International Journal of Computer Vision and Signal Processing, 1(1), 1-8(2020) ORIGINAL ARTICLE

Multi-Level Feature and Context Pyramid Network for Object Detection

Xia Wang, Yingdong Ma*

College of Computer Science,Inner Mongolia University
Hohhot, China.

Abstract

Robust object detection require fine details to represent object structure and high-level semantic knowledge extracted from deep feature maps. Besides, contextual information is also important for exact location of multiple scale objects. However, it is difficult to meet these demands simultaneously in the top-down CNN structure. In this work, we present the Multi-Level Feature and Context Pyramid Network (MLFCP Net) to tackle this problem. The proposed MLFCP Net consist of two main modules. To utilize advantages of multiple level features, the Multi-level Feature Fusion (MFF) module combines different layer feature maps to form enhanced multi-level features. The Context Pooling Aggregation module combines local and global context features to further improve detection accuracy. Our method achieves 84.9% mAP on PASCAL VOC2007 test at 16.7FPS with 320×320 input and 42.5% AP on MS COCO. Experimental results demonstrate effectiveness of the proposed feature fusion method and the context aggregation scheme.

Contribution of the Paper: we propose a novel Multi-Level Feature and Context Pyramid Network (MLFCP Net).

Keywords: multi-scale feature fusion, contextual information, object detection

IJCVSP

ISSN: 2186-1390 (Online)
http://cennser.org/IJCVSP

Article History:
Received: 27 December 2019
Revised: 25 May 2020
Accepted: 21 June 2020
Published Online: 25 June 2020

1. INTRODUCTION

Multi-scale object detection is a difficult and fundamental task in the field of computer vision. Due to the fast development of convolutional neural networks (CNNs), significant progress has been made towards improving object detection performance. Some works on object detection [1, 2] and semantic segmentation [3, 4] have shown that prediction from high-level feature maps only may lead to lower accuracy. While deep features provide rich semantic information, the top-down CNN structure suffers fine-level object information loss. The problem becomes more serious in practice due to scale variation across object instances.

The feature pyramid network has been widely used to utilize advantages of multiple level features [5]. Feature pyramid-based object detectors integrate different layer features using the lateral connections between bottom-up and top-down layers. The reverse connection introduced in [6] combines upsampled high-level feature maps with low-level features to improve multi-scale object detection performance. In the proposed RON network, large-scale objects are detected using deep layer feature maps and shallow layer feature maps are mainly used to locate small objects. The RefineDet with two inter-connected modules uses variable-size anchors to determine object positions and combines low-level information with high-level semantic features [7]. The MDSSD [8] and DFP [9] also utilize high-level semantic features with low-level object information. These methods indicate that the lower-level and upper-level features are complementary and their combination is necessary for object detection.

Recently, some object detection and semantic segmentation approaches explore the usage of context information for improving detection and segmentation performance [2, 10, 11, 12]. The DSSD adopts deconvolutional operators to construct an encoder-decoder structure for extracting additional context information [12]. The dilated convolution is also widely used to incorporate multi-scale context information [13]. The contrast prior and fluid pyramid integration method introduce a contrast enhanced network to improve depth information [11]. The enhanced depth maps are then used as a global context clue for salient ob-

*Corresponding author
Email addresses: 31809016@mail.imu.edu.cn (Xia Wang), csmyd@imu.edu.cn (Yingdong Ma)

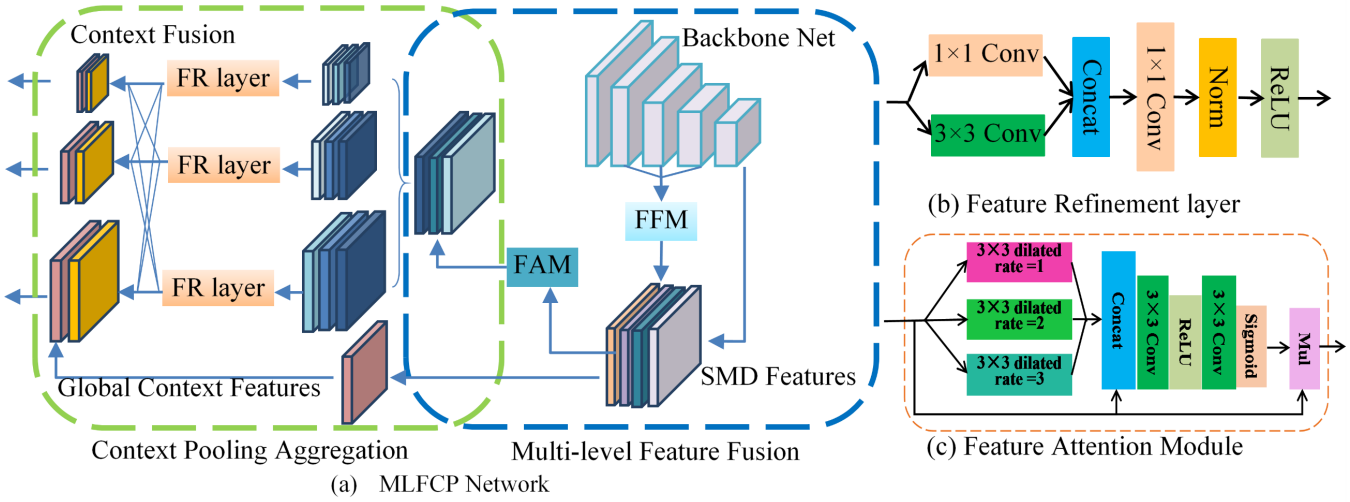

Figure 1: An overview of the proposed network. (a)Overview of the MLFCP Network. FFM: feature fusion module. FAM: feature attention module. FR layer: feature refinement layer. SMD: Shallow, medium and depth features. (b) Feature Refinement Layer (FR layer). (c)Feature Attention Module (FAM). (three pyramid level is shown in context pooling aggregation module for simplicity; actual pyramid level is four)

ject detection. In [14], context information outside regions of interesting is integrated by using the spatial recurrent neural network. These works illustrate that utilizing context information makes the final prediction more reliable. Although these methods achieve better object detection with the help of context information, the major issue of these models is lack of suitable strategy to integrate global and local context information with the top-down CNN structure. In general, deep layer features correspond to global context clues, which are important for the large object recognition. On the other hand, local context information can be extracted from low level features, which facilitates exact location of different size objects.

Motivated by these works, we propose a new network structure, the Multi-level Feature and Context Pyramid Network to alleviate these problems. The MLFCP Net consists of two main stages. In the first stage, the Multi-level Feature Fusion (MFF) module integrates different level feature maps to enhance multiple level features (the SMD features). To better utilize the local and global context information, a Context Pooling Aggregation (CPA) module is proposed in the second stage. The CPA module collects multi-scale context information using pyramid pooling. The context information is further combined with SMD features to improve object detection performance. The overall framework is shown in Figure 1(a). The main contributions of this paper are summarized as follows:

- We propose the multi-level feature and context pyramid network that aggregates different level features to improve detection performance. The new model enhances network representation ability for both fine-level features and object-level semantic features.
- A multi-scale context pooling aggregation module is developed to tackle the problem of missing context information in the top-down CNN structure. The CPA module combines local and global context features under different pyramid scales.

2. RELATED WORKS

CNNs based object detection. Because of the powerful feature representation ability, deep convolutional neural networks have been successfully applied to various computer vision applications, such as image classification [15] and semantic segmentation [3, 4]. In the past decade, CNNs have also been widely adopted in object detection. To utilize different level features, some detectors make predictions on multi-scale features [10, 11, 14, 16]. The Trident Net generates scale-specific features through multiple parallel branches [10]. The dilated convolution is employed in the multi-branch architecture with different dilation rates to adapt the receptive fields for multiple scale objects. The single-shot detector introduced in [17] adopts the feature pyramid to yield reliable prediction. The SSD uses multiple scale features to predict class scores and bounding boxes regression. In [18], the RetinaNet also uses a feature pyramid network as backbone model. A new focal loss is introduced to address the foreground-background class imbalance problem. In [19], multi-resolution prediction maps are densely connected to achieve deeply supervised object detectors.

Multiple layer feature fusion. Some approaches combine multi-layer feature maps to make better use of different level features [8, 9, 19, 20]. The multi-scale deconvolutional single shot detector integrates high-level features with low-level details using a deconvolution fusion block. The method adds semantic information to the low-level details to generate features with strong representational power for small object instances [8]. The NAS-FPN, which is discovered by using a neural architecture search algorithm,

consists of a combination of top-down and bottom-up connections to fuse different scale features so that it has high resolution and rich semantic information [20]. In [14], the ION concatenates low-level and high-level features from different layers to generate better feature maps for object prediction. In the DSSD, several deconvolutional layers are added to the top of SSD network to up-sample feature maps [12]. These feature maps are then combined with different scale feature maps to improve detection accuracy. The PANet [21] adds an extra bottom-up path augmentation on FPN structure to enhance feature pyramid. The new framework strengthens network localization capability by propagating spatial features from the low level to top level feature maps.

Object detection with context Information. Object detection not only relies on fine-level features and object-level semantic clues, but also requires context information to make reliable prediction. The dilated convolution is a common method to generate context information. In [13], dilated convolution is adopted to aggregate multi-scale context information. Methods proposed in [16, 21, 22] extract context information with different resolutions from multiple regions. For example, the Global pooling is used in PSP Net [16] to generate global context clues for semantic segmentation and the DFANet [22] combines features of different stages in a cross-level feature aggregation structure to incorporate multi-level context. The DSSD network appends deconvolutional layers to SSD model to generate additional context information [12]. The enriched feature guided refinement network [23] introduce a feature enrichment scheme to produce multi-scale context features to implement object detection.

3. MULTI-LEVEL FEATURE AND CONTEXT PYRAMID NETWORK

We introduce a multi-level feature and context pyramid network for multi-scale object detection. The proposed MLFCP network has two main parts, the multi-level feature fusion module and the context pooling aggregation module. The feature fusion module provides combined features by aggregating fine-level features and object-level semantic features. The context pooling aggregation module integrates different scales context information to facilitate exact location of objects in complex scene. The method introduces an efficient scheme to integrate multiple level semantic and context information with the top-down CNN structure. It is worth to note that not only multi-scale object detection, other computer vision applications, such as semantic segmentation, can also benefit from the proposed architecture.

3.1. Multi-level Feature Fusion Module

Most widely used CNNs provide feature maps with large receptive field for object-level information. However, both fine-level spatial information and object-level semantic information are necessary for object detection applications. Recently, lots of object detection approaches have been developed to fuse multiple level features in the top-down structure. Some detectors encode spatial information with dilated convolution method [2, 10]. The U-shape structure is also widely utilized, in which deep features are combined with features of shallow layers to increase spatial details [24].

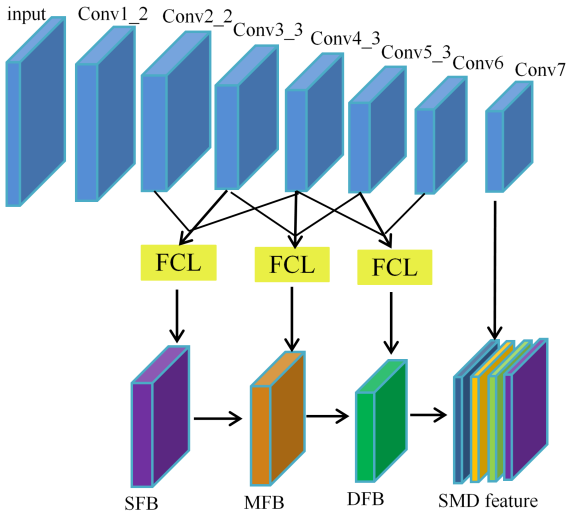

Figure 2: Overview of the multi-level feature fusion module. FCL: feature combination layer. SFB: shallow feature block, MFB: medium feature block, DFB: deep feature block.

Motivated by these works, we propose the multi-level feature fusion module to enhance network capability for both fine-level features and object-level semantic features. Figure 2 illustrates the MFF module structure. In the feature fusion module, we apply a feature combination layer to integrate features of multiple layers. In each feature combination layer, shallow features are down-sampled to match the size of deep features. The output of MFF modules consists of shallow feature block, medium feature block and deep feature block.

The structure of feature combination layer is shown in Figure 3. Instead of feature summation, we adopt concatenation to combine features of different layers. As input feature maps of the feature combination layer have different resolutions, before feature concatenation, larger size feature maps are down-sampled by using the average pooling to get uniform size feature maps. The shallow feature block is computed as:

$$DS = Concat(DS_4(conv2\_2), DS_2(conv3\_3), conv4\_3) \quad (1)$$

$$F_s = ReLU(BN(Conv(DS))) \quad (2)$$

where F_s is the shallow feature block, DS_4 and DS_2 are 4× down-sample and 2× down-sample. We use average pooling with stride 4 and stride 2 to implement down-sampling for two larger size input feature maps. These resized feature maps are then concatenated and followed by a 3×3 convolution. The batch normalization and the rectified

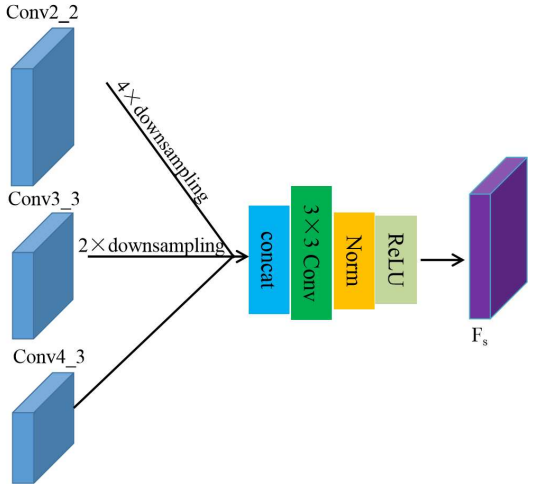

Figure 3: Example of the feature combination layer. Fs: shallow feature block.

linear unit (ReLU) [25] are utilized for normalization and activation. The medium feature block and deep feature block are computed by the same method. The dimensions of shallow, medium, and deep feature blocks are 512, 512 and 1024, respectively. Finally, all feature blocks and the conv7 layer feature maps are concatenated to generate the multi-level feature fusion module features (the SMD features).

3.2. Feature Attention Module

The attention mechanism plays an important role in the process of human perception as we have the ability to pay more attention to salient regions in a cluttered visual scene. Recently, the attention mechanism has been successfully applied in many computer vision tasks [15, 25]. In [25], Hu proposed a Squeeze and Excitation structure to adjust output response by modeling the relationship between channel features.

We utilize a Feature Attention Module to refine the SMD features. The dilated convolution is adopted in the FA module to expand receptive field and gather context information. Specifically, three dilated convolutions with different dilation rates are used to expand the receptive field. Then, dilated features are concatenated and followed by a convolution layer with batch normalization and the rectified linear unit. Finally, the sigmoid function is used to compute attention vector to guide the feature weight learning. Figure 1(c) shows the structure of the feature attention module.

3.3. Context Pooling Aggregation Module

Some recent visual recognition works explored the capability of context information in object detection and semantic segmentation [13, 14]. These works have shown that utilizing scene context clues make the final prediction more reliable. To further improve detection performance, we present a context pooling aggregation module to facilitate exact location of objects in complex scene (For convenience, we refer the module as context pyramid module for short). As combination of multiple-scale feature maps is a common approach for collecting context information, we adopt the spatial pyramid pooling with various sizes of pooling sub-regions to construct different pyramid scales, as shown in Figure 1(a).

The context pyramid module fuses feature maps from four different pyramid layers. Let the weighted SMD features have D channels. F=$\{f_1, f_2, f_3\ f_4\}$ represents feature maps of different pyramid layers generated by the spatial pyramid pooling and each of them has D channels. After a feature refinement layer, feature maps in each pyramid layer has 1/2D channels. We obtain down-sampled feature maps in different pyramid layers by successively decreasing the spatial size by half.

A feature refinement layer is applied in the context pyramid module to enhance feature maps of different pyramid layer and reduce the channel number. Figure 1(b) shows the details of feature refinement layer. We adopt the inception layers to refine features in each pyramid layer. In the inception layer, a 3×3 convolution and a 1×1 convolution are applied in parallel to extract context features for each pyramid scale. After feature concatenation, we use a 1×1 convolution to reduce feature map channels to 1/2D in each pyramid layer.

To combine local and global context information, we propose the context fusion method as shown in Fig.1(a). Specifically, let feature maps in four pyramid scales have spatial size of W×H×D, 1/2W×1/2H×D, 1/4W×1/4H×D, and 1/8W×1/8H×D, where D is the number of feature channel. After the feature refinement layer, we use feature maps in the same scale as the primary features and feature maps in other three scales as the supplementary features to form a four-level feature pyramid. Then, features in each level use a 1×1 convolution to reduce the number of channels to 1/2D. To match the sizes of different level feature maps, we up-sample the smaller size feature maps via bilinear interpolation and down-sample larger size feature maps by using the average pooling.

3.4. Details of the MLFCP Network

We use different aspect ratios 1,2,3,1/3,1/2 for the default boxes. Thus, at each pyramid layer, we use anchors at five aspect ratios to cover objects of different sizes. In the prediction stage, we use a 3×3 convolution layer to refine feature maps firstly. Then, a 3×3 convolution layer is applied to predict locations of objects and their class labels. For the box regression sub-network, a 3×3 Convolution layer with 4×A filters is applied to each level to calculate the relative offset between the anchor and the predicted bounding box, where A is the number of anchors per location of the feature maps. Another 3×3 Convolution layer with (K+1) × A filters is applied to predict the probability of an object being present at each spatial position for each of the A anchors and K object classes.

4. EXPERIMENT

We train and evaluate the proposed MLFCP network on two datasets: the PASCAL VOC 2007 [26] and the MS COCO [27].

4.1. PASCAL VOC 2007

We train the MLFCP network on PASCAL VOC 2007 and PASCAL VOC 2012 trainval (16551 images), and test on VOC 2007 test set (4952 images). The network is pre-trained on the ILSVRC dataset [28]. We train MLFCP network for 120k iterations with an initial learning rate of 10^{-3}. The learning rate changes to 10^{-4} at 80k iterations and 10^{-5} at 100k iterations. We use the SGD with a mini-batch size of 10 and 8 for MLFCP network with 320×320 input images and 512×512 input images, respectively. We use a momentum of 0.9 and 0.0005 weight decay. The evaluation metric of this work is the mean average precision (mAP).

Ablation study of different dilation rates. We conduct following experiments to verify the effectiveness of different dilation rates in the feature attention module. As listed in Table 1, we obtain the best performance when setting the dilation rates to 1, 2, and 3.

Table 1: Ablation study of different dilation rates. D3 and D5: 3×3 and 5×5 dilated convolutions. (xxx): dilation rates used in dilated convolutions.

Method	mAP(%)
VGG16+D5(123)	81.1
VGG16+D5(234)	80.7
VGG16+D5(134)	80.9
VGG16+D3(123)	81.9
VGG16+D3(234)	81.5
VGG16+D3(134)	81.7

Ablation study of the feature fusion module and the feature attention module. Experimental results with and without the feature fusion module and the feature attention module are shown in Table 2. The context fusion module is removed in all these experiments for better performance comparison. Comparing to the VGG-16 baseline network, utilizing of the feature fusion module increases detection performance by 3.5%. In the next experiment, the Output feature maps of the baseline model are connected to the feature attention module directly. This experiment yields 70.6% mAP with 0.6% accuracy drop compared with the FFM module. It demonstrates the important of the feature fusion module as some spatial details are lost in the deep feature maps and only part of them can be recovered in the post-processing steps. We obtain the highest performance (72.6%) with the feature fusion module and the feature attention module. As we can see from the last experiment, with the help of feature fusion module, the FAM module increases detection accuracy by 1.4%.

Ablation study of the context pooling aggregation module

Table 2: Ablation study of the feature fusion module and the feature attention module.

VGG	FFM	FAM	Parameters(M)	mAP(%)
√			31	67.7
√	√		61	71.2
√		√	50	70.6
√	√	√	65	72.6

The pyramid levels: Experimental results of using different number of pyramid levels are shown in Table 3. The proposed MLFCP network achieves the best result of 81.9% mAP with four levels feature pyramid. In the case of five-layer feature pyramid with 1, 2, 4, 8, 16 pooling rates, we obtain 81.0% mAP.

Table 3: Experiment result of various pyramid levels.

	Pyramid levels		
	3	4	5
mAP(%)	80.3	81.9	81.0

Feature refinement layer and context fusion method: Table 4 illustrates results of experiments with and without the feature refinement layer and the context fusion method. To verify the effectiveness of various pooling method, the avg-pooling and max-pooling are used separately in the context fusion method. In the first experiment (the second row of Table 4), output feature maps of the feature refinement layer are used directly in the prediction stage. As listed in Table 4, without the context pooling aggregation module, the multi-level feature fusion module yields 72.6% mAP. Detection performance is improved to 74.2% by using the pyramid pooling and the feature refinement layer. The proposed context fusion method further improves performance to 78.9% (with max-pooling) and 79.7% (with avg-pooling). In these experiments, we observe that the avg-pooling has better performance than the max-pooling. In the last two experiments, utilizing of the feature refinement layer and the context fusion method with avg-pooling achieves the best performance of 81.9%, about 9.3% improvement compared to the first experiment. These experiments indicate that multiple scale context information is necessary for object detection.

Experiments on PASCAL VOC 2007. Table 5 shows experimental results of the proposed MLFCP network and state-of-the-art on PASCAL VOC 2007 dataset. The MLFCP network achieves 81.9% mAP when using 320×320 input images. Detection performance is further increased to 83.2% with 512×512 input images. The proposed method outperforms other detectors using input im-

Table 4: Performance comparisons of the feature refinement lever (FR) and the Context Fusion Method (CFM). Avg:Avg-Pooling. Max:Max-Pooling.

SMD features	CFM+ Avg	CFM+ Max	FR	Parameter (M)	mAP(%)
√				73	72.6
√			√	111	74.2
√		√		109	78.9
√	√			109	79.7
√		√	√	113	80.5
√	√		√	113	81.9

Table 5: Experiment results on PASCAL VOC 2007 with PASCAL VOC 2007 trainval and PASCAL VOC 2012 trainval as the training set,PASCAL VOC 2007 test as the test set.

Method	Backbone	GPU /Number	Input size	Speed (FPS)	mAP 2007(%)
ION [14]	VGG-16	K40 /-	∼1000×600	1.3	79.2
DSOD [19]	DS/64-192-48-1	Tian X/1	300×300	17.4	77.7
RON320 [6]	VGG-16	Tian X/1	320×320	15	71.7
SSD300 [17]	VGG-16	Tian X /1	300×300	46	77.5
SSD512 [17]	VGG-16	Tian X /1	512×512	19	79.8
DSSD321 [12]	ResNet-101	Tian X /1	321×321	9.5	78.6
DSSD513 [12]	ResNet-101	Tian X /1	513×513	5.5	81.5
RefineDet320 [7]	VGG-16	Tian X /1	320×320	40.3	80.0
RefineDet512 [7]	VGG-16	Tian X /1	512×512	24.1	81.8
RefineDet320+ [7]	VGG-16	Tian X /1	-	-	83.1
RefineDet512+ [7]	VGG-16	Tian X /1	-	-	83.8
MDSSD300 [8]	VGG-16	1080Ti /1	300×300	38.5	78.6
MDSSD512 [8]	VGG-16	1080Ti /1	512×512	17.3	80.3
EFGRNet320 [23]	VGG-16	Tian XP/1	320×320	-	81.4
EFGRNet512 [23]	VGG-16	Tian XP/1	512×512	-	82.7
PFPNet-R320 [29]	VGG-16	Tian X /1	320×320	33	80.7
PFPNet-R512 [29]	VGG-16	Tian X /1	512×512	24	82.3
PFPNet-R320+ [29]	VGG-16	Tian X /1	-	-	83.5
PFPNet-R512+ [29]	VGG-16	Tian X /1	-	-	84.1
MLFCP320	VGG-16	1080Ti /1	320×320	16.7	81.9
MLFCP512	VGG-16	1080Ti /1	512×512	10.1	83.2
MLFCP320+	VGG-16	1080Ti /1	-	-	83.9
MLFCP512+	VGG-16	1080Ti /1	-	-	**84.9**

ages of the same size. To reduce the impact of various input sizes for a fair comparison, we conduct multi-scale testing with different sizes input images. As shown in Table 5, the MLFCP320+ and MLFCP512+ achieve 83.9% mAP and 84.9% mAP, respectively. The proposed method exhibits the best mAP among other detectors with the same multi-scale input images, such as PFPNet-R512+ and RefineDet512+. The PFPNet utilizes pyramid pooling to extract context information from the output features of base network. Different from the PFPNet, in our method, multiple level features are integrated into shallow, medium and deep features (the SMD features) firstly. The SMD features are further enhanced by using the attention module. Finally, context information is extracted from the enhanced multi-level SMD features.

4.2. MS COCO

To further validate our model, we conduct experiments using the MS COCO dataset with 320×320 images (MLFCP320) and 512×512 images (MLFCP512). we use the trainval (123187 images) for training and evaluate the results on the standard test-dev2015 split (20288 images). The performance evaluation metric for the COCO dataset is slightly different from that of the VOC dataset. The average precision over different IoU thresholds from 0.5 to 0.95 ($AP_{50:95}$) is adopted to report overall performance. The APs with IoU thresholds of 0.5 and 0.75 are denoted as AP_{50} and AP_{75}, respectively. The batch size is set to 10 for 320×320 input images and 8 for 512×512 input images. We train the model with an initial learning rate of 10^{-3} for the first 160k iterations, and then decreasing it to 10^{-4} and 10^{-5} for the next 120k and 40k iterations. The total number of training iterations is 320k. Other settings are the same as PASCAL VOC dataset.

As shown in Table 6, MLFCP320 has the APs of 33.0%, which outperforms most VGG-16-based detectors using input images with 320×320 pixels. For the input size of 512×512, Our MLFCP512 model achieves 37.1% accuracy, outperforms other hourglass models, such as RetinaNet500, DSSD513, and RefineDet512. The proposed MLFCP512 shows the result similar to M2Det512, which uses multiple U-shape networks to collect multi-level features. We also employed the multi-scale testing on the MS COCO dataset. The MLFCP320+ obtains 38.3% accuracy and the MLFCP512+ shows 42.5% accuracy. Our model achieves similar result to M2Det512+ and better than other state-of-the-art works. For example, MLFCP512+ outperforms the RefineDet512+ which adopts the ResNet-101 as baseline model. The M2Det utilizes decoder layers of each U-shape module as the features for detecting objects. Different from the M2Det, in our method, multiple scale context features are used to detect objects.

4.3. From MS COCO to PASCAL VOC

Generally, deep convolutional neural networks achieve better accuracy with large-scale training data. In this experiment, we explore the contribution of large-scale training data on object detection. The MLFCP model is pre-trained by using MS COCO datasets and fine-tuned using VOC07+12 dataset and MS COCO dataset. The network is tested on the PASCAL VOC 2007 test data set.

Table 6: Tests the structure using test-dev 2015.

Method	Train Data	network	Average Precision(%)		
			AP_{50}	AP_{75}	$AP_{50:95}$
ION [14]	train	VGG-16	55.7	34.6	33.1
SSD300 [17]	trainval35k	VGG-16	43.1	25.8	25.1
SSD500 [17]	trainval35k	VGG-16	48.5	30.3	28.8
RON320 [6]	trainval	VGG-16	47.5	25.9	26.2
RON384 [6]	trainval	VGG-16	49.5	27.1	27.4
MDSSD300 [8]	trainval35k	VGG-16	46.0	27.7	26.8
MDSSD512 [8]	trainval35k	VGG-16	50.5	31.4	30.1
PFPNet320+ [29]	trainval35k	VGG-16	60.0	40.7	37.8
PFPNet512+ [29]	trainval35k	VGG-16	61.5	42.6	39.4
RefineDet320+ [7]	trainval35k	VGG-16	56.1	37.7	35.2
RefineDet512+ [7]	trainval35k	VGG-16	58.7	40.8	37.6
RetinaNet400 [18]	trainval35k	Residual-50	47.8	32.7	30.5
RetinaNet400+ [18]	trainval35k	Residual-101	49.5	34.1	31.9
RetinaNet500 [18]	trainval35k	Residual-50	50.9	32.5	34.8
RetinaNet500+ [18]	trainval35k	Residual-101	53.1	36.8	34.4
RetinaNet800 [18]	trainval35k	Residual-50-FPN	59.1	42.3	39.1
RetinaNet800+ [18]	trainval35k	Residual-101-FPN	61.1	44.1	40.8
DSSD513 [12]	trainval35k	Residual-101	53.3	35.2	33.2
RefineDet320+ [7]	trainval35k	Residual-101	59.9	41.7	38.6
RefineDet512+ [7]	trainval35k	Residual-101	62.9	45.7	41.8
M2Det320 [30]	trainval35k	VGG-16	52.4	35.6	33.5
M2Det512 [30]	trainval35k	VGG-16	56.6	40.5	37.6
M2Det320+ [30]	trainval35k	VGG-16	59.1	42.4	38.9
M2Det512+ [30]	trainval35k	VGG-16	62.5	47.2	42.9
MLFCP320	trainval	VGG-16	53.6	34.4	33.0
MLFCP512	trainval	VGG-16	57.5	39.8	37.1
MLFCP320+	trainval	VGG-16	60.4	41.2	38.3
MLFCP512+	trainval	VGG-16	**62.1**	**46.5**	**42.5**

Table 7: Detection results on PASCAL VOC test dataset.The model is Pre-trained using the MS COCO dataset.Using MS COCO and PASCAL VOC 0712 datasets train this network.

Method	Backbone	2007 test (%)
DSOD300 [19]	DS/64-192-48-1	81.7
SSD300 [17]	VGG-16	81.2
SSD512 [17]	VGG-16	83.2
RON320++ [6]	VGG-16	80.3
RON384++ [6]	VGG-16	81.3
RefineDet320+ [7]	VGG-16	85.6
RefineDet512+ [7]	VGG-16	85.8
MLFCP320	VGG-16	85.4
MLFCP512	VGG-16	86.7
MLFCP320+	VGG-16	86.9
MLFCP512+	VGG-16	87.5

Table 7 shows the experimental results. The MLFCP320 achieves 85.4% mAP with MS COCO training data. We observe 3.5% performance improvement compared to the VOC training set. As the input size increasing to 512×512, the MLFCP512 has 86.7% mAP, about 1.3% better than the MLFCP320. We also conduct multi-size testing. The MLFCP320+ has 86.9% mAP and the MLFCP512+ achieves 87.5% mAP, better than most detectors with the same size input images.

4.4. Performance Comparison

Table 5 shows the performance comparison of the MLFCP network and other state-of-the-art models. We evaluate the inference speed of MLFCP network on a machine with a 1080Ti GPU. For input images with size 320×320 and 512×512, the proposed MLFCP network operates at a speed of 16.7FPS and 10.1FPS, respectively.

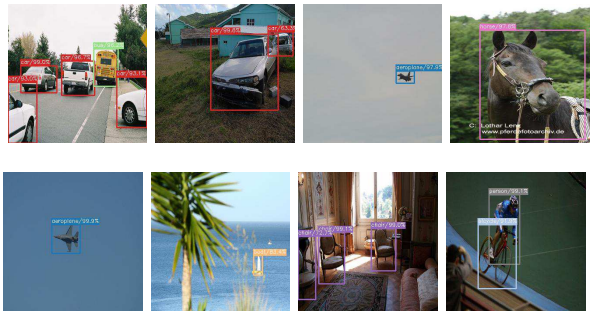

Figure 4: Examples of MLFCPNet detection results.

Some object detection examples are illustrated in Figure 4. As shown in the first row, MLFCPNet works well with a wide range of objects including crowded, overlapped, small and large objects. Examples in the second row show that in the case of blur and occlusion, MLFCPNet might fail to detect objects.

5. CONCLUSIONS

In this paper, a Multi-Level Feature and Context Pyramid Network is proposed to detect multiple scale objects. The new model has two main stages. In the first stage, it

applies a feature fusion module to strengthen network representation ability for both fine-level features and object-level semantic features. The feature fusion module provides combined multiple level features by aggregating different layer feature maps. A context pooling aggregation module is introduced in the second stage. The context module integrates local and global context information to further improve detection performance. Experimental results on the PASCAL VOC 2007 test and the MS COCO datasets demonstrate superior object detection performance of the new framework as compared with state-of-the-art. Ablation studies further demonstrate the effectiveness of the proposed architecture. In future work, we plan to implement a real-time MLFCP object detection system by utilizing a light-weight backbone.

References

[1] J. Redmon, S. K. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection (2016) 779–788.

[2] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, J. Sun, Detnet: Design backbone for object detection (2018) 339–354.

[3] Z. Tian, T. He, C. Shen, Y. Yan, Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation (2019) 3126–3135.

[4] J. Z. H. Wu, K. Huang, Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation, in: IEEE conference on computer vision and pattern recognition (CVPR), arXiv: Computer Vision and Pattern Recognition.

[5] R. G. K. H. B. H. T. Y. Lin, P. Dollar, S. Belongie, Feature pyramid networks for object detection, in: IEEE conference on computer vision and pattern recognition (CVPR), 2017, pp. 936–944.

[6] A. Y. H. L. M. L. T. Kong, F. Sun, Y. Chen, Ron: reverse connection with objectness prior networks for object detection, in: IEEE conference on computer vision and pattern recognition (CVPR), 2017, pp. 5244–5252.

[7] X. B. Z. L. S. Zhang, L. Wen, S. Z. Li, Single-shot refinement neural network for object detection, in: IEEE conference on computer vision and pattern recognition (CVPR), 2018, pp. 4203–4212.

[8] L. Cui, Mdssd: Multi-scale deconvolutional single shot detector for small objects, arXiv: Computer Vision and Pattern Recognition.

[9] T. Kong, F. Sun, W. Huang, H. Liu, Deep feature pyramid reconfiguration for object detection (2018) 172–188.

[10] Y. Li, Y. Chen, N. Wang, Z. Zhang, Scale-aware trident networks for object detection, arXiv: Computer Vision and Pattern Recognition.

[11] J. Zhao, Y. Cao, D. Fan, M. Cheng, X. Li, L. Zhang, Contrast prior and fluid pyramid integration for rgbd salient object detection (2019) 3927–3936.

[12] C. Fu, W. Liu, A. Ranga, A. Tyagi, A. C. Berg, Dssd : Deconvolutional single shot detector., arXiv: Computer Vision and Pattern Recognition.

[13] M. Yang, K. Yu, C. Zhang, Z. Li, K. Yang, Denseaspp for semantic segmentation in street scenes (2018) 3684–3692.

[14] S. Bell, C. L. Zitnick, K. Bala, R. Girshick, Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks (2016) 2874–2883.

[15] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, Residual attention network for image classification (2017) 6450–6458.

[16] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network (2017) 6230–6239.

[17] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, A. C. Berg, Ssd: Single shot multibox detector (2016) 21–37.

[18] T. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal loss for dense object detection (2017) 2999–3007.

[19] Z. Shen, Z. Liu, J. Li, Y. Jiang, Y. Chen, X. Xue, Dsod: Learning deeply supervised object detectors from scratch (2017) 1937–1945.

[20] G. Ghiasi, T. Lin, Q. V. Le, Nas-fpn: Learning scalable feature pyramid architecture for object detection (2019) 7036–7045.

[21] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation (2018) 8759–8768.

[22] H. Li, P. Xiong, H. Fan, J. Sun, Dfanet: Deep feature aggregation for real-time semantic segmentation, arXiv: Computer Vision and Pattern Recognition.

[23] J. Nie, R. M. Anwer, H. Cholakkal, F. S. Khan, Y. Pang, L. Shao, Enriched feature guided refinement network for object detection (2019) 9537–9546.

[24] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, N. Sang, Learning a discriminative feature network for semantic segmentation (2018) 1857–1866.

[25] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, IEEE Transactions on Pattern Analysis and Machine Intelligence (2019) 1–1.

[26] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks 15.

[27] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, C. L. Zitnick, Microsoft coco: Common objects in context (2014) 740–755.

[28] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition.

[29] S. Kim, H. K. Kook, J. Y. Sun, M. C. Kang, S. Ko, Parallel feature pyramid network for object detection (2018) 239–256.

[30] Q. Zhao, T. Sheng, Y. Wang, Z. Tang, L. Cai, Y. Chen, H. Ling, M2det: A single-shot object detector based on multi-level feature pyramid network 33 (01) (2019) 9259–9266.