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

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2022)06-1768-31

论文引用格式: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]

视觉弱监督学习研究进展

任冬伟[1],王旗龙[2],魏云超[3],孟德宇[4],左旺孟[1]*

1. 哈尔滨工业大学,哈尔滨　150001; 2. 天津大学,天津　300350;
3. 北京交通大学,北京　100091; 4. 西安交通大学,西安　710049

摘　要:视觉理解,如物体检测、语义和实例分割以及动作识别等,在人机交互和自动驾驶等领域中有着广泛的应用并发挥着至关重要的作用。近年来,基于全监督学习的深度视觉理解网络取得了显著的性能提升。然而,物体检测、语义和实例分割以及视频动作识别等任务的数据标注往往需要耗费大量的人力和时间成本,已成为限制其广泛应用的一个关键因素。弱监督学习作为一种降低数据标注成本的有效方式,有望对缓解这一问题提供可行的解决方案,因而获得了较多的关注。围绕视觉弱监督学习,本文将以物体检测、语义和实例分割以及动作识别为例综述国内外研究进展,并对其发展方向和应用前景加以讨论分析。在简单回顾通用弱监督学习模型,如多示例学习(multiple instance learning, MIL)和期望—最大化(expectation-maximization, EM)算法的基础上,针对物体检测和定位,从多示例学习、类注意力图机制等方面分别进行总结,并重点回顾了自训练和监督形式转换等方法;针对语义分割任务,根据不同粒度的弱监督形式,如边界框标注、图像级类别标注、线标注或点标注等,对语义分割研究进展进行总结分析,并主要回顾了基于图像级别类别标注和边界框标注的弱监督实例分割方法;针对视频动作识别,从电影脚本、动作序列、视频级类别标签和单帧标签等弱监督形式,对弱监督视频动作识别的模型与算法进行回顾,并讨论了各种弱监督形式在实际应用中的可行性。在此基础上,进一步讨论视觉弱监督学习面临的挑战和发展趋势,旨在为相关研究提供参考。

关键词:弱监督学习;目标定位;目标检测;语义分割;实例分割;动作识别

Progress in weakly supervised learning for visual understanding

Ren Dongwei[1], Wang Qilong[2], Wei Yunchao[3], Meng Deyu[4], Zuo Wangmeng[1]*

1. Harbin Institute of Technology, Harbin　150001, China; 2. Tianjin University, Tianjin　300350, China;
3. Beijing Jiaotong University, Beijing　100091, China; 4. Xi'an Jiaotong University, Xi'an　710049, China

Abstract: Visual understanding, e. g. , object detection, semantic/instance segmentation, and action recognition, plays a crucial role in many real-world applications including human-machine interaction, autonomous driving, etc. Recently, deep networks have made great progress in these tasks under the full supervision regime. Based on convolutional neural network (CNN), a series of representative deep models have been developed for these visual understanding tasks, e. g. , you only look once (YOLO) and Fast/Faster R-CNN (region CNN) for object detection, fully convolutional networks (FCN) and DeepLab for semantic segmentation, Mask R-CNN and you only look at coefficients (YOLACT) for instance segmentation.

收稿日期:2022-03-05;**修回日期**:2022-03-07;**预印本日期**:2022-03-15

*通信作者:左旺孟　wmzuo@hit.edu.cn

基金项目:科技创新2030——"新一代人工智能"重大项目(2021ZD0112100);国家自然科学基金项目(62172127,U19A2073)

Supported by: National Key R&D Program of China(2021ZD0112100); National Natural Science Foundation of China (62172127, U19A2073)

Recently, driven by novel network backbone, e. g. , Transformer, the performance of these tasks have been further boosted under full supervision regime. However, supervised learning relies on massive accurate annotations, which are usually laborious and costly. By taking semantic segmentation as an example, it is very laborious and costly for collecting dense annotations, i. e. , pixel-wise segmentation masks, while weak supervision annotations, e. g. , bounding box annotations, point annotations, are much easier to collect. Moreover, for video action recognition, the scenes in videos are very complicated, and it is very likely to be impossible to annotate all the actions with accurate time intervals. Alternatively, weakly supervised learning is effective in reducing the cost of data annotations, and thus is very important to the development and applications of visual understanding. Taking object detection, semantic/instance segmentation, and action recognition as examples, this article aims to provide a survey on recent progress in weakly supervised visual understanding, while pointing out several challenges and opportunities. To begin with, we first introduce two representative weakly supervised learning methods, including multiple instance learning (MIL) and expectation-maximization (EM) algorithms. Despite of different network architectures in recent weakly supervised learning methods, most existing methods can be categorized into the family of MIL or EM. As for object localization and detection, we respectively review the methods based on MIL and class attention map (CAM), where self-training and switching between supervision settings are specifically introduced. By formulating weakly supervised object detection (WSOD) as the problem of MIL-based proposal classification, WSOD methods tend to focus on discriminative parts of object, e. g. , head for human or animals may be simply detected to represent the entire object, yielding significant performance drops in comparison to fully supervised object detection. To address this issue, self-training and switching between supervision settings have been respectively developed, and transfer learning has also been introduced to exploit auxiliary information from other tasks, e. g. , semantic segmentation. As for weakly supervised object localization, CAM is a popular solution to predict the object position where objects from one class with the highest activation value can be found. Similarly, CAM based localization methods are also facing the issue of discriminative parts, and several solutions, e. g. , suppressing the most discriminative parts and attention-based self-produced guidance, have been proposed. Based on pattern analysis, statistical modeling and computational learning visual object classes (PASCAL VOC) and Microsoft common objects in context (MS COCO) datasets, several representative weakly supervised object localization and detection methods have been evaluated, showing performance gaps between fully supervised methods. As for semantic segmentation, we consider several representative weak supervision settings including bounding box annotations, image-level class annotations, point or scribble annotations. In comparison to segmentation mask annotations, these weak annotations cannot provide accurate pixel-wise supervision. Image-level class annotations are the most convenient and easiest way, and the key issue of image-level weakly supervised semantic segmentation methods is to exploit the correlation between class labels and segmentation masks. Based on CAM, coarse segmentation results can be obtained, while facing inaccurate segmentation masks and focusing on discriminative parts. To refine segmentation masks, several strategies are introduced including iterative erasing, learning similarity between pixels, and joint learning of saliency detection and weakly supervised semantic segmentation. Point or scribble annotations and bounding box annotations can provide more accurate localization information than image-level class annotations. Among them, bounding box annotations is likely to be a good solution to balance the annotation cost and performance of weakly supervised semantic segmentation under EM framework. Moreover, weakly supervised instance segmentation is more challenging than weakly supervised semantic segmentation, since a pixel is not only assigned to an object class but also is accurately assigned to one specific object. In this article, we consider bounding box annotations and image-level annotations for weakly supervised instance segmentation. Based on image-level class annotations, peak response map in CAM is highly correlated with object instances, and can be adopted in weakly supervised instance segmentation. Based on bounding box annotations, weakly supervised instance segmentation can be formulated as MIL, where instance masks are usually more accurate than those based on image-level class annotations. Besides, in these weakly supervised segmentation methods, post-processing techniques, e. g. , dense conditional random field, are usually adopted to further refine the segmentation masks. On PASCAL VOC and MS COCO datasets, representative weakly supervised semantic and instance segmentation methods with different levels of annotations are evaluated. As for video action recognition, it is much more difficult to collect accurate annotations of all the actions due to complicated scenes in videos, and thus weakly supervised action recognition is attracting research attention in recent years. In this article, we introduce the

models and algorithms for different weak supervision settings including film scripts, action sequences, video-level class labels and single-frame labels. Finally, the challenges and opportunities are analyzed and discussed. For these visual understanding tasks, the performance of weakly supervised methods still has improvement room in comparison to fully supervised methods. When applying in the wild, it is also a valuable and challenging topic to exploit large amount of unlabeled and noisy data. In future, weakly supervised visual understanding methods also will benefit from multi-task learning and large-scale pre-trained models. For an example, vision and language pre-trained models, e. g. , contrastive language-image pre-training (CLIP), is potential to provide knowledge to significantly improve the performance of weakly supervised visual understanding tasks.

Key words: weakly supervised learning; object localization; object detection; semantic segmentation; instance segmentation; action recognition

0 弱监督学习方法

近年来,视觉理解技术的快速发展仍主要建立在监督学习的基础上。然而,物体检测、语义和实例分割以及视频动作识别等视觉理解任务往往需要大量的全标注数据集 $\boldsymbol{D}^{\mathrm{s}}=\{(\boldsymbol{x},\boldsymbol{y}^{\mathrm{s}})\}$,其中 $\boldsymbol{x}$ 为输入样本,$\boldsymbol{y}^{\mathrm{s}}$是全标注。例如,对于分割任务,图像 $\boldsymbol{x}$ 中包含各种物体,全标注 $\boldsymbol{y}^{\mathrm{s}}$为像素级别的类别标签。得益于深度学习技术的快速发展,全监督学习在各种视觉理解任务中已经取得了巨大成功,如目标检测任务中的 YOLO (you only look once) (Redmon 等,2016)、Fast R-CNN (fast region-based convolutional neural network) (Girshick,2015;He 等,2017;Ren 等,2017)、FCOS (fully convolutional one-stage object detector) (Tian 等,2019)和 DETR (detection transformer) (Carion 等,2020)等,语义分割中的 DeepLab (Chen 等,2018)、PSPNet (pyramid scene parsing network) (Zhao 等,2017)等。然而,大规模数据集的全标注往往要耗费大量的人力成本和时间成本。仍以图像分割为例,为获取全监督训练数据集如 CityScapes (Cordts 等,2016),需人工对图像进行像素级别标注。该数据集中包含 5 000 幅图像,使用 LabelMe 软件(Russell 等,2008)进行精细的像素级标注,每幅图像平均需要花费超过 1.5 h 来保证标注的质量,人工标注的成本大约是每小时 6 ~ 10 美元。显然,过高的标注成本制约了语义分割在其他类别上的更快发展。无监督生成学习和自监督学习虽然在很大程度上能够缓解标注代价,但仍需要一定数量的全标注数据用于模型微调。数据标注作为一种典型劳动密集型工作,已成为关乎当前整个 AI(artificial intelligence)产业的基础。如何有效地降低标注成本或者避免数据标注,同时保证视觉理解模型的性能,不仅是深度学习未来发展应用的关键问题,同时也是机器学习乃至于人工智能领域的重要开放问题,在经济和社会层面上均具有重要的研究意义。

弱监督学习作为一种降低数据标注成本的有效方式,有望对缓解这一问题提供可行的解决方案,因而获得了较多的关注。在视觉弱监督学习方法中,对于样本 $\boldsymbol{x}$ 仅需提供弱标注 $\boldsymbol{y}^{\mathrm{w}}$ 构成弱监督数据集 $\boldsymbol{D}^{\mathrm{w}}=\{(\boldsymbol{x},\boldsymbol{y}^{\mathrm{w}})\}$。如对于图像分割任务,图像级别和标注框的弱监督标注,相较于像素级别的标注的代价显著降低。仍以 CityScapes 数据库为例,一个边界框的标注需要 7 s,一个图像的类别标注只需要 1 s,弱标注相较于像素级的全标注显著降低了成本。视觉弱监督学习旨在利用弱标注数据集 $\boldsymbol{D}^{\mathrm{w}}$,通过发展有效的学习模型以缩小与全监督模型的性能差距。视觉弱监督方法能够显著降低标注成本,且期望接近全监督视觉模型的性能。因而如何结合深度学习和视觉数据任务特点发展视觉弱监督学习模型方法,成为近年来计算机视觉领域的一个研究热点。

0.1 多示例学习

作为一种典型的弱监督学习问题,多示例学习(multi-instance learning, MIL) (Zhou,2004)中的多个示例形成一个“包” $\boldsymbol{O}$,整个“包”被赋予一个标签 Y。由于许多视觉弱监督任务可以建模为多示例学习问题,MIL 获得了较为广泛的研究和应用。MIL 假设“负包”中只存在负示例,而只要存在至少一个正示例则为“正包”,即 MIL 中“包”的标签可定义为

$$Y=\begin{cases}+1 & \exists y_i=+1\\ -1 & \forall y_i=-1\end{cases}$$

MIL 算法通常包括两个交替迭代的步骤:示例

选择和模型学习。在示例选择中，示例选择器会计算每一个示例的目标得分，并从包中选择一个最大得分的示例作为正示例。然后，模型学习利用选择正示例和负包中的负示例来更新模型。为了更好地提取正示例，Erez 和 Maron（1998）将每个包视为一个流形，其中一个新形成的包被判为正包当且仅当其与所有的正特征流形相交，且与所有的负特征流形不相交。在某些问题中，同一个包中的示例之间具有一定的相似性，并且与其他包中的示例之间没有相似性。miGrapa 算法（Zhou 等，2009）对每个包建立图模型，相似的示例被划为到一个组中，从而通过组大小一起调整它们的相对重要性。在 CCE（constructive clustering based ensemble）（Zhou 和 Zhang，2007）中，每个示例都被划分到一个聚类中，一个包被表示为一个二元向量，用每个值表示是否包含某个聚类。相似的示例在同一个聚类中，对包内相似性具有一定的鲁棒性。

许多视觉弱监督学习任务均可建模为 MIL 问题，MIL 因而成为弱监督物体检测、语义分割以及动作识别中应用最为广泛的方法。由于 MIL 通常关注包中最具区别性的示例，在检测、分割以及动作识别中均存在一定的局限性，需要针对具体的视觉任务进行正负包和 MIL 模型设计。

0.2　期望—最大化算法

期望—最大化（expectation-maximization，EM）（Dempster 等，1977）旨在利用最大似然估计解决缺失数据问题。给定观测数据 $\boldsymbol{X}$，模型依赖于无法观测的隐变量 $\boldsymbol{Z}$，即缺失数据，似然函数可定义为

$$L(\theta;\boldsymbol{X}) = p(\boldsymbol{X}\mid\theta) = \int p(\boldsymbol{X},\boldsymbol{Z}\mid\theta)\,\mathrm{d}\boldsymbol{Z}$$

EM 算法交替执行两个步骤，第 1 步是计算期望（E 步），利用对隐藏变量的现有估计值，计算其最大似然估计值

$$Q(\theta\mid\theta^{(t)}) = E_{\boldsymbol{Z}\mid\boldsymbol{X},\theta^{(t)}}[\log L(\theta;\boldsymbol{X},\boldsymbol{Z})]$$

第 2 步（M 步）通过最大化在 E 步上求得的最大似然值来计算参数的值，即

$$\theta^{(t)} = \arg\max_{\theta} Q(\theta\mid\theta^{(t)})$$

M 步上找到的参数估计值被用于下一个 E 步计算。基于 Jensen 不等式构造，可证明 EM 算法是在反复地构造新的下界，能够收敛到局部极值。传统 EM 算法对初始值较为敏感，算法收敛的优劣很大程度上取决于其初始参数（Blömer 和 Bujna，2013；Chen，2013）。对于视觉弱监督学习，可以对每一个观测数据 $\boldsymbol{x}$ 及其弱标注 $\boldsymbol{y}^{\mathrm{w}}$，把全标注 $\boldsymbol{y}^{\mathrm{s}}$ 当做隐变量，从弱标注数据中学习目标任务模型 $f(\boldsymbol{x};\boldsymbol{\Theta})$。其中 E 步估计伪强标注信息，M 步从估计的伪强标注信息训练目标视觉模型。

然而，当使用 EM 算法进行视觉弱监督学习时，E 步通常采用一些经验性准则将当前训练模型的估计结果转化为伪强标注信息，在较大程度上限制了 EM 算法的弱监督学习性能。因而，基于伪标签的视觉弱监督方法可归纳为 EM 框架，根据弱监督标记获取更强的伪标签，以缩小与全监督模型间的区别。如目标检测方法（Dong 等，2021）和实例分割方法（Xie 等，2021），在 E 步中分别根据图像类别标注预测伪边界框和边界框标注预测伪实例分割图，在 M 步中利用伪标签提升弱监督模型的性能。如何获取更好的伪标签是基于 EM 的视觉弱监督方法的关键。

1　弱监督物体检测与定位

1.1　典型问题设置

物体检测（Masita 等，2020）不仅是计算机视觉的一个重要的基础性任务，在许多应用领域中也发挥着至关重要的作用。物体检测旨在学习一个物体检测器，使其能够对图像中的感兴趣物体进行精确的分类和定位。随着深度卷积神经网络（convolutional neural network，CNN）（Simonyan 和 Zisserman，2015）的发展，以及大规模数据集（Russakovsky 等，2015）相继出现，全监督物体检测近年来已经取得了引人注目的发展，并涌现了一批代表性的深度物体检测模型（Cai 等，2020；He 等，2017；Law 和 Deng，2018；Redmon 等，2016；Ren 等，2017）。然而，大规模全标注数据集的构建往往需要耗费昂贵的人工和时间成本。为此，研究人员提出了弱监督物体检测方法（weakly-supervised object detection，WSOD）（Shao 等，2021）。

不同于全监督物体检测，弱监督物体检测方法旨在仅利用图像级的类别标签（image-level labels）训练物体检测器，以完成物体的分类和定位任务，如图 1 所示。弱监督物体检测方法不依赖于实例级的边界框标注信息（instance-level labels），同时随着互联网和搜索引擎的发展，带类别标注的图像更容易

图 1　物体检测不同监督形式标注示例

Fig. 1　Annotation examples with different supervision types for object detection

获得,进一步降低了弱监督物体检测的标注成本。因而,近年来弱监督物体检测得到了较多的关注和发展。

弱监督物体定位(weakly-supervised object localization,WSOL)与弱监督物体检测任务的定义相似,旨在利用图像集类别标签训练一个可检测图像中单一物体的检测器。弱监督物体检测和定位任务的区别为前者检测图像中多个物体,而后者只需检测单个物体,因此弱监督定位视为一种特殊的弱监督检测任务。针对两个任务不同的特性,弱监督物体检测通常采用多示例学习,而弱监督物体定位则主要基于类别激活图(class activation map,CAM)。

1.2　弱监督物体检测模型

主流方法通常将弱监督物体检测建模为 MIL(Dietterich 等,1997)问题。通过在待检测图像中提取若干候选框,则该幅图像以及生成的候选框则可视为一个包。如果该图像的类别标签中存在某个类别,则意味着该图像中至少存在着该类别的物体,即正示例,则在生成的候选框中,有一个或多个候选框包含该物体。反之,若该图像标签中没有某个类别,则所有的候选框中都不会涵盖该类别的物体,即负示例。基于上述表示,弱监督物体检测需要利用图像级的标签学习每一个候选框到对应类别的映射,然后将模型的预测结果中分数较高的候选框作为检测结果。因此,弱监督物体检测可以转化为基于 MIL 的候选框分类问题。具体而言,目前的弱监督物体检测模型主要基于 WSDDN(weakly supervised deep detection networks)(Bilen 和 Vedaldi,2016),由候选框生成器、特征提取器和检测头等模块构成。

1)特征提取器。该结构的作用是提取原始图像的特征,生成对应的特征图。通常采用经典的卷积网络结构,例如 AlexNet(Krizhevsky 等,2012),VGG(Visual Geometry Group)(Simonyan 和 Zisserman,2015),ResNet(He 等,2016),SENet(Hu 等,2018)等。首先利用大规模分类数据集进行网络预训练,然后去掉分类头,保留剩余的网络层作为后续物体检测模型的特征提取器。

2)候选框生成器。对每幅训练图像生成若干个候选框,通常为 1 000 ～ 4 000 个。一般采用 Selective search,Edge boxes 或 MCG(multiscale combinatorial grouping)(Uijlings 等,2013;Zitnick 和 Dollár,2014)等经典的候选框提取算法。给定候选框,采用全监督物体检测中的 RoI Pooling 和 RoI Align 操作,提取每一个候选框的特征表示。

3)检测头。由多层全连接层和一层输出层组成,形式上和全监督物体检测方法中的 Fast/Faster R-CNN 相似。然而,由于弱监督数据集没有提供物体级的边界框监督信息,因此在 WSDDN 网络设计中,采用双分支结构。其中分类分支预测每一个候选框对每一个类别的可能性,而定位分支预测每一个候选框对单一类别的贡献度。最后,这两个结果矩阵进行哈达马积操作,作为每一个候选框的最终分类分数。由于监督信息只有图像级的类别标签,因此将每一个类别按照候选框的维度累加,作为这幅图像的类别预测结果。

1.3　弱监督物体检测算法

WSDDN 是一种较早的基于深度网络的弱监督物体检测方法。在训练阶段,WSDDN 首先利用 Selective search(Uijlings 等,2013)或 Edge boxes(Zitnick 和 Dollár,2014)等对每一幅训练图像提取候选框,使用骨干网络提取训练图像的特征图表示。

随后,将 RoI Pooling (He 等,2015)用于把候选框映射到特征图的对应位置中,提取每一个候选框的特征向量。进而,将候选框特征输入到检测头中,输出该图像的分类预测结果,与图像级标签计算交叉熵损失,通过反向传播算法更新网络的参数。在推理阶段,网络根据每一个候选框的分类预测结果输出检测到的物体和边界框。

由于缺少示例级标注,仅依赖图像级标注去指导候选框的分类任务,WSDDN 的性能与全监督模型仍有较大差距。此外,WSDDN 会倾向于将物体中最具判别能力的部件,如动物和人的头部 (Zhou 等,2016)检测为物体。为了解决上述问题,近年来国内外学者在 WSDDN 的基础上做了更为深入的探索和改进。具体而言,自训练方法通过加大训练的优化次数以改进候选框筛选,而监督形式转换的方法以弱监督学习提供的伪标注为基础,进一步利用全监督或半监督学习改善学习性能。进而,迁移学习和多任务学习方法则通过引入与目标任务相关的源数据集和其他相关学习任务(如:语义分割)来改善弱监督物体检测性能。

1.3.1　自训练方法

为了缓解弱监督模型关注于判别性区域的问题, OICR (online instance classifier refinement) (Tang 等,2017)使用 WSDDN 作为基础架构,并在其之后进行 3 次在线实例分类器调整。每个分类器插入到骨干网络中,由两个全连接层外加一个分类层组成。在训练过程中,将原始 WSDDN 的输出结果作为中间结果,对其每个类别取分数最高的候选框作为示例级的伪监督信息,然后用其监督后续的一级示例分类器。第 1 级示例分类器的结果继续生成伪标签,以监督第 2 级的示例分类器,以此类推。该伪监督所计算的损失函数的形式类似于全监督方法中的 Fast R-CNN (Girshick,2015),最终由各个分类器所计算的损失函数共同优化网络。在推理阶段,每个分类器分别对测试图像及候选框输出分类分数,对每个候选框计算 3 个分类器的均值得到其最终得分。通过该多阶段学习的方式,使模型可以学习到更精准的特征表示。同时,华中科技大学相关学者(Tang 等,2020)基于 OICR 的思想,提出了候选框空间聚类的方式,将每一个聚类簇当成一个子 MIL 任务,并以此生成更好的伪标签以及对候选框的标签进行分配。

Zeng 等人(2019)利用底层视觉特征,通过自顶向下和自底向上的知识互蒸馏,来学习物体级信息更好的特征表示。自下而上的物体挖掘可以测量包含完整物体的边界框的概率,用于自适应训练中,在 CNN 中提取边界特征,并添加了边界框回归器来进一步改善模型的定位性能。Ren 等人(2020)对 OICR 的每一级分类器都添加了一个回归分支同时改善模型对候选框的定位能力,并设计了一个 Dropblock 去除每个候选框特征中的部分值,以改善模型会关注于判别区域的问题。此外,他们还设计了一个顺序批量反向传播算法,在内存消耗最大的阶段批量处理数据,以缓解内存消耗问题。Kosugi 等人(2019)进一步研究了正负样本挖掘问题,发现通过候选框的上下文分类损失大小可以初步判断该候选框是否准确地覆盖住了物体,因此设计了 CAP (context-aware positive)模块来挑选正样本。另外 Kosugi 等人(2019)提出了 SRN (spatially restricted negative)模块来避免将图像中部分待挖掘的物体标记为负样本。Yang 等人(2019)设计了一个单一的端到端弱监督对象检测网络,可以联合优化候选框分类和回归,从而显著提高物体的定位性能。同时,他们还设计了一个分类引导的注意力模块来增强特征学习的定位能力,有助于后续的分类和回归任务的学习。Yan 等人(2019)提出了一个 C-MIDN (coupled multiple instance detection network)模型,使用两个基于 MIL 的弱监督物体检测网络,通过交互的方式来进行低质量候选框的删除工作。同时,进一步引入了一种新颖的利用分割信息引导的候选框去除算法。最后,将两个 WSOD 的输出聚合到一起,以获得更精准的检测结果。Chen 等人(2020)为了解决模型聚集在判别区域的问题,提出一个空间似然投票(spatial likelihood voting,SLV)模块来生成位置更加准确的候选框。SLV 模块统计每一个像素点上所累加的候选框的分数,从分数高于一定阈值的区域中生成新的候选框,以此来消除判别区域的影响。此外,Chen 等人(2020)在模型中同样引入了回归分支,以进一步提升模型的定位性能。Kantorov 等人(2016)通过引入两种类型的上下文感知指导模型来解决弱监督检测问题,利用其周围的上下文区域的特征信息来改善模型的定位性能。对于加性模型,利用候选框的上下文信息来支持该候选框对某一类别的置信度。而对比模型则倾向于利用该候

选框和上下文信息的不同点来增强其置信度。Huang 等人(2020)提出了 CASD(comprehensive attention self-distillation)模型,计算从给定一幅图像的不同图像增强的副本,以及不同特征层来计算注意力。为了对候选框实施一致的空间约束,CASD 对 WSOD 网络进行自蒸馏,通过同一个候选框的不同视角的注意力对网络进行约束。Yin 等人(2021)提出了一个基于类别特征库的实例挖掘框架,由类别特征库(class filter bank,CFB)和特征引导的实例挖掘(feature-guided instance mining,FGIM)两部分组成。CFB 由每个类别的高质量物体特征组成,用于从更广泛的多角度集合每类物体的信息。在训练阶段,记录质量较高的候选框特征并更新在 CFB 中,然后,FGIM 利用 CFB 中记录的高质量特征来筛选其他候选框以改善 MIL 分支的训练。Jia 等人(2021)提出了一个两阶段的 WSOD 训练模型 GradingNet。具体来说,GradingNet 由两个模块组成:边界框打分模块和信息增强模块。边界框打分模块使用标准的单阶段弱监督方法生成候选框,然后利用"包含原则"来挑选出高度可靠的框并评估每个框的等级。有了边界框打分模块初步预测的候选框及其等级信息,Jia 等人(2021)又设计了一个有效的锚生成器和一个等级感知损失来训练信息增强模块,以提升网络的性能,可即插即用到任意主流的单阶段弱监督方法中。

1.3.2 监督形式转换训练方法

一些工作尝试将弱监督物体检测任务转化为全监督物体检测任务。一般而言,这类方法先从弱监督物体检测模型的输出中筛选检测框并生成伪标签,然后用其监督去训练一个全监督物体检测网络,例如 Fast/Faster R-CNN。Zhang 等人(2018c)提出了 W2F(weakly-supervised to fully-supervised framework)算法,该算法中包含伪标注挖掘和伪标注适应模块。具体而言,伪标注挖掘模块可以挖掘更准确和更紧密的伪标签,而伪标注适应模块在 FSOD(fully supervised object detection)训练时可以动态地调节伪标签,使得 FSOD 充分地学习到更准确的知识。

Sui 等人(2021)将弱监督物体检测任务转化为半监督物体检测任务,提出了 SoS-WSOD(salvage of supervision WSOD)方法,该方法包含 3 个步骤:1)传统的弱监督检测模型的训练;2)利用 WSOD 的输出结果作为伪标签来监督 FSOD 的训练;3)利用 FSOD 对训练图像计算全监督的损失函数,通过损失的大小来判断 WSOD 生成的伪监督信息的准确性,并以此将伪监督标签分成标注集和未标注集,然后使用半监督物体检测框架(Liu 等,2021b)来学习检测器。检测结果如图 2 所示。

图 2 SoS-WSOD(Sui 等,2021)在 MS-COCO 数据集上的检测结果

Fig. 2 Detecton results by SoS-WSOD on MS-COCO dataset

((a) ground truth; (b) results of stage 1); (c) results of stage 2); (d) results of stage 3))

1.3.3　基于迁移学习的弱监督检测方法

由于弱监督数据集缺少物体级的监督信息，使得弱监督检测模型和全监督检测模型的性能仍有较大差距。一些工作尝试引入全监督辅助数据集，在保证该数据集的类别与弱监督数据集的类别无重合的情况下，尝试设计对应的迁移学习算法，使得网络可以借助辅助数据集上的知识去完成弱监督目标数据集上的任务。Li 等人(2019b)利用辅助数据集上的物体级信息，提出了一个混合监督模型。首先利用辅助数据集和目标数据集一个二分类网络，同时利用梯度反转的操作来强化网络在目标数据集上的泛化能力。然后用该检测器去除弱监督网络中低质量的候选框。Zhong 等人(2020)设计了一个基于迁移学习的迭代训练框架，交替训练基于辅助数据集的全监督网络和基于弱监督目标数据集的弱监督网络，同时，全监督网络为弱监督网络的训练提供候选框并改善定位性能，而弱监督网络则挖掘辅助数据集中更多的弱监督类别物体，使得全监督网络学到更多丰富目标域的知识。Uijlings 等人(2018)探索了不同层次语义下的物体级知识的迁移效果，该工作为全监督物体检测网络添加多个检测头，每一个检测头学习辅助数据集中不同层次的类别，然后探究不同层次的源知识是否能对弱监督数据集上的相关类别产生影响。Lee 等人(2018)提出了一个通用的类别无关的边界框回归器，该回归器首先在辅助数据集上训练至收敛，然后可对弱监督物体检测模型的原始输出做修正，以进一步提高弱监督模型的预测精度。Dong 等人(2021)从改善模型定位性能的角度出发，利用全监督辅助数据集训练了一个边界框校正器，该校正器可学得辅助数据集上类别无关的定位知识，然后通过知识蒸馏的方式，在弱监督检测网络训练时提供额外的监督信息，以提升弱监督检测网络的定位性能。Cao 等人(2021)提出一种双平均教师的混合监督模型，通过多个检测头间 EMA (exponential moving average)更新的方式来整合辅助数据集和目标数据集之前的领域差异。另外，还提出了一种新颖的语义图卷积网络来解决非重叠类别转移，该网络促进了相关类别之间语义特征的聚合。Liu 等人(2021a)提出了 TraMaS(transferving mask)迁移学习框架，在辅助数据集上学习分割先验，并用其对网络提取到的特征图进行增强，以使检测头学习到更强大的特征表示。同时，利用辅助数据集中的物体训练了一个类别无关的示例相似度判别器，对迁移到弱监督数据集上候选框的质量进行判断。

1.3.4　分割—检测联合训练方法

一些工作探索利用语义分割任务的特性来增强弱监督物体检测模型的精度。Shen 等人(2019)提出一个弱监督物体检测与语义分割联合训练框架 WS-JDS (weakly supervised joint detection and segmentation)。在该模型中，有两个分支分别完成检测任务和分割任务。同时，提出 CGL (cyclic guidance learning)训练方式，结合检测和分割输出结果，并生成质量更优的伪标签来优化网络，采用生成和对抗策略来生成更加准确的分割结果和检测结果。

1.4　弱监督物体定位模型

弱监督物体定位是一个与弱监督物体检测相似但又有所不同的任务。二者的目的都是利用仅提供类别标签的图像数据集实现对物体边界框的检测定位。不同的是，弱监督物体检测中每幅图像可能会有不同种类和不同数目的物体，而弱监督物体定位进一步要求每幅图像只允许有一个种类的一个物体。因此，弱监督物体定位可视为弱监督物体检测任务的一个较为简单的特例。

当前大多数弱监督物体定位方法是基于类别激活图 CAM (Zhou 等，2016)的。通常将分类卷积网络末端的分类层和卷积层之间的全连接层替换为全局平均池化层 (global average pooling, GAP)，以此将特征图映射为一个 1 维的特征向量，然后通过分类层得到该输入图像的图像级类别预测分数。充分训练后，可由分类层每一个类别对应的权重与最后一层卷积层输出的特征图计算加权平均，即得到了该类别的激活图。该激活图中热点位置则表示该位置的物体最有可能对应着预测类别。因此，可通过设定阈值的方式，选择激活值大于阈值的区域，生成检测框。

1.5　弱监督物体定位算法

与弱监督物体检测类似，在 CAM 中判别区域的激活值往往大于整个物体本身，因此网络预测的检测框往往只关注于物体的一部分。为了解决这个问题，Diba 等人(2017)提出了一个端到端的级联卷积网络 WCCN (weak cascaded convolutional networks)，该网络包含 3 个级联阶段：第 1 阶段是 CAM 网络，生成 CAM 激活图和初始的候选框；第 2 阶段是一个分割网络，使用 CAM 来训练分割网络并

筛选高质量的候选框;第 3 阶段是一个 MIL 网络,对第 2 阶段提取的多个候选框进行多示例学习,以学习到每个候选框的类别信息。Zhang 等人(2018a)引入两个并行分类器并采用对抗性补充学习的方式,提出了 AcoL(adversarial complementary learning)模型。该模型首先利用第 1 个分类器定位判别区域,并消除该区域对应的特征图,然后输入到第 2 个网络中,以强制网络关注物体的其他区域。最后,AcoL 结合两个分类器的 CAM 来生成检测框。Selvaraju 等人(2020)提出了 Grad-CAM,该模型可使用任何目标形式的梯度,利用这些梯度信息流入到卷积层以生成粗略的定位图,突出显示图像中的重要区域以预测目标。该方法可以很方便地扩展到目前的任意一个训练完好的 CNN 模型中,通过合并引导反向传播的结果,Grad-CAM 可以做到更细粒度的可视化分析,并尝试解释了模型将原图分类到某一类的原因。此外,Grad-CAM 可以推广到其他视觉任务的分析中,例如图像分类、图像描述和视觉问答等任务。Durand 等人(2016)提出了 WILDCAT(weakly supervised learning of deep convolutional neural networks)模型,该模型可以同时完成弱监督物体检测和分割任务。在该方法中,作者将单一 CAM 预测修改为多 CAM 热度图预测,以使其可以识别多个局部区域,同时,该工作提出了 WILDCAT 池化来整合每一个 CAM 的特征,供之后的分类和分割分支使用。Kim 等人(2017)提出了 TP-WSL(two-phase learning for weakly supervised object localization)模型,该模型包含两个阶段。在第 1 阶段,图像输入到全卷积网络,通过 CAM 方法生成判别区域的热度图。在第 2 阶段,网络通过预测反馈去抑制最显著部分的激活,然后进行第 2 次学习以找到下一个最重要部分的区域。Zhang 等人(2018b)提出了 SPG(self-produced guidance)方法,利用多阶段训练的方式,结合注意力生成了 SPG 掩膜,以将需要关注的前景物体从背景中分离。具体而言,SPG 由骨干网络、SPG-A、SPG-B 和 SPG-C 共 4 部分组成。首先将图像输入到骨干网络提取特征图,输入到 SPG-A 进行分类,然后 SPG-B 根据 SPG-A 的分类特征预测 SPG 掩膜,最后 SPG-C 模块将 SPG 图作为额外的监督信息强化注意力图。

1.6 数据集及模型性能评估

弱监督物体检测任务的常用评价数据集有 PASCAL VOC(pattern analysis, statistical modeling and computational learning visual object classes)(Everingham 等,2015)物体检测数据集,MS-COCO(Microsoft common objects in context)(Lin 等,2014)物体检测数据集,ILSVRC(imagenet large scale visual recognition challenge)(Russakovsky 等,2015)物体检测数据集及 CUB-200-2011(Welinder 等,2010)数据集。PASCAL VOC 数据集是一个多任务数据集,包含 20 个类别的待检测物体。它包含多个版本,目前研究人员通常使用 PASCAL VOC 2007 及 2012 这两个版本进行模型的训练与测试。在 PASCAL VOC 2007 中,有 2 501 幅训练图像,2 510 幅验证图像,及 4 952 幅测试图像;在 PASCAL VOC 2012 中,有 5 717 幅训练图像,5 823 幅验证图像,及 10 991 幅测试图像。MS-COCO 是一个更大规模的多任务数据集,在检测任务标注中,该数据集包含 80 个类别的待检测物体,有 118 287 幅训练图像及 5 000 幅测试图像。ILSVRC 是一个大型的多任务数据集,包含多个版本。通常人们采用 ILSVRC 2013 物体检测数据集做弱监督物体检测任务,该数据集包含 200 个类别的待检测物体,约 40 万幅训练图像,2 万幅验证图像及 4 万幅测试图像;用 ILSVRC 2016 数据集做弱监督物体定位任务,该数据集包含 1 000 个类别的待定位物体,约 120 万幅训练图像和 5 000 幅测试图像。CUB-200-2011 数据集包含 200 个不同品种的鸟类,该数据集包含 5 994 幅训练图像和 5 794 幅验证图像。

对于弱监督物体检测问题,通常采用 mAP(mean average precision)及 CorLoc(correct localization)评价模型性能;对于弱监督物体定位问题,通常采用 Top error 评价模型性能。表 1 展示了部分弱监督物体检测方法在 PASCAL VOC 2007 和 2012 上的性能表现,目前性能综合表现最佳的方法是 SoS-WSOD,它在 PASCAL VOC 2007 和 2012 上的 mAP 分别为 64.4% 和 61.9%,在性能上相比最经典的 WSDDN 已有较大提升,但与全监督方法(例如 Faster R-CNN)仍有差距,这也说明了弱监督设置下的物体检测方法仍有研究的意义和空间。

2 弱监督语义分割与实例分割

2.1 典型问题设置

图像分割技术作为其他图像处理方法的基础,

表 1　不同 WSOD 方法在 PASCAL VOC 2007/2012 上的性能表现

Table 1　Comparison of WSOD methods on PASCAL VOC 2007/2012

/%

方法	VOC 07		VOC 12	
	mAP	CorLoc	mAP	CorLoc
WSDDN(Bilen 和 Vedaldi,2016)	39.3	58.0	–	–
ContextLocNet(Kantorov 等,2016)	36.3	55.1	35.3	54.8
OICR(Tang 等,2017)	42.0	61.2	38.2	63.5
WS-JDS(Shen 等,2019)	45.6	64.5	39.1	63.5
PCL(Tang 等,2020)	45.8	63.0	41.6	65.0
Kosugi 等人(2019)	47.6	66.7	43.4	66.7
SDCN(Li 等,2019a)	50.2	68.6	43.5	67.9
MSD(Li 等,2019b)	51.1	66.8	43.4	–
OICR + UBBR(Lee 等,2018)	52.0	–	–	–
W2F(Zhang 等,2018c)	52.4	70.3	47.8	69.4
C-MIDN(Yan 等,2019)	52.6	68.7	50.2	71.2
SLV(Chen 等,2020)	53.5	71.0	49.2	69.2
WSOD2(Zeng 等,2019)	53.6	–	47.2	–
GradingNet-C-MIL(Jia 等,2021)	54.3	72.1	50.5	71.9
Yang 等人(2019)	54.5	68.0	49.5	69.5
Ren 等人(2020)	54.9	68.8	52.1	70.9
IM-CFB(Yin 等,2021)	55.8	72.2	49.4	69.6
LBBA(Dong 等,2021)	56.6	72.5	55.4	73.7
CaT(Cao 等,2021)	59.2	75.9	–	–
CASD(Huang 等,2020)	56.8	–	53.6	–
Zhong 等人(2020)	60.2	75.2	–	–
TraMaS(Liu 等,2021a)	62.9	77.7	–	–
SoS-WSOD(Sui 等,2021)	64.4	–	61.9	–

注:“–”表示原文献未在该数据集进行实验或未给出该指标;SDCN 为 segmentation-detection collaborative network,MSD 为 mixed supervised detection,UBBR 为 universal bounding box regression,IM-CFB 为 instance mining with class feature banks,LBBA 为 learning bounding box adjusters,CaT 为 category transfer,PCL 为 proposal cluster learning,C-MIDN 为 coupled multiple instance detection network。

一直是计算机视觉研究领域的热点与难点,也在人机交互、自动驾驶等领域内发挥着重要的作用。早期,图像分割方法往往存在着精度不足的问题,随着人工智能技术的兴起,通过引入卷积神经网络 CNN,并使用大量的像素级标注数据,基于深度学习的语义分割与实例分割方法已经能获得很高的可靠性。

语义分割方面,Long 等人(2015)通过提出全卷积网络(fully convolutional network, FCN),首次尝试采用深度 CNN 进行语义分割。随后,研究者们提出了许多变体以提高 FCN 的性能。Chen 等人(2018)通过使用空洞卷积来扩大特征图的感受野,提出了 DeepLab 模型。此外考虑到图像内的长程连接,DeepLab 还使用 Krähenbühl 和 Koltun(2011)提出的稠密条件随机场算法 DenseCRF(dense conditional random field)作为后处理来细化最终的预测结果。Zhao 等人(2017)提出了 PSPNet(pyramid scene parsing network)来利用金字塔池化多尺度特征,同时利用全局上下文信息来提高基于 FCN 方法的分割性能。Wang 等人(2021a)进一步探索了多尺度特征的思想,提出了 HRNetV2 架构,其目的是在低分辨率特征表示的帮助下提升高分辨率特征表示能力。

实例分割方面,He 等人(2020)在 Faster R-CNN 目标检测框架的基础上,添加了用于实例分割的分支,通过多任务学习来完成实例分割任务,同时基于双线性插值引入了 ROI Align 以解决 ROI Pooling 导致的特征图与原始图像无法对齐的问题。Bolya 等人(2022)提出的 YOLACT(you only look at coefficients)在一阶段目标检测模型基础上加入了分割预测分支,并将实例分割任务拆分为并行的两个子任务:通过一个原型网络为每幅图像生成多个原型掩膜;为每个实例预测多个掩膜的线性组合系数,最终通过线性组合得到实例分割结果。Xie 等人(2020)提出的 PolarMask 基于极坐标对分割结果进行建模,把实例分割问题转化为实例中心点分类问题与密集距离回归问题,能够在不预测物体边界框的情况下得到每个实例的分割结果。Tian 等人(2020)提出的 CondInst 使用了全卷积网络解决实例分割问题,不需要使用 ROI Align 来进行特征对齐,同时使用了动态卷积,结合原型掩膜与相对坐标图生成每个实例对应的分割结果。

尽管这些基于深度学习的语义分割或实例分割方法已经能够取得不错的分割性能,但是这些方法所使用像素级分割标注有着很高的成本。以 MS-COCO 数据集(Lin 等,2014)为例,如果只标注每幅图像包含的物体类别,仅需 8 个工作人员标注 17 751 h 就能完成,而每标注 1 000 个分割对象则

需要超过22 h,标注整个数据集中约2 500 000个实例共花费约55 000 h。

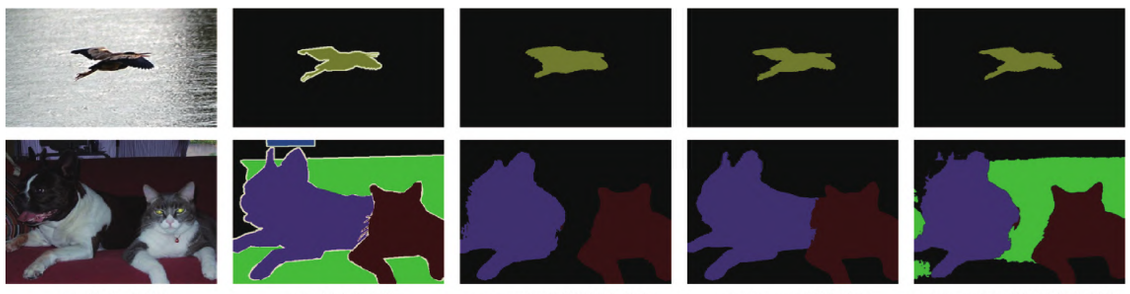
为了解决像素级标注成本过高的问题,近几年涌现出许多针对图像分割的弱监督方法。与使用像素级标注数据的全监督图像分割方法不同,弱监督图像分割方法往往采用成本低廉的物体边界框标注、图像级的物体类别标注数据以及线标签或者点标注作为监督,如图3所示。其中边界框标注会给出物体在图像中的位置框及物体所属类别,而物体类别标注则是指给出图像中所包含的物体类别的图像级标注。Lin等人(2014)的研究表明,为一幅图像标注像素级的标签的时间成本约是标注物体边界框的15倍,标注物体类别标签的60倍。因此,使用边界框标注或图像级别的类别标注能显著降低数据标注的成本,与此同时,如果能够在使用这些成本低廉的标注的同时尽可能得到精确的图像分割结果,将是非常有意义的,也能够在自动驾驶、机器人等行业内发挥重要的作用。图4展示了一些典型的弱监督分割方法的分割结果,可以看到仅使用图像级标注得到的分割结果已相对不错,而如果使用成本更高的线标注或边界框标注,所得到的分割结果精度也会更高。选用哪种标注形式本质上是在标注成本与分割精度上进行权衡,如在自动驾驶等更加强调系统的安全性、对分割精度要求较高的领域,可以尽量使用边界框标注或像素级标注来保证分割的精度;而在人机交互等对分割精度要求较低的领域,可以使用图像级标注或者点/线标注来节省数据标注的成本。

(a) 图像级标注　(b) 点标注　(c) 线标注　(d) 边界框标注

图3　弱监督分割标注示意图

Fig. 3　Examples of weak annotations for segmentation

((a) image-level annotation; (b) points annotation; (c) scribble annotation; (d) bounding box annotation)



(a) 输入图像　(b) 真实标签　(c) EPS(图像级标注)　(d) SPML(线标注)　(e) BBAM(边界框标注)

图4　典型弱监督分割方法结果示意图

Fig. 4　Segmentation results by representative weakly supervised segmentation methods((a) input images; (b) ground truths; (c) image-level annotation; (d) scribble annotation; (e) bounding box annotation)

2.2　基于边界框标注的弱监督语义分割

由于边界框标注相比图像级类别标注额外包含了物体的数量、大小和位置等对于图像分割而言比较关键的语义信息,基于边界框标注的弱监督语义分割方法往往能够获得更好的分割效果。目前该领域的主流方法使用的是两阶段处理的思路,首先,通过对物体边界框标注进行预处理,包括使用传统分割方法或者深度学习方法,来生成相对准确的伪标签;然后,使用已生成的伪标签对分割模型进行训练,通过迭代式训练或者考虑边界框标注先验等方式使分割模型最终取得较好的分割效果。

早期,Dai等人(2015)提出了BoxSup方法,这一方法首先使用一些提取候选区域的算法,如MCG(Arbeláez等,2014)等来提取一系列候选区域,然后从中选择一个与边界框有最大重叠部分的候选区域作为初始阶段的伪标签,然后通过迭代训练的方式

逐渐优化伪标签。在每轮迭代中,首先使用伪标签训练以 VGG（Simonyan 和 Zisserman,2015）作为骨干网络的 DeepLab-CRF（Chen 等,2015）分割模型,然后再使用本轮迭代中训练的分割模型进行预测,从每个边界框得分最高的几个候选区域中随机选出一个作为新的伪标签参与下一轮的迭代过程,这种引入随机性的迭代训练的方式,能够保证网络在不陷入相对较差的局部最优解的同时,不断提高伪标签的准确率。Papandreou 等人（2015）则提出了 WSSL（weakly-and semi-supervised learning）方法,将 EM 算法与弱监督语义分割问题相结合,并在弱监督与半监督配置下探索了 3 种基于边界框标注的弱监督语义分割方法:Bbox-Rect、Bbox-Seg 与 Bbox-EM-Fixed。其中,Bbox-Rect 直接将边界框中的所有像素当做正样本,如果某个像素同时处于多个边界框内部,则将其归为最小边界框所属的类别;Bbox-Seg 先对边界框标注使用 DenseCRF 得到伪标签,并使用交叉验证的方法对 DenseCRF 中的超参数进行选择,尽可能提高伪标签的精度,然后再使用伪标签训练同样以 VGG 作为骨干网络的 DeepLab-LargeFOV 分割模型,也取得了较好的分割效果;Bbox-EM-Fixed 使用 EM 算法,通过迭代的方式细化分割的预测结果。实验表明,在仅使用边界框标注的条件下,Bbox-Seg 的性能最好。而在使用少量精细的像素级标注,其余标注仍使用边界框标注的情况下,Bbox-EM-Fixed 方法的性能更高。Khoreva 等人（2017）提出了 SDI（simply does it）方法,他们首先基于 GrabCut（Rother 等,2004）提出了一种变体 GrabCut +,使用 HED（holistically-nested edge detection）（Xie 和 Tu,2015）预测的边缘来代替传统的 RGB 二元项,得到了质量更好的分割结果。考虑到单独使用 MCG 或 GrabCut + 生成伪标签的精度不高,他们创新性地将两种方法结合起来,选取 MCG 与 GrabCut + 所得候选区域的交集作为伪标签,即 M∩G +,之后使用生成的伪标签训练 DeepLab-LargeFOV 分割模型,然后再使用 DenseCRF 作为后处理提高分割的精度。同时,他们还尝试使用 COCO 数据集作为辅助,先在 COCO 数据集上使用同样的 M∩G + 方法生成伪标签,并对模型进行预训练,然后再使用 PASCAL VOC 2012 数据集（Everingham 等,2015）上生成的伪标签训练该模型。除此之外,他们还尝试使用了 DeepLabv2-ResNet-101 作为分割模型。这些实验都得到了不错的结果,也很有借鉴意义。Song 等人（2019）提出了 BCM（box-driven class-wise masking）方法,这一方法为每一类别的物体单独学习一个掩膜来帮助模型消除无关区域对预测分割结果的影响,同时为模型预测提供各类物体在形状与位置方面的提示。同时,考虑到物体在边界框中的填充率在获取伪标签过程中也有着指导意义,Song 等人（2019）还提出了一种自适应的填充率指导损失,进一步考虑到同类物体因为形状与姿态等原因在边界框中的填充率存在着差异,Song 等人（2019）又使用聚类的方式将每类物体细分为几个子类。最终,这一方法首先通过 DenseCRF 得到预选标签,用来计算物体在边界框内的平均填充率,并且输入一个基于全卷积网络的分割网络。然后将分割网络的输出中每个类别的特征图乘上对应类别的掩膜,再根据所属类别的平均填充率约束最后的网络输出。

Xie 等人（2021）提出了基于学习的伪标签生成器 LPG（learned proposal generator）方法,用来为边界框标注提取精度更高的候选区域。使用 COCO 数据集作为辅助数据集,使用其中与 PASCAL VOC 2012 数据集中不相交的 60 类物体来学习一个类别无关的能够用在任意给出边界框标注的数据集上提取候选区域的模型。引入了两层优化模型来对分割模型与候选区域提取器进行训练,其中训练分割模型定义为下层问题,训练基于边界框的候选区域提取器定义为上层问题,使用 EM 算法将 LPG 与分割模型联合训练,在多个阶段的训练过程中不断优化伪标签,并用于分割模型的训练,最终取得了较好的结果。Oh 等人（2021）提出了背景感知池化 BAP（background-aware pooling）和噪声感知损失 NAL（noise-aware loss）,其中 BAP 方法使用注意力图来区分边界框的前景与背景部分,通过聚合前景区域的特征并去除背景区域,获得更为精确的类激活图 CAM,再结合注意力图与 CAM 使用 DenseCRF 后处理生成伪标签;NAL 方法通过自适应地使用网络输出、中间层特征以及分类器的权重计算交叉熵损失,使用中间层特征的相似度产生第 2 种伪标签,并在两种伪标签不同的区域使用特征相似度对交叉熵损失加权来抑制伪标签中错误信息的影响。Lee 等人（2021a）则提出了 BBAM（bounding box attribution map）方法,针对伪标签的生成方式做出了创新,主

要思路是先使用边界框标注数据训练目标检测模型,然后使用训练好的目标检测模型获取边界框的属性图,即能使目标检测模型与原输入图像相同结果的最小图像区域,这些区域就表明了边界框中的前景区域,基于这一区域使用 DenseCRF 获取伪标签,用于第 2 阶段对分割模型的训练。

基于边界框标注的弱监督语义分割方法的分割效果越来越好,但是这其中也存在着一些问题,比如:第 1 阶段中生成的伪标签的精度还不够高,其中包含的错误分类信息往往导致第 2 阶段中分割模型训练的效果不够好,难以使用性能更好的模型来做弱监督方法的分割模型。未来该领域可能更会聚焦在通过挖掘边界框标注的先验信息,抑制伪标签中错误信息的影响以及将分割结果的先验信息与网络结构相融合,从而应用性能更好的分割网络来进一步提升基于边界框的弱监督语义分割方法的性能。

2.3 基于图像级标注的弱监督语义分割

相比物体边界框标注,图像级标注的成本更低,相应的这种标注中所包含的信息也更少,仅包含图像中所含有的物体类别信息,因此基于图像级标注的弱监督语义分割相比基于边界框的弱监督语义分割问题更具挑战性。

Pinheiro 和 Collobert(2015)提出了一种从图像类别标签获得像素级分割结果的方法,该方法认为弱监督语义分割问题的本质是一种多示例学习问题,每幅图像就是一个包,图像中的每个像素是这个包中的一个示例。如果这个图像中包含某类物体,则至少含有该类物体的一个像素,即正示例,而如果图像不包含某类物体,则每个像素一定不属于该类物体,即负示例,为解决这一问题,该方法首先将图像经过一个 OverFeat (Sermanet 等,2014)卷积神经网络,得到分数图,再对这个分数图通过一个聚合层聚合,再用物体类别标注来训练模型,使得分类正确的像素能有更高的权重,同时还会使用 3 种平滑策略来对输出的预测结果进行平滑,实验表明,使用 LSE (log-sum-exp)进行聚合、基于 MCG 方法进行平滑具有最好的分割效果。Zhou 等人(2016)提出了类激活图 CAM 方法,指出在诸如 ImageNet 数据集上训练之后的分类模型在物体的可区分部分往往具有更大的响应值,其使用了全局平均池化 GAP 来将网络输出的每个类别的特征图聚合为一个标量,并使用图像级标注对模型进行训练,从而通过多分类任务的梯度反向传播来使网络分割层的不同通道具有定位物体的可分辨区域的能力,这一方法也在后续许多方法中得以应用及发展。

以 CAM 方法为基础,一些方法(Hou 等,2018;Mai 等,2020;Wei 等,2017a)在迭代训练的过程中每次将网络分类得到的响应值大的区域先删除,然后对剩余区域再次进行多类学习,将新得到的响应值大的区域加入到分割预测结果中,直到分割预测结果不再变化为止。其中 Wei 等人(2017a)提出的 PSL (prohibitive segmentation learning)方法引入了对抗擦除的思想,对网络进行多阶段训练,将输入图像中当前阶段网络所关注的区域擦除以作为下一个阶段网络的输入,从而使这些网络关注到物体的不同区域,避免了 CAM 仅关注图像中最具判别性区域的问题。Hou 等人(2018)提出的 SeeNet 对于给定的初始注意力图,将输入图像分为 3 个区域:内部注意力区域、外部的背景区域与交界部分的潜在区域,作者设计了一个包含 3 个分支的自擦除网络,通过引入背景先验约束了注意力区域的范围,解决了 PSL 容易导致 CAM 扩散到背景区域的问题。

Mai 等人(2020)提出了 EIL (erasing integrated learning)方法,将对抗擦除与 CNN 网络相结合,通过共享参数减少了计算量,同时还提出可以在 CNN 的不同层中插入多个 EIL 模块挖掘物体关键区域的多尺度特征。在 CAM 方法的基础上,一些方法(Lee 等,2021c;Wei 等,2017b;Yu 等,2019b)将显著性检测任务与弱监督语义分割任务相结合。Wei 等人(2017b)使用 CAM 定位目标、显著性检测选取背景来解决弱监督语义分割问题。提出的 STC (simple to complex)方法考虑了 3 个阶段的渐进式学习方式,第 1 阶段,以显著性检测模型预测的显著性图为约束,训练网络学习只含一类物体的图像的前景与背景分类;第 2 阶段,训练网络预测这些图像的语义分割结果;第 3 阶段,训练网络预测复杂图像的语义分割结果,最终使网络具有对复杂图像进行语义分割的能力。Yu 等人(2019b)提出的 SSNet (saliency and segmentation network)考虑到之前的弱监督语义分割方法往往直接利用训练好的显著性检测模型,而没有明确建模显著性检测与弱监督语义分割任务之间的关系,提出了一个多任务学习框架,设计了一个显著性聚合模块,使用一个端到端的网络来联合解决显著性检测任务与弱监督语义分割任

务。Lee 等人(2021c)提出的 EPS (explicit pseudo-pixel supervision)方法同样结合了 CAM 与显著性图,考虑到 CAM 可以区分物体,但是边界不够清晰,而显著性图的边界足够清晰但是无法区分具体的物体类别,EPS 引入了显著性损失来使用显著性图为 CAM 中的前景部分获取准确的边界信息。

Ahn 等人(2019)以及 Ahn 和 Kwak(2018)在 CAM 方法的基础上通过训练一个额外的网络来捕捉像素间的相似度,然后通过随机游走来精细化 CAM,使其能够覆盖整个物体,提高生成的伪标签质量。Ahn 和 Kwak(2018)提出的 PSA (pixel-level semantic affinity)方法,设计了亲和度网络 AffinityNet 来预测相邻像素的语义相似度,在 CAM 的可信区域中抽取相邻像素作为训练数据,然后再以训练好的亲和度预测网络精细化 CAM,提升 CAM 的质量,生成了更为准确的分割伪标签。随后,Ahn 等人(2019) 又设计了一个具有两个分支的 IRNet,能够同时解决基于图像级物体类别标注的弱监督语义分割与实例分割问题。这一网络共有两个分支,其中一个分支用来预测物体的像素位移场,并以此为基础为每个实例生成 CAM;而另一个分支用来预测物体的边界,得到像素对间的相似度关系。通过对两个分支所得结果进行融合和处理得到最终的分割结果。同时该方法还引入了随机游走,根据亲和度矩阵对所得预测结果进行传播来优化网络预测的分割结果。

Fan 等人(2020)考虑到 CAM 往往只关注物体的局部区域,导致物体区域估计不完整的问题,设计了 CIAN(cross-image affinity net)网络框架,在含有同类物体的不同图像间对每个像素的特征建立关系矩阵,从而挖掘不同图像中同类物体的语义相似性,并使用这个相似性来优化网络提取到的特征,最终提高 CAM 对物体区域的识别能力。Li 等(2021a)在此基础上使用了图神经网络,不只针对两幅含有同类物体的图像,而是针对多幅含有同类物体的图像挖掘其中包含的语义相似性。在迭代的过程中,不断对含有同类物体的一组图像中的图像进行聚合,使用同组内的其他图像特征来提高图像特征的表示能力,同时还提出了图神经网络中的 dropout 方法,使得网络提高对只有少量样本的物体类别的学习。Wang 等人(2020)提出的 SEAM (self-supervised equivariant attention mechanism)方法考虑了图像的多尺度变换或者各种仿射变换对 CAM 目标定位能力的增强,同时引入了等变正则化项约束分割模型,即不同尺寸的图像作为网络输入得到的分割结果应具有一致性。这一约束使预测的结果能够更好地覆盖到整个物体上,而不仅仅只是覆盖住一些判别性较强的小部分区域。Zhang 等人(2020)提出的 CONTA (context adjustment)方法考虑到不同类别间的物体以及背景与前景物体间可能存在关联关系,会误导图像级的分类模型学习到错误的像素与物体类别间的关系,因此将因果模型引入了弱监督语义分割问题。通过因果干预的方法来消除不属于物体的上下文特征对 CAM 的影响,从而提高了伪标签的精度。

除了基于 CAM 的方法外,Pathak 等人(2015)提出的使用多示例学习的方法也是很有借鉴意义的。这一方法也使用 FCN 作为分割网络,考虑到具有更大输出值的像素更可能属于图像的前景部分,该方法关注网络输出中的最大值部分,并仅对这些像素计算损失函数,从而对模型进行更新。

现有基于图像级标注的弱监督语义分割方法主要是对 CAM 方法的继承与发展,包括迭代擦除、学习像素间相似度以及使用显著性图来约束网络等。CAM 方法的问题在于其往往只关注到对图像分类起到重要作用的像素区域,而这些像素常常是物体像素的一个子集,难以覆盖整个物体。未来该领域仍会聚焦如何优化 CAM,使其覆盖整个物体以及使用一些如多示例学习或半监督的方法来使伪标签的生成更加精确。

2.4　基于线/点标注的弱监督语义分割

除了基于边界框标注以及图像级标注的弱监督语义分割方法外,研究者们还提出了基于点标注或线标注的弱监督语义分割方法,其中线标注是指在物体内部画一条连续的线并表明物体类别作为监督信息;而点标注包括使用极点,即物体掩膜的最上、最下、最左和最右 4 个点,或者使用物体内任意一点作为监督信息。线标注包含了图像中物体的位置、类别、数量以及不精确的形状信息;极点标注与物体的边界框标注类似,包含有图像中物体的位置、类别、数量以及大小信息;任意点标注则仅包含图像中的类别、数量以及位置信息。因此,从解决弱监督语义分割问题的难易程度来看,极点标注最简单,线标注其次,而包含信息最少的任意点标注的难度最高。

2.4.1 线标注

Lin 等人(2016)提出了 ScribbleSup 方法,首先基于线标注生成超像素,然后再基于 GrabCut 算法通过能量函数来对超像素进行传播,并且采用了 EM 算法的思想,交替更新分割模型以及训练分割模型的标签,直至模型收敛。Vernaza 和 Chandraker (2017)提出的 RAWKS (random-walk weakly-supervised segmentation)方法通过训练一个一致性网络将稀疏的线标注传播到物体区域,提出了使用一个特定的概率模型用于稀疏标签传播,且这一模型在进行语义边缘检测训练时是可微的,即随机游走,极大地提高了使用线标注获取伪标签的精度。Ke 等人(2021)提出的 SPML (semi-supervised pixel-wise metric learning)使用 SegSort (Hwang 等,2019)作为骨干网络,同时使用 HED 轮廓检测器作为额外的监督信息,引入了一个像素级的特征一致性损失,使得具有相近语义的像素映射到特征空间后具有较近的距离,而属于不同语义的像素映射到特征空间后相互远离。现有方法通常约束颜色相近像素具有相似的特征,可能会误导网络的学习。针对这一问题,Zhang 等人(2021a)提出了动态特征正则化(dynamic feature regularization, DFR)损失,仅在一个小的窗口内约束颜色相近的像素对特征相似,同时还设计了特征一致性模块,通过选取模型预测置信度高的像素级特征作为监督,要求窗口与其属于同类别的像素具有与其相似的特征,DFR 损失与特征一致性模块均可以直接应用到其他弱监督语义分割方法中。

2.4.2 点标注

Qian 等人(2019)首次提出采用点标注信息进行场景解析。Papadopoulos 等人(2017)提出的点标注方法使用物体的极点作为监督信息,使用边缘检测器找到物体的轮廓,通过最大化概率的方式在极点间寻找一条最优路径,将这条路径作为物体的轮廓输入 GrabCut 算法,得到伪标签。之后再使用 DeepLab 作为分割模型训练伪标签,可以得到比单纯使用边界框标注进行 GrabCut 的伪标签训练模型更好的效果。Bearman 等人(2016)使用物体上的任意点作为监督信息,并结合预训练的通用物体对象性(objectness)检测器为模型设计了对象性先验损失,要求对象性较高的区域存在物体的概率更高。Li 等人(2021b)为全景分割设计了一个全卷积网络框架,即 Panoptic FCN,这一网络将每个实例编码为一个卷积核,并用来生成对应的预测结果,从而使网络可以以同样的方式预测可数与不可数物体。通过使用基于点标注的物体检测器中的定位头来区分前景与背景区域,Li 等人(2021b)选取物体上的随机个点作为监督信息,并且将这些点集连接成一条曲线,使用曲线内部的区域作为训练分割模型的伪标签,从而通过点标注获取到了较为准确的物体形状信息,取得了较好的分割性能。

相比于边界框标注以及图像级标注,线标注以及点标注所包含的信息往往并不直观,这就更需要研究者们进一步深入挖掘这些标注中所包含的深层语义信息,例如使用线标注结合提取候选区域的方法将监督信息向没有监督的区域进行传播以及将点标注之间相互连接,以得到物体大致的形状信息等。

2.5 弱监督实例分割

相比语义分割,实例分割不仅要区分出每个像素所属的物体类别,还要区分出每个像素具体属于哪个物体,因此实例分割也更具有挑战性。在弱监督实例分割任务中,常用的伪标签类型主要是物体的边界框标注以及图像级的物体类别标注。

2.5.1 基于图像级标注的弱监督实例分割

Zhou 等人(2018)观察到类别响应图中的峰值与物体实例有着很强的关联性,因此提出了峰值响应图 PRM(peak response map)方法,将类峰值响应作为图像级标注,用来训练卷积神经网络,使其具备实例分割的能力。这一方法在全卷积网络的最顶层加入了峰值激活层,并证明了峰值激活能够刺激类响应图中出现相应的峰值,同时提出了峰值的反向传播进一步细化峰值响应。这一方法以自上而下的方式将每个类峰值响应定位到最相关的区域,从而生成详细的实例感知视觉线索,即峰值响应图,并将其与类别响应图以及现有的生成候选物体的方法相结合来完成弱监督实例分割任务。Laradji 等人(2019)在 PRM 方法的基础上提出了 WISE (weakly-supervised instance segmentation)方法,这一方法首先使用图像级类别标注训练了一个带有峰值激活层的图像分类器,以得到图像的 PRM,然后再使用 MCG 等提取图像候选区,并以 PRM 作为选择候选区域的权重来得到训练实例分割模型的伪标签,最后以 Mask R-CNN 作为实例分割模型进行训练,由于使用了 MCG 方法提取图像中的候选区域,这一方

法生成的伪标签比 PRM 方法所得到的更加准确，从而提升了弱监督实例分割的性能。由于 WISE 生成伪标签时没有考虑同类别物体间伪标签应该具有的一致性，Arun 等人（2020）提出了 LACI（learning annotation consistent instances）方法，为生成物体伪标签设计了条件网络，使得在为每个实例生成伪标签时，都与其类别保持一致性。论文对伪标签的不确定性使用了条件分布进行建模，其中包含 3 项：用来为每个候选分割预测分数的类别相关一元项；用来使分割结果更完整地覆盖物体的物体边缘二元项；用来约束所得的属于同一物体类别的候选分割具有一致性的标注一致性高阶项。通过组合这 3 项，能够保证条件分布提供的候选分割的准确性与一致性。

Ge 等人（2019）设计了 Label-PENet，这一网络除用来提取特征的骨干网络外，共包含 4 个级联的模块：多标签分类模块、物体检测模块、实例精细化模块以及实例分割模块，并通过这 4 个模块的级联，对物体包含的实例语义信息进行反复挖掘与优化。网络设计了两阶段的训练方法，在第 1 阶段，将骨干网络固定，首先使用图像级标注训练多标签分类模块，结合候选区域矫正来得到实例的位置以及掩膜伪标签，然后再使用得到的实例位置伪标签训练目标检测模块，并再次结合候选区域矫正来优化实例分割伪标签，之后结合前两个步骤得到的边界框与实例分割伪标签训练 Mask R-CNN 实例分割模型以得到最终的伪标签，最后使用该伪标签再次训练实例分割模型；而在第 2 阶段，考虑到第 1 阶段的训练容易使模型陷入局部最优值，因此选择使用了循环学习的方式，又进行了反向验证，按第 1 阶段训练顺序的倒序对 4 个模块进行训练，对物体实例信息进行了反复的挖掘。

2.5.2　基于边界框标注的弱监督实例分割

Hsu 等人（2019）提出的 BBTP（bounding box tightness prior）方法将弱监督实例分割问题视为多示例学习任务，考虑到边界框是包含物体的最小矩形，认为边界框中的每一行或列都应该至少包含一个属于物体类别的像素，而物体边界框外的像素则应该均为背景像素，将边界框中的行或列看做正包，边界框外的行或列看做负包，属于物体的像素即为正示例，不属于物体的像素为负示例，通过约束正包内最大的网络输出趋近 1、负包中最大的网络输出趋近 0 来保证正包中包含至少一个正示例，同时负包中仅包含负示例。考虑到这种基于最大值的损失可能导致网络输出丢失连续性，Hsu 等人（2019）还提出了二元项来对网络预测的实例分割结果进行平滑，并对网络输出使用了 DenseCRF 进行后处理，达到了较好的弱监督实例分割性能。Lan 等人（2021）考虑到实例分割能够显著提升在多物体场景下挖掘物体间语义相关性的能力，从而可以在实例分割的基础上定义更有挑战性的挖掘物体间语义相关性的任务，更加关注物体相关性的质量以及物体定位的准确性，提出了 DiscoBox 框架，将弱监督实例分割任务与目标检测任务、挖掘语义相关性的任务相结合，在图像与图像间利用多层次的结构化知识与自监督学习来提高模型完成相应任务的能力。DiscoBox 考虑到 BBTP 所生成的示例分割结果仍然不够精确，使用了自我补偿的思想，引入退化模型与原模型间的一致性约束作为自监督来提高模型的表示能力，即在模型训练过程中，要求图像中物体的特征与历史模型所得到的特征具有相关性，最终显著提高了弱监督实例分割的性能。

Tian 等人（2021）基于能够动态调整分割头部权重的 CondInst 提出了 BoxInst 方法，考虑到边界框的掩膜与实际物体掩膜在横纵坐标轴具有相同的投影结果，提出了投影损失，形式上与 BBTP 中的一元项类似，即要求边界框每行或列的最大值趋近于 1，而边界框外的行或列最大值应该趋近于 0，同时也考虑到相邻的具有相近颜色的像素对很可能属于同一类别，提出了一个二元项来降低监督信息中存在的噪声，约束颜色相似度大于某一阈值的相邻相似度具有相同的类别，并通过统计与实验的方式对设定的阈值进行了选择与验证。通过使用这两种损失替换 CondInst 中的分割损失，BoxInst 可以使用边界框标注端到端地训练实例分割模型，并取得了不错的分割性能。

Wang 等人（2021c）提出了 BoxCaseg 方法，将边界框监督信息与显著性检测相结合，其认为基于边界框的弱监督实例分割问题关键在于基于边界框的类别无关物体分割问题 BoxCaseg（box-supervised class-agnostic object detection），即在给定目标物体的边界框的情况下，获取到边界框内物体的像素级掩码。Wang 等人（2021c）设计了一个针对 BoxCaseg 的联合训练框架，同时针对基于边界框标注的多示例学习任务、显著图像的多示例学习任务以及显著

图像的像素级分割任务。其中显著图像及其显著性检测标签来源于一个小的外部数据集,然后再使用这个训练好的 BoxCaseg 模型生成伪标签,使用边界框标签对伪标签进行裁剪,忽略边界框外的区域,并使用最终生成的伪标签训练 Mask R-CNN 实例分割模型来解决基于边界框标注的弱监督实例分割问题。除此之外,Cheng 等人(2021)提出的 Point-sup 方法结合了边界框标注以及在其内部随机采样的点标注进行弱监督实例分割,这种标注方式依然能够比像素级的图像标注工作快 5 倍。在 PointRend 模块基础上针对弱监督实例分割任务设计了 Implicit PointRend 模块,能够产生可以用于区分物体的区域级别的上下文信息,为每个物体生成不同的参数来预测最终的点级别掩膜,从而使模型在经过边界框与点标注端到端训练后具有实例分割的能力。

目前,针对弱监督实例分割任务的方法还相对较少,这可能是由于实例分割相比语义分割更具挑战性,在区分像素所属物体类别的基础上还需要区分具体属于哪个物体,而这一问题也是弱监督实例分割方法的关键。对于基于边界框标注的弱监督实例分割而言,已经给出了图像中的各个实例的标签,研究者们更需要关注在基于边界框标注的弱监督语义分割基础上,如何使用含有噪声的边界框标注来区分每个像素属于哪一个物体。对于基于图像级标签的弱监督实例分割问题,如何仅通过图像级标签挖掘图像中的实例信息可能是更具挑战性的问题。

2.6 数据集及模型性能评估

弱监督语义和实例分割任务的常用评价数据集有 PASCAL VOC (Everingham 等,2015)数据集,MS COCO (Lin 等,2014)数据集。通常采用 MS COCO 作为训练集,PASCAL VOC 作为测试集验证弱监督方法的性能。评价指标通常采用交并比(intersection over union, IoU)和平均精度(average precision, AP)等。表 2 展示了部分弱监督语义分割和示例分割方法在 PASCAL VOC 2012 验证集和测试集的性能表现。随着网络架构的发展,基于 Transformer 的方法(Liu 等,2021c)在语义分割任务中取得了较大的提升。相较于图像级的类别标注,边界框和线标注能够提供更多的空间信息,相应的弱监督方法性能更优。同时,由于实例分割任务较语义分割难度更大,弱监督实例分割方法的性能仍存在较大的差距。

表 2 不同弱监督分割方法在 PASCAL VOC 2012 上的性能表现

Table 2 Quantitative comparison of weakly supervised segmentation methods on PASCAL VOC 2012 dataset

语义分割	标注形式	mIoU/%	
		验证集	测试集
Bearman 等人(2016)	P	46.1	–
PSL(Wei 等,2017a)	I	55.0	55.7
Papadopoulos 等人(2017)	P	58.4	–
RAWKS (Vernaza 和 Chandraker,2017)	S	61.4	–
BoxSup(Dai 等,2015)	B	62.0	64.6
ScribbleSup(Lin 等,2016)	S	63.1	–
IRNet(Ahn 等,2019)	I	63.5	64.8
CIAN(Fan 等,2020)	I	64.3	65.3
SEAM(Wang 等,2020)	I	64.5	65.7
CONTA(Zhang 等,2020)	I	66.1	66.7
SDI(Khoreva 等,2017)	B	69.4	–
BCM(Song 等,2019)	B	70.2	–
EPS(Lee 等,2021c)	I + Saliency	71.0	71.8
LPG(Xie 等,2021)	B	73.3	–
BAP(Oh 等,2021)	B	74.6	76.1
SPML(Ke 等,2021)	S	76.1	–
Swin(Liu 等,2021c)	S	82.8	82.9
实例分割	标注形式	AP_{50}	AP_{75}
		验证集	
PRM(Zhou 等,2018)	I	26.8	9.0
Label-PENet(Ge 等,2019)	I	30.2	12.9
LACI(Arun 等,2020)	I	50.9	28.5
BBTP(Hsu 等,2019)	B	58.9	21.6
BoxInst(Tian 等,2021)	B	61.4	37.0
DiscoBox(Lan 等,2021)	B	63.6	34.1
BoxCaseg(Wang 等,2021c)	B + Saliency	67.6	42.4

注:B 表示边界框标注,I 表示图像级标注,S 表示线标注,P 表示点标注。mIoU: mean IoU, IRNet: inter-pixel relation network, CIAN: Cross-image affinity net AP_{50}:IoU 取值大于 0.5 时的平均准确率,AP_{75}:IoU 取值大于 0.75 时的平均准确率。

3 弱监督动作检测

随着视频数据在现实生活中的应用越来越广

泛，与图像场景理解相似，人们对于视频场景理解的需求也在逐渐增加。其中，视频动作检测是视频场景理解相关任务中的一个主要任务。传统的全监督动作检测任务与图像目标检测类似，标记视频中的一个动作需要同时给出该动作的分类，以及该动作的起止时间。但是对于大规模视频数据集，标注所有动作需要大量的时间与资源。因此，基于弱监督学习的动作检测逐渐成为近几年视频理解领域的研究热点。本节首先介绍弱监督目标检测领域不同的问题设置，然后给出常用的弱监督动作检测模型的核心结构与计算流程，最后将根据不同的问题设置，分别介绍近几年提出的弱监督动作检测算法及其特点。

3.1　典型问题设置

3.1.1　电影脚本作为监督

早期的动作检测任务可以追溯到 21 世纪初针对电影中演员动作的检测。电影脚本包括了演员的动作信息，相比带有精确时间轴的字幕，通常来讲电影脚本不包含精确时间信息，且需要额外的对齐步骤才能够与视频信息对应起来，因此监督信息相对较弱，但相对全监督标注更易获得。Laptev 等人(2008)最先开始使用电影脚本来进行弱监督动作检测任务，其标注来源于 3 个知名电影脚本网站的数百份电影脚本。但是这类监督信息存在缺点，其缺陷在于只有电影类别的视频才能够有脚本标注，而现实生活场景中的视频很难有这样的标注，因此较难泛化到现实生活中通用场景下的弱监督动作检测任务。

3.1.2　动作序列作为监督

考虑到基于电影脚本监督的动作检测存在的问题，Bojanowski 等人(2014)尝试将电影脚本替换成一个动作序列，例如“开门→起立→握手”，基于这样的标注信息来训练一个对应的弱监督动作检测模型。一个典型的帧序列和对应的动作序列如图 5 所示，该帧序列包含了连续的 5 个动作，上述动作构成了一个有序的动作序列，但是不包含时间信息。动作序列作为监督有几个好处：一是相比电影脚本，动作序列的标注可以面向所有视频片段，因此这类标注足够通用；二是动作序列标注包含了一定程度的动作的先后顺序信息，给出了较强的先验知识，因此在一定程度上可以提高弱监督动作检测的精度，同时并没有引入很多的标注代价。

3.1.3　类别标签作为监督

动作序列本身仍然包含了动作先后顺序的监督信号，这样的先验可能不利于模型在不同视频上的泛化能力，因此，Wang 等人(2017)参照早期的弱监督目标检测方法的经验，只使用视频中动作的分类标签作为监督信息。这一类别的训练设置可概述为：给定一个视频数据集，该数据集中，每段视频中包含的动作类别已知，使用给定的视频和对应的类别标签训练一个弱监督动作检测模型。如图 5 所示，该帧序列只包含 4 种不同的动作类型，其他的动作类型并不存在，可视为负标签。相比电影脚本与动作序列，类别标签不包含任何动作先后顺序的先验知识，因此使用这样的监督信息来训练检测模型，理论上可以有效保证模型的泛化能力。使用类别标签作为监督也成为近几年弱监督动作检测方法的主流问题设置，大量的工作围绕着类别标签监督信息展开，并且取得了可观的动作检测精度。本节中将要介绍的大多数后续工作都是基于类别标签监督信息展开。

3.1.4　单帧标签作为监督

除了已经广泛采用的基于类别标签的弱监督动作检测方法外，近期一些基于单帧标注的弱监督动作检测方法(Ju 等，2020；Lee 和 Byun，2021；Ma 等，2020；Moltisanti 等，2019)也相继提出。与弱监督图像分割中的单点标注类似，动作检测中的单帧标注要求，对于视频中的每一个动作，只有其中一帧被标注该动作的类别信息，且要求该帧的时间信息已知，即被标注帧的时间位置。如图 5 所示，给定的帧序列中包含了 5 个动作，其中每个动作都包含一个时间戳和给定的动作标签，因此可以用 5 个时间戳与对应的动作标签构成一组点标注。给定这样的标注，训练一个弱监督动作检测模型。基于上述表述，可以总结出单帧标签作为监督信息的特性：1)一段视频中单帧标签的数量等于视频中的动作数量，给出了较强的数量先验；2)单帧标签的时间信息已知，给出了较强的位置先验。这两点特性有效地提升了弱监督动作检测的标注信息质量。此外，单帧标签表面上看起来需要很长的标注时间，但是根据 Ma 等人(2020)的研究，标注一段 1 min 视频中的动作，给出类别标签平均需要 45 s，而给出单帧标签平均只需要 50 s，仅比标注类别标签多 5 s，且远远低于给出全监督标注的 300 s。因此单帧标签还是相对比较高效的，有可能是未来最适合工业界的标注方法，具有足够多的研究与发展空间。

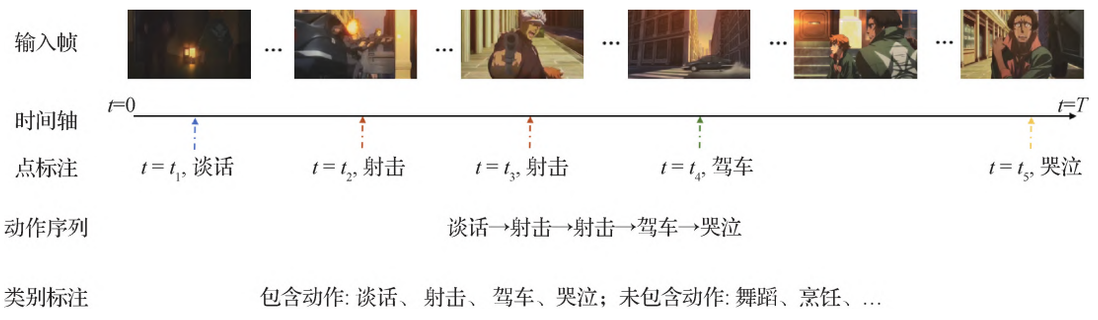


图 5　弱监督动作检测 3 种常见动作标注方式的示意图

Fig. 5　Examples of common action annotations for weakly supervised action recognition

3.2　弱监督动作检测模型

随着深度学习技术的发展,近期的弱监督动作检测模型基本沿用 Wang 等人(2017)提出的 UntrimmedNet 的网络架构和(Nguyen 等,2018)提出的 STPN (sparse temporal pooling network)网络架构。具体来讲,给定一段视频以及使用 TV-L1 算法生成的光流,UntrimmedNet 首先通过候选区采样模块将视频以及光流抽取成等长且不重叠的连续帧片段,构成多个候选区。然后对每个候选区,UntrimmedNet 使用在 Kinectics 数据集上预训练的双分支卷积神经网络或者时间片段网络计算每个候选区的特征,其中该特征为一个向量。对每个候选区的特征向量,该网络的分类模块通过一个线性变换将特征向量转化为该候选区的分类置信度,并通过 softmax 函数转换成类别概率。给定所有候选区的类别概率,该网络的选择模块通过基于多示例学习的硬选择模块或者基于注意力机制的软选择模块,输出每个候选区是否包含一个动作的置信度,并挑选出置信度最高的多个候选区。给定分类概率和包含动作的置信度,该网络计算两组分数的乘法,得到最终的动作定位结果,并使用预设阈值的后处理算法给出动作检测信息。而对于 STPN 网络,该网络在将特征提取器从时间片段网络替换成预训练的 I3D (inflated 3D ConvNet)网络的基础上,还引入了如下的改进:1)对于每个候选区,STPN 会通过一个参数共享的注意力模块生成一个值域为 0 ～1 的注意力分数,该分数不仅会作为片段注意力权重乘以候选区向量,还会用于计算一个正则项损失函数来优化网络训练;2)STPN 引入了时间维度的类别激活图 T-CAM 的机制,STPN 分别计算视频帧的 T-CAM 以及光流的 T-CAM,通过后处理算法生成最终的候选区用于给出最终的动作定位结果。后续基于深度学习方法的弱监督动作检测网络基本基于这两种网络的主要结构,加入新的网络模块或者损失函数来改进弱监督动作检测方法的性能。

3.3　弱监督动作检测算法

由于在不同的时间段,弱监督动作检测算法选用的监督信息不同,因此,根据监督信息对现有的弱监督动作检测算法进行归纳总结。

3.3.1　基于电影脚本的弱监督动作检测

Laptev 等人(2008)最先开始使用电影脚本来进行弱监督动作检测任务。该方法首先在视频帧中基于时空多尺度特征通过 Harris 运算符抽取出兴趣点,然后使用 K-Means 算法进行聚类,构成一组时空特征包,然后规整成可用于分类的形式。最后该方法利用前述特征包和脚本标注训练一个非线性支持向量机,使得该分类器可以对特征包进行分类,完成弱监督动作分类。Duchenne 等人(2009)在这一工作的基础上,针对给定的视频片段特征提出了一个显著性聚类算法,该方法通过一个支持向量机分类器来区分出背景视频片段和包含动作的视频片段,并最小化一个特定的代价函数,从而可以训练出一个理想的分类器用于挑选出包含动作的视频片段,从而真正开始支持视频中的动作检测任务。

3.3.2　基于动作序列的弱监督动作检测

Bojanowski 等人(2014)提出了第 1 种使用动作序列作为监督信息的弱监督动作检测算法。该算法在早期基于显著性聚类算法的基础上,针对包含动作执行顺序信息的动作序列,引入了一个动作分配矩阵,该矩阵由一个非降序的动作分配序列生成,保证动作的执行顺序与动作序列一致。为了学到这个动作分配矩阵,引入了动态规划算法和 Frank-Wolfe

优化算法，反复搜索更优的动作分配矩阵，直到收敛。

在上述工作的基础上，Huang 等人（2016）提出了一个扩展连接时序分类框架，该框架相比标准的连接时序分类方法，考虑到连续帧之间存在的语义信息的高度关联，利用帧间的视觉相似性，显式地强化了动作检测算法的标签对齐能力。

上述方法显著改善了弱监督目标检测能力，但是因为引入了包括循环神经网络（recurrent neural network，RNN）在内的计算量很大的结构，前述方法很难扩展到大规模数据集上。为了能够显著降低训练与测试阶段的计算复杂度并提升算法可扩展性，Xu 和 Ding（2018）提出了时间卷积特征金字塔网络，该网络首先引入了一个基于 U-Net 的特征金字塔结构来强化视频序列特征中的时序特征融合，然后提出了一个迭代边界分配结构，该结构可以更准确地找到有区分度的特征用于构造动作的边界。这两个改进显著提高了基于动作序列监督的弱监督动作检测方法的精度。

3.3.3　基于类别标签的弱监督动作检测

随着深度学习方法的大规模应用，与弱监督目标检测方法类似，Wang 等人（2017）开始将多示例学习方法与深度神经网络结合，提出针对类别标签监督的动作检测学习算法。不同于弱监督目标检测方法需要依靠外部的预计算候选框生成候选区特征，完整的视频序列可以通过等间隔切分构成一组不重叠的视频片段，不失一般性，可以假设每个视频片段只包含一个动作或者为背景视频。此时可以假设对于视频的每个分类标签，至少有一个视频片段可以被分类为该类别，那么所有的视频片段构成了一个包，要通过多示例学习算法执行包分类的任务。因此，基于前述的 UntrimmedNet 输出的片段分类结果，该网络将分类结果沿着时间维度求平均，可以得到该视频的分类结果，此时可以计算视频分类标签与分类结果的交叉熵来优化网络参数。

直接应用多示例学习方法来训练弱监督动作检测网络会导致检测网络关注到每个动作最显著的区域，从而可能忽视掉动作的开始与结束部分，影响最终的检测精度。为了解决这一问题，Singh 和 Lee（2017）从数据增广的角度提出了一个改进策略，该策略要求在训练弱监督动作检测网络时，每次随机遮盖掉一部分视频片段，然后按照正常的多示例学习算法去训练网络。这一方法可以有一定概率遮盖掉每个动作最显著的区域，使得网络更多地关注每个动作的其他部分，迫使网络尽可能准确地分类其他片段，从而改善动作定位性能。

同样是为了解决动作检测网络的分类器会关注到最显著部分的问题，Zhong 等人（2018）提出了一个训练多个分类器进行多阶段处理的方式。该算法要求多个分类器串行输出分数，前一个分类器中给出的高置信度结果会先进行遮盖处理，然后送到下一个分类器中输出新的分数，迫使网络尽可能多地感受到每个动作显著性偏度的部分，尤其是动作的开始与结束片段。将多个分类器的输出结果取并集即可得到最终的定位结果。该方法相比在输入阶段进行随机遮盖处理，能够更有效地让网络关注到所有包含动作片段的区域，进而提高检测精度。

前述的弱监督动作检测网络与全监督检测方法不同，这类方法通常需要预设阈值的后处理方法来检测高于某个阈值的连续动作片段来作为最终的动作定位结果，并不能直接输出动作的起点与终点。为了解决这一问题，Shou 等人（2018）提出了 AutoLoc 网络，该网络提出了一个全新的定位分支，该分支同时回归每个动作的中心位置和动作持续长度，基于这些结果可以计算出一个动作片段的内边界和外边界。同时提出了一个全新的内外对照损失函数，该损失函数要求对于包含动作的片段，内边界中视频片段的平均置信度和外边界中的视频片段平均置信度之间的差距尽可能大。基于这样的优化原则，在网络训练完毕后，可以选取动作片段的内边界直接作为动作定位的输出结果。

进一步，如果网络可以同时输出每个视频片段的分类结果，同时回归出视频片段的起止时间，那么参照弱监督目标检测领域的在线优化策略（Tang 等，2017），弱监督动作检测也可以通过类似的思路提升检测精度。基于这一思路，Pardo 等人（2021）提出了一个迭代优化算法，引入多个动作检测分支，通过在线的伪标签生成策略和自学习方法，逐步提升每个分支的检测精度，从而提升弱监督动作检测网络的整体性能。

除了同一段视频内所有片段的特征相似性，不同视频样本间的视频片段也有一些相似性可以利用，例如两段同时包含跑步的视频，各自对应跑步的视频片段特征与分类结果会趋近于相似，而这一片

段的特征与分类结果与不包含跑步的视频片段差别较大。基于这一发现，Paul 等人（2018）提出了 Co-Activity 损失，目的在于强化样本对之间同类片段的相似性，这个损失函数可以与多示例学习损失函数组成联合学习算法，有效提升动作检测精度。

为了避免动作检测网络错误地从背景类或无关动作中学到关联，并且强化连续动作间的相关性，Xu 等人（2019）提出了 STAR（segregated temporal assembly recurrent networks），将 RNN 结构引入到弱监督动作检测网络中来强化视频片段间的时序关联，同时他们提出了包含 ST-GradCAM 模块的标签生成算法指导动作定位器输出优化的定位结果。

同样是为了建模并强化视频片段中的时序关联，Yu 等人（2019a）提出了时序结构挖掘算法，该算法将每个动作分解成起始、进行和结束 3 个阶段，同时将原有的类别置信度细化为阶段分类置信度。为了优化阶段分类置信度，还提出了一个最大循环路径搜索算法，通过动态规划算法搜索出对应的路径并优化网络。

此外，上述方法虽然尝试建立了视频中的时序关联，但是现有的方法均没有直接建模一段视频中的动作数量。为了建模动作数量从而改善弱监督动作检测方法性能，Narayan 等人（2019）提出了 3C-Net，该方法在标准的分类损失的基础上，引入了一个中心损失来强化特征的可区分度，同时引入了一个计数损失来强化相邻的动作的可区分度，由于动作数量这一监督信息的引入，可以有效地建模一段视频中有多少动作该输出为高置信度，因此可以显著改善弱监督动作检测的识别能力。

随着弱监督动作检测可以更精确地识别出非显著区域的动作，研究人员开始逐渐关注弱监督学习范式下每个动作的边界如何界定。通常来讲，一个动作的起始与停止阶段的视频帧结构特征与其上下文的结构特征相似，但是包含着不同的动作信息，例如起止片段有动作而上下文部分没有动作。为了解决这一问题，Liu 等人（2019）提出了一个多分支分类网络，对于不同的分类器，希望对应的类别激活序列能够关注到不同的区域，因此提出了一个区分度损失，迫使不同分类器关注不同的时序区域，尽可能多地覆盖每个动作的不同阶段，使得动作的每个阶段都可能被一个分类器所检测到，并在最后相加进行特征融合，构成完整的动作检测结果。

为了建模背景，Lee 等人（2020）提出了背景抑制网络，该网络引入了一个背景抑制分支和一个对照损失，该分支包含一个过滤模块，通过给定特征生成前景权重来抑制掉背景部分，减少检测网络的误检率。此外，Lee 等人（2021b）又提出了不确定性建模的方法，通过建模动作间和背景片段间的不确定性，以及将背景类的视频片段建模为不服从分布的样本，来克服直接建模背景类时因为不一致性造成的训练困难，从而提高背景类的感知能力，进一步提升动作检测精度。

同时，为了构建视频数据中的不确定性，Yang 等人（2021b）又提出了一种不确定性联合学习的训练策略。首先引入了一个在线伪标签生成模块，该模块可以通过指数滑动平均的方法维护一个标签生成网络，来为 RGB 数据和光流数据生成另一种数据的伪标签，互相指导进行联合学习。然后作者引入了一个不确定性相关学习模块，该模块结合伪标签损失函数可以减少生成的伪标签中的噪声，减少因为噪声带来的训练阶段负面影响。

同样是来自背景上下文的干扰，如何高效地建模前景与背景片段也是近期弱监督动作检测网络算法急需解决的一个问题。因此，Nguyen 等人（2019）提出了一个能够自动进行背景先验建模的弱监督动作检测网络，通过引入背景类损失、自注意力损失和一个聚类损失，使得背景类片段可以更高效地被挖掘出来，进而提升动作片段和背景上下文之间的区分度，进而提升弱监督检测性能。此外，同样是为了建模视频片段上下文，Ma 等人（2021）提出了一种动作挑选学习的训练算法，该算法引入了一个类别无关的动作检测模型，专门用来学习哪些视频片段会被分类器选进前 K 高分的视频片段集合。而在测试阶段，该动作检测网络只需要整合标准动作检测网络分支的输出和该类别无关模型的输出结果，获得最终的检测结果，这一简单的改进能够带来可观的检测性能提升。

为了更好地建模动作与背景上下文之间的区分度，Shi 等人（2020）提出了一种生成式建模的方法，通过迭代训练一个用来重建视频片段特征的 CVAE（conditional variational auto-encoder）和最终想要的弱监督检测网络，分别建模生成式注意力能力和区分度注意力能力，进而可以有效改善检测网络对动作片段及其上下文的感知与区分能力。

此外，Zhai 等人(2020)为了进一步解决弱监督检测网络输出假阳性片段的问题，提出了 TSCN(two-stream consensus network)，该方法包含一个双分支融合模块，用来整合 RGB 和光流的类激活序列，再生成伪标签，通过前述迭代自学习的方法独立训练 RGB 数据和光流数据的分类器。实验表明，将两类数据分治处理，在分类结果部分整合的效果要优于直接进行特征融合再做自学习迭代优化的方法。

在建模动作边界和背景上下文的任务上，对照学习也可以起到一定的作用。Zhang 等人(2021b)提出了一种基于对照学习来定位较难区分的视频片段的动作定位算法。具体地，提出了片段对照损失函数，在特征空间来优化较难区分的片段的特征表示。然后提出了一个分治算法，利用一个难分类片段挖掘算法先挑选出潜在的难分类样本，优化难分类片段的内外边界，最后从难分类样本和易分类样本中分别选择视频片段，应用前述片段对照损失函数进行网络训练。

Gong 等人(2020)则选择从建立动作的共同定位这一角度分析，提出了一种基于共同注意力的无监督动作定位算法。该方法引入了一个聚类 + 定位的解决思路，来解决无监督动作检测问题。为了优化网络，提出了一个动作—背景区分损失函数和一个基于聚类思想的三元组损失函数，可以有效提升动作和背景的区分度，提高动作检测能力。

3.3.4　基于单帧标签的弱监督动作检测

基于单帧标签作为监督的弱监督动作检测算法首先由 Moltisanti 等人(2019)提出。该方法首先提出了两种不同的单帧标签生成方式：第 1 种是在每个动作片段上按照均匀分布的方式随机选择一帧作为标注点；第 2 种是取每个动作片段的中点为均值点构造一个高斯分布，基于高斯分布随机采样一帧作为标注点。根据类似的采样方式，Moltisanti 等人(2019)希望以每个标注点为中心，各自初始化一个抽样分布，并用这样的抽样分布代替原始的起始点动作标注，来采样训练帧用来表示对应类别的特征。同时为了避免采到背景帧引入噪声，Moltisanti 等人(2019)提出了一个帧排序策略，对每个类别，选择置信度最高的多个帧参与训练。在完成一次训练后，Moltisanti 等人(2019)对采样分布进行更新，丢弃置信度低的候选分布并更新高置信度候选分布的参数。基于上述类似期望最大化方法的优化过程，反复迭代更新抽样分布与更新参数，最终获得目标弱监督动作检测网络。

类似地，Ma 等人(2020)在前述工作的基础上，提出了一组联合伪标签挖掘方式：第 1 部分是动作帧挖掘，将有标注的帧视为锚点，随着训练进行逐步向锚点两侧扩张，将分类结果不变且置信度上升的帧构造为包含动作的伪标签，作为逐帧分类的监督信息加入到训练中；第 2 部分是背景帧挖掘，由于背景帧没有标注信息，为了能够初始化背景帧类别的信息，首先引入了背景类分类器直接参与训练，尝试自发学到一些背景类别的信息，然后对所有无标注的帧排序，选择背景类置信度最高的多个帧，最大化这些帧在背景类的概率，从而挖掘出最有可能是背景帧的内容。在挖掘出动作帧与背景帧的基础上，提出了一个动作/背景类显著性模块，通过一个非线性变换和一个 sigmoid 运算得到每个候选区是否包含动作的置信度，并通过挖掘出的标签进行二分类监督，使得动作检测网络可以感知到类别无关的动作信息，提高泛化能力。同时，他们也首次进行了人工单帧标注的相关实验，实验结果发现，相比从原始全监督标注中随机采样，人工标注能够带来更高的单帧检测精度以及相同的平均动作定位精度。

进一步地，Ju 等人(2020)针对单帧标签监督信息的动作检测问题，提出了一个两阶段学习方法。在第 1 阶段，使用单帧标签训练一个关键帧检测器，关键帧检测器检出的关键帧会作为锚点用来将视频分割为多个视频片段。第 2 阶段，提出了一个独立的位置检测器，用来对前述的视频片段生成候选区，进而再通过一个映射器将候选区转换成分割掩码的形式，指导网络将视频片段划分为前景与背景，送入分类器中进行分类，并使用类别标签计算损失函数完成训练。该算法的好处在于，在测试阶段可以直接使用已经训练好的关键点检测器、位置检测器和分类器输出最终的动作检测结果，省去了复杂的后处理过程，有利于在工业界的应用。

但是，前述方法并不能基于单帧标注学到一个动作的完整性，为了解决这一问题，Lee 和 Byun(2021)在前述工作的基础上，首先通过一个最优序列搜索算法，在给定某一类的单帧标签和整个序列的动作置信度的前提下，搜索出一个稠密的伪标签

序列,该伪标签序列保证所有动作的完整性优于输入的动作置信度。所述伪标签序列只在网络训练阶段进行计算并参与优化,在测试阶段,由于网络已经学到了一定程度的动作完整性,因此不需要额外搜索最优序列,提高了测试阶段的执行效率。为了使动作检测网络可以学到动作的完整性,Lee 和 Byun(2021)首先提出了分数对照损失函数,使得网络更多地关注到与最优序列不一致的实例并改善这种不一致,消除将背景类误判成动作区域的影响。同时,提出了特征对照损失函数,该损失函数的目的与传统的对照学习类似,要将标注为同一类别的动作帧特征学到尽可能相似,且与其他背景类帧特征尽可能不同。基于上述对照损失函数与基础的弱监督学习损失函数,Lee 和 Byun(2021)构建出了一个可以学到动作完整性的联合学习算法。相比之前的方法,该算法在不同规模的数据集上均取得了更优的动作检测性能。

类似地,在前述类别标签监督信息的基础上,如果可以给出一些背景帧的具体位置,也能构造另一种基于帧标注监督的弱监督动作检测方法。其中,Yang 等人(2021a)首先基于这种思路提出了一种新的弱监督动作检测方法。具体来讲,给定背景帧监督信息,该网络可以通过背景帧的交叉熵损失函数高效地对背景类进行建模。同时,提出了一个分数区分模块和一个近似模块,来高效地建模位置信息和特征信息,提高弱监督模型对于动作类和背景类的区分度,提高网络检测精度。

3.4 动作检测数据集及相关算法性能

近几年基于深度学习的弱监督动作检测方法通常围绕 THUMOS 14 数据集(Idrees 等,2017)和 ActivityNet 1.2 数据集(Heilbron 等,2015)开展实验,因此简要介绍这两个数据集。

在弱监督动作检测任务中使用的 THUMOS 14 数据集通常为该数据集的一个子集,包含 20 类不同的动作。其中,训练集包含 200 个不同的视频,测试集包含 213 段不同的视频,这些视频均包含精确的动作片段起止时间标注,但上述标注仅在测试集的检测精度计算中得到应用。ActivityNet 1.2 数据集包含 4 819 段用来训练的视频和 2 383 段用来测试的视频。这些视频包含了 100 种不同的动作,每段视频平均包含 1.5 个动作片段。相比 THUMOS 14 数据集,ActivityNet1.2 数据集规模更大,更接近现实场景。

通常来讲,通过计算估计出的动作片段与真实标注的动作片段在不同 IoU 阈值下的分类精度 AP 来计算最终的平均检测精度 mAP(mean AP),并用该指标来度量检测算法的优劣。近期典型的弱监督动作检测算法的性能如表 3 所示。可以看出,随着建模动作边界和动作上下文信息挖掘的程度逐渐提升,弱监督动作检测算法的精度也在逐渐提高。但是,上述弱监督动作检测算法相对于全监督动作检测算法在 mAP 指标上仍然存在一定的差距。

表 3 典型弱监督动作检测算法在 THUMOS 14 数据集和 ActivityNet1.2 数据集上的平均精度

Table 3 Averege precision of weakly supervised action recognition methods on THUMOS 14 and ActivityNet1.2 datasets

方法	标注方式	mAP/%	
		THUMOS 14	ActivityNet1.2
Wang 等人(2017)	类别	29.0	3.6
Liu 等人(2019)	类别	40.9	22.4
Shi 等人(2020)	类别	45.6	24.4
Lee 等人(2021b)	类别	51.6	25.9
Luo 等人(2021)	类别	52.1	25.5
Zhang 等人(2021b)	类别	50.3	26.1
Ma 等人(2020)	点	51.2	22.8
Lee 和 Byun(2021)	点	62.7	26.8

4　视觉弱监督学习挑战及趋势

综上，视觉弱监督学习已经在物体检测、语义和实例分割以及视频动作识别等任务上取得了显著的进展。随着网络架构的发展，以 Transformer 为代表的弱监督方法已在语义分割等任务上取得了较大的性能提升。然而，视觉弱监督学习模型的性能与全监督学习仍存在较大的差距。另外，近年来兴起的自监督学习和大规模预训练模型也将为视觉弱监督学习带来新的挑战与机遇。此外，动态开放环境下的实际应用中往往涉及多样化的任务和标注形式，多种任务视觉弱监督学习以及多种监督形式的结合和转换将显得尤为必要。

4.1　存在的问题和挑战

相对于全监督学习，视觉弱监督学习由于仅依靠弱标注，因而会面临局部聚焦，难以准确地挖掘出所有的物体和精确地定位物体。如何进一步缩小视觉弱监督学习与全监督学习的性能差异，并将其进一步应用于实际的视觉理解应用，仍然是未来视觉弱监督学习研究中亟待解决的问题。另外，当视觉弱监督学习应用于开放动态环境时，标注噪声和异常样本往往难以避免。如何利用开放环境的大量未标注数据或噪声数据，并有效抑制异常样本的不利影响，是视觉弱监督学习研究走向应用的过程中亟待解决的挑战。此外，弱监督学习并不是降低标注成本的唯一方式。视觉弱监督学习如何在与其他低标注成本学习方式的竞争中脱颖而出，或者结合其他方式形成更为有效的低标注成本学习解决方案，仍然是一个值得深入研究的问题。

4.2　视觉弱监督学习的发展趋势

视觉弱监督学习近年来获得了较多的关注并获得了长足的发展。然而，如何提升弱监督学习性能、开放环境下的稳健性和连续性，仍然是未来视觉弱监督学习研究中值得关注的几个重要趋势。

模型性能的提升不仅依赖于视觉弱监督学习方法自身的进步，还需要考虑与现有数据集、任务和模型的结合问题。例如，受益于大规模预训练模型，利用知识迁移和蒸馏等方式将辅助数据集或更大的开放数据集的知识用以提升视觉弱监督学习性能。在目标检测与语义分割方向，已有工作尝试引入全监督辅助数据集（Dong 等，2021；Xie 等，2021），在保证辅助数据集与弱监督数据集无重合类别的情况下，通过知识蒸馏与迁移学习方法，借助辅助数据集上的知识来提升弱标注数据集上检测与分割的性能。近年来开始出现的 CLIP（contrastive language-image pre-training）等视觉—语言大规模预训练模型也可以视为一种重要的知识来源，有助于大幅提升视觉弱监督学习性能。对于大量未标注样本，自监督学习已经开始展现出良好的学习和泛化性能，因而也有望与弱监督学习有机结合以取得更高的性能。此外，检测和分割任务具有较强的相关性，不同学习任务的结合也有望为提升弱监督学习性能提供新的研究思路。

当视觉弱监督学习应用于开放动态环境时，无标注样本、标注噪声和异常样本往往难以避免。为在开放环境下实现稳健的模型学习，需要发展更为有效的学习方法，能够在标注噪声和异常样本的情况下实现稳健的视觉弱监督学习，并充分利用无标注样本提升学习性能。

此外，开放动态环境下的视觉弱监督学习往往涉及新类别的发现、增加和调整等。因而，需要结合视觉弱监督学习和连续学习，使得模型在学习和适应到新的类别时，在已有类别上仍然能够保持原有性能。另一方面，为适应开放动态环境，还需要发展开放域视觉弱监督学习方法，提升模型对未知类别的发现能力。

致　谢　本文由中国图象图形学学会机器视觉专委会组织撰写，该专委会更多详情请见链接：http://www.csig.org.cn/detail/2386。

参考文献（References）

Ahn J, Cho S and Kwak S. 2019. Weakly supervised learning of instance segmentation with inter-pixel relations//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 2204-2213 [DOI: 10.1109/CVPR.2019.00231]

Ahn J and Kwak S. 2018. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4981-4990 [DOI: 10.1109/CVPR.2018.00523]

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,

USA: IEEE: 328-335 [DOI: 10.1109/CVPR.2014.49]

Arun A, Jawahar C V and Kumar M P. 2020. Weakly supervised instance segmentation by learning annotation consistent instances//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 254-270 [DOI: 10.1007/978-3-030-58604-1_16]

Bearman A, Russakovsky O, Ferrari V and Fei-Fei L. 2016. What's the point: semantic segmentation with point supervision//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 549-565 [DOI: 10.1007/978-3-319-46478-7_34]

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

Blömer J and Bujna K. 2013. Simple methods for initializing the EM algorithm for Gaussian mixture models [EB/OL]. [2022-03-05]. https://arxiv.org/pdf/1312.5946.pdf

Bojanowski P, Lajugie R, Bach F, Laptev I, Ponce J, Schmid C and Sivic J. 2014. Weakly supervised action labeling in videos under ordering constraints//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer: 628-643 [DOI: 10.1007/978-3-319-10602-1_41]

Bolya D, Zhou C, Xiao F Y and Lee Y J. 2022. YOLACT++ better real-time instance segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(2): 1108-1121 [DOI: 10.1109/TPAMI.2020.3014297]

Cai L, Dong F, Chen K, Yu K H, Qu W and Jiang J F. 2020. An FPGA based heterogeneous accelerator for single shot multibox detector (SSD)//Proceedings of the 15th IEEE International Conference on Solid-State and Integrated Circuit Technology. Kunming, China: IEEE: 1-3 [DOI: 10.1109/ICSICT49897.2020.9278177]

Cao T Y, Du L Y, Zhang X Y, Chen S H, Zhang Y and Wang Y F. 2021. CaT: weakly supervised object detection with category transfer//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 3050-3059 [DOI: 10.1109/ICCV48922.2021.00306]

Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A and Zagoruyko S. 2020. End-to-end object detection with transformers//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 213-229 [DOI: 10.1007/978-3-030-58452-8_13]

Chen F Q. 2013. An improved EM algorithm [EB/OL]. [2022-03-05]. https://arxiv.org/pdf/1305.0626.pdf

Chen L C, Papandreou G, Kokkinos I, Murphy K and Yuille A L. 2015. Semantic image segmentation with deep convolutional nets and fully connected CRFs//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: ICLR

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 Z, Fu Z H, Jiang R X, Chen Y W and Hua X S. 2020. SLV: spatial likelihood voting for weakly supervised object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 12992-13001 [DOI: 10.1109/CVPR42600.2020.01301]

Cheng B W, Parkhi O and Kirillov A. 2021. Pointly-supervised instance segmentation [EB/OL]. [2022-03-05]. https://arxiv.org/pdf/2104.06404.pdf

Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S and Schiele B. 2016. The cityscapes dataset for semantic urban scene understanding//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 3213-3223 [DOI: 10.1109/CVPR.2016.350]

Dai J F, He K M and Sun J. 2015. BoxSup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 1635-1643 [DOI: 10.1109/ICCV.2015.191]

Dempster A P, Laird N M and Rubin D B. 1977. Maximum likelihood from incomplete data via the *EM* algorithm. Journal of the Royal Statistical Society, 39(1): 1-22 [DOI: 10.1111/j.2517-6161.1977.tb01600.x]

Diba A, Sharma V, Pazandeh A, Pirsiavash H and Van Gool L. 2017. Weakly supervised cascaded convolutional networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5131-5139 [DOI: 10.1109/CVPR.2017.545]

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]

Dong B W, Huang Z T, Guo Y L, Wang Q L, Niu Z X and Zuo W M. 2021. Boosting weakly supervised object detection via learning bounding box adjusters//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 2856-2865 [DOI: 10.1109/ICCV48922.2021.00287]

Duchenne O, Laptev I, Sivic J, Bach F and Ponce J. 2009. Automatic annotation of human actions in video//Proceedings of the 12th IEEE International Conference on Computer Vision. Kyoto, Japan: IEEE: 1491-1498 [DOI: 10.1109/ICCV.2009.5459279]

Durand T, Thome N and Cord M. 2016. WELDON: weakly supervised learning of deep convolutional neural networks//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 4743-4752 [DOI: 10.1109/CVPR.2016.

51]

Erez O and Maron T. 1998. A framework for multiple-instance learning. Advances in Neural Information Processing Systems [EB/OL]. [2022-03-05]. https://papers.nips.cc/paper/1997/file/82965d4ed8150294d4330ace00821d77-Paper.pdf

Everingham M, Eslami S M A, Van Gool L, Williams C K I, Winn J and Zisserman A. 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]

Fan J S, Zhang Z X, Tan T N, Song C F and Xiao J. 2020. CIAN: Cross-image affinity net for weakly supervised semantic segmentation//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 10762-10769

Ge W F, Huang W L, Guo S and Scott M. 2019. Label-PEnet: sequential label propagation and enhancement networks for weakly supervised instance segmentation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 3344-3353 [DOI: 10.1109/ICCV.2019.00344]

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

Gong G Q, Wang X H, Mu Y D and Tian Q. 2020. Learning temporal co-attention models for unsupervised video action localization//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 9816-9825 [DOI: 10.1109/CVPR42600.2020.00984]

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

He K M, Gkioxari G, Dollár P and Girshick R. 2020. Mask R-CNN. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2): 386-397 [DOI: 10.1109/TPAMI.2018.2844175]

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 (CVPR). Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Heilbron F C, Escorcia V, Ghanem B and Niebles J C. 2015. ActivityNet: a large-scale video benchmark for human activity understanding//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE: 961-970 [DOI: 10.1109/CVPR.2015.7298698]

Hou Q B, Jiang P T, Wei Y C and Cheng M M. 2018. Self-erasing network for integral object attention//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: Curran Associates Inc.: 547-557

Hsu C C, Hsu K J, Tsai C C, Lin Y Y and Chuang Y Y. 2019. Weakly supervised instance segmentation using the bounding box tightness prior//Proceedings of Advances in Neural Information Processing Systems. Vancouver, Canada: Neural Information Processing Systems Foundation: 6582-6593

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, USA: IEEE: 7132-7141 [DOI: 10.1109/CVPR.2018.00745]

Huang D A, Li F F and Niebles J C. 2016. Connectionist temporal modeling for weakly supervised action labeling//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 137-153 [DOI: 10.1007/978-3-319-46493-0_9]

Huang Z Y, Zou Y, Bhagavatula V and Huang D. 2020. Comprehensive attention self-distillation for weakly-supervised object detection [EB/OL]. [2022-03-05]. https://arxiv.org/pdf/2010.12023.pdf

Hwang J J, Yu S, Shi J B, Collins M, Yang T J, Zhang X and Chen L C. 2019. SegSort: segmentation by discriminative sorting of segments//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea(South): IEEE: 7333-7343 [DOI: 10.1109/ICCV.2019.00743]

Idrees H, Zamir A R, Jiang Y G, Gorban A, Laptev I, Sukthankar R and Shah M. 2017. The THUMOS challenge on action recognition for videos "in the wild". Computer Vision and Image Understanding, 155: 1-23 [DOI: 10.1016/j.cviu.2016.10.018]

Jia Q F, Wei S K, Ruan T, Zhao Y F and Zhao Y. 2021. GradingNet: towards providing reliable supervisions for weakly supervised object detection by grading the box candidates//Proceedings of the 35th AAAI Conference on Artificial Intelligence. [s.l.]: AAAI

Ju C, Zhao P S, Zhang Y, Wang Y F and Tian Q. 2020. Point-level temporal action localization: bridging fully-supervised proposals to weakly-supervised losses [EB/OL]. [2022-03-05]. https://arxiv.org/pdf/2012.08236.pdf

Kantorov V, Oquab M, Cho M and Laptev I. 2016. ContextLocNet: context-aware deep network models for weakly supervised localization//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 350-365 [DOI: 10.1007/978-3-319-46454-1_22]

Ke T W, Hwang J J and Yu S X. 2021. Universal weakly supervised segmentation by pixel-to-segment contrastive learning//Proceedings of the 9th International Conference on Learning Representations. [s.l.]: ICLR

Khoreva A, Benenson R, Hosang J, Hein M and Schiele B. 2017. Simple does it: weakly supervised instance and semantic segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1665-1674 [DOI: 10.1109/CVPR.2017.181]

Kim D, Cho D, Yoo D and Kweon I S. 2017. Two-phase learning for weakly supervised object localization//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 3554-3563 [DOI: 10.1109/ICCV.2017.382]

Kosugi S, Yamasaki T and Aizawa K. 2019. Object-aware instance labeling for weakly supervised object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea(South): IEEE: 6063-6071 [DOI: 10.1109/ICCV.2019.00616]

Krähenbühl P and Koltun V. 2011. Efficient inference in fully connected CRFs with Gaussian edge potentials//Proceedings of the 25th Advances in Neural Information Processing Systems. Granada, Spain: NIPS: 109-117

Krizhevsky A, Sutskever I and Hinton G E. 2012. Imagenet classification with deep convolutional neural networks//Proceedings of the 26th Neural Information Processing Systems. Lake Tahoe, USA: NIPS: 1106-1114

Lan S Y, Yu Z D, Choy C, Radhakrishnan S, Liu G L, Zhu Y K and Anandkumar A. 2021. DiscoBox: weakly supervised instance segmentation and semantic correspondence from box supervision//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 3386-3396

Laptev I, Marszalek M, Schmid C and Rozenfeld B. 2008. Learning realistic human actions from movies//Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, USA: IEEE: 1-8 [DOI: 10.1109/CVPR.2008.4587756]

Laradji I H, Vázquez D and Schmidt M. 2019. Where are the masks: instance segmentation with image-level supervision//Proceedings of the 30th British Machine Vision Conference. Cardiff, UK: BMVA: #255

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

Lee J, Yi J H, Shin C and Yoon S. 2021a. BBAM: bounding box attribution map for weakly supervised semantic and instance segmentation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 2643-2651 [DOI: 10.1109/CVPR46437.2021.00267]

Lee P and Byun H. 2021. Learning action completeness from points for weakly-supervised temporal action localization//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 13628-13637 [DOI: 10.1109/ICCV48922.2021.01339]

Lee P, Uh Y and Byun H. 2020. Background suppression network for weakly-supervised temporal action localization//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 11320-11327

Lee P, Wang J L, Lu Y and Byun H. 2021b. Weakly-supervised temporal action localization by uncertainty modeling//Proceedings of the 35th AAAI Conference on Artificial Intelligence. [s. l.]: AAAI: 1854-1862

Lee S, Kwak S and Cho M. 2018. Universal bounding box regression and its applications//Proceedings of the 14th Asian Conference on Computer Vision. Perth, Australia: Springer: 373-387 [DOI: 10.1007/978-3-030-20876-9_24]

Lee S, Lee M, Lee J and Shim H. 2021c. Railroad is not a train: saliency as pseudo-pixel supervision for weakly supervised semantic segmentation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 5491-5501 [DOI: 10.1109/CVPR46437.2021.00545]

Li X Y, Kan M N, Shan S G and Chen X L. 2019a. Weakly supervised object detection with segmentation collaboration//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea(South): IEEE: 9734-9743 [DOI: 10.1109/ICCV.2019.00983]

Li X Y, Zhou T F, Li J W, Zhou Y and Zhang Z X. 2021a. Group-wise semantic mining for weakly supervised semantic segmentation//Proceedings of the 35th AAAI Conference on Artificial Intelligence. [s. l.]: AAAI: 1984-1992

Li Y, Zhang J G, Huang K Q and Zhang J G. 2019b. Mixed supervised object detection with robust objectness transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(3): 639-653 [DOI: 10.1109/TPAMI.2018.2810288]

Li Y W, Zhao H S, Qi X J, Chen Y K, Qi L, Wang L W, Li Z M, Sun J and Jia J Y. 2021b. Fully convolutional networks for panoptic segmentation with point-based supervision [EB/OL]. [2022-03-05]. https://arxiv.org/pdf/2108.07682.pdf

Lin D, Dai J F, Jia J Y, He K M and Sun J. 2016. ScribbleSup: scribble-supervised convolutional networks for semantic segmentation//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 3159-3167 [DOI: 10.1109/CVPR.2016.344]

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

Liu D C, Jiang T T and Wang Y Z. 2019. Completeness modeling and context separation for weakly supervised temporal action localization//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 1298-1307 [DOI: 10.1109/CVPR.2019.00139]

Liu Y, Zhang Z J, Niu L, Chen J J and Zhang L Q. 2021a. Mixed supervised object detection by transferring mask prior and semantic similarity [EB/OL]. [2022-03-05]. https://arxiv.org/pdf/2110.14191.pdf

Liu Y C, Ma C Y, He Z J, Kuo C W, Chen K, Zhang P Z, Wu B C,

Kira Z and Vajda P. 2021b. Unbiased teacher for semi-supervised object detection//Proceedings of the 9th International Conference on Learning Representations. [s. l.]: ICLR

Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, Lin S and Guo B. 2021c. Swin transformer: hierarchical vision transformer using shifted windows//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 9992-10002 [DOI: 10.1109/ICCV48922.2021.00986]

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, USA: IEEE: 3431-3440 [DOI: 10.1109/CVPR.2015.7298965]

Luo W, Zhang T Z, Yang W F, Liu J E, Mei T, Wu F and Zhang Y D. 2021. Action unit memory network for weakly supervised temporal action localization//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 9964-9974 [DOI: 10.1109/CVPR46437.2021.00984]

Ma F, Zhu L C, Yang Y, Zha S X, Kundu G, Feiszli M and Shou Z. 2020. SF-Net: single-frame supervision for temporal action localization//Proceedings of Computer Vision – ECCV 2020 – 16th European Conference. Glasgow, UK: ECCV: 420-437

Ma J W, Gorti S K, Volkovs M and Yu G W. 2021. Weakly supervised action selection learning in video//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 7583-7592 [DOI: 10.1109/CVPR46437.2021.00750]

Mai J J, Yang M and Luo W F. 2020. Erasing integrated learning: a simple yet effective approach for weakly supervised object localization//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8763-8772 [DOI: 10.1109/CVPR42600.2020.00879]

Masita K L, Hasan A N and Shongwe T. 2020. Deep learning in object detection: a review//Proceedings of 2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems. Durban, South Africa: IEEE: 1-11 [DOI: 10.1109/icABCD49160.2020.9183866]

Moltisanti D, Fidler S and Damen D. 2019. Action recognition from single timestamp supervision in untrimmed videos//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 9907-9916 [DOI: 10.1109/CVPR.2019.01015]

Narayan S, Cholakkal H, Khan F S and Shao L. 2019. 3C-Net: category count and center loss for weakly-supervised action localization//Proceedigns of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea(South): IEEE: 8678-8686 [DOI: 10.1109/ICCV.2019.00877]

Nguyen P, Han B, Liu T and Prasad G. 2018. Weakly supervised action localization by sparse temporal pooling network//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 6752-6761 [DOI: 10.1109/CVPR.2018.00706]

Nguyen P, Ramanan D and Fowlkes C. 2019. Weakly-supervised action localization with background modeling//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 5501-5510 [DOI: 10.1109/ICCV.2019.00560]

Oh Y, Kim B and Ham B. 2021. Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 6909-6918 [DOI: 10.1109/CVPR46437.2021.00684]

Papadopoulos D P, Uijlings J R R, Keller F and Ferrari V. 2017. Extreme clicking for efficient object annotation//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 4940-4949 [DOI: 10.1109/ICCV.2017.528]

Papandreou G, Chen L C, Murphy K P and Yuille A L. 2015. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 1742-1750 [DOI: 10.1109/ICCV.2015.203]

Pardo A, Alwassel H, Heilbron F C, Thabet A and Ghanem B. 2021. RefineLoc: iterative refinement for weakly-supervised action localization//Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE: 3318-3327 [DOI: 10.1109/WACV48630.2021.00336]

Pathak D, Shelhamer E, Long J and Darrell T. 2015. Fully convolutional multi-class multiple instance learning//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: ICLR

Paul S, Roy S and Roy-Chowdhury A K. 2018. W-TALC: weakly-supervised temporal activity localization and classification//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 588-607 [DOI: 10.1007/978-3-030-01225-0_35]

Pinheiro P O and Collobert R. 2015. From image-level to pixel-level labeling with convolutional networks//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 1713-1721 [DOI: 10.1109/CVPR.2015.7298780]

Qian R, Wei Y C, Shi H H, Li J C, Liu J Y and Huang T S. 2019. Weakly supervised scene parsing with point-based distance metric learning//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI: 8843-8850

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 (CVPR). Las Vegas, USA: IEEE: 779-788 [DOI: 10.1109/CVPR.2016.91]

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 Z Z, Yu Z D, Yang X D, Liu M Y, Lee Y J, Schwing A G and Kautz J. 2020. Instance-aware, context-focused, and memory-efficient weakly supervised object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10595-10604 [DOI: 10.1109/CVPR42600.2020.01061]

Rother C, Kolmogorov V and Blake A. 2004. "GrabCut": interactive foreground extraction using iterated graph cuts. ACM Transactions on Graphics, 23(3): 309-314 [DOI: 10.1145/1015706.1015720]

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, 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]

Russell B C, Torralba A, Murphy K P and Freeman W T. 2008. LabelMe: a database and web-based tool for image annotation. International Journal of Computer Vision, 77(1/3): 157-173 [DOI: 10.1007/s11263-007-0090-8]

Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D and Batra D. 2020. Grad-CAM: visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision, 128(2): 336-359 [DOI: 10.1007/s11263-019-01228-7]

Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R and LeCun Y. 2014. OverFeat: integrated recognition, localization and detection using convolutional networks//Proceedings of the 2nd International Conference on Learning Representations. Banff, Canada: ICLR

Shao F F, Chen L, Shao J, Ji W, Xiao S N, Ye L, Zhuang Y T and Xiao J. 2021. Deep learning for weakly-supervised object detection and object localization: a survey [EB/OL]. [2022-03-05]. https://arxiv.org/pdf/2105.12694.pdf

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

Shi B F, Dai Q, Mu Y D and Wang J D. 2020. Weakly-supervised action localization by generative attention modeling//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1006-1016 [DOI: 10.1109/CVPR42600.2020.0010]

Shou Z, Gao H, Zhang L, Miyazawa K and Chang S F. 2018. AutoLoc: weakly-supervised temporal action localization in untrimmed videos//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 162-179 [DOI: 10.1007/978-3-030-01270-0_10]

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: ICLR

Singh K K and Lee Y J. 2017. Hide-and-seek: forcing a network to be meticulous for weakly-supervised object and action localization//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 3544-3553 [DOI: 10.1109/ICCV.2017.381]

Song C F, Huang Y, Ouyang W L and Wang L. 2019. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3131-3140 [DOI: 10.1109/CVPR.2019.00325]

Sui L, Zhang C L and Wu J X. 2021. Salvage of supervision in weakly supervised detection [EB/OL]. [2022-03-05]. https://arxiv.org/pdf/2106.04073.pdf

Tang P, Wang X G, Bai S, Shen W, Bai X, Liu W Y and Yuille A. 2020. 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]

Tang P, Wang X G, Bai X and Liu W Y. 2017. Multiple instance detection network with online instance classifier refinement//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 3059-3067 [DOI: 10.1109/CVPR.2017.326]

Tian Z, Shen C H and Chen H. 2020. Conditional convolutions for instance segmentation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 282-298 [DOI: 10.1007/978-3-030-58452-8_17]

Tian Z, Shen C H, Chen H and He T. 2019. FCOS: fully convolutional one-stage object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 9626-9635 [DOI: 10.1109/ICCV.2019.00972]

Tian Z, Shen C H, Wang X L and Chen H. 2021. BoxInst: high-performance instance segmentation with box annotations//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 5439-5448 [DOI: 10.1109/CVPR46437.2021.00540]

Uijlings J R R, Popov S and Ferrari V. 2018. Revisiting knowledge transfer for training object class detectors//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1101-1110 [DOI: 10.1109/CVPR.2018.00121]

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]

Vernaza P and Chandraker M. 2017. Learning random-walk label propagation for weakly-supervised semantic segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2953-2961 [DOI: 10.1109/CVPR.2017.315]

Wang J D, Sun K, Cheng T H, Jiang B R, Deng C R, Zhao Y, Liu D, Mu Y D, Tan M K, Wang X G, Liu W Y and Xiao B. 2021a. Deep high-resolution representation learning for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(10): 3349-3364 [DOI: 10.1109/TPAMI.2020.2983686]

Wang L M, Xiong Y J, Lin D H and Van Gool L. 2017. UntrimmedNets for weakly supervised action recognition and detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 6402-6411 [DOI: 10.1109/CVPR.2017.678]

Wang X G, Feng J P, Hu B, Ding Q, Ran L J, Chen X X and Liu W Y. 2021c. Weakly-supervised instance segmentation via class-agnostic learning with salient images//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 10220-10230 [DOI: 10.1109/CVPR46437.2021.01009]

Wang Y D, Zhang J, Kan M N, Shan S G and Chen X L. 2020. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 12272-12281 [DOI: 10.1109/CVPR42600.2020.01229]

Wei Y C, Feng J S, Liang X D, Cheng M M, Zhao Y and Yan S C. 2017a. Object region mining with adversarial erasing: a simple classification to semantic segmentation approach//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6488-6496 [DOI: 10.1109/CVPR.2017.687]

Wei Y C, Liang X D, Chen Y P, Shen X H, Cheng M M, Feng J S, Zhao Y and Yan S C. 2017b. 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]

Welinder P, Branson S, Mita T, Wah C, Schroff F, Belongie S and Perona P. 2010. Caltech-ucsd birds 200. California Institute of Technology. CNS-TR-2010-001

Xie C H, Ren D W, Wang L, Hu Q H, Lin L and Zuo W M. 2021. Learning class-agnostic pseudo mask generation for box-supervised semantic segmentation. [EB/OL]. [2022-03-05]. https://arxiv.org/pdf/2103.05463.pdf

Xie E Z, Sun P Z, Song X G, Wang W H, Liu X B, Liang D, Shen C H and Luo P. 2020. PolarMask: single shot instance segmentation with polar representation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 12190-12199 [DOI: 10.1109/CVPR42600.2020.01221]

Xie S N and Tu Z W. 2015. Holistically-nested edge detection//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 1395-1403 [DOI: 10.1109/ICCV.2015.164]

Xu C L and Ding L. 2018. Weakly-supervised action segmentation with iterative soft boundary assignment//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 6508-6516 [DOI: 10.1109/CVPR.2018.00681]

Xu Y L, Zhang C W, Cheng Z Z, Xie J W, Niu Y, Pu S L and Wu F. 2019. Segregated temporal assembly recurrent networks for weakly supervised multiple action detection//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI: 9070-9078

Yan G, Liu B X, Guo N, Ye X C, Wang 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 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea(South): IEEE: 9833-9842 [DOI: 10.1109/ICCV.2019.00993]

Yang K, Li D S and Dou Y. 2019. Towards precise end-to-end weakly supervised object detection network//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea(South): IEEE: 8371-8380 [DOI: 10.1109/ICCV.2019.00846]

Yang L, Han J W, Zhao T, Lin T W, Zhang D W and Chen J X. 2021a. Background-click supervision for temporal action localization. [J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://ieeexplore.ieee.org/document/9633199 [DOI: 10.1109/TPAMI.2021.3132058]

Yang W F, Zhang T Z, Yu X Y, Qi T, Zhang Y D and Wu F. 2021b. Uncertainty guided collaborative training for weakly supervised temporal action detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 53-63 [DOI: 10.1109/CVPR46437.2021.00012]

Yin Y F, Deng J J, Zhou W G and Li H Q. 2021. Instance mining with class feature banks for weakly supervised object detection//Proceedings of the 35th AAAI Conference on Artificial Intelligence. [s.l.]: AAAI: 3190-3198

Yu T, Ren Z, Li Y C, Yan E X, Xu N and Yuan J S. 2019a. Temporal structure mining for weakly supervised action detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea(South): IEEE: 5521-5530 [DOI: 10.1109/ICCV.2019.00562]

Yu Z, Zhuge Y, Lu H C and Zhang L H. 2019b. Joint learning of saliency detection and weakly supervised semantic segmentation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea(South): IEEE: 7222-7232 [DOI: 10.1109/ICCV.2019.00732]

Zeng Z Y, Liu B, Fu J L, Chao H Y and Zhang L. 2019. WSOD2: learning bottom-up and top-down objectness distillation for weakly-

supervised object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 8291-8299 [DOI: 10.1109/ICCV.2019.00838]

Zhai Y H, Wang L, Tang W, Zhang Q L, Yuan J S and Hua G. 2020. Two-stream consensus network for weakly-supervised temporal action localization//Proceedings of Computer Vision – ECCV 2020 – 16th European Conference. Glasgow, UK: ECCV: 37-54

Zhang B F, Xiao J M and Zhao Y. 2021a. Dynamic feature regularized loss for weakly supervised semantic segmentation [EB/OL]. [2022-03-05]. https://arxiv.org/pdf/2108.01296.pdf

Zhang C, Cao M, Yang D M, Chen J and Zou Y X. 2021b. CoLA: weakly-supervised temporal action localization with snippet contrastive learning//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 16005-16014 [DOI: 10.1109/CVPR46437.2021.01575]

Zhang D, Zhang H W, Tang J H, Hua X S and Sun Q R. 2020. Causal Intervention for Weakly-Supervised Semantic Segmentation//Advances in Neural Information Processing Systems. [s.l.]: NeurIPS

Zhang D W, Han J W, Cheng G and Yang M H. 2021c. Weakly supervised object localization and detection: a survey. [J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence [DOI: 10.1109/TPAMI.2021.3074313]

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

Zhang X L, Wei Y C, Kang G L, Yang Y and Huang T. 2018b. Self-produced guidance for weakly-supervised object localization//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 610-625 [DOI: 10.1007/978-3-030-01258-8_37]

Zhang Y Q, Bai Y C, Ding M L, Li Y Q and Ghanem B. 2018c. W2F: a weakly-supervised to fully-supervised framework for object detection//Proceedings of 2018 IEEE/CVF conference on computer vision and pattern recognition. Salt Lake City, USA: IEEE: 928-936 [DOI: 10.1109/CVPR.2018.00103]

Zhao H S, Shi J P, Qi X J, Wang X G and Jia J Y. 2017. Pyramid scene parsing network//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6230-6239 [DOI: 10.1109/CVPR.2017.660]

Zhong J X, Li N N, Kong W J, Zhang T, Li T H and Li G. 2018. Step-by-step erasion, one-by-one collection: a weakly supervised temporal action detector//Proceedings of the 26th ACM International Conference on Multimedia. Seoul, Korea(South): Association for Computing Machinery: 35-44 [DOI: 10.1145/3240508.3240511]

Zhong Y Y, Wang J F, Peng J and Zhang L. 2020. Boosting weakly supervised object detection with progressive knowledge transfer//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 615-631 [DOI: 10.1007/978-3-030-58574-7_37]

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

Zhou Y Z, Zhu Y, Ye Q X, Qiu Q and Jiao J B. 2018. Weakly supervised instance segmentation using class peak response//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 3791-3800 [DOI: 10.1109/CVPR.2018.00399]

Zhou Z H. 2004. Multi-instance learning: a survey. AI Lab, Department of Computer Science and Technology: 1-31

Zhou Z H, Sun Y Y and Li Y F. 2009. Multi-instance learning by treating instances as non-I. I. D. samples//Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Canada: ACM: 1249-1256 [DOI: 10.1145/1553374.1553534]

Zhou Z H, Zhang M L. 2007. Solving multi-instance problems with classifier ensemble based on constructive clustering. [J/OL] Knowledge and Information Systems. https://ieeexplore.ieee.org/document/9409690

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

作者简介

任冬伟,1988 年生,男,副教授,主要研究方向为视频图像复原、物体检测。
E-mail: csdren@hit.edu.cn

左旺孟,通信作者,男,教授,主要研究方向为底层视觉、图像视频生成、图像分类、物体检测。
E-mail:wmzuo@hit.edu.cn

王旗龙,男,副教授,主要研究方向为图像视频分类、物体检测、深层概率分布建模。E-mail:qlwang@tju.edu.cn
魏云超,男,教授,主要研究方向为弱监督学习、计算机视觉。
E-mail:yunchao.wei@bjtu.edu.cn
孟德宇,男,教授,主要研究方向为计算机视觉、机器学习。
E-mail:dymeng@mail.xjtu.edu.cn