Enhanced Single Shot Multiple Detection for Real-Time Object Detection in Multiple Scenes

Divine Njengwie Achinek*
Information Science and Technology College, Dalian Maritime University, Dalian 116026, China
achinekdivine002@dlmu.edu.cn

Ibrahim Shehi Shehu
Information Science and Technology College, Dalian Maritime University, Dalian 116026, China
shehu@dlmu.edu.cn

Athuman Mohamed Athuman
Information Science and Technology College, Dalian Maritime University, Dalian 116026, China
realathu@dlmu.edu.cn

Xianping Fu
Information Science and Technology College, Dalian Maritime University, Dalian 116026, China and Pengcheng Laboratory, Shenzhen 518055, China
fxp@dlmu.edu.cn

ABSTRACT

CNN-based object detection architectures have achieved great performances in recent times using SSD, YOLO, and R-CNN. However, using these algorithms for real-time detection suffer from low FPS and accuracy. In this paper, we enhanced the conventional SSD as research has shown that it has higher FPS and accuracy compared to others making it more suitable for real-time object detection. However, this conventional SSD suffers computational complexity and low accuracy for small objects detection. We proposed an enhanced SSD for real-time object detection to improve the accuracy of conventional SSD and reduce its computational complexity with a higher FPS. Our main contribution is at the level of the multi-scale object detection, where we implemented PIV layers for enhanced localization and detection of objects in the feature layers. Furthermore, we introduced extended dilated convolutions with different dilation operations thereby increasing the receptive field and improved the detection of objects. To demonstrate the effectiveness of our proposed method, we first carried out experiments on PASCAL VOC 2007 and PASCAL VOC 2012 and achieved improved performances in mAP of 82.0 and mAP of 85.6 on PASCAL VOC 2007 and PASCAL VOC 2012 respectively at 63 FPS, with input size of 300x300 for a batch size of 8. Using the same experimental approach, we further demonstrated the versatility of the proposed method on the underwater image dataset where we achieved also improved performance in mAP of 79.1. Our experimental results have shown to be an effective alternative for real-time objection detection to the conventional SSD and other state-of-the-art architectures.

CCS CONCEPTS

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

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

ACM ISBN 978-1-4503-8985-3/21/10…$15.00
https://doi.org/10.1145/3487075.3487082

KEYWORDS

Convolutional neural network, Object detection, Inceptions, Extended dilated convolution, Object localization

ACM Reference Format:
Divine Njengwie Achinek*, Ibrahim Shehi Shehu, Athuman Mohamed Athuman, and Xianping Fu. 2021. Enhanced Single Shot Multiple Detection for Real-Time Object Detection in Multiple Scenes. In *The 5th International Conference on Computer Science and Application Engineering (CSAE 2021), October 19–21, 2021, Sanya, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3487075.3487082

1 INTRODUCTION

In recent years, Convolutional Neural Networks (CNNs) have shown great achievements in image processing, natural language processing, sequences, time series, and prediction. In images processing tasks such as object detection, these CNNs are generally categorized into two; that is region-based proposal networks and non-region-based proposal networks. Unlike region-based proposal networks, single-shot multiple detectors (SSD) [1], can be looked as a deep CNNs based object detector that does not resample features of pixels for bounding box hypotheses but predicts the boundary boxes and the classes directly from feature maps in one single pass. Meanwhile, it also achieves great accuracy without using the approaches that do resampling. As a result, this conventional SSD achieves significant improvement in speed and accuracy over other related architectures [2], like You Only Look Once (YOLOv1) [3] and Region Convolutional Neural Network (R-CNN) [4]. This is the motivation for our enhanced SSD proposed in this paper for real-time object detection.

Conventional SSD fundamentally have two categories of convolutional network, the base and auxiliary or additional networks. These network categories are the basis for our focus, since the conventional SSD suffers significantly from low accuracy and computational complexity with relatively low frames per second (FPS) for small objects detection. We decided to overcome this difficulty and further enhance the performance of the conventional SSD for general object detection tasks. We basically implemented an extended dilated convolution with a dynamic dilated operator and pure inception variant (PIV) on the conventional SSD architecture. This generally improves the accuracy of our proposed method with a higher FPS with the following specifics as our main contributions:

•Implementing an extended dilated convolution on the base network and auxiliary network increases the receptive field to gain more semantic data at the level of the feature layers.
•Adding PIV in the auxiliary network reduces complexity and increases the overall speed for real-time object detection.
Our proposed method further describes that in normal dilated convolution, the pixel spacing is minus the length of the dilated factor and it also applies to normal convolution where the dilated rate is one and there is no pixel spacing in the receptive field during convolution. The extended dilated convolution on the other hand uses a caterer of the dilated factor for dilated factors greater than one to maximize the amount of semantic information received from the receptive field, and then used with inception in the auxiliary layers for object detection.
CNN architectures in general have been successful in large-scale image and video detection [5], which has led to series of breakthroughs in the classification and detection of images [6] [7] [8]. Based on this, we evaluated the strength of our proposed method on a large collection of public datasets. We first evaluated on renowned large-scale image repositories, PASCAL VOC-2007 and PASCAL VOC-2012 [9]. We then further evaluated on a more complex dataset of underwater images made available by the National Natural Science Foundation of China (NSFC) [10]. This evaluation proves the versatility of our proposed method because underwater images are mostly characterized by low contrast thus makes its detection challenging.
The rest of the work is arranged as follows: section two talks about related works, section three gives a review of dilated convolution and inception then shows the architecture of our proposed method. Experimental results are discussed in section four to show what we achieved, then finally a conclusion in section five that summarized our research, proposed method and future work.

2 RELATED WORKS

Early object detection architectures before the establishment of neural networks were some profound architectures that combined robust low-level features and compositional models that are elastic to object deformation such as the Deformable Parts Model (DPM) [11][12]. This model around the year 2010 represented the state-of-the-art methods for object detection and later other conventional methods like Selective Search [13]. However, looking into the speed and accuracy of these models it is obvious their performance requires some enhancement to be applicable in real-world applications. DPM may use different compositional templates for different object classes, which are handcrafted making the model difficult to generalize. Recent models, such as CNNs and Recurrent Neural Networks (RNNs), which have not received much attention in object detection, have been used for solving key computer vision problems [14]. In the field of object detection, the recent use of deep neural networks has been demonstrated to perform better than the traditional methods like DPM and Selective Search. The recent deep learning attempt for object detection can be divided into two categories, which are region-proposal based architecture and regression-based ones.
Region-proposal based architecture includes R-CNN, SPP-net [15], fast R-CNN [16], faster R-CNN [17], PVANET [18] and R-FCN [19]. These architectures in the first stage generate object boxes and use a deep neural network for classification and location regression in the second stage. R-CNN, SPP-net and fast R-CNN use the Selective Search method to generate region proposals which is the logjam of the whole algorithm. Later faster R-CNN was introduced to overcome the logjam and abandon Selective Search architecture but instead uses Region-proposal Network (RPN) [17] to generate regional proposals. PVANET enhance VGG16 [20] on R-CNN with the inception block and the RFN was ranked the first on the VOC2012 dataset of PASCAL Challenge, which was based on R-CNN and the deep residual network. These architectures cannot be used to achieve real-time processing because of the low FPS generated by them even though they have achieved great performance. Other scholars used conventional machine learning methods to extract target features such as Haar-like and SIFT in [21], and [22] respectively. Then these extracted features are been used for detection by classifying them into their different trained categories described in [23], and [24]. Great improvements have been reached in recent years, which saw to the improved performance with lesser errors with the convolutional architectures on like these conventional algorithms. Furthermore, just by exploiting multiple layers within a convolutional network, there is a number of ways to improve its detection performance. One of the ways is to use a combined feature map, from different layers of the convolutional network to do prediction. ION in [25] uses L2 normalization [26] for object proposals from the combination of multiple layers from VGG16 and its pooling feature. Another method that is similar to this and uses a combined layer to learn object proposals and to pool features is Hypernet [27], this is because combined layers have feature maps from different layers and the pooled feature is more preferred for localization and classification as it is more descriptive but losses some salient information during pooling. A different set of methods predict objects of different scales by using multichannel dilated convolution [28] to improve the salient information received from feature layers.
The regression-based architectures on the other hand include YOLO [29] [3] and Inception SSD [30] that uses only a single network to generate bounding boxes and classification simultaneously, which makes it suitable for real-time processing on a high-performance processor. YOLO generally deals with object detection by using a smart process of dividing the image into a grid and each mesh predicts the confidence and the locations of two object boxes, the performance is limited when there are multiple objects and achieves high FPS with low accuracy. SSD is generally faster than Faster RCNN and it associates a set of default boxes with feature maps at the top of the network, which can identify multiple objects within the image with various scales and aspect ratio. YOLOv1 and SSD both suffer from low accuracy in detecting small objects in images. A multi-scale convolutional neural network (MS-CNN) [31] that helps to improve accuracy especially in small object detection applies deconvolution just like SSD on multiple layers of a convolutional network to increase the feature map resolution. To detect small objects, MS-CNN needs to use information from shallow layers with small receptive field and dense feature maps and may cause low performance.
In comparison with the region-based architectures, the regression-based architectures achieved higher FPS but the detection accuracy

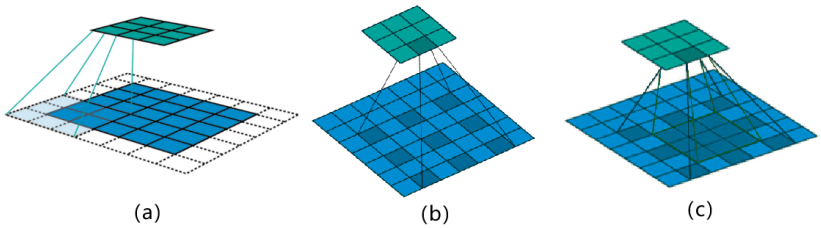

Figure 1: (a) Normal Convolution [32]; (b) Dilated Convolution [32]; (c) Extended Dilated Convolution.

is slightly low. To overcome these issues, we proposed an enhanced SSD method that achieved a relative upsurge in accuracy and speed.

3 METHODOLOGY

We introduced extended dilated convolution on both the base and auxiliary networks such that we are able to increase the semantic information in successive feature layers by increasing the receptive field on different layers during convolution. Then we describe how they are used with the pure inception blocks to enhanced detection both on Pascal VOC image datasets, and NSFC underwater image datasets.

3.1 Dilated Convolution

Dilated convolutions are convolutions that increase receptive fields on images during convolution by inflating the kernel [32]. Holes are being added to the kernel determined by an additional parameter between the kernel elements to increase the pixel spacing in the feature layers.

An increase in the receptive field on feature layers turn to increase the amount of semantic information that can be captured during convolution. But some of the detailed information will be missing caused by inflation of holes on the kernel, consequently leading to skipping of some pixels in dilated convolutions. These pixels that are being skipped carry some detailed information. Our proposed extended dilated convolution reserves this detailed information from a pixel that are been skipped during normal dilated convolution thereby not only increasing the receptive field but also increasing the semantic information that is been captured. Extended dilation is an extension of dilated convolution in which the dilated rates and its subset, for dilated rates greater than one, are used across convolutional layer blocks to get more semantic information from the increased receptive field. The receptive field can be seen as the region in the input space that a given CNN feature is looking at.

Figure 1 shows normal convolution, and dilated convolution and extended dilated convolution. We used a convolutional kernel of sizes 3x3 and we illustrate our extended dilation in Figure 1(c).

Using the same dilated rate, a dilated convolution will yield the same feature layer as an extended dilated convolution considering the same number of strides in the convolution. Since it has been established that these dilated convolutions are not contiguous, there can be seen as discrete function with a discrete filter size having a discrete filter operator. The dilated operator we used in our architecture is motivated by [28]:

Let F: Z2 $\rightarrow$ R be a discrete function. Let $\Omega_r = [-r, r]2 \cap$ Z2 and let $k : \Omega_r \rightarrow$ R be a discrete filter of size $(2r + 1)2$. The discrete convolution operator $*$ can be defined as:

$$(F * k)(P) = \sum_{s+t=P} F(s)\, k(t). \tag{1}$$

Where F is a discrete function, k the filter, P *is* the feature map, s the receptive field and t the padding.

Let's consider l be a dilation factor then $*l$ will be defined as:

$$(F *_l k)(P) = \sum_{s+lt=P} F(s)\, k(t). \tag{2}$$

We used a 3x3 convolutional kernel for extended dilation. Considering this convolutional filter as a k x k filter and a dilated rate l, the size of the dilated rate k_l x k_l can be defined as:

$$k_l = k + (k-1)(l-1) \tag{3}$$

Where k is the length or width of the convolutional filter. For extended dilation with the same parameters as above to get more information from the pixel spacing caused by dilation, we proposed a subset of the dilated rate and the size of the dilated rate can be defined as:

$$k_l = \sum_{l=l-1}^{l} 1 + l(k-1), \quad for\ l > 1 \tag{4}$$

Normal convolution can be seen in equation 1, and equation 2 shows convolution with a dilated rate l as a mathematical expression of Figure 1(b). Figure 1(c) shows that the region where semantic information is more highlighted in localization maps as it minimizes the amount of pixel skipped in dilated convolution. Its mathematical representation is shown in equation 4.

The convolutional operator itself is been modified to use filter parameters in diverse ways in our approach across different convolutional layers i.e., the dilated convolution operator can apply the same filter at different ranges using different dilation rates, l in equations 2 and 4. We demonstrate a proper implementation of the extended dilated convolution which does not involve creating dilated filters in each of the layers in our architecture by using the dilated operator. Usually, the dilated operator is a factor of two but we extended it by using different dilated rates on the base network convolutional layers that turn to learn more deep essential features during convolution.

3.2 Inceptions

Generally deep neural network has a more complex model structure unlike conventional networks, which when trained with a large amount of training data yields better performance. The performance can be improved by increasing network depth and width, however

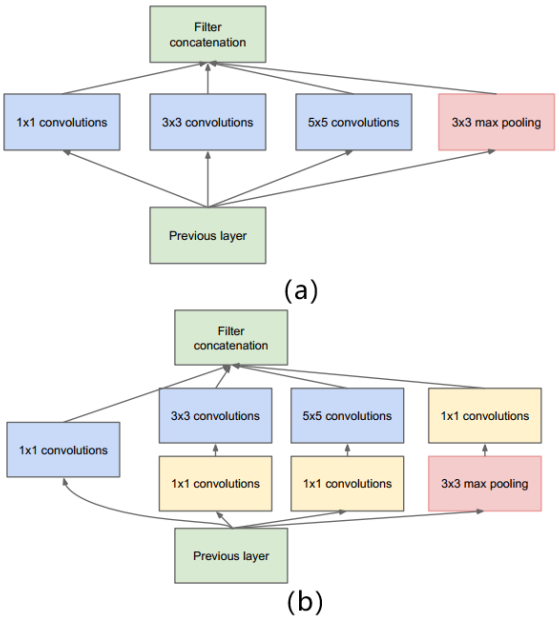

Figure 2: Inception Module [13] (a) Inception Module, Naive Version; (b) Inception Building Block with Dimensionality Reduction and the 5x5 Convolution can Further Be Reduced.

an increase in depth of a convolutional network can lead to a delay in convergence and also takes a long time to train the network. We introduced PIVs with some residual connections [33] in the auxiliary convolutional layer for faster convergence. Thus, reducing the computational complexity of the GPU or CPU, because the memory will take less time to compute ML operations since the parameters have been reduced. Then we apply a single shot as a multi-scale sliding window detector that leverages these auxiliary convolutional layers for both classification and localization. The main idea of inception is to use dense components to approximate the optimal local sparse structure [34].

Elementary foundations of inception [13] enabled us to develop a pure inception alternate with residual connections used in our architecture. These elementary pioneers of inception blocks are shown in Figure 2. A different version of the pure inception blocks was used with extended dilation. For the residual connection on the inception blocks, scaling down the residual before adding them to previous layer activation help in stabilizing training [35]. We trained the pure inception alternate without partitioning the replicas on the additional convolutional network to deal with the deepness of our proposed method, this further reduces the problem of computational complexity as the general number of parameters is reduced. A higher level of the inception block is shown in Figure 3

On the auxiliary network where the pure inception blocks are added we used a 3 x3 and 1 x1 convolutional filters without activation with the extended dilated convolutions as shown in Figure 4. For a 1-dilated convolution each element has a normal convolutional receptive field and for a 2-dilated convolution each element has a receptive field of 7 x 7 and for a 3-dilited convolution each element has a receptive field of 11 x 11 thereby causing the number

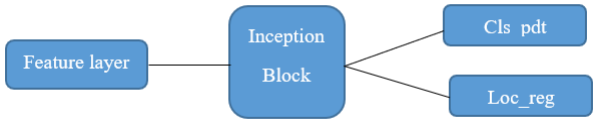

Figure 3: Pure Inception Alternate.

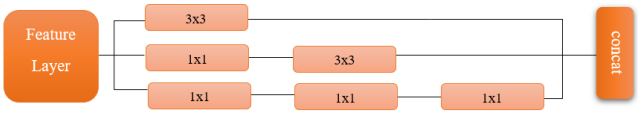

Figure 4: Pure Inception Blocks Variant Used in the Auxiliary Network.

of parameters to grow linearly while the receptive field grows exponentially giving space for more information to be learned. The receptive fields are the same for dilated convolution and extended dilated convolution but extended dilated convolution increase with essential deep features learned. Different dilation rates are used on auxiliary blocks 8, 9 and 10 (see Figure 5) with PIVs thereby causing the inception blocks to get more semantic information without necessarily increasing computational parameters and complexity. Ordinarily the auxiliary layers of conventional SSD use only a 3 x 3 filters and performance at these layers can be increased by going deeper and wider. This can lead the network to take a longer time to train and increase computational complexity. Inception blocks can speed up the training process and reduce any computational complexity. Replacing these filters with inception blocks variant will turn to increase receptive field features as well, since replacing these convolutional filters with PIV in different convolutional blocks affect the total number of contiguous feature layers [30]. We reduced the total amount of feature map to be as same as that of the conventional SSD when replaced with the PIV. Moreover, they are used at the top of the convolutional layers for object detection as Softmax is used for class predictions, and bounding box regression is used to predict the offsets for some predefined default bounding boxes from the datasets.

3.3 Combining Extended Dilated Convolution and the Pure Inception Variants

The base network has five convolutional layers with filters of sizes 3 x 3. We used equation 4 with an extended dilated factor of two for block1 and block2, and a factor of two and four on block 4 in our proposed method. The number of channels is the same both for the input image and the filters. Same as the conventional SSD, we also used MS-CNN to point out that improving the sub-network of each layer can better the accuracy in the auxiliary network where convolutional layers are been added with a decrease in size progressively. Here we introduced extended dilation alongside the PIVs convolution. Specifically, we replaced the convolutional filters in block 8, block 10 with each having a dilation rate of 2, 3, 4 and 5 with block 9, and a rate of 2 and 3 with the PIVs. Our experiments show that

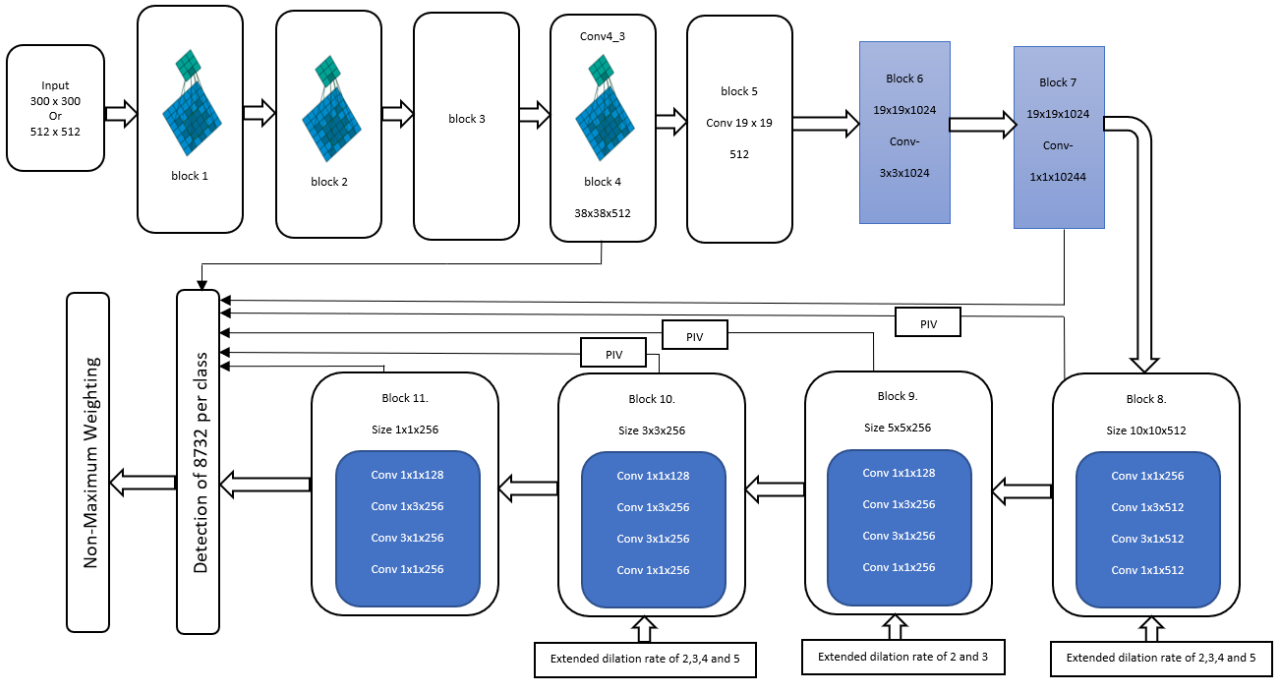

Figure 5: Architecture of Our Proposed Method with Extended Dilated Convolution and PIVs.

we achieved reasonable performance with these PIVs and dilation rates than other dilation rates. The fourth convolutional block in the base network is also used for prediction and we experimented with different extended dilated factors and conclude that dilated factors of 3 and 4 achieve reasonable performance.

The overall architecture of the proposed method is shown in Figure 5. Blocks 4, 7, 8, 9, 10, and 11 are used for detection with some of the blocks modified with a set of extended dilated convolution and inception block. Residual connections with the inception module lessen the problem of convergence in the deep neural networks we concatenate the output of block 6 with its input and used it as the input to block 7. The feature maps of these extra layers will produce the objects' location offset and confidence by small convolution operations as seen in Figure 3In the final prediction, we used maximum weighting (NMW) [36] that maximizes object detection information coming from a bounding box by considering the non-maximum results unlike conventional SSD that uses maximum suppression [37]. The NMW is based on a feed forward convolutional network and the final detection is being determined by the non-maximum suppression by using a confidence threshold of 0.01, it can filter out most boxes. For the default anchor boxes and aspect ratios, we associate a set of default bounding boxes with each grid cell or feature map cell, for multiple feature maps at the top of the network as [1]. The algorithm of the default boxes and the way they are been assign to the feature map is in a convolutional manner, so that their position relative to the feature map is fixed.

We then predicted the offsets relative to the default box shapes in the cell and in each of those boxes we predicted the per class scores that indicate the presence of a class instance and the object localization with the pure inception block at the top of the auxiliary network. The pure inception alternates on blocks 8, 9 and 10 are used for detection as shown in Figure 5 allowing different default box shapes in several feature maps gives us the capacity to competently discretize the space of possible output box shape.

4 RESEARCH RESULTS

We carried out the evaluation of our experiment on the Pascal VOC dataset, and the NSFC underwater image dataset to establish and validate the performance of our proposed method. Similar to the conventional SSD, we adopted the same training process to establish the effectiveness of our proposed method and for a fair comparison. We use a matching phase while training to match the appropriate default box with the bounding boxes of each ground truth object within an image to target the ground truth box. The default box with the highest degree of overlap with an object is responsible for predicting the object class and its location. For instance, we take each ground truth box and match it with the best overlapped default box and consider the ones whose default boxes Jaccard overlap is larger than a threshold which simplifies the learning problem. Then for the non-matched default boxes, hard negative mining is been used on them. That is, certain boxes are been selected to be negative samples based on the confidence loss so that the ratio with matched ones is 3:1 [2].

We then calculated and minimized the cost function using adaptive moment estimation, the joint localization loss and confidence loss. The smooth L1 loss [38] between the predicted box and the ground truth box is used for localization loss which combines the

Table 1: PASCAL VOC2007 Test Detection Results Compared with Our Proposed Method. The Input Resolution for Fast and Faster R-CNN Use Input Images Whose Minimum Dimension Is 600x600. Conventional SSD and Our Proposed Method Use Input Sizes of 300x300 and 512x512 with the Same Settings. We Trained on VOC2007 Trainval and VOC2012 Trainval, and Used VOC2007 Test for Testing

Method	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
Fast RCNN [16]	70.0	77.0	78.1	69.3	59.4	38.3	81.6	78.6	86.7	42.8	78.8	68.9	84.7	82.0	76.6	69.9	31.8	70.1	74.8	80.4	70.4
Faster RCNN [17]	73.2	76.5	79.0	70.9	65.5	52.1	83.1	84.7	86.4	52.0	81.9	65.7	84.8	84.6	77.5	76.7	38.8	73.6	73.9	83.0	72.6
SSD 300 [1]	74.3	75.5	80.2	72.3	66.3	47.6	83.0	84.2	86.1	54.7	78.3	73.9	84.5	85.3	82.6	76.2	48.6	73.9	76.0	83.4	74.0
SSD 512 [1]	76.8	82.4	84.7	78.4	73.8	53.2	86.2	87.5	86.0	57.8	83.1	70.2	84.9	85.2	83.9	79.7	50.3	77.9	73.9	82.5	75.3
Our Method 300	**82.0**	**90.6**	**85.5**	**81.4**	**73.1**	**59.3**	**87.1**	**84.8**	**90.7**	**68.3**	**85.3**	**77.9**	**90.6**	**90.1**	**87.6**	**82.2**	**65.7**	**77.8**	**86.5**	**90.2**	**85.3**
Our Method 512	**85.2**	**93.8**	**89.8**	**88.3**	**91.1**	**75.5**	**91.0**	**53.5**	**93.4**	**73.6**	**80.0**	**89.4**	**98.6**	**94.8**	**96.4**	**86.6**	**65.8**	**67.7**	**89.8**	**94.6**	**81.8**

Table 2: PASCAL VOC2012 Test Detection Results. The Minimum Dimension of Fast and Faster R-CNN is 600x600 and for YOLOv1 is 448 x 448. We Trained on VOC2007 Trainval and Test and VOC2012 Trainval and VOC2012 Test for Testing.

Method	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
Fast RCNN [16]	68.4	82.3	78.4	70.8	52.3	38.7	77.8	71.6	89.3	44.2	73.0	55.0	87.5	80.5	80.8	72.0	35.1	68.3	65.7	80.4	64.2
Faster RCNN [17]	70.4	84.9	79.8	74.3	53.9	49.8	77.5	75.9	88.5	45.6	77.1	55.3	86.9	81.7	80.9	79.6	40.1	72.6	60.9	81.2	61.5
YOLOv1 [3]	57.9	77.0	67.2	57.7	38.3	22.7	68.3	55.9	81.4	36.2	60.8	48.5	77.2	72.3	71.3	63.5	28.9	52.2	54.8	73.9	50.8
SSD 300 [1]	72.4	75.5	80.2	72.3	66.3	47.6	83.0	84.2	86.1	54.7	78.3	73.9	84.5	85.3	82.6	76.2	48.6	73.9	76.0	83.4	74.0
SSD 512 [1]	74.9	87.4	82.3	75.8	59.0	52.6	81.7	81.5	90.0	55.4	79.0	59.8	88.4	84.3	84.7	83.3	50.2	78.0	66.3	86.3	72.0
Our Method 300	**85.6**	**96.4**	**88.5**	**84.5**	**75.2**	**60.3**	**90.4**	**88.1**	**97.5**	**70.1**	**89.3**	**81.1**	**96.6**	**94.2**	**91.5**	**85.1**	**67.4**	**80.4**	**91.0**	**94.9**	**88.9**
Our Method 512	**87.3**	**97.0**	**92.1**	**90.7**	**94.2**	**80.4**	**94.1**	**56.1**	**98.3**	**76.3**	**81.3**	**89.6**	**98.6**	**96.2**	**97.4**	**88.5**	**62.6**	**71.0**	**91.2**	**98.2**	**89.0**

advantages of L1-loss and L2-loss, and softmax loss that is a combination of softmax loss and cross entropy loss for confidence loss. We used adaptive moment estimation as our optimizing weight function with an initial learning rate set to 0.001, momentum set to 0.9, weight decay set to 0.0005, and batch size set to 32. Our based network is pre-trained on the ILSVRC CLS-LOC dataset [39] [40]. Training also includes data augmentation to make the model more robust to various input object sizes and shapes. We used batch normalization for the output of previous layers and input to the next layer for smooth training to eliminate the internal covariate shift problem that is to maintain a constant distribution of the inputs during training as the inputs of different neurons in the network change. We also used randomization of the original input image size and random photometric distortion as well as random flipping of the cropped patch.

4.1 Experimental Results

Our experiments are based on the enhanced VGG backed end and the auxiliary network, in order to train the new network from pre-trained data, we had to exclude some layers and later render them trainable during the training process.

Evaluation Metrics. The commonly used methods to compare detection and recognition performance in real-time are the mean average precision (mAP), and FPS. We have used both in our evaluation, the mAP denotes the average value of all category average precision (AP), meanwhile the AP over the interval of recall = 0 to 1 computes the average value of the precision. Where precision and recall are calculated from the detections true positive (TP), false positive (FP), and false negative (FN) values as:

$$AP = \int_0^1 p\,(\mathrm{r})\,\mathrm{d}r \tag{5}$$

Where p and r stand for precision and recall respectively

$$precision = \frac{TP}{TP + FP}\ ,\ \ \text{recall} = \frac{TP}{TP + FN} \tag{6}$$

The method we used for calculating FPS was the same approach used in the conventional SSD and other different models which was setting the batch size to 1 and 8, took the predicted time of all images used and the sum of the feature extraction time, and divide it by the total number of the images to calculate the detection time of a single image.

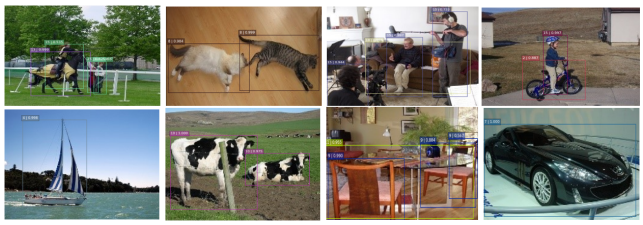

Figure 6: Samples of PASCAL VOC Images Showing Detection of Different Image Categories by Our Method.

4.2 Pascal VOC2007 and 2012

4.2.1 Pascal VOC2007. Since the Pascal VOC dataset is good for deep learning and other machine learning architectures and was also used by the conventional SSD, we used it for our training. Pascal VOC 2007 has 9,963 train, validation and test images containing 24,640 annotated objects and 4952 test images (VOC2007 test) over 20 categories [20]. We used the same approach of fine-tuning on the pre-trained VGG16 network and using this dataset we compare the detection results of our proposed methods with state-of-the-art architectures like SSD, Fast R-CNN [27] and Faster R-CNN [20] to demonstrate the performance of our proposed method in Table 1We start training our model with 10^{-3} learning rate for the first 90k iterations and then continue training for 60k iterations with a learning rate of 10^{-4} for the first stage then 20k iteration with 10^{-4}, and another 20k iterations with 10^{-5} learning rate.
Our proposed method achieved an improved performance over other state-of-the-art architectures with mAP of 82.0 for 300x300 input size, and mAP of 85.2 for 512x512 input size.

4.2.2 Pascal VOC2012. Pascal VOC 2012 has 11,530 train, validation and test images containing 27,450 ROI annotated objects and 4952 test images over 20 categories [20] some samples are shown in Figure 6. We used the same approach of fine-tuning on the pre-trained VGG16 network and using this dataset we compare detection results of our proposed method with state-of-the-art architectures shown in Table 2We start training our model with 10^{-3} learning rate for the first 90k iterations and then continue training for 30k iterations with a learning rate of 10^{-4} for the first stage. Then we fine-tune the entire network with a learning rate of 10^{-3} for the first 80k iteration and continue training for 20k iteration with 10^{-4}, and another 20k iteration with 10^{-5} learning rate.
Our proposed method achieved an improved performance over other state-of-the-art architectures with mAP of 85.6 for 300x300 input size, and mAP of 87.3 for 512x512 input size.

4.3 NSFC Underwater Image Dataset

The underwater image dataset we used in this experiment was extracted at our lab from underwater videos provided by the NSFC. Basically, frames were captured from the underwater video using a MATLAB code for frame extraction from video. Samples of these images are shown in Figure 7
There are 18,164 images in total containing four different categories of seafood that is sea cucumber, sea urchin, scallop and starfish. Our test set was one-tenth of this dataset. In this experiment we start training our model with 10^{-4} learning rate for the first 30k iterations and then continue training for 80k iterations with a learning rate of 10^{-4} for the first stage. Then we fine-tune the entire network with a learning rate of 10^{-3} for the first 20k iteration and continue training for 20k iteration with 10^{-4}, and another 20k iteration with 10^{-5} learning rate.
Table 3 shows that we achieved an improved performance over other state-of-the-art architectures with mAP 79.1 for 300x300 input size, and mAP of 80.6 for 512x512 input size.

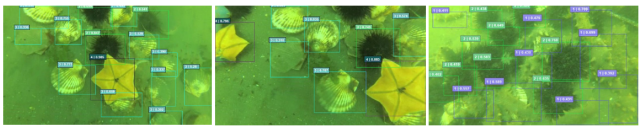

Figure 7: Samples of the Underwater Image Contained in the USFC Dataset with Our Detection Method Showing Four Different Categories of Underwater Seafood [10].

Table 3: Underwater Image Dataset Detection Results Compare to Our Model

Method	mAP	sea cucumber	sea urchin	scallop	starfish
Fast RCNN [16]	63.0	70.1	63.2	60.3	58.5
Faster RCNN [17]	68.4	75.6	68.3	60.9	68.9
SSD 300 [1]	72.8	80.4	70.7	64.9	75.6
SSD 512 [1]	74.6	83.7	72.9	65.3	76.8
Our Method 300	**79.1**	83.4	**77.3**	**74.5**	**81.3**
Our Method 512	**80.6**	**84.7**	**79.5**	**75.3**	**82.8**

4.4 Inference

We used NMW [36] as mention earlier for inference. It maximizes object detection information coming from a bounding box with a Jaccard of 0.45 overlap per class by considering the non-maximum results using a NVIDIA Tesla T4, cnDNN v 7.0 and CUDA v 9.0. Real-time performance results are shown in Table 4
Our results show that higher input resolution achieved higher accuracy but with a decrease in FPS.

5 CONCLUSIONS

This paper investigates the strategies to improve the conventional SSD for real-time object detection. We basically exploited extended dilated convolution in the base and auxiliary network of our proposed architecture and later introduced inception along with extended dilated convolution on the auxiliary network. Our proposed method relatively increased the amount of semantic information captured and learned during convolution thereby minimizing the computational complexity. The parameters as inception building blocks increase the relative speed of the network. We achieved

Table 4: Result on FPS for Real-Time Detection with Pascal VOC2007 Test

Method	Input resolution	Batch size	Number of boxes	mAP	FPS
Faster R-CNN [17]	~1000 x 600	1	~ 6000	73.2	7
YOLOv1 [3]	448 x 448	1	98	66.4	22
Fast YOLO [3]	448 x 448	1	98	52.7	**155**
YOLOv2 [29]	288 x 288	1	1445	69.0	**91**
YOLOv2 [41]	544x544	1	1445	78.6	40
SSD 300 [1]	300 x 300	1	8732	74.3	48
SSD 300 [1]	300 x 300	8	8732	74.3	60
SSD 512 [1]	512 x 512	1	24564	76.8	20
SSD 512 [1]	512 x 512	8	24564	76.8	23
YOLOv3 (Darknet-53) [41]	320 x 320	1	-	79.2	51
YOLOv3 (Darknet-53) [41]	416 x 416	1	-	82.0	41
YOLOv4 (CSPDarknet-53) [42]	416 x 416	1	-	85.1	44
Our Method 300	**300 x 300**	**1**	**8732**	**82.0**	**49**
Our Method 300	**300 x 300**	**8**	**8732**	**82.0**	**63**
Our Method 512	**512 x 512**	**1**	**24564**	**85.2**	**30**
Our Method 512	**512 x 512**	**8**	**24564**	**85.2**	**34**

improved performance over the conventional SSD and other state-of-the-art architectures. Experimental results show that our proposed method yields relatively better performance on PASCAL VOC 2007 and 2012. Besides that, it also achieved improved performance with the NSFC underwater images dataset, which demonstrates the versatility of our proposed method.

Future research on this will involve the detection of objects in video since video object detection is affected by the fuzziness of the objects in the video that losses focus especially in fast motion and going deeper with small object detection in videos.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China Grant 61802043, by the Liaoning Revitalization Talents Program Grant XLYC1908007, by the Foundation of Liaoning Key Research and Development Program Grant 201801728, by the Fundamental Research Funds for the Central Universities Grant 3132016352 and Grant 3132020215, by the Dalian Science and Technology Innovation Fund 2018J12GX037 and 2019J11CY001.

REFERENCES

[1] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). Ssd: Single shot multibox detector. In European conference on computer vision (pp. 21-37). Springer, Cham.

[2] SermanetP, E., & FergusR, L. O. (2018). Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229.

[3] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

[4] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[5] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.

[6] Chen, X., Fang, H., Lin, T. Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325.

[7] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4), 541-551.

[8] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.

[9] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. International journal of computer vision, 88(2), 303-338.

[10] Underwater Image Dataset, National Natural Science Foundation of China (NSFC). Online, retrieved from http://www.cnurpc.org/index.html

[11] Felzenszwalb, P., McAllester, D., & Ramanan, D. (2008, June). A discriminatively trained, multiscale, deformable part model. In 2008 IEEE conference on computer vision and pattern recognition (pp. 1-8). Ieee.

[12] Girshick, R. B., Felzenszwalb, P. F., & McAllester, D. (2012). Discriminatively trained deformable part models, release 5.

[13] Uijlings, J. R., Van De Sande, K. E., Gevers, T., & Smeulders, A. W. (2013). Selective search for object recognition. International journal of computer vision, 104(2), 154-171.

[14] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

[15] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 37(9), 1904-1916.

[16] Girshick, R. (2015). Fast R-CNN. computer science. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV) (pp. 1440-1448).

[17] Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence, 39(6), 1137-1149.

[18] Kim, K. H., Hong, S., Roh, B., Cheon, Y., & Park, M. (2016). Pvanet: Deep but lightweight neural networks for real-time object detection. arXiv preprint arXiv: 1608.08021.

[19] Dai, J., Li, Y., He, K., & Sun, J. (2016). R-fcn: Object detection via region-based fully convolutional networks. In Advances in neural information processing systems (pp. 379-387).

[20] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[21] Mita, T., Kaneko, T., & Hori, O. (2005, October). Joint haar-like features for face detection. In Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1 (Vol. 2, pp. 1619-1626). IEEE.

[22] Ma, X., & Grimson, W. E. L. (2005, October). Edge-based rich representation for vehicle classification. In Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1 (Vol. 2, pp. 1185-1192). IEEE.

[23] Kazemi, F. M., Samadi, S., Poorreza, H. R., & Akbarzadeh-T, M. R. (2007, April). Vehicle recognition using curvelet transform and SVM. In Fourth International Conference on Information Technology (ITNG'07) (pp. 516-521). IEEE.

[24] Lan, R., Lu, H., Zhou, Y., Liu, Z., & Luo, X. (2020). An LBP encoding scheme jointly using quaternionic representation and angular information. Neural Computing and Applications, 32(9), 4317-4323.

[25] Bell, S., Zitnick, C. L., Bala, K., & Girshick, R. (2016). Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp.

2874-2883).
[26] Liu, W., Rabinovich, A., & Berg, A. C. (2015). Parsenet: Looking wider to see better. arXiv preprint arXiv:1506.04579.
[27] Kong, T., Yao, A., Chen, Y., & Sun, F. (2016). Hypernet: Towards accurate region proposal generation and joint object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 845-853).
[28] Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122.
[29] Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7263-7271).
[30] Ning, C., Zhou, H., Song, Y., & Tang, J. (2017, July). Inception single shot multibox detector for object detection. In 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW) (pp. 549-554). IEEE.
[31] Cai, Z., Fan, Q., Feris, R. S., & Vasconcelos, N. (2016, October). A unified multi-scale deep convolutional neural network for fast object detection. In European conference on computer vision (pp. 354-370). Springer, Cham.
[32] Louwill, (2021). Deep Learning 100 Q -18: how to calculate CNN's receptive field? Online retrieved from https://www.programmersought.com/article/84213363050/
[33] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
[34] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).
[35] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017, February). Inception-v4, inception-resnet and the impact of residual connections on learning. In Thirty-first AAAI conference on artificial intelligence.
[36] Fu, C. Y., Liu, W., Ranga, A., Tyagi, A., & Berg, A. C. (2017). Dssd: Deconvolutional single shot detector. arXiv preprint arXiv:1701.06659.
[37] Neubeck, A., & Van Gool, L. (2006, August). Efficient non-maximum suppression. In 18th International Conference on Pattern Recognition (ICPR'06) (Vol. 3, pp. 850-855). IEEE.
[38] Chen, Y., Li, W., Sakaridis, C., Dai, D., & Van Gool, L. (2018). Domain adaptive faster r-cnn for object detection in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3339-3348).
[39] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge IJCV. arXiv preprint arXiv:1409.0575.
[40] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062.
[41] Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.
[42] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.