDSSD300 and feature-fused SSD respectively, and 0.6% higher than FSSD300 and RSSD300 on average and 1.6% higher than DSOD300, indicating that the proposed detection model in this paper has better performance.

Table 3. Results of CKP-DC and other advanced algorithms on PASCAL VOC2007 test set

Method	mAP	bus	cat	chair	table	Sheep	sofa	train
Fast RCNN	70.0	81.6	86.7	42.8	68.9	70.1	74.8	80.4
Faster RCNN	73.2	83.1	86.4	52.0	65.7	73.6	73.9	83.0
ION	75.6	85.4	87.0	54.4	73.8	75.8	72.7	84.2

MRCNN	78.2	88.0	87.8	60.3	73.7	76.3	75.5	85.0
SSD300	77.5	87.0	88.1	60.3	77.0	77.9	79.5	87.6
DSSD300	78.4	86.0	88.3	62.2	78.0	79.8	79.0	87.5
Feature-fused	78.9	86.6	88.3	63.2	76.8	80.6	79.5	88.2
MDSSD300	78.6	86.9	88.1	58.5	73.4	78.6	74.5	86.8
CKP-DC	**79.3**	**88.0**	**88.6**	**64.8**	**80.2**	**82.2**	**83.7**	**89.0**

Note that: the bold values are the best.

Table 4 shows the detection results with different methods on ResNet-101 for the PASCAL VOC2007 test set. The results of the compared algorithms are from the original literature. CKP-DC is 2.9% higher than Faster R-CNN, 2.2% than SSD, and 0.7% than DSSD. The results show that increasing the depth of the model is helpful to improve the detection accuracy, but the detection accuracy does not improve further when the layer number increases to a certain extent. CKP-DC has a slightly lower accuracy than RFCN and DSSD. However, according to the comparison of test times in figure 9, the test speed of RFCN is 9 frames/s, while that of DSSD is 5.5 frames/s, which is far lower than the 21 frames/s of the proposed model in this paper.

Table 4. Results of other advanced methods of ResNet-101 on PASCAL VOC2007 test set

Method	mAP	aero	bike	cow	table	dog	mbike	train
Faster RCNN	76.4	79.8	80.7	87.8	69.4	88.3	80.9	85.3
R-FCN	**80.5**	79.9	**87.2**	**88.1**	74.5	**89.8**	79.9	85.9
SSD321	77.1	76.3	84.6	82.6	76.9	86.7	85.4	**87.3**
DSSD321	78.7	**81.9**	84.9	83.5	**78.7**	86.7	**86.7**	87.2

1000 images are used and batch-size is set to 1 to evaluate the test speed of CKP-DC, as shown in figure 9. It compares some current advanced methods, the results are from the original literature. The GPU model used for each experimental result is TitanX 1060 [35]. All methods here are trained on the PASCAL VOC2007 and VOC2012 and tested on the PASCAL VOC2007 test set. CKP-DC has a speed of 21 frames/s on 300×300 pixel input images, it is faster than the two-stage methods, slower than SSD, but faster than DSSD. It has higher detection accuracy than SSD and DSSD and related improved single-stage detection methods.

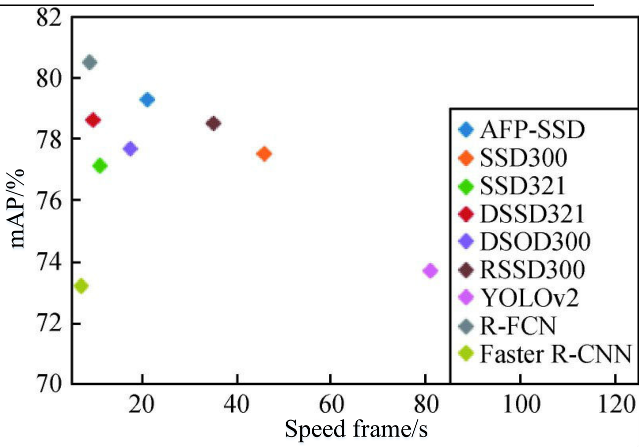


Figure 9. Comparison of accuracy and speed on PASCAL VOC2007 test set

4.2. Experiments on UCAS-AOD

The UCAS-AOD remote sensing dataset contains 1000 images with aircraft, which are randomly divided into training sets and test sets in a 7:3 ratio. Each of the 1000 remote sensing images has a resolution of 1 280×659 pixels.

When converting all data to an imdb format that Caffe can recognize, the image sizes and bounding box labels are also fully scaled so that the model's input image size is 300×300 pixels. SSD, DSSD and CKP-DC models are initialized with parameters trained on VOC datasets and fine-tuned on remote sensing training datasets. The first 6×10^4 iterations are trained with learning rate of 0.001, and the next 2×10^4 iterations are trained with learning rate of 0.0001. The object detection results of each approach are shown in table 5. The average precision (AP) of CKP-DC is 2.8% and 1.9% higher than that of SSD and DSSD, respectively.

Table 5. Results of CKP-DC and other methods on remote sensing dataset

Method	Base network	size	AP
SSD300	VGG	300×300 pixel	88.3
DSSD300	VGG	300×300 pixel	89.2
CKP-DC	VGG	300×300 pixel	**91.5**

Figure 10 shows some detection results of SSDS and CKP-DC on the PASCAL VOC2007 test set. Row 1 and row 3 are the results of SSD model detection, and rows 2 and 4 are the results of CKP-DC model detection. Figure 11 shows some detection results of the two models on the UCAS-AOD remote sensing dataset. Figure 11(a) and (b) are the results of SSD and CKP-DC detection respectively. Only test results with a confidence level higher than 0.8 are displayed in the figure. It can be seen that CKP-DC does a better detection than SSD in detecting object overlap and small objects.

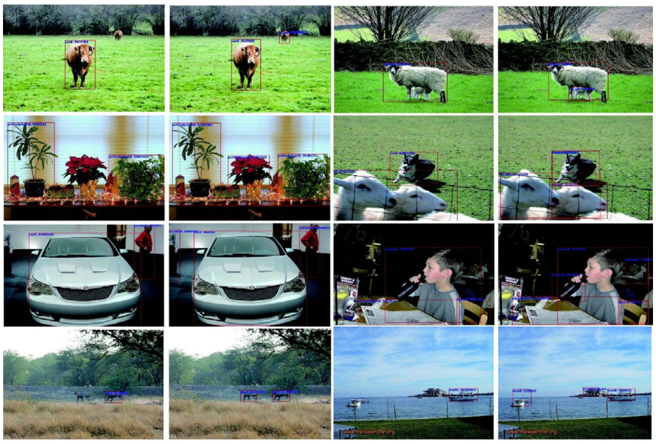


Figure 10. Results of SSD and CKP-DC on PASCAL VOC2007 test set

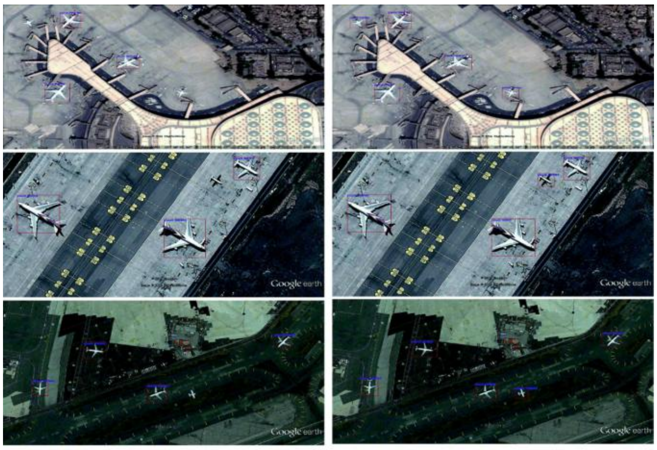


Figure 11. Results of SSD and CKP-DC in UCAS-AOD dataset

4.3. Experiments on Wushu data sets

We select Jab, push, brace, lift, elbow actions in Wushu dataset to conduct experiment [36]. The results are shown in table 6. The time is also the shortest. It shows that the proposed has better effect.

Table 6. AP results of CKP-DC and other methods on Wushu dataset/%

Method	jab	push	brace	lift	elbow	Average time
SSD300	69.3	58.1	71.2	69.3	66.2	10.6s
DSSD300	71.2	63.7	77.4	74.9	77.1	5.8s

CKP-DC	**78.6**	**73.4**	**81.1**	**82.5**	**83.6**	**2.3s**

5. Conclusion

A convolution kernel pyramid and dilated convolution model for Wushu object detection is proposed in this paper. Firstly, the feature information is fused by adding pixel by pixel and splicing channel to form a feature map with rich semantic information and detail information, which is used as a prediction feature map to provide rich feature information for predicting the category and position of boundary boxes. Then it introduces convolution kernels in the mechanism of anchor box of pyramid structure, overcomes the problem that the anchor box area does not match the corresponding features. In order to more accurately detect multi-scale object, at the same time, shallow convolution is increased. Due to the effective information fusion and the introduction of convolution kernel pyramid structure in the anchor frame mechanism, compared with the current advanced methods, the model has faster detection speed and higher detection accuracy, especially better solves the detection problems of small objects and overlapping objects. The future works will focus on the practical application and it will produce enormous economic benefits.

Acknowledgements.

The authors gratefully acknowledge the anonymous reviewers for their valuable comments.

References

[1] Yang, Y., Kurnianggoro, L., Jo, K. H. (2019) Moving Object Detection for a Moving Camera Based on Global Motion Compensation and Adaptive Background Model. International Journal of Control Automation and Systems 17(2).

[2] Zhao, N., Wang, X., Yin, S. (2021) Research of Fire Smoke Detection Algorithm Based on Video. International Journal of Electronics and Information Engineering 13(1): 1-9.

[3] Ju, J., Xing, J. (2019) Moving object detection based on smoothing three frame difference method fused with RPCA[J]. Multimedia tools and applications 78(21): 29937-29951.

[4] Montero, V., Jung, Y., Jeong, Y. (2021) Fast background subtraction with adaptive block learning using expectation value suitable for real-time moving object detection. Journal of Real-Time Image Processing 18(1):1-15.

[5] A, J., Yin, S. (2021) A New Feature Fusion Network for Student Behavior Recognition in Education. Journal of Applied Science and Engineering 24(2): 133-140.

[6] Cuevas, C., Berjon, D., Moran, F., et al. (2012) Moving object detection for real-time augmented reality applications in a GPGPU. IEEE Transactions on Consumer Electronics 58(1):117-125.

[7] Xi, C., Chen, X., Cao, J. (2015) Research on Moving Object Detection Based on Improved Mixture Gaussian Model. Optik - International Journal for Light and Electron Optics 126(20):2256-2259.

[8] An, L., Maunder, R. G., Al-Hashimi, B., et al. (2016) Implementation of a Fully-Parallel Turbo Decoder on a General-Purpose Graphics Processing Unit. IEEE Access 4:5624-5639.

[9] Yin, S., Li, H. (2020) Hot Region Selection Based on Selective Search and Modified Fuzzy C-Means in Remote Sensing Images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13: 5862-5871.

[10] Yin, S., Li, H., Teng, L. (2020) Airport Detection Based on Improved Faster RCNN in Large Scale Remote Sensing Images. Sensing and Imaging 21.

[11] Wu, B., Wan, A., Iandola, F., et al. (2017) SqueezeDet: Unified, Small, Low Power Fully Convolutional Neural Networks for Real-Time Object Detection for Autonomous Driving. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 446-454, doi: 10.1109/CVPRW.2017.60.

[12] Haque, M F., Kang, D S. (2021) PPCNN: Object Detection using Fine-grained Feature Extraction and Localization. The Journal of Korean Institute of Information Technology 19(2):29-37.

[13] Wang, X., Yin, S., Li, H. (2020) A Network Intrusion Detection Method Based on Deep Multi-scale Convolutional Neural Network. International Journal of Wireless Information Networks 27(4): 503-517.

[14] Yin, S., Li, H., Teng, L., et al. (2020) An optimised multi-scale fusion method for airport detection in large-scale optical remote sensing images. International Journal of Image and Data Fusion 11(2): 201-214.

[15] Thai, V X., Jang, G C., Jeong, S Y., et al. (2020) Symmetric Sensing Coil Design for the Blind-Zone Free Metal Object Detection of a Stationary Wireless Electric Vehicles Charger. IEEE Transactions on Power Electronics 35(4):3466-3477.

[16] Wang, X., Shrivastava, A., Gupta, A. (2017) A-Fast-RCNN: Hard Positive Generation via Adversary for Object Detection," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3039-3048, doi: 10.1109/CVPR.2017.324.

[17] Li, C., Zhou, P. (2020) Improved Faster RCNN Object Detection. World Scientific Research Journal 6(3):74-81.

[18] Xiang, Z., Seeling, P., Fitzek, F. (2021) You Only Look Once, But Compute Twice: Service Function Chaining for Low-Latency Object Detection in Softwarized Networks. Applied Sciences 11(5):2177.

[19] Sang, J., Wu, Z., Guo, P., et al. (2018) An Improved YOLOv2 for Vehicle Detection. Sensors 18(12).

[20] Kumar, A., Zhang, Z J., Lyu, H. (2020) Object detection in real time based on improved single shot multi-box detector algorithm. EURASIP Journal on Wireless Communications and Networking, 2020 (204).

[21] Zhang, H., Wang, K., Tian, Y., et al. (2018) MFR-CNN: Incorporating Multi-Scale Features and Global Information for Traffic Object Detection. IEEE Transactions on Vehicular Technology 67(9):8019-8030.

[22] Bell, S., Zitnick, C. L., Bala, K., and Girshick, R. (2016) Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2874-2883, doi: 10.1109/CVPR.2016.314.

[23] Cui, L., Ma, R., Lv, P., et al. (2020) MDSSD: multi-scale deconvolutional single shot detector for small objects. Sciece China. Information Sciences 63(2).

[24] Chung, Y L., Lin, C K. (2020) Application of a Model that Combines the YOLOv3 Object Detection Algorithm and Canny Edge Detection Algorithm to Detect Highway Accidents. Symmetry 12(11):1875.

[25] Karim, S., Zhang, Y., Yin, S. et al. (2019) Impact of compressed and down-scaled training images on vehicle detection in remote sensing imagery. Multimedia Tools and Applications 78: 32565-32583.

[26] Loh, D R., Wen, X Y., Yapeter, J., et al. (2021) A Deep Learning Approach to the Screening of Malaria Infection: Automated and Rapid Cell Counting, Object Detection and Instance Segmentation using Mask R-CNN. Computerized Medical Imaging and Graphics 88.

[27] Sahilliolu, Y., Yemez, Y. (2013) Scale Normalization for Isometric Shape Matching. Computer Graphics Forum 31(7):2233-2240.

[28] Xu, C., Hong, X., Yao, Y., et al. (2020) Multi-Scale Region-based Fully Convolutional Networks. Neurocomputing 391: 220-226.

[29] Wang, J., Li, H., Yin S., and Sun, Y. (2019) Research on Improved Pedestrian Detection Algorithm Based on Convolutional Neural Network. 2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), 2019, pp. 254-258, doi: 10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00063.

[30] Lakshmi, T N., Janaki, N., In, T. (2018) Deep learning based detection and recognition of objects using mobile nets and SSDs. International Journal of Development Research 3(2).

[31] Jiang, W., Liu, M., Peng, Y., et al. (2021) HDCB-Net: A Neural Network With the Hybrid Dilated Convolution for Pixel-Level Crack Detection on Concrete Bridges. IEEE Transactions on Industrial Informatics 17(8): 5485-5494.

[32] Yin, S., Zhang, Y., and Karim, S. (2019) Region search based on hybrid convolutional neural network in optical remote sensing images. International Journal of Distributed Sensor Networks 15(5).

[33] Xu, Y., Zhou, C., Yu, X., et al. (2021) Pyramidal Multiple Instance Detection Network With Mask Guided Self-Correction for Weakly Supervised Object Detection. IEEE Transactions on Image Processing 30: 3029-3040.

[34] Ji, J., Zhang, T., Yang, Z., et al. (2019) Aircraft Detection from Remote Sensing Image Based on A Weakly Supervised Attention Model. IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, 2019, pp. 322-325, doi: 10.1109/IGARSS.2019.8899864.

[35] Yin, S., Liu, J., Teng, L. (2018) Strategic Target Classification with Transfer Learning. International Journal of Electronics and Information Engineering 9(1): 22-28.

[36] Galyaev, A A., Lysenko, P V., Yakhno, V P. (2021) Evading a Single Detector by an Object Moving at a Given Speed. Automation and Remote Control 82(7):1281-1291.