基于特征金字塔网络的改进算法

陈景明，金杰*，王伟锋
天津大学电气自动化与信息工程学院，天津 300072

摘要 针对小目标检测提出了一种基于特征金字塔网络改进的算法。通过引入预测优化模块，并结合感兴趣区域的上下文信息，使得特征信息具有更强的稳健性，同时通过内部级联的多阈值预测网络进行预测，最终实现多尺度多阶段的预测，在保证网络参数基本不变的前提下准确率得到提升。实验结果表明，经标准数据集 VOC07+12 训练后，所提算法在 VOC2007 测试中的准确率达到 80.9%，具有很好的检测性能。

关键词 机器视觉；特征金字塔；目标检测；多尺度检测；级联网络
中图分类号 TP391 **文献标识码** A **doi**：10.3788/LOP56.211505

Improved Algorithm Based on Feature Pyramid Networks

Chen Jingming, Jin Jie*, Wang Weifeng
School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, *China*

Abstract An improved algorithm based on feature pyramid networks is proposed for small target detection. A prediction optimization module is introduced, which is combined with the context information of the region of interest to make the feature information more robust, multi-threshold prediction networks with internal cascade are predicted, and the multi-scale and multi-stage prediction is realized finally. On the premise that the network parameters are basically unchanged, the accuracy is further improved. The experimental results show that the accuracy of the proposed algorithm reaches 80.9% in the VOC2007 test after the training of the standard data set VOC07+12, which has good detection performance.

Key words machine vision; feature pyramid; object detection; multi-scale detection; cascade network
OCIS codes 150.1135; 100.3008; 040.1880

1 引 言

作为计算机视觉的三大分支之一的目标检测，一直是科研的热门课题。无论是实时性良好的单阶段检测算法（如 YOLO[1]、改进的 SSD[2]、SSD[3] 和 RetinaNet[4]），还是准确率普遍较高的基于区域提议的算法（如 RCNN[5]、改进的 Faster R-CNN[6] 和 R-FCN[7]），多尺度的目标识别一直都是算法改进的一个挑战。在进行目标检测时，直接在最后一层特征图上进行预测，YOLO、RCNN 等采用此方式，获得了较好的检测结果，但由于忽略了其他层的特征信息，在进行目标检测时极易丢失小目标，对小目标的检测效果明显变差。于是，人们尝试改进多尺度的目标检测。目前常用的改进方法主要有两种：1）使用不同特征图进行不同尺度的预测；2）结合多层特征进行预测。

一般而言，使用不同特征图进行不同尺度的预测时，在底层的特征图上预测小目标，在顶层的特征图上预测大目标。文献[8]通过结合底层、中层和顶层的特征图来缓解小目标的检测问题。文献[9]提出缩小与放大网络，其结构类似于卷积/反卷积结构，通过特征图注意力决策单元在两条分流中自适应激活某些通道的特征，调整权重以优化整体的训练损失。虽然此方法对小目标的检测效果有一定的改善，但是没有考虑特征图之间的联系，这使得在进行小目标检测时容易丢失语义信息，在进行大目标

收稿日期：2019-03-21；**修回日期**：2019-04-18；**录用日期**：2019-04-30
基金项目：国家自然科学基金（61571320）
* **E-mail**：jinjie@tju.edu.cn

检测时容易丢失位置信息。

结合多层特征进行预测，这种不同特征的结合使得最终预测的特征具有更强的语义信息，检测结果也更加准确。文献[10]提出了特征金字塔网络(FPN)算法，该算法有两条分支：1)自上而下的反馈网络；2)自下而上的前馈网络。通过横向连接将二者结合，实现多尺度的目标检测。文献[11]采用类似 FPN 的结构，该结构同样由三部分组成，但该算法是在最后结合的特征图上进行预测。文献[12]中的算法是通过引入的反向连接来解决多尺度的问题。文献[13]通过类似的尺度转化模块结合多层特征进行多尺度预测。文献[14]则是通过传输连接块来融合高层特征，其结构也类似于 FPN。

虽然 FPN 通过结合多个特征图来进行多尺度的检测，在小目标的检测方面有所改善，但是该算法没有结合上下文信息，当重叠度(IOU)阈值增大后，检测效果会下降。针对以上不足，文献[15]中的算法引入定向特征金字塔结构，即在 FPN 的基础上又增加了反向特征金字塔，进一步使得高语义的特征利用底层的位置信息，同时也引入了一个级联锚优化来提高其定位能力，但是反向特征金字塔的存在使得网络参数大大增加，检测速率也随之下降。

本文针对多目标检测提出了新的改进算法，即 Refine-FPN。主要内容包括：

1)引入预测优化模块，其实质上是结合上下文信息的级联网络。该模块首先结合了感兴趣区域的上下文信息，使特征更具有表达能力，然后通过级联的预测网络进行预测，从而增强了网络的定位能力，其中每个阶段预测网络的阈值都不相同。该模块最终实现了多尺度多阶段的预测。

2)采用新型的网络模型 DetNet-59 进行训练，DetNet 模型是专门为目标检测而设计的，该模型在保证特征图分辨率的基础上增加了网络模型的深度，相比于传统的 VGGNet[16]和 ResNet[17]，其准确率得到进一步的提升。

2 网络结构与算法原理

Refine-FPN 算法的整体网络结构如图 1 所示，网络主要由两部分构成。第一部分是用于特征提取的特征金字塔，其结构与文献[18]中的结构一致，左支路是自下而上的特征金字塔，金字塔层分别为{C2，C3，C4，C5，C6}，右支路则是自上而下的特征金字塔，其特征映射集为{P2，P3，P4，P5，P6}，且与左支路一一对应，其中顶层是新增的金字塔层；第二部分是预测优化模块，该模块是在感兴趣区域的基础上结合上下文信息进行级联的目标检测。

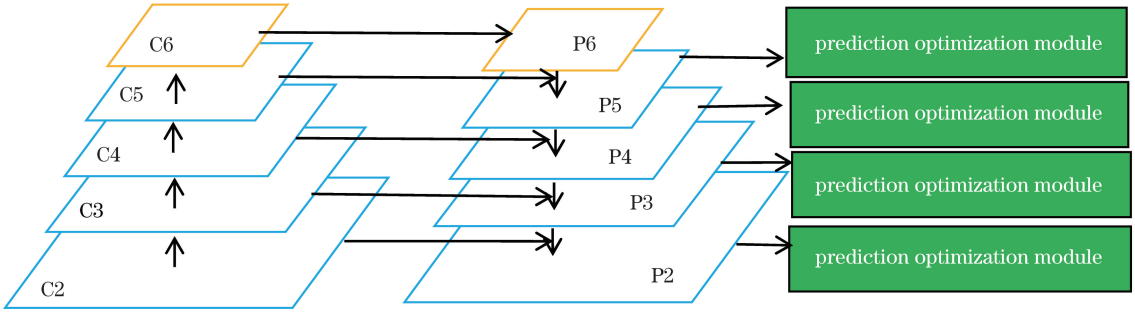

图 1 Refine-FPN 算法的整体网络结构

Fig. 1 Overall network structure of Refine-FPN algorithm

2.1 特征金字塔网络结构

传统的 FPN 算法在进行目标检测时，虽然有较好的检测效果，但仍存在一些不足，如：1)预训练模型与网络不一致。经典的网络模型 ResNet 或者 VGGNet 的步长一般为 32，因此通常会有 5 个阶段(P1～P5)，而在 FPN 中又额外增加了一个阶段 P6，但是该阶段没有被预训练过；2)大目标的回归能力较差。在 FPN 等目标检测网络中，大目标是在较深的特征图上预测，此时目标边界已经模糊，很难准确回归。

本文采用与文献[18]中相同的 FPN 结构，其中前 4 个阶段与原始的 FPN 结构一致，第 5 阶段是在原始 FPN 的基础上进行了改进。将空间分辨率定义为 16 倍的降采样，并使通道数仍为 256，最后在主干网络上新增第 6 阶段，该阶段的通道数以及空间分辨率与第 4、5 阶段相同。总体而言，其基本结构主要是由一个自下而上的支路、自上而下的支路和横向连接组成。

1)自下而上的支路。自下而上的支路是神经网络的前馈计算，其计算是由步长为 2 的多尺度特征映射组成的特征通道。通常有许多特征通道产生相同大小的输出映射，这些通道位于相同的网络阶段，并定义每个阶段为一个金字塔层，将 conv2、conv3、

conv4、conv5 和 conv6 的输出表示为 C2、C3、C4、C5、C6，相对于输入图像，它们的步长分别为 4、8、16、16、16 pixel。

2）自上而下的支路和横向连接。自上而下的支路上采样较高金字塔等级的特征，将其映射为更高分辨率的特征，从而产生一系列分辨率逐渐增大的特征金字塔层，这些特征随后经由横向连接和自下而上支路上的特征进行结合。每个横向连接将两条支路中具有相同空间大小的特征映射进行了合并。自下而上的特征映射具有较低级别的语义，但其可以更精确地定位，而自上而下的特征映射位置信息模糊，但其具有较高的语义信息，二者能很好地互补。最终的特征映射集为{P2，P3，P4，P5，P6}，对应之前的{C2，C3，C4，C5，C6}。

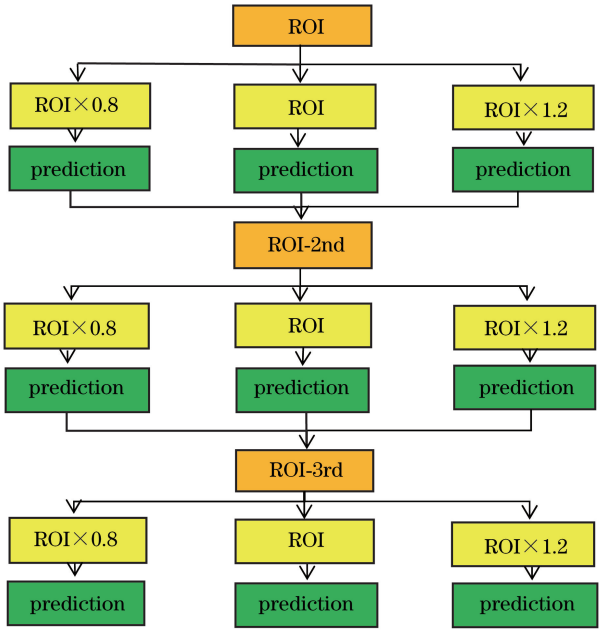

图 2 预测优化模块的基本组成结构

Fig. 2 Basic structure of prediction optimization module

2.2 预测优化模块

预测优化模块的基本组成结构如图 2 所示，该模块主要通过结合感兴趣区域（ROI）的上下文信息获得高语义的特征表达，并通过级联的多阈值预测网络来提高其定位能力。该模块分为三个阶段。在第一阶段中，首先将在特征金字塔中产生的 ROI 区域进行缩放（×0.8，×1，×1.2），由此分为 3 条支路，每条支路后都跟着相同的预测网络，用于计算分类和边界框损失，经过预测网络后，将各个支路的边界框损失相加，并取平均作为最终损失用于边界框回归，然后将产生的新感兴趣区域送入第二阶段进行相同操作。其中每个阶段预测网络的阈值都不同，后一阶段的阈值要比前一阶段的大，最后可得到 3 个阶段的预测结果。在进行网络预测时，采用级联的不同阈值的边界框回归算法，其基本原理如下：

边界框 $b=\{b_x,b_y,b_w,b_h\}$ 包含目标图像 g 的 4 个坐标，其中 b_x,b_y,b_w,b_h 分别代表边界框左上角的横坐标、纵坐标以及边界框的宽和高，边界框回归的任务是通过使用回归器 $f(x,b)$ 将候选边界框 b 回归到目标边界框 t 中。其中训练样本为 $\{t_i,b_i\}$，网络训练的目的是将边界框风险损失降到最低。

$$R_{loc}[f]=\sum_{i=1}^{N}L_{loc}(f(x_i,b_i),t_i), \tag{1}$$

式中：L_{loc} 为一个 $L1$ 损失函数，通过计算最小绝对值偏差得到；R_{loc} 为进行边界框回归得到的损失函数；N 为进行边界框回归的总数；b_i 和 t_i 分别为第 i 个候选边界框及其对应的目标边界框。为了增加对尺寸和位置的回归不变性，L_{loc} 对距离向量 $\boldsymbol{\Delta}=(\delta_x,\delta_y,\delta_w,\delta_h)$ 作出调整：

$$\begin{cases}\delta_x=(t_x-b_y)/b_w\\ \delta_y=(t_y-b_y)/b_h\\ \delta_w=\lg(t_w/b_w)\\ \delta_h=\lg(t_h/b_h)\end{cases}, \tag{2}$$

式中：t_x,t_y,t_w,t_h 分别为目标边界框左上角的横、纵坐标以及目标边界框的宽和高。本文采用的是级联的不同阈值的边界框回归，它是一个级联回归问题，依赖于一系列专门的回归量：

$$f(x,b)=f_T\circ f_{T-1}\circ\cdots\circ f_1(x,b), \tag{3}$$

式中：T 为级联级数的总数；$\circ$ 为迭代操作，用于将每个阶段回归器产生的候选框进行迭代。级联中的每个回归量 f_T 都是针对 T 阶段下的样本分布的优化，从而使得级联回归成为重新采样的过程，最终优化每个阶段的预测。每一阶段 T 都包括一个分类 h_T 和一个针对 IOU 阈值 u_T 优化的回归量 f_T，其中 $u_T>u_{T-1}$。这可以通过公式最小化损失得到：

$$L(x_T,t)=L_{cls}(h_T(x_T),y_T)+\lambda[y_T\geqslant 1]L_{loc}(f_T(x_T,b_T),t), \tag{4}$$

式中：$b_T=f_T(x_{T-1},b_{T-1})$ 表示在 T 阶段下的边界框分布；$h(x)=p(y=k\mid x)$；p 为条件概率；k 为 IOU 的阈值；t 为 x_T 的基本真值对象；$\lambda=1$ 为权衡系数；$[\cdot]$ 为指标函数；y_T 为 x_T 的标签；L_{cls} 为经典的交叉熵损失。

3 实验与结果分析

3.1 实验设置

本文实验模型是基于 PyTorch 的深度学习框

架，实验使用 1 块 GPU，型号为 GTX 1080 Ti，显卡型号为 NVIDIA Quadro P4000。在标准的数据集 Pascal VOC 上进行训练检测，Pascal VOC 数据集共有 20 类，常用的数据集主要是 VOC2007 和 VOC2012。其中 VOC2007 有 9963 张图片，VOC2012 则有 17125 张图片。本实验训练样本采用 VOC07+12 数据集，其中训练集 8218 张，验证集 8333 张，测试集 4952 张。输入图像的尺寸为 600 pixel×1000 pixel，采用随机梯度下降的优化方式，批次大小是 1 块 GPU 上 1 张图片，初始学习率为 0.001，学习率的衰减因子为 10，IOU 阈值设置为 {0.5，0.6，0.75}，max_epoch 设为 12，训练过程共用时 108 h。

3.2 实验结果

实验不仅对算法性能进行横向分析比较，而且还具体分析了预测优化模块。表 1 为 Refine-FPN 与其他经典算法的比较分析。由表 1 可知，Refine-FPN 算法优于其他大多数经典的算法。当 IOU 阈值为 0.5 时，原始的 FPN 算法准确率为 80.5%，Refine-FPN 算法在没有采用在线难例挖掘(OHEM)等优化方法的情况下准确率提升了0.4%，相比 SSD 算法则增加了 3.6%，当阈值增大时，Refine-FPN 算法的准确率进一步提升，相比于原始的 FPN 算法，IOU 为 0.6 时的准确率提升了1.2%，IOU 为 0.75 时的准确率提升了 3.5%。同时，该算法相对于以 DetNet-59 为训练模型的 FPN* 算法，准确率提升了 1.1%，进而说明了预测优化模块的良好性能。但 Refine-FPN 算法的准确率相对于 BPN[15] 算法、Couplenet[19] 算法的准确率还存在差距，这主要是因为在 BPN 中不仅有类似级联网络的锚优化网络，而且在原始特征金字塔的基础上额外增加了新的支路，使得特征具有更强的语义信息，而 Couplenet 算法则采用更深的训练模型。这两种算法的网络参数较多，使得实时检测的速率受限，因此，Refine-FPN 算法具有很好的性能。

表 2 为在标准数据集 VOC2007 上具体检测的部分结果。由表 2 可知，Refine-FPN 算法与 FPN 算法相比，对各个尺度目标检测的准确率都有提升，尤其是针对小目标的检测。例如：对数据集中的大目标(如火车等)进行检测时，FPN 算法的准确率为 87.4%，Refine-FPN 算法的准确率为 87.6%，仅提升了 0.2%。而对于小目标(如椅子、瓶子等)，Refine-FPN 算法相比于 FPN 算法，准确率分别提升了 1.8%和 3.1%，说明此方法对小目标检测有很好的改进效果。

表 3 为 FPN* 算法中级联不同阈值的预测网络的检测结果。由表 3 可知，FPN* 算法初始检测的准确率为 79.8%。当级联了一个 IOU 阈值为 0.6 的预测网络进行训练后，准确率提升了 0.5%，继续级联一个 IOU 阈值为 0.75 的预测网络，准确率仅提升了 0.2%。而当继续级联一个 IOU 阈值为 0.8 的预测网络后，准确率出现了一定的下降，检测效果变差。

表 4 为 FPN* 算法中结合不同上下文信息的检测结果。当结合了 ROI(×0.8)的上下文信息后，准确率提升了 0.2%；当结合 ROI(×1.2)的上下文信息后，准确率提升了0.3%。本文中预测优化模块则同时结合两支路的上下文信息，使得准确率相比于初始的 FPN* 算法提升了 0.5%。表 3 和表 4 的分析说明证实了预测优化模块的良好性能。图 3 为检测结果样本。

表 1 Refine-FPN 与其他经典算法的比较

Table 1 Comparison between Refine-FPN and other classical algorithms

Method	Training set	Backbone	Accuracy /%		
			IOU of 0.5	IOU of 0.6	IOU of 0.75
SSD	VOC07+12	VGG-16	77.3	72.3	61.3
RFCN	VOC07+12	ResNet-101	80.5	73.2	61.8
Faster Rcnn	VOC07+12	ResNet-101	76.4	69.5	57.3
YOLOv2	VOC07+12	Darknet-19	78.6	69.1	56.5
FPN	VOC07+12	ResNet-101	80.5	—	—
Couplenet	VOC07+12	ResNet-101	81.7	—	—
Res101-RFCN-Cascade	VOC07+12	ResNet-101	79.6	—	59.2
BPN512	VOC07+12	VGG-16	81.9	77.6	68.3
Refine-FPN	**VOC07+12**	**Detnet-59**	**80.9**	**74.4**	**65.3**
FPN*	VOC07+12	Detnet-59	79.8	—	—

表 2 VOC2007 上的具体检测结果

Table 2 Specific test results on VOC2007 %

Method	mAP	Bike	Boat	Bottle	TV	Chair	Table	Sheep	Train	Cat
SSD	77.3	83.9	69.6	50.5	76.8	60.3	77.0	77.9	87.6	88.1
RFCN	80.5	89.6	69.0	69.2	79.5	65.4	72.1	79.6	87.1	88.4
Faster Rcnn	76.4	80.7	68.3	55.9	72.0	56.7	69.4	78.6	85.3	85.3
FPN	80.5	80.1	72.9	67.4	72.3	61.5	68.7	78.3	87.4	87.3
Couplenet	81.7	86.0	74.5	72.3	80.1	68.8	75.6	81.9	86.7	88.5
Refine-FPN	**80.9**	**80.4**	**73.8**	**70.5**	**73.1**	**63.3**	**69.8**	**79.1**	**87.6**	**87.8**

表 3 FPN* 算法中级联不同预测网络的比较结果

Table 3 Comparison results when different prediction networks are cascaded in FPN* algorithm

Stage	Accuracy (IOU of 0.5) /%
First stage (IOU of 0.5)	79.8
Second stage(IOU of 0.6)	80.3
Third stage(IOU of 0.75)	80.5
Fourth stage(IOU of 0.8)	80.4

表 4 FPN* 算法中结合不同上下文信息的比较结果

Table 4 Comparison results when different contextual information are combined in FPN* algorithm

Context information	Accuracy (IOU of 0.5) /%
ROI(×1)	79.8
ROI(×0.8, ×1)	80.0
ROI(×1, ×1.2)	80.1
ROI(×0.8, ×1, ×1.2)	80.3

图 3 检测结果

Fig. 3 Test results

4 结 论

提出了一种基于特征金字塔网络改进的算法，该算法采用更适合目标检测的 DetNet-59 网络模型进行训练，通过引入预测优化模块并结合感兴趣区域的上下文信息，获得了更具表达能力的特征信息，同时也提高了算法的定位能力。相比于 FPN 算法，Refine-FPN 算法在保证检测速度基本不变的情况下，实现了更高的准确率。

参 考 文 献

[1] Redmon J, Divvala S, Girshick R, *et al*. You only look once: unified, real-time object detection[C]// The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 26-July 1, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 779-788.

[2] Hua X, Wang X Q, Wang D, *et al*. Multi-objective detection of traffic scenes based on improved SSD[J]. Acta Optica Sinica, 2018, 38(12): 1215003.
华夏, 王新晴, 王东, 等. 基于改进 SSD 的交通大场景多目标检测[J]. 光学学报, 2018, 38(12): 1215003.

[3] Liu W, Anguelov D, Erhan D, *et al*. SSD: single shot MultiBox detector[M]// Leibe B, Matas J, Sebe N, *et al*. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9905: 21-37.

[4] Lin T Y, Goyal P, Girshick R, *et al*. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018: 1.

[5] Girshick R, Donahue J, Darrell T, *et al*. Region-based convolutional networks for accurate object detection and segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(1): 142-158.

[6] Feng X Y, Mei W, Hu D S. Aerial target detection based on improved faster R-CNN[J]. Acta Optica Sinica, 2018, 38(6): 0615004.

冯小雨, 梅卫, 胡大帅. 基于改进 Faster R-CNN 的空中目标检测[J]. 光学学报, 2018, 38(6): 0615004.

[7] Dai J F, Li Y, He K M, *et al*. R-FCN: object detection via region-based fully convolutional networks[C]// Proceedings of the 30th International Conference on Neural Information Processing Systems, December 5-10, 2016, Barcelona, Spain. New York: ACM, 2016: 379-387.

[8] Kong T, Yao A B, Chen Y R, *et al*. HyperNet: towards accurate region proposal generation and joint object detection [C] // The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 26-July 1, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 845-853.

[9] Li H Y, Liu Y, Ouyang W L, *et al*. Zoom out-and-in network with map attention decision for region proposal and object detection [J]. International Journal of Computer Vision, 2019, 127(3): 225-238.

[10] Lin T Y, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection [C] // The IEEE Conference on Computer Vision and Pattern Recognition(CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 2117-2125.

[11] Shrivastava A, Sukthankar R, Malik J, *et al*. Beyond skip connections: top-down modulation for object detection[J/OL]. (2017-09-19)[2019-04-26]. https://arxiv.org/abs/1612.06851.

[12] Kong T, Sun F C, Yao A B, *et al*. RON: reverse connection with objectness prior networks for object detection[C] // The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 5936-5944.

[13] Zhou P, Ni B B, Geng C, *et al*. Scale-transferrable object detection [C] // The IEEE Conference on Computer Vision and Pattern Recognition(CVPR), June 18-22, 2018, Salt Lake City, UT. New York: IEEE, 2018: 528-537.

[14] Zhang S F, Wen L Y, Bian X, *et al*. Single-shot refinement neural network for object detection[C] // The IEEE Conference on Computer Vision and Pattern Recognition(CVPR), June 18-22, 2018, Salt Lake City, UT. New York: IEEE, 2018: 4203-4212.

[15] Wu X W, Zhang D X, Zhu J K, *et al*. Single-shot bidirectional pyramid networks for high-quality object detection [J/OL]. (2018-03-22) [2019-04-26]. https://arxiv.org/abs/ 1803.08208.

[16] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J/OL]. (2015-04-10)[2019-04-26]. https://arxiv.org/abs/1409.1556.

[17] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition [C] // The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 26-July 1, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[18] Li Z M, Chao P, Yu G, *et al*. DetNet: a backbone network for object detection[J/OL]. (2018-04-19)[2019-04-26]. https://arxiv.org/abs/1804.06215.

[19] Zhu Y S, Zhao C Y, Wang J Q, *et al*. CoupleNet: coupling global structure with local parts for object detection[C]// The IEEE International Conference on Computer Vision (ICCV), October 21-26, 2017, Honolulu, Hawaii. New York: IEEE, 2017: 4126-4134.