中图分类号：中图法分类号:TP183　**文献标识码：**A　**文章编号：**1006-2022 -
论文引用格式：（论文引用格式：）[DOI:]

图像级标记弱监督目标检测综述

陈震元[1，2，3]，王振东[1，2，3]，宫辰[1，2，3*]

1. 南京理工大学计算机科学与工程学院, 江苏南京 210094 ; 2. 高维信息智能感知与系统教育部重点实验室, 江苏南京 210094;
3. 江苏省社会安全图像与视频理解重点实验室, 江苏南京 210094

摘　要：目标检测是计算机视觉领域的基本任务之一，根据标签信息的不同，其可分为全监督目标检测、半监督目标检测、弱监督目标检测等。弱监督目标检测旨在仅利用图像级别的类别标记信息训练检测器，从而完成对测试图像中所有目标物体的定位和分类。因能够显著降低数据标记成本，弱监督目标检测愈发受到关注且已取得了令人瞩目的进展。本文由弱监督目标检测的研究意义引入，首先介绍了弱监督目标检测的标签设置及问题定义、基于多示例学习的基础框架和面临的局部主导、实例歧义和计算消耗这三大难题，接着按核心网络架构将该领域的典型算法归纳为三大类，分别是：基于优化候选框生成的算法、结合图像分割的算法和基于自训练的算法，并分别阐述各类算法的核心贡献。进一步，本文通过实验在多种评估指标上对比了各类弱监督目标检测算法的检测效果。在 VOC2007 数据集中，平均精度均值(mean Average Precision，mAP)最高的方法为 54.9%的 MIST 算法，正确定位率(correct localization，CorLoc)最高的方法为 71.1%的 SLV 算法。在 VOC2012 数据集中，mAP 最高的方法为 53.9%的 NDI-WSOD 算法，CorLor 最高的方法为 73.3%的 P-MIDN 算法。在 COCO 数据集中，在交并比(intersection over union, IoU)阈值为 50%时验证集上的平均精度 $ValAP_{50}$ 最高的方法为 27.4%的 P-MIDN。最后探讨了弱监督目标检测未来的研究发展方向。本文所总结的弱监督目标检测算框架，对后续研究人员的网络设计、模型探究和优化方向等都具有一定的参考价值。

关键词：弱监督目标检测；弱监督语义分割；候选框生成器；自训练；综述

Image-level Labeled Weakly Supervised Object Detection: A Survey

Zhenyuan Chen[1，2，3], Zhendong Wang[1，2，3], Chen Gong[1，2，3*]

1. School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu, 210094，China;
2．Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Nanjing, Jiangsu, 210094, China;
3．Jiangsu Key Lab of Image and Video Understanding for Social Security, Nanjing, Jiangsu, 210094，China

Abstract: Object detection is a fundamental problem in computer vision and image processing, which can be divided into fully supervised, semi-supervised, and weakly supervised object detection from the perspective of supervision. Recently, object detection has been playing an important role in various areas, showing great application value. Precise object detection depends on the accurate region or instance-level image labeling during detector training. However, the complexity of the background and the diversity of objects in real scenes make accurate image labeling extremely time-consuming and laborious. Specifically, traditional fully supervised object detection algorithms need to mark the position and category of each object in the image manually with a minimum rectangular box, increasing the cost of acquiring a training label. By contrast, weakly-supervised object detection (WSOD) algorithms only require the category labels of the whole image for training. As a result, a large number of training samples can be easily obtained by searching the category labels on some image websites. Due to the ability to significantly reduce the labor cost of labeling, WSOD has received increasing attention and achieved encouraging

收稿日期：；　修回日期：
*通讯作者：宫辰 chen.gong@njust.edu.cn
基金项目：国家自然科学基金资助(项目编号：61973162)
Supported by: National Natural Science Foundation of P. R. China (61973162)

progress. Therefore, researchers began to focus on WSOD algorithms based on image-level coarse labeling, which are less dependent on supervised information, compared with supervised, WSOD aims to localize and classify objects in an image by only using image-level category annotations. Starting from the research significance of WSOD, this paper firstly introduces the definition, basic framework, and main challenges of WSOD: 1) WSOD is performed in the training phase and test phase with standard detectors. The whole problem of WSOD can be understood as learning a mapping relationship from several candidate boxes contained in an image to image category markers. 2) It is worth noting that the problem setup of WSOD is consistent with that of multi-example learning in weakly supervised learning, so WSOD can be treated as a multi-example learning problem by taking each candidate box as an example and the image that contains all the candidate boxes as a "package" itself. For each category, if the image contains at least one target object of this category, the image is a positive packet, otherwise, it is a negative packet. Therefore, detector parameters can be learned based on candidate boxes in images. If an image is predicted to be a positive packet of a certain class, it indicates that the image contains the target of that class, and thus the target can be identified using a rectangular candidate box. 3) WSOD faces three major problems: local dominance problem, instance ambiguity problem, and conspicuous memory consumption problem. After that, advanced WSOD algorithms are classified into three categories according to the network architectures, i.e., optimization candidate box generation-based algorithms, segmentation-based algorithms, and self-training-based algorithms. Among them, the core of optimized candidate box generation-based algorithms is the improved candidate box generator in the basic framework. The core of segmentation-based algorithms and self-training-based algorithms is to improve the detector in the basic framework. The difference is that the former algorithms aim to add a segmentation branch and guide detection through segmentation, while the latter algorithms aim to optimize the detection network. Further, the detection results of various WSOD algorithms are compared under several evaluation metrics through extensive experiments. This paper selects and compares the current mainstream weakly supervised object detection algorithms on PASCAL VOC2007 and VOC2012 datasets. In order to ensure the fairness of comparison, all algorithms use the VGG16 network that has been pre-trained on the ILSVRC dataset as the backbone for feature extraction, and only evaluate the performance of the model itself, without considering the effect of fully supervised models such as Fast R-CNN. In the mean Average Precision (mAP) comparison on the VOC2007 dataset, MIST is considered the best, with the single model getting 54.9% mAP. To now, the mAP of the existing advanced WSOD algorithms is between 50% and 60%. Compared with the OICR algorithm, which is often used as the baseline method, the mAP of MIST has been improved by less than 15%, indicating that there is still a large room for improvement in this field. In the comparison of mAP and Correct Localization (CorLoc) on the VOC2012 dataset, NDI-WSOD achieves better performance, reaching 53.9%, which is 16% higher than OICR. The best algorithm for the CorLoc dataset is P-MIDN, which reaches 73.3% in terms of CorLoc, and is 11.2% higher than OICR. In addition, various algorithms are adopted for comparison on MS COCO datasets. The algorithm with the highest Val-AP50 is still P-MIDN, which achieved 27.4%. MIST combines optimized pseudo notation generation, regularization technique, and bounding box regression in the self-training process, and thus it can continue to be superior to the competitors on different datasets. Even though the research of WSOD algorithm based on image level labeling has made a great breakthrough due to the vigorous development of deep learning, WSOD still faces many challenges, and there is still a certain gap between it and fully supervised object detection. Finally, some valuable future research directions in this field have been discussed: 1) Generating fewer candidate boxes with higher quality, 2) Designing a more reasonable and efficient cooperative framework for detection and segmentation, 3) Designing a more reasonable strategy or digging out more and better positive samples through the network itself, 4) Designing lightweight network models that can be applied to mobile terminals.

Key words: weakly-supervised object detection; weakly-supervised semantic segmentation; proposal generator; self-training; survey

0 引言

目标检测是计算机视觉领域的基本任务之一，其旨在使用矩形框定位图像中的每个目标物体并预测其类别。目标检测已在生活的各个领域发挥着重要作用，体现出巨大的应用价值(曹家乐等，2022)。

例如，在自动驾驶中，需要实时对周围环境进行分析，检测出可能存在的障碍物从而辅助无人系统及时做出反应（徐歆恺等，2021）；在遥感图像识别中，需要在高分辨率图像上检测出目标（如道路、植被、水体等）分布，从而提供更准确的地理位置信息（赵文清等，2021）。随着卷积神经网络的高速发展，目标检测在一些实际应用中已能达到较高的精度。然而，目标检测的高精度依赖于检测器训练时精确的区域或实例级别的图像标记，但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精确标注极为费时费力。因此，研究人员开始将目光转移到对于监督信息依赖程度较低的、基于图像级别粗标记的弱监督目标检测算法之上。

弱监督目标检测旨在降低对标记的要求，从而有利于更便捷地获取大量已标记训练样本，使模型达到接近于全监督目标检测的效果。具体地，传统的全监督目标检测算法需要人工用最小矩形框标记出图像中各物体的位置及其类别，因此训练样本的获取代价较高；而弱监督目标检测算法只需要整体图像的类别标记即可进行训练，所以通过一些图像检索网站上的类别标签，就可以轻松获取大量训练样本。因此，弱监督目标检测算法具有较高的研究意义和应用价值，对该领域的进展进行归纳和综述也有很大的必要性。

然而，现有弱监督目标检测相关综述仍存在一些不足之处。比如，杨辉等人(2021)按照不同的特征处理方法对典型算法进行分类，该分类依据的边界较为模糊，且不能直观体现弱监督目标检测算法的特点。周小龙等人(2019)的综述发表的时间较早，因此没有囊括近几年的新进展。近年来Shao等人(2021)和Zhang等人(2022)都是将目标检测与目标定位相结合，统一描述两者的发展历程，没有细致的区分深度弱监督目标检测的方法类别。任冬伟等人(2022)综合介绍了弱监督视觉领域的研究进展，对弱监督目标检测的介绍还不够细致。针对上述问题，本文首次根据核心网络架构对弱监督目标检测领域的经典及最新算法进行了全面且清晰的分类归纳与对比分析，并提出多个有价值的未来研究方向。

全文其余部分安排如下。第一节将介绍弱监督目标检测的问题定义、基础框架和面临的主要难题；第二节按核心网络架构将现有典型算法分为三大类并分别阐述各类算法的核心贡献；第三节通过实验对比了各类主流算法的检测效果；最后在第四节中，本文简要探讨了弱监督目标检测领域未来的研究方向。

1 弱监督目标检测简介

本节将从弱监督目标检测的问题定义出发，介绍基于多示例学习的通用基础框架(Bilen和Vedaldi，2016)，并阐述该领域所面临的三大主要难题。

1.1 问题定义

弱监督目标检测训练阶段和测试阶段的示意图见图1。其中，训练阶段的输入是训练图像及其类别标记，输出是训练好的目标检测器。测试阶段的输入是测试图像，输出是在该图像中的目标检测结果。训练阶段中，由于目标检测需要使用矩形框框出图像中每个目标物体的位置，因此一般需要先在输入图像上生成大量目标候选框，然后对目标候选框提取特征并预测其类别，最后将预测结果与输入的图像类别标记计算损失并以此更新模型参数。所以，整个弱监督目标检测问题可理解为学习一个从图像包含的若干候选框到图像类别标记的映射关系。

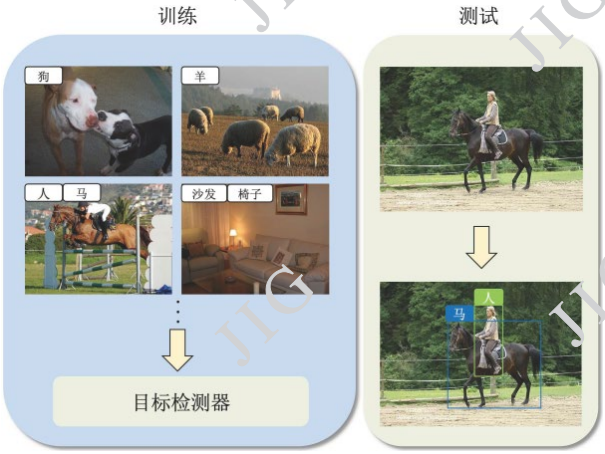


图 1 弱监督目标检测训练和测试示意图

Fig.1 Illustration of training and test phases in weakly-supervised object detection

1.2 基础框架

弱监督目标检测所需解决的问题与弱监督学习中的多示例学习(Dietterich 等，1997)研究目标相吻合，因此通常将弱监督目标检测视为多示例学习问题来处理。具体地，将每个候选框看作一个示例，将包含所有候选框的图像本身看作一个“包”。对于每个类别，图像中若含有至少一个该类的目标物体，则该图像为一个正包，否则为一个负包。因此，可基于图像中的候选框进行检测器参数学习。如果某张图像被预测为某类的一个正包，则表明了该图像

中包含该类目标，从而可以使用矩形候选框标识出该目标。

Bilen和Vedaldi(2016)首次提出基于多示例学习的弱监督目标检测框架。该框架的核心贡献是解决了将实例级别的候选框得分映射到图像级别的类别标记这一重要难题。具体地，该框架将经过空间金字塔池化(Spatial Pyramid Pooling, He等，2015)之后的候选框特征矩阵输入一个识别分支和一个检测分支。进而，在识别分支中将候选框特征矩阵使用Softmax操作映射到类别维度，在检测分支中将候选框特征矩阵使用Softmax操作映射到候选框维度。最后，将得到的两个矩阵按位相乘并计算所有候选框关于每个类别的得分之和，从而得到维度为类别数的向量，完成从候选框得分到类别标记的映射。完整框架包含以下三个主要部分：

1) 候选框生成器。该部分一般采用Selective search (Uijlings等，2013)或Edge boxes(Zitnick等，2014) 算法在输入图像上生成大量的目标候选框；

2) 特征提取。该部分一般采用VggNet(Simonyan和Zisserman，2015)对输入图像进行特征提取，再通过空间金字塔池化(He等，2015)或感兴趣区域池化(Region-of-Interest Pooling，Girshick等，2015)生成固定尺寸的候选框特征矩阵。

3) 检测器。如前文所述，将候选框特征映射到图像类别标记，计算多示例学习损失函数，完成对图像中目标物体的定位和分类。

尽管上述弱监督目标检测框架有效且易于实现，其检测精度较之于全监督目标检测算法仍有较大的提升空间，二者之间的差距主要归结于下述1.3节中所介绍的三大难题。

1.3 主要难题

尽管弱监督目标检测算法已经取得了较大的进步，但是与目前效果最好的全监督目标检测算法相比还是有较大差距。在公开数据集VOC2007上，目前效果最好的弱监督目标检测算法的精度达到58.1%，然而全监督目标检测算法能达到89.3%。造成如此之大的差距主要归结于弱监督目标检测所面临的三大难题：

1) 局部主导问题。模型更关注图像中辨识度较高的部分，而不关注整体。如图2的第三行所示，以第二幅图像为例，模型只能检测出人和马的头部，而无法检测出人和马的身体部分，原因在于头部往往更有辨识度。

2) 实例歧义问题。对于图像中含有多个目标物体的情形，算法容易遗漏目标物体，且难以区分同类别的不同实例。遗漏物体实例的情况如图2的第一行所示，以第一幅图像为例，该图像中包含数量较多的天鹅，但是只有个别天鹅能够被检测出来。难以区分物体实例的情况如图2的第二行所示，以第三幅图像为例，该图像中存在多只羊相互遮挡。对于这种遮挡的情形模型容易将相互挨着的多个物体实例检测为一个物体实例。

3) 显存消耗问题。图像级别的标记信息决定了弱监督目标检测必须生成并处理大量的候选框，因此模型训练对于显存的消耗程度较大，导致训练和预测速度较慢。同时，由于显存消耗大，用于提取特征的主干网络往往只能采用规模较小的VggNet(Simonyan 和 Zisserman，2015)，而难以采用ResNet(He等，2016)等更深、更先进的复杂网络。

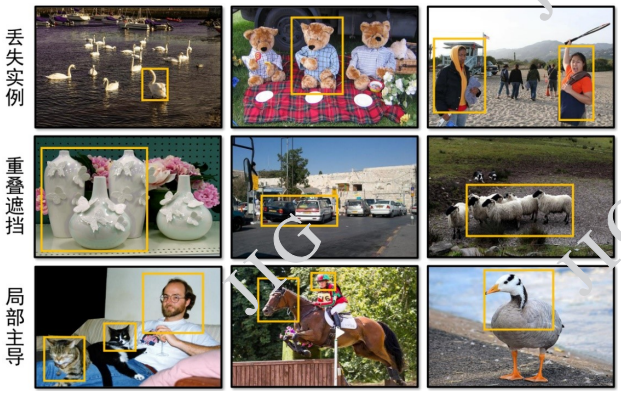


图 2 实例歧义问题和局部主导问题示意图

Fig.2 Illustration of instance ambiguity problem and local dominance problem

为了解决三大难题并进一步提高检测精度，研究人员以1.2节中所介绍的弱监督目标检测框架为基础，从该框架的各个部分切入开展了大量的研究工作。

表 1 弱监督目标检测算法优缺点对比

Table 1 Comparison of weakly-supervised object detection algorithms

算法类别	代表算法	方法特点	优点	缺点
基于优化候选框生成的算法	PG-PS(Cheng等，2020)	将选择性搜索与相似性区域生成结合，生成更多高质量的候选框	能挖掘出更多高质量的正样本和困难负样本	候选框数量过多的问题仍然存在，训练与推理速度仍然较慢

算法类别		代表算法	方法特点	优点	缺点
结合分割的算法	利用分割提供先验知识的算法	P-MIDN(Xu等，2021)	借助弱监督语义分割来保持对候选框对抗擦除后的多示例学习约束	1. 利用分割结果为检测任务提供额外的指导； 2. 检测和分割任务共享主干网络，通过协作互利帮助彼此跳出局部极小	1. 检测任务生成的检测热力图不足以作为分割标记，导致分割任务提供的指导有限； 2.模型复杂度较高，训练收敛困难且耗时严重
	检测和分割相互协作的算法	SDCN(Li 等，2019)	弱监督目标检测与弱监督语义分割相互协作，彼此学习互补知识		
基于自训练的算法	优化伪标记生成的算法	WSOD2(Zeng等，2019)	结合自顶向下的候选框得分和自底向上的似物性评分来赋予候选框伪标记	1. 通过自训练的方式不断细化分类和定位结果； 2. 伪标记生成过程充分考虑了空间上下文信息； 3. 通过添加正则化约束使得模型关注目标整体	1. 伪标记生成结果存在大量遗漏的正样本和错误标记的困难负样本，因此仍然具有较大的优化空间； 2. 模型初始化对于检测结果的影响较大，模型易于陷入局部极小
	结合正则化技术的算法	D-MIL(Gao等，2022)	通过构建综合注意力图对多个特征图施加正则化约束；同时加入了指导机制		
	结合边界框回归的算法	MIST(Ren等，2020)	在自训练过程中结合边界框回归；同时也加入了正则化技术		

2 弱监督目标检测算法

本节按核心网络架构将现有弱监督目标检测典型算法分为基于优化候选框生成的算法、结合分割的算法和基于自训练的算法。其中，基于优化候选框生成的算法的核心在于改进1.2节所介绍的基础框架中的候选框生成器。结合分割的算法和基于自训练的算法的核心皆在于改进基础框架中的检测器，区别在于前者旨在添加一个分割分支并通过分割指导检测，而后者旨在优化检测网络本身。特别地，基础框架的第二部分，即特征提取部分，由于都是采用现有的主干网络(Simonyan和Zisserman，2015；He等，2016)，因此不属于核心网络架构上的创新。三类算法的优缺点对比如表1所示。

2.1 基于优化候选框生成的算法

大部分弱监督目标检测算法都是使用Selective search (Uijlings等，2013)或Edge boxes (Zitnick等，2014) 算法来生成目标候选框，通过在一幅图像上生成数以千计的候选框来确保召回率，然而其中绝大多数候选框都属于负例，因此十分影响检测效果。同时，大量候选框的处理严重消耗显存，不仅难以采用ResNet (He等，2016) 等更先进的复杂网络提取特征，还会导致训练和预测速度低下。全监督目标检测中一般使用区域建议网络(Region Proposal Network)代替传统方法 (Ren等，2015)，该网络通过最小化一个前背景二分类损失和一个边界框回归损失对初始生成的候选框进行筛选和优化，从而将原本数以千计的候选框减少到数十个，大幅度提高了算法效率。然而，该方法需要借助实例级别的标记，因此无法应用于弱监督目标检测任务。针对此问题，一些学者提出了适用于弱监督目标检测的候选框生成的优化算法。Zhu等人(2017)提出了一个软建议网络(Soft Proposal Network，SPN)，首次将候选框生成集成在一个端到端的卷积神经网络里。作者定义了一个软建议(Soft Proposal)模块，可以插入到卷积神经网络的任意一层，并且额外时间消耗几乎可以忽略不计。借助该模块，模型可以在迭代中不断优化候选区域，然后再将其映射回特征图上，最后实现网络参数的整体优化。Tang等人(2018)提出了一个基于弱监督的区域建议网络(Weakly Supervised Region Proposal Network,WSRPN)，该网络由三个阶段组成，第一个阶段利用卷积神经网络的底层语义信息来评估滑动窗口的似物性分数(Objectness Score)，第二个阶段通过一个基于区域的卷积神经网络分类器来优化第一阶段的候选框，最后一个阶段完成目标检测。Cheng等人(2020)提出了一种高质量候选框的生成算法(Proposal Generation and Proposal Selection, PG-PS)，作者将选择性搜索 (Selective Search, Bilen和Vedaldi，2016)与基于梯度的类激活图(Gradient-based Class Activation Map，Selvaraju等，2017) 相结合，从而

生成比基于贪婪搜索的方法更多高交并比的候选框。针对候选框筛选，该方法对于每一个目标类别，在选取尽可能多的正样本的同时，只选取类别明确的困难负样本，并通过上调它们的权重，使模型在训练中关注更具辨识度的负例候选框，从而提高检测精度。周明非和汪西莉(2018)提出一种候选框融合算法，合并重叠候选框的同时调整候选框的位置，以此优化候选框。Jia等人(2021)提出了一种新颖的两阶段框架，其包含一个候选框评分模块(Boxes Grading Module)和一个信息增强模块(Informative Boosting Module)。具体地，候选框评分模块通过训练一个弱监督目标检测模型来生成候选框并对其进行筛选和评分；信息增强模块利用候选框评分模块生成的定位监督信息训练增强的候选框生成器和检测器，从而进一步提升检测效果。区别于前人基于多示例学习的算法框架，Song等人(2021)提出了一种基于分组标签的上下文实例特征梯度和掩码预测的方法(Weakly Supervised Group Mask Network, WSGMN)，利用这些掩码动态的选择最有价值的实例特征信息来识别特定的对象。图3为弱监督目标检测基础范式，通过似物性分数、选择性搜索、基于梯度的类激活图选择、掩码预测选择等方法，以此达到优化候选框生成的目的。

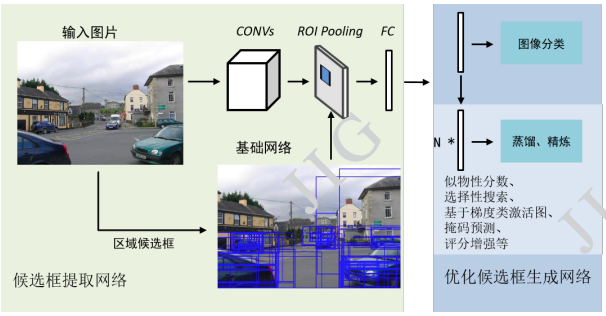


图 3 弱监督目标检测基础范式

Fig.3 Illustration of the basic framework of weakly supervised object detection

2.2 结合分割的算法

如下图4所示，结合分割的弱监督目标检测算法的特点是在主干网络的基础上添加一个分割分支，希望借助分割结果来指导检测任务。结合分割做弱监督目标检测包含单向和双向两种策略，单向策略指仅利用分割给弱监督目标检测提供先验知识；双向策略是让检测和分割相互协作、共同进步。这里的分割指弱监督语义分割，分割所需的伪标记在两种策略中由不同方式生成。

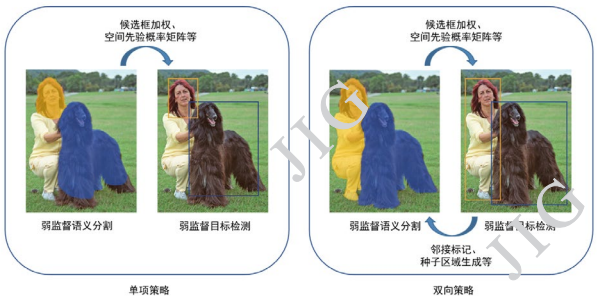


图 4 结合分割算法的弱监督目标检测范式

Fig.4 Illustration of weakly supervised object detection paradigm combined with segmentation algorithm

2.2.1 利用分割提供先验知识的算法

该类算法由于单方面通过语义分割来指导目标检测，因此分割所需的伪标记是由在弱监督语义分割中被普遍用于生成分割标记的类激活图(Class Activation Map, Zhou等，2016)提供的。Wei等人(2018)提出了一个利用周围语境分割上下文的紧贴框挖掘算法(Tight Box Mining with Surrounding Segmentation Context, TS2C)，该算法分为三部分：1）训练一个分类网络生成类激活图；2）将类激活图作为语义分割伪标记来训练分割网络，生成分割置信图；3）利用分割置信图来挖掘更紧贴目标物体的候选框，从而提升检测网络的效果。Gao等人(2019)提出了一个利用一对多示例检测网(Coupled Multiple Instance Detection Network, C-MIDN)来对候选框进行对抗擦除的方法，其同时借助弱监督分割结果来保持对候选框对抗擦除后的多示例学习约束，最后通过组合两个多示例检测网络的结果，有效缓解了1.3节中所介绍的局部主导问题。Xu等人(2021)在MIDN的基础上提出了一种多尺度空间金字塔融合的方法(Pyramidal Multiple Instance Detection Network, P-MIDN)，对不同尺度的候选框检测结果进行融合，生成更高质量、更全面的伪标签。

2.2.2 检测和分割相互协作的算法

该类算法旨在将弱监督目标检测和弱监督语义分割结合到一个多任务学习框架。具体而言，检测分支和分割分支相互提供指导，最终在两个任务上同时达到更好的效果。由于检测和分割之间的影响是双向的，因此分割所需的伪标记是由检测分支生成的检测热力图(Detection Heat map)提供的。Shen等人(2019)提出了一种结合弱监督目标检测和弱监督语义分割的多任务学习框架(Weakly Supervised Join Detection and Segmentation, WS-

JDS)。作者发现检测任务能发现更多的目标物体，而分割任务能挖掘出更完整的目标物体。为了充分利用这两种任务学习到的互补知识，作者提出了一个循环引导学习(Cyclic Guidance Learning) 框架，其中检测分支为分割分支提供较好的像素种子，分割分支学习到的分割图帮助检测分支跳出局部极小值。Li等人(2019)提出了一种分割检测协作网络(Segmentation Detection Collaboration Network, SDCN)。在该网络中，检测分支生成检测热力图为分割分支提供实例级别的监督信息，分割分支生成分割图反过来为检测分支提供空间先验概率矩阵，以指导候选框筛选。最终，检测分支和分割分支彼此紧密地相互作用并形成动态协作循环，从而相辅相成获得更好的效果。

2.3 基于自训练的算法

基于自训练的弱监督目标检测算法旨在优化检测网络本身。在1.2节所介绍的弱监督目标检测基础框架中，检测器负责将实例级别的候选框特征矩阵映射到图像级别的类别标记，所以实际上只是一个图像级别的分类器。然而图像级别的分类器难以学习到准确的实例级别的候选框得分，为此研究人员考虑利用半监督学习中的自训练(Self-training)思想来解决该问题。具体地，如下图5所示，将基础框架中的图像级分类器作为初始分类器，利用其输出的候选框得分为每个候选框生成伪标记并训练实例级分类器，最后重复该过程训练多个实例级分类器，通过这种知识蒸馏(Knowledge Distillation)的方式不断提炼出更准确的候选框得分。

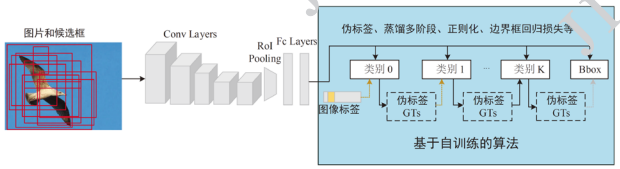


图 5 结合自训练的弱监督目标检测范式

Fig.5 Illustration of weakly supervised object detection paradigm combined with self-training

基于自训练的算法是弱监督目标检测领域的研究热点，按其优化检测效果的技术特点又可进一步分为优化伪标记生成的算法、结合正则化技术的算法和结合边界框回归的算法。

2.3.1 优化伪标记生成的算法

为每个候选框生成更加准确的类别伪标记可以减轻负样本过多的影响，有效缓解1.3节所介绍的局部主导和实例歧义问题，因此该步骤是自训练过程中尤为重要的一步。Tang等人(2017)提出了一个在线实例分类器优化算法(Online Instance Classifier Refinement，OICR)。该算法首次在基础框架之上添加了k个串行的实例分类器调整模块，每个实例分类器所需的伪标记由它的前一个模块提供，其中候选框伪标记按以下规则生成：为每个类别得分最高的候选框及与其交并比高于某一阈值的候选框赋予相应的类别标记，其余候选框标记为背景。Tang等人(2018)在OICR的基础上提出了一种候选框聚类学习算法(Proposal Cluster Learning, PCL)，该算法将空间上毗邻且与同一物体有关联的候选框划分到同一个簇，根据聚类结果给整个簇打上类别伪标记。这种将每一个簇作为一个小的多示例学习中的“包”的方式比直接给候选框打伪标记所产生的歧义更少。Kosugi等人(2019)提出了一种新的候选框标记算法，该算法利用上下文分类损失来找到包含更完整物体的候选框并赋予其正标记，同时对负标记施加额外的空间约束。Kosugi等人(2019)提出了一种目标物体挖掘算法(Object Instance Mining)，该算法通过对候选框建立空间图和外观图来挖掘图像中所有可能的目标物体，并设计了一个目标物体权重重调损失函数来平衡置信度最高的候选框和区分度较低的候选框的权重。Zeng等人(2019)提出了一种结合自底向上和自顶向下的似物性蒸馏算法(Objectness Distillation for Weakly Supervised Object Detection, WSOD2)。该算法首先通过图像级分类器计算每个候选框的得分，再利用每个候选框区域的低级特征来计算其似物性得分，最后将两个得分矩阵点乘得到最终的候选框得分矩阵，并据此赋予候选框伪标记。Ren等人(2020)提出了一种新颖的多示例自训练算法(Multiple Instance Self Training，MIST)，该算法中的候选框伪标记生成过程同时考虑了候选框的得分、上下文以及前人没有考虑到的空间多样性约束。Zhang等人(2020)提出的弱监督学习框架包含了对视觉表象的认知过程(Proposal and Semantic Level Relationships, PSLR)，和对提议层和语义层关系的推理过程，从而形成了新的深度多实例推理框架。具体而言，该框架基于传统的基于cnn的网络架构，增加了两个基于图卷积网络的推理模型，在一个端到端网络训练过程中实现目标位置推理和多标签推理。Wang等人(2022)提出了一种基于负确定性信息(Negative Deterministic Information，NDI-WSOD)的WSOD改进方法，该方法包含NOI收集和开发两个阶段，在收集阶段，设计了几个流程来在线识别和提取负面实例中的NDI。在开发阶段，

利用抽取的NDI构建了一种新的消极对比学习机制和消极引导实例选择策略，分别处理部分支配和缺失实例的问题。上述算法都在一定程度上缓解了局部主导问题以及更加困难的实例歧义问题。

2.3.2 结合正则化技术的算法

由于缺乏实例级别的监督信息，弱监督目标检测模型在训练时容易陷入局部极小，导致模型只关注物体辨识度较高的部分。为此，可利用正则化技术为模型引入一些额外的约束，使得模型更加平等地对待每个区域，从而缓解局部主导问题。Ren等人(2020)在其提出的多示例自训练算法MIST的基础上，进一步提出了一种参数化且可微分的特征空间随机失活模块(Concrete Drop Block)，该模块通过端到端的学习来鼓励模型考虑上下文而不是局限于辨识度较高的部分，从而实现目标的完整检测。Huang等人(2020)提出了一种综合注意力自提炼算法(Comprehensive Attention Self-Distillation)，该算法在OICR的基础上(Tang等，2020)，从网络的多个层和图像的多个变换特征图上分别获得注意力图(Attention Map)，并将这些注意力图整合为综合注意力图，然后利用这个综合注意力图对多个层和多个变换的特征图施加正则化约束，最终综合注意力图中的信息被提炼到各个特征图上，从而实现完整物体和小物体的检测。Gao等人(2022)将差异协同模块引入多示例学习(Discrepant Multiple Instance Learning, D-MIL)中，采用多个MIL学习器来寻找不同但互补的目标部分，并将其与协作模块融合，实现目标的精确定位。与此同时，D-MIL实施了一种新的教师-学生模式(Teacher-Student)，MIL学习者扮演教师，物体探测器扮演学生。多名教师提供丰富而互补的信息，这些信息被学生吸收并传递回来，以强化教师的绩效。

2.3.3 结合边界框回归的算法

边界框回归是全监督目标检测中用于细化定位结果的一种常用手段。虽然弱监督目标检测没有实例级别的检测框标记，但是仍可以通过生成实例级伪标记来进行边界框回归。Yang等人(2019)提出了一种注意力引导的结合边界框回归的目标检测算法(Towards Precise end-to-end Weakly Supervised Object Detection, TPWSD)，该算法包含一个OICR(Tang等，2020)分支和一个边界框回归分支。两个分支共享特征提取网络，其中OICR分支为边界框回归分支提供监督信息，同时特征提取网络通过添加一个注意力模块来为两个分支提供增强的特征图。Chen等人(2020)提出了一种空间似然投票算法(Spatial Likelihood Voting, SLV)，该算法在OICR的基础上添加了一个串行的空间似然投票模块，该模块以OICR的输出为输入，进行实例挑选、空间概率积累和高似然区域投票，并将投票结果用于后续的重分类和重定位（即边界框回归）。文献(Ren等，2020)在其提出的多示例自训练算法MIST中也结合了边界框回归。该算法中的实例级分类器不仅包含一个分类分支，还包含一个边界框回归分支，每个实例级分类器所需的伪标记由前一个模块提供。该文献还进一步通过实验验证了在自训练过程中结合边界框回归有助于提高检测的鲁棒性和泛化性。

3 实验分析

3.1 常用数据集

弱监督目标检测任务的常用数据集如下：

1) PASCAL VOC数据集(EveringhamM等，2010；EveringhamM等，2015)，总共分为4个大类（交通工具、房屋设施、动物、人），并可进一步分为20个小类。该数据集包含多个版本，其中VOC2007和VOC2012是弱监督目标检测领域最常用的数据集。VOC2007训练集包含2501个样本，验证集包含2510个样本，测试集包含4952个样本，共计9963个样本。VOC2012训练集包含5717个样本，验证集包含5823个样本，测试集包含11540个样本，共计23080个样本。

2) MSCOCO数据集(Lin等，2014)，总共包含80个类别。COCO数据集拥有33万个样本，有标记样本超过20万个。这个数据集因样本数量和类别数量较多，所以难度比VOC数据集要大。

3) ILSVRC数据集(Russakovsky等，2015)，包含用于目标检测任务的200个类别，涉及到大部分生活中会见到的物体。该数据集包含多个版本，其中最常用于目标检测的是ILSVRC2013，训练集包含12125个样本，验证集包含20121个样本，测试集包含40152个样本。该数据集的难度比VOC数据集和COCO数据集都要大。

3.2 评价指标

弱监督目标检测领域的常用评价指标如下：

1) 平均精度均值(mean Average Precision，简称mAP)。准确率和召回率计算公式如下：

$$PR=\frac{TP}{TP+FP} \tag{1}$$

$$RE=\frac{TP}{TP+FN} \tag{2}$$

式中，平均精度(Average Precision, AP)由准确率(Precision, PR)和召回率(Recall, RE)构成。常用

于图像分类和目标检测任务。使用True Positive (*TP*) 表示正例样本中预测正确的样本数量，False Positive (*FP*) 表示正例样本中预测错误的样本数量，False Negative (*FN*) 表示负例样本中预测错误的样本数量。

样本预测正确是指预测框与真实框的交并比(intersection over union，*IoU*)大于等于0.5。*IoU*的计算公式如下：

$$IoU\left(b,b^g\right)=\frac{AR\left(b\cap b^g\right)}{AR\left(b\cup b^g\right)} \tag{3}$$

式中，b 表示预测框，b^g 表示预测框所对应的真实框，*AR* 表示区域大小。平均精度 *AP* 的具体计算过程如下：设定一组阈值，如$[0,0.1,0.2,...,1]$，对于召回率大于每一个阈值分别得到一个对应的最大精确率，*AP* 就是这组精确率的平均值。最终，平均精度均值 *mAP* 就是关于所有类别的 *AP* 的均值。

2) 正确定位率(correct localization，简称*CorLoc*)。*CorLoc* 表示每个类别中至少有一个预测框与真实框的 *IoU* 大于等于50%的样本占所有样本的百分比。*CorLoc* 是在数据集上进行评估的重要指标。

3) top错误率(top error)。top error包含top-1分类错误率、top-5分类错误率、top-1定位错误率和top-5定位错误率。top-1分类错误率是指预测得分最高的候选框被错误分类的样本占所有样本的百分比，top-5分类错误率是指预测得分前5的候选框被错误分类（预测得分前5的候选框里至少有一个分类正确就算作正确）的样本占所有样本的百分比。定位错误率与分类错误率类似，不同点在于前者通过*IoU*来判断定位是否正确。

3.3 实验结果对比

本文选取了当前主流的弱监督目标检测算法，在PASCAL VOC2007和VOC2012数据集上进行了对比。为了确保对比的公平性，所有算法均采用在ILSVRC数据集上进行过预训练的VGG16网络作为用于提取特征的主干网络，且全部只考虑模型自身的效果，不考虑集成Fast R-CNN等全监督模型的效果(Girshick等，2015)。其中，WSRPN(Tang等，2018)、PG-PS(Cheng等，2020)、WSGMN(Song等，2021)属于基于优化候选框生成的算法；TS2C(Wei等，2018)、C-MIDN(Gao等，2019)、P-MIDN(Xu等，2021)属于结合分割的算法中的利用分割提供先验知识的算法，WS-JDS(Shen等，2019)、SDCN(Li等，2019)属于结合分割的算法中的检测和分割相互协作的算法；OICR(Tang等，2017)、PCL(Tang等，2018)、WSOD2(Zeng等，2019)、PSLR(Zhang等，2020)、NDI-WSOD(Wang等，2022)属于基于自训练的算法中的优化伪标记生成的算法，TPWSD(Yang等，2019)、SLV(Chen等，2020)、D-MIL(Gao等，2022)属于基于自训练的算法中的结合边界框回归的算法，MIST(Ren等，2020)涵盖了基于自训练的算法中的全部三种技术。

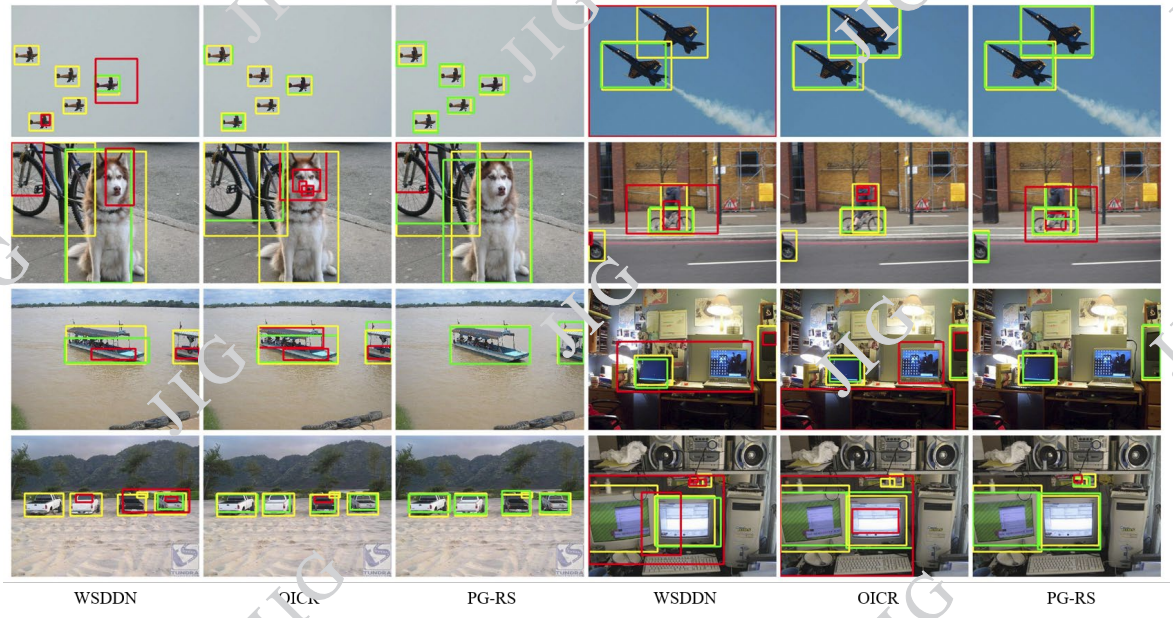


图 6 WSDDN、OICR 和 PG-RS 算法在 VOC2007 测试数据集的可视化结果对比图

Fig.6 Comparison of visualization results of WSDDN, OICR and PG-RS algorithms on the VOC 2007 test dataset

图6展示了WSDDN、OICR和PG-RS算法在VOC2007 测试数据集的可视化结果，黄色矩形表示地面真实边界框。成功检测(*IoU* >0.5)用绿色边框标示，失败检测(*IoU* <0.5)用红色边框标示。根据该图可以看出，PG-RS算法可以生成更紧密的边界框，实现更精确的定位，而其他两种方法总是不能生成过大的框或只包含对象的一部分。特别是当同一类别的多个物体出现在一幅图像中时，PG-RS算法可以用较大的*IoU*准确地检测到它们，但其他两种方法通常会有一些漏检。

表2展示了主流算法在VOC2007数据集上的*mAP*对比，其中效果最好的算法是MIST(Ren等，2020)，单模型*mAP*达到了54.9%。目前先进的弱监督目标检测算法*mAP*都在50%到60%之间，难以超过60%，与常被用作基线的OICR (Tang等，2017)算法相比提高了不到15%，可见该领域尚有较大的提升空间。观察各个类别的最高*AP*，不难发现在船、瓶子、椅子、人、植物这几类物体上，效果最好方法的*AP*依然难以超过40%，原因在于这几类物体存在更复杂的形变、遮挡等问题。同时，大部分类别的最高*AP*出现在基于自训练的算法中，说明自训练过程是弱监督目标检测缩小与全监督目标检测之间差距的重要环节。综合整个表1可知，三类算法各有所长，并无明显的优劣，因此本文归纳的三类算法都能有效缓解弱监督目标检测所面临的难题，提高检测精度。

表 2 主流算法在 VOC2007 数据集上的 mAP 对比

Table2 mAP comparison of popular algorithms on VOC2007 dataset

类别	优化候选框生成			结合分割				自训练							
	WSRPN	PG-PS	WSGMN	TS2C	C-MIDN	WS-JDS	SDCN	OICR	PCL	PSLR	WSOD2	D-MIL	TPWSD	SLV	MIST
aero	57.9	63.0	55.6	59.3	53.3	52.0	59.4	58.0	54.4	62.2	65.1	60.4	57.6	65.6	**68.8**
bike	70.5	64.4	60.3	57.5	71.5	64.5	71.3	62.4	69.0	61.1	64.8	71.3	70.8	71.4	**77.7**
bird	37.8	50.1	50.3	43.7	49.8	45.5	38.9	31.1	39.3	51.1	**57.2**	51.1	50.7	49.0	57.0
boat	5.7	27.5	37.3	27.3	26.1	26.7	32.2	19.4	19.2	33.8	**39.2**	25.4	28.3	37.1	27.7
bottle	21.0	17.1	24.5	13.5	20.3	27.9	21.5	13.0	15.7	18.0	24.3	23.8	27.2	24.6	**28.9**
bus	66.1	70.6	69.3	63.9	70.3	60.5	67.7	65.1	62.9	66.7	69.8	70.4	**72.5**	69.6	69.1
car	69.2	66.0	66.2	61.7	69.9	47.8	64.5	62.2	64.4	66.5	66.2	70.3	69.1	70.3	**74.5**
cat	59.4	71.1	63.3	59.9	68.3	59.7	68.9	28.4	30.0	65.0	61.0	**71.9**	65.0	70.6	67.0
chair	3.4	25.8	13.5	24.1	28.7	13.0	20.4	24.8	25.1	18.5	29.8	25.2	26.9	30.8	**32.1**
cow	57.1	55.9	60.5	46.9	65.3	50.4	49.2	44.7	52.5	59.4	64.6	63.4	64.5	63.1	**72.2**
table	**57.3**	43.2	41.2	36.7	45.1	46.4	47.6	30.6	44.4	44.8	42.5	42.6	47.4	36.0	48.1
dog	35.2	62.7	46.1	45.6	64.6	56.3	60.9	25.3	19.6	60.9	60.1	**67.1**	47.7	61.4	45.2
horse	64.2	65.9	49.2	39.9	58.0	49.6	55.9	37.8	39.3	65.6	**71.2**	57.7	53.5	65.3	54.4
mbike	68.6	64.1	63.5	62.6	71.2	60.7	67.4	65.5	67.7	66.9	70.7	70.1	66.9	68.4	**73.7**
person	32.8	10.2	21.3	10.3	20.0	25.4	31.2	15.7	17.8	24.7	21.9	15.5	13.7	12.4	**35.0**
plant	28.6	22.5	26.2	23.6	27.5	28.2	22.9	24.1	22.9	26.0	28.1	26.6	29.3	**29.9**	29.3
sheep	50.8	48.1	52.2	41.7	54.9	50.0	45.3	41.7	46.6	51.0	58.6	58.7	56.0	52.4	**64.1**
sofa	49.5	53.8	60.3	52.4	54.9	51.4	53.2	46.9	57.5	53.2	59.7	**63.3**	54.9	60.0	53.3
train	41.1	**72.2**	60.8	58.7	69.4	66.5	60.9	64.3	58.6	66.0	52.2	66.9	63.4	67.6	65.3
tv	30.0	67.4	61.3	56.6	63.5	29.7	64.4	62.6	63.0	62.2	64.8	**67.6**	65.2	64.5	65.2
mAP	45.3	51.1	49.5	44.3	52.6	45.6	50.2	41.2	43.5	51.2	53.6	53.5	51.5	53.5	**54.9**

(注)：黑色字体表示最优结果。

表3展示了主流算法在VOC2007数据集上的*CorLoc*指标对比，其中效果最好的算法是SLV(Chen等，2020)，其*CorLoc*达到了71.0%，与常被用作基线的OICR(Tang等，2017)算法相比提升了10.4%。各个算法之间的*CorLoc*差距并不明显，尤其是较为先进的算法，大多都在68%到71%之间。观察各个类别的最高*CorLoc*，不难发现上限较低的仍然是瓶子、椅子、人等类别，原因同样在于这

几类物体存在更复杂的形变、遮挡等问题

表 3 主流算法在 VOC2007 数据集上的 CorLoc 对比

Table3 CorLoc comparison of popular algorithms on VOC2007 dataset

类别	优化候选框生成				结合分割			自训练							
	WSRPN	PG-PS	WSGMN	TS2C	C-MIDN	WS-JDS	SDCN	OICR	PCL	PSLR	WSOD2	D-MIL	TPWSD	SLV	MIST
aero	77.5	85.4	82.9	84.2	-	82.9	85.0	81.7	79.6	86.3	**87.1**	81.3	80.0	84.6	-
bike	81.2	80.4	78.7	74.1	-	74.0	83.9	80.4	**85.5**	72.9	80.0	82.0	83.9	84.3	-
bird	55.3	69.1	73.5	61.3	-	73.4	58.9	48.7	62.2	71.2	**74.8**	72.7	74.2	73.3	-
boat	19.7	58.0	48.6	52.1	-	47.1	59.6	49.5	47.9	59.0	**60.1**	48.9	53.2	58.5	-
bottle	44.3	35.9	50.7	32.1	-	**60.9**	43.1	32.8	37.0	36.3	36.6	42.0	48.5	49.2	-
bus	80.2	82.7	**84.5**	76.7	-	80.4	79.7	81.7	83.8	80.2	79.2	80.2	82.7	80.2	-
car	86.6	86.7	82.3	82.9	-	77.5	85.2	85.4	83.4	84.4	83.8	86.1	86.2	**87.0**	-
cat	69.5	**82.6**	61.1	66.6	-	78.8	77.9	40.1	43.0	75.6	70.6	78.5	69.5	79.4	-
chair	10.1	45.5	35.0	42.3	-	18.6	31.3	40.6	38.3	30.8	43.5	43.9	39.3	**46.8**	-
cow	87.7	84.9	73.9	70.6	-	70.0	78.1	79.5	80.1	83.6	**88.4**	80.2	82.9	83.6	-
table	**68.4**	44.1	48.8	39.5	-	56.7	50.6	35.7	50.6	53.2	46.0	42.2	53.6	41.8	-
dog	52.1	**80.2**	61.0	57.0	-	67.0	75.6	33.7	30.9	75.1	74.7	76.5	61.4	79.3	-
horse	84.4	84.0	72.7	61.2	-	64.5	76.2	60.5	57.8	82.7	87.4	68.7	72.4	**88.8**	-
mbike	**91.6**	89.2	86.5	88.4	-	84.0	88.4	88.8	90.8	87.1	90.8	91.2	91.2	90.4	-
person	**57.4**	12.3	37.3	9.3	-	47.0	43.7	21.8	27.0	37.7	44.2	32.7	22.4	19.5	-
plant	**63.4**	55.7	53.6	54.6	-	50.1	56.4	57.9	58.2	54.6	52.4	56.0	57.5	59.7	-
sheep	77.3	79.4	82.1	72.2	-	71.9	73.2	76.3	75.3	74.2	81.4	81.4	**83.5**	79.4	-
sofa	58.1	63.4	65.5	60.0	-	57.6	62.6	59.9	68.5	59.1	61.8	**69.6**	64.8	67.7	-
train	57.0	82.1	75.9	65.0	-	**83.3**	77.2	75.3	75.7	79.8	67.7	78.7	75.7	82.9	-
tv	53.8	82.1	70.2	70.3	-	43.5	79.9	81.4	78.9	78.9	79.9	79.9	77.1	**83.2**	-
mean	63.8	69.2	66.2	61.0	68.7	64.5	68.6	60.6	62.7	68.1	69.5	68.7	68.0	**71.0**	68.8

(注)：“-”表示相关实验结果原文未提供，黑色字体表示最优结果。

表 4 主流算法在 VOC2012 数据集上 mAP、CorLoc 对比

Table4 mAP comparison and CorLoc comparison of popular algorithms on VOC2012 dataset

算法类别	算法	mAP	CorLoc
优化候选框生成	WSRPN	40.8	64.9
	PG-PS	48.3	68.7
	WSGMN	45.7	66.1
结合分割	TS2C	40.0	64.4
	C-MIDN	50.2	68.7
	P-MIDN	52.8	**73.3**
	WS-JDS	39.1	63.5
	SDCN	43.5	67.9
自训练	OICR	37.9	62.1
	PCL	40.6	63.2
	WSOD2	47.2	71.9
	D-MIL	49.6	70.1
	TPWSD	45.6	68.7
	SLV	49.2	69.2
	MIST	52.1	70.9
	PSLR	46.3	68.7
	NDI-WSOD	**53.9**	72.2

(注)：黑色字体表示最优结果。

表4展示了主流算法在VOC2012数据集上的*mAP*和*CorLoc*对比，其中*mAP*最高的算法是NDI-WSOD(Wang等，2022)，达到了53.9%，较之于OICR(Tang等，2017)提高了16%。*CorLoc*最高的算法是P-MIDN(Xu等，2021)，达到了73.3%，较之于OICR(Tang等，2017)提高了11.2%。由于VOC2012数据集较VOC2007数据集样本更多更复杂，因此在这个数据集上各方法的*mAP*普遍降低，但是在VOC2007数据集上检测精度较高的算法在VOC2012数据集上依然具有优势。

表 5 主流算法在 COCO 数据集上 ValAP、$ValAP_{50}$ 对比

Table5 Val-AP comparison and $ValAP_{50}$ comparison of popular algorithms on COCO dataset

算法类别	算法	ValAP	$ValAP_{50}$
优化候选框生成	PG-PS	-	20.7
结合分割	C-MIDN	9.6	21.4
	P-MIDN	**13.1**	**27.4**
	WS-JDS	10.5	20.3
自训练	PCL	8.5	19.4
	WSOD2	10.8	22.7
	D-MIL	11.3	24.7
	MIST	11.4	24.3
	PSLR	11.1	23.6
	NDI-WSOD	12.1	26.2

(注)：“-”表示相关实验结果原文未提供；黑色字体表示最优结果。

此外如表5所示，本文还选取了部分算法在MS COCO数据集上进行了对比。由于COCO数据集样本数量大、种类多，因此现有算法很难获得较高的检测精度。如表5所示，$ValAP_{50}$ 最高的算法是P-MIDN(Xu等，2021)，达到了27.4%。其中 $ValAP$ 表示验证集上的平均精度，$ValAP_{50}$ 表示在 IoU 阈值为50%时验证集上的平均精度。

4 未来研究方向

得益于深度学习的蓬勃发展，基于图像级别标记的弱监督目标检测算法研究取得了较大突破。然而弱监督目标检测仍然面临诸多难题，其与全监督目标检测相比还有一定的差距。本领域一些有价值的未来研究方向包括：

1) 现有算法大多采用Selective search (Uijlings等，2013)或Edge boxes(Zitnick等，2014)来生成初始候选框，然而这两种方法极为耗时且生成的绝大多数候选框属于负例。因此，如何生成数量更少、质量更高的候选框，是一个亟待解决的问题。

2) 由于检测热力图过于粗糙，不足以作为分割标记，所以现有的检测和分割相互协作的弱监督目标检测算法并不能很好地利用检测指导分割。因此，可以考虑设计更合理、更高效的检测分割协作机制。

3) 自训练过程中的伪标记生成是基于人工设计的策略实现的。尽管现有算法已借助多种手段来优化伪标记生成，但该步骤仍会遗漏大量正样本和错误标记大量负样本。因此，怎样设计更合理的策略或通过网络本身来挖掘出更多、更好的正样本，是一个值得深究的问题。

4) 现有的弱监督目标检测算法的网络模型复杂度较高。由于只有图像级别的监督信息，导致网络模型不得不通过增加复杂度来换取更高的精度，从而大大增加对硬件的需求。因此，设计轻量级的、能够应用于移动端的网络模型同样具有重要的研究价值。

5. 结语

基于图像级别标记的弱监督目标检测算法对于标记信息的要求较低，能够显著降低训练样本的获取代价，因此具有重要的研究意义。本文首先介绍了弱监督目标检测的问题定义、基础框架和面临的主要难题。然后按核心网络架构将现有典型算法归纳为基于优化候选框生成的算法、结合分割的算法和基于自训练的算法，并分析了各种算法的特点及其优缺点。进一步，在多个公共数据集和多种指标上对主流算法进行了效果验证和比较，得出结论：本文归纳的三大类算法均可在一定程度上缓解该领域所面临的主要难题并提高检测效果，其中目前效果最为显著是基于自训练的算法。最后，根据现有算法的不足，并以进一步解决主要难题为目标，提出了该领域的一些有价值的未来研究方向，供相关研究人员参考借鉴。

参考文献（References）

ArunA, Jawahar C V and Kumar M P.2019. Dissimilarity coefficient based weakly supervised object detection// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Long Beach, USA: IEEE, 9432 - 9441.[DOI: 10.1109/CVPR.2019.00966]

Bilen H, Pedersoli M and Tuytelaars T.2015. Weakly supervised object detection with convex clustering// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Boston, USA: IEEE, 1081 - 1089.[DOI: 10.1109/CVPR.2015.7298711]

Bilen H, Vedaldi A.2016. A Weakly supervised deep detection networks// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Las Vegas, USA: IEEE, 2846 - 2854.[DOI: 10.1109/CVPR.2016.311]

Cao J L, Li Y L, Sun H Q, Xie J, Huang K Q and Pang Y W. 2022. A survey on deep learning based visual object detection. Journal of

Image and Graphics,27(06):1697-1722(曹家乐,李亚利,孙汉卿,谢今,黄凯奇,庞彦伟. 2022. 基于深度学习的视觉目标检测技术综述. 中国图象图形学报,27(06):1697-1722)[DOI:10.11834/ jig. 220069]

Cao T, Du L, Zhang X, Chen S H, Zhang Y and Wang Y F.2021.CaT: Weakly supervised object detection with category Transfer// Proceedings of the International Conference on Computer Vision. Montreal, Canada: IEEE, 3070 - 3079. [DOI: 10.1109/ICCV48922.2021.00306]

Cheng G, Yang J, Gao D, Guo L and Han J W.2020. High-quality proposals for weakly supervised object detection. IEEE Transactions on Image Processing, 29: 5794 - 5804. [DOI: 10.1109/TIP.2020.2987161]

Chen Z, Fu Z, Jiang R, Chen Y W and Hua X S.2020.Slv: Spatial likelihood voting for weakly supervised object detection// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Seattle, USA: IEEE, 12995 - 13004. [DOI: 10.1109/CVPR42600.2020.01301]

Dong B, Huang Z, Guo Y, Wang Q L, Niu Z X and Zuo W M.2021.Boosting weakly supervised object detection via learning bounding box adjusters// Proceedings of the International Conference on Computer Vision. Montreal, Canada: IEEE, 2876 - 2885. [DOI: 10.1109/ICCV48922.2021.00287]

Ding W Z, Jun W H, Gong C and Ming H Y.2022. Weakly supervised object localization and detection: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence.44(9): 5866-5885 [DOI: 10.1109/TPAMI.2021.3074313]

Dietterich T G, Lathrop R H and Lozano-pérez T.1997.Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence, 89(1-2): 31 - 71. [DOI: 10.1016/S0004-3702(96)00034-3]

Durand T, Mordan T, Thome N and Matthieu C.2017.Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Honolulu, USA: IEEE, 642-651. [DOI: 10.1109/CVPR.2017.631]

Dong X, Meng D, Ma F and Yang Y.2017.A dual-network progressive approach to weakly supervised object detection// Proceedings of the 25th ACM International Conference on Multimedia.Mountain View: USA, 279-287.[DOI: 10.1145/3123266.3123455]

Diba A, Sharma V, Pazandeh A, Hamed P and Luc V G.2017.Weakly supervised cascaded convolutional networks// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE,914 - 922.[DOI: 10.1109/CVPR.2017.545]

EveringhamM, Eslami S M A, VanGool L, Christopher K. I. Williams, John M W and Andrew Z.2015.The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision, 111(1): 98 - 136.[DOI: 10.1007/s11263-014-0733-5]

Everingham M, VanGool L, Williams C K I, John M W and Andrew Z.2010.The pascal visual object classes (voc) challenge. International Journal of Computer Vision, 88(2): 303 - 338.[DOI: 10.1007/s11263-009-0275-4]

Fei F S, Long C, Jian S, Wei J, Shao N X, Lu Y, Yue T Z and Jun X.2022. Deep learning for weakly-supervised object detection and localization: a survey. Neurocomputing 496: 192-207 [DOI: 10.1016/j.neucom.2022.01.095]

Feng X, Han J, Yao X and Cheng G.2020. Progressive contextual instance refinement for weakly supervised object detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 58(11): 8002-8012.[DOI: 10.1109/TGRS.2020.2985989]

Girshick R.2015.Fast r-cnn// Proceeding of the International Conference on Computer Vision[C]. Santiago, Chile: IEEE, 1440 - 1448.[DOI: 10.1109/ICCV.2015.169]

Gao Y, Liu B, Guo N, Ye X C, Wan F, You H H and Fan D R.2019.C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection// Proceedings of the International Conference on Computer Vision.Seoul, Korea: IEEE, 9834 - 9843.[DOI: 10.1109/ICCV.2019.00993]

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

Gonthier N, Gousseau Y, Ladjal S and Olivier B.2018.Weakly supervised object detection in artworks// Proceedings of the European Conference on Computer Vision Workshops.Munich, Germany: Springer. [DOI: 10.1007/978-3-030-11012-3_53]

Gonthier N, Ladjal S and Gousseau Y.2022.Multiple instance learning on deep features for weakly supervised object detection with extreme domain shifts. Computer Vision and Image Understanding, 214: 103299.[DOI: 10.1016/j.cviu.2021.103299]

Gao W, Wan F, Yue J, Xu S C and Ye Q X.2022.Discrepant multiple instance learning for weakly supervised object detection. Pattern Recognition, 122: 108233.[DOI: 10.1016/j.patcog.2021.108233]

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

Huang Z, Zou Y, B.V.K.Vijaya Kumar and Huang D.2020.Comprehensive attention self-distillation for weakly-supervised object detection// Proceedings of the Neural Information Processing Systems.Online: MIT Press, 33: 16797 - 16807.[DOI: 10.48550/arXiv.2010.12023]

He K, Gkioxari G, Dollár P and Ross B G.2017. Mask r-cnn// Proceedings of the International Conference on Computer Vision.Venice, Italy: IEEE, 2961 - 2969. [DOI: 10.1109/ICCV.2017.322]

Haußmann M, Hamprecht F A and Kandemir M.2017.Variational bayesian multiple instance learning with gaussian processes// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Honolulu, USA: IEEE, 6570-6579. [DOI: 10.1109/CVPR.2017.93]

Hou L, Zhang Y, Fu K and Li J.2021.Informative and consistent correspondence mining for cross-domain weakly supervised object detection// Proceedings of the Computer Vision and Pattern Recognition.Nashville, USA: IEEE, 9924 - 9933. [DOI: 10.1109/CVPR46437.2021.00980]

Inoue N, Furuta R, Yamasaki T and Kiyoharu A.2018.Cross-domain weakly-supervised object detection through progressive domain adaptation// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Salt Lake, USA: IEEE, 5001 - 5009. [DOI: 10.1109/CVPR.2018.00525]

Iqbal J, Munir M A, Mahmood A, Afsheen R A and Mohsen A.2021. Leveraging orientation for weakly supervised object detection with application to firearm localization. Neurocomputing, 440: 310-320.[DOI: 10.1016/j.neucom.2021.01.075]

Jia Q, Wei S, Ruan T, Zhao Y, Zhao Y.2021.GradingNet: Towards providing reliable supervisions for weakly supervised object detection by grading the box candidates// Proceedings of the AAAI Conference on Artificial Intelligence.Vancouver, Canada: AAAI, 35(2): 1682 - 1690. [DOI: 10.1108/AAAI.16261]

Kosugi S, Yamasaki T and Aizawa K.2019.Object-aware instance labeling for weakly supervised object detection// Proceedings of the International Conference on Computer Vision.Seoul, Korea: IEEE, 6064 - 6072. [DOI: 10.1109/ICCV.2019.00616]

KantorovV, Oquab M, Cho M and Ivan L.2016.Contextlocnet: Context-aware deep network models for weakly supervised localization// Proceedings of the European Conference on Computer Vision.Amsterdam, Netherlands: Springer, 350 - 365. [DOI: 10.1007/978-3-319-46454-1_22]

Liu L, Ouyang W, Wang X, Fieguth P, Chen J, Liu X, and Pietikainen M.2020. Deep learning for generic object detection: A survey. International Journal of Computer Vision.128(2): 261-318 [DOI: 10.1007/s11263-019-01247-4]

Li X, Kan M, Shan S and Chen X L.2019. Weakly supervised object detection with segmentation collaboration// Proceedings of the International Conference on Computer Vision. Seoul. Korea: IEEE, 9735 - 9744. [DOI: 10.1109/ICCV.2019.00983]

Lin C, Wang S, Xu D, Liu Y and Zhang W Y.2020.Object instance mining for weakly supervised object detection// Proceedings of the AAAI Conference on Artificial Intelligence.New York, USA: AAAI, 34(07): 11482 – 11489. [DOI: 10.48550/arXiv.2002.01087]

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Doll´ar P, and Zitnick C L.2014.Microsoft coco: Common objects in context// Proceedings of the European Conference on Computer Vision.Zurich, Switzerland: Springer. [DOI: 10.1007/978-3-319-10602-1_48]

Lin T Y, DollárP, Girshick R, He K M, Bharath H and Serge J B.2017.Feature pyramid networks for object detection// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Honolulu, USA: IEEE, 2117 - 2125. [DOI: 10.1109/CVPR.2017.106]

Mahajan D, Girshick R, Ramanathan V, He K M, Manohar P, Li Y X, Ashwin B and Laurens v d M.2018.Exploring the limits of weakly supervised pretraining// Proceedings of thr European Conference on Computer Vision.Munich, Germany: Springer, 181 - 196. [DOI: 10.1007/978-3-030-01216-8_12]

Nguyen D K, Tseng W L and Shuai H H.2020. Domain-adaptive object detection via uncertainty-aware distribution alignment// Proceedings of the 28th ACM International Conference on Multimedia.Seattle: USA, 2499-2507. [DOI: 10.1145/3394171.3413953]

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

Ren Z, Yu Z, Yang X, Liu M Y, Lee Y J, Alexander G S and Jan K.2020.Instance-aware, context-focused, and memory-efficient weakly supervised object detection// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Seattle,

USA: IEEE, 10598 - 10607. [DOI: 10.1109/CVPR42600.2020.01061]

Ren S, He K, Girshick R and Sun J.2015.Faster r-cnn: Towards real-time object detection with region proposal networks// Proceedings of the Neural Information Processing Systems.Montreal, Canada: MIT Press, 28: 91 - 99. [DOI: 10.1109/TPAMI.2016.2577031]

Ren D W, Wang Q L, Wei Y C, Meng D Y and Zuo W M. 2022. Progress in weakly supervised learning for visual understanding. Journal of Image and Graphics,27(06):1768-1798(任冬伟,王旗龙,魏云超,孟德宇,左旺孟. 2022. 视觉弱监督学习研究进展. 中国图象图形学报,27(06):1768-1798)[DOI:10. 11834/ jig. 220178]

Song, L., Liu, J., Sun, M., and Shang, X. 2021. Weakly supervised group mask network for object detection. International Journal of Computer Vision, 129(3), 681-702. [DOI:10.1007/s11263-020-01397-w]

Selvaraju R R, Cogswell M, Das A, Ramakrishna V, Devi P and Dhruv Batra.2017. Grad-cam: Visual explanations from deep networks via gradient-based localization// Proceedings of the International Conference on Computer Vision.Venice, Italy: IEEE, 618 - 626. [DOI: 10.1007/s11263-019-01228-7]

Simonyan K and Zisserman A.2015.Very deep convolutional networks for large-scale image recognition// Proceedings of the International Conference on Learning Representations.San Diego, USA: IEEE, 714 - 723. [DOI: 10.48550/arXiv.1409.1556]

Shen Y, Ji R, Wang Y, Wu Y J and Cao L J.2019.Cyclic guidance for weakly supervised joint detection and segmentation// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Long Beach, USA: IEEE, 697 - 707. [DOI: 10.1109/CVPR.2019.00079]

Shen Y, Ji R, Zhang S, Zuo W M and Wang Y.2018.Generative adversarial learning towards fast weakly supervised detection// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Salt Lake, USA: IEEE, 5764 - 5773. [DOI: 10.1109/CVPR.2018.00604]

Shen Y, Ji R, Wang C, Li X and Li X L.2018.Weakly supervised object detection via object-specific pixel gradient. IEEE Transactions on Neural Networks and Learning Systems, 29(12): 5960-5970.[DOI: 10.1109/TNNLS.2018.2816021]

Sangineto E, Nabi M, Culibrk D and Nicu S.2018.Self paced deep learning for weakly supervised object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(3): 712-725.[DOI: 10.1109/TPAMI.2018.2804907]

Shen Y, Ji R, Wang Y, Chen Z W, Zheng F, Huang F Y and Wu Y S.2020.Enabling deep residual networks for weakly supervised object detection// Proceedings of the European Conference on Computer Vision.Online: Springer: 118-136.[DOI: 10.1007/978-3-030-58598-3_8]

Shen Y, Ji R, Chen Z, Wu Y J and Huang F Y. 2020. UWSOD: Toward fully-supervised-level capacity weakly supervised object detection// Proceedings of the Neural Information Processing Systems, 33: 7005-7019[EB/OL]. [2021-10-14] https://proceedings.neurips.cc/paper/2020/file/4e0928de075538c593fbdabb0c5ef2c3-Paper. pdf

Tang P, Wang X, Wang A, Yan Y L, Liu W Y, Huang J Z and Alan L.Y.2018.Weakly supervised region proposal network and object detection// Proceedings of the European Conference on Computer Vision.Munich, Germany: Springer, 352 - 368. [DOI: 10.1007/978-3-030-01252-6_22]

Tang P, Wang X, Bai X, Liu W Y.2017.Multiple instance detection network with online instance classifier refinement// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Honolulu, USA: IEEE, 2843 - 2851. [DOI: 10.1109/CVPR.2017.326]

Tang P, Wang X, Bai S, Shen W, Bai X, Liu W Y and Alan L Y.2018.Pcl: Proposal cluster learning for weakly supervised object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(1): 176 - 191. [DOI: 10.1109/TPAMI.2018.2876304]

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

Wan F, Liu C, Ke W, Ji X Y, Jiao J B and Ye Q X.2019.C-mil: Continuation multiple instance learning for weakly supervised object detection// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Long Beach, USA: IEEE, 2199 - 2208. [DOI: 10.1109/CVPR.2019.00230]

Wan F, Wei P, Jiao J and Ye Q X.2018.Min-entropy latent model for weakly supervised object detection// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Salt Lake, USA: IEEE, 1297 - 1306. [DOI: 10.1109/CVPR.2018.00141]

Wei Y, Feng J, Liang X, Cheng M M, Zhao Y and Yan S C.2017.Object region mining with adversarial erasing: A simple classification to semantic segmentation approach// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Honolulu, USA: IEEE, 1568 - 1576. [DOI: 10.1109/CVPR.2017.687]

Wei Y, Shen Z, Cheng B, Shi H H, Xiong J J, Feng J S and Thomas

S.Huang. 2018.Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection// Proceedings of the European Conference on Computer Vision.Munich, Germany: Springer, 434 - 450.[DOI: 10.48550/arXiv.1807.04897]

Wei Y C, Liang X, Chen Y, Shen X H, Cheng M M, Feng J S, Zhao Y and Yan S C.2016.Stc: A simple to complex framework for weakly-supervised semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(11): 2314 - 2320. [DOI: 10.1109/TPAMI.2016.2636150]

Wei Y C, Xiao H, Shi H, Jie Z Q, Feng J S and Thomas S H.2018. Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Salt Lake, USA: IEEE, 7268 – 7277. [DOI: 10.1109/CVPR.2018.00759]

Wang X, You S, Li X and Ma H M.2018.Weakly-supervised semantic segmentation by iteratively mining common object features// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Salt Lake, USA: IEEE, 1354 - 1362. [DOI: 10.1109/CVPR.2018.00147]

Wang, G., Zhang, X., Peng, Z., Tang, X., Zhou, H., and Jiao, L. 2022. Absolute Wrong Makes Better: Boosting Weakly Supervised Object Detection via Negative Deterministic Information// Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, (IJCAI), Vienna, Austria, 1378—1384. [DOI: 10.24963/ijcai.2022/192]

Xu, Y., Zhou, C., Yu, X., Xiao, B., and Yang, Y. 2021. Pyramidal multiple instance detection network with mask guided self-correction for weakly supervised object detection. IEEE Transactions on Image Processing, 30, 3029-3040. [DOI: 10.1109/TIP.2021.3056887]

Xu X K，Ma Y，Qian X and Zhang Y． 2021． Scale-aware EfficientDet: real-time pedestrian detection algorithm for automated driving．Journal of Image and Graphics，26(01) : 0093-0100(徐歆恺，马岩，钱旭，张龑．2021．自动驾驶场景的尺度感知实时行人检测．中国图象图形学报，26(01) : 0093-0100)［DOI: 10. 11834 /jig. 200445］

Yang H, Quan J C, Liang X Y and Wei Z W.2021.Research Progress of Object Detection Based on Weakly Supervised Learning. Computer Engineering and Applications, ,57(16): 40-49 (杨辉, 权冀川, 梁新宇, 王中伟.2021.基于弱监督学习的目标检测研究进展.计算机工程与应,57(16): 40-49) [DOI: 10.3778/j.issn.1002-8331.2103-0306]

Yang K, Li D and Dou Y.2019.Towards precise end-to-end weakly supervised object detection network// Proceedings of the International Conference on Computer Vision.Seoul, Korea: IEEE, 8372 - 8381. [DOI: 10.1109/ICCV.2019.00846]

Yin Y, Deng J, Zhou W and Li H Q. 2021. Instance mining with class feature banks for weakly supervised object detection// Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI, 35(4): 3190 – 3198[EB/OL]. [2021-06-02] https://ojs. aaai. org/index. php/AAAI/article/view/16429

Yang Z, Mahajan D, Ghadiyaram D, Ram N and Vignesh R.2019.Activity driven weakly supervised object detection// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Long Beach, USA: IEEE, 2917-2926. [DOI: 10.1109/CVPR.2019.00303]

Yao X, Feng X, Han J, Cheng G and Guo L.2020.Automatic weakly supervised object detection from high spatial resolution remote sensing images via dynamic curriculum learning. IEEE Transactions on Geoscience and Remote Sensing, 59(1): 675-685. [DOI: 10.1109/TGRS.2020.2991407]

Zhou X L, Chen X J, Chen S Y and Jun B.2019.Weakly supervised learning-based object detection: A survey. Computer Science, 46(11): 49-57 (周小龙, 陈小佳, 陈胜勇, 帮军.2019.弱监督学习下的目标检测算法综述.计算机科学, 46(11): 49-57) [DOI: 10.11896/jsjkx.181001899]

Zhou M F, Wang X L. 2018. Object detection models of remote sensing images using deep neural networks with weakly supervised training method (in Chinese). SciSin Inform, 2018, 48: 1022-1034(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. 中国科学: 信息科学, 2018, 48: 1022-1034) [DOI: 10.1360/N112017-00208]

Zhao W Q，Kong Z X，Zhou Z D and Zhao Z B． 2021． Target detection algorithm of aerial remote sensing based on feature enhancement technology． Journal of Image and Graphics，26(03) : 0644-0653(赵文清，孔子旭，周震东，赵振兵． 2021．增强小目标特征的航空遥感目标检测．中国图象图形学报，26(03) : 0644-0653)［DOI: 10. 11834 /jig. 190612］

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

Zhu Y, Zhou Y, Ye Q, Qiu Q and Jiao J B.2017.Soft proposal networks for weakly supervised object localization// Proceedings of the International Conference on Computer Vision.Venice, Italy: IEEE, 1841 - 1850. [DOI: 10.1109/ICCV.2017.204]

Zeng Z, Liu B, Fu J, Chao H Y and Zhang L.2019.Wsod2: Learning bottom-up and top-down objectness distillation for weakly-

supervised object detection// Proceedings of the International Conference on Computer Vision.Seoul, Korea: IEEE, 8292 - 8300. [DOI: 10.1109/ICCV.2019.00838]

Zhou B, Khosla A, Lapedriza A, Aude O and Antonio T.2016. Learning deep features for discriminative localization// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Las Vegas, USA: IEEE, 2921 - 2929. [DOI: 10.1109/CVPR.2016.319]

Zhang X, Feng J, Xiong H, Tian Q.2018.Zigzag learning for weakly supervised object detection// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Salt Lake, USA: IEEE, 4262 - 4270. [DOI: 10.1109/CVPR.2018.00448]

Zhang Y, Bai Y, Ding M, Li Y Q and Bernard G.2018.W2f: A weakly-supervised to fully-supervised framework for object detection// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Salt Lake, USA: IEEE, 928 - 936. [DOI: 10.1109/CVPR.2018.00103]

Zhang B, Xiao J, Wei Y, Sun M J and Huang K Z. 2020. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach// Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA: AAAI, 34(07): 12765 – 12772[EB/OL]. [2020-02-12] https://ojs. aaai. org/index. php/AAAI/article/view/6971

Zhang Y, Bai Y, Ding M, Li Y Q and Bernard G.2018, Weakly-supervised object detection via mining pseudo ground truth bounding-boxes. Pattern Recognition, 84: 68-81. [DOI: 10.1016/j.patcog.2018.07.005]

Zhang D, Han J, Zhao L, Zhao L and Meng D Y.2019.Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework. International Journal of Computer Vision, 127(4): 363-380. [DOI: 10.1007/s11263-018-1112-4]

Zhang D, Han J, Zhao L, Zhao T.2020.From discriminant to complete: Reinforcement searching-agent learning for weakly supervised object detection. IEEE Transactions on Neural Networks and Learning Systems, 31(12): 5549-5560. [DOI: 10.1109/TNNLS.2020.2969483]

Zhang, D., Zeng, W., Yao, J., and Han, J. 2020. Weakly supervised object detection using proposal-and semantic-level relationships. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(6),3349-3363. [DOI: 10.1109/TPAMI.2020.3046647]

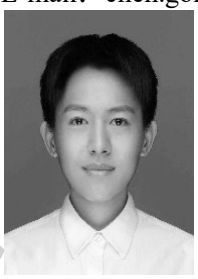
Zhang X, Wei Y, Feng J, Yang Y and Thomas S H.2018.Adversarial complementary learning for weakly supervised object localization// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Salt Lake, USA: IEEE: 1325-1334. [DOI: 10.1109/CVPR.2018.00144]

Zhong Y, Wang J, Peng J and Zhang L.2020.Boosting weakly supervised object detection with progressive knowledge transfer// Proceedings of the European Conference on Computer Vision.Online: Springer: 615-631. [DOI: 10.1007/978-3-030-58574-7_37]

作者简介：

陈震元，1998 年生，男，南京理工大学计算机科学与工程学院硕士研究生。主要研究方向为图像处理，语义分割和目标检测。
E-mail: zhenyuanchen@njust.edu.cn

宫辰，通信作者，男，南京理工大学计算机科学与工程学院教授。主要研究方向为机器学习与数据挖掘。
E-mail：chen.gong@njust.edu.cn



王振东，1996 年生，男，南京理工大学计算机科学与工程学院硕士研究生。主要研究方向为图像处理与目标检测。
E-mail：aes1758@163.com