基于聚类式区域生成的全卷积目标检测网络

潘志浩*, 陈莹**
江南大学物联网工程学院, 轻工过程先进控制教育部重点实验室, 江苏 无锡 214122

摘要 基于区域的全卷积网络(R-FCN)的区域生成网络(RPN)沿用了更快速区域卷积神经网络(Faster R-CNN)的 RPN。针对 RPN 先验框的大小与数量均需人为固定,生成的建议区域过多等问题,将聚类思想应用到 RPN 中,改进先验框的生成方式,提出了基于聚类式区域生成的全卷积目标检测网络。通过对训练样本的真实框进行 *K*-Means 聚类得到先验框的最适大小和最佳数量,取代原本人为固定选取先验框的方式。此外,为增强模型的泛化能力,在改进后的 R-FCN 上使用 ResNet 基础网络,采用困难样本挖掘方法进行训练。实验结果表明,相较于 R-FCN 等方法,该聚类区域全卷积目标检测网络得到的检测结果在精度和速度上都得到了明显的提升。

关键词 图像处理; 目标检测; 聚类算法; *K*-Means; 先验框; 区域生成网络
中图分类号 TP391 **文献标识码** A **doi**: 10.3788/LOP56.151001

Full-Convolution Object Detection Network Based on Clustering Region Generation

Pan Zhihao*, Chen Ying**
Key Laboratory of Advanced Process Control for Light Industry of Ministry of Education, School of Internet of Things Engineering, Jiangnan University, Wuxi, Jiangsu 214122, *China*

Abstract Region proposal networks (RPN) in region-based full-convolutional networks (R-FCN) follow the RPN of faster region convolutional neural networks. In this paper, a full-convolution object detection network based on clustering region generation is proposed to solve the problems of the artificially fixed sizes and quantities of anchor boxes and excessively generated proposals. *K*-means clustering on the ground-truth box of the training samples is used to optimize the sizes and numbers of the anchor boxes in order to replace the fixed boxes in the R-FCN. Furthermore, to enhance the generalization ability of the model, an online hard example mining is used to train the datasets based on the backbone network of ResNet. The experimental results show that the accuracy of the detection results of the proposed algorithm is significantly higher than that of the R-FCN.

Key words image processing; object detection; clustering algorithm; *K*-Means; anchor box; region proposal network
OCIS codes 100.4996; 100.3008; 100.5010

1 引 言

近年来,随着深度学习的兴起,卷积神经网络在目标检测领域得到了广泛的应用。卷积神经网络通过卷积操作和激活函数将图像中的细节信息映射到高维空间,使得一些任务能够在高维空间简单实现。卷积神经网络通过对图像进行卷积操作,提取图像中的重要特征及细节信息,再让网络学习该特征并进行训练。目前,目标检测算法已被应用于多个领域,用于检测各种复杂环境及特殊要求下的目标类别,与传统方法相比,不仅提高了检测准确率和检测速度,也能够达到实时性。叶国林等[1]将卷积神经网络用于夜间行人检测研究,有效地检测出夜间的行人;陆永帅等[2]将深度残差网络应用于霾监测,提高了霾识别精度;王正林等[3]提出了一种并行深度残差网络,将其用于检测堆场烟雾,在提升了烟雾检

收稿日期: 2019-01-04; **修回日期**: 2019-02-15; **录用日期**: 2019-03-05
基金项目: 国家自然科学基金(61573168)
* **E-mail**: 543152026@qq.com; ** **E-mail**: chenying@jiangnan.edu.cn

测准确率的同时，还降低了类烟物体产生的误报率。

当前目标检测算法有很多，但主要分为两类：一类是二段探测器算法[4-6]，该算法先生成一系列作为样本的候选框，再通过卷积神经网络进行样本分类，如快速区域卷积神经网络(Fast R-CNN)[7]、更快速区域卷积神经网络(Faster R-CNN)[8]、基于区域的全卷积网络(R-FCN)[9]、基于掩模的快速区域卷积神经网络(Mask R-CNN)[10]等；另一类则是一段探测器算法[11-12]，该算法不产生候选框，直接将目标边框定位的问题转化为回归问题进行处理，如单段多框目标检测器(SSD)[13]、统一的实时目标检测方法(YOLO)[14]、反卷积单段检测器(DSSD)[15]等。二段探测器算法相较于一段探测器算法，检测准确率和定位精度更高，但算法速度略低，这是由于二段探测器算法先使用区域生成网络(RPN)生成建议区域，再对建议区域进行分类和定位。然而，RPN 中的先验框是人为固定选取的(3 种尺度和 3 种长宽比)，存在一定的人为主观性，且生成的建议区域中存在很多不必要的重叠，增加了检测时间。

本文在 R-FCN 的基础上进行改进，在 R-FCN 的 RPN 中使用 K 均值(K-Means)聚类算法，改变先验框的生成方式，通过对训练样本中的真实框进行聚类分析，得到先验框的最适大小及最佳数量，由此遍历特征映射图得到建议区域，并将其送入后续网络进行困难样本挖掘训练。改进后的 R-FCN 在 ResNet-50、ResNet-101 基础网络对 Pascal VOC 07+12 trainval 训练集进行训练，使用 Pascal VOC 07 test 进行测试，实验结果表明，与 Faster R-CNN 和原 R-FCN 相比，所提方法的检测精度和速度都有所提升。

2 R-FCN 结构

R-FCN 结构[9]主要分为 3 部分：

1) 卷积网络。图像归一化后经过卷积网络，提取得到图像共享特征映射图。

2) RPN。通过 RPN 在共享特征映射图上提取感兴趣区域(RoI)。

3)分类网络。R-FCN 通过卷积操作为每类生成位置敏感分数图，经过位置敏感 RoI 池化层进行投票打分及分类定位。

R-FCN 结构图如图 1 所示。

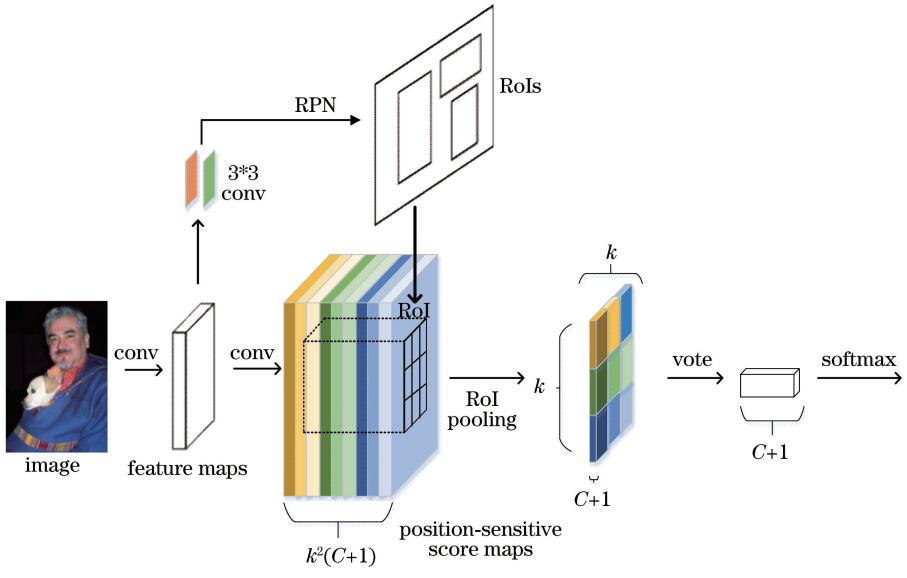

图 1 R-FCN 结构图

Fig. 1 Structure of R-FCN

2.1 卷积网络

将网络读取的任意尺寸的图像经预处理操作后归一化为 1000×600，再将图像输入网络，经过一组卷积层、激活层和池化层得到共享特征映射图，最后送入后续的 RPN 和分类网络。

2.2 RPN

RPN 将输入的共享特征映射图经过一个 3×3 的滑动窗口生成长度为 512 维的全连接特征，RPN 结构图如图 2 所示。该全连接特征产生两个分支：定位层(reg-layer)和分类层(cls-layer)。定位层用于预测先验框对应的建议区域的坐标 x、y，宽 w 和高 h。分类层用于判定建议区域是前景还是背景。

在滑动窗口位置预测多个建议区域，定义每个位置最大可能的建议区域数量为 t，默认使用 3 种

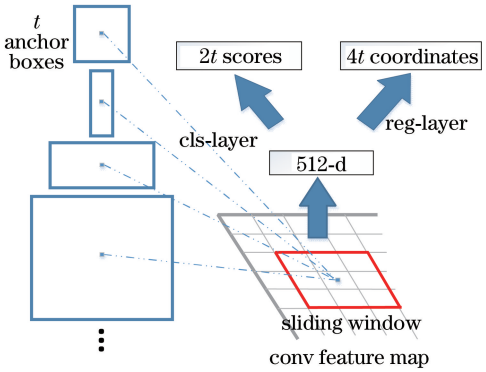

图 2 RPN 结构图

Fig. 2 Structure of RPN

比例(64^2,128^2,256^2)和 3 种长宽比(1∶1,1∶2,2∶1)的 9 个固定大小的先验框(即 anchor boxes)并通过滑动窗口产生建议区域。则 reg-layer 有 $4t$ 个输出,cls-layer 有 $2t$ 个输出。

2.3 分类网络

R-FCN 分类网络通过卷积操作为每类生成 $k\times k$(k 的取值为自定义,图 1 中 $k=3$,表示将一个 RoI 划分成 3×3)个位置敏感分数图。对于 C 个种类,有 $k^2(C+1)$个通道输出(C 种类别+背景),$k\times k$ 个子区域反映了对应位置的空间网格,保存目标的空间位置信息。对于任意子区域(i,j),$0\leqslant i,j\leqslant k-1$,通过 RoI 池化层进行位置敏感 RoI 池化操作,可表示为

$$r_c(i,j\,|\,\Theta)=\frac{1}{n}\sum_{(x,y)\in b_{\text{in}}(i,j)} z_{i,j,c}(x+x_0,y+y_0\,|\,\Theta),\tag{1}$$

式中:$r_c(i,j\,|\,\Theta)$为第 c 个种类的第(i,j)个网格的池化响应;$z_{i,j,c}$ 为 $k^2(C+1)$个分数图中的一个输出;(x_0,y_0)表示 RoI 的左上角部分坐标;n 为网格中像素的数量;Θ 为网络中可学习的参数;b_{in}为网格单元。故每个类别共有 k^2 个分数图(图 1),RoI 对于每个类别可以得到 k^2 个值,由于类别总数是 $C+1$ 个,因此一个 RoI 可以得到 $k^2(C+1)$个值。对于每个类别,该类别的 k^2 个值都表示该 RoI 属于该类别的响应值,将 k^2 个数相加就得到该类的分数,则共有 $C+1$ 个分数。最后通过 softmax 计算出各个类别的概率。

类似于位置敏感分数图,在最后一个共享卷积层上增加一个并行分数图用于回归 x、y、w、h 的偏移量,回归的维度是 $4k^2$,经过位置敏感 RoI 池化操作后,每个 RoI 得到 4 个值作为该 RoI 的 x、y、w、h 的偏移量,以此进行定位。

3 改进的 R-FCN

在 R-FCN 现有的基础上,将 RPN 与 K-Means 聚类算法相结合,提出了一种改进的 R-FCN。针对先验框人为选取不客观,生成的建议区域重叠部分过多的问题,对 RPN 进行改进和优化,不仅得到更加适合模型的先验框大小和数量,使得模型更加容易学习,还去掉了许多不必要的建议区域,减少了网络的检测时间。

3.1 问题分析

R-FCN 沿用了 Faster R-CNN 的 RPN 网络,仍使用 3 种比例(64,128,256)和 3 种长宽比(1∶1,1∶2,2∶1)的 9 个人为固定先验框,这会导致生成的建议区域过多、计算损耗较大和定位不精准的问题,具体原因如下:

1) 从图 2 可以看出,R-FCN 的 RPN 网络中使用 9 个人为固定的先验框遍历共享特征映射图(大小通常为 60×40),为特征映射图上的每个点都生成 9 个先验框作为初始的建议区域,数量约为 21600(≈60×40×9)。由于后续的分类网络使用 softmax 层计算 RPN 生成的所有建议区域的类别概率,因此分类网络 softmax 层计算量增加,降低了模型的检测效率。

2) 先验框的尺寸(长和宽)都是人为预先设定的,在一定程度上带有人为主观性,同时这种固定的选取策略可能会使生成的候选区域不适用于数据集,导致 R-FCN 的检测结果存在定位偏差及不准确。R-FCN 检测结果如图 3 所示,其中红色框为 R-FCN 的检测结果。由结果可知,图 3(a)中虽然定位基本准确,但由于先验框固定的长宽比,车尾部分无法进入定位框;图 3(b)中左上方人物具有较严重的自遮挡,目标比例异于人为主观,检测误差较大。

由上述分析可知,RPN 网络若采用人为固定的先验框,对检测网络的定位精度和速度均会产生不利影响。因此,考虑通过统计方法将训练集中的目标框进行聚类分析,以获取大小合适、数量适宜的先验框代替原 R-FCN 中的固定先验框,不仅可以生成适合该数据集的建议区域,增加目标定位的准确度,还能减少建议区域的数量,减少检测时间,使得模型更容易学习,做出更好的预测。

3.2 改进网络的框架

改进后的网络框架如图 4 所示,黑色虚线框部分为卷积网络和分类网络,红色虚线框部分为 RPN。改进的 R-FCN 主要分为 3 部分:

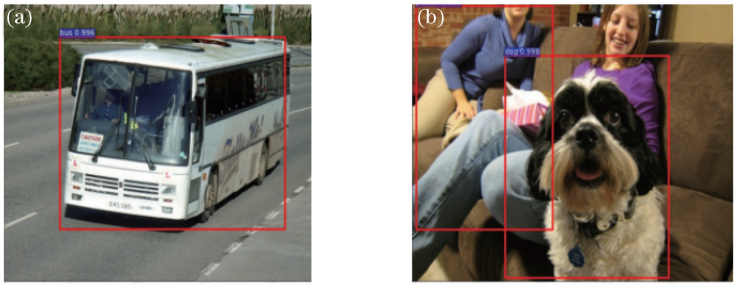

图 3 R-FCN 检测结果。(a)定位不精确;(b)检测误差严重

Fig. 3 Detection results of R-FCN. (a) Inaccurate locating; (b) serious detection error

1) 卷积网络。图像归一化后经过卷积网络,提取得到图像共享特征映射图。为达到与文献[9]相同的实验环境和条件,以深度残差网络[16]为基础网络(主要是 ResNet-50,ResNet-101 网络)。

2) RPN 聚类网络。将训练样本图像归一化后提取出真实框,使用 K-Means 算法对真实框的尺寸进行维度聚类,得到先验框的最佳大小和数量,再通过 RPN 在共享特征映射图上提取感兴趣区域。

3) 分类网络。R-FCN 通过卷积操作为每类生成位置敏感分数图,经过位置敏感 RoI 池化层进行投票打分,完成分类定位。

3.3 RPN 聚类网络

为了得到先验框的最适大小及最佳数量,通过将 K-Means 聚类算法应用到 RPN 网络中对 RPN 网络做出改进。通过 K-Means 聚类算法来生成先验框,取代原本人为选取方式。K-Means 聚类算法快速、简单,具有可伸缩性,且对于大数据集具有较高的效率,能够根据相似性原则将数据集中具有较高相似度的真实框通过聚类划分为同一类簇,将具有较高相异度的真实框聚类划分至不同类簇,将得到的紧凑且独立的真实框簇应用于 RPN 的建议区域生成。

RPN 聚类网络基本结构如图 5 所示。首先,将网络提取训练集中的真实目标框的宽高,作为 K-Means 聚类算法的输入,以 K 为先验框的个数;接着,根据K-Means算法中的相似性原则,将尺寸相

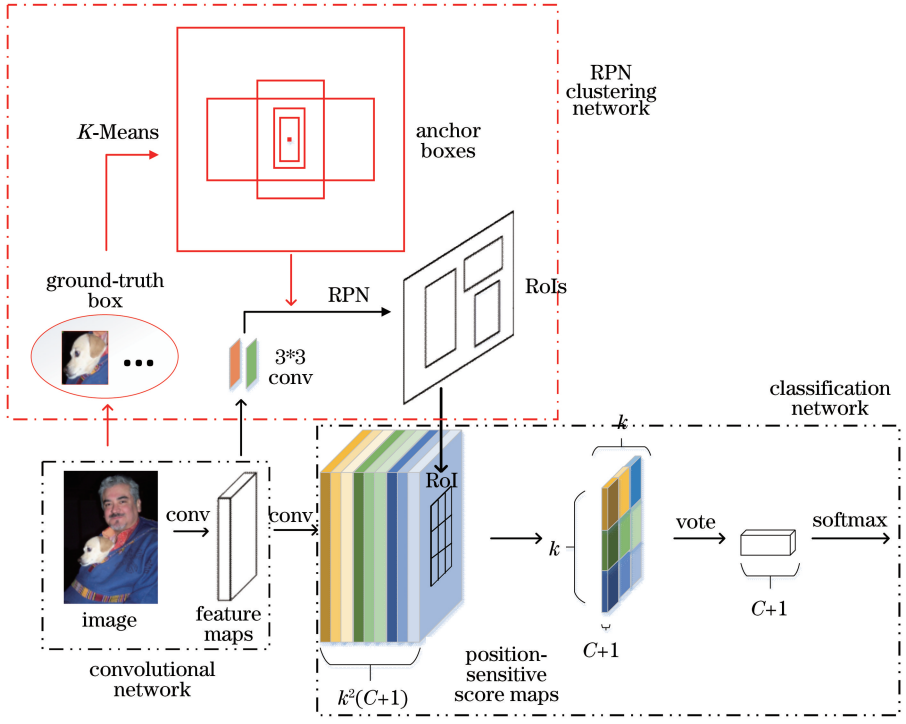

图 4 改进后的网络框架

Fig. 4 Improved network frame

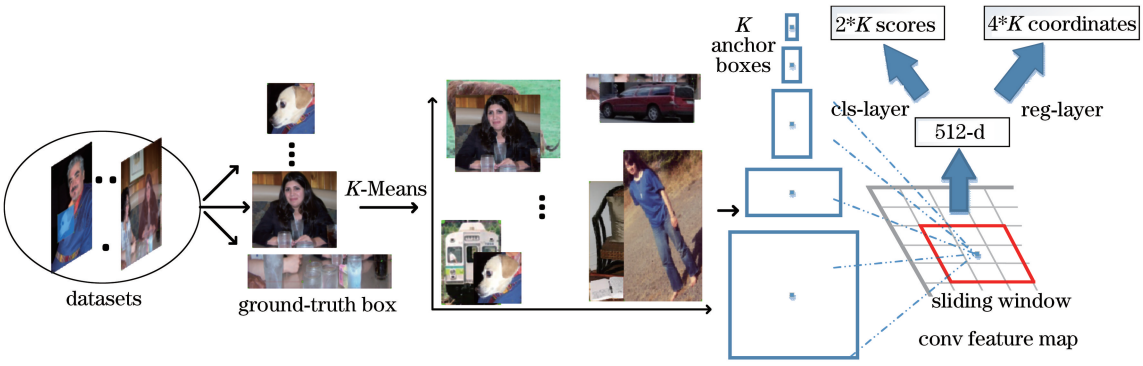

图 5 RPN 聚类网络基本结构

Fig. 5 Basic structure of RPN clustering network

似度较高的真实框划为同簇，尺寸较高相异度的真实框划分到不同类簇，取各个不同簇的中心框为先验框，计算每个不同簇中真实框宽高的平均值，得到先验框的合适尺寸。

在聚类过程中，使用欧氏距离会导致大框比小框产生更大的误差，而框与框的交集除以它们的并集(IOU)与框尺寸无关。为了减小误差和框尺寸之间的关系，更好地计算出合适的先验框，使得预测框与真实框的 IOU 值更高，在聚类分析时选用真实框与聚类中心框之间的 IOU 值作为距离指标：

$$d(\text{box},\text{centroid})=1-I(\text{box},\text{centroid}),\quad(2)$$

式中：box 代表训练样本的真实框；centroid 代表聚类中心的框；$I(\text{box},\text{centroid})$ 为真实框与聚类中心框两者的交集与两者的并集之比；$d(\text{box},\text{centroid})$ 为真实框与聚类中心框之间的距离。

另一方面，RPN 聚类网络在原始特征提取网络的基础上，额外添加一个3×3 的卷积层(图 4)。3×3 卷积相当于一个大小为 3×3 的滑动窗口在卷积特征图上滑动，每一个滑动窗口映射到一个低维向量(对应 ResNet 网络均是 512 维)，RPN 聚类网络基本结构如图 5 所示。与原 RPN 相似，网络通过两个 1×1 卷积将该向量分为两路，其中一路输出目标和非目标的概率，另一路输出检测定位框相关的 4 个参数$[x,y,w,h]$。就卷积运算本身而言，该向量将作为分类层和定位回归层的输入。

网络训练损失函数为一个多任务损失函数，可分为两部分，对应 RPN 两条支路，即目标与否的分类误差和候选框的回归误差，定义为

$$L(s,t_{x,y,w,h})=L_{\text{cls}}(s_{c^*})+\lambda(c^*>0)L_{\text{reg}}(t,t^*),\quad(3)$$

式中：s 为 softmax 响应；$t_{x,y,w,h}$ 为预测框相对于真实框中心坐标及宽高的偏移量；λ 为平衡权重；$L_{\text{cls}}(s_{c^*})=-\ln s_{c^*}$ 为分类的交叉熵损失；s_{c^*} 为真实标签类别的 $softmax$ 响应；L_{reg} 为边界框回归损失；$[c^*>0]$为一个指标，如果参数为真，则等于 1，否则为 0；t^* 表示真实框。$c^*=0$ 表明 RoI 的真实标签为背景；L_{reg} 与 $Faster\ R\text{-}CNN$[8] 中的边界框损失回归相同，和 $Faster\ R\text{-}CNN$[8] 一样，设定 $\lambda=1$。

3.4 困难样本挖掘

为了增强网络模型的泛化能力，采用困难样本挖掘(OHEM)方法[17]计算区域生成网络生成的各个 RoI 的损失值，通过对每个 RoI 损失值进行排序，选取损失值最高的前 128 个样本组成迭代训练中一个批次的训练网络。在网络训练过程中利用困难样本更新网络参数，使网络训练更充分，从而达到更好的检测效果。

4 实验结果及分析

为了验证所提方法的可行性，并将所提算法与现有算法 Faster R-CNN[8]、R-FCN[9] 进行分析对比，采用和文献[8-9]相同的数据集进行实验，训练集为 Pascal VOC 07+12trainval，测试集为 Pascal VOC 07 test，目标检测的精度以平均精确度(mAP)为标准。

4.1 实验设置

实验机器的图形处理器(GPU)为 TITAN XP，深度学习的框架为 Caffe，Ubuntu 系统版本为 14.04，Cudnn 版本为 8.0。所研究的 R-FCN 网络以深度残差网络 ResNet-50、ResNet-101 为网络基础。网络参数更新方法均为随机梯度下降法，基础学习率为 0.001，步长为 50000，冲量为 0.9，权重衰减项为 0.0005，训练的最大迭代次数为 50 万，分两种情况训练检测，一种是未采用困难样本挖掘方法，另一种是采用困难样本挖掘方法。

4.2 聚类参数分析

K-Means 聚类算法中 K 的取值通过多次实验得到，对 Pascal VOC 07+12 数据集进行聚类检测，取最优值。图 6 为聚类检测结果图，其中图 6(a)为不同 K 值下的平均 IOU，图 6 (b)为在 ResNet-101 基础网络下，采用困难样本挖掘方法得到的不同 K 值的检测精度，图 6 (c)为不同 K 值下聚类耗费时间。

从图 6(a)可以看出，随着聚类次数 K 的增大，平均 IOU(即各个边界框与聚类中心的 IOU 的平均值)也增大。从图 6(b)可以看出，K 取 5 和 7 时能获得较高的检测精度。从图 6(c)可以看出，当 K 值不断增大时，聚类真实框得到先验框耗费的时间越来越长，模型的复杂度在不断的增加。与文献[18]中综合考虑 K-Means 聚类算法所耗费的时间以及精度之间的平衡关系一致，为了平衡平均 IOU、精度、聚类所用时间和模型复杂度之间的关系，本文最终选取 K 为 5。另外，由图 4 可以看出，聚类数据集真实框得到 K 个先验框操作独立于整个网络，可离线进行，RPN 聚类网络遍历最后一层卷积层生成特征映射图上的每个点，并为每个点生成 K 个先验框作为初始建议区域，再送入后续的分类网络进行分类定位，整体模型的检测效率并没有受到影响。相反，由于重叠建议区域数量的减少，网络的整体检测时间相较原网络有所缩短。

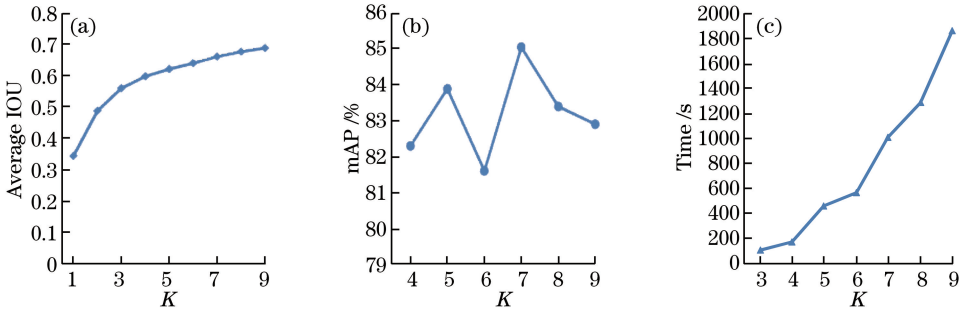

图 6 聚类检测结果。(a)不同 K 值下的平均 IOU；(b)不同 K 值下的检测精度；(c)不同 K 值下聚类耗费时间

Fig. 6 Clustering detection results. (a) Average IOU with different K values; (b) detection accuracy with different K values; (c) clustering consuming with different K values

4.3 结果对比分析

4.3.1 检测精度分析

数据集的数量总计约为 27000 张，测试样本的数量约为 4950 张。表 1、2 分别为 ResNet-50 和 ResNet-101 下，Faster R-CNN、R-FCN 以及改进的 R-FCN 的检测结果对比，N/A 表示数据缺失，原论文中无该实验。表中 OHEM 表示采用困难样本挖掘的方法。表 3 为 ResNet-101 下不同方法的各类检测结果。

从表 1 可以看出，以 ResNet-50 为基础网络进行训练，所提方法得到的mAP值为79.04%，相较于 Faster R-CNN 的 76.60%提升了 2.44%；在困难样本挖掘情况下，所提方法得到的 mAP 值为 83.36%，相较于 R-FCN 的 77.4%提升了 5.96%。

表 1 ResNet-50 下，不同方法的检测结果比较

Table 1 Detection results with different methods based on ResNet-50

Backbone network	Method	mAP /%	Detection time /s
ResNet-50	Faster R-CNN	76.60	0.420
	R-FCN	N/A	N/A
	Proposed	79.04	0.031
	Faster R-CNN (OHEM)	N/A	N/A
	R-FCN (OHEM)	77.40	0.099
	Proposed (OHEM)	83.36	0.031

表 2 ResNet-101 下，不同方法的检测结果比较

Table 2 Detection results with different methods based on RseNet-101

Backbone network	Method	mAP /%	Detection time /s
ResNet-101	Faster R-CNN	76.40	0.420
	R-FCN	76.60	0.170
	Proposed	81.01	0.046
	Faster R-CNN (OHEM)	79.44	0.042
	R-FCN (OHEM)	79.50	0.170
	Proposed (OHEM)	84.64	0.046

从表 2 可以看出，以 ResNet-101 为基础网络进行训练，所提方法得到的 mAP 值为 81.01%，Faster R-CNN 的 mAP 值为 76.40%，R-FCN 的 mAP 值为 76.60%，相较于这两种方法，所提方法的 mAP 值分别提升了 4.61%和 4.41%；在使用困难样本挖掘方法时，所提方法训练得到的 mAP 值为84.64%，Faster R-CNN 和 R-FCN 训练得到的 mAP 值分别为79.44%和79.5%，相较于这两种方法，所提方法的

表 3 ResNet-101 下不同方法的各类检测结果

Table 3 All kinds of detection results with different methods based on RseNet-101

Method	mAP/%	Areo	Cat	Bird	Boat	Bottle	Bus	Plant	Bike	Chair	Cow
R-FCN	79.5	82.5	88.4	83.7	69.0	69.2	87.5	54.1	83.7	65.4	87.3
Ours	81.01	82.0	90.8	82.8	79.3	59.2	89.4	58.3	82.8	59.6	88.9
Proposed (OHEM)	84.64	82.5	90.7	88.5	81.3	71.4	89.9	66.7	88.5	72.3	89.7
Method	mAP/%	Table	Dog	Horse	Bike	Person	Car	Sheep	Sofa	Train	TV
R-FCN	79.5	72.1	87.9	88.3	81.3	79.8	88.4	79.6	78.8	87.1	79.5
Ours	81.01	75.1	90.8	89.9	84.2	79.2	85.5	86.3	84.3	90.2	77.9
Proposed (OHEM)	84.64	81.1	90.6	90.3	88.1	80.3	88.3	89.2	85.1	90.4	86.0

mAP 值分别提升了5.2%和 5.14%。

从表 3 可以看出，boat、plant、chair、table、sheep、sofa、TV 这些种类的检测精度提升幅度较为明显，精度提升约为 10%，而其他如 cat、bottle、bus、cow、dog、horse 等种类的精度虽然也有所提升，但精度提升只有约 2%～3%，提升幅度相对较小。

由表 1、2 可以看出，相比于 Faster R-CNN 和 R-FCN 的检测效果，在 RPN 子网络中使用 *K*-Means 聚类算法得到的检测效果最好，这证明了所提方法的有效性。此外，由表 3 可以看出，改进网络对于 boat、table、sofa 等大目标类别的检测效果提升较大，对于 cat、dog 等小目标类别的检测效果提升较小，说明使用聚类算法得到的先验框大小比原来人为固定选取的先验框更加适合原网络。

算法改进前后检测效果的对比如图 7 所示，其中图 7 (a)为 R-FCN 检测结果，图 7 (b)为改进 R-FCN 的检测结果。由图 7 可以看出，图 7 (a)的前两张图片存在定位框太大或太小以及定位不准确的情况，后 3 张图片则存在漏检的情况。而图 7 (b)的前两张检测图片相较于图 7 (a)定位框大小合适且定位准确，并且后 3 张检测出了图 7 (a)中漏检的种类。相比之下，改进的 R-FCN 的检测结果优于 R-FCN。

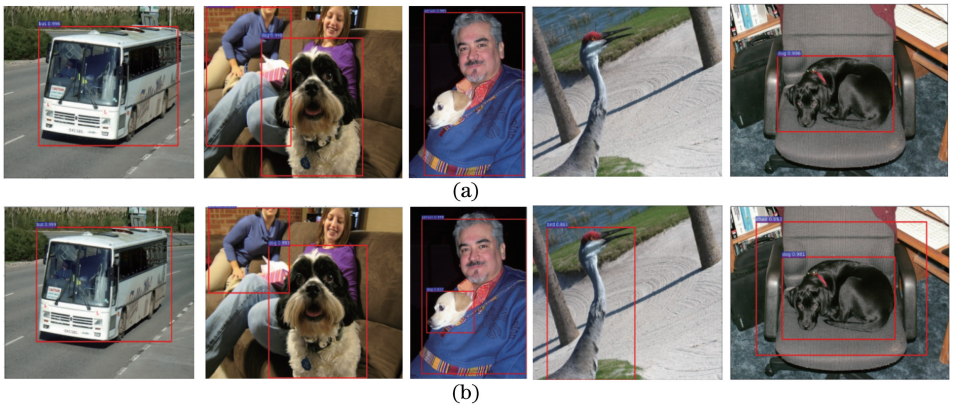

图 7 算法改进前后检测效果对比图。(a) R-FCN 检测结果；(b)改进算法检测结果

Fig. 7 Comparison of detection results before and after algorithm improvement. (a) Detection results of R-FCN; (b) detection results of improved algorithm

4.3.2 检测速度分析

先验框的数量由原来的 9 个减少为现在的 5 个，通过先验框在滑动窗口处生成的建议区域数量也减少为 12000，剔除太小和超出边界的建议区域，则建议区域数量变得更少，重叠部分也比原来少，后续操作根据建议区域进行非极大值抑制，进一步减少建议区域的重叠数量(进行非极大值抑制的 IOU 阈值设定为 0.7)。从表 1 可以看出，以 ResNet-50 为基础网络进行训练时，检测时间由原本的 0.099 s 缩短为 0.031 s；从表 2 可以看出，以 ResNet-101 为基础网络训练时，检测时间由原本的 0.170 s 缩短为 0.046 s。整个网络的检测时间整体上缩短了约 1/3，大大提升了检测效率。

实验结果表明，在不改变 R-FCN 分类网络的前提下，相较于人为选取先验框的大小和数量，使用 *K*-Means 聚类算法改进先验框的生成方式，通过对训练集的真实框尺寸进行聚类得到先验框合适的大小和数量，可以减少建议区域的数量，提升检测速

度，缩短网络的检测时间，同时先验框的大小由训练样本的真实框聚类得到，更加适合网络学习，显著提高网络的检测精度。

4.3.3 K 值普遍性分析

为验证 $K=5$ 是否具有普遍性，从 MS COCO 数据集中提取出与 Pascal VOC 数据集不同的 10 种类别约 20000 张图片形成新数据集，仍采用 $K=5$ 进行聚类式区域生成。实验结果如表 4 所示。

从表 4 可以看出，R-FCN[9]方法在新的数据集上得到的检测精度为 40.06%，而所提方法得到的检测精度为 41.16%，检测精度提升了 1.1%。检测精度及提升幅度相较于 Pascal VOC 数据集都较小。这是由于 MS COCO 数据集中每张图片中的物体数目较多，且含有很多小目标物体，增加了网络检测难度，因此相对于 Pascal VOC 数据集，整体的检测精度、提升的幅度都较小。但该实验表明，当 $K=5$ 时得到的精度仍高于 R-FCN[9]得到的精度，所以在相近的目标检测任务中该 K 值也适用，因此具有普遍性。

表 4 K 值普适性实验结果

Table 4 Experimental results of generalization of K value

Backbone network	Method	mAP /%	Detection time /s
ResNet-101	R-FCN (OHEM)	40.06	0.170
	Proposed (OHEM)	41.16	0.046

5 结 论

随着深度学习在机器学习领域不断热门化，目标检测算法被应用到越来越多的识别领域。卷积神经网络作为一种高效的识别方法在目标检测领域得到广泛应用。采用基于区域的全卷积网络进行改进，通过对 RPN 网络中先验框生成方式及数量进行改进，得到合适的先验框大小及数量，将其应用到后续的 R-FCN 分类网络中，并在 Pascal VOC 07+12 数据集进行训练和测试。相较于 R-FCN、Faster R-CNN 的检测结果，所提方法不仅显著提升了检测精度，也缩短了网络检测时间，提升了检测速度。实验结果验证了加入聚类方法的有效性和可行性。

参 考 文 献

[1] Ye G L, Sun S Y, Gao K J, *et al*. Nighttime pedestrian detection based on faster region convolution neural network [J]. Laser & Optoelectronics Progress, 2017, 54(8): 081003.
叶国林, 孙韶媛, 高凯珺, 等. 基于加速区域卷积神经网络的夜间行人检测研究[J]. 激光与光电子学进展, 2017, 54(8): 081003.

[2] Lu Y S, Li Y X, Liu B, *et al*. Hyperspectral data haze monitoring based on deep residual network[J]. Acta Optica Sinica, 2017, 37(11): 1128001.
陆永帅, 李元祥, 刘波, 等. 基于深度残差网络的高光谱遥感数据霾监测[J]. 光学学报, 2017, 37(11): 1128001.

[3] Wang Z L, Huang M, Zhu Q B, *et al*. Smoke detection in storage yard based on parallel deep residual network [J]. Laser & Optoelectronics Progress, 2018, 55(5): 051008.
王正来, 黄敏, 朱启兵, 等. 基于并行深度残差网络的堆场烟雾检测方法[J]. 激光与光电子学进展, 2018, 55(5): 051008.

[4] 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.

[5] He K M, Zhang X Y, Ren S Q, *et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[6] Lin T Y, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 17355379.

[7] Girshick R. Fast R-CNN [C] // 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 1440-1448.

[8] Ren S Q, He K M, Girshick R, *et al*. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[9] Dai J F, Li Y, He K M, *et al*. R-FCN: object detection via region-based fully convolutional networks [C] // 30th Conference on Neural Information Processing Systems (NIPS 2016), December 5-10, 2016, Barcelona, Spain. USA: Curran Associates Inc., 2016: 379-387.

[10] He K M, Gkioxari G, Dollár P, *et al*. Mask R-CNN[C] // 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 2980-2988.

[11] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 6517-6525.

[12] Redmon J, Farhadi A. Yolov3: an incremental improvement[J/OL]. (2018-04-08)[2018-11-16]. https://arxiv.org/abs/1804.02767.

[13] 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.

[14] Redmon J, Divvala S, Girshick R, *et al*. You only look once: unified, real-time object detection[C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 779-788.

[15] Fu C Y, Liu W, Ranga A, *et al*. DSSD: deconvolutional single shot detector[J/OL]. (2017-01-23)[2018-11-16]. https: // arxiv. org/abs/1701.06659.

[16] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[17] Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining[C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 761-769.

[18] Shrivastava M, Matlani P. A smoke detection algorithm based on K-means segmentation[C] // 2016 International Conference on Audio, Language and Image Processing (ICALIP), July 11-12, 2016, Shanghai, China. New York: IEEE, 2016: 301-305.