计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(8):170−175 [doi: 10.15888/j.cnki.csa.007008]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于目标检测算法的 FashionAI 服装属性识别①

陈亚亚, 孟朝晖

(河海大学 计算机与信息学院, 南京 211100)
通讯作者: 陈亚亚, E-mail: yychentracy@126.com

摘　要: 随着网络上服装图片数量的快速增长, 对于大量的服装进行分类的需求与日俱增. 传统的使用手工进行服装图像的语义属性标注并不能完全的表达服装图像中的丰富信息, 并且传统的手工设计的特征已经不能满足现实的精度和速度的需求. 近年来, 深度学习已经应用到计算机视觉方方面面, 为基于深度学习的服装分类识别技术奠定了坚实的基础. 本文根据已有的数据集 DeepFashion 构建了三个新的子数据集, 进行分类训练的 deepfashionkid 数据集和进行 Faster R-CNN 训练的 deepfashionVoc 数据集和进行 Mask R-CNN 训练的 deepfashionMask 数据集. 使用 deepfashionkid 数据集在 VGG16 上进行预训练得到 clothNet 模型, 进而改进 Faster R-CNN 的损失函数. 并且各自对比了这两种算法使用 clothNet 预训练的模型与不使用的区别. 另外, 本文了采用一种新的类似嫁接学习的预训练策略. 实验表明, 这些训练技巧对于检测精度的提高具有一定的帮助.

关键词: 深度学习; 目标检测; 嫁接学习; 卷积网络; Mask R-CNN

引用格式: 陈亚亚,孟朝晖.基于目标检测算法的 FashionAI 服装属性识别.计算机系统应用,2019,28(8):170–175. http://www.c-s-a.org.cn/1003-3254/7008.html

FashionAI Clothes Recognition Based on Object Detection Algorithm

CHEN Ya-Ya, MENG Zhao-Hui

(College of Computer and Information, Hohai University, Nanjing 211100, China)

Abstract: With the rapid growth in the number of clothing pictures on the Internet, the demand for classification of a large number of clothing is increasing. The traditional use of manual semantic attribute annotation of clothing images does not fully express the rich information in the clothing image, and the traditional hand-designed features can no longer meet the requirements of real precision and speed. In recent years, deep learning has been applied to all aspects of computer vision, laying a solid foundation for clothing classification and recognition technology based on deep learning. In this study, three new sub-datasets are constructed according to the existing dataset deepfashion, the deepfashionkid dataset for classification training, the deepfashionVoc dataset for training with Faster R-CNN, and the deepfashionMask dataset for Mask R-CNN training. The clothNet model is pre-trained on the VGG16 using the deepfashionkid dataset to obtain the clothNet model, which in turn improves the loss function of the Faster R-CNN. And each compares the difference between the two algorithms using clothNet pre-trained model and not used. In addition, this study adopts a new pre-training strategy to adopt a training method similar to grafting learning. Experiments show that these training techniques are helpful for improving the detection accuracy.

Key words: deep learning; object detection; grafting learning; convolutional network; Mask R-CNN

① 收稿时间: 2019-01-30; 修改时间: 2019-02-21; 采用时间: 2019-03-04; csa 在线出版时间: 2019-08-08

近年来, 随着网购的发展, 越来越多的服装图片出现在购物网站上, 对于这些服装进行分类和属性识别变的越来越困难, 以前对于服装属性识别大部分是基于关键字的, 但是这种检索的方法就需要提前给数据库里面的服装打标签, 手工打标签耗时耗力, 并且这种标签都是基于语义特征的, 忽略了图像中的很多细节特征和一些难以描述的特征, 准确率不是很高. 随着图像处理技术和卷积神经网络逐渐发展, 网络直接学习图像的特征, 获取到特征表示之后再度量特征[1]之间的距离和相似度来分类, 可以很好的提高分类的速度和精度.

Pan RR[2]等人采用了 FCM 算法从针织物中提取纹理特征和结构特征送入到 BP 网络进行花纹识别. Salem YB[3]对于服装在纹理分析的基础上进行属性识别, 使用 SVM 分类器进行分类. Liu S[4]等人提出了一种新的 colorfashion 数据集, 结合了人体姿态估计模块, 颜色和类别推断模块和超像素类别分类器学习模块使用多个分类器, 直接解析分类服装图像的多个类别. Bourdev L[5]等人使用了基于 poselet 的方法, 将多个属性分解, 然后训练每个属性的分类器, 将这些分类器组合在一个判别模型中. Hu ZL[6]等人采用了 CDT 的服装分割算法, 对于前景和背景分别建模, 可以从背景和光照条件不同的静态图像中提取到不同的服装特征. 这些方法是基于人工设计的特征, 包括 sift 特征, hog 特征, 即使采用特征融合也很难达到一定的准确率. 近年来, 卷积神经网络[7–9]得到了快速的发展, 利用卷积网路提取图像的特征已经成为研究的热点, 因为卷积可以通过学习和训练提取到图像的特征, 并且卷积神经网络快速地被利用到服装属性识别和分类中, Liu ZW[10]等人引入了 deepfashion 数据集提出了基于关键点的服装属性识别, 取得了很好的效果. 在 2018 年阿里举办了 FashionAI 比赛, 分为服装关键点定位和属性标签识别两个任务. 总之, 深度学习[11]促进了服装属性识别技术发展, 提高了服装属性识别的准确率. 然而在基于深度学习的服装属性识别中, 由于服装的细节信息比较丰富, 属性之间的相似度比较高, 加上一些例如光照的外界干扰因素严重影响识别和分类的精度和准确度. 首先需要一个种类丰富的数据集, 另外网络框架的一些细节的设计直接影响到了服装属性识别和分类的效果. 本文研究分析现有的服装属性识别框架, 并做了对比实验, 在已有网络的基础上改进损失函数, 采取更有效的区域选择方法.

1 基于局部关键点检测的卷积神经网络

将原来的网路分为单个网络分支, 全局网络, 局部网络, 和关键点检测网络. 加入了一个新的层, 这个层叫做 landmark. 首先得到一个局部特征, 然后将这些局部特征按通道进行组合在一起, 使用全连接层进行编码输出, 得到固定维数的局部特征. landmark 代表的是一种局部的细粒度的特征, 将这些特征和全局特征进行组合能较好的对图像特征进行表征, 基于这一点设计了一种关键点回归网络, 基本部分与主网络共享权值. 输出的关键点信息作为输入被网络 2 吸收. 当有足够多的有标注的数据时, 深度学习可以同时学习图像特征和度量函数, 其背后的思想就是通过给定的度量函数, 学习特征在该度量空间下具有很好的判别性, 端到端的特征学习重要研究方向就是如何构建更好的特征表示形式和损失函数, 网络结构如图 1 所示.

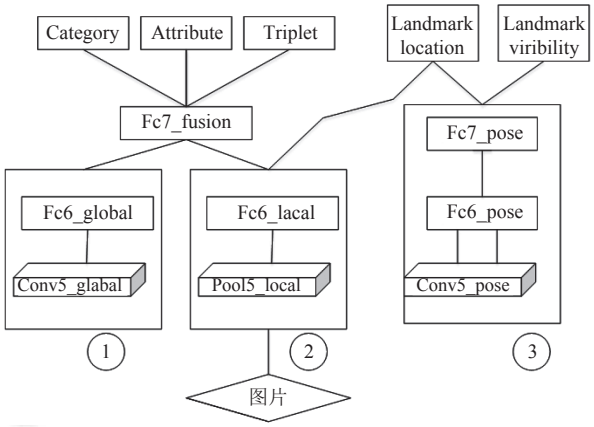

图 1 fashionnet 网络结构

2 目标检测与 R-CNN 系列

2.1 目标检测

目标检测算法就是给定一张图片, 对于图形中的物体进行分类和定位, 传统的目标检测采用的是基于滑动窗口的, 并且图像的特征是人工设计的特征, 不仅计算复杂度高, 而且检测精度比较低. 随着深度学习的发展, 基于深度学习目标检测算法已经逐渐取代了传统的目标检测算法, RCNN[12]系列的目标检测算法是 Two-Stage[13]的检测算法, 首先在图片上选出 Proposals Region, 然后在建议区域上进行特征提取, 随后在最后面加上分类头和回归头进行分类和定位. 随着检测算

法的发展, SPP-Net[14]的提出解决了全连接层必须输入固定尺度大小的特征图导致的原始图像经过 wrap 和 crop 带来的拉伸变形问题, 它在最后一个卷积层上加入了金字塔池化层, 使得全连接层的输入固定. 这样输入图像就可以是任意尺寸的, 输出还是一个固定维数的特征向量然后送到全连接层. 这种结构随后被加入到 RCNN 中, Fast R-CNN 就是对于整张图片进行特征提取, 然后找到每个 Proposals Region 在特征图上的映射 patch, 将 patch 作为每个 Proposals Region 的卷积特征输入到 SPP 层和后面的层, Faster R-CNN[15]提出了新的 RPN 结构, 已经成为目标检测的主流框架, 被很多检测网络使用.

2.2 Faster R-CNN 与 Mask R-CNN

Faster R-CNN 是 Girshick RB 在他设计的 R-CNN 和 Fast R-CNN 基础上改进出来的. Faster R-CNN 无论在速度还是精度上都都比之前的框架有明显的提升, 它是一种基于双阶段的检测算法, 之前的 Fast R-CNN 首先使用 selective search 选出 2000 个 Proposals Region, 然而速度的运算瓶颈在于候选区域的生成时间. 候选区域的生成时间严重制约了目标检测的时间, 并且候选区域的好坏直接影响了后续目标检测任务的结果. 设计一种含有少量的候选区域而且质量很高的结构进行候选区域提取就显得很重要. Faster R-CNN 就设计了 RPN 结构, RPN 直接利用了 CNN 产生候选区域, 不再使用 selective search, 而是使用卷积网络在整张图上进行特征提取, 并在卷积层的最后一层利用 RPN 网络滑动, 利用网络的边框回归机制和 Anchor 机制来得到多个面积不一, 多个不同的长宽比的候选区域. Faster R-CNN 真正意义上实现了端到端的网络训练, 将目标候选框的查找和目标的分类两个步骤结合起来, 在速度和精度上都有不错的提高. Cascade R-CNN[16]在此网络的基础上在后面的检测器堆叠了几个级联模块, 并采用不同的 IOU 阈值训练, 进一步提高了准确率. IOU-Net[17]改进 NMS 算法, 本文就是采用了这个网络进行训练, 得到了很好的效果.

Mask-R-CNN[18]扩展了 Faster R-CNN 结构, 在后面添加了一个和分类头, 回归头并行的机制 mask 层, 这一层是一种二值的 mask, 可以实现逐像素的检测. 并且在细节上为了提高特征表达能力, 使用 ResNet 替换了 Faster R-CNN 中的 VGG 基础网络. 为了挖掘多尺度的特征, 作者使用了 FPN 网络, FPN 网络可以将金字塔高级特征传递到底层, 每一层都是高层特征与底层特征的组合. 使得网络的特征图同时具备细节和局部信息. 由于此网络的 mask 层是逐像素的检测, 所以会出现 misalignment 问题, 于是作者提出了 ROIAlign 结构, 之前的 ROI Pooling 会经历两个量化的过程, 第一个就是从 ROI Proposals 到特征图进行映射的过程, 第二个就是从特征图划分成 7×7 的 bin 的过程, 由于四舍五入将会导致量化的误差. ROIAlign 结构可以使用双线性插值, 不再进行四舍五入, 将候选区域分割成 $k\times k$ 个单元, 在每个单元中计算固定的 4 个坐标的位置, 用双线性内插的方法计算出 4 个位置的值, 然后进行最大池化操作. Mask-RCNN 损失函数就是分类加上回归和 Mask 的预测的损失之和. Mask 使用了 sigmoid 的函数, 这样避免了类间的竞争. 基于 Mask R-CNN 的这些优点, 本篇文章将 mask-rcnn 运用到服装检测中, 建立了服装检测模型.

3 基于 R-CNN 的服装属性识别

Faster R-CNN 候选区域的生成使用了 RPN 网络, 采用一种 Anchor 机制, 既保证了精度又提高了速度. 在候选区域的生成过程中, 采用的主要算法如下: 以 IOU 为分类指标, 将候选框分为正负样本两类.

$$F_{\mathrm{IOU}}=\frac{S_a\cap S_b}{S_a\cup S_b} \tag{1}$$

其中, S_a代表预测框, S_b代表是标记框, F_{IOU}代表的这两者之间的重叠率, 这个比值越高, 预测效果越好. 值大于 0.7, 就标记为正样本, 如果小于 0.3 就标记为负样本, 但是在进行候选框的选取时候, 传统的 NMS 算法采用的是分类得分作为评判标准, 然而在实验中就会发现这种方式不能很好地表征预测框的准确性, 得分高的预测框并不一定比得分低的预测框好, 就会使一些分类得分低但是与S_b的值很接近的框被抑制过滤掉, 导致最终的得分指标下降. 本文采用了 IOU-Net 的思想, 引入了回归得分作为评判标准. 每次计算预被剔除的框的的分类得分和之前的计算的最高的分类得分的最高值, 保留最高的分类得分, 这样返回的 IOU 的预测框的得分不一定是此框的原始得分, 而是以该预测框为依据提出框的最高分类得分.

Mask R-CNN 不再使用原始的 ROI 池化, 因为 ROI 池化将对应的候选区域池化成固定尺寸的特征图的时候, 两次量化结果的输出导致最后得到的特征图

和原来的特征图像素位置偏移, 不能很好的匹配. Mask R-CNN 使用了 ROIAlign 方法, 不再取整, 而是采用双线性差值的方法固定四个点坐标的像素值, 使得不连续的操作变得连续起来. 且定义了三个阶段的损失, 总的损失函数定义为:

$$L_{\text{loss}} = L_{\text{categories}} + L_{\text{BBox}} + L_{\text{mask}} \quad (2)$$

4 实验

本文实验目的主要有两个, 其一是在原有的数据集上面进行微调训练自己的数据集, 得到一个新的模型 ClothNet, 然后将这个模型作为检测网络的基础模型, 继续微调送入到 Faster R-CNN 检测网络. 其二在预训练模型的基础上使用了 Mask R-CNN 结构训练检测网络, 对比这两种算法的结果.

4.1 数据集介绍

本节实验所使用到的数据集有 deepfashion, coco 数据集. DeepFashion 是香港中文大学开放的一个 large-scale 数据集, 包含 80 万图片, 包含了不同角度, 不同场景, 买家秀, 卖家秀的图片, 总共有四个任务, 分别是服装类别和属性预测, in-shop 和 c2s 服装检测点, 外接矩形检测. 每张图片具有丰富的标注信息, 包括类别, 属性, bbox, 特征点等信息. 本文主要做服装的属性识别与分类, 所选的五个数据集中的一个, category-attribute, 里面包含了 289 222 张服装图片, 50 个类别, 1000 个服装属性, 每张图片是否标注了 Bounding Box 和服装类型 clothing type. Bounding Box 标注了左上角位置和右下角位置, 每张图片的最长边 resize 到 300, 保持原图片的长宽比. 服装类型分为上身服装, 下身服装, 全身服装这三个类别, 属性包括五类, 纹理, 面料, 形状, 部分和风格. 数据的可视化结果如图 2 所示.

我们在实验中抽取了这个数据集单个类别个数超过 2000 的 20 个类别, 每个类别抽取 2000 张图像, 构建子数据集 deepfashionkid 在 VGG16[19]上进行微调.

4.2 实验总体框架

由于本实验解决的问题是服装的目标检测, 本文做了一组对比实验, 并在改进的网络结构上做了实验. 实验如下:

首先对于 Faster R-CNN, 首先使用 VGG16 在 imagenet 的预训练模型上进行 fine-tuning 使用数据集 deepfashion 的子数据集 deepfashionkid. 将得到的模型 clothNet 保存下来, 方便后面的目标检测进行迁移学习. 算法框架如图 3 所示.

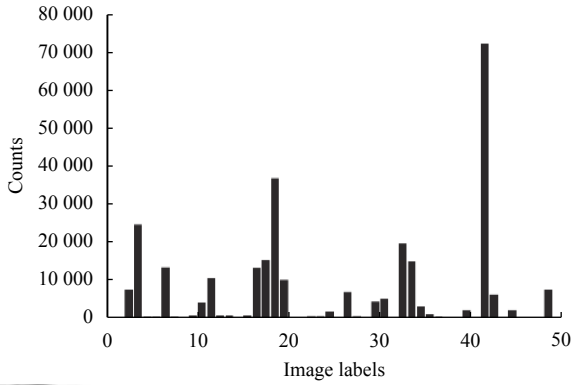

图 2 数据的可视化结果

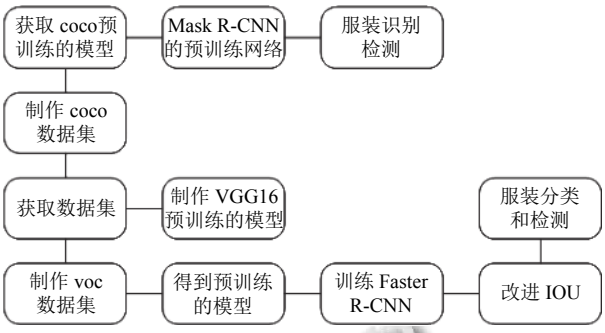

图 3 服装检测系统整体框架图

Deepfashionkid 数据集内容如表 1 所示: 又在 deepfashionkid 数据集中每类抽取 200 张构建目标检测的标记数据集, 分别是 deepfashionVoc 和 deepfashion Mask 两个数据集, 如表 2 和表 3 所示.

4.3 实验结果分析

嫁接学习的训练方式将数据集分成两部分, part1 和 part2, part1 部分的数据量远大于数据 part2, 首先训练 part1 部分得到的模型基础上训练 part2. Faster R-CNN 采用的是 clothNet 作为预训练的模型, 然后使用了 deepfashionVoc 数据集进行目标检测. Mask R-CNN 采用了 coco 数据集的与训练的模型, 在这个基础上进行微调, 进而得到目标检测的结果. 使用 labelme 进行标注, 因为数据集中的服装基本上都是穿在人身上的, coco 数据集也有关于人的关键点检测模型, 可以拿来作为预训练模型, 在这个基础上从头训练, 得到的实验结果.

表 1 各算法采用的数据集个数

	数据集	VGG16	Faster R-CNN	Mask R-CNN
1	Blouse	2000	200	200
2	Blazer	2000	200	200
3	Cardigan	2000	200	200
4	Coat	2000	200	200
5	Dress	2000	200	200
6	Hoodie	2000	200	200
7	Jacket	2000	200	200
8	Jeans	2000	200	200
9	Joggers	2000	200	200
10	Jumpsuit	2000	200	200
11	Kimono	2000	200	200
12	Leggings	2000	200	200
13	Romper	2000	200	200
14	Shorts	2000	200	200
15	Skirt	2000	200	200
16	Sweater	2000	200	200
17	Sweatpant	2000	200	200
18	Tank	2000	200	200
19	Tee	2000	200	200
20	Top	2000	200	200

本节主要评估服装属性识别在 deepfashion 数据集上对比实验, 并且在此数据集上改进实验.

本次的实验方案参照 BTIC[20]的训练 trick 进行训练的, 线性扩展学习率, 随着 batchsize 的增大, 学习速率也在增大. 在刚开始训练的时候使用一个小的学习率训练几个 epoch, 然后增大学习率, 并且采用 16 位的浮点型参数进行训练. 在训练的过程中, 学习的初始速率为 0.02, 每经过 5000 次迭代就将, 学习的速率就会相应的减少 0.1.

表 2 deepfashionVoc 数据集上实验结果对比

	Mean acc.	Mean IoU
Faster R-CNN	78.3	38.5
Faster R-CNN +clothNet	80.2	39.2

在自己手动标注的数据集上进行 Mask R-CNN 的实验, 根据不同的样本数据, 实验结果如表 3.

表 3 deepfashionMask 数据集上实验结果对比

	Pixel acc.	Mean acc.	Mean IoU
Mask R-CNN	85.2	51.7	39.5
Mask R-CNN +clothNet	88.3	53.5	45.2

由此发现, 随着样本数量的增加, mAP 的值也会提升, 随着网络模型深度的增加, mAP 的值相对于浅层模型提高的更为明显, 增加数量可以提高模型的准确度, 但是增加一定的数量就会导致过拟合, 目标检测多框或者框的位置不够准确. 使用新的数据在原有的网络上进行微调, 并且在检测网络上更换数据, 进一步微调对于检测效果的提高作用很明显. 并且损失函数是影响模型检测的一个重要方面.

5 总结与展望

本文自己构建了一个新的数据集 deepfashionkid, 使用这个数据集在 VGG16 上面进行微调得到了一种新的服装检测模型叫做 clothNet, 并且在这个数据集上进一步选取 deepfashionVoc 数据集和 deepfashion-Mask 数据集. 使用 clothNet 预训练的模型加上 deepfashionVoc 进行 Faster R-CNN 的目标检测. 接着采用了 coco 数据集预训练的模型进行微调, 训练 Mask R-CNN 网络. 得到服装检测模型. 本文在训练网络的时候, 采取了 BTIC 论文中提到的很多训练技巧. 本文仍有改进之处, 在 IOU 的处理上可以参照 Cascade R-CNN 和 IOUNet 的结构和想法运用到这个网络上, 结果可能会更好. 本文只是做了服装的分类和定位, 数据集也是选取了部分数据集, 对于这个数据集, 上面仍有很多的关键点标记数据和一些特定属性类别的数据, 完全可以采用多任务和多标签学习的方式, 重新定义损失函数, 进一步提高服装属性和类别的准确度, 扩张整个算法框架, 运用到更多的应用中, 比如智能衣柜, 提高“以图搜图”的准确度.

参考文献

1 包青平, 孙志锋. 基于度量学习的服装图像分类和检索. 计算机应用与软件, 2017, 34(4): 255–259. [doi: 10.3969/j.issn.1000-386x.2017.04.043]

2 Pan RR, Gao WD, Liu JH, *et al*. Automatic recognition of woven fabric pattern based on image processing and BP neural network. The Journal of the Textile, 2011, 102(1): 19–30.

3 Salem YB, Nasri S. Automatic recognition of woven fabrics based on texture and using SVM. Signal, Image and Video Processing, 2010, 4(4): 429–434. [doi: 10.1007/s11760-009-0132-5]

4 Liu S, Feng JS, Domokos C, *et al*. Fashion parsing with weak color-category labels. IEEE Transactions on Multimedia, 2014, 16(1): 253–265. [doi: 10.1109/TMM.2013.2285526]

5 Bourdev L, Maji S, Malik J. Describing people: A poselet-based approach to attribute classification. Proceedings of 2011 IEEE International Conference on Computer Vision.

Barcelona, Spain, 2011: 1543–1550.
6 Hu ZL, Yan H, Lin XG. Clothing segmentation using foreground and background estimation based on the constrained Delaunay triangulation. Pattern Recognition, 2008, 41(5): 1581–1592. [doi: 10.1016/j.patcog.2007.10.005]
7 王林, 张鹤鹤. Faster R-CNN 模型在车辆检测中的应用. 计算机应用, 2018, 38(3): 666–670. [doi: 10.11772/j.issn.1001-9081.2017082025]
8 Szegedy C, Ioffe S, Vanhoucke V, *et al*. Inception-v4, inception-ResNet and the impact of residual connections on learning. arXiv: 1602.07261, 2016.
9 Lin TY, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection. arXiv: 1612.03144, 2016.
10 Liu ZW, Luo P, Qiu S, *et al*. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA, 2016: 1096–1104.
11 尹宝才, 王文通, 王立春. 深度学习研究综述. 北京工业大学学报, 2015, 41(1): 48–59.
12 Girshick R. Fast R-CNN. arXiv: 1504.08083, 2015.
13 Uijlings JRR, Van De Sande KEA, Gevers T, *et al*. Selective search for object recognition. International Journal of Computer Vision, 2013, 104(2): 154–171. [doi: 10.1007/s11263-013-0620-5]
14 He KM, Zhang XY, Ren SQ, *et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904–1916. [doi: 10.1109/TPAMI.2015.2389824]
15 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]
16 Cai ZW, Vasconcelos N. Cascade R-CNN: Delving into high quality object detection. arXiv: 1712.00726, 2017.
17 Jiang BR, Luo RX, Mao JY, *et al*. Acquisition of localization confidence for accurate object detection. arXiv: 1807.11590, 2018.
18 He KM, Gkioxari G, Dollar P, *et al*. Mask R-CNN. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018. [doi: 10.1109/TPAMI.2018.2844175]
19 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556, 2014.
20 He T, Zhang Z, Zhang H, *et al*. Bag of tricks for image classification with convolutional neural networks. arXiv: 1812.01187, 2018.