弱监督学习下的目标检测算法综述

周小龙[1,2] 陈小佳[1] 陈胜勇[1] 雷帮军[2]
(浙江工业大学计算机科学与技术学院 杭州 310023)[1]
(湖北省水电工程智能视觉监测重点实验室(三峡大学) 湖北 宜昌 443002)[2]

摘 要 目标检测是计算机视觉领域的基本问题之一,基于监督学习的目标检测算法是当前目标检测的主流算法。在现有的研究中,高精度的图像标记是强监督学习目标检测能够获得良好性能的前提。然而,实际场景中背景的复杂性以及目标的多样性等因素,使得图像标注任务非常费时费力。随着深度学习的不断发展,如何通过低成本的图像标注获得良好的训练结果成为当前的研究重点。文中主要综述了基于图像级别标签的弱监督目标检测算法,首先介绍了目标检测的发展历程,主要基于强监督学习对目标检测算法进行了阐述并指出其训练数据的局限性;然后从图像分割、多示例学习以及卷积神经网络 3 个方面对弱监督目标检测方法进行了分析,从显著性学习、多网络协作学习等角度对多示例学习和卷积神经网络进行了详细的描述;最后通过实验对弱监督学习下的多种主流方法进行了横向比较,并且将其与当前主流的强监督目标检测算法进行了比较。实验结果表明:弱监督学习已经取得了很大的进步,卷积神经网络的应用极大地促进了弱监督目标检测算法的发展,逐步替代了传统的多示例学习方法,尤其是采用了联合算法之后在 Pascal VOC 2007 上的准确率有了显著提高,达到了 79.3%。但是由于其性能依然低于强监督学习下的目标检测算法,因此弱监督目标检测依然有很大的发展空间。基于卷积神经网络的联合算法逐渐成为当前基于弱监督学习的目标检测的主流方法。

关键词 目标检测,弱监督学习,图像分割,多示例学习,卷积神经网络
中图法分类号 TP391.4 **文献标识码** A **DOI** 10.11896/jsjkx.181001899

Weakly Supervised Learning-based Object Detection: A Survey

ZHOU Xiao-long[1,2] CHEN Xiao-jia[1] CHEN Sheng-yong[1] LEI Bang-jun[2]
(College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China)[1]
(Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering (China Three Gorges University), Yichang, Hubei 443002, China)[2]

Abstract Object detection is one of the fundamental problems in the field of computer vision. Currently, supervised learning-based object detection algorithm is one of the mainstream algorithms for object detection. In the existing researches, high-precision image labels are the precondition of supervised object detection to gain good performance. However, it becomes more and more difficult to gain accurate labels due to the complexity of background and variety of objects in a real scenario. With the development of deep learning, how to receive good performance with the low-cast image labels becomes the key point in this field. This paper mainly introduced object detection algorithms based on weakly supervised learning with image-level labels. Firstly, this paper described the background of object detection and pointed out the shortcomings of training data. Then, it reviewed weakly supervised object detection algorithm based on image-level labels from three aspects: image segmentation, multi-instance learning and convolutional neural network. The multi-instance learning and convolutional neural network were comprehensively illustrated in several ways like saliency learning and collaborative learning. Finally, this paper compared mainstream algorithms based on weakly supervised learning horizontally and compared them with object detection algorithms based on supervised learning. The results prove that weakly supervised object detection algorithm has achieved great progress, especially the convolutional neural network has greatly promoted the development and gradually replaced multi-instance learning. After taking fusion algorithm, its accuracy rate is remarkably increased to 79.3% on Pascal VOC 2007. However, it still performs worse than

到稿日期:2018-10-12 返修日期:2019-04-12 本文受国家自然科学基金(61876168,61403342,61325019,61603341,61402415),浙江省自然科学基金(LY18F030020),湖北省水电工程智能视觉监测重点实验室开放基金(2017SDSJ03),湖北省创新群体项目(2015CFA025)资助。
周小龙(1986-),男,博士,副教授,CCF 会员,主要研究方向为目标跟踪、视线跟踪、模式识别,E-mail:zxl@zjut.edu.cn;**陈小佳**(1994-),男,硕士,主要研究方向为目标检测;**陈胜勇**(1973-),男,博士,教授,博士生导师,CCF 会员,主要研究方向为计算机视觉、图像处理,E-mail:csy@zjut.edu.cn;**雷帮军**(1973-),男,博士,教授,博士生导师,CCF 会员,主要研究方向为智能监控、计算机视觉。
万方数据

supervised object detection algorithm. To achieve better performance, the fusion algorithm based on convolutional neural network is becoming a mainstream algorithm in weakly supervised object detection.

Keywords Object detection, Weakly supervised learning, Image segmentation, Multi-instance learning, Convolutional neural network

1 引言

目标检测作为计算机视觉领域最基本的问题之一，被广泛应用于目标跟踪[1]、行为理解[2]、人机交互[3]、人脸识别[4]等诸多领域，在 20 世纪初就吸引了众多学者的广泛关注和研究。目标检测的目的是判断图中是否包含目标并且找出其位置，其算法框架如图 1 所示。

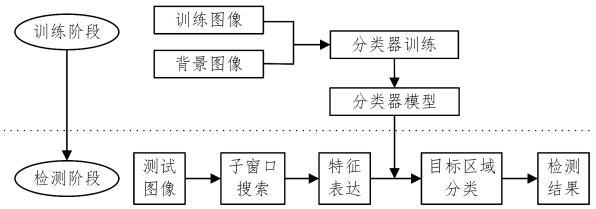


图 1 目标检测框架

Fig. 1 Object detection framework

由于缺乏对图像的有效表达且计算机资源相对落后，早期的目标检测算法主要是基于手工特征的。Viola 等[5]提出基于 Haar 特征和 Adaboost 级联算法的 Viola-Jones(VJ)人脸检测器，极大地促进了目标检测的发展。Dalal 等[6]提出了梯度方向直方图特征(HOG)，并且采用支持向量机[7](SVM)作为分类器，进一步提高了目标检测的精度，在行人检测上取得了重大突破。在 HOG 检测器的基础上，Felzenszwalb 等[8]提出可变形部件模型(DPM)，将基于手工特征的目标检测算法推到顶峰。其主要方法为：将对目标的检测拆分为对目标各个部分的检测，然后再通过聚合得到最终结果。该算法模型并没有给出目标各部分的标注，因此采用了基于弱监督的策略。然而，在基层特征表达的基础上构建复杂模型的方法已经不能适应高精度和高速度的要求，而深度学习的引入为目标检测带来了新的发展方向。最初，基于卷积网络提出的 OverFeat[9]使用分类滑动窗口机制进行回归，相对于传统方法提升不大。与此同时，Girshick 等[10]提出了基于区域卷积神经网络的目标检测模型(R-CNN)，打开了深度学习目标检测的大门。然而，该模型的检测结果严重依赖于区域的数量和质量，检测速度相当缓慢。He 等[11]提出了 SPPNet，其在卷积层和全连接层之间增加了空间金字塔池化层(Spatial Pyramid Pooling，SPP)，避免了缩放候选区域的操作。Girshick 等[12]提出了 Fast R-CNN 检测算法，实现了多任务学习的方式，大大提高了检测速度。随后，Ren 等[13]又提出了 FasterR-CNN 检测算法，设计了候选区域生成网络，首次实现了端到端的深度学习算法，速度和精度都得到了极大的提升，成为目标检测的基本框架，影响了很多后续研究[14-15]。Lin 等[16]在 Faster R-CNN 的基础上提出了 Feature Pyramid Networks(FPN)检测算法，将网络最顶层的特征图像逐层地进行反馈并且与前层的特征图像进行融合，所提算法在难度很大的 MSCOCO 数据集上取得了很好的检测结果。除了基于区域的目标检测算法，近两年提出的 YOLO[17]算法、SSD[18]算法、Retina-Net[19]算法等基于一体化卷积神经网络的检测算法抛弃了传统的检测流程，取得了更快的检测速度。

然而，基于监督学习的目标检测结果严重依赖于目标标注的准确性，而图像标注的结果很容易受到主观判断的影响。随着深度学习的不断发展，图像标注的成本变得越来越高，如何利用低成本的图像标注取得良好的检测结果成为了当下研究的热点，因此研究者开始研究基于弱监督学习的目标检测算法。与强监督需要大量人工标注不同，弱监督只需要提供图像级别的标签，即指出图像中是否包含某类目标，从而实现目标检测。

弱监督学习下的目标检测算法主要分为 3 类：基于分割的弱监督目标检测算法、基于多示例学习的弱监督目标检测算法以及深度学习下的弱监督目标检测算法。本文第 1 节为弱监督目标检测的相关背景介绍；第 2 节详细介绍弱监督目标检测的各种方法；第 3 节将不同的方法进行对比，并通过实验得出各种方法的优缺点；最后对弱监督学习下的目标检测方法进行总结与展望。

2 研究现状分析

2.1 基于分割的弱监督目标检测方法

最初，弱监督都是基于分割[20]的方法。图像分割的目的是把图像分割成若干个独立区域，并且提取出感兴趣的区域。因此，最初的图像分割是检测和识别的基础，没有正确的分割就没有正确的识别。然而，分割又会受到各种因素的影响，例如光照、噪声、阴影等。在深度学习之前，研究者主要通过设计巧妙的数学模型来进行弱监督学习下的图像分割与识别。Alexe 等[21]提出了分割与识别同时进行的方法，通过设计一个全局的分割能量函数，在每次分割之前对其进行优化，从而使模型有效地学习到物体的形状和位置。Joulin 等[22]在现有分割方法的基础上，采用正则化以及核方法实现目标识别。Vicent 等[23]利用前景分割大大改善了分割的结果。

然而，随着图像分割和目标检测在行为理解、自动驾驶、遥感识别等领域的广泛应用，传统的方法已经不能满足越来越高的快速性和准确性要求。近年来，深度学习的发展为该领域带来了极大的突破，弱监督学习开始有了更加明确的定义。一般来说，弱监督信息是指比人工标注的标签信息更弱的信息。对目标检测而言，图像的标签信息相对于目标的标签信息是一种弱监督信息；对图像分割而言，图像以及目标的标签信息相对于像素的标签信息是一种弱监督信息。而本文所说的弱监督信息一般都是基于图像级别的，这种带有图像标签的图片可以从互联网上大量获得。深度学习的引入，将基于分割的弱监督目标检测归类为了以下问题：如何找到图像级别的标签与图像像素之间的联系，从而进行分割，然后利用卷积神经网络来学习分类。Li 等[24]通过图像级别标签的

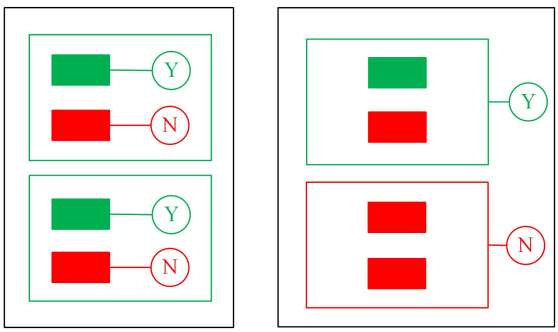
万方数据

数据集进行弱监督分割与识别，提出了 CCNN(Constrained Convolutional Neural Network)，将训练迭代过程看作是有线性限制条件的最优化。该实验在 Pascal VOC2012test 数据集上的最优 mloU 为 0.54，最优 FCN-8s 为 0.622。Liu 等[25]通过条件随机场(CRF)对图像的底层特征进行学习，并进一步计算了各个特征所占的比重，从而得到显著图。Yang 等[26]在 Liu 等的基础上将 CRF 和字典算法相结合，通过梯度下降法不断更新模型，最终实现目标检测。Deselasers 等[27]利用 CRF 同时学习目标特征，并进行目标检测，通过能量性函数得到目标显著图。Xu 等[28]进一步研究了弱监督学习下的图像分割，采用了期望值最大化算法(EM)来估计未标记的像素类别以及 CNN 的参数，再利用边界框进行自动分割和监督学习。对于只有图像级别标签的图像，文中提出观测图像的像素值和图像级别标记的方法，把每个像素的标号当作隐变量。对于给定的训练图片，首先使用 CRF 对图像进行自动分割，然后再进行全监督学习。该方法在 Pascal VOC2012test 数据集上的 mloU 为 0.622，若结合全监督学习其 mloU 可达到 0.69。由此可见，单纯的弱监督图像分割的性能有限，只有结合全监督学习才能达到较好的效果。因此，基于分割的弱监督学习下的目标检测算法依然是未来研究的难点。

2.2　基于多示例学习的弱监督目标检测方法

2.2.1　多示例学习下的检测算法

对于给定的只有弱标签信息的训练集，可以将每张图片形容为一个“包”，包中包含若干个示例，若目标包含在这些示例当中，则该图片是一个正包，否则为一个负包。这种表达往往被视为“多示例学习”[29-30]。该算法则是通过包含正包和负包的训练集得到一个检测器，从而确定最有可能是目标的候选框。强监督学习标签与弱监督学习标签的主要区别如图 2 所示。



(a)强监督学习标签　　(b)弱监督学习标签

图 2　强监督学习标签与弱监督学习标签的对比

Fig. 2　Comparison between supervised learning labels and weakly supervised learning labels

图 2 中，最外围的边框表示图像训练集，中间的边框表示图像，实心矩形表示图像的若干示例。图 2(a)给出了每个目标的标签(Y 表示该示例为某类目标，N 表示该示例不是任何一类目标)；而图 2(b)并没有对示例进行标注，仅仅指出了该图片中是否包含某类目标(Y 表示该图片中包含某类目标，N 表示该图片不包含任何目标)。Felzenszwalb 等[31]详细介绍了如何将多示例学习应用到目标检测领域。类似于监督学习下候选区域的重要性，多示例学习的检测结果也依赖于图像“包”的生成。Wei 等[32]总结了关于弱监督学习下包生成的各种算法。此外，不少学者将局部区域之间的相似性融入多示例学习。Wang 等[33]提出了一种构造图算法，选择一系列的窗口，将每个窗口和正包中最近的窗口连接起来。Bilen 等[34]提出了学习辨别性模型，并且通过级联算法得到目标区域的相似度。

尽管大多数方法将弱监督归类为多示例学习，但是多示例学习会导致非凸优化的问题。很多学者面临着局部优化的问题，因此最终的结果很大程度上取决于初始化的情况。一些研究致力于发展多样性初始化策略。Kumar 等[35]提出逐层把困难样本融入到初始化的训练集中的自步学习策略。Deselaers 等[36]通过目标得分来对目标区域进行初始化。Cinbis 等[37]提出数据集的多重分离模型，避免了局部优化。而另一些研究则尝试解决正则优化问题[38]。Cabral 等[39]把整张图片表示成不同部分的加权，通过这种方法可以将一张图片表示成不同的加权类别和背景损失，再通过低秩进行优化。

2.2.2　结合显著性学习的多示例检测算法

显著性目标检测作为一种视觉注意力机制，在视觉图像处理领域有着很重要的作用。对于计算机而言，传统的方法采用滑动窗口机制对图像进行详尽而盲目的搜索[40-41]，如图 3 所示。而对于图 4，当人们看到这张图片时往往会注意到图中的人，计算机通过模拟这种感知行为可以有效避免滑动窗口机制带来的大量低准确度数据。为了模拟人类对一张图片中的目标区域关注的感知行为，有关学者提出了基于显著性学习的目标检测模型[42-43]。对于给定的带有弱标签信息的图片，显著性检测可以得到大量最有可能包含目标的区域，然后再对这些区域进行训练及优化，从而得到目标区域。显著性学习的提出，极大地促进了弱监督学习下目标检测的发展。

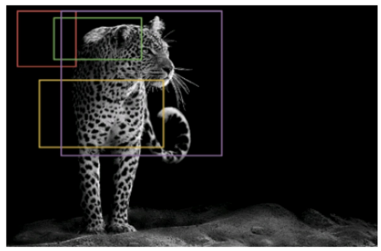


图 3　滑动窗口机制

Fig. 3　Sliding window mechanism

图 4　视觉注意力机制

Fig. 4　Visual attention mechanism

计算机在带有特定的经验或者目的的情况下处理一幅图

万方数据

片时，往往会比较关注图片的某些性质。为了获得与之相关的特征，往往会用监督学习的方式来获取感兴趣的目标的特征或者标签。图像的显著性检测往往可能得到几百张感兴趣的区域，通过这些区域的监督学习，最终可以达到目标的边框。Navalpakkam 等[44]提出通过学习感兴趣的目标表象特征来形成特征图，再将这些特征图自下而上地进行融合，然后计算各特征图的权重，从而得到显著图。Borji 等[45]在前人的基础上做了一个融合，结合了很多底层特征和高级特征，并通过多种分类器最终得到目标的效果图。Shi 等[46]提出主题模型，将多标记学习、半监督学习和自适应学习相结合，大大提高了显著性检测的准确性。

最初的显著性检测都是从图像底层出发，根据图像的颜色、纹理、灰度以及梯度等信息进行提取，找到与周围明显不同的显著性目标区域。Itti 等[47]提出 Itti 模型，通过颜色、方向和亮度 3 个特征对图像进行分层，计算不同特征对应的显著性区域，通过加权得到最终结果，其算法框架如图 5 所示。Itti 作为较早提出的视觉注意机制，主要是对图像进行多个特征通道和多尺度的分解，然后通过滤波得到特征图，最后再通过融合计算得到显著图。该模型已经成为自下而上的视觉注意模型的标准。Tsauro 等[48]首先提取图像的特征，然后通过算法得到每个特征下的显著图，最后通过归一化得到图像的显著图。Hou 等[49]提出将图像先转换到频域，对图像的频谱进行分析，以得到图像在频域下的显著图，从而实现图像检测。Goferman 等[50]首先得到图像局部以及全局的显著图，然后再结合局部显著图附近的区域得到可以代表图像场景的区域。

显著性学习结合多示例学习作为弱监督方向的经典算法，极大程度地促进了弱监督学习的发展。Wei 等[51]认为两个包之间的距离是两个包之间最近的两个示例之间的距离。Andrews 等[52]利用传统的支持向量机和多示例核方法对示例进行预测，并再次训练支持向量机进行迭代。Chen 等[53]利用相似性衡量，将每个图片映射到向量空间，并且排除了大量的冗余信息。Li 等[54]在传统的支持向量机的基础上提出 KI-SVM，将问题看作一个凸优化问题，并采用核学习方法进行学习。Russakovsky 等[55]提出了一个空间池模型，用来判断目标物体的位置，并把前景和背景的特征分开。上述方法都在弱监督学习中取得了不错的效果。

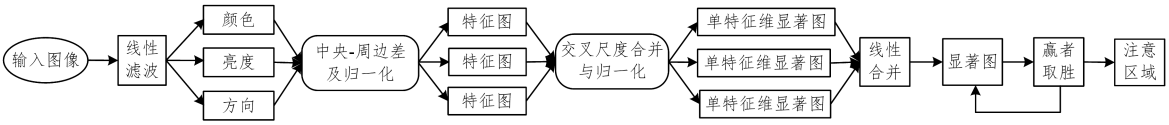


图 5 Itti 算法框架

Fig. 5 Itti algorithm framework

2.3 基于卷积神经网络的弱监督目标检测方法

2.3.1 基于单一卷积神经网络的训练方法

随着神经网络的不断发展，目标的标注成本变得越来越高。尽管深度学习基本都是基于强监督学习的，但是神经网络极强的非线性映射能力可以很好地描述向量的空间特征，卷积神经网络已经成为目标检测最为主流的框架。由于网络学习到的特征可以很好地迁移到其他的检测任务中，因此卷积神经网络同样适用于弱监督学习。Zhou 等[56]提出，全局平均池化层不仅仅可以用来正则化，其对于带有图像级别标签的样本同样具有很好的目标定位能力。如图 6 所示，该算法使用的网络结构大部分都是卷积层，只在输出层之前使用了全局平均池化层，并且将它们作为得出分类的全连接层，此结构可以通过输出层权重映射回卷积层特征的方式把图像中的重要区域标记出来。该方法为很多后来的研究提供了参考。

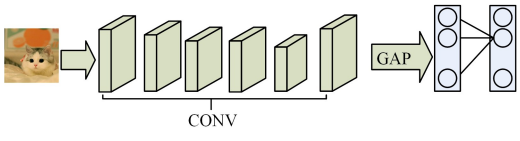


图 6 用 CNN 的全局平均池化层生成 CAM 的过程

Fig. 6 Process of generating CAM by global average pooling

然而，事实上基于深度卷积网络的弱监督目标检测算法还是将问题归类于多示例学习问题。Cinbis 等[37]将多示例学习和卷积网络模型相结合，实现了目标定位，结果证明卷积神经网络学习的图像模型可以更好地描述图像模型。Oquab 等[57]提出转换图像代表区域，表明一些目标定位可以通过评估卷积神经网络在重叠区域的输出结果来实现。Oquab 等[58]在之前的基础上，从给定的训练集中准确地预测了图像标签。随后，该团队[59]又使用了一个预训练卷积神经网络模型来计算带有弱标签信息的图片，并且在 Pascal VOC 上取得了不错的成果。紧接着，Hong 等[60]修改了网络架构，使其可以在预测目标标签的同时对示例进行粗略定位。Bilen 等[61]提出了一个弱监督深度检测网络，通过修改网络使其能够训练带有弱标签的区域。也有一些学者从标签信息入手，因为弱监督学习仅仅依赖于图像标签，即指出图像中是否包含某类目标，所以这些研究的目标是从基于图像标签的图像中得到目标的位置和类别。

2.3.2 基于卷积神经网络的融合算法

尽管卷积神经网络已经在弱监督问题上取得了很大的进步，但是单纯地对带有弱标签的训练集进行训练的结果依然不及纯粹的监督学习。之前，基于卷积神经网络的弱监督检测方法绝大部分都是通过提取大量目标区域来表示图片。虽然以上方法都取得了一定的效果，但是由于卷积神经网络仅仅被用于候选区域的提取和分类，因此依然存在两个问题：1）由于图像没有目标级别的标签信息，因此候选区的选择往往比较困难，通常都伴随有大量噪音，几千个候选区域中可能只有几个是包含目标的，增加了大量的时间成本，同时也提高了错误率；2）传统的检测模型无法适应更加复杂的深度模型，导致弱监督模型的训练难度很大，因此，很多学者在深度学习的基础上，创新性地融入了诸如监督学习等方法，使得传统的

万方数据

深度模型能够更好地迁移到弱监督检测任务上。

最近的研究发现，人类的认识过程并不是一次性放在某个场景上，而是逐渐从整个场景的不同区域同时提取信息。学者将这种认知学习融入到很多高难度的学习任务中[62]。张文等[63]引入同时提取不同区域进行学习的注意力机制，结合弱监督网络提出了一种弱监督多标号分类算法，该算法把卷积神经网络的特征作为递归神经网络的输入，这样每一次递归神经网络只会关注图像的局部区域，在下一个步骤重新关注新的区域，如此多步之后的最终结果就是预测结果。该方法提出的基于注意力机制并且结合递归神经网络的模型在结果上取得了很大的提升。类似于结合递归神经网络，周明非等[64]提出了一种基于遥感图像的目标检测框架，解决了遥感图像中目标过小、背景复杂的问题。其提出的检测框架包括卷积神经网络和全卷积网络两种深层神经网络。全卷积网络主要是提取遥感图像中可能存在目标的候选区域，卷积神经网络则用来对候选区域进行分类。此外，他们还提出了一种融合算法，用于调整候选框的位置。该算法提高了检测的准确性，并且具有更快的检测速度。

近两年，有学者提出了弱监督多层协作学习框架（见图 7），将监督学习和弱监督学习相结合。给定两个相关模型，一个是弱监督模型，一个是监督模型，通过弱监督多层协作模型将它们结合起来[38,61]。然而，之前的很多研究都仅考虑了两个检测器之间单方面的联系，使得监督训练的检测结果严重依赖于弱监督检测器。为了解决这个问题，Wang 等[33]提出了弱监督协作学习（WSCL）框架，将弱监督学习网络和强监督学习网络连接成一个整体，通过一致性损失约束强监督网络和弱监督学习网络，使其具有相似的结果，通过部分特征共享来保证两个网络的一致性，从而实现弱监督协作学习。协作学习的提出为目标检测提供了一个重要的发展方向。

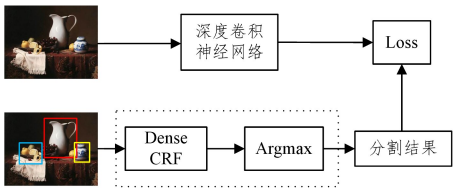


图 7　弱监督协同学习的框架

Fig. 7　Collaborative learning framework for weakly supervised object detection

除了以上方法，近两年还有学者提出了基于域适应的检测框架。Tzeng 等[65]提出了一种对抗判别域适应方法，其通过固定源域映射是目标映射来靠近源映射，通过损失函数进行优化来完成域适应，优化后的目标模型将被用于目标数据中的无监督学习。Inoue 等[66]主要通过域迁移技术即图像转换技术将域源数据转换为带有示例标签的图像，或者通过伪标记对目标域数据产生伪示例标注。由于深度学习具有强大的迁移能力，因此基于卷积神经网络的弱监督学习的目标检测算法依然有很大的发展空间。

3　实验对比

本文主要选取了当前弱监督目标检测比较主流的方法，并在 Pascal VOC2007 数据集上对其进行了比较。其中，Bilen 等[61]提出了两种方法，方法 1 使用了单个带有加权池的神经网络，方法 2 引入了目标边界框和空间正则化进行协作学习；Kantorov[69]在卷积神经网络的基础上增加了上下文信息；Cinbis 等[37]采用了传统的多示例学习方法；Tang 等[38]简化了 Bilen 等[61]提出的方法 1 的网络结构；Li 等[67]提出了一个打分机制，通过联合一个能选择最高得分区域的监督学习检测模型；Jie[70]和 Wang[33]通过模型通道在线分享区域结果，并且训练了一个基于目标级别标签的模型进行协作学习；Tzeng 等[68]和 Inoue 等[66]基于域适应提出了新的解决方案；Zhang[63]则联合了卷积神经网络和递归神经网络，通过逐步视觉机制提升了其性能。

此外，本文还在 Pascal VOC2012 数据集上将以上算法与监督学习下的目标检测算法进行了比较，选取了 Tang 等[38]和 Wang[33]提出的两种弱监督目标检测算法，以及 FasterR-CNN[13]和 YOLO[17]两种经典的强监督学习算法。

3.1　弱监督学习下目标检测算法的比较

当前弱监督目标检测主要偏向于两个方向，一个是将其视为多示例学习，结合深度卷积神经网络进行比较；另一个是将弱监督与强监督通过共享特征相结合来进行协作学习。本文主要选取了两个评判标准：平均精确率均值（mAP）和定位准确率（CorLoc）。因为目标检测模型中的分类和定位都需要进行评估，但是每个图像都可能具有不同类别的不同目标，所以图像分类问题中的标准度量不能用于目标检测问题上，故我们选用 mAP。此外，对于弱监督学习，因为没有给出目标的边界框，所以 CorLoc 也是衡量检测结果的重要标准。

表 1 列出了当前较为主流的弱监督目标检测算法的平均精确率，其中的前 4 种方法是基于单个弱监督模型的，中间 4 种方法则是采用多个网络进行协作学习。采用单个训练模型的平均精确率均值依次是 34.9%，36.3%，27.4% 以及 41.2%，而采用了协作学习方式的结果则是 39.3%，39.5%，41.7% 和 48.3%。后 2 种则是基于域适应的检测算法，其结果依此为 27.4% 和 55.4%。最后一种方法是联合了逐层视觉显著机制的检测方法，其结果为 79.3%。

结果表明，使用了协作学习方式后，弱监督目标检测的结果得到了明显的提升。以文献[61]中的单个卷积网络算法为例，其在使用了单个带有加权池的卷积神经网络之后的 *mAP* 为 34.9%，但是在引入了目标边界框和空间正则化进行协作学习之后，其结果达到了 39.3%，性能得到了提升。Tang 简化了 Bilen 的网络结构，*mAP* 为 41.2%，这个结果已经达到了协作学习的性能，接近于目前单模型弱监督目标检测的最高水平。Wang 在 Jie 的基础上将结果提升了 6.5%。Tzeng 提出了基于域适应的对抗网络模型，然而损失函数导致的震荡问题使得其结果只达到了 27.4%，但其对后面 Inoue 等提出的方法的产生了一定的影响。Inoue 等所提方法的平均精确率均值达到了55.4%，在当前主流的弱监督检测方法中取得了较好的成绩。Zhang 等提出将逐层视觉机制和卷积神经网络相融合，使其平均精确率均值达到了 79.3%，效果最优，基本上与当前的强监督学习方式处于同一个水平。

万方数据

表 1 Pascal VOC 2007 上平均精确率的比较

Table 1 Comparison on Pascal VOC 2007 test set in terms of average precision

(单位:%)

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Bilen[61]	43.6	50.4	32.2	26.0	9.8	58.5	50.4	30.9	7.9	36.1	18.2	31.7	41.4	52.6	8.8	14.0	27.8	46.9	53.4	47.9	34.9
Kantorov[60]	57.1	52.0	31.5	7.6	11.5	55.0	53.1	34.1	1.7	33.1	49.2	42.0	47.3	56.6	15.3	12.8	24.8	48.9	44.4	47.8	36.3
Cinbis[37]	38.1	47.6	28.2	13.9	13.2	45.2	48.0	19.3	17.1	27.7	17.3	19.0	30.1	45.4	13.5	17.0	28.8	24.8	38.2	15.0	27.4
Tang[38]	58.0	62.4	31.1	19.4	13.0	65.1	62.2	28.4	24.8	44.7	30.6	25.3	27.8	65.5	15.7	24.1	41.7	46.9	64.3	62.6	41.2
Bilen[61]	46.4	58.3	35.5	25.9	14.0	66.7	53.0	39.2	8.9	41.8	26.6	38.6	44.7	59.0	10.8	17.3	40.7	49.6	56.9	50.8	39.3
Li[67]	54.5	47.4	41.3	20.8	17.7	51.9	63.5	46.1	21.8	57.1	22.1	34.4	50.5	61.8	16.2	29.9	40.7	15.9	55.3	40.2	39.5
Jie[70]	52.2	47.1	35.0	26.7	15.4	61.3	66.0	54.3	3.0	53.6	24.7	43.6	48.4	65.8	6.6	18.8	51.9	43.6	53.6	62.4	41.7
Wang[33]	61.2	66.6	48.3	26.0	15.8	66.5	65.4	53.9	24.7	61.2	46.2	53.5	48.5	66.1	12.1	22.0	49.2	53.2	66.2	59.4	48.3
Tzeng[68]	20.1	50.2	20.5	23.6	11.4	40.5	34.9	2.3	39.7	22.3	27.1	10.4	31.7	53.6	46.6	32.1	18.0	21.1	23.6	18.3	27.4
Inoue[66]	50.5	60.3	40.1	55.9	34.8	**79.7**	61.9	13.5	56.2	**76.1**	57.7	36.8	63.5	**92.3**	76.2	49.8	40.2	**28.1**	60.3	74.4	55.4
Zhang[63]	**90.2**	**89.0**	**93.1**	**88.3**	**48.2**	79.3	**93.8**	**86.4**	**60.3**	72.5	**71.5**	**83.2**	**90.1**	81.2	**90.6**	59.2	**78.4**	**65.3**	**91.5**	**73.6**	**79.3**

由表 1 可知,在 bike,bus,car,cow,mbike,tv 的检测上往往会取得较好的结果,而在 bottle,person 等检测上的精度较低。这是由于前者在图片中所占的比例较大,所对比的方法绝大部分都采用了基于全局的训练方法,因此对较大的目标具有更好的鲁棒性。此外,采用了 Tzeng 以及 Inoue 的方法后,小目标的精度得到了明显的提高。由此可知,利用基于域适应的训练方法后,小目标的噪音得到了明显的降低。而 Zhang 的方法之所以取得了如此优越的性能,主要还是因为其采用了递归神经网络逐层学习以及验证集的先验知识。除了以上问题之外,遮挡、光照等问题同样会对检测结果造成很大的影响。

表 2 列出了定位准确率(CorLoc)的比较结果。单模型的 CorLoc 结果依次为 56.1%,55.1%,47.3%和 60.0%,而采用了协作学习之后的结果为 58.0%,52.4%,56.1%和 64.7%。与表 1 结果不同的是,在采用了协作学习之后,CorLoc 的效果并没有得到明显的提升。由此可知,尽管强监督模型和弱监督模型是有共享通道的,但是强监督模型并不能对弱监督模型的结果产生特别大的影响。

表 2 Pascal VOC 2007 上定位准确率的比较

Table 2 Comparison on Pascal VOC 2007 test set in terms of correct localization

(单位:%)

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	CorLoc
Bilen[61]	65.1	63.4	59.7	45.9	38.5	69.4	77.0	50.7	30.1	58.8	34.0	37.3	61.0	82.9	25.1	42.9	79.2	59.4	68.2	64.1	56.1
Kantorov[60]	83.3	68.6	54.7	23.4	18.3	73.6	74.1	54.1	8.6	65.1	**47.1**	59.5	67.0	83.5	**35.3**	39.9	67.0	49.7	63.5	65.2	55.1
Cinbis[37]	57.2	62.2	50.9	37.9	23.9	64.8	74.4	24.8	29.7	64.1	40.8	37.3	55.6	68.1	25.5	38.5	65.2	35.8	56.6	33.5	47.3
Tang[38]	81.7	**80.4**	48.7	**49.5**	32.8	**81.7**	**85.4**	40.1	**40.6**	**79.5**	35.7	33.7	60.7	88.8	21.8	**57.9**	76.3	59.9	**75.3**	**81.4**	60.6
Bilen[61]	68.9	68.7	65.2	42.5	40.6	72.6	75.2	53.7	29.7	68.1	33.5	45.6	65.9	86.1	27.5	44.9	76.0	62.4	66.3	66.8	58.0
Li[67]	78.2	67.1	61.8	38.1	36.1	61.8	78.8	55.2	28.5	68.8	18.5	49.2	64.1	73.5	21.4	47.4	64.6	22.3	60.9	52.3	52.4
Jie[70]	72.7	55.3	53.0	27.8	35.2	68.6	81.9	60.7	11.6	71.6	29.7	54.3	64.3	88.2	22.2	53.7	72.2	52.6	68.9	75.5	56.1
Wang[33]	**85.8**	**80.4**	**73.0**	42.6	**36.6**	79.7	82.8	**66.0**	34.1	78.1	36.9	**68.6**	**72.4**	**91.6**	22.2	51.3	**79.4**	**63.7**	74.5	74.6	**64.7**

3.2 弱监督学习与强监督学习的比较

本文主要选取了 Tang 等[38]和 Wang 等[33]提出的两种弱监督算法,其在弱监督学习中均取得了很好的成绩。另外,本文选取了目标检测领域两种经典的深度学习算法,即 Ren 等[13]提出的 FasterR-CNN 和 Redmon 等[17]提出的 YOLO,在 Pascal VOC2012 数据集上将其检测的平均精确率均值进行比较,结果如表 3 所列。

由表 3 可知,Tang 和 Wang 的方法的平均精确率均值分别为 41.2%和 48.3%,与之相应的是 FasterR-CNN 取得的 70.4%和 57.9%,而 Zhang 的方法则取得了 79.3%的成绩。可以看到,尽管大部分弱监督学习可以摆脱算法对目标标签的要求,但是其总体检测结果依然远低于强监督学习的结果。表 3 中,尽管 Zhang 的结果好于强监督学习的结果,但是其网络框架是在强监督学习的基础上进行弱监督学习的,并且通过视觉机制逐层感知,对于弱监督学习有很大的启示作用。除了 Zhang 所提出的算法外,在检测精度方面显然是强监督学习下的检测算法更胜一筹。此外,强监督学习解决了弱监督学习中行人等目标 mAP 较低的情况。总的来说,尽管近年来弱监督学习不断发展,并且取得了不错的成绩,但是相对于强监督学习,其依然有很大的发展空间。

表 3 弱监督学习与强监督学习结果的比较

Table 3 Comparison between weakly supervised learning and supervised learning

(单位:%)

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Tang[38]	58.0	62.4	31.1	19.4	13.0	65.1	62.2	28.4	24.8	44.7	30.6	25.3	27.8	65.5	15.7	24.1	41.7	46.9	64.3	62.6	41.2
Wang[33]	61.2	66.6	48.3	26.0	15.8	66.5	65.4	53.9	24.7	61.2	46.2	53.5	48.5	66.1	12.1	22.0	49.2	53.2	66.2	59.4	48.3
Zhang[63]	**90.2**	**89.0**	**93.1**	**88.3**	48.2	**79.3**	**93.8**	86.4	**60.3**	72.5	**71.5**	83.2	**90.1**	**81.2**	**90.6**	**59.2**	**78.4**	**65.3**	**91.5**	**73.6**	**79.3**
Ren[13]	84.9	79.8	74.3	53.9	**49.8**	77.5	75.9	**88.5**	45.6	**77.1**	55.3	**86.9**	81.7	80.9	79.6	40.1	72.6	60.9	81.2	61.5	70.4
Redmon[17]	77.0	67.2	57.7	38.3	22.7	68.3	55.9	81.4	36.2	60.8	48.5	77.2	72.3	71.3	63.5	28.9	52.2	54.8	73.9	50.8	57.9

万方数据

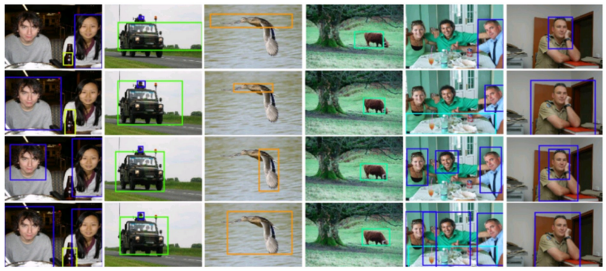


图 8　4 种检测器的检测结果比较[33]

Fig. 8　Visualization of detection results of four detectors[33]

图 8 中，第一行是单个的弱监督检测模型进行训练的结果，第二行和第四行是采用了协作学习方式后的结果，第三行则是强监督学习迭代训练的结果。从图中可知，只采用了弱监督学习框架的检测结果表现最差，产生的包围框不够全面，而且不能覆盖 bird，car 等目标；而采用了弱监督与强监督协作学习之后，检测效果明显得到了提升，基本可以对目标进行准确的覆盖，但是依然不够全面；表现最好的是强监督学习，其可以得到更加紧凑、全面的包围框。

以上的实验对比充分说明了一个问题：单纯的弱监督目标检测框架很难达到目标检测的准确度要求，弱监督学习的性能依然远远低于强监督学习的性能。因此，将弱监督学习和强监督学习相结合进行协作学习，是弱监督目标检测未来发展的趋势。根据以上实验对比分析，提出弱监督目标检测的未来研究方向：

(1)实验表明，强监督学习框架并不能明显地提升弱监督学习目标定位的准确率，因此可以将强监督学习网络中的位置信息共享给弱监督网络，并且可以在弱监督学习网络中增加一个反馈环节，将得到的目标定位信息再次送回网络进行处理，从而提高弱监督目标检测的定位准确率。

(2)弱监督学习和强监督学习相结合尽管可以取得更好的结果，但依然没有避免对大量高精度样本的要求。对于很少数的高精度样本，是否可以通过机器学习模型进行学习并且与弱监督学习相融合，值得进一步的探讨。

结束语　本文主要介绍了弱监督学习下的目标检测算法，其大致可分为基于分割的算法、基于多示例学习的算法和基于卷积神经网络的算法。近年来，随着深度学习的发展，弱监督学习在卷积神经网络的基础上不断进步，成为了当前的主流方法。通过实验对比发现，尽管目前弱监督学习已经取得了不错的性能，但是其与强监督学习相比依然有很大的差距。单纯的弱监督学习很难达到强监督学习的性能，然而强监督学习又太依赖于大量精确的标注数据，因此基于卷积神经网络的联合算法是目标检测领域发展的一条重要途径。

参 考 文 献

[1] ZHOU X，LI Y，HE B，et al. GM-PHD-Based Multi-Target Visual Tracking Using Entropy Distribution and Game Theory [J]. IEEE Transactions on Industrial Informatics，2014，10(2)：1064-1076.

[2] SHAO Z，LI Y. On Integral Invariants for Effective 3-D Motion Trajectory Matching and Recognition [J]. IEEE Transactions on Cybernetics，2016，46(2)：511-523.

[3] ZHOU X，CAI H，LI Y，et al. Two-Eye Model-Based Gaze Estimation from A Kinect Sensor[C]// Proceedings of IEEE International Conference on Robotics and Automation. New York：IEEE Press，2017：1646-1653.

[4] ZHENG J，YANG P，CHEN S，et al. Iterative Reconstrained Group Sparse Face Recognition with Adaptive Weights Learning [J]. IEEE Transactions on Image Processing，2017，26(5)：2408-2423.

[5] VIOLA P，JONES M. Rapid Object Detection Using a Boosted Cascade of Simple Features[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2001：511-518.

[6] DALAL N，TRIGGS B. Histograms of Oriented Gradients for Human Detection [C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2014：886-893.

[7] LIAO S，JAIN A，LI S. A Fast and Accurate Unconstrained Face Detector [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2016，38(2)：211-223.

[8] FELZENSZWALB P，MCALLESTER D，RAMANAN D. A Discriminatively Trained，Multiscale，Deformable Part Model[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2008：1-8.

[9] HE K，ZHANG X，REN S，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.

[10] GIRSHICK R，DONAHUE J，DARRELL T，et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2013：580-587.

[11] HE K，ZHANG X，REN S，et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2015，37(9)：1904-1916.

[12] GIRSHICK R. Fast R-CNN[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2015：1440-1448.

[13] REN S，HE K，GIRSHICK R，et al. Faster R-CNN：Towards Real-Time Object Detection with Region Proposal Networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2017，39(6)：1137-1149.

[14] ZHANG H，KYAW Z，YU J，et al. PPR-FCN：Weakly Supervised Visual Relation Detection via Parallel Pairwise R-FCN[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2017：4243-4251.

[15] ZHANG L，LIN L，LIANG X，et al. Is Faster R-CNN Doing Well for Pedestrian Detection? [C]// Proceedings of European Conference on Computer Vision. Berlin：Springer，2016：443-457.

[16] LIN T，DOLLAR P，GIRSHICK R，et al. Feature Pyramid Networks for Object Detection [C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York：

万方数据

IEEE Press,2017:936-944.
[17] REDMON J,DIVVALA S,GIRSHICK R,et al. You Only Look Once:Unified,Real-Time Object Detection[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York:IEEE Press,2016:779-788.
[18] LIU W,ANGUELOV D,ERHAN D,et al. SSD:Single Shot MultiBox Detector[C]// Proceedings of European Conference on Computer Vision. Berlin:Springer,2016:21-37.
[19] LIN T,GOYAL P,GIRSHICK R,et al. Focal Loss for Dense Object Detection[C]// Proceedings of IEEE International Conference on Computer Vision. New York: IEEE Press, 2017: 2999-3007.
[20] FELZENSZWALB P, HUTTENLOCHER D. Efficient Graph-Based Image Segmentation [J]. International Journey of Computer Vision,2004,59(2):167-181.
[21] ALEXE B, DESELAERS T, FERRARI V. Classcut for Unsupervised Class Segmentation [C] // Proceedings of European Conference on Computer Vision. Berlin: Springer, 2010: 380-393.
[22] JOULIN A, BACH F, PONCE J. Discriminative Clustering for Image Co-segmentation[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York:IEEE Press,2010:1943-1950.
[23] VICENTE S,ROTHER C,KOLMOGOROV V. Object Cosegmentation[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York:IEEE Press,2011: 2217-2224.
[24] LI J,LI X,YANG B,et al. Segmentation-Based Image Copy-Move Forgery Detection Scheme[J]. IEEE Transactions on Information Forensics and Security,2017,10(3):507-518.
[25] LIU T, YUAN Z, SUN J, et al. Learning to Detect A Salient Object [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2011,33(2):353-367.
[26] YANG M, YANG J. Top-down Visual Saliency Via Joint CRF and Dictionary Learning[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York:IEEE Press,2012:2296-2303.
[27] DESELASERS T. Weakly Supervised Localization and Learning with Generic Knowledge [J]. International Journal of Computer Vision,2012,100(3):275-293.
[28] XU J,SCHWING A,URTASUN R. Learning to Segment Under Various Forms of Weak Supervision[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York:IEEE Press,2015:3781-3790.
[29] ZHOU Z. A Brief Introduction to Weakly Supervised Learning [J]. National Science Review,2018,5(1):44-53.
[30] ZHOU Z. Multi-instance Learning From Supervised View [J]. Journal of Computer Science and Technology,2006,21(5):800-809.
[31] FELZENSZWALB P,GIRSHICK R,MCALLESTER D,et al. Object Detection with Discriminatively Trained Part-based Models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2010,32(9):1627-1645.
[32] WEI X,ZHOU Z. An Empirical Study on Image Bag Generators for Multi-instance Learning [J]. Kluwer Academic Publishers, 2016,105(2):1-44.
[33] WANG C,REN W,HUANG K,et al. Weakly Supervised Object Localization with Latent Category Learning[C]// Proceedings of European Conference on Computer Vision. Berlin: Springer, 2014:431-445.
[34] BILEN H,PEDERSOLI M,TUYTELAARS T. Weakly Supervised Object Detection with Convex Clustering[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York:IEEE Press,2015:1081-1089.
[35] KUMAR M,PACKER B,KOLLER D. Self-paced Learning for Latent Variable Models[C]// Proceedings of International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc,2010:1189-1197.
[36] DESELAERS T,ALEXE B,FERRARI V. Localizing Objects While Learning Their Appearance[C]// Proceedings of European Conference on Computer Vision. Berlin: Springer, 2010: 452-466.
[37] CINBIS R,VERBEEK J,SCHMID C. Weakly Supervised Object Localization with Multi-fold Multiple Instance Learning [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(1):189-203.
[38] TANG P,WANG X,BAI X,et al. Multiple Instance Detection Network with Online Instance Classifier Refinement[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York:IEEE Press,2017:3059-3067.
[39] CABRAL R,TORRE F,COSTEIRA J,et al. Matrix Completion for Weakly-supervised Multi-label Image Classification [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2015,37(1):121-135.
[40] DALAL N,TRIGGS B. Histograms of Oriented Gradients for Human Detection [C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press,2014:886-893.
[41] VIOLA P,JONES M. Robust Real-time Face Detection [J]. Internal Journey of Computer Vision,2004,57(2):137-154.
[42] CHENG M,ZHANG G,MITRA N,et al. Global Contrast Based Salient Region Detection[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York:IEEE Press,2011:409-416.
[43] JIANG H,WANG J,YUAN Z,et al. Salient Object Detection:A Discriminative Regional Feature Integration Approach [C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York:IEEE Press,2013:2083-2090.
[44] NAVALPAKKAM V,ITTI L. Modeling the Influence Of Task on Attention [J]. Vision Research,2005,45(2):205-231.
[45] BORJI A. Boosting Bottom-up and Top-down Visual Features for Saliency Estimation[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York:IEEE Press,2012:438-445.
[46] SHI Z,HOSPEDALES T,XIANG T. Bayesian Joint Modeling for Object Localization in Weakly Labeled Images [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015,37(10):1959-1972.

万方数据

[47] ITTI L，KOCH C，NIEBUR E. A Model of Saliency-based Visual Attention for Rapid Analysis [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2002，20(11)：1254-1259.

[48] TSAURO G，TOURCTZKY D，LN T，et al. Advances in Neural Information Processing Systems [J]. Morgan Kaufmann Publishers，2009，2(4)：368-374.

[49] HOU X，ZHANG L. Saliency Detection：A Spectral Residual Approach[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2007：1-8.

[50] GOFERMAN S，ZELNIKMANOR L，TAL A. Context-Aware Saliency Detection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2012，34(10)：1915-1926.

[51] WEI Y，ZHOU Y，LI H. Spectral-Spatial Response for Hyperspectral Image Classification [J]. Remote Sensing，2017，9(3)：203-234.

[52] ANDREWS S，TSOCHANTARIDIS I，HOFMANN T. Support Vector Machines for Multiple-instance Learning [J]. Advances in Neural Information Processing Systems，2003，15(2)：561-568.

[53] CHEN Y，BI J，WANG J. MILES：Multiple-instance Learning via Embedded Instance Selection [J]. IEEE Transactions on Pattern Anlaysis and Machine Intelligence，2006，28(12)：1931-1947.

[54] LI Y，KWOK J，TSANG I，et al. A Convex Method for Locating Regions of Interest with Multi-instance Learning[C]// Proceedings of European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin：Springer，2009：15-30.

[55] RUSSAKOVSKY O，LIN Y，YU K，et al. Object-centric Spatial Pooling for Image Classification[C]// Proceedings of European Conference on Computer Vision. Berlin：Springer，2012：1-15.

[56] ZHOU B，KHOSLA A，LAPEDRIZA A，et al. Learning Deep Features for Discriminative Localization [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2015：2921-2929.

[57] OQUAB M，BOTTOU L，LAPTEV I，et al. Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2014：1717-1724.

[58] OQUAB M，BOTTOU L，LAPTEV I，et al. Is Object Localization for Free? Weakly-supervised Learning with Convolutional Neural Networks [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2015：685-694.

[59] OQUAB M，BOTTOU L，LAPTEV I，et al. Weakly supervised object recognition with Convolutional Neural Networks[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2014：1-38.

[60] HONG S，KWAK S，HAN B. Weakly Supervised Learning with Deep Convolutional Neural Networks for Semantic Segmentation：Understanding Semantic Layout of Images with Minimum Human Supervision [J]. IEEE Signal Processing Magazine，2017，34(6)：39-49.

[61] BILEN H，VEDALDI A. Weakly Supervised Deep Detection Networks[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2016：2846-2854.

[62] XU J，LEI B，KIROS R，et al. Show，Attend and Tell：Neural Image Caption Generation with Visual Attention[C]// Proceedings of IEEE Conference on Machine Learning. Lille：JMLR org，2015：2048-2057.

[63] ZHANG W，TAN X Y. Weakly-Supervised multi-label-Classification-Based attention mechanism [J]. Journey of Data Acquisition and Processing，2018，33(5)：801-808. (in Chinese)
张文，谭晓阳. 基于 Attention 的弱监督多标号图像分类[J]. 数据采集与处理，2018，33(5)：801-808.

[64] ZHOU M F，WANG X L. Object detection models of remote sensing images using deep neural networks with weakly supervised training methods [J]. Science China，2018，48(8)：1022-1034. (in Chinese)
周明非，汪西莉. 弱监督深层神经网络遥感图像目标检测模型[J]. 中国科学，2018，48(8)：1022-1034.

[65] TZENG E，HOFFMAN J，SAENKO K，et al. Adversarial Discriminative Domain Adaptation[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2017：2962-2971.

[66] INOUE N，FURUTA R，YAMASAKI T，et al. Cross-Domain Weakly-Supervised Object Detection through Progressive Domain Adaptation[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City：IEEE Press，2018：5001-5009.

[67] LI D，HUANG J，LI Y，et al. Weakly Supervised Object Localization with Progressive Domain adaptation[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2016：3512-3620.

[68] TZENG E，HOFFMAN J，SAENKO K，et al. Adversarial Discriminative Domain Adaptation[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2017：2962-2971.

[69] KANTOROV V，OQUAB M，CHO M，et al. ContextLocNet：Context-aware Deep Network Models for Weakly Supervised Localization[C]// Proceedings of European Conference on Computer Vision. Berlin：Springer，2016：350-365.

[70] JIE Z，WEI Y，JIN X，et al. Deep Self-taught Learning for Weakly Supervised Object Localization [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2017：4294-4302.

万方数据