文章编号 1004-924X(2017)增-0221-07

# 基于预分割和回归的深度学习目标检测

潘 蓉,孙 伟*
(西安电子科技大学 空间科学与技术学院,陕西 西安 710071)

**摘要**:针对高分辨率图像中的小目标检测检测难的问题,结合基于候选区域的目标检测方法中的感兴趣区域提取策略和基于回归的目标检测算法中的回归策略,提出了基于预分割和回归的深度学习目标检测算法。因此使用四叉树对原始图像兴趣目标提取,使用基于回归的目标检测方法对感兴趣区域的目标进行细致的再定位和分类。与传统的 Fast-RCNN 方法和 YOLO 系列的基于回归的深度学习方法相比,基于四叉树的深度学习的目标检测算法在精度和速度上有明显优势。经过实验结果分析表明,与 Fast-RCNN 相比,Quad-ssd 算法在目标检测时精度提高了 6.5%,达到了 74.9%,检测速度大幅提高,达到 45 帧每秒,完全满足实时性的要求。

**关 键 词**:四叉树;深度学习;小目标检测;Quad-ssd
**中图分类号**:TP752.1 **文献标识码**:A **doi**:10.3788/OPE.20172513.0221

# Deep learning target detection based on pre-segmentation and regression

PAN Rong, SUN Wei*

(*School of Aerospace Science and Technology, Xidian University, Xi'an* 710071, *China*)
* *Corresponding author, E-mail: wsun@xidian.edu.cn*

**Abstract**: Aiming at the problem of difficult small target detection in high resolution images, combined with region-of-interest (ROI) extraction strategy in target detection method based on candidate region and regression strategy in target detection algorithm based on regression, deep learning target detection algorithm based on pre-segmentation and regression (Quad-ssd) was proposed. As fast-RCNN series implement image location and classification separately, small targets could be detected but detection time was too long. YOLO series method used regression method to implement classification and location for targets in images at the same time. As only high-level features were used, detection accuracy for small target was not enough. Therefore, quad tree was used to extract interest target of original images, and target detection method based on regression was used to implement detailed relocation and classification for targets in interested region. Compared with traditional Fast-RCNN method and deep learning method based on regression of YOLO series, target detection algorithm of deep learning based on quad tree has obvious advantages in accuracy and speed. The experimental results show that

**收稿日期**:2017-06-01;**修订日期**:2017-07-04.
**基金项目**:国家自然科学基金青年基金资助项目(No. 61201290),国家自然科学基金面上资助项目(No. 61671356),中央高校基本科研业务资助项目(No. JB161301,No. JBG161307)

compared with Fast-RCNN, accuracy of Quad-ssd algorithm is improved by 6.5% and reaches 74.9% at the time of target detection. The detection speed is improved greatly; reaching 45 f/s, and can satisfyrequirements of timeliness.

**Key words**: quadtree; deep learning; small target detection; Quad-ssd

## 1 引　言

目标检测是计算机视觉领域中一个富有挑战性的课题,其核心任务是在静态图片或者视频中使用某种目标识别算法和搜索策略,获取特定目标在图像或视频中的位置和类别。目前目标检测的方法主要分为基于特征及机器学习的目标检测算法和基于深度学习的检测方法。其中基于特征及机器学习的方法是通过对目标进行区域选择,特征提取、分类器分类等过程实现目标检测。区域选择是为了对目标的位置进行定位,一般通过滑动窗口对整幅图像进行遍历选择可能存在目标的图像边框,但时间复杂度太高,冗余窗口过多,直接影响后续的特征提取和分类的速度和性能。特征提取中常用的特征有 Haar[1] 小波特征 、HOG[2] 特征、SIFT[3] 特征和混合特征等,由于图像的光照条件,背景和目标的形态等的多样性,对特征的鲁棒性要求比较高,提取的特征好坏直接影响分类的准确性。传统的分类器主要包括支持向量机 SVM[4] 和迭代器 Adaboost[5]。由于是针对某个特征的识别任务,且数据量不大,模型泛化能力差,很难在实际应用中对目标精准识别。从 2014 年开始,基于深度学习的目标检测算法取得了重大的突破,克服了传统的目标检测算法中的缺点。目前主流的基于深度学习的目标检测算法主要分为两类:基于候选区域(Region Proposal)的深度学习目标检测算法和基于回归的深度学习目标检测算法。基于候选区域的目标检测算法的代表是 R Girshick 提出的 R-CNN 算法,该算法的检测框架结合候选区域(Region Proposal)和卷积神经网络(CNN)分类。由 R-CNN[6] 逐步优化提速产生了 SPP-NET[7], Fast R-CNN[8] 和 Faster R-CNN[9],目标检测的精度和速度都由很大的提高,但由于此类方法进行目标检测时分为定位和分类两个步骤且定位耗时太长,因此还不能实时地进行目标检测。后来有人提出了基于回归方法的深度学目标检测算法。具有代表性的基于回归方法的深度学习目标检测算法:YOLO[10] 和 SSD[11],这类算法主要是通过回归的方法直接从图像中回归出目标的位置和类别,这种方法的目标检测速度大大加快,可以达到目标实时检测的要求,但对输入图像的大小有严格要求并且目标检测最后的目标的位置定位精度略差,无法检测图像中的小目标。文献[11]中的 SSD300 要求输入图像尺寸必须是 300×300,SSD500 要求输入图像必须是 500×500。针对高分辨率的图像中的小目标检测,本文提出了基于四叉树和卷积神经网络的目标检测算法,首先通过快速四叉树分割粗略提取出图像中的感兴趣区域,其次将感兴趣区域输入到检测网络中,最后基于回归的卷积神经网络对其进行目标检测并将结果映射到输入图像中。本文提出的 quad-ssd 算法模型不受图像的大小的限制。在保证精度的同时速度也可以达到实时性的要求。

## 2 Quad-ssd 算法

Quad-ssd 算法可以分为两部分:基于四叉树的感兴趣区域提取和基于回归的目标检测。

### 2.1 彩色图像四叉树分割算法

彩色图像的四叉树分割目的是将原始图像(如图 1(a)、(d)、(g)、(j))逐步分成小块,操作的原理是将具有一致性(像素间的灰度差值小于给定的阈值)的像素分到同一小块中。

四叉树分解过程如下:

(1)令 R 表示当前进行分解的图像区域,并选择一个属性 Q(区域的像素间满足一致性标准)。

(2)如果有 Q(R)=FALSE ,则将该图像区域分割为四个象限区域。

(3)分割后的任意一个区域,如果仍有 Q(R)=FALSE,则将该象限区域再次细分为四个象限区域

(4)以此类推,直到图像中的所有区域都满足一致性标准才停止。

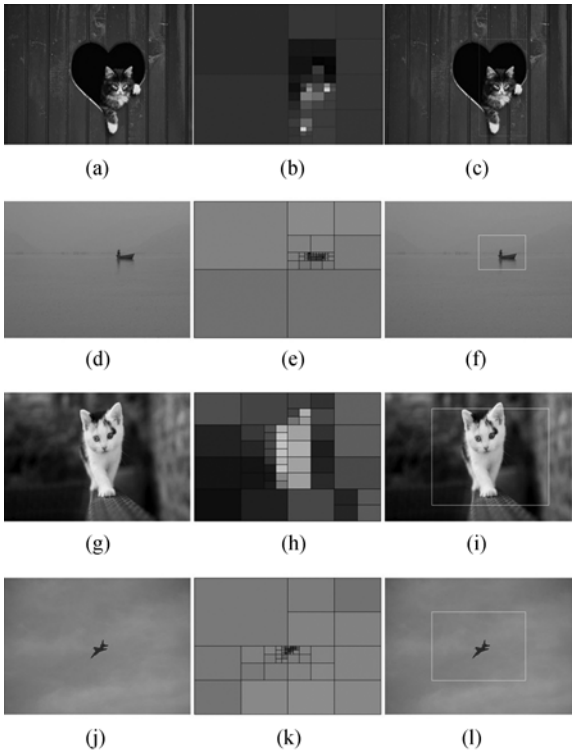


图 1　彩色图像四叉树提取感兴趣区域

Fig. 1　Color image quadtree extraction region of interest

最后，四叉树分解的结果可能包括多种不同尺寸的方块如图 1 中的(b)、(e)、(h)、(k)所示。通过对合适深度的分割图像进行分析，得到分割后的感兴趣区域如图 1 中的(c)、(f)、(i)、(l)所示，图中白色边框为提取的感兴趣区域。为后续的目标检测提供了感兴趣区域，同时根据感兴趣区域的尺寸选择合适的深度学习网络结构，有利于小目标的检测和识别。通过四叉树预分割，可以充分利用图像的色彩信息对图像进行感兴趣区域提取。同时四叉树提取的感兴趣目标区域使得后面的深度卷积神经网络提取的特征能恰当的结合目标周围的上下文信息，而不是像 YOLO 使用整幅图像的信息，避免了过多的特征提取。对于一幅图像中的目标，显然使用目标周围的特征信息更为合理。

### 2.2　深度卷积神经网络(CNN)

基于候选区域的深度学习框架对小目标的检测效果比较好，但目前现有此类框架的代表方法速度达不到实时性的要求，而基于回归方法的深度学习框架在目标检测速度上完全符合实时性要求。SSD 结合 Faster-RCNN 算法中的 anchor 机制和 YOLO 算法中的使用回归的方法进行目标类别和位置预测机制，在速度和精度上都明显优于 Faster-RCNN 和 YOLO 算法。本文用于目标检测的卷积神经网络是一种基于回归方法的快速 ssd 网络，其模型如图 1 所示。Ssd 网络是基于一个前向传播的卷积神经网络，用于生成一系列固定尺寸的边框以及每个边框中存在目标的概率，最后经过一个非极大值抑制得到最终的预测结果(边框位置和类别)。本文使用的 ssd 框架的模型由基础网络和辅助结构组成，基础网络是缩减了的 VGG-16 网络，本文使用前 conv5_3 层，用于高分辨率图像分类的标准框架。

辅助结构位于缩减了的 VGG-16 网络之后，用于生成多种尺度的特征图。辅助结构如图 2 所示，包括两部分，一是 VGG-16 网络中的两个全连接层转换后的卷积层 conv6 和 conv7，另一是新增的 4 个递减的卷积层 conv8_2、conv9_2、conv10_2、conv11_2。在多尺度的特征图上进行边框预测和类别预测的精度高于 YOLO 中的在一种尺度的特征图上进行预测的精度。SSD 框架是在特征图上使用卷积滤波器产生预测结果，而 YOLO 框架则是使用一个全连接层来替代辅助结构中新增的卷积层；SSD 框架中预测的是边框相对于真实边框的相对位置，而 YOLO 框架预测的是边框的位置，因此前者在不同尺度的特征层上进行预测的精度高于后者，尤其是对于小目标。在卷积神经网络中越是靠前的特征层包含的图像细节信息就越多，因此不能只使用尺寸较小的特征层做检测，否则会丢失图像中的大量细节信息。

在辅助结构中新增的卷积层，使用一系列卷积滤波器产生相应的固定大小的预测结果[12]。假如特征层的尺寸为 $m\times n$，通道数为 $p$，则使用的卷积核大小为 $3\times3\times p$。每个特征层上的每个特征点对应 k 个默认边框，物体的类别数为 classes，则每一个特征层都需要使用 k(classes+4)个这样的卷积滤波器。对于一个 $m\times n$ 的特征层，则得到 $(m\times n)\times k\times(\text{classes}+4)$ 个输出。默认边框类似于 Fast-RCNN 中的 Anchor boxes，但是不同于 Fast-RCNN，默认边框被用在了不同分辨率的特征层，在不同的特征层上默认边框的长宽比可以不同，这样有利于高效离散化可能的输出边框的空间[13]。

### 2.3 模型训练

基于候选区域的方法通常是先训练 RPN 网络，之后再将 RPN 网络训练好的参数输入到 Fast-RCNN 网络进行训练直到网络收敛，最后将所有参数输入到整个网络模型，训练过程很复杂且不易调整[13-15]。不同于基于候选区域的方法，本文使用一种端到端的训练方法。首先将训练图像中的真实边框通过匹配策略对应到固定的输出边框上，这些输出边框是事先定义好的一系列固定大小的输出边框。将训练图像中的真实边框与固定输出的边框对应之后，就可以端对端的进行损失函数的计算以及反向传播的计算更新了，这种端到端的训练模式更简单更易于调整网络。

本文在训练时的目标损失函数源自 Multi-Box 的目标损失函数，将其扩展到多类别目标。本文中的目标损失函数由定位损失函数和分类损失函数两部分组成如公式(1)所示，其中定位损失函数是 Fast R-CNN 中 $\text{smooth}_{L1}$ Loss，分类损失函数是 softmax 的损失函数。

$$L(x,l,c,g)=\frac{1}{N}(L_{conf}(x,c))+\alpha L_{loc}(x,l,g). \tag{1}$$

其中权重系数 $\alpha$ 通过交叉验证设置为 1，N 为与真实标记的边框相匹配的默认边框个数。定位损失函数如式(2)所示，其中 $\text{smooth}_{L1}$ 为 Fast R-CNN 中的 $\text{smooth}_{L1}$ Loss，用在预测边框 $l$ 与真实标记的边框 $g$ 的参数(即中心坐标位置，边框宽和高)中，最终回归得到预测边框的中心位置，以及预测边框宽和高。

$$L_{loc}(x,l,g)=\sum_{i\in Pos}^{N}\sum_{m\in\{cx,cy,w,h\}}x_{ij}^{k}\,smooth_{L1}(l_i^m-\hat{g}_j^m). \tag{2}$$

其中 $\hat{g}_j^{cx}=(g_j^{cx}-d_i^{cx})/d_i^w$，$\hat{g}_j^{cy}=(g_j^{cy}-d_i^{cy})/d_i^h$，$\hat{g}_j^w=\log\left(\frac{g_j^w}{d_i^w}\right)$，$\hat{g}_j^h=\log\left(\frac{g_j^h}{d_i^h}\right)$。

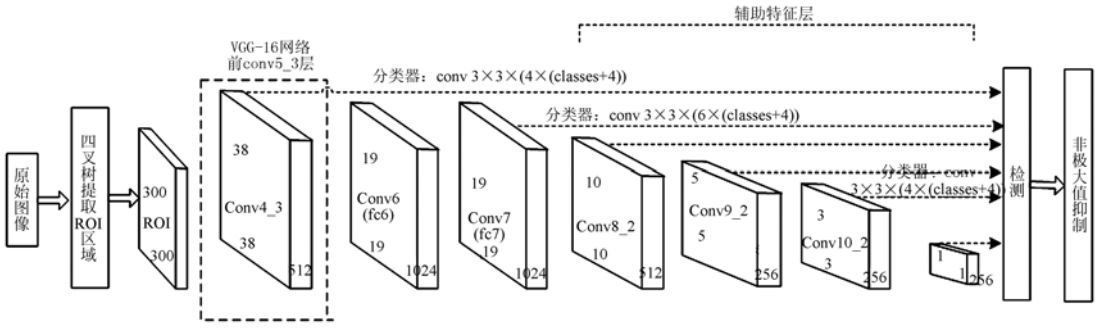


图 2　quad-ssd 算法结构框图

Fig. 2　Block diagram of Quad-ssd algorithm

在分类的损失函数中，用 $x_{ij}^p=1$ 表示第 $i$ 个默认边框与类别 $p$ 的第 $j$ 个真实标记的边框相匹配，反之，$x_{ij}^p=0$。根据匹配策略，则 $\sum x_{ij}^p>1$，即第 $j$ 个真实标记的边框，有可能有多个默认边框与其相匹配。在多类别的置信度 c 上的置信度损失函数是 softmax 损失函数：

$$L_{conf}(x,c)=\sum_{i\in Pos}^{N}x_{ij}^p\log(\hat{c}_i^p)-\sum_{i\in Neg}\log(\hat{c}_i^0), \tag{3}$$

其中 $\hat{c}_i^p=\dfrac{\exp(c_i^p)}{\sum_p\exp(c_i^p)}$.

在深度卷积神经网络中，随着卷积网络的深入，特征图的尺寸越来越小，越顶层的特征图，提取的特征尺度不变性和平移不变性越好[16]如图 3(d)中的特征图所示，而底层的特征图如图 3(a)所示，拥有更多的细节信息，比如图像的边缘信息。同时使用顶层特征和底层特征进行预测，既可以保持原始图像中的细节信息又可以保留一些鲁棒性强的特征。

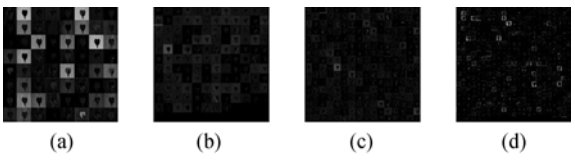


图 3　Quad-ssd 网络中的特征层可视化图

Fig. 3　Feature layer visualization graphs in Quad-ssd networks

输出的特征图上的一个节点，对应输入图像上尺寸的大小称为感受野如图 4 中左边的输入图像所示，网络中不同的卷积层有着不同的感受野。在本文中，特征图中特定的位置，来负责图像中特定的区域，以及物体特定的尺寸，因此默认边框不必与每一层的特征图中的感受野对应。用来做预测的特征图中，每一个特征图中默认边框的尺寸大小的取值范围为 0.2～0.95，每个默认边框使用不同的长宽比 $\left\{1,2,3,\frac{1}{2},\frac{1}{3}\right\}$，在结合特征图上，所有不同尺度、不同长宽比的默认边框预测多个预测边框如图 4 所示，其中包含了目标的不同尺寸、形状。在生成的一系列的预测边框中，会产生很多个符合真实标记边框的预测边框，但不符合的预测边框远多于符合的，这会造成正预测边框和负预测边框之间的不均衡，训练时网络难以收敛[17-18]。因此，先将每一个目标位置上对应的负预测边框按照边框的置信度大小进行排序，选择最高的几个，保证最终的正负预测边框的比例为 3:1。这样的正负预测边框比例使得网络在训练时更稳定。

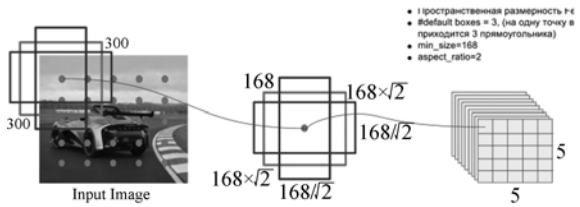


图 4　感受野

Fig. 4　Receptive field

## 3　实验结论与分析

本文使用的是基于 ILSVRC CLS-LOC 上预训练好的 VGG-16 网络。使用 pascal voc2007 和 pascal voc2012 的训练数据集做训练，在 pascal voc2007 的测试集上进行测试。训练样本的主要硬件配置如下：处理器为 Intel® Core™ i7-6700K CPU@ 4.00 GHz × 8，内存 15.6 GB，GPU 卡：TITAN X (Pascal)/PCIe/SSE2。在测试集上进行测试的结果如表 1 所示。

**表 1　各种深度学习算法在测试集上的测试结果**

Tab. 1　Results of the various depth learning algorithms on the test set

| 算法模型 | 数据集 | 检测精度 | 检测速度 |
|---|---|---|---|
| Fast-RCNN | 07＋＋12 | 68.4 | 3 |
| Faster-RCNN | 07＋＋12 | 70.4 | 5 |
| YOLO | 07＋＋12 | 57.9 | 47 |
| SSD300 | 07＋＋12 | 72.4 | 59 |
| Quad-ssd | 07＋＋12 | 74.9 | 45 |

基于候选区域的方法例如 Fast-RCNN，Faster-RCNN 等，虽然精度很高，但是检测时间过长，不能满足实时检测的要求，基于回归的方法比如 YOLO、SSD300 等，虽然可以满足实时性检测的要求，但是目标检测精度过低，对于小目标的检测效果一般，相比之下本文的算法结合候选区域和回归的方法进行目标检测，在满足精度的同时，实时性也可以达到要求，多层特征图和默认边框的设计对于小目标的检测效果更好。单幅图像的测试效果如图 5 所示。图(a)、(b)、(c)、(d)为 Fast-rcnn 算法的测试结果；图(e)、(f)、(g)、(h)为 SSD300 算法的测试结果；图(i)、(j)、(k)、(l)为 Quad-ssd 算法的测试结果 Faster-RCNN 处理的结果分类精度和定位都很准确，但是耗时很长，SSD300 对于一些小目标检测是无效的，相比之下本文的 Quad-ssd 的检测结果精度和时间都比前两者高。通过上述实验可以看出通过四叉树提取感兴趣区域，继而做目标检测时的结果边框更

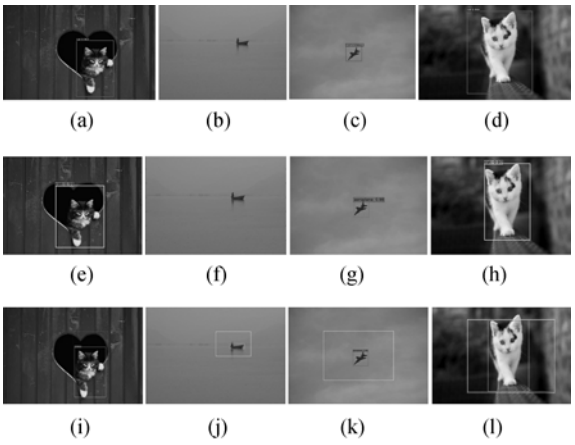


图 5　单幅图像实验结果

Fig. 5　Single image experiment results

贴合目标，分类精度也更准确。

上述实验结果表明：在进行目标检测前进行感兴趣区域提取，更利于小目标的定位和分类；由于基于回归的深度学习方法做目标检测时同时进行定位和分类，因此检测时间可以满足实时性的要求。本文所提算法既有类似于基于候选区域的目标检测方法中的感兴趣区域提取，同时又使用基于回归的方法对感兴趣区域的目标进一步进行高精度的定位和分类，因此本文的方法可以在满足精度的同时也可以保证实时性。

## 4 结　论

本文提出的基于四叉树分割的深度学习的目标检测算法，使用四叉树对原始彩色图像进行感兴趣区域提取，结合基于回归的目标检测方法对感兴趣区域中的目标进行更高精度的检测，从而提高目标检测算法的精度和速度。四叉树提取感兴趣区域可以充分利用图像的颜色信息，提出的模型对输入图像的尺寸并没有要求，YOLO和SSD的方法都需要对输入的图像进行缩放，对于一些大型图像缩放的过程会遗失很多细节信息。基于回归的目标检测方法使用多种尺度的特征图并结合多种默认边框，使得最后得到的检测精度明显高于只使用一种特征图的目标检测算法。通过实验也验证了提出的基于四叉树分割的深度学习目标检测算法在精度和速度上的优势。

**参考文献：**


[1] 宋燕星，袁峰，丁振良，等. 使用形态Haar小波法检测目标感兴趣区域[J]. 光学 精密工程，2009，17(7)：1752-1758.

[2] TAIGMAN Y，YANG M，RANZATO M，*et al*.. DeepFace：Closing the Gap to Human-Level Performance in Face Verification[C]. *IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society*，2014：1701-1708.

[3] MA X，GRIMSON W E L. Edge-based rich representation for vehicle classification[C]. *Tenth IEEE International Conference on Computer Vision. IEEE*，2005：1185-1192 Vol. 2.

[4] KAZEMI F M，SAMADI S，POORREZA H R，*et al*.. Vehicle Recognition Using Curvelet Transform and SVM[C]. *International Conference on Information Technology. IEEE*，2007：516-521.

[5] FREUND Y，SCHAPIRE R E. A desicion-theoretic generalization of on-line learning and an application to boosting[C]. *European Conference on Computational Learning Theory. Springer Berlin Heidelberg*，1995：23-37.

[6] FELZENSZWALB P F，GIRSHICK R B，MCALLESTER D，*et al*.. Object detection with discriminatively trained part-based models.[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*，2010，32(9)：1627-45.

[7] SZEGEDY C，TOSHEV A，ERHAN D. Deep Neural Networks for Object Detection.[C]. *Advances in Neural Information Processing Systems*[S. *l*.]：*NIPS Press*，2013：1673-1675.

[8] SERMANET P，EIGEN D，ZHANG X，*et al*.. OverFeat：Integrated Recognition，Localization and Detection using Convolutional Networks[J]. *Eprint Arxiv*，2013.

[9] GIRSHICK R，DONAHUE J，DARRELL T，*et al*.. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[J]. 2013：580-587.

[10] UIJLINGS J R，SANDE K E，GEVERS T，*et al*.. Selective Search for Object Recognition[J]. *International Journal of Computer Vision*，2013，104(2)：154-171.

[11] HE K，ZHANG X，REN S，*et al*.. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*，2015，37(9)：1904.

[12] GIRSHICK R. Fast R-CNN[J]. *Computer Science*，2015.

[13] REN S，HE K，GIRSHICK R，*et al*.. Faster R-CNN：Towards Real-Time Object Detection with Region Proposal Networks[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*，2016，39(6)：1137.

[14] REDMON J，DIVVALA S，GIRSHICK R，*et al*.. You only look once：unified，real-time object detection[J]. *IEEE Computer Society*，2015：

779-788.

[15] LIU W, ANGUELOV D, ERHAN D, *et al.*. SSD: Single Shot MultiBox Detector[C]. *European Conference on Computer Vision. Springer International Publishing*, 2016:21-37.

[16] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation [C]. *Computer Vision and Pattern Recognition. IEEE*, 2015:3431-3440.

[17] FELZENSZWALB P F, GIRSHICK R B, MCALLESTER D, *et al.*. Object detection with discriminatively trained part-based models.[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2010, 32(9):1627-1645.

[18] RUSSAKOVSKY O, DENG J, SU H, *et al.*. ImageNet Large Scale Visual Recognition Challenge [J]. *International Journal of Computer Vision*, 2015, 115(3):211-252.

**作者简介：**

**孙　伟**(1980－)，男，安徽砀山人，博士，教授，博士生导师。主要研究方向为高性能视觉信息计算及嵌入式系统设计。E-mail：wsun@xidian.edu.cn

**潘　蓉**(1992－)，女，山西运城人，研究生，研究方向为数字图像处理。E-mail:1621980248@qq.com