基于卷积神经网络的目标检测模型综述

许必宵[1,2],宫 婧[2,3],孙知信[2,3]
(1. 南京邮电大学 物联网学院,江苏 南京 210003;
2. 南京邮电大学 宽带无线通信与传感器网络技术重点实验室,江苏 南京 210003;
3. 南京邮电大学 现代邮政学院,江苏 南京 210003)

摘 要:目标检测一直是计算机视觉领域中的研究热点。随着深度学习技术的迅猛发展,基于卷积神经网络的目标检测模型逐渐被广泛关注。文中主要对基于卷积神经网络的目标检测模型的现状进行综述。首先,介绍了目标检测的相关基础,特别罗列了一些目标检测模型中常用的卷积神经网络结构,也介绍了检测模型常用的梯度下降法训练方式。然后,重点从候选区域和回归方法两类对近年来提出的优秀模型进行综述,候选区域一类也创新地使用特征尺度进行区分,说明了多尺度特征能够有效提高小尺度目标检测精度。对于每一类检测模型,根据同一数据集上的检测结果分析这些模型的优势与缺陷,最后根据分析的结果总结一些基于卷积神经网络的目标检测模型的优化方案。

关键词:卷积神经网络;目标检测;深度学习;计算机视觉
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2019)12-0087-06
doi:10.3969/j.issn.1673-629X.2019.12.016

A Survey of Object Detection Models Based on Convolutional Neural Networks

XU Bi-xiao[1,2],GONG Jing[2,3],SUN Zhi-xin[2,3]
(1. School of Internet of Things,Nanjing University of Posts and Telecommunications,Nanjing 210003,China;
2. Key Laboratory of Broadband Wireless Communication and Sensor Network Technology,Nanjing University of Posts and Telecommunications,Nanjing 210003,China;
3. School of Modern Posts,Nanjing University of Posts and Telecommunications,Nanjing 210003,China)

Abstract:Object detection has always been a research hotspot in the field of computer vision. With the rapid development of deep learning technology,the object detection model based on convolutional neural network is widely concerned. We mainly review the current status of object detection models based on convolutional neural networks. First of all,we introduce the relevant basis of target detection,especially the convolutional neural network structure commonly used in some object detection models,and also introduce the gradient descent training method commonly used in detection models. Then,we summarize the excellent models proposed in recent years from region-based and region-free and compare the test results. The region-based models are distinguished with feature scales intelligently,which shows that multi-scale features can effectively improve the accuracy of small-scale object detection. For each type of detection model,we analyze the advantages and disadvantages of these models based on the results on the same data set. Finally,based on the analysis results,some optimization schemes based on the convolutional neural network are proposed.

Key words:convolutional neural network;object detection;deep learning;computer vision

0 引 言

目标检测是一种利用算法在图像中搜索感兴趣目标对象的计算机视觉技术[1]。检测过程主要分为两步,首先对目标类别进行检测,然后使用边框对目标所在位置进行标注[2]。图像按照像素矩阵存储,需从中抽象出目标类别和边框位置有关的语义信息才能进行目标检测,这种语义信息即图像的特征。由于传统特征提取方法泛化能力差并且精度较低,随着卷积神经

收稿日期:2018-12-21 **修回日期**:2019-04-23 **网络出版时间**:2019-06-27
基金项目:国家自然科学基金(61373135);江苏省研究生科研与实践创新计划项目(KYCX17_0775)
作者简介:许必宵(1993-),男,在读硕士,工程师,研究方向为目标检测技术;宫 婧,博士,副教授,研究方向为深度学习、计算机视觉等;孙知信,博士后,教授,通信作者,研究方向为信息安全、人工智能与计算机视觉。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190627.1105.050.html

网络(convolutional neural networks,CNN)的出现,特征提取方法开始被深度学习替代。

CNN 通过在具有标签的训练数据集上进行学习,形成能够提取数据特征的复杂网络结构,然后就可以提取各种相似数据的特征[3]。一开始,CNN 被设计用来识别手写签字即 LeNet-5[4],该模型首次使用梯度反向传播算法(back propagation,BP)进行监督式训练。2010 年后,随着计算机 GPU 性能的提升,CNN 在图像处理中变得不可替代,在计算机视觉场景中应用广泛[5]。

基于卷积神经网络的目标检测模型使用卷积神经网络提取图像特征,然后根据特征进行目标分类和边框回归,无需复杂的人工特征设计过程。所以深度学习和 CNN 的提出不仅促进了神经网络学的发展,更是促进目标检测等计算机视觉技术的发展。

1 基础介绍

本节将介绍目标检测模型的发展历史,其中会着重介绍目标检测模型中优秀的 CNN 结构和训练方法。

1.1 传统的目标检测模型

传统目标检测模型的主要流程是先对输入图像进行预处理,即去噪、增强以及剪裁伸缩等,然后利用滑动窗口方法对图像进行候选区域筛选,接着使用特征提取算法包括 Sift、HOG、DPM 等对候选区域进行特征提取[6],最后利用分类算法对提取的特征进行分类。分类方法包括 AdaBoost、SVM[7] 等,分类结果用来判断候选区域中的目标所属类别,通过目标类别再对目标进行边框回归。传统的目标检测模型有很多缺陷,例如人工特征存在性能问题,不同特征需要选择合适的分类器,因此鲁棒性不是太强。

相对于传统的目标检测模型,基于深度学习的目标检测模型具有更强大的特征表达能力,泛化能力强、鲁棒性较好[8]。

1.2 卷积神经网络结构

目前深度学习技术被用于计算机视觉领域主要依靠三种神经网络结构:卷积神经网络、深度信念网络(deep belief network,DBN)和堆叠自动编码器(stacked belief network,SAE)[9]。而 CNN 由于其精度高和速度快的优势相比其他两种使用更为广泛[10]。

目前在目标检测模型中应用广泛的 CNN 结构有很多,2012 年 Hinton 等提出 Alex-Net[11] 模型,主要用于图像分类领域,并在 ImageNet 数据集上将结果错误率降低到 15.3%,一举成为 ILAVRC 挑战赛第一名。再后来,牛津大学提出的 VGGNet[12] 由于窄而深的标准卷积结构成为主流卷积神经网络结构,实验证明其迁移学习能力很强;用于目标检测模型的还有两种主流卷积神经网络 ResNet[13] 和 GoogLeNet[14],ResNet 由于其很深的层次结构同时使用残差节点使其精度有了较高的提升,GoogLeNet 复杂的 Inception 结构具有多卷积核的特征,结果表明其能够有效提升计算资源利用率。后来为了提升模型应对特殊场景的性能,也有其他卷积神经网络结构被应用,例如 DarkNet[15]、DenseNet[16] 等等。

1.3 目标检测模型的训练

基于深度学习的目标检测模型中卷积核权重 W 和偏置项 b 等参数是通过梯度下降和反向传播算法求解的。这种方法的主要思想是利用输出和标签值之间的损失误差在 CNN 结构中以梯度下降法从输出层向输入层逐层传播,然后根据梯度值来迭代优化参数。例如式 1 表示误差函数:

$$J(W,b;x,y)=\frac{1}{2}\|h_{w,b}(x)-y\|^2 \tag{1}$$

训练目的是使得误差代价越来越小,因此更新参数时都会以误差函数 J 的梯度方向进行迭代,具体如式 2:

$$\begin{cases} W=W-\eta\dfrac{\partial}{\partial W}J(W,b;x,y) \\ b=b-\eta\dfrac{\partial}{\partial b}J(W,b;x,y) \end{cases} \tag{2}$$

其中,η 是学习效率。基本所有目标检测模型中的卷积神经网络训练都是采取这种方法。

2 基于卷积神经网络的目标检测模型

本节将具体介绍近年来提出的基于卷积神经网络的目标检测模型,可以分成候选区域和回归方法两类。

2.1 基于候选区域的目标检测模型

2.1.1 基础单尺度特征模型

这类检测模型真正起源应该是由 Girshick 等提出的 R-CNN[17],后来成为此类模型的基准,模型结构如图 1 所示,基本结构使用 AlexNet。R-CNN 中使用了一种 selective search (SS)的方法提取候选框,相比滑动窗口,这种先分离再聚合的筛选方式获取候选框更准确。另外,R-CNN 采用独立训练的 SVM 分类器和线性回归模型分别进行目标分类和边框预测。在 VOC 2012 数据集上的实验结果相比当时最佳检测模型的 mAP 提高了 30%。

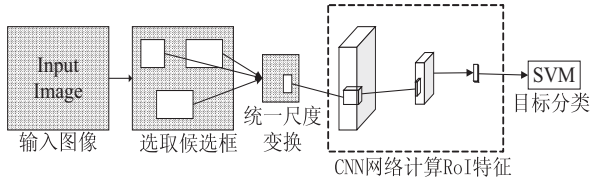

图 1 R-CNN 网络结构示意

R-CNN 同样具有很多问题,例如固定的输入尺

寸使得图像需要被拉伸和裁剪,使得细节特征受到损害;多环节流程使得模型精度由于环节之间接口复杂而降低;另外每个候选框都单独进行特征提取,使得该网络模型的实时性大大降低。针对 R-CNN 的缺陷,一系列基于 R-CNN 改进的模型被提出。

He 等[18]从卷积可视化的角度去分析如何避免图像尺度对卷积计算的影响。SPP-Net 因此诞生,以 ZF-Net 作为特征提取网络,引入图像金字塔的概念,使用一种多卷积核的金字塔特征池化方式。相比 R-CNN,SPP-Net 可以实现不同尺度输入的多层卷积计算,不需要进行裁剪伸缩等预处理操作,保存了底层一些细节特征。

Girshick 等结合 R-CNN 和 SPP-Net 的优点,提出直接将目标分类和边框回归都交给 CNN 来完成,形成端到端的目标检测模型 Fast R-CNN[19]。Fast R-CNN 设置了一个多目标损失函数,使得分类和回归两个任务共享卷积特征,不需要分为多个阶段,节省了存储空间和计算时间。和 SPP-Net 一样,采取了最大值池化层 ROI Pooling 将卷积特征变成固定大小的 ROI 特征,送入全连接层。Fast R-CNN 加快了检测速度,在 VOC 2007 上的结果比 SPP-Net 也提高了 4%。

Ren 等[20]提出了 Faster R-CNN 模型,汲取 Fast R-CNN 的经验继续将候选框推荐部分也由 CNN 完成,图像输入到 VGG 中获取特征后通过一个区域推荐网络(region proposal network,RPN)获取候选框,然后将每个候选框对应的特征送到池化层。RPN 利用滑动窗口和不同尺度的 anchor 筛选出 1K ~ 2K 的候选框。Faster R-CNN 利用 RPN 不仅加快了检测速度,在 VOC 2012 上 mAP 达到了 70%。但是,Faster R-CNN 仍然还是由多个阶段构建而成。

以上的模型都是使用单尺度特征即最高卷积层的输出作为 Feature Map,也有很多文献提出类似的模型。文献[21]中提出一种 MR-CNN 模型,对 SPP-Net 改进后提出了一种多区域形式的 CNN 结构,对每个 proposal 进行尺度形变获取不同的 region 来增强特征,从而提高检测精度;文献[22]中提出 OHEM 算法,模型根据输入样本的损失进行评估,筛选出对结果影响较大的难样本,然后融入到随机梯度下降法中进行训练。实验证明这种算法能够有效提高 Fast R-CNN 的检测精度。

Dai 等[23]提出了 R-FCN 模型,他们认为 Faster R-CNN 第一部分卷积提取网络是 ROIs 共享的,存在位置不敏感性问题,第二部分是每个 proposal 的分类和回归层,每个 ROI 都不共享。因此,模型检测精度较低,而且这样做不能充分利用分类网络。R-FCN 模型能够解决位置敏感问题。

R-FCN 的关键部分如图 2 所示。在预训练最后一个卷积层获得 Feature Map 后进行 3 个分支流程,首先在该 Feature Map 上使用 RPN 获取 ROIs,然后从 Feature Map 上获取一个 $K \times K \times (C+1)$ 维的位置敏感得分映射(position-sensitive score map)用来分类;接着,再从 Feature Map 上获得一个 $K \times K \times 4$ 维的位置敏感得分映射用来回归;最后,在两个映射上面分别执行位置敏感对应池化方法获得对应的类别和位置信息。

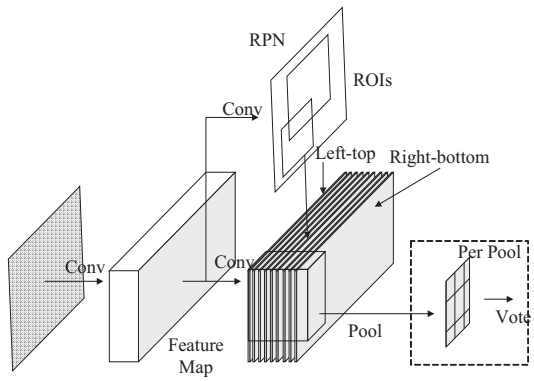

图 2　R-FCN 结构

R-FCN 考虑了目标位置对检测精度的影响,利用敏感位置得分来求解最后的结果,虽然计算得分矩阵花费了较多时间,但有效克服了位置平移带来的误差,在 VOC2012 上 mAP 最高达到 82%。

下面将对上面几种流行的模型对于同一数据集进行实验对比,结果如表 1 所示。'-'表示模型没有在该数据集上测试过,2007 表示 VOC 2007,2012 表示 VOC2012。

表 1　单尺度特征目标检测模型结果对比　%

模型	结构	2007	2012	COCO
R-CNN	AlexNet	58.5	53.3	-
SPP-Net	ZF-5	60.9	-	-
Fast R-CNN	VGGNet	70.0	68.4	19.7
MR-CNN	ZF-5	78.2	73.9	-
FasterR-CNN	VGGNet	78.8	75.9	21.9
OHEM	VGGNet	78.9	76.3	25.5
R-FCN	ResNet	83.6	82.0	31.5

单尺度特征下的目标检测模型存在很多问题,选取的特征属于高语义输出,在底层一些细节信息在高层特征容易损失,因此一些尺度较小的目标检测精度非常低,后来很多文献都提出多尺度特征的检测模型。

2.1.2　多尺度特征下的目标检测模型

根据感受野的理论,很多局部微细的条纹以及形状的变化经过多层卷积的处理变得越来越不明显。因此,大部分单尺度特征目标检测模型对小尺度目标的检测精度较低。多尺度特征不再单一选择最后一层卷

积输出作为图像的特征而进行多层特征融合。

较早开始考虑融合多尺度特征方法的是 Bell 等[24]提出的 inside-outside net (ION),优化小目标物体的检测精度,提高对目标遮挡等环境的适应能力。ION 设计了两个子网络 Outside Net 和 Inside Net,Outside 在 ROI 区域之外利用两个循环卷积神经网络层完成上下文特征提取,充分提取全局上下文信息。Inside 在 ROI 区域之内从 Conv3、Conv4 以及 Conv5 三个卷积层提取 ROI 对应的三个尺度特征,最后和上下文特征融合,提高小目标检测的精度。

Kong 等[25]提出了 HyperNet,对特征融合策略做了优化,不同层采样的方式不一样,在层次较低的卷积层使用最大池化技术,对较深层次的卷积层通过添加一个反卷积操作来进行上采样操作,并且还需要使用局部响应正则化手段来归一化多个 Feature Maps。模型和 Faster R-CNN 前向推测时间近似,但在结果上 mAP 却高出 1.6%。

Kim 等[26]提出的 PVANET,基于 Faster R-CNN 优化得到,该网络使用了 C.ReLU、Inception、HyperNet 以及残差网络模块等技术,最主要是采用了 Inception 的结构,在不同层使用多个卷积核完成特征提取,提高感受野对多尺度的适应能力,并且在模型中融合了 HyperNet 的部分结构提高特征的多尺度性。

Lin 等[27]提出将特征金字塔网络模型 FPN 用于多尺度特征的融合。FPN 主体网络采取 ResNet,模型中最核心的部分如图 3 所示。利用自底向上的线路完成图像的初步特征提取,利用自顶向下的线路完成特征的语义传递,利用横向连接完成多尺度特征的融合。文献中将 FPN 用在 Faster R-CNN 的特征提取网络中和区域推荐网络 RPN 中,使得不同尺度的候选区域对应不同层不同尺度的特征。结果证明这种模型在 COCO 数据集上提高了小尺度目标检测精度。

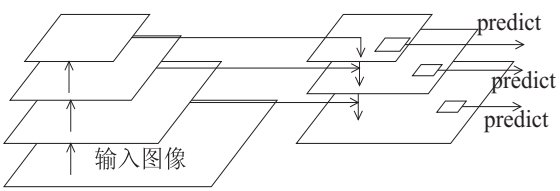

图 3 FPN 结构示意

He 等[28]提出的 Mask R-CNN 模型是利用 FPN 完成目标检测任务,Mask R-CNN 主要目的是实现高精度的实例分割任务,相比目标检测模型仅仅多了一个使用 FCN 网络完成目标 Mask 提取的分支。Mask R-CNN 也创新地使用 ROI Align 池化技术来避免量化造成的精度失配等问题,提高了目标检测精度。

下面将上述模型对同一数据集进行实验结果对比,如表 2 所示。其中‘-’表示模型没有在数据集上测试过,2007 表示 VOC 2007,2012 表示 VOC2012。

表 2 多尺度特征目标检测模型结果对比 %

模型	结构	2007	2012	COCO
ION	VGGNET	79.2	76.4	33.1
HyperNet	VGGNET	76.3	71.4	-
FPN	ResNet	-	-	36.2
MaskRCNN	ResNet	-	-	37.1

将多尺度特征的检测模型在目标尺度复杂的 COCO 数据集上进行结果对比,从数据分析看出多尺度特征的融合能够有效提高检测精度。

2.2 基于回归方法的目标检测模型

相比基于候选区域的目标检测模型,基于回归方法的目标检测模型省去提取候选框的步骤,采取一阶段式模型,即直接在特征图上采取回归方法。目前有很多文献在研究这类模型。由于大部分采取多尺度特征融合的方式,所以不进行特征尺度的区分。

Redmon 等[29]提出 YOLO,他们认为基于候选区域的检测模型效率较低的原因是其复杂的区域推荐方法。因此,提出将整张图作为输入,取得结果后直接在输出层上回归边框的位置并且判断其属于哪一类物体,这样模型的计算效率会大幅度提高。YOLO 首先将一幅图像进行网格划分,分成 $s \times s$ 个网格,如果检测目标的中心位置在某个网格中,那么此网格就负责预测此目标。YOLO 使用多任务损失函数,同时负责目标分类和边框回归。YOLO 虽然效率较高,但是存在很多问题,如对于紧邻的目标以及小尺度目标的检测效果不是特别好,只能预测一类目标和泛化能力弱等等。

Najibi 等[30]提出类似于 YOLO 的 G-CNN 模型,与 YOLO 的区别在于,笔者认为边框搜索不是一个简单的线性搜索过程,如果直接使用传统线性的搜索方法无法取得最优解,因此使用一种迭代回归的方式,然后逐步逼近最优解,实验证明检测结果优于 YOLO。

Liu 等[31]提出一种多尺度特征融合的模型 SSD,和 YOLO 类似的地方在于 SSD 将分类的过程和回归的过程转化为一个目标函数。但是,SSD 融合了 RPN 的 anchor 思想,提出类似的 piror box 方法产生目标的预选框。SSD 使用底层 Feature Map 实现小目标检测,高层 Feature Map 完成大目标检测。SSD 相比 YOLO 精度更高,检测效率也比 Faster R-CNN 更高。

基于 SSD 优化的模型有很多,Jeong 等[32]提出了 RSSD 模型,认为 SSD 有两个问题,首先是相同物体会被大小不同的边框同时检测出来,然后 SSD 对小尺度的物体检测结果比较差。RSSD 模型改善了 SSD 的特征融合方式,利用分类网络增加获取的 Feature Map 之

间的联系,减少了重复框的出现,同时增加了 Feature Map 的个数使得模型能够检测到更多尺度的目标;Fu 等[33]提出 DSSD,使用 Resnet - 101 替代原来的 VGGNet,在分类回归之前使用残差模块,在辅助的 SSD 额外特征卷积层后面添加了几层反卷积层,整个网络就像沙漏一样呈现出宽窄宽的结构。结果显示 DSSD 提高了小目标的检测精度。

当然,在 YOLO 基础上改进的模型也相当多。Redmon 等[34]提出了 YOLO v2 模型和 YOLO 9000 模型。YOLO v2 对 YOLO 模型进行版本优化,YOLO 9000 使用数量较大的分类数据集来协助检测模型的训练。相比 YOLO 具有三部分优化,首先对网络每一层的输入都做 BN 归一化操作,提高模型的收敛速度,使用 anchor 机制和聚类方法完成尺度聚类,然后为了提高模型的鲁棒性,还引入了多尺度训练方式。最后,模型使用 WordTree 原理来融合分类和检测的数据集,以达到增加检测类别的目的。

其实,基于回归方法的目标检测模型也在不断汲取基于候选区域的目标检测模型的优点,因此也有文献提出将两类模型相结合,其中比较典型的就是 Kong 等[35]提出的 RON 模型,如图 4 所示。

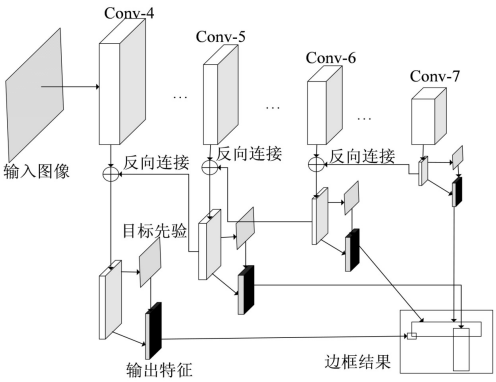

图 4　RON 模型结构示意

RON 有两个比较重要的创新,采用多尺度特征同时设计了反向连接的结构,能够使其检测更多尺度的目标;融入了负样本挖掘,使用目标先验来引导模型对目标的搜索,训练时根据目标的先验结果来更新类别标签,预测时产生目标先验再进行类别检测。RON 不仅精度高而且比较通用。

文中对已有的该类模型进行对比,由于这类模型很少在 MS COCO 数据集上进行实验,所以在 VOC 2007 和 VOC 2012 上进行对比,此外对相同配置同样分辨率输入情况下的检测效率 FPS 进行对比,结果见表 3。基于回归方法的检测模型的检测效率大部分非常可观,较早的 YOLO 系列和 SSD 系列结果精度有很大上升空间,经过不断优化,目前这类模型在精度上能够满足要求,检测基本已经达到 40 fps 以上。

表 3　回归方法类目标检测模型结果对比　%

模型	结构	2007	2012	FPS
YOLO v1	GoogLeNet	66.4	57.9	45
G-CNN	GoogLeNet	66.8	66.4	3
SSD	VGGNet	76.8	74.9	46
RSSD	VGGNet	80.8	76.4	35
DSSD	ResNet-101	81.5	80.0	6.6
YOLO v2	Darknet-19	78.6	73.4	40
RON	VGGNet	77.6	75.4	-

3　结束语

文中主要对现有基于卷积神经网络的目标检测模型进行综述,总结了其优缺点,分析现有目标检测模型的发展空间。总体来看,尚有很多改进空间,例如类内局部语境信息和类间全局语境信息在检测过程中没有考虑,因此这些检测模型在应对复杂背景和多类遮挡环境下显得鲁棒性较差。还有如何优化特征提取网络的基本结构,在不增加计算负担的情况下提高精度等,都有待进一步研究。

参考文献:

[1] 田合雷,丁　胜,于长伟,等. 基于目标检测及跟踪的视频摘要技术研究[J]. 计算机科学,2016,43(11):297-299.

[2] SCHMIDHUBER J. Deep learning in neural networks:an overview[J]. Neural Network,2015,61:85-117.

[3] BENGIO Y,COURVILLE A,VINCENT P. Representation learning:a review and new perspectives[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2013,35(8):1798-1828.

[4] YU Naigong,JIAO Panna,ZHENG Yuling. Handwritten digits recognition base on improved LeNet5[C]//27th Chinese control & decision conference. Qingdao,China:IEEE,2015:4871-4875.

[5] 施泽浩,赵启军. 基于全卷积网络的目标检测算法[J]. 计算机技术与发展,2018,28(5):55-58.

[6] LE M H,WOO B S,JO K H. A comparison of SIFT and Harrisconner features for correspondence points matching[C]//17th Korea-Japan joint workshop on frontiers of computer vision. Ulsan,South Korea:IEEE,2011:1-4.

[7] ZHANG Ying,LI Baohua,LU Huchuan,et al. Sample-specific SVM learning for person re-identification[C]//IEEE conference on computer vision & pattern recognition. Las Vegas,NV,USA:IEEE,2016:1278-1287.

[8] 张　娟,汪西莉,杨建功. 基于深度学习的形状建模方法[J]. 计算机学报,2018,41(1):132-144.

[9] TONG Guofeng,YONG Li,CAO Lihao,et al. A DBN for hyperspectral remote sensing image classification[C]//12th IEEE conference on industrial electronics and applications. Siem Reap,Cambodian:IEEE,2018:1757-1762.

[10] CUESTA-INFANTE A, GARCÍA F J, PANTRIGOJ, et al. Pedestrian detection with LeNet like convolutional networks [J]. Neural Computing & Applications, 2017, 8(3): 1-7.

[11] YUAN Zhengwu, ZHANG Jun. Feature extraction and image retrieval based on AlexNet[C]//Eighth international conference on digital image processing. Chengu, China: [s. n.], 2016: 83-92.

[12] WANG Limin, GUO Sheng, HUANG Weilin, et al. Places 205-VGGNet models for scene recognition [EB/OL]. 2015. https://arxiv. org/abs/1508. 01667.

[13] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//IEEE conference on computer vision and pattern recognition. [s. l.]: IEEE, 2016: 770-778.

[14] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning [EB/OL]. 2016. https://arxiv. org/abs/1602. 07261.

[15] NUNES E, DIAB A, GUNN A, et al. Darknet and deepnet mining for proactive cybersecurity threat intelligence[C]// IEEE conference on intelligence and security informatics. Tucson, AZ, USA: IEEE, 2016: 7-12.

[16] HUANG Gao, LIU Shichen, LAURENS V D M, et al. CondenseNet: an efficient densenet using learned group convolutions[EB/OL]. 2017. https://arxiv. org/abs/1711. 09224.

[17] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//IEEE conference on computer vision and pattern recognition. Columbus, USA: IEEE, 2014: 580-587.

[18] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 37(9): 1904-1916.

[19] GIRSHICK R. Fast R-CNN[C]//IEEE international conference on computer vision. Washington, USA: IEEE, 2015: 1440-1448.

[20] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 39(6): 1137-1149.

[21] GIDARIS S, KOMODAKIS N. Object detection via a multi-region and semantic segmentation-aware CNN model[C]// IEEE international conference on computer vision. Santiago, Chile: IEEE, 2015: 1134-1142.

[22] SHRIVASTAVA A, GUPTA A, GIRSHICK R. Training region-based object detectors with online hard example mining [C]//IEEE conference on computer vision and pattern recognition. Las Vegas, NV, USA: IEEE, 2016: 761-769.

[23] DAI Jifeng, LI Yi, HE Kaiming, et al. R-FCN: object detection via region-based fully convolutional networks[C]// Proceedings of the 30th international conference on neural information processing systems. Barcelona, Spain: Curran Associates Inc., 2016: 379-387.

[24] BELL S, ZITNICK C L, BALA K, et al. Inside-outside net: detecting objects in context with skip pooling and recurrent neural networks[C]//IEEE conference on computer vision and pattern recognition. Las Vegas, NV, USA: IEEE, 2016: 2874-2883.

[25] KONG Tao, YAO Anbang, CHEN Yurong, et al. Hyper-net: towards accurate region proposal generation and joint object detection[C]//2016 IEEE conference on computer vision and patternrecognition. Las Vegas, NV, USA: IEEE, 2016: 845-853.

[26] HONG S, ROH B, KIM K H, et al. PVANet: lightweight deep neural networks for real-time object detection[C]// 2016 IEEE conference on compuer vision and pattern recognition. Las Vegas, NV, USA: IEEE, 2016: 412-425.

[27] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//2017 IEEE conference on compuer vision and pattern recognition. Las Vegas, NV, USA: IEEE, 2017: 936-944

[28] HE Kaiming, GKIOXARI G, DOLLAR P, et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 22(17): 12-24.

[29] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//2016 IEEE conference on compuer vision and pattern recognition. Las Vegas, NV, USA: IEEE, 2016: 429-442.

[30] NAJIBI M, RASTEGARI M, DAVIS L S. G-CNN: an iterative grid based object detector[C]//2016 IEEE conference on computer vision and pattern recognition. Las Vegas, NV, USA: IEEE, 2016: 2369-2377.

[31] LIU Wei, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//European conference on computer vision. Amsterdam, The Netherlands: Springer, 2016: 21-37.

[32] JEONG J, PARK H, KWAK N. Enhancement of SSD by concatenating feature maps for object detection[C]//2017 IEEE conference on computer vision and pattern recognition. Hawaii, USA: IEEE, 2017: 236-246.

[33] FU C Y, LIU W, RANGA A, et al. DSSD: deconvolutional single shot detector[C]//2017 IEEE conference on computer vision and pattern recognition. Hawaii, USA: IEEE, 2017: 2372-2376.

[34] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//2017 IEEE conference on computer vision and pattern recognition. Honolulu, HI, USA: IEEE, 2017: 6517-6525.

[35] KONG Tao, SUN Fuchun, YAO Anbang, et al. RON: reverse connection with objectness prior networks for object detection[C]//2017 IEEE international conference on computer vision. Venice, Italy: IEEE, 2017: 5244-5252.