E-mail: jig@radi.ac.cn
Website: www.cjig.cn
Tel: 010-58887035

中图法分类号: TP751.1　　**文献标识码:** A　　**文章编号:** 1006-8961(2020)09-1754-19

论文引用格式: Shen F C, Zhang P, Luo J, Liu S Y and Feng S J. 2020. Scale changing in general object detection: a survey. Journal of Image and Graphics, 25(09): 1754-1772(申奉璨, 张萍, 罗金, 刘松阳, 冯世杰. 2020. 目标检测中的尺度变换应用综述. 中国图象图形学报, 25(09): 1754-1772) [DOI: 10.11834/jig.190624]

目标检测中的尺度变换应用综述

申奉璨, 张萍, 罗金, 刘松阳, 冯世杰

电子科技大学光电科学与工程学院, 成都　610054

摘　要:　目标检测试图用给定的标签标记自然图像中出现的对象实例, 已经广泛用于自动驾驶、监控安防等领域。随着深度学习技术的普及, 基于卷积神经网络的通用目标检测框架获得了远好于其他方法的目标检测结果。然而, 由于卷积神经网络的特性限制, 通用目标检测依然面临尺度、光照和遮挡等许多问题的挑战。本文的目的是对卷积神经网络架构中针对尺度的目标检测策略进行全面综述。首先, 介绍通用目标检测的发展概况及使用的主要数据集, 包括通用目标检测框架的两种类别及发展, 详述基于候选区域的两阶段目标检测算法的沿革和结构层面的创新, 以及基于一次回归的目标检测算法的3个不同的流派。其次, 对针对检测问题中影响效果的尺度问题的优化思路进行简单分类, 包括多特征融合策略、针对感受野的卷积变形和训练策略的设计等。最后, 给出了各个不同检测框架在通用数据集上对不同尺寸目标的检测准确度, 以及未来可能的针对尺度变换的发展方向。

关键词: 图像语义理解; 通用目标检测; 卷积神经网络; 尺度变换; 小目标检测

Scale changing in general object detection: a survey

Shen Fengcan, Zhang Ping, Luo Jin, Liu Songyang, Feng Shijie

School of Optoelectronic Science and Engineering, University of Electronic Science and Technology of China, Chengdu　610054, China

Abstract: General object detection has been one of most important research topics in the field of computer vision. This task attempts to locate and mark an object instance that appears in a natural image using a series of given labels. The technique has been widely used in actual application scenarios, such as automatic driving and security monitoring. With the development and popularization of deep learning technology, the acquisition of the semantic information of images has become easier; thus, the general object detection framework based on convolutional neural networks (CNNs) has obtained better results compared with other target detection methods. Given that the large-scale dataset of the task is relatively better than datasets designed for other vision tasks and the metrics are well defined, this task rapidly evolves in CNN-based computer vision tasks. However, general object detection tasks still face many problems, such as scale and illumination changes and occlusions, due to the limitations of the CNN structure. Given that the features extracted by CNNs are sensitive to the scale, multiscale detection is often valuable but challenging in the field of CNN-based target detection. Research on scale transformation also has reference value for other scales in small target- or pixel-level tasks, such as the semantic segmentation and pose detection of images. This study mainly aims to provide a comprehensive overview of object detection strategies for scales in CNN architectures, that is, how to locate and classify different sizes of targets robustly. First, we introduce the development of general target detection problems and the main datasets used. Then, we introduce two categories of the gen-

收稿日期: 2019-12-13; **修回日期:** 2020-03-02; **预印本日期:** 2020-03-09

基金项目: 四川省科技计划项目(2018GZ0166, 2019YFG0307)

Supported by: Science and Technology Planning Project of Sichuan Province, China (2018GZ0166, 2019YFG0307)

eral object detection framework. One of the categories, i. e. , two-stage strategies, first obtains the region proposals and then selects the proposals by points of classification confidence; it mostly takes region-based convolutional neural networks (RCNN) as the baseline. With the development of the RCNN structure, all the links are transformed into specific convolution layers, thus forming an end-to-end structure. In addition, several tricks are designed for the baseline to solve specific problems, thus improving the robustness of the baseline for all kinds of object regions. The other category, i. e. , one-stage strategies, obtains the region location and category by regressing once; it starts with a structure named "you only look once" which regresses the information of the object for every block divided. Then, the baseline becomes convolutional and end to end and uses deep and effective features. This baseline has also become popular since focal loss has been proposed because it solves the problem in which regression may cause an unbalance of positive and negative samples. Besides, some other methods, which detect objects via point location and learn from pose estimation tasks, also obtain satisfactory results in general target detection. We then introduce a simple classification of the optimization ideas for scale problems; these ideas include multi-feature fusion strategies, convolution deformations for receptive fields, and training strategy designs. Multi-feature fusion strategies are used to detect the classes of objects that are not always performed in a small scale. Multi-feature fusion can obtain semantic information from different image scales and fuse them to attain the most suitable scale. It can also effectively identify the different sizes of one-class objects. Widely used structures can be divided as follows: those that use single-shot detection and those with feature pyramid networks. Some structures have a jump layer fusion design. In a receptive field, every feature corresponds with an image or lower-level feature. The specific design can solve a target that always appears small in the image. The general receptive field of a convolution is the same as the size of the kernel; another special convolution kernel is designed. Dilated kernels are the most deformed kernels, which are used with the designed pooling layer to obtain a dense high-level feature. Some scholars have designed an offset layer to attain the most useful deformation information automatically for the convolution kernel. A training strategy can also be designed for small targets. A dataset that only includes small objects can be designed, and different sizes of the image can be trained in the structure in an orderly manner. Resampling images is also a common strategy. We provide the detection accuracy results for different sizes of targets on common datasets for different detection frameworks. Results are obtained from the Microsoft common objects in context (MS COCO) dataset. We use average precision (AP) to measure the result of the detection, and the result set includes results for small, medium, and large targets and those for different intersection-over-union thresholds. It shows the influence of the changes for scale. This study provides a set of possible future development directions for scale transformation. It also includes strategies on how to obtain robust features and detection modules and how to design a training dataset.

Key words: image semantic understanding; general object detection; convolutional neural network (CNN); scale changing; small target detection

0 引 言

目标检测是计算机视觉领域具有挑战性和实际意义的底层问题。检测准确度的提高可以极大提升计算机视觉领域中其他高级语义检测任务的准确程度,如图像中特定类别的检测、显著物体的检测、姿态检测等。同时,目标检测已广泛应用于监控安全、无人驾驶、无人机场景监控和交通监控等实际任务中,因此目标检测问题的研究具有很重要的现实意义。

一般来说,目标检测包括目标定位和目标分类两个目的。目标检测想要获得的信息为类别及该类别的目标位置,包括定位坐标和目标框的长宽。坐标位置由算法决定,一般为左上角点或中心点。两种信息有其一时,获取另一种的准确度较高。如应用较为广泛的人脸识别就是典型的已知分类寻求定位的问题,准确度可以达到 99% 以上(Schroff 等,2015; Sun 等,2015; Deng 等,2018)。但通过特定结构无先验地同时获得分类和定位信息是一个大的挑战。

传统目标检测方法的主要流程是输入图像预处理、筛选优化候选区域、候选框特征提取、候选框特征分类和目标位置边框回归。首先,对输入图像进行预处理,包括去噪、增强、剪裁以及尺寸伸缩重采样等。其次,利用优化后的滑动窗口方法筛选和优化候选区域。包括对象法(Alexe 等,2012)、选择性

搜索法(Uijlings 等,2013)、有限制的最小剪裁法(Carreira 和 Sminchisescu,2012)、多尺度图聚类法(Arbeláez 等,2014)、Edgebox(Zitnick 和 Dollár,2014)和贝叶斯算法(Zhang 等,2015)等。然后,对候选框进行特征提取。较鲁棒的特征包括尺度不变特征(scale invariant feature transform, SIFT)(Le 等,2011)、方向梯度直方图(histogram of oriented gradients,HOG)(Dalal 和 Triggs,2005; Ren 和 Ramanan,2013)和可变组件模型特征(deformable part model, DPM)(Felzenszwalb 等,2008; Dean 等,2013)等。最后,使用分类算法对提取的候选框中的特征进行分类,通过分类结果得到目标框中的内容所属的类别,依据该类别进行目标位置的边框准确回归。常见的分类方法包括有监督的分类方法,如支持向量机(support vector machine, SVM)(Zhang 等,2016)、Adaboost 等,和无监督的分类方法(Sermanet 等,2013)。然而,传统目标检测方法存在许多缺陷,如提取的特征不够普适和鲁棒,不同的特征需要适配不同的分类器,以及任务不可以迁移等。

随着计算能力的极大提升,对图像高阶特征理解十分友好的卷积神经网络(convolutional neural network,CNN)应运而生,图像的抽象特征的提取逐渐由手动设置的特征转向自动提取的特征。这种抽象特征能够在各种维度对图像进行抽象化,可以用来完成许多图像理解任务(张顺 等,2019),包括图像分类(Simonyan 和 Zisserman,2015; He 等,2016; Huang 等,2017; Xie 等,2017)、图像语义分割(Chen 等,2017a; Chen 等,2017b)、图像标题生成(Xu 等,2015; Jia 等,2015)、人脸识别(Schroff 等,2015; Sun 等,2015; Deng 等,2018)和图像融合(Li 等,2019; Du 等,2019; Tang 等,2017)等。图像分类是 CNN 的首项任务,常见的方法包括 VGG(visual geometry group)-Net(Simonyan 和 Zisserman,2015)、ResNet(He 等,2016)、ResNeXt(Xie 等,2017)、DenseNet(Huang 等,2017)、SE-Net(Hu 等,2018)、Xception(Chollet,2017)和 MobileNet(Howard 等,2017; Sandler 等,2018)等。

在计算机视觉任务中,目标检测发展较快,原因有两点。1)训练集易于获取。对输入与输出都是图像的任务,目标检测任务训练集中每幅图像的标签一般用向量表示,对图像只需要标定而不需要处理,因此产出了许多大规模数据集,训练出的大型特征提取结构也较多。2)衡量标准易于界定。对目标框来说,确认框定正确与否很容易量化,因此可以设计出有效的损失函数。相对其他计算机视觉任务,目标检测的发展十分蓬勃。

但目标检测任务中还存在很多问题,如无法应对遮挡、如何将目标进行细分类以及如何将目标精确定位等。在这些问题中,尺度问题一直是影响检测效果的关键问题。由于技术上基本使用卷积神经网络作为特征提取结构,而卷积神经网络提取的特征又存在尺度敏感的特性,因此多尺度检测在基于深度学习的目标检测领域具有很大挑战。

目标检测领域有很多大规模数据集,包括 ILSVRC(ImageNet large scale visual recognition challenge)(Russakovsky 等,2015)、Pascal VOC(visual object class)(Everingham 等,2010)、SUN(scene understanding)(Xiao 等,2016)等通用目标数据集。此外,还有一些针对某类物体的特定数据集。VisDrone 2018(Zhu 等,2018)是基于无人机拍摄的视频进行标定的数据集、Caltech(Dollar 等,2012)和 CityPersons(Zhang 等,2017)是包含行人设计的数据集、KITTI(Karlsruhe Institute of Technology & Toyota Technological Institute)(Geiger 等,2012)是自动驾驶场景下的算法评价数据集。应用较为广泛的是 MS COCO(Microsoft common objects in context)数据集(Lin 等,2014),其包括大量具有挑战性的标注图像,并设有检测挑战赛。训练特定网络时,一般需要对这些大规模数据集内的数据进行筛选(Chen 等,2017b)。部分上述数据集包含的数据量如表 1 所示。其中测试图像量包括验证集图像量和测试集图像量,在表 1 中分别给出。

表 1 主要的通用目标数据集比较

Table 1 Comparation of datasets for general object detection

数据集	类别数	训练集图像量/KB	测试集图像量/KB
PASCAL VOC 2007	20	7.5	2.5+5
PASCAL VOC 2012	20	17	5.8+11.5
ILSVRC 2014	200	450	20+40
MS COCO 2014	91	82.8	40.5+40.8
MS COCO 2015	91	165.5	81+81

在 MS COCO 给出的衡量标准中,准确率在尺度层面细分为小目标准确率、中等目标准确率和大目标准确率。小目标指真值框小于 32 × 32 像素的目标,大目标指真值框大于 96 × 96 像素的目标,其余为中等目标。这种分类方式为通用目标中不同尺度的目标提供了精确定义。

卷积神经网络使用多特征的融合、感受野设计及训练策略设计等方式对多尺度目标检测进行改进,对不同类型的多尺度目标产生了针对性的效果。

从结果来看,基于尺度变换问题的结构改善对目标检测结果的提高十分明显,因此更好地设计模块将尺度检测更加鲁棒是十分必要的。尺度问题需要更加细密的特征,也需要更加有效的融合特征,因此有效解决尺度问题可以使更多领域的问题得到解决,如超分辨问题和目标细分类问题等。

1　基于卷积神经网络的目标检测模型的演变

随着计算机计算能力的显著提升,基于卷积神经网络的通用目标检测算法开始兴起。相对于传统的基于手动设定特征提取的目标检测模型,卷积神经网络具有更强大的特征表达能力和更优秀的泛化能力,鲁棒性强,且计算速度下降到合理范围之内。应用深度学习框架的目标检测算法逐步替代了其他方法,成为主流算法。基于卷积神经网络的目标检测算法主要分为两种: 第 1 种是基于候选区域的方法,即先选取候选区域,后对候选区域进行分类和回归的算法,称为两阶段算法; 第 2 种是基于回归的目标检测方法,同时对图像进行分类和候选框参数的回归,摒弃了多次回归的步骤,称为一阶段算法。本文介绍的方法如图 1 所示。

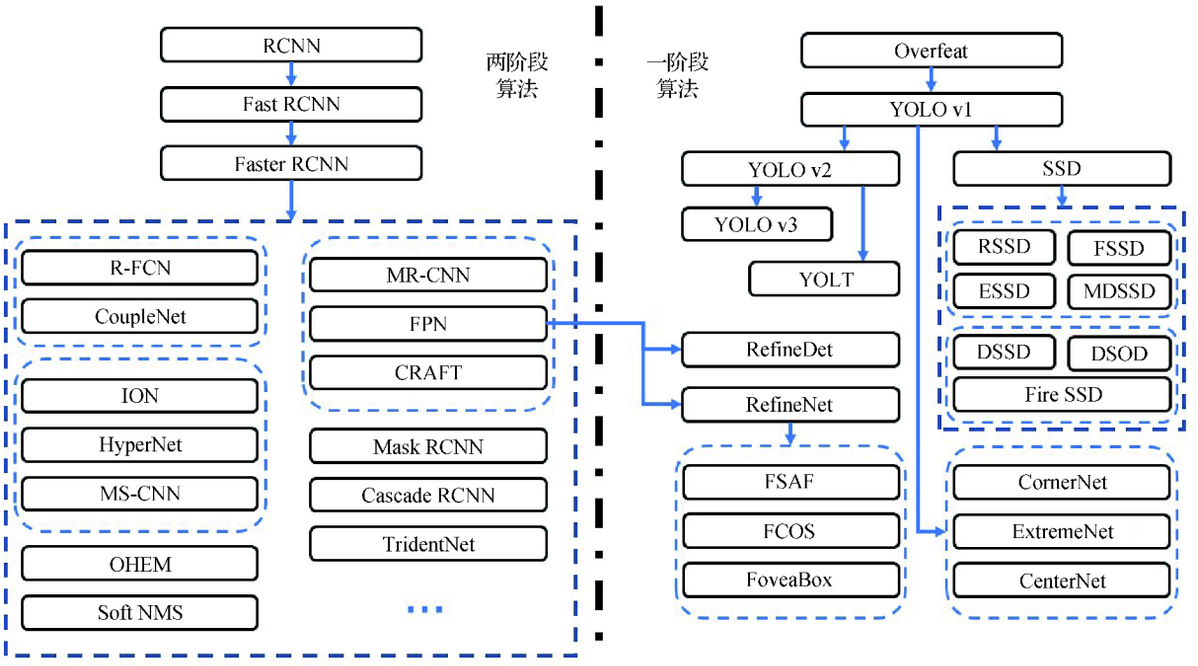


图 1　本文提及方法的发展拓扑图

Fig. 1　Development map for the methods mentioned in the paper

1.1　基于候选区域的目标检测算法

基于候选区域的目标检测算法分为两步。第 1 步是对图像进行目标框的初步框定,获得一系列候选区域; 第 2 步是对得到的候选区域进行分类和调整,得到更加准确的结果。

1.1.1　RCNN 系列结构的演变

Girshick 等人(2014) 提出的候选区域卷积网络(region convolutional neural network,RCNN) 算法是 RCNN 系列算法的奠基之作,在此基础上,RCNN 系列结构不断优化,形成了一整套效果可观的目标检测方案。算法具体如下: 首先通过选择性搜索算法生成 2 000 个候选区域,然后使用 CNN 网络进行特征的提取,最后使用类线性支持向量机分类器对这些区域进行分类处理,效果比传统的目标检测算法

有大幅提升，但在速度、精度和鲁棒性上仍然有很大的提升空间。

在 RCNN 方法中，需要将输入 CNN 的候选区域归一化为统一尺寸，这样会失去较多信息。为此，He 等人(2015) 在 CNN 中加入了空间金字塔池化层，减少了信息损失，这个结构称为 SPPNet(spatial pyramid pooling network)。该层输出的结果是从不同大小的特征图中提取出的相同长度的特征向量，因此只需要对原图做一次 CNN 处理，而不是对每一个候选区域进行卷积处理，可以节省大量的计算时间。

在 SPPNet 算法的基础上，Girshick(2015) 提出改进版的 Fast RCNN 算法，将特征提取部分和分类部分融合为一个网络，对每个感兴趣区域(region of interest，ROI) 都输出多分类结果和边框回归两个向量。该方法使用多任务损失，训练出一个端到端的网络，进一步加快了运算速度。

Fast RCNN 依然是候选框提取和特征分析两部分分离的结构，在此基础上，Ren 等人(2017) 提出了 Faster RCNN 算法，使用候选区域生成网络(region proposal network，RPN) 代替选择性搜索算法，将结构分为两部分，分别是候选区域提取的全卷积网络以及后续的感兴趣区域分类器。RPN 本质是一个无标签的目标框生成器。目标框生成后，用分类器对目标框是否属于特定类进行判定，最后使用回归器进一步调整框定位置。整个网络共享 CNN 提取的特征信息，提高了算法速度和准确率。

RCNN 在发展过程中，可见的趋势是算法结构虽然分为两部分，但所有的处理全部卷积化，即逐渐使用卷积层代替传统的分类、聚类或回归的过程，使得网络可以一体化端到端的训练以及梯度回传。因此在 Faster RCNN 之后，很少有加入非卷积神经网络固有的模块，因为那样既会拖慢运行的速度，又不利于一体化训练。因此在效果不可控的情况下，加入的模块一般使用卷积化模块。在可见范围内，一体化过程可以提高包含 CNN 的方法的运行速度以及准确率。

1.1.2 结构细节的改动

在 Faster RCNN 之后，相较于 RCNN 系列结构本身的改进，更多的是对细节的修改。这些改进致力于对系统的某个细节进行更鲁棒的改进，解决某些特定问题，以获得更好的效果。在 Faster RCNN 的基础上，各种方法对结构的优化如图 2 所示。

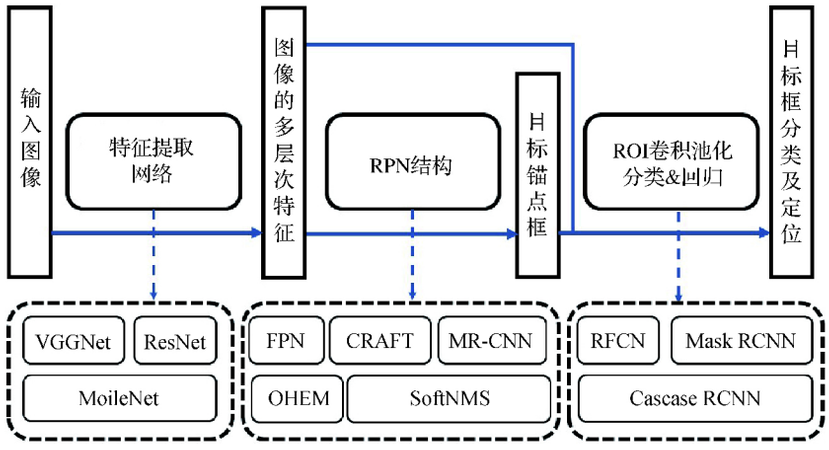


图 2 RCNN 的基本结构及改进结构图

Fig. 2 Basic structure of RCNN and its improvement diagram

RPN 的运算实际上只使用了最后一层的特征，因此很多时候无法覆盖大多数目标，只能覆盖较大或语义较为明显的目标框。为此，Gidaris 和 Komodakis(2015) 提出了 MR-CNN 结构，在 RPN 获取区域后加入剪切层，融合预特征分析层获得的特征，进行空间自适应池化和全连接网络，以得到更准确的候选区域。Lin 等人(2017b) 提出特征金字塔网络(feature pyramid network，FPN) 结构，设计了一个侧路结构，将高层低分辨率的特征与底层高分辨率的特征结合，各层输出各层自己的融合结果。Yang 等人(2016) 提出 CRAFT(cascade region proposal network and fast RCNN) 结构，在 RPN 中加入一个二类检测网络，进行误定位区域筛选，使 RPN 的输出包含一个初步的标签，以便后续分类时获取的

分类精度更加准确。RPN 的改进可以使得目标框(也称锚点框)在开始时便有一个相对准确的分类和回归,以使得后面的处理更加有效。

ROI 分类也存在许多问题,最严重的是平移结果变化较大。为了使 ROI 获得平移不变性,Dai 等人(2016)提出基于区域提取的全卷积网络(region based fully convolutional network,RFCN)结构,在 ROI 池化前加入卷积层先做预测,获得位置敏感特征图,并将得到的位置信息通过得分 ROI 加入池化层。在此基础上,Zhu 等人(2017)提出 CoupleNet,并联了传统 ROI 和得分 ROI,结合全局信息和局部信息以提升效果。通过引入或全局或周边位置信息的方式,平移不变性便得以实现。

除此以外,ROI 还有其他改进。He 等人(2017)提出了 Mask RCNN,在分类回归层中,在分类分支和回归分支之外加入物体模板分支,通过 FCN 子网络获得物体实例的分割,以提高重叠物体的检测准确率。同时,用 ROIAlign 代替 ROI 池化的部分,通过双线性差值代替重采样,消除了采样误差,显著提升了效果。

针对交并比(interestion over union,IoU)阈值的不同设定导致训练欠拟合及过拟合现象,Cai 和 Vasconcelos(2018)提出了 Cascade RCNN 结构,设置级联目标检测器代替单目标检测器,使用不同的 IoU 训练每一个级联块,目标框多次回归使得结果更加接近真实值,因此显著提升了效果,并很好地适配了其他部分的优化。

在候选框预处理方面,RPN 出现之前,Shrivastava 等人(2016)提出针对选择性搜索获取候选框的优化算法 OHEM(online hard example mining),在 ROI 层面通过回传优化得到有效的无效框抑制效果。RPN 出现之后,Bodla 等人(2017)在训练时用软性非极大值抑制算法代替实际的非极大值抑制算法,使用阈值的置信度而不是其本身来对结果进行筛选。

基于候选区域的目标检测算法的优缺点都十分明显,优点:1)特征被两次利用,也就是说会回归两次来获得精准定位和大小,使得准确率相对较高;2)由于在样本框定阶段限制了正负样本的数量,因而在分类时避免了样本的不均衡性,使得分类器得以完善训练;3)由于第 1 次回归产生一系列无标签目标框,因此网络在进行针对性修正时只需要在第 2 步中微调,即结构对任务的迁移性较强。缺点:1)运用了两次回归,导致速度不及一次回归的方法;2)对整个结构来说,两阶段需要的参数较多,因而需要更大的数据集使结构回归至较优的情况;3)这类结构的准确率和两个阶段都有关系,因此每一步准确率都会影响最后的结论,即鲁棒性较弱。

对于这一类结构,尺度变换上的挑战比较清晰,因为这类结构中往往将尺度上的分层单立为一个模块加入在候选区域提取模块和检测模块,这种方式称为多参考检测(multi-reference detection)。因此在尺度层面最大的挑战在于如何精确设计这一部分的结构使得所有不同尺寸的候选区域在这个模块得到对应精准的分类和回归。

1.2 基于一次回归的目标检测算法

在两阶段算法中,输入图像无论是选择性搜索算法,还是 RPN 结构,以及其一系列变种,都是先对目标框进行初步回归,筛选后再对框内的物体进行分类和精细化回归。一阶段算法和两阶段算法的本质差别是摒弃多步回归的思路,通过直接回归目标框进行物体的定位搜索,同时给出类别信息,将目标识别分类和目标定位视为一个问题,回归出一个同时包含两种信息的向量。

虽然一阶段算法比两阶段算法更快速,但相对来说,依然需要更加有效和准确的回归想法或思路。Sermanet 等人(2014)首先使用了这种一体化结构,提出了名为 Overfeat 的结构,采用多尺度滑窗卷积的方式,同时进行分类、定位和检测。该尺度回归的算法较现在的算法较为粗糙,后续相继出现了一些方案对其优化。现阶段的主流方案有采用整图离散采样回归思想的 YOLO 系列、针对多尺度特征回归思想的 SSD 系列以及在 2018 年中段左右开始流行的非锚点框(anchor-free)算法。

1.2.1 YOLO 系列目标检测算法

YOLO(you only look once)是 Redmon 等人(2016)提出的基于单整体神经网络的目标检测算法。YOLO 将目标检测问题转化为一个回归问题,通过输入的图像在多个位置输出回归框位置和对应类别。这样做的好处减少了大量的候选锚点框,可以更好地避免误检区域,对物体的精准分类能力要好于两步算法。

YOLO 的第 1 代结构在 2015 年提出(Redmon

等,2016),在第1代中,算法将图像分成7×7个子区域,对每个区域回归该区域的两个可能的目标框及可能的类别,共1 470个参数(设定为20类),最后进行非极大值抑制,筛选出结果目标框。第1代的不足有两点:1)定位上由于只有7×7个子区域,很容易产生物体的定位错误;2)对小物体,尤其是密集的小物体,漏检率较高,原因是一个栅格只能预测两个物体。

经过改进,YOLO系列于2017年推出第2代(Redmon和Farhadi,2017)。在第2代中,设计了19层DarkNet网络,加入了批量归一化的预处理,采用两种尺寸的ImageNet数据集图像对网络进行10轮预训练后,再进行网络微调,以适应高分辨率的图像。此外,借鉴anchor的思想,重点解决了召回率和定位精度的问题。将图像栅格从7×7个子区域细化为13×13个子区域,每个区域获取5个不同尺度的anchor。在每个anchor内回归置信度、目标位置及类别参数。全连接层会严重拖慢网络,因此该方法采用卷积层降采样的方式,将图像从416×416的输入大小,降采样为13×13的特征图大小,获取anchor。这样处理后,准确率虽有小幅度下降,但召回率上升明显。算法结构中还加入了基于IoU的聚类方法来调整目标区域的精度,以及加入了强约束方法来保证中心点不发生震荡。在训练阶段,加入了多尺度特征跃迁层,训练时也加入多尺度输入训练法。这些设计相较于第1代效果都有一些提升。

2018年,YOLO系列推出了第3代结构(Redmon和Farhadi,2018),在第2代基础上,设计了一个新的特征提取网络结构,从19层网络扩展成53层网络,准确率好于19层DarkNet,速度快过残差网络。此外,在分类预测上,使用二元交叉熵损失函数代替softmax函数进行类别预测。

从YOLO系列算法的变迁可以看出,一阶段算法具有特征可迁移性强和速度快的优点。但在算法的精确度上,YOLO算法对于目标的分类回归准确度低于同时期的两阶段目标检测算法,原因在于YOLO在回归前使用的区域是手动设定的单区域,不及其他系列中回归出的区域精度高。

1.2.2 SSD系列目标检测算法

相比YOLO使用全连接层后再检测,Liu等人(2016)提出了SSD(single-shot detector)算法,直接使用卷积操作进行检测,使得框架的速度比YOLO更快。主要思路是从不同尺寸的特征中提取先验区域,通过卷积层对这些先验区域进行分类回归。

在最初的SSD算法中,将多尺度特征图作为检测图像。类似于RPN的形式,SSD算法对每个特征点都进行锚点框的框定,并将其作为先验区域。对每个先验区域通过卷积方式同时进行分类和回归,使其一次回归便尽量精确。此外,引入了空洞卷积来获得更加致密的特征图,使其相对粗糙划分区域的YOLO第1代效果有了很大提升。

然而SSD对小目标的回归并不完全鲁棒。问题之一便是特征提取网络的特征不够细致。针对这个问题,有两种解决方法:一种是对特征处理层中不同层的特征进行更加有效的融合;另一种是使用更加先进的特征处理网络进行特征处理。这两种方案都对效果有所提升。前一种方法的改进将在第3节详细阐述。

在特征提取模块的整体替换方面,承接当时较新的残差网络,Fu等人(2017)提出了使用ResNet-101的检测网络DSSD(deconvolutional SSD)。在这个结构中,除了在特征方案上使用残差之外,还在分类卷积层后加入了一系列的反卷积层,参照FCN网络(Long等,2015)设计了一个先下采样再上采样的模式,并在残差模块输出分类结果。而Shen等人(2017)则是将特征网络改为改进后的DenseNet网络(Huang等,2017),放弃微调方案,从头开始训练一个完整的网络,提出了DSOD(deeply supervised object detection)算法。Liau等人(2018)提出了Fire SSD结构,将SqueezeNet(Iandola等,2017)作为主网络,并将其中的残差模块中的卷积改得更小更适合检测使用。

SSD系列算法使用聚类得到的固定尺寸的锚点框将特征图量化,使用卷积操作进行分类回归。这个过程中由于引入了锚点框,不可避免地会由于回归量的增大导致速度减缓。但因使用单个检测器检测多类物体,速度不逊于YOLO,且准确度高于YOLO。但这个结构仍然有其局限性,一是共有的问题,即尺度无法把控,另一个是正样本过多导致的大量无效回归轮次。

1.2.3 其他一阶段目标检测算法

基于锚点框的算法会产生大量的负样本,与正样本存在着数量级的差距,这会导致数据失衡影响检测效果。2018年兴起的anchor free算法放弃了

大量锚点框分类回归的思路,而是通过回归关键点来进行目标检测。

最早的 anchor free 方法是 YOLO 第 1 代,以及 Huang 等人(2017)提出的基于 DenseNet 构造的人脸识别算法。DenseNet 算法使用了像素级回归网络,只是回归的值代表着距离真值框上下左右 4 个坐标的坐标差,再通过非极大值抑制筛选目标框的位置。该算法结合 Unitbox(Yu 等,2016)中提及的 IoU 损失函数,为最新的 anchor free 算法奠定了基础。

最初的 anchor free 方法应用在字符识别中(Zhong 等,2019),通过一个角点分支来降低 FPN 带来的锚点框过多导致速度较慢的问题,但其对目标角点的精确定位为其他方法提供了思路。

Law 和 Deng(2018)在目标的角点检测方面,提出了 corner pooling 的池化方式,通过横纵向扫描峰值获取潜在的目标框角点坐标。使用一个等尺寸特征提取网络 Hourglass(Newell 等,2016)进行特征图的提取,再使用角点池化的方法获取目标框角点信息(包括坐标信息、配对信息以及偏移信息),获得了较为鲁棒的效果。在此基础上,更换了特征提取的网络(Law 等,2019),在不改变效果的基础上,进一步加快了速度,甚至达到了实时效果。相似地,Zhou 等人(2019)也是通过关键点定位,但提出的 ExtremeNet 网络检测的是上下左右边界位置,对每个位置选取 40 个最优解合并后,寻找其中心点,计算置信度并阈值筛选,这种方法在单一尺度下效果比其他方法差,但多尺度下效果较好。Duan 等人(2019)使用的是左上、右下及中心 3 个点的置信度作为筛选标准来进行最优框的获取。Zhou 等人(2019)通过获取一个更加致密的特征图,获取到了局部峰值,将这个局部峰值作为中心点回归目标的大小,也获取了较好的效果,且可移植到很多其他任务中。

除了上述提到的算法之外,还有很多一阶段的算法通过压缩回归次数的方式进行加速。针对一阶段算法中类别不均衡的问题,Lin 等人(2017b)提出了一种新的损失函数 focal loss。这种损失函数通过在类别前加入权重系数的形式,极大地压制了大量负样本引起的梯度失衡。基于该损失函数,还设计了一个名为 RetinaNet 的网络,通过对类 FPN 结构后接子网络(代替原来的 RPN)的方式,将这种损失函数的效用达到最大。另外,Zhang 等人(2018)将 SSD 与 RPN 结构直接融合,设计了 ARM(anchor refinement module)和 ODM(object detection module)模块,并加入类 FPN 的 TCB(transfer connection block)模块来解决类别失衡问题,这个结构命名为 RefineDet。

一阶段算法有很多特色和优点。首先,在这个过程中,只用到了一次回归,因此运行速度相对更快。其次,由于在回归时定位、大小和类别信息同时输出,因此这种方法强调目标位置信息和类别信息的一体性,逻辑上更适用先有类别再寻求类别对应位置的生成式任务。

同样,一阶段算法也具有生成式任务的缺点。首先,由于缺少了二次回归位置微调的步骤,其定位精准度相对较低。其次,相较于其他生成式任务,这个结构中标签都是分析给出的,因此标签本身的准确度也会在相当程度上影响检测整体效果。

对这一类结构,尺度变化的挑战更多地融合在一体化结构方面。由于一阶段算法应用的更多是多分辨率检测(multi-resolution detection),因此在尺度层面最大的挑战在于如何设计更有效的模块,能够在多分辨率的情况下区分每种分辨率的特征。

此外,另一个很明显的特征是针对特定问题的提升手段对基于候选区域和基于回归的方法是共通的。在演化过程中都逐渐使用更深的网络,全卷积化、损失函数也使用相同的类型。因此从结构本身来看,对一阶段和两阶段检测算法的一些改进是共通的,很多方案可以同时讨论,实现共通。

2 目标检测算法中针对尺度变换问题的优化

在目标检测算法框架下,存在许多未解决的问题。其中,小目标检测由于小目标占图像的面积过小,特征过于模糊,一直是难点问题。对小目标检测主要在特征和分辨率两个层面进行改进,检测效果有了一系列提高。

一般来说,小目标分为两种,一种是目标本身并不是单纯的小尺度,只是因为拍摄角度使其在图像中占比较小;另一种是目标本身属于小尺度,不会在图像中出现大尺度的情况。对这两种情况,应对的策略也不同。对第 1 种情况,一般使用特征融合的方法,使用不同层次的特征的融合可以很好地解决这个问题;第 2 种情况使用更加特制的感受野网络

可以提取更鲁棒的特征。此外还可以进行训练集的针对化处理，特征提取网络的优化也能细化特征尺度的对应性。

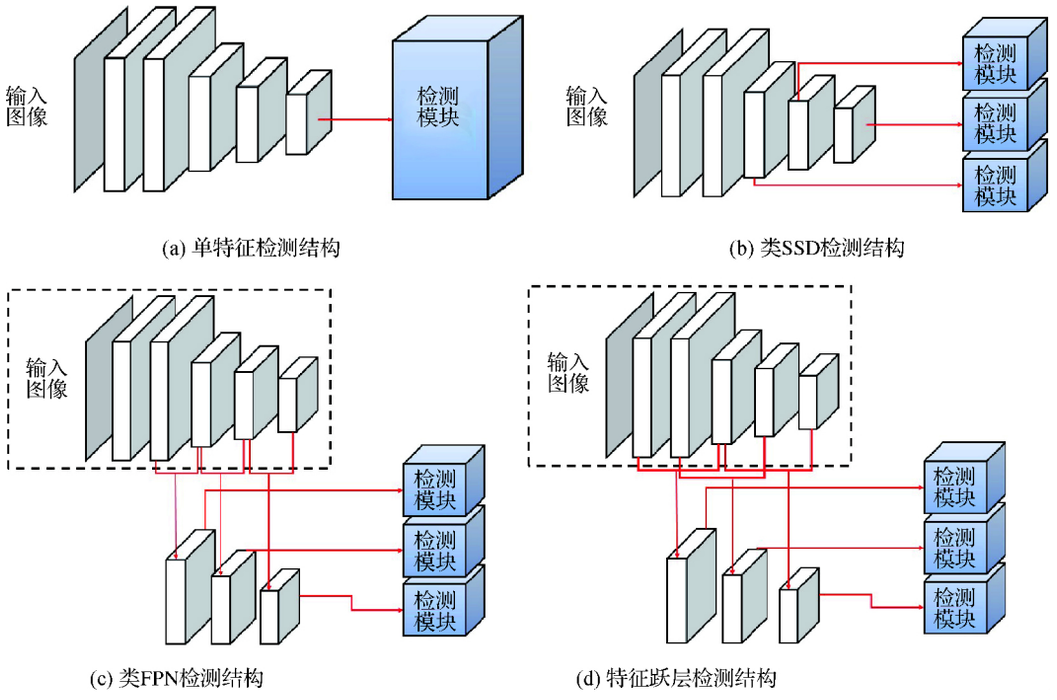
2.1 基于特征提取网络的优化

神经网络在高阶语义特征的应用中，特征提取网络一直是影响效果的关键因素之一。特征提取网络在 VGG-Net（Simonyan 和 Zisserman，2015）、ResNet（He 等，2016）到 ResNeXt（Xie 等，2017）、Xception（Chollet，2017）等网络的变迁过程中，出现了很多平衡速度和精度的网络版本。如 VGG-Net 网络出现了 11 层、16 层和 19 层的版本，ResNet 网络出现了 22 层、50 层、101 层和 152 层的版本。不同的网络对于尺度的敏感度是不同的。此外，对每个网络也有很多优化和改变，包括向量化卷积（即用向量相乘代替矩阵相乘）、向量化通道（即用向量来压缩通道数）、模块化卷积（即特征分块分别卷积再合并）和深度可分离卷积（即对不同输入通道采取不同卷积）。由于这些操作是对结构效果、结构操作速度和存储的衡量，因此对目标尺度变换的效果也是不同的。

2.2 基于多特征尺度融合的优化

在卷积神经网络结构中，不同层的结果代表不同层次的特征。卷积次数少得到的是接近细节的底层特征，卷积次数多得到的是接近语义的高层特征。相对于大目标检测，小目标由于尺寸较小，因此更加依赖细节特征，同时也依赖高层特征进行定位和分类。因此不同层不同尺寸的特征融合可以有效提高小目标检测的结果。

特征融合的发展如图 3 所示。Faster RCNN、YOLO 等未考虑目标尺度，使用的结构如图 3（a）所示，会出现最高层次的特征无法表征的状况。在 SSD 算法中，检测结构如图 3（b）所示，将检测模块分别接在后 3 层中。这样一来，接在较浅层的特征的检测模块会显现出对小目标回归的友好，但因为缺少高层次特征无法做更精准的分类操作。因此，FPN 将相邻层进行特征级联操作，检测结构如图 3（c）所示。对应地，在发展中发现跃层级联能更有效地融合特征，出现了如图 3（d）的特征级联模式。



(a) 单特征检测结构　(b) 类SSD检测结构

(c) 类FPN检测结构　(d) 特征跃层检测结构

图 3　特征融合检测发展结构图

Fig. 3　Feature fusion structures for detection

((a) single feature; (b) feature fusion as SSD; (c) feature fusion as FPN; (d) multi-layer feature fusion)

在 Faster RCNN 算法结构中，问题是比较明显的，即其候选区域提取结构以及 ROI 结构都只基于 VGG-Net 或 ResNet 中的最后一层特征来进行区域提取和分类处理。这样会忽略底层的特征信息，对小目标的检测不利。因此很有必要结合不同层的特征，进行归一化处理，再进行 RPN 或 ROI 处理。实

验表明,小目标的特征在最后 3 层即可得到,无需获取更底层的特征。

对此,Bell 等人(2016)提出 ION(inside-outside network)结构,将 VGG-Net 中的第 3、4、5 块卷积模块的结果进行尺寸和取值归一化,级联后作为联合特征进行 ROI 池化操作。此外,应用空间关联信息,分析每个特征的上下左右的特征,同样级联并入池化。这样处理后,提高了小目标检测的分类准确度和回归框精度。Kong 等人(2016)提出 HyperNet 结构,在最后的特征部分并未取最高层信息,而是针对 VGGNet,将第 1 层特征池化、第 5 层特征反卷积、与第 3 层特征级联再池化,得到候选区域提取和分类的特征,获得了较好效果。Cai 等人(2016)提出 MS-CNN 结构,认为针对大小目标的算法实际是在复杂度与准确度之间寻找中和。因此在第 3 层卷积后,直接在每一层接入分类部分,进行多次检测,并设计了多任务回归损失函数来整体训练。这些想法为后来的算法提供了很好的研究方向。

在传统的图像处理算法中,图像金字塔方法通过不同分辨率的图像得到不同大小的特征图。借鉴这种思路,Lin 等人(2017a)提出了 FPN 结构,设计了一个获取多尺度候选区域的算法,将相邻尺寸的特征通过降采样和使用 1 × 1 大小的卷积核卷积来得到不同维度不同高度的特征图,再对每个特征图进行候选区域的提取。这样一来,提取出的候选区域覆盖了多个尺度,进而可以得到更好的小目标检测结果。目前,这种方法有了基于 VGG-Net、ResNet 和 Inception 等特征网络的变种结构,逐渐成为了两步法提升效果的标配插件,其变种也逐渐应用于新的结构中。

对 SSD 式多特征检测也可以用特征跳跃融合的方法改善本身对小尺寸高阶特征的丢失。Jeong 等人(2017)将特征网络不同层的特征通过池化的方式进行了级联,提出了 RSSD(rainbow SSD)结构。Huang 等人(2017)则是将不同层先进行卷积,上采样级联之后再次卷积,提出 FSSD(feature fusion SSD)结构(Li 和 Zhou,2017)。Zheng 等人(2018)提出 ESSD(extended SSD)算法,对 VGG-16 网络的第 7 ~ 10 层特征进行上采样,与其对应的上一层进行残差计算,得到的结果连接检测模块。Cui 等人(2018)在 ESSD 的基础上通过实验得出残差计算的最优间隔,优化了上采样结构,得到了 MDSSD(multi-scale deconvolutional SSD)算法。这几种结构的特征融合模式如图 4 所示。

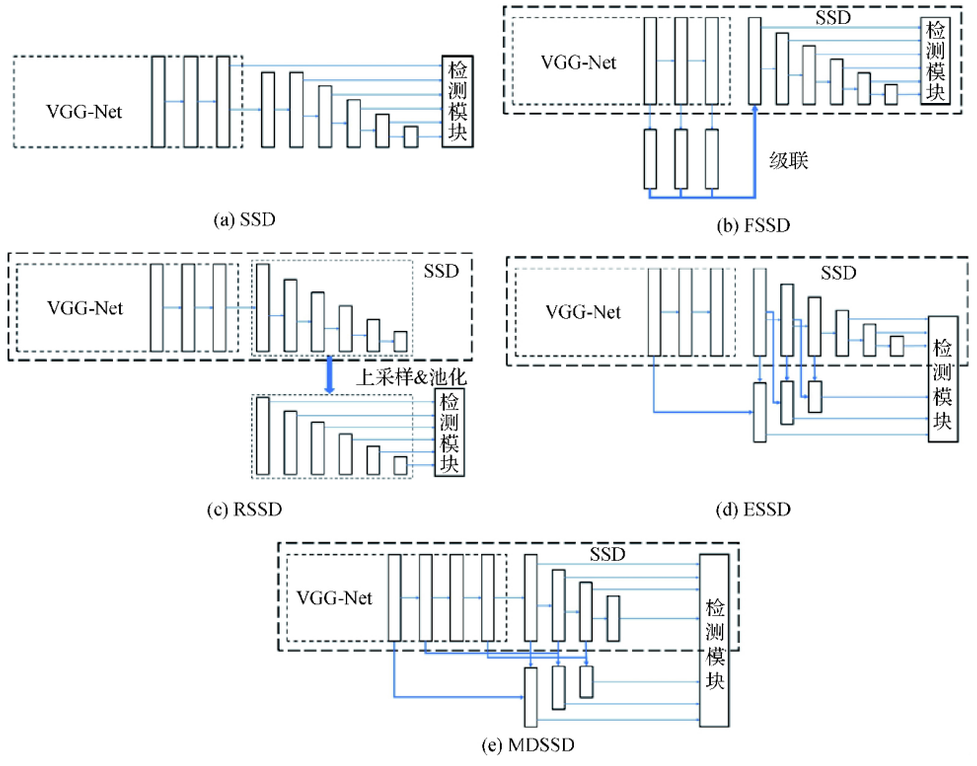


图 4　SSD 系列结构特征融合模式示意图

Fig. 4　Schematic diagram of SSD series structural feature fusion model((a) SSD; (b) FSSD; (c) RSSD; (d) ESSD; (e) MDSSD)

基于 SSD 类方法进行多尺度优化的思路除了上述方法之外，还有在 RetinaNet 的基础上进行的改动。Zhu 等人(2019)提出 FSAF(feature selective anchor-free module)结构，在 RetinaNet 的基础上，在每一个特征级分支的后面加入了无需锚点框的分类和回归分支。在结构应用时，对一个目标合理的尺寸特征图会自然地回归出高信赖值的目标框。Tian 等人(2019)提出 FCOS(fully convolutional one-stage object detection)结构，使用类似语义分割的训练方法，进行点对点的训练，通过限制回归范围和中心增强分支进行效果增强。Kong 等人(2019)提出 FoveaBox 结构，认为中央凹的信息较多，因此通过 FPN 回归出中央凹的信息，再通过坐标转换的方式获取区域位置。这 3 种方法都是基于 FPN 的多尺度目标检测，通过跳跃的层间特征的级联以及对其密集回归来提升多种不同尺寸的目标的检测效果。

以上 3 种方法都属于 anchor-free 的范畴，都使用了 SSD 结构中最重要的尺度自适应性质。由于特征金字塔的特点，不同层的特征输出可以表征不同尺寸目标的特征。因此在自适应的层面上，不同尺度的目标可以从多个特征层中检测到，每个检测模块负责不同的尺寸。由于检测模块本身并不大，因此多个模块的并列并不会显著拖慢结构的速度。这种方法使得这些结构对小目标的检测准确度大幅增大。

现阶段对特征的处理，除了结构上选择更优更深的网络之外，效果最优的是基于层间跳跃级联的特征融合方法。通过不同层的特征融合，可以做到底层细节特征与高层语义特征的结合，用以匹配不同尺度目标的特征细腻度和语义完整度。

2.3 基于特征感受野变化的优化

对每个特征点，由于来源于卷积，因此可以回溯得到每个特征点对应的数据来源，称为该特征点的感受野。高层特征单点感受野较大，底层特征单点感受野相对小。因此除了使用不同层次特征之外，还可以通过不同卷积方式获得更加致密的、感受野可变的特征图。

在一阶段算法中，YOLO 系列对大图像小目标的多角度目标检测不鲁棒。对此，van Etten(2018)提出了 YOLT(you only look twice)算法，用于检测遥感图像中小于 26 × 26 像素的目标。由于遥感图像一般较大，因此首先使用 416 × 416 像素的滑窗将图像划分成多个子区域，再使用优化后可得到更致密特征的 DarkNet 进行特征筛选和优化。YOLT 在小目标检测上的主要贡献是将无法检测的小于 26 × 26 像素的目标的 DarkNet19 网络优化到更加致密。

Papandreou 等人(2015)首先应用空洞卷积获取更加致密的特征图。之后，RFCN(region-based fully convolutional network)(Dai 等，2016)和 SSD(Liu 等，2016)等算法采用了这种卷积方法。通过对不同间隔的像素点进行取样和卷积，以及特定的池化模式，可以得到稀疏采样的结果，以做到更改感受野的目的，使同级特征相对更加致密，不同级特征相对感受野大小差距较小。

然而在空洞卷积中，感受野只可能扩大不可能缩小。换句话说，其对小目标检测的优化只在于获取的特征图十分致密。导致应用空洞卷积检测小目标时，感受野扩大带来的弊端中和了特征图致密带来的利端。在 2015 年，Jaderberg 等人(2015)提出在网络中加入训练旋转分量的思想，整个网络多训练一个旋转参数并应用旋转参数对图像进行预处理。在这个思路的基础上，2017 年微软亚洲研究院提出可变卷积核的思路(Dai 等，2017)，在训练卷积核的同时，在每个位置训练一个偏移量，对卷积核进行致密偏移。这样可以使卷积过程中的感受野更加可塑，也更加匹配物体本身。这样一来，可以使卷积过程更加专注于小目标本身及其周边大小的区域。

图 5 是感受野的示意图，红色块为卷积中心位置，蓝色块为卷积感受野。图 5(a)展示了正常卷积在特征图中对应的感受野。图 5(b)展示的是空洞卷积对应的感受野，可以看出其感受野相对更大。图 5(c)则是为每个卷积过程学习了偏移量，对应每个箭头向量，这种感受野更加多变，使得图像中尺度一直很小的目标更易抓取。

空洞卷积的结合会使特征图的表达效果变好。对多种空洞卷积的级联来说，有多种策略融合所有信息。在 DeepLabV2(Chen 等，2018)中，使用 ASPP(atrous spatial pyramid pooling)增强特征信息。不同分支的空洞卷积提取不同感受野的特征，再用最大值池化融合通道。而 RFBNet 的网络结构(Liu 等，2018)则是基于 SSD 将不同层特征空洞卷积至相应尺寸再进行级联。在 trident network(Li 等，2019)中，直接将不同尺寸的空洞卷积分成不同分支，通过

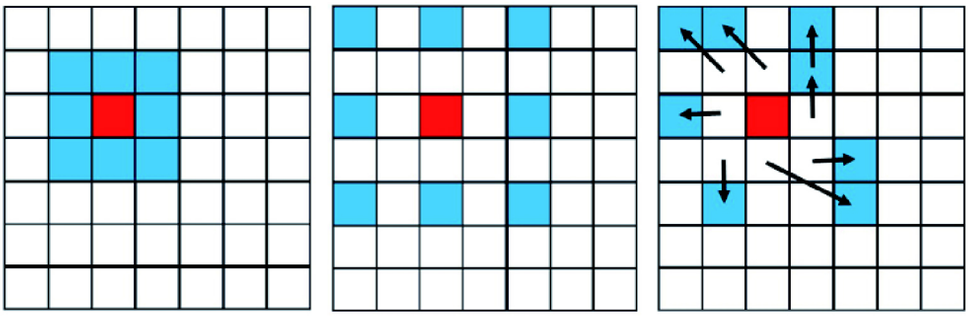


图 5　不同卷积过程感受野的变化示意图

Fig. 5　Schematic diagram of the change of receptive field in different convolution processes

((a) receptive field of simple convolution; (b) receptive field of dilated convolution; (c) receptive field of deformable convolution)

自适应在每个分支输出不同尺寸目标,获取了对小目标检测的较好效果。

2.4　基于不同训练策略的研究

对特定尺度的目标来说,使用特定的经过设计的训练集,并设计针对性的网络结构对效果提升是比较显著的。Chen 等人(2017a)基于 RCNN 结构,筛选了几个大数据集中的小物体,并将其压缩到一定大小以下,将 RPN 中锚点框的大小减到 16^2、40^2 和 100^2 等 3 个尺寸。在分类时又将锚点框做上采样,以获得足够数量的分类回归特征。这个过程使得对小目标的检测效果得到了较大提升。

除去特定的网络设计之外,合理的训练策略和训练集也可以有效提高小目标检测的效率。在 YOLO第 2 代(Redmon 等,2017)的结构训练过程中,采取随机选择图像尺寸的训练策略,即每经过 10 个周期的训练就会随机选择新尺寸,以此获得多尺度鲁棒效果。

针对训练过程中的输入图像尺度问题,2018 年 Singh 和 Davis(2018)设计了 3 个网络,进行了一系列实验。CNN-B 和 CNN-B-FT 测试图像是先下采样成低分辨率图像,再上采样到尺寸为单边长 224 的方形图像; CNN-S 在 48 × 48 像素、96 × 96 像素的图像上测试。CNN-B 是在分辨率为 224 × 224 像素的样本上训练; CNN-S 是在低分辨率图像上训练; CNN-B-FT 是先在 224 × 224 像素的样本上进行训练,再使用低分辨率上采样图像进行网络微调。这些网络都使用了数据增强,如裁剪、颜色增强。这样设计的结构效果可见,训练和检测分辨率差的越大,效果也越差; 上采样图像对于训练有效,并在检测小目标上好于设计特定尺寸的分类器。基于这样的结论,提出了一个基于不同尺寸的训练输入的结构,通过不同尺度的掩模来筛选 RPN 获得的不同大小的目标框。设定了一个不同有效范围的真值框,进行选择性训练得到不同尺度的网络结构。这个结构名为 SNIP(scale normalization for image pyramids),虽然运行速度较低但效果较好。

在此基础上,2018 年提出的改进结构(Singh 等,2018)中,同样将图像缩放到了 3 个尺度,但并非直接训练,而是将其通过滑窗提取成一定大小的块模型,然后将块模型归一化后再训练。负样本则是通过 RPN 来生成,去掉与正样本重叠的区域后,选择最大覆盖的区域框进行训练。这样使得正负样本平衡。使用网络时,将原图规整至 3 种不同尺寸并用软极大值抑制来控制合并。这样处理后的效果相对 SNIP 有较大的提升。

对于训练过程中训练集的多尺度的规整,主要思路依旧是将训练集中的不同目标进行不同程度的重采样改变大小。这样的好处是对不好检测的尺度(一般是小目标)的效果较好,坏处是结构基于多尺度产生的冗余计算较多,结构运行速度较慢。

3　不同方法对尺度的鲁棒性比较

通用目标检测需要大量数据进行结构训练和数据微调,因此需要依赖大数据库进行模型的训练和验证。较为常用的大数据库包括 PASCAL VOC 数据集(Everingham 等,2010)和 Microsoft COCO(MS COCO)数据集。Lin 等人(2014)基于 MS COCO 数

据集，提供了一系列验证效果的参数，包括不同尺寸目标的检测效果的参数，通过这些数据，可以清楚地看出不同方法对于尺寸的鲁棒性。

在检测任务中，一般使用平均查准率(average precision，AP)评价特征提取网络的效果。以某一重合率(一般是50%和75%)为界，认为大于这个数值的目标框为查准目标框。查准率指查准目标框占所有预测目标框的比例。

表2是不同方法使用不同特征提取网络的效果比较，ms表示多尺度模块，AP为阈值采样平均的查准率，AP_{50}和AP_{75}表示为阈值为0.5和0.75时的平均查准率，AP_S、AP_M和AP_L分别为小、中和大目标的平均查准率。从表2可以看出：1)网络深度的加深对各尺度目标的检测效果都有提升。2)在VGG-16与ResNet的比较中，若不包含多尺度模块，二者对小目标的准确度差距基本不大，效果的提升都在

表2 同一算法不同特征提取网络在MS COCO数据集验证集上的运行结果比较

Table 2 Comparison of running results of different feature extraction networks of the same algorithm on validation set of MS COCO dataset

/%

算法	输入尺寸	特征提取网络	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RFCN	321	ResNet 50	27.1	49.0	26.9	10.4	29.7	39.2
	321	ResNet 101	30.52	52.9	31.2	12.0	33.9	43.8
Cascade RCNN	321	ResNet 50 + FPN	40.6	59.9	44.0	22.6	42.7	52.1
	321	ResNet 101 + FPN	42.8	62.1	46.3	23.7	45.5	55.2
SSD	300	VGG-16	25.1	43.1	25.8	6.6	22.4	35.5
	321	ResNet 101	28.0	45.4	29.3	6.2	28.3	49.3
	512	VGG-16	28.8	48.5	30.3	10.9	31.8	43.5
	513	ResNet 101	31.2	50.4	33.3	10.2	34.5	49.8
DSSD	321	ResNet 101	28.0	46.1	29.2	7.4	28.1	47.6
	513	ResNet 101	33.2	53.3	35.2	13.0	35.4	51.1
RetinaNet	500	ResNet 101	34.4	53.1	36.8	14.7	38.5	49.1
	800	ResNet 101 + FPN	39.1	59.1	42.3	21.8	42.7	50.2
RefineDet	320	VGG-16	29.4	49.2	31.3	10.0	32.0	44.4
	320	ResNet 101	32.0	51.4	34.2	10.5	34.7	50.4
	320	VGG-16 + ms	35.2	56.1	37.7	19.5	37.2	47
	320	ResNet 101 + ms	38.6	59.9	41.7	21.1	41.7	52.3
	512	VGG-16	33.0	54.5	35.5	16.3	36.3	44.3
	512	ResNet 101	36.4	57.5	39.5	16.6	39.9	51.4
	512	VGG-16 + ms	37.6	58.7	40.8	22.7	40.3	48.3
	512	ResNet 101 + ms	41.8	62.9	45.7	25.6	45.1	54.1
CenterNet	511	Hourglass 52	41.6	59.4	44.2	22.5	43.1	54.1
	511	Hourglass 104	44.9	62.4	48.1	25.6	47.4	57.4
	511	Hourglass 52 + ms	43.5	61.3	46.7	25.3	45.3	55.0
	511	Hourglass 104 + ms	**47.0**	**64.5**	**50.7**	**28.9**	**49.9**	**58.9**

注：加粗字体表示每列最优结果。

大目标上。若包含多尺度模块,如 FPN 模块,则 ResNet 对小目标的效果更好,显然 ResNet 对多尺度模块更加友好。3) 小目标的检测精度远没有大目标的检测精度高,有时甚至不足大目标的一半。4) 一般只有小目标的准确度低于平均的 AP 值,因此提高小目标的检测水准一定会提高平均检测水平。

从检测方法的发展来看,没有一种方法是只提高某一尺度的检测精度。因此,对小目标设计的提升模块通常也可以提升大目标的检测效果。同时可以看出,检测算法在全尺寸上有较大提升空间。

较为具有结构创新的两阶段和一阶段算法的运行结果如表 3 和表 4 所示。其中,表 4 中的 SSD321 算法由 Liu 等人(2016) 提出,但该算法使用 VGG-16 作为特征提取网络,因此采用 Fu 等人(2017) 给出的结论。所有算法均选取同样的 101 层残差网络(去除一些使用独立网络的方法) 进行特征提取。

表 3　两阶段算法在 MS COCO 数据集验证集上的运行结果

Table 3　The results of two-stage algorithms on validation set of MS COCO dataset

/%

算法	特征提取网络	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
OHEM(Shrivastava 等,2016)	VGG-16	22.6	42.5	22.2	5.0	23.7	37.9
OHEM ++(Shrivastava 等,2016)	VGG-16	25.5	45.9	26.1	7.4	27.7	40.3
RFCN(Dai 等,2016)	ResNet 101	29.9	51.9	–	10.8	32.8	45.0
CoupleNet(Zhu 等,2017)	ResNet 101	34.4	54.8	37.2	13.4	38.1	52.0
Faster RCNN +++(He 等,2016)	ResNet 101	34.9	55.7	37.4	15.6	38.7	50.9
Faster RCNN + FPN(Lin 等,2017)	ResNet 101	36.2	59.1	39.0	18.2	39	48.2
Deformable R-FCN(Dai 等,2017)	Inception-ResNet	37.5	58.0	40.8	19.4	40.1	52.5
Mask RCNN(He 等,2017)	ResNeXt 101	39.8	62.3	43.4	22.1	43.2	51.2
Cascade RCNN(Cai 和 Vasconcelos,2018)	ResNet 101	42.8	62.1	46.3	23.7	45.5	55.2
Fitness NMS(Tychsen-Smith 等,2018)	ResNet 101	41.8	60.9	44.9	21.5	45.0	57.5
SNIP(Singh 等,2018)	DPN 98	45.7	67.3	51.1	29.3	48.8	57.1
SNIPER(Singh 等,2018)	ResNet 101	**47.62**	**68.5**	**53.4**	**30.9**	**50.6**	**60.7**

注: 加粗字体表示每列最优结果,“–”表示无法考据结果。

4　结　语

随着特征选取网络的深入、优化和更新,目标检测的算法精度在逐步提升,但是依然存在很多需要解决的问题。其中,空间层面上的目标尺度问题是一个绕不开的大问题。对这类尺度问题的研究可以根据现有的提升方式窥见可行的发展方向。

1) 使用更加鲁棒的模型获得的特征对图像中的小目标特征进行提取是一个比较重要的研究方向。有些网络使用中高层特征结合级联的方式提升鲁棒性,实现位置的局部信息与类别的高级信息的融合; 有些网络使用残差模块或跃层模块进行信息的交错衡量。如果能提高两种特征之间的联系,将对尺度不同的目标的检测效果有较大提升。

2) 将获取的特征用于不同尺度方向的检测模块是另一个较为可行的研究方向,包括对精度的研究和对速度的研究。提高精度需要有更加多元或尺度友好的检测模块,或者使用点检测代替框检测的方法进行检测; 加快速度需要有更加快速的尺度适应法则,相较于尺度放缩,加快速度有更好的运行速度的模块。在这些思路上,应该尝试引入传统的检测分类算法,可能会有更加优秀的效果。

3) 样本的精确设计也可以作为提升多尺度目标检测效果的方案之一。对样本来说,现阶段主要的思路是将样本中的目标进行放大和缩小,通过这

表 4　一阶段算法在 MS COCO 数据集验证集上的运行结果

Table 4　The results of one-stage algorithms on validation set of MS COCO dataset

/%

算法	特征提取网络	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
YOLO V2(Redmon 和 Farhadi,2017)	DarkNet 19	21.6	44.0	19.2	5.0	22.4	35.5
YOLO V3(Redmon 和 Farhadi,2018)	DarkNet 53	33.0	57.9	34.4	18.3	35.4	41.9
SSD321(Liu 等,2016; Fu 等,2017)	ResNet 101	28.0	45.4	29.3	6.2	28.3	49.3
SSD513(Fu 等,2017)	ResNet 101	31.2	50.4	33.3	10.2	34.5	49.8
DSSD321(Fu 等,2017)	ResNet 101	28.0	46.1	29.2	7.4	28.1	47.6
DSSD513(Fu 等,2017)	ResNet 101	33.2	53.3	35.2	13.0	35.4	51.1
RetinaNet500(Lin 等,2017)	ResNet 101	34.4	53.1	36.8	14.7	38.5	49.1
RetinaNet800(Lin 等,2017)	ResNet 101	39.1	59.1	42.3	21.8	42.7	50.2
RefineDet320 +(Xie 等,2017)	ResNet 101	38.6	59.9	41.7	21.1	41.7	52.3
RefineDet512 +(Xie 等,2017)	ResNet 101	41.8	62.9	45.7	25.6	45.1	54.1
FoveaBox(Kong 等,2019)	ResNet 101	40.6	60.1	43.5	23.3	45.2	54.5
CornetNet(Law 和 Deng,2018)	Hourglass104	40.5	57.8	45.3	20.8	44.8	56.7
CenterNet(Zhou 等,2019)	Hourglass104	**47.0**	**64.5**	**50.7**	**28.9**	**49.9**	**58.9**

注: 加粗字体表示每列最优结果。

种方式来进行不同尺度同一目标的分类检测。实际上还可以设计更有效的样本来提升训练效果。

参考文献(References)

Alexe B, Deselaers T and Ferrari V. 2012. Measuring the objectness of image windows. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(11): 2189-2202 [DOI: 10.1109/TPAMI.2012.28]

Arbeláez P, Pont-Tuset J, Barron J, Marques F and Malik J. 2014. Multiscale combinatorial grouping//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE: 328-335 [DOI: 10.1109/CVPR.2014.49]

Bell S, Zitnick C L, Bala K and Girshick R. 2016. Inside-outside net: detecting objects in context with skip pooling and recurrent neural networks//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE: 2874-2883 [DOI: 10.1109/CVPR.2016.314]

Bodla N, Singh B, Chellappa R and Davis L S. 2017. Soft-NMS—improving object detection with one line of code//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE: 5562-5570 [DOI: 10.1109/iccv.2017.593]

Cai Z W, Fan Q F, Feris R S and Vasconcelos N. 2016. A unified multi-scale deep convolutional neural network for fast object detection//Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer: 354-370 [DOI: 10.1007/978-3-319-46493-0_22]

Cai Z W and Vasconcelos N. 2018. Cascade R-CNN: delving into high quality object detection//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE: 6154-6162 [DOI: 10.1109/cvpr.2018.00644]

Carreira J and Sminchisescu C. 2012. C CPMC: automatic object segmentation using constrained parametric Min-Cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(7): 1312-1328 [DOI: 10.1109/TPAMI.2011.231]

Chen C Y, Liu M Y, Tuzel O and Xiao J X. 2017a. R-CNN for small object detection//Proceedings of the 13th Asian Conference on Computer Vision. Taipei, China: Springer: 214-230 [DOI: 10.1007/978-3-319-54193-8_14]

Chen L C, Papandreou G, Kokkinos I, Murphy K and Yuille A L. 2018. DeepLab: semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4): 834-848 [DOI: 10.1109/TPAMI.2017.2699184]

Chen L C, Papandreou G, Schroff F and Adam H. 2017b. Rethinking atrous convolution for semantic image segmentation [EB/OL]. [2019-12-01]. https://arxiv.org/pdf/1706.05587.pdf

Chollet F. 2017. Xception: deep learning with depthwise separable convolutions//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE: 1800-1807 [DOI: 10.1109/CVPR.2017.195]

Cui L S, Ma R, Lv P, Jiang X H, Gao Z M, Zhou B and Xu M L. 2018.

MDSSD: multi-scale deconvolutional single shot detector for small objects [EB/OL]. [2019-12-01]. https://arxiv.org/pdf/1805.07009.pdf

Dai J F, Li Y, He K M and Sun J. 2016. R-FCN: object detection via region-based fully convolutional networks//Proceedings of the 30th Conference on Neural Information Processing Systems. Barcelona: NIPS: 379-387

Dai J F, Qi H Z, Xiong Y W, Li Y, Zhang G D, Hu H and Wei Y C. 2017. Deformable convolutional networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE: 764-773 [DOI: 10.1109/iccv.2017.89]

Dalal N and Triggs B. 2005. Histograms of oriented gradients for human detection//Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego: IEEE: 886-893 [DOI: 10.1109/CVPR.2005.177]

Dean T, Ruzon M A, Segal M, Shlens J, Vijayanarasimhan S and Yagnik J. 2013. Fast, accurate detection of 100 000 object classes on a single machine//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland: IEEE: 1814-1821 [DOI: 10.1109/CVPR.2013.237]

Deng J K, Guo J, Xue N N and Zafeiriou S. 2018. Arcface: additive angular margin loss for deep face recognition [EB/OL]. [2019-12-01]. https://arxiv.org/pdf/1801.07698.pdf

Dollar P, Wojek C, Schiele B and Perona P. 2012. Pedestrian detection: an evaluation of the state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(4): 743-761 [DOI: 10.1109/TPAMI.2011.155]

Du C B, Gao S S, Liu Y and Gao B B. 2019. Multi-focus image fusion using deep support value convolutional neural network. Optik, 176: 567-578 [DOI: 10.1016/j.ijleo.2018.09.089]

Duan K W, Bai S, Xie L X, Qi H G, Huang Q M and Tian Q. 2019. CenterNet: keypoint triplets for object detection [EB/OL]. [2019-12-01]. https://arxiv.org/pdf/1904.08189.pdf

Everingham M, van Gool L, Williams C K I, Winn J and Zisserman A. 2010. The Pascal Visual Object Classes (VOC) challenge. International Journal of Computer Vision, 88(2): 303-338 [DOI: 10.1007/s11263-009-0275-4]

Felzenszwalb P, Mcallester D and Ramanan D. 2008. A discriminatively trained, multiscale, deformable part model//Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage: IEEE: 1-8 [DOI: 10.1109/CVPR.2008.4587597]

Fu C Y, Liu W, Ranga A, Tyagi A and Berg A C. 2017. DSSD: deconvolutional single shot detector [EB/OL]. [2019-12-01]. https://arxiv.org/pdf/1701.06659.pdf

Geiger A, Lenz P and Urtasun R. 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence: IEEE: 3354-3361 [DOI: 10.1109/CVPR.2012.6248074]

Gidaris S and Komodakis N. 2015. Object detection via a multi-region and semantic segmentation-aware CNN model//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago: IEEE: 1134-1142 [DOI: 10.1109/ICCV.2015.135]

Girshick R. 2015. Fast R-CNN//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago: IEEE: 1440-1448 [DOI: 10.1109/ICCV.2015.169]

Girshick R, Donahue J, Darrell T and Malik J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE: 580-587 [DOI: 10.1109/CVPR.2014.81]

He K M, Gkioxari G, Dollár P and Girshick R. 2017. Mask R-CNN//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE: 2980-2988 [DOI: 10.1109/iccv.2017.322]

He K M, Zhang X Y, Ren S Q and Sun J. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9): 1904-1916 [DOI: 10.1109/TPAMI.2015.2389824]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Howard A G, Zhu M L, Chen B, Kalenichenko D, Wang W J, Weyand T, Andreetto M and Adam H. 2017. Mobilenets: efficient convolutional neural networks for mobile vision applications [EB/OL]. [2019-12-01]. https://arxiv.org/pdf/1704.04861.pdf

Hu J, Shen L and Sun G. 2018. Squeeze-and-excitation networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE: 7132-7141 [DOI: 10.1109/cvpr.2018.00745]

Huang G, Liu Z, van der Maaten L and Weinberger K Q. 2017. Densely connected convolutional networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE: 2261-2269 [DOI: 10.1109/CVPR.2017.243]

Iandola F N, Han S, Moskewicz M W, Ashraf K, Dally W J and Keutzer K. 2017. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size [EB/OL]. [2019-12-01]. https://arxiv.org/pdf/1602.07360.pdf

Jaderberg M, Simonyan K, Zisserman A and Kavukcuoglu K. 2015. Spatial transformer networks//Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge: MIT Press: 2017-2025

Jeong J, Park H and Kwak N. 2017. Enhancement of SSD by concatenating feature maps for object detection//Proceedings of 2007 British Machine Vision Conference. London: BMVA Press: #22514709 [DOI: 10.5244/c.31.76]

Jia X, Gavves E, Fernando B and Tuytelaars T. 2015. Guiding long-short

term memory for image caption generation [EB/OL]. [2019-12-01]. https://arxiv.org/pdf/1509.04942.pdf

Kong T, Yao A B, Chen Y R and Sun F C. 2016. HyperNet: towards accurate region proposal generation and joint object detection//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE: 845-853 [DOI: 10.1109/CVPR.2016.98]

Kong T, Sun F C, Liu H P, Jiang Y N and Shi J B. 2019. FoveaBox: beyond anchor-based object detector [EB/OL]. [2019-12-01]. https://arxiv.org/pdf/1904.03797.pdf

Law H and Deng J. 2018. CornerNet: detecting objects as paired keypoints//Proceedings of the 15th European Conference on Computer Vision. Munich: Springer: 765-781 [DOI: 10.1007/978-3-030-01264-9_45]

Law H, Teng Y. Russakovsky O and Deng J. 2019. CornerNet-Lite: efficient keypoint based object detection [EB/OL]. [2019-12-01]. https://arxiv.org/pdf/1904.08900.pdf

Le M H, Woo B S and Jo K H. 2011. A Comparison of SIFT and Harris Conner features for correspondence points matching//Proceedings of the 17th Korea-Japan Joint Workshop on Frontiers of Computer Vision. Ulsan: IEEE: 1-4 [DOI: 10.1109/FCV.2011.5739748]

Li Y H, Chen Y T, Wang N Y and Zhang Z X. 2019. Scale-aware trident networks for object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, IEEE: 6053-6062 [DOI: 10.1109/ICCV.2019.00615]

Li Z X and Zhou F Q. 2017. FSSD: feature fusion single shot multibox detector [EB/OL]. [2019-12-01]. https://arxiv.org/pdf/1712.00960.pdf

Liau H, Nimmagadda Y and Wong Y. 2018. Fire SSD: wide fire modules based single shot detector on edge device [EB/OL]. [2019-12-01]. https://arxiv.org/pdf/1806.05363.pdf

Lin T Y, Dollár P, Girshick R, He K M, Hariharan B and Belongie S. 2017a. Feature pyramid networks for object detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE: 936-944 [DOI: 10.1109/CVPR.2017.106]

Lin T Y, Goyal P, Girshick R, He K M and Dollár P. 2017b. Focal loss for dense object detection//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE: 2999-3007 [DOI: 10.1109/iccv.2017.324]

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L. 2014. Microsoft COCO: common objects in context//Proceedings of the 13th European Conference on Computer Vision. Zürich: Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48]

Liu S T, Huang D and Wang Y H. 2018. Receptive field block net for accurate and fast object detection//Proceedings of the 15th European Conference on Computer Vision. Munich: Springer: 404-419 [DOI: 10.1007/978-3-030-01252-6_24]

Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y and Berg A C. 2016. SSD: single shot MultiBox detector//Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer: 21-37 [DOI: 10.1007/978-3-319-46448-0_2]

Long J, Shelhamer E and Darrell T. 2015. Fully convolutional networks for semantic segmentation//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE: 3431-3440 [DOI: 10.1109/CVPR.2015.7298965]

Newell A, Yang K Y and Deng J. 2016. Stacked hourglass networks for human pose estimation//Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer: 483-499 [DOI: 10.1007/978-3-319-46484-8_29]

Papandreou G, Kokkinos I and Savalle P A. 2015. Modeling local and global deformations in deep learning: epitomic convolution, multiple instance learning, and sliding window detection//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE: 390-399 [DOI: 10.1109/CVPR.2015.7298636]

Redmon J, Divvala S, Girshick R and Farhadi A. 2016. You only look once: unified, real-time object detection//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE: 779-788 [DOI: 10.1109/CVPR.2016.91]

Redmon J and Farhadi A. 2017. YOLO9000: better, faster, stronger//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE: 6517-6525 [DOI: 10.1109/CVPR.2017.690]

Redmon J and Farhadi A. 2018. YOLOv3: an incremental improvement [EB/OL]. [2019-12-01]. https://arxiv.org/pdf/1804.02767.pdf

Ren S Q, He K M, Girshick R and Sun J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149 [DOI: 10.1109/TPAMI.2016.2577031]

Ren X F and Ramanan D. 2013. Histograms of sparse codes for object detection//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland: IEEE: 3246-3253 [DOI: 10.1109/CVPR.2013.417]

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S A, Huang Z H, Karpathy A, Khosla A, Bernstein M, Berg A C and Li F F. 2015. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 115(3): 211-252 [DOI: 10.1007/s11263-015-0816-y]

Sandler M, Howard A, Zhu M L, Zhmoginov A and Chen L C. 2018. MobileNetV2: inverted residuals and linear bottlenecks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE: 4510-4520 [DOI: 10.1109/CVPR.2018.00474]

Schroff F, Kalenichenko D and Philbin J. 2015. FaceNet: a unified embedding for face recognition and clustering//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Bos-

ton: IEEE: 815-823 [DOI: 10.1109/CVPR.2015.7298682].

Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R and LeCun Y. 2014. OverFeat: integrated recognition, localization and detection using convolutional networks [EB/OL]. [2019-12-01]. https://arxiv.org/pdf/1312.6229.pdf

Sermanet P, Kavukcuoglu K, Chintala S and Lecun Y. 2013. Pedestrian detection with unsupervised multi-stage feature learning//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland: IEEE: 3626-3633 [DOI: 10.1109/CVPR.2013.465]

Shen Z Q, Liu Z, Li J G, Jiang Y G, Chen Y R and Xue X Y. 2017. DSOD: learning deeply supervised object detectors from scratch//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE: 1937-1945 [DOI: 10.1109/iccv.2017.212]

Shrivastava A, Gupta A and Girshick R. 2016. Training region-based object detectors with online hard example mining [EB/OL]. [2019-12-01]. https://arxiv.org/pdf/1409.1556.pdf

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2019-12-01]. https://arxiv.org/pdf/1409.1556.pdf

Singh B and Davis L S. 2018. An analysis of scale invariance in object detection-SNIP//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE: 3578-3587 [DOI: 10.1109/cvpr.2018.00377]

Singh B, Najibi M and Davis L S. 2018. SNIPER: efficient multi-scale training//Proceedings of the 32nd Conference on Neural Information Processing Systems. Montréal: NeurIPS: 9310-9320

Sun Y, Liang D, Wang X G and Tang X O. 2015. DeepID3: face recognition with very deep neural networks [EB/OL]. [2019-12-01]. https://arxiv.org/pdf/1502.00873.pdf

Tang H, Xiao B, Li W S and Wang G Y. 2017. Pixel convolutional neural network for multi-focus image fusion. Information Sciences, 433-434: 125-141 [DOI: 10.1016/j.ins.2017.12.043]

Tian Z, Shen C H, Chen H and He T. 2019. FCOS: fully convolutional one-stage object detection [EB/OL]. [2019-12-01]. https://arxiv.org/pdf/1904.01355.pdf

Tychsen-Smith L, Petersson L. 2018. Improving object localization with fitness NMS and bounded IoU loss//Proceedings of IEEE computer vision and pattern recognition. Salt Lake City: IEEE: 6877-6885. [DOI: 10.1109/cvpr.2018.00719]

Uijlings J R R, van de Sande K E A, Gevers T and Smeulders A W M. 2013. Selective search for object recognition. International Journal of Computer Vision, 104(2): 154-171 [DOI: 10.1007/s11263-013-0620-5]

van Etten A. 2018. You only look twice: rapid multi-scale object detection in satellite imagery [EB/OL]. [2019-12-01]. https://arxiv.org/pdf/1805.09512.pdf

Xiao J X, Ehinger K A, Hays J, Torralba A and Oliva A. 2016. SUN database: exploring a large collection of scene categories. International Journal of Computer Vision, 119(1): 3-22 [DOI: 10.1007/s11263-014-0748-y]

Xie S N, Girshick R, Dollár P, Tu Z W and He K M. 2017. Aggregated residual transformations for deep neural networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE: 5987-5995 [DOI: 10.1109/CVPR.2017.634]

Xu K, Ba J L, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel R S and Bengio Y. 2015. Show, attend and tell: neural image caption generation with visual attention//Proceedings of the 32nd International Conference on Machine Learning. Lille: JMLR: 2048-2057.

Yang B, Yan J J, Lei Z and Li S Z. 2016. CRAFT objects from images//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE: 6043-6051 [DOI: 10.1109/CVPR.2016.650]

Yu J H, Jiang Y N, Wang Z Y, Cao Z M and Huang T. 2016. UnitBox: an advanced object detection network//Proceedings of the 24th ACM International Conference on Multimedia. Amsterdam: ACM: 516-520 [DOI: 10.1145/2964284.2967274]

Zhang S, Gong Y H and Wang J J. 2019. The development of deep convolution neural network and its applications on computer vision. Chinese Journal of Computers 42(3): 453-482 (**张顺, 龚怡宏, 王进军**. 2019. **深度卷积神经网络的发展及其在计算机视觉领域的应用**. **计算机学报**, 42(3): 453-482 [DOI: 10.11897/SP.J.1016.2019.00453]

Zhang S F, Wen L Y, Bian X, Lei Z and Li S Z. 2018. Single-shot refinement neural network for object detection//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE: 4203-4212 [DOI: 10.1109/cvpr.2018.00442]

Zhang S S, Benenson R and Schiele B. 2017. CityPersons: a diverse dataset for pedestrian detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE: 4457-4465 [DOI: 10.1109/CVPR.2017.474]

Zhang Y, Li B H, Lu H C, Irie A and Ruan X. 2016. Sample-specific SVM learning for person re-identification//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE: 1278-1287 [DOI: 10.1109/CVPR.2016.143]

Zhang Y T, Sohn K, Villegas R, Pan G and Lee H. 2015. Improving object detection with deep convolutional networks via Bayesian optimization and structured prediction//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE: 249-258 [DOI: 10.1109/CVPR.2015.7298621]

Zheng L W, Fu C M and Zhao Y. 2018. Extend the shallow part of single shot multibox detector via convolutional neural network//Proceedings of SPIE 10806, 10th International Conference on Digital Image Processing. Shanghai: SPIE: # 1080613 [DOI: 10.1117/12.

2503001]

Zhong Z Y, Sun L and Huo Q. 2019. An anchor-free region proposal network for Faster R-CNN-based text detection approaches. International Journal on Document Analysis and Recognition, 22(3): 315-327 [DOI: 10.1007/s10032-019-00335-y]

Zhou X Y, Wang D Q and Krähenbuhl P. 2019. Objects as points [EB/OL]. [2019-12-01]. https://arxiv.org/pdf/1904.07850.pdf

Zhu C C, He Y H and Savvides M. 2019. Feature selective anchor-free module for single-shot object detection//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE: 840-849 [DOI: 10.1109/CVPR.2019.00093]

Zhu P F, Wen L Y, Bian X, Ling H B and Hu Q H. 2018. Vision meets drones: a challenge [EB/OL]. [2019-12-01]. https://arxiv.org/pdf/1804.07437.pdf

Zhu Y S, Zhao C Y, Wang J Q, Zhao X, Wu Y and Lu H Q. 2017. CoupleNet: coupling global structure with local parts for object detection//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE: 4146-4154 [DOI: 10.1109/iccv.2017.444]

Zitnick C L and Dollár P. 2014. Edge boxes: locating object proposals from edges//Proceedings of the 13th European Conference on Computer Vision. Zurich: Springer: 391-405 [DOI: 10.1007/978-3-319-10602-1_26]

作者简介

申奉璨，1997 年生，男，硕士研究生，主要研究方向为可见光及红外目标的检测与跟踪。E-mail: xiaoshenrobert@outlook.com

张萍，通信作者，女，副教授，主要研究方向为图像与视频信号的处理、目标检测与跟踪、机器学习与人工智能、计算机视觉以及深度学习。E-mail: pingzh@uestc.edu.cn

罗金，男，硕士研究生，主要研究方向为可视化目标跟踪算法。E-mail: l.jim@foxmail.com

刘松阳，男，本科生，主要研究方向为基于机器视觉的物体特征识别。E-mail: 15520730127@163.com

冯世杰，男，本科生，主要研究方向为仪表盘器件的刻度识别与自动读数算法。E-mail: ShijieFeng97@163.com