IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 21, 2019, accepted August 30, 2019, date of publication September 11, 2019, date of current version October 4, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2940626

Scale Pyramid Attention for Single Shot MultiBox Detector

JIE HAO[1], (Member, IEEE), FENG JIANG[2], RUFEI ZHANG[2], XIPENG LIN[1], BIAO LENG[3,4,5], AND GUANGLU SONG[4]

[1]School of Electronics and Information Engineering, Beihang University, Beijing 100191, China
[2]Beijing Institute of Control and Electronic Technology, Beijing 100038, China
[3]Shenzhen Institute of Beihang University, Shenzhen 518057, China
[4]School of Computer Science and Engineering, Beihang University, Beijing 100191, China
[5]Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100191, China

Corresponding author: Biao Leng (lengbiao@buaa.edu.cn)

This work was supported by the Science, Technology and Innovation Commission of Shenzhen Municipality Foundation (JCYJ20170307100750340), and the Beijing Municipal Natural Science Foundation (No. 4182034).

ABSTRACT Feature pyramid mechanism has improved the performance of object detectors by a large margin, especially for the objects with small scale. As one of the first attempts to use pyramidal feature hierarchy, Single Shot MultiBox Detector (SSD) has largely accelerated the pipeline of detection with the competitive performance. Even if the feature pyramid mechanism is used for better detecting objects with small scale, the complex background information still misleads the network to focus on invalid areas. Image areas where small objects exist are more easily to be misjudged which influences the performance of lower-level layer in feature pyramid. In addition, a great deal of false positives may also be introduced into each pyramid layer because of the complex background. In this paper, we propose a novel method named scale pyramid attention hierarchy to better assist the SSD. The general feature pyramid detector can be guided to focus on the valid image areas as well as neglect the influence of complex background with the assist of this proposed module. Furthermore, the end-to-end training can highlight the foreground information which reduces the pressure of detectors, especially for detecting small objects. Flexible implementation makes it easily embedded in the feature pyramid mechanism. Experimental results on PASCAL VOC 2007, VOC 2012 and MS COCO confirm the effectiveness of the proposed module.

INDEX TERMS Object detection, pyramidal attention hierarchy, attention mechanism.

I. INTRODUCTION

Object detection has attracted a wide range of attention in the field of computer vision recently. In comparison with image classification, the object detection is more challenging, which requires more well-designed approaches to solve location and classification problems simultaneously.

Previous works are mostly region-based approaches, popularized by R-CNN [1]. The extended versions Fast R-CNN [2] and Fater R-CNN [3] can achieve the outstanding performance. These two-stage detection algorithms first propose candidate regions, then further classify and regress candidate boxes. Each stage of training and testing takes a certain amount of time. Accordingly, the speed of these pipelines is largely limited by the two-stage mechanism. Therefore, In order to accelerate the detection pipeline while maintaining the acceptable performance, a series of one-stage algorithms are proposed [4]–[7]. These one-stage methods classify and regress candidate boxes just after presenting them without using additional networks. So it effectively improves the detection speed. However, it is difficult to detect objects with variant scales using the single-level layer because of the limitation of receptive field. The development of one-stage detection algorithm is greatly restricted by this problem. Fortunately, thanks to the feature pyramid mechanism [8], this issue is greatly alleviated.

The feature pyramid mechanism utilizes the property that different-level layers have different receptive fields which is suitable for detecting objects with different scales. Using the information in different-level layers to detect specific scale of objects, and it makes the detector more friendly to objects with variant scales. SSD [7] is one of the first attempts to use pyramidal feature to improve the performance of object detectors, while largely accelerating the detection pipeline. Although the feature pyramid can be of great benefit to detect objects with variant scales, the different levels in the detector

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/



FIGURE 1. Illustration of detection results through SSD detector baseline.

are always easily disturbed by different kinds of factors. In feature pyramid, the lower-level layers have a smaller receptive field and tend to preserve the fine-grained information of the input image, so it is more suitable to detect small targets for low-lever of layers. On the contrary, the deeper network layer has larger receptive field, which is more suitable to detect large-scale targets. Moreover, the different receptive fields also introduce many false positive samples into each pyramid layer due to the complicated background. As is shown in FIGURE 1, the first line is a group of images with ground truth bounding boxes, and the the second line is a group of images with detected bounding boxes. The red boxes in detection results represent the center of coordinate is within valid areas and green boxes represent the center of coordinate is within invalid areas. FIGURE 1 shows that the complex background information will mislead the detector to focus more on invalid areas than the valid areas.

In order to deal with the problem that the network is more likely to be disturbed by complex background, we propose a supervision mechanism to make the network pay more attention to the target regions. And for the sake of coping with the issue that different levels of network layer are easily disturbed by different scales of targets, we propose a multi-scale mechanism to enable different network layers to detect targets of different scales. Furthermore, we embed the attention and multi-scale mechanism into one unified network architecture, named scale pyramid attention (SPA) to better guide the SSD. Each pyramid layer generates a supervised feature map to rectify the features of the current layer, and each layer of the network processes targets of different scales. More precisely, the SPA consists of multi-level sub-SPAs, which can be embedded into corresponding layers of SSD. Every sub-SPA is devised using lightweight ConvNet and can be trained with a well-designed strategy to learn the spatial information of foreground in corresponding layers. Different from other state-of-the-art SSD-variants, such as [9]–[13], where [9], [10] and [11] fuse the feature of different levels in order to improve the expressive ability of network, [12] and [13] re-design the loss function or classification strategies to cope with the imbalance of samples, SPA focuses on the detailed semantic representation by highlighting the valid regions where objects exist and reducing the influence of complex background.

To verify the effectiveness of our proposed SPA, we integrate it with the SSD. As is shown in FIGURE 2, the first column on the left shows the input image and the locations of ground truth bounding boxes. The other four columns show the feature maps extracted from lowest four layers. The value of each pixel is indicated by the color map, from blue to red represents from small to large representatively. FIGURE 2 shows that the lowest three layers are mostly affected by the complex background information. Thus, the SPA is embedded into the lowest three layers to guide SSD detector to focus on foreground areas. Experimental results in PASCAL VOC 2007, VOC 2012 [14] and MS COCO [15] demonstrate the SPA-based detector could achieve more competitive results.

To summarize, the contributions of our work are as follow:

- We propose a novel method named scale pyramid attention hierarchy(SPA), which can guide the ConvNet's feature pyramids to focus on valid areas and neglect the complex background.
- Our methods make full use of the field of perception of different network layers, so that the different layers can deal with different scales of targets;
- We device a specific training strategy to train every sub-SPA in associated layers with a multi-loss function, which guarantees the accurate prediction of spatial information in each pyramidal layers.
- Adaptive experiments prove the effectiveness of the proposed method.

II. RELATED WORK

A. CLASSICAL OBJECT DETECTORS

The sliding windows strategy is widely adopted in early detecting algorithm, where a high-quality classifier and hand-crafted feature are applied over dense image grids to

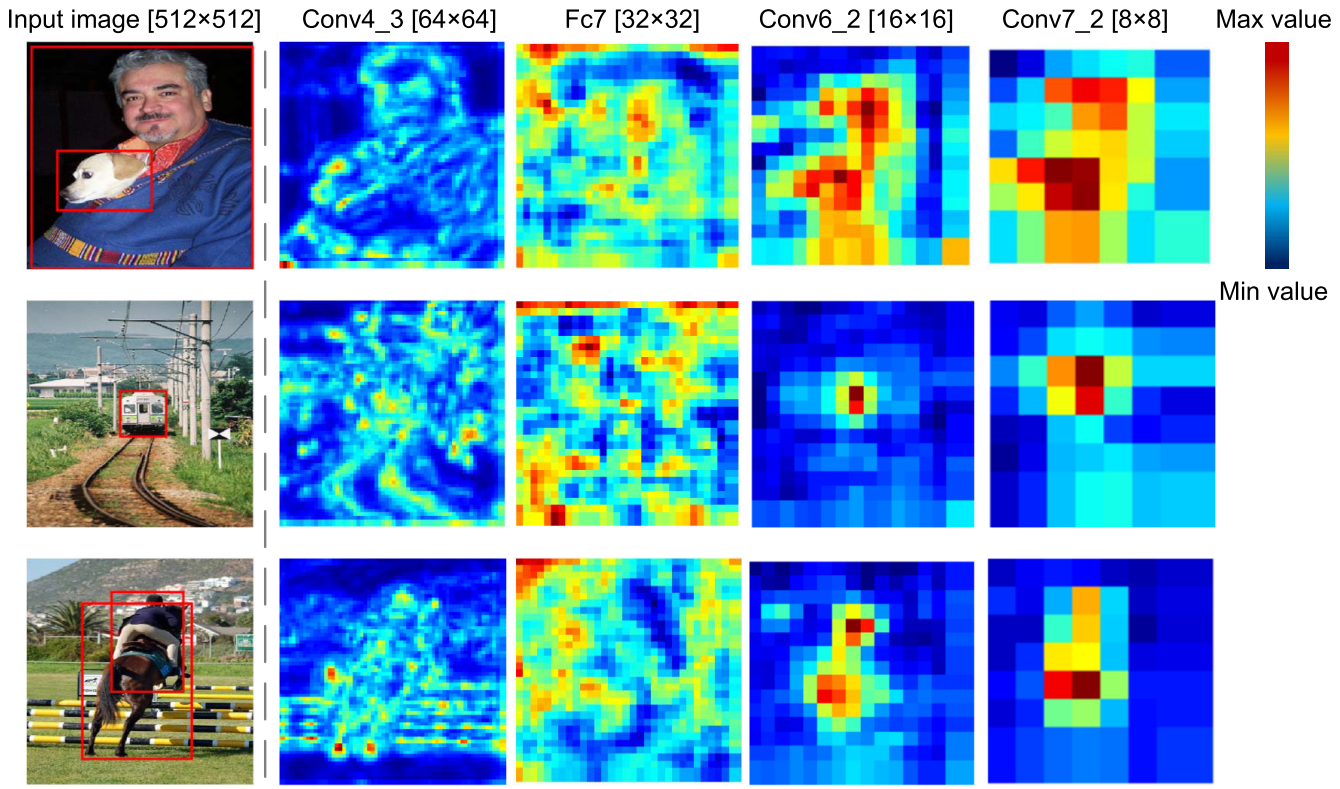


FIGURE 2. Illustration of feature maps in SSD's feature pyramids.

locate objects. In order to improve performance, hand-crafted features, such as SIFT [16] and HOG [17] are introduced, which are both computed densely over image pyramid. As one of successful early methods, Viola and Jones [18] train a series of cascade classifiers for face detection, and achieve competitive performance with high speed. DPM [19] is another successful method, which is an extension of HOG and maintained top performance on benchmark Pascal VOC [14] for many years. But with the resurgence of convolution neural network [20], the object detection is gradually dominated by ConvNet-based methods.

B. CONVNET-BASED OBJECT DETECTORS

Currently, the ConvNet-based methods can be roughly divided into two categories, including two-stage detectors and one-stage detectors. As popularized by R-CNN [3], the first stage of two-stage approaches generates a sparse set of candidate object proposals which may contain the most targets. Then these proposals are classified and regressed to yield more accurate results in the second stage. Based on two-stage mechanism, further detectors [2], [3], [21]–[24] and [25]–[27] are proposed and make two-stage methods achieve state-of-the-art performance on several popular datasets (Pascal VOC [14] and MS COCO [15]). Although two-stage methods can achieve excellent accuracy, the efficiency is largely limited by the two-stage design. Hence, a series of one stage methods [4]–[7] are put forward to accelerate the speed of pipeline. The one-stage methods make prediction on dense sampling proposals and can advocate a trade-off between accuracy and speed. However, due to the fixed receptive field, it is difficult to handle various scales of objects by making prediction on a single-scale layer. To solve this problem, feature pyramids hierarchy computed by FCN [28] was proposed.

C. FEATURE PYRAMIDS HIERARCHY

Recently, numerous approaches have improved performance by using information from different layers in a ConvNet's feature pyramids. As one of the first attempts to use ConvNet's feature pyramids, SSD accelerates the speed largely with maintaining high accuracy. [6], [8], [29] integrate contextual information across multi-scale layers to get higher recall. [9]–[11] focus on better integrating the features of different levels to improve the feature representation. Although one-stage methods can achieve better performance through feature pyramids, the lower-level layers are inclined to be influenced by complex background information. The lower-level layers capture more fine-grained information and usually mislead the detector to concentrate on invalid area.

III. PROPOSED METHOD

In this section, we depict the details of the proposed SPA-based SSD (Sec. III-A) architecture and the associated training methodology for every sub-SPA concretely in (Sec. III-B). Afterwards, we present the experimental results in PASCAL VOC2007, VOC2012 [14] and MS COCO [15]

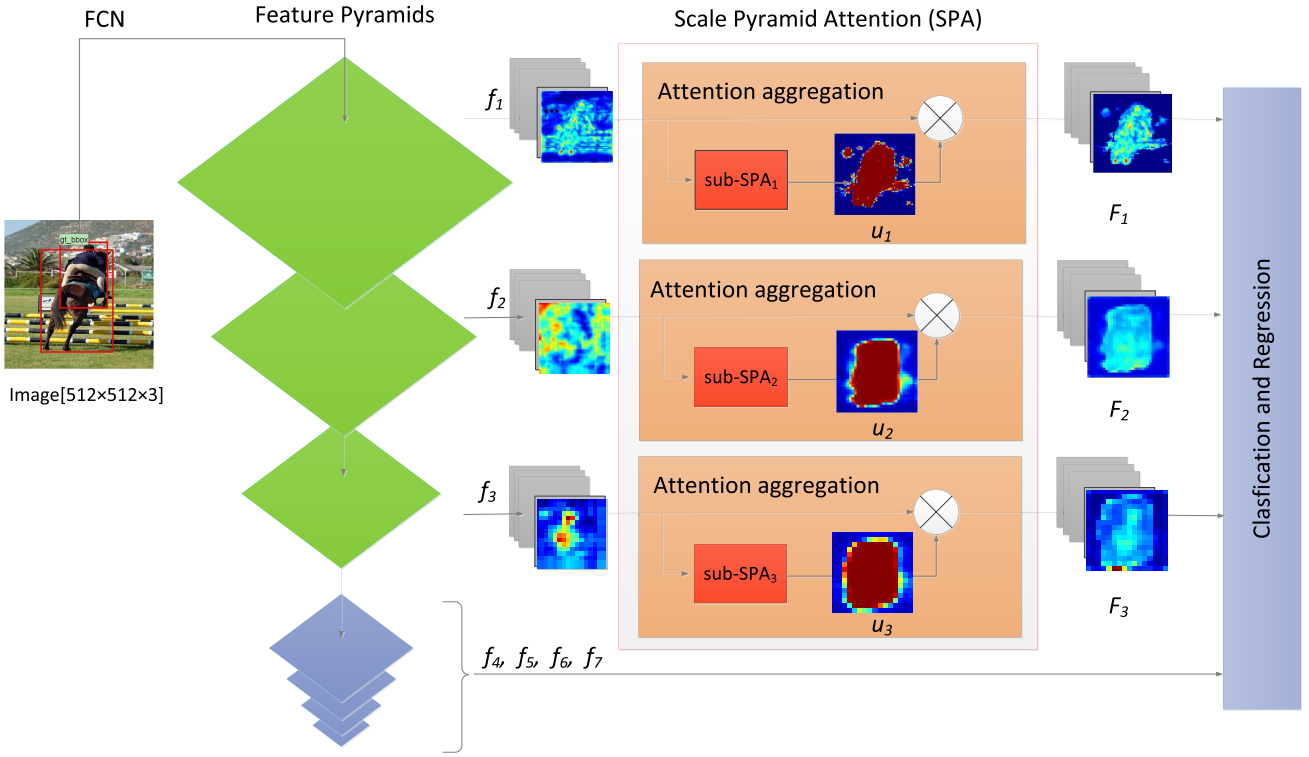


FIGURE 3. Architecture of the proposed SPA.

in Sec. III-C with comparisons of different method details and published works, which strongly validate the advantage of our method.

A. ARCHITECTURE OVERVIEW

As shown in FIGURE 3, the size of input image is fixed to 512 × 512. Then the image is sent to the fully convolution network(FCN) to generate the feature pyramids. Feature maps from the lowest three layers are fed into SPA, while the other feature maps are directly sent to classification and regression unit. Every sub-SPA$_i$ unit generates an attention map. Then the attention maps and associated feature maps will be aggregated by the attention aggregation unit to produce the final feature maps. We will introduce the concrete structure of sub-SPA$_i$ unit in the next section. Given the input image I, the fully convolution network (FCN [28]) will generate feature pyramids, composed of m different-scale layers (the value of m is 7 in SSD). We define $\{f_i(I)\}$, $i \in [1, m]$ for feature maps from different layers. FIGURE 2 shows some feature maps examples in SSD's feature pyramids. In SSD baseline, the lower layers are more seriously affected by background. Accordingly, the SPA is embedded in the lowest three layers. All sub-SPA$_i$ units will learn to generate attention maps $\{\mu_1(I), \mu_2(I), \mu_3(I)\}$. Eventually, the attention maps and their corresponding feature maps generate the refined feature maps $\{F_1(I), F_2(I), F_3(I)\}$ by the attention aggregation unit. We will discuss these units and their functions in the following sections.

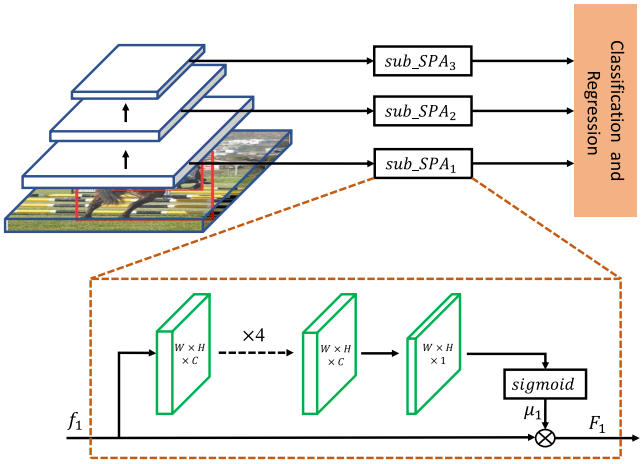


FIGURE 4. The specific structure of sub-SPA$_i$.

B. THE SPECIFIC STRUCTURE OF TRAINING SUB-SPA$_I$

In FIGURE 4, we simplify the overall network architecture and show the structure of sub-SPA$_i$ in details. For better illustration, we take sub-SPA$_1$ in the first layer for example. Every sub-SPA$_i$ is aiming at fully mining the spatial information of foreground. Given the input feature map f_i, a tiny feed-forward network is applied to learn an attention map μ_i of the same size as the input feature. The output of the last convolution layer of every tiny feed-forward network is scaled to [0, 1] by executing sigmoid function $\sigma(\cdot)$. After passing through these layers, we get corresponding

attention map. In order to ensure more accurate prediction of attention maps, sub-SPA$_i$ are supervised by well-designed foreground labels $\mathcal{L}_i$.

For each object j in image I, we define $B_t^{j_local}$ to represent particular localization-dependent regions in associated feature map, where $t \in \{0, 1, 2\}$ indicates the first three layers in feature pyramids and $t = 0$ represents the lowest level of the pyramid feature. Suppose that (x_{jl}, y_{jl}), (x_{jr}, y_{jr}) represent the top left and bottom right corners of ground truth bounding box, then the attention boundary of the corresponding object j can be formulated as $(x_{t,j}^{min}, y_{t,j}^{min})$, $(x_{t,j}^{max}, y_{t,j}^{max})$:

$$\begin{aligned} x_{t,j}^{min} &= max(\lceil \frac{x_{jl}}{N_s} \rceil - t, 0) \\ y_{t,j}^{min} &= max(\lceil \frac{y_{jl}}{N_s} \rceil - t, 0) \\ x_{t,j}^{max} &= min(\lceil \frac{x_{jr}}{N_s} \rceil + t, W_i) \\ y_{t,j}^{max} &= min(\lceil \frac{y_{jr}}{N_s} \rceil + t, H_i) \end{aligned} \tag{1}$$

where N_s means the stride of the current pyramid feature t and $\lceil \frac{x_{jl}}{N_s} \rceil$ signifies the corresponding coordinates of bounding box in feature map. We associate the feature pyramid level t with the attention boundary and hope that the high-level t with larger receptive field can be assigned the enhanced boundary. Based above, we can easily get the localization-based enhanced attention regions for object j in pyramid layer t as follows:

$$\begin{aligned} B_t^{j_local} = \{(x, y) | & x_{t,j}^{min} \leqslant x \leqslant x_{t,j}^{max}, \\ & y_{t,j}^{min} \leqslant y \leqslant y_{t,j}^{max}\} \end{aligned} \tag{2}$$

Furthermore, this paper also limits the scale of the attention regions for different pyramid feature layers. Suppose that S_j is the area of the ground truth object j, it can be expressed as:

$$S_j = (x_{jr} - x_{jl}) \times (y_{jr} - y_{jl}) \tag{3}$$

In the pyramid feature t, the scale-related attention region $B_t^{j_scale}$ can be expressed as follows:

$$B_t^{j_scale} = \{(x, y) | S_t^{min} \leqslant S_j \leqslant S_t^{max}\} \tag{4}$$

We refer to [7] for scale boundary S_t^{min} and S_t^{max} settings. The higher the level of pyramid feature layer, the larger the scale of settings. Based above, the final attention region set B_t^j can be get as the intersection of $B_t^{j_local}$ and $B_t^{j_scale}$:

$$B_t^j = B_t^{j_local} \bigcap B_t^{j_scale} \tag{5}$$

More specifically, as a function of the ground truth objects' coordinates and areas, the attention region set B_t^j can be expressed as:

$$\begin{aligned} B_t^j = \{(x, y) | & x_{t,j}^{min} \leqslant x \leqslant x_{t,j}^{max}, \\ & y_{t,j}^{min} \leqslant y \leqslant y_{t,j}^{max} \\ & S_t^{min} \leqslant S_j \leqslant S_t^{max}\} \end{aligned} \tag{6}$$

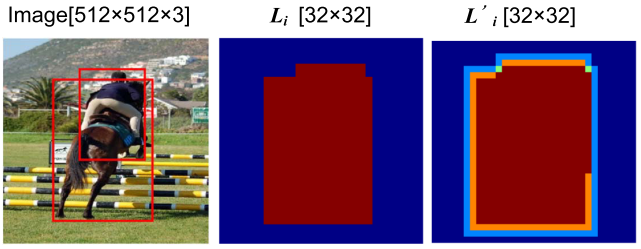


FIGURE 5. An example of difference between $\mathcal{L}_i$ and $\mathcal{L}'_i$.

If the number of objects is J, The corresponding label set B_t on pyramid layer t can be expressed as:

$$B_t = \bigcup_{j=1}^{J} B_t^j \tag{7}$$

If $t = 0$, B_0 means the valid areas within ground truth bounding boxes in the associated feature map.

Following this, the value of each coordinate in the $\mathcal{L}_i$ is formulated as:

$$\mathcal{L}_i(x, y) = \begin{cases} 1 & (x, y) \in B_0 \\ 0 & else \end{cases} \tag{8}$$

In $\mathcal{L}_i$, the edge of boxes is a mutation from 1 to 0. Considering every point in attention labels has a particular receptive field, we devise a more smooth label $\mathcal{L}'_i$, where the edge is a gradual transition from 1 to 0.

For every object j, $\mathcal{L}'^j_i$ can be formulated as:

$$\mathcal{L}'^j_i(x, y) = \begin{cases} 1 & (x, y) \in B_0^j \\ 0.5 & (x, y) \in B_1^j \cap \overline{B}_0^j \\ 0.25 & (x, y) \in B_2^j \cap \overline{B}_1^j \\ 0 & else \end{cases} \tag{9}$$

Following this, we can get the value of each coordinate in the $\mathcal{L}'_i(x, y)$:

$$\mathcal{L}'_i(x, y) = min(\sum_{j=1}^{J} \mathcal{L}'^j_i(x, y), 1) \tag{10}$$

FIGURE 5 displays a pair of example on $\mathcal{L}_i$ and $\mathcal{L}'_i$. For better illustration, we take labels generated in second layers for example, whose resolution is 32×32. From the picture, we can see the difference between $\mathcal{L}_i$ and $\mathcal{L}'_i$ intuitively. Compared with sudden change in edge of $\mathcal{L}_i$, the edges of bounding boxes are more smooth in $\mathcal{L}'_i$. Specific experiments comparing these two kinds of labels will be presented in the following section IV-A.2.

Training objective We use sigmoid cross entropy loss as loss function. The loss of sub-SPA$_i$ can be formulated as follow:

$$\begin{aligned} L_{sub-SPA_i} = -\frac{1}{N} \sum_{n=1}^{N} & [p_n(x, y) log \widehat{p}_n(x, y) \\ & + (1 - p_n(x, y)) log(1 - \widehat{p}_n(x, y))] \end{aligned} \tag{11}$$

TABLE 1. Comparison with previous published two-stage and one-stage works on Pascal VOC dateset. The * in the table means the results we tested on our own machine, and the rest are published in [30].

Methods	Backbone	Input size	FPS	mAP@0.5(%)	
				VOC 2007	VOC 2012
$two-stage:$					
Fast R-CNN [2]	VGG-16 [31]	$\sim 1000 \times 600$	0.5	70.0	68.4
Faster R-CNN [3]	VGG-16	$\sim 1000 \times 600$	7	73.2	70.4
OHEM [26]	VGG-16	$\sim 1000 \times 600$	7	74.6	71.9
HyperNet [32]	VGG-16	$\sim 1000 \times 600$	0.88	76.3	71.4
Faster R-CNN [3]	ResNet-101 [33]	$\sim 1000 \times 600$	2.4	76.4	73.8
$one-stage:$					
YOLO [4]	GoogleNet [34]	448×448	45	63.4	57.9
RON384 [12]	VGG-16	384×384	15	75.4	73.0
YOLOv2 [5]	Darknet-19 [5]	544×544	40	78.6	73.4
SSD512 [7]	VGG-16	512×512	21.5*	76.8	74.9
DSOD [35]	DS/64-192-48-1	300×300	17.4	77.7	76.3
SPA(ours)	VGG-16	512×512	18.8*	77.4	77.48

where N signifies the total number of coordinates in $\mathcal{L}'_i$, $\widehat{p}_n(x, y)$ is the computed probability in coordinate (x, y) normalized by sigmoid function, and $p_n(x, y)$ is the ground truth in labels.

C. ATTENTION AGGREGATION UNIT

The attention aggregation is adopted to highlight the foreground information and weaken the background influence in lower-level layers. For an image I, fully convolution network generates the scale feature pyramids, including seven feature maps $\{f_i(I)\}, i \in [1, 7]$. The lowest three layers were sent to corresponding sub-SPA$_i$ to produce associated attention maps $\{\mu_1(I), \mu_2(I), \mu_3(I)\}$. The final refined feature maps can be denoted as $F_i = \{\Phi_1(I), \Phi_2(I), \Phi_3(I)\}$, where $\Phi(\cdot)$ is the attention aggregation function which generates refined feature maps by incorporate attention maps with every channel of feature maps. The aggregation is formulated as:

$$F_i^k(x, y) = f_i^k(x, y) \times \mu_i(x, y), \quad k \in [1, K] \tag{12}$$

where x and y are the horizontal and vertical positions and k is the channel of feature maps.

During training, the SPA can not only serve as a feature selector during forward inference, but also as a gradient update filter during back propagation. The gradient of SPA's loss for input feature is:

$$\frac{\partial L_{SPA}}{\partial f_i} = \frac{\partial L_{SPA}}{\partial F_i} \cdot \frac{\partial F_i}{\partial f_i} = \mu_i \cdot \frac{\partial L_{SPA}}{\partial F_i} \tag{13}$$

During each training iteration, the gradient will be back-propagated to each coordinates in F_i. Then, the gradient within invalid areas will be killed under the guide of attention map. Only the parameters within foreground areas will be updated. Therefore, the SPA will make base network be more robust to noise background and alleviated the pressure to detect small object significantly.

IV. EXPERIMENTS

A. PASCAL VOC 2007

We first perform experiments on the PASCAL VOC 2007 detection benchmarks, which contain objects from 20 categories. All models are trained on the union set of VOC 2007 *trainval* and VOC 2012 *trainval* (VOC 07 + 12 *trainval*, 16,551 images), and test on VOC 2007 *test* set (VOC 07 test), including 4952 images. We use VGG16 [31] as backbone network, which is pre-trained on ILSVRC CLS-LOC dataset [36]. Then the network is fine-tuned using SGD with base learning rate 10^{-3} for the first 80k iterations, and being decayed to 10^{-4} and 10^{-5} in the following twice 20k iterations, respectively. The batch size is fixed to 32 and momentum is set to 0.9. We make SSD and the proposed method under the above configuration and the training time is both around 18 hours. For the sake of generality, we report Average Precision using IOU threshold at 0.5 (AP@0.5) to evaluate the accuracy. The whole implementations are based on the deep learning framework Caffe [37].

We compare our proposed method SPA with the previous related works in table 1. As shown in the table, our proposed SPA achieves better performance than all previous two-stage detectors. Even though SPA uses smaller image size and is based on VGG-16 backbone, it achieves 1.0% above the top-performing Faster R-CNN based on ResNet101.

Compared to other one-stage detectors, our method outperforms most works by a large margin, but has a big gap below YOLOv2 [5]. We also make other experiments on VOC 2007 Dataset.

1) RUN TIME PERFORMANCE

We present the inference speed of proposed method and the state of-the-art methods in Table 1. The speed is evaluated with batch size 1 on a machine with NVIDIA Titan X, CUDA 8.0 and cuDNN v6. The * in the table means the results we tested on our own machine, and the rest are published in [30].

TABLE 2. MS COCO test-dev2015 result.

Methods	Data	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
$two-stage$:								
Fast R-CNN [2]	train	VGG-16 [31]	19.7	35.9	-	-	-	-
Faster R-CNN [3]	trainval	VGG-16	21.9	42.7	-	-	-	-
OHEM [26]	trainval	VGG-16	22.6	42.5	22.2	5.0	23.7	37.9
R-FCN [23]	trainval	VGG-16	29.9	51.9	-	10.8	32.8	45.0
Faster R-CNN with FPN [33]	trainval35k	ResNet-101	36.2	59.1	39.0	18.2	39.0	48.2
$one-stage$:								
RON384 [12]	trainval	VGG-16	27.4	49.5	27.1	-	-	-
YOLOv2 [5]	trainval35k	Darknet-19 [5]	21.6	44.0	19.2	5.0	22.4	35.5
SSD300 [7]	trainval35k	VGG-16	23.2	41.2	23.4	5.3	23.2	39.6
SSD512 [7]	trainval35k	VGG-16	26.8	46.5	27.8	9.0	28.9	41.9
SPA(ours)	trainval35k	VGG-16	29.8	48.9	31.5	10.8	34.1	45.1

TABLE 3. Effects of embedding SPA in different layers of feature pyramids.

Methods	f_1	f_2	f_3	f_4	mAP@0.5(%)
SPA-4	√	√	√	√	76.9
SPA-3	√	√	√		77.4
SPA-2	√	√			77.2
SPA-1	√				76.8
Baseline					76.8

TABLE 4. Comparation between $\mathcal{L}_i$ and $\mathcal{L}'_i$.

Methods	Label	mAP@0.5(%)
SPA-2	$\mathcal{L}_i$	76.7
	$\mathcal{L}'_i$	77.2
SPA-3	$\mathcal{L}_i$	76.5
	$\mathcal{L}'_i$	77.4
Baseline		76.8

As shown in Table 1, SPA can improve the performance of basic SSD by a large margin with negligible inference burden.

2) ALATION STUDY ON SPA

Table 3 compares the influence of SPA in different layers of feature pyramids. For fair comparison, all these results shown in the table are trained from the same pre-trained model and using the same configuration and training strategy. To measure the improvement gain SPA brought to baseline, we progressively add single sub-SPA from the lowest layers of feature pyramid. The higher-level layers have larger receptive field and hardly affected by background. Thus, we only perform experiments on the lowest four layers.

Compared to SSD baseline, the mAP of SPA-1 almost remain unchanged. When the SPA is embedded into the lowest two layers, the SPA-2 can achieve better performance with 0.4% improvement. When feature pyramids add SPA in the lowest three layers (SPA-3), the results is higher than baseline by 0.6%. As for SPA-4, the mAP drops, but still higher than baseline by 0.1%. This demonstrates that ours proposed PA can guide the feature pyramids to focus on valid areas in lower-level layers and reduce the influence of complex background.

3) COMPARISON BETWEEN $\mathcal{L}_i$ AND $\mathcal{L}'_i$

In previous section III-B, we put forward two kinds of attention label, namely $\mathcal{L}_i$ and $\mathcal{L}'_i$. The bounding boxes edge in attention label $\mathcal{L}'_i$ is gradual change, which is more suitable for supervising training process. To verify the advantage of $\mathcal{L}'_i$, we conduct related experiments based on SPA-2 and SPA-3, respectively. In table 4, we can notice that $\mathcal{L}_i$ hardly improve the performance, even hurt the mAP of baseline, which demonstrates the mutation in edge of attention label is harmful for the accuracy of attention maps.

B. PASCAL VOC 2012

According to VOC 12 evaluation criteria, we submit the detection results of the proposed method to the testing server for evaluation. We use VOC 2007 *trainval*, *test* set and VOC 2012 *trainval* set (VOC 07++12 *trainval*, 21,503 images) for training, and test on VOC 2012 *test* set (10,991 images). The network is fine-tuned using SGD with base learning rate 10^{-3} for the first 160k iterations, and reducing to 10^{-4} and 10^{-5} in the following twice 40k iterations, respectively. The training time of SSD and the proposed method is both about 24 hours. Other experimental settings are the same as PASCAL VOC 2007.

Table 1 shows the accuracy of the proposed algorithm, as well as the other state-of-the-art methods. As can be seen from the table, the accuracy on VOC 2012 is generally slightly lower than the result on VOC 2007 for other methods. But the result of the proposed method on VOC 2012 is slightly higher than that on VOC 2007. This fully proves the robustness of proposed SPA-SSD method. In the vertical direction, the comparison between the proposed method and others is roughly the same as that of VOC 2007 in VOC 2012 data set,

surpassing other single-stage and two-stage methods. This shows that scale pyramid attention hierachy does have a stable enhancement to the neural network methods.

C. MS COCO

In addition to PASCAL VOC, we also evaluate our proposed method on MS COCO [15]. Following the protocol in MS COCO, we use the *trainval*35*k* set [38] for training and evaluate the results from *test*−*dev*2015 evaluation server. The backbone network is VGG16 [31] which is also pre-trained on ILSVRC CLS-LOC dataset [36]. The batch size is fixed to 32 and the learning rate is set to 10^{-3} for the first 280*k* iterations, then 10^{-4} and 10^{-5} for another 40*k* and 80*k* iterations, respectively.

As can be seen from Table 2, compared to most single-stage detection methods, the proposed SPA algorithm can get better accuracy, especially on small scales. This shows that the SPA method is really helpful to improve the detection performance. However, the proposed method is still not as good as the two-stage detection method Faster R-CNN with FPN [33], indicating that the single-stage approach still has a lot of room for improvement. The training time is about 75 hours.

V. CONCLUSION

In this paper, we put forward a novel scale pyramid attention hierarchy(SPA) to guide feature pyramids based detectors. With the assistance of SPA, the lower-level layers are more easily to focus on valid areas and get rid of complex background details. The SPA can be trained to predict accurate foreground information by a novel training strategy. To verify the effectiveness of our methods, we make numerous experiments on PASCAL VOC and MS COCO to show that our SPA method can improve the performance of SSD dramatically, especially for objects with small scales. Later, we will explore the performance of SPA module on other networks and tasks.

ACKNOWLEDGMENT

This work was supported by the Science, Technology and Innovation Commission of Shenzhen Municipality Foundation (JCYJ20170307100750340), and the Beijing Municipal Natural Science Foundation (No. 4182034).

REFERENCES

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 580–587.

[2] B. R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)* Santiago, Chile, Dec. 2015, pp. 1440–1448.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2015, pp. 91–99.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 779–788.

[5] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 7263–7271.

[6] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," Jan. 2017, *arXiv:1701.06659*. [Online]. Available: https://arxiv.org/abs/1701.06659#

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[8] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2117–2125.

[9] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," 2017, *arXiv:1701.06659*. [Online]. Available: https://arxiv.org/abs/1701.06659

[10] G. Cao, X. Xie, W. Yang, Q. Liao, G. Shi, and J. Wu, "Feature-fused SSD: Fast detection for small objects," *Proc. SPIE*, vol. 10615, Apr. 2018, Art. no. 106151E.

[11] Q. Zhao, T. Sheng, Y. Wang, Z. Tang, Y. Chen, L. Cai, and H. Ling, "M2Det: A single-shot object detector based on multi-level feature pyramid network," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 9259–9266.

[12] T. Kong, F. Sun, A. Yao, H. Liu, M. Lu, and Y. Chen, "RON: Reverse connection with objectness prior networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 5244–5252.

[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2980–2988,

[14] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, and P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 740–755.

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, Jun. 2005, pp. 886–893.

[18] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2001, pp. 511–518.

[19] P. Felzenszwalb, D. Mcallester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. CVPR*, Jun. 2008, pp. 1–8.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, 2012, pp. 1106–1114.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proc. Eur. Conf. Comput. Vis.*, Zurich, Switzerland, Sep. 2014, pp. 346–361.

[22] K. He, G. Gkioxari, and P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2961–2969.

[23] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Barcelona, Spain, Dec. 2016, pp. 379–387.

[24] Y. Zhu, C. Zhao, J. Wang, X. Zhao, Y. Wu, and H. Lu, "CoupleNet: Coupling global structure with local parts for object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 4126–4134.

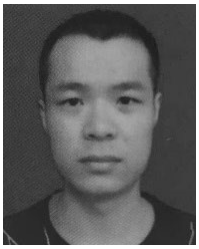
[25] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Apr. 2013.

[26] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 761–769.

[27] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Computer Vision—ECCV*. Zürich, Switzerland: Springer, Sep. 2014, pp. 391–405.

[28] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3431–3440.

[29] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.
[30] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 4203–4212.
[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Comput. Sci.*, to be published.
[32] T. Kong, A. Yao, Y. Chen, and F. Sun, "HyperNet: Towards accurate region proposal generation and joint object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 845–853.
[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.
[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9.
[35] Z. Shen, Z. Liu, J. Li, Y.-G. Jiang, Y. Chen, and X. Xue, "DSOD: Learning deeply supervised object detectors from scratch," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1919–1927.
[36] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[37] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd Int. Conf. Multimedia*, 2014, pp. 675–678.
[38] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2874–2883.



JIE HAO (M'93) received the bachelor's degree from the College of Electronic Information Engineering, Beihang University, Beijing, China, in 2017, where he is currently pursuing the master's degree. His research interests include object detection and video target tracking.



FENG JIANG was born in 1995. He received the master's degree from the Beijing Institute of Technology. He is currently an Engineer. His research interests include intelligent perception and collaborative control.



RUFEI ZHANG was born in 1981. He received the Ph.D. degree from Northwestern Polytechnical University. He is currently a Senior Engineer. His research interest includes guidance, navigation, and control.



XIPENG LIN received the bachelor's degree from the College of Electronic Information Engineering, Beihang University, Beijing, China, in 2017, where he is currently pursuing the master's degree. His research interests include object detection and re-identification.



BIAO LENG received the B.S. degree from the National University of Defense Technology, in 2004, and the Ph.D. degree form Tsinghua University, in 2009. He is currently an Associated Professor with Beihang University. His main research interests include: 1) multimedia information processing, image and 3D model retrieval, video analysis, and understanding; 2) large data analysis, data mining, and data activation; and 3) complex network and so on.



GUANGLU SONG received the B.S. degree from Beihang University, Beijing, China, where he is currently pursuing the M.S. degree with the School of Computer Science and Engineering. His research interests include face detection, pedestrian detection, and re-identification.