第 60 卷 第 4 期/2023 年 2 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

特征自蒸馏机制下的弱监督目标检测

高文龙，陈莹*，彭勇

江南大学物联网工程学院轻工过程先进控制教育部重点实验室，江苏 无锡 214122

摘要 目前基于图像级注释信息的主流弱监督目标检测算法常常出现局部定位问题，仅仅关注图像中局部高辨别性的区域，却忽略了完整的目标。为了解决这种问题，提出了一种端对端的基于特征自蒸馏的弱监督目标检测网络(FSD-Net)，其中可拆卸的特征自蒸馏模块充分利用不同层级特征表示中的细节信息和语义信息，并通过特征自蒸馏损失约束网络训练，在未增加测试期计算代价的前提下增强了检测器综合性能；同时构造回归分支简单却有效地提取并利用特征中隐性位置信息，并通过改善监督信息生成算法、平衡优化损失等策略进一步改善了弱监督目标检测器的局部定位问题。在Pascal VOC 2007、VOC 2012、MS-COCO等大规模公开数据集上的实验结果表明，FSD-Net拥有比Baseline及近年主流方法更好的检测性能，有效地缓解了局部定位问题。

关键词 图像处理；目标检测；深度学习；弱监督学习；特征自蒸馏

中图分类号 TP391 **文献标志码** A **DOI**：10.3788/LOP212868

Weakly Supervised Object Detection Based on Feature Self-Distillation Mechanism

Gao Wenlong, Chen Ying*, Peng Yong

Key Laboratory of Advanced Process Control for Light Industry of Ministry of Education, School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, *Jiangsu, China*

Abstract The current mainstream weakly supervised object detection methods based on image-level annotation often occur local localization problem, tend to overfit the most discriminative regions, and ignore object integrity. To solve these existing problems, an end-to-end weakly supervised object detection network based on feature self-distillation (FSD-Net), in which the detachable feature self-distillation module fully uses the semantic and detailed information in the representation of different hierarchical features, is proposed. Additionally, through feature self-distillation loss constraint network training, the comprehensive performance of the detector is enhanced without increasing the calculation cost during the test period. Moreover, the regression branches are constructed to simply extract and effectively utilize the implicit location information in the features, improves the original supervision information generation algorithm, and balances optimization loss and other strategies to further improve the local localization problem of the weakly supervised object detector. Experiments on large-scale public datasets, such as Pascal VOC 2007, VOC 2012, and MS-COCO, show that FSD-Net has a better detection performance than the Baseline and other existing mainstream methods, effectively alleviating the local localization problem in weakly supervised object detection.

Key words image processing; object detection; deep learning; weakly supervised learning; feature self-distillation

1 引 言

计算机硬件和卷积神经网络的进步极大地促进了目标检测领域的发展，如SSD[1]、YOLO[2]、Faster RCNN[3]等优异的目标检测器不断涌现。这些目标检测算法通常依赖于大量强监督的训练样本(带有边界框标注信息)，而在现实生活中，有时难以提供足够多的样本以支撑目标检测网络的训练[4]。比如肿瘤、癌症等医疗图像领域，只有具有一定医学知识的工作人员乃至医学专家才可完成特定图像的标记(如病灶形状等)。因此，为了完成资源受限条件下的高精度目标检测，许多研究人员开始使用图像级标签(只有类别注

收稿日期：2021-11-03；**修回日期**：2021-12-06；**录用日期**：2021-12-21；**网络首发日期**：2022-01-11

基金项目：国家自然科学基金(62173160)

通信作者：*chenying@jiangnan.edu.cn

释信息）训练目标检测器，即弱监督目标检测（WSOD）[5]。如今，成熟的压缩超快摄影[6]、高带宽成像系统[7]、超分辨率[8]等技术使高质量图片样本的获取更加容易，相信仅需要弱监督标注信息的目标检测器有着日益光明的发展前景。

当今，弱监督目标检测算法[9-11]通常将弱监督检测看作一种多实例学习（MIL）任务；不同于元学习，多实例学习网络的训练在选择区域建议和预测区域建议得分间迭代进行。然而，弱监督检测器倾向于仅学习输入图像中最具辨别性的局部特征（如一张测试图像中动物的头部），而忽略了更加完整的目标特征，从而造成局部定位问题。

近年来，为了解决局部定位问题，从持续优化网络方向出发，Tang 等[12]提出了一种在线实例分类器优化网络（OICR），创新性地将基础多实例分类网络和多级实例优化器集成到单个网络中，将前一个阶段的输出作为下一个阶段的监督信息；Wan 等[13]将实例划分为多个空间和类别相关的子集，并设计了一系列平滑的损失函数来定位完整的目标范围。从优化区域建议方向出发，Lin 等[14]提出了一种基于空间外观图的目标实例挖掘算法（OIM），Zeni 等[15]通过平均 K 级优化分支的输出弥补实例优化过程中的信息损失；Cheng 等[16]结合选择性搜索和梯度加权类激活映射图生成目标定位更准确的区域建议框；Jiang 等[17]设计了一种动态区域建议采样策略（DPS）来渐进性消除背景对目标对象的影响，从而改善检测器的局部定位问题。也有学者开始构建检测任务和其他任务间的桥梁，如 Shen 等[18]通过多任务学习方案共同训练弱监督目标检测和分割任务，通过协同合作和循环训练来促进两个子网络跳出各自的局部最小值；李阳等[19]基于分类网络生成的类别显著图构建伪标注信息，并利用伪标注信息训练实时目标检测器。此外，Yin 等[20]设计了一种类特征库实例挖掘框架（IM-CFB），从更广泛的角度收集目标多样性信息，从而提高目标定位预测的准确性。

上述方法，均没有充分利用卷积神经网络不同层级的特征信息，从而导致局部定位问题依旧存在。因此，本文提出了一种基于特征自蒸馏的弱监督目标检测网络（FSD-Net），充分利用不同层级特征表示中的细节信息和抽象语义信息，并通过改进伪监督信息生成算法、平衡优化分支损失、引入回归分支等策略，进一步改善了弱监督目标检测器的综合性能。

2　特征自蒸馏机制下的弱监督目标检测网络

2.1　网络整体架构

本文构建了一个全新的端对端弱监督目标检测网络 FSD-Net，其整体架构如图 1 所示。首先，可拆卸的特征自蒸馏模块读取原始输入图像 I，基于卷积层、选择性搜索、感兴趣区域（ROI）池化、特征自蒸馏损失等提取丰富的特征信息；然后，将提取到的特征输送给全连接层 FC6、FC7 生成固定维度的特征向量 $\boldsymbol{V}$。最后，多实例学习分支和回归分支分别对输入特征向量 $\boldsymbol{V}$ 进行处理，生成目标分类和定位预测结果。

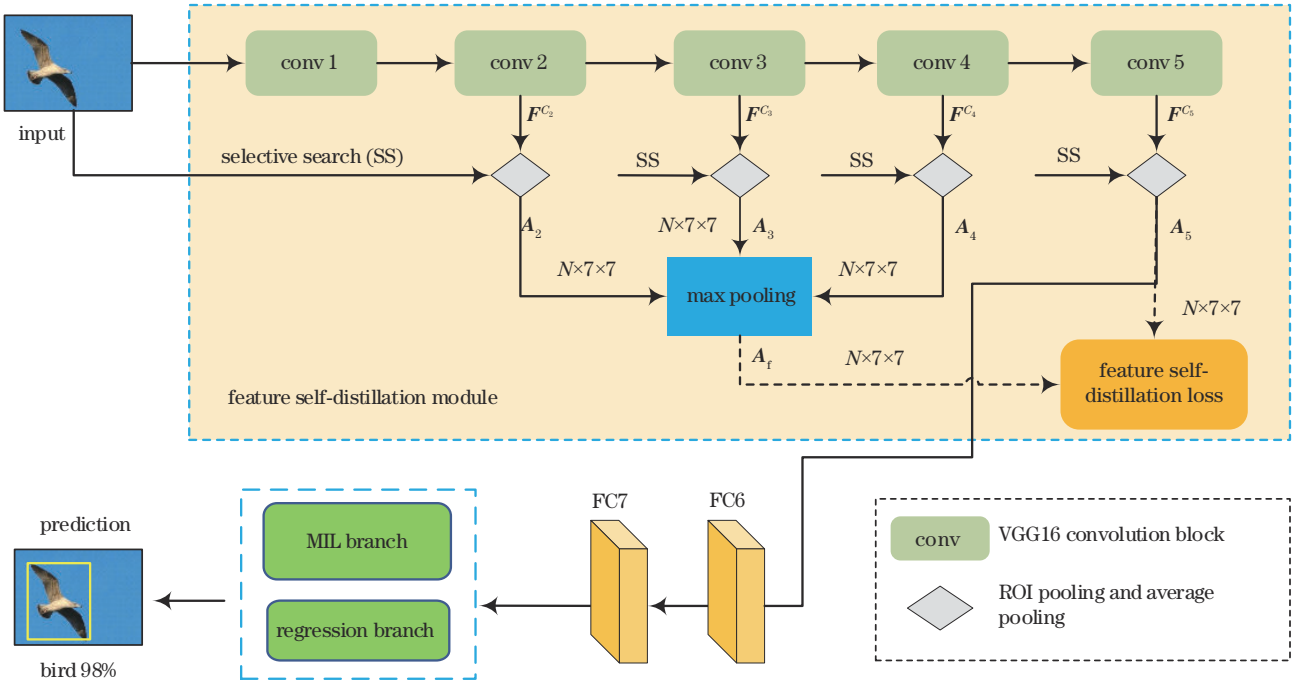

图 1　网络整体架构

Fig. 1　Overall structure of network

2.2　特征自蒸馏模块

卷积神经网络中，浅层网络感受野较小，网络提取到的特征和输入信息距离较近，特征中包含更多的细节信息；而深层网络，随着卷积次数的增加和池化等操作，网络感受野逐渐增大，不同感受野之间的重叠区域也不断增加，网络提取到的特征和输入信息距离较远，特征中包含更多的抽象语义信息。为了充分利用网络中的深层信息和浅层信息，本文对初始特征提取网络（VGG16）不同层级的特征信息进行自蒸馏，丰富VGG16的特征表示，以缓解弱监督目标检测器的局部定位问题。

如图1所示，将目标图像$I\in R^{W\times H\times D}$输入特征自蒸馏模块，其中$W$、$H$和$D$分别表示输入目标图像的宽度、高度和通道数，VGG16中的各个卷积块分别记为conv 1～conv 5，每个卷积块输出的特征图为$F^{C_1}\sim F^{C_5}$。本文通过选择性搜索生成区域建议P，并将区域建议P和不同层级输出的特征图F^{C_n}送入属性相同的ROI池化器中，获得长度和宽度相同的特征表示信息$F_{\mathrm{ROI}}\in R^{7\times 7\times D\times N}$，其中$N$为输入图像通过选择性搜索生成的区域建议的数量，卷积块conv $n[n\in(2,3,4)]$对应的特征表示F_{ROI}的通道数D分别为128、256和512。

为了充分利用不同层级的特征表示，本文在通道维度对多个F_{ROI}进行平均值池化获取维度相同的注意力图$A_n\in R^{7\times 7\times N}$，接着将这些注意力图进行最大值池化以获得特征信息更丰富的综合注意力图$A_{\mathrm{f}}\in R^{7\times 7\times N}$：

$$A_{\mathrm{f}}=\max(A_2,A_3,A_4)。\tag{1}$$

网络训练初期特征提取能力较差，此时获得的A_{f}噪声信息较多，直接使用可能会损害目标检测器的性能；基于此，本文未直接将A_{f}作为FC6、FC7等网络层的输入特征图，而是构造特征自蒸馏损失L_{fsd}作为目标蒸馏器，使得综合注意力图A_{f}和VGG16的原始输出$A_5\in R^{7\times 7\times N}$互相吸收彼此丰富的细节信息和抽象语义信息，特征信息更丰富的A_5促进A_{f}的发展，A_{f}驱动F^{C_n}进化，进化后的F^{C_n}再次促进A_5的发展，不断正反馈下，避免VGG16输出的仅关注图像中的高辨别性区域（如鸟的头部）却忽略完整的目标。假设特征图A_{f}与A_5在位置(x,y)的像素值差异记为$e(x,y)=A_{\mathrm{f}}(x,y)-A_5(x,y)$，特征自蒸馏损失$L_{\mathrm{fsd}}$可表示为

$$L_{\mathrm{fsd}}=\sum_x\sum_y l_{\mathrm{fsd}}[e(x,y)],\ \text{where}\ l_{\mathrm{fsd}}(e)=\begin{cases}0.5e^2 & |e|<1\\ |e|-0.5 & \text{otherwise}\end{cases}。\tag{2}$$

此外，由于先前的特征自蒸馏约束使得A_5已经拥有丰富的细节信息和语义信息，因此如图2所示，测试阶段仅使用VGG16基础网络提取输入图像的特征信息即可，未增加测试期计算代价却增强了弱监督目标检测器的综合性能。

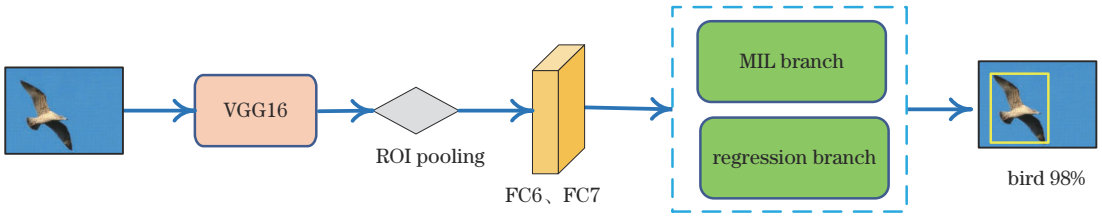

图2　测试期网络架构

Fig. 2　Network architecture during test period

2.3　多实例学习分支

由于弱监督目标检测任务中只有图像级注释信息（该图片中是否包含某个类别的目标），本文也基于OICR[12]构建了多实例学习分支来预测目标类别及位置。如图3左半部分所示，该多实例学习分支由基础检测器和K级优化器组成。

基础多实例检测器，通过两层全连接（FC）层和Softmax层获得粗糙的检测得分$\sigma(x^{\mathrm{det}})$和分类得分$\sigma(x^{\mathrm{cls}})$，接着通过两者逐元素相乘获得第r个区域建议的类别预测得分$x_r=\sigma(x^{\mathrm{det}})\cdot\sigma(x^{\mathrm{cls}})$，并对所有的区域建议求和以获取图像级预测得分$\varphi_c=\sum_{r=1}^{R}\boldsymbol{x}_{cr}$，最终计算预测得分$\varphi_c$和数据集真实类别标注信息$y_c$间的损失来约束基础多实例检测器的训练：

$$L_{\mathrm{cls}}=-\left[y_c\log\varphi_c+(1-y_c)\log(1-\varphi_c)\right],\tag{3}$$

式中：$y_c=1$表示该输入图像包含第c个类别目标；$y_c=0$则表示不包含此类别目标。

尽管基础多实例检测器有时可能会高度关注局部目标，但多个检测器检测到的局部目标可能会覆盖完整的目标对象，或者目标对象的较大部分。基于此，本文构建了多级优化分支，其中每级优化分支中都包含一个FC层和Softmax层，上一级优化器的输出作为下一级优化器的$C+1$维类别监督信息（如图3中多实例学习分支虚线所示），并用优化损失L_{ref}^k来指导优化器的训练：

$$L_{\mathrm{ref}}^k=-\frac{1}{|R|}\sum_{r=1}^{|R|}\sum_{c=1}^{C+1}w_r^k\,y_{cr}^k\log x_{cr}^k,\tag{4}$$

式中：k表示第k级优化器；y_{cr}^k为来自第$k-1$级优化器

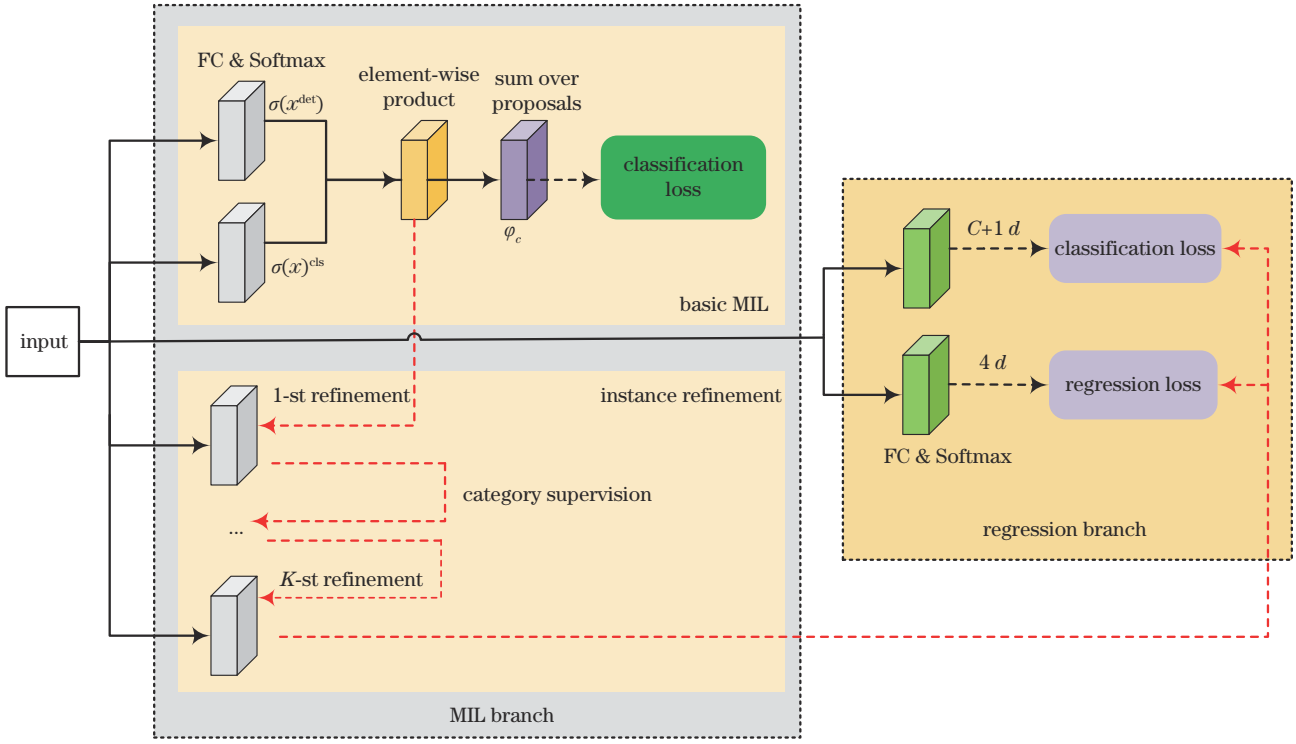

图 3　多实例学习和回归分支

Fig. 3　Multi-instance learning (MIL) and regression branch

的类别监督信息；x_{cr}^{k}为第r个区域建议关于类别c的预测得分；$w_r^k = x_{cj_c^{k-1}}^{k-1}$为抑制训练噪声的权重项，和OICR一样，取上一级优化器对最高得分区域建议j_c^{k-1}的类别预测得分。

改进监督信息生成算法：OICR的类别监督信息生成算法中，若区域建议r和最高得分区域建议j_c间的交并比（IoU）大于阈值λ，将区域建议r设置为类别c（即$y_{cr}^{k}=1$），否则直接设置为背景类[即$y_{(C+1)r}^{k}=1$]。一个图像中可能包含多个目标，若目标A和目标B（假设目标B为最高得分区域建议）相距较远（此时IoU很小），此算法会将目标A错标为背景类。为了修正这种问题，本文新增阈值λ_{ign}，若区域建议r和最高得分区域建议j_c间的IoU小于阈值λ_{ign}时，则将其损失权重w_r^k设为0，忽略此区域建议对优化损失函数L_{ref}^{k}的影响。

平衡优化分支损失：由于OICR中第1级优化分支的监督信息来自包含了两层Softmax逐元素相乘的基础多实例检测器，相较于只有1级Softmax的优化器，第1级优化分支的损失权重较小。为了平衡各优化分支间的损失，本文对第1级优化分支进行了额外的加权，最终K级优化分支（本文$K=3$）的损失L_{ref}为

$$L_{\text{ref}} = wL_{\text{ref}}^{1} + L_{\text{ref}}^{2} + L_{\text{ref}}^{3}。\tag{5}$$

2.4　回归分支

由于弱监督环境中不存在强监督信息，OICR单纯地基于目标分类结果预测目标位置，即多实例学习分支中得分最高区域建议的位置，这种粗糙的预测结果是非常不准确的。为了改善目标定位准确度，本文构建了类似于Fast RCNN[21]的伪强监督回归分支，如图3所示，其包含双路FC层及Softmax层。对于每个区域建议特征向量，上支路输出$C+1$维的目标类别预测结果$p=(p_0, p_1, \cdots, p_C)$及每个目标类别的边界框偏移结果$t^c=\{t_x^c, t_y^c, t_w^c, t_h^c\}$。其中，$t^c$表示相对于区域建议的尺度不变平移及对数空间的高度/宽度偏移量，具体细节可参见文献[21]。

即使无法从外部数据集获取实例监督信息，但MIL分支优化过的每个区域建议均对应着一个分类预测结果y及自身的边界框信息t。基于此，本文将第K级优化器输出最高得分的区域建议的类别信息、边界框信息作为“伪强监督信息”，并构造多任务损失L_{multi}指导回归分支的训练：

$$L_{\text{multi}} = L'_{\text{cls}}(p, y^*) + L_{\text{loc}}(t, t^*),\tag{6}$$

$$L_{\text{cls}}(p, y^*) = -\frac{1}{|R|}\sum_{r=1}^{|R|}\sum_{c=1}^{C+1} w_r y_{cr}^{*} \log p_{cr},\tag{7}$$

式中：分类损失L'_{cls}采用标准的多分类交叉熵格式约束目标分类预测，并取最高得分区域建议j_c的类别预测得分作为权重w_r抑制网络训练初期的噪声信息；r为区域建议索引；c为目标类别，y_{cr}^{*}为第K级优化器输出的类别监督信息；p_{cr}为第r个区域建议关于类别c的预测得分。

回归损失L_{loc}主要由$\text{Smooth}_{\text{L1}}$损失函数和$y_{cr}^{*}$组成，$y_r^{*}$保证了只有图像中包含目标对象（$y_{cr}^{*}=1$）时才激活回归损失，否则禁用回归损失。图片中的背景类不包含目标，对其进行边界框回归是毫无意义的，设

置 $y^*_{(C+1)r}=0$。

$$L_{\mathrm{loc}}(t,t^*)=\sum_{r=1}^{|R|}\sum_{c=1}^{C+1}\sum_{i}y^*_{cr}\cdot\mathrm{Smmoth}_{\mathrm{L1}}(t_i,t_i^*)。\quad(8)$$

因此，整个弱监督目标检测网络的最终约束 L 可表示为

$$L=L_{\mathrm{fsd}}+L_{\mathrm{cls}}+wL^1_{\mathrm{ref}}+L^2_{\mathrm{ref}}+L^3_{\mathrm{ref}}+L_{\mathrm{multi}}。\quad(9)$$

3　实验验证和分析

3.1　数据集介绍

本文在最常见的目标检测数据集 Pascal VOC 2007 和 Pascal VOC 2012 上对本文网络结构进行了广泛的实验，并将其性能与近年主流的方法进行了比较。

Pascal VOC 2007 数据集总共由 23080 张图片组成，并包含了人、交通工具（摩托车、飞机等）、动物（鸟、牛等）、室内物品（桌子、电视、沙发等）等 20 个常见的目标类别；Pascal VOC 2012 同样包含了 20 个类别的目标，但图像总数量为 54900 张，检测难度更高。

由于是弱监督环境（无边界框注释信息），因此仅使用 Pascal VOC 2007 和 VOC 2012 数据集中的图像级注释信息（该图像中是否包含某个类别的目标）训练目标检测器。无论输入的训练图像样本包含一个还是多个此类别的目标，该类别的图像级标签均为 1，不包含此类别目标时方为 0；基于此，无需对图像中的多个目标进行分割裁切获取单实例样本。

3.2　实验细节及评价标准

本文采用 OICR[12] 作为 Baseline，所有实验基于 Pytorch 深度学习框架实现，优化策略为小批量随机梯度下降法（mini-batch SGD），momentum 设置为 0.9，权重衰减为 5×10^{-4}，batchsize 为 4。模型训练时，总共迭代 25 K 次，前 10 K 次迭代的学习率为 5×10^{-4}，后 15 K 次迭代时学习率逐渐衰减为 5×10^{-5}。服务器环境为 Ubuntu 16.04（搭载有一张 Nvidia Geforce RTX 3090 显卡）。

弱监督目标检测中，最常用的检测指标为平均精度（mAP），其次为正确定位（CorLoc）。本文所有实验获得的 mAP 和 CorLoc 均遵循 Pascal VOC 所规定的计算标准：预测结果与 ground truth 间的 IoU 大于 0.5。

3.3　消融实验

3.3.1　各个改进模块的贡献

本文首先对各个改进模块的贡献，进行了相关的消融实验，实验结果如表 1 所示。本文 Baseline 取得了 46.5% 的 mAP，通过改进优化分支中监督信息生成算法，mAP 提高了 2.7 个百分点，达到 49.2%；通过引入权重 w 平衡优化分支间的损失，检测精度又提升了 1.8 个百分点，达到 51.0% 的 mAP；通过构造回归分支，mAP 再次提高了 1.6 个百分点；特征自蒸馏模块的加入，促使弱监督目标检测网络的 mAP 从 52.6% 提升至 54.8%。上述实验结果，有力地证明了本文各个模块的有效性。

表 1　各个改进模块对检测精度的贡献

Table 1　Contribution of each improved module to detection accuracy

Improved supervision generation algorithm	Balancing optimization loss	Regression branch	Feature self-distillation	mAP /%
−	−	−	−	46.5
+	−	−	−	49.2
+	+	−	−	51.0
+	+	+	−	52.6
+	+	+	+	**54.8**

3.3.2　特征自蒸馏

从 2.2 节描述可以知道，卷积神经网络浅层更关注于目标细节信息，而深层更关注于抽象语义信息。具体选择哪些层级的信息进行特征自蒸馏，可能对目标检测网络的最终精度造成不同的影响，实验结果如表 2 所示。若不进行特征自蒸馏，目标检测的 mAP 为 52.6%，当选取 conv2、conv3 和 conv4 输出的信息进行特征自蒸馏，丰富 conv5 输出特征表示时，取得了当前最高的检测 mAP 54.8%。

表 2　自蒸馏层级的影响

Table 2　Influence of feature self-distillation layers

Layer	−	conv 2+ conv 3	conv 2+ conv 4	conv 2+ conv 3+ conv 4
mAP /%	52.6	53.4	52.9	**54.8**

3.3.3　改进监督信息生成算法

K 级优化分支中，第 k 级输出为第 $k+1$ 级分支生成类别监督信息。其中，阈值 λ_{ign} 的大小决定了忽略目标的多少，从而对优化损失和检测精度造成影响，具体实验结果如表 3 所示。当阈值 λ_{ign} 为 0 时，完全不忽略和真实目标有较小 IoU 的目标，目标检测的 mAP 为 52.5%；随着阈值 λ_{ign} 的变大，被检测网络忽略的目标逐渐增多。当阈值 λ_{ign} 为 0.05 时，取得了当前最高的检测 mAP 55.3%；当阈值 λ_{ign} 为 0.2 时，由于忽略的目标过多，输送到下一级的类别监督信息非常不准确，从而严重破坏了检测精度。

3.3.4　平衡优化损失权重分析

为了平衡第 1 级优化损失和后面的 $K-1$ 级损

表 3 监督信息生成算法中不同阈值对检测精度的影响

Table 3 Influence of threshold on detection accuracy in supervision information generation algorithm

Layer	0.00	0.05	0.10	0.15	0.20
mAP /%	52.5	**55.3**	54.8	52.7	48.2

失，本文对第 1 级优化损失进行了加权，并通过广泛的实验探究了不同权重对目标检测精度的影响，具体精度如表 4 所示。可以看到，不加权（权重为 1.0）时，目标检测的 mAP 为 52.6%；权重为 3.0 时，目标检测精度明显优于其他几个权重值，mAP 达到了 55.3%。

表 4 平衡优化损失中不同权重对检测精度的影响

Table 4 Influence of weight on detection accuracy in balancing optimization loss

Weight w	-	2.0	2.5	3.0	3.5
mAP /%	52.6	52.8	54.5	**55.3**	53.0

3.4 与主流方法比较

在本小节中，将展示 FSD-Net 与近几年主流的弱监督目标检测方法的比较结果，进一步证明本文所提出网络 FSD-Net 的有效性。

表 5 和表 6 分别展示了本文提出的 FSD-Net 及近几年 SOTAs 方法在 Pascal VOC 2007 数据集 20 个类别上的检测性能和定位性能，其中粗体和下划线标注的数字分别表示本类别的最高准确度和次高准确度。由表 5 可知，FSD-Net 在自行车、瓶子、猫等 9 个类别中取得了最高准确度，在船、椅子、桌子等 4 个类别中取得了次高准确度；由表 6 可知，FSD-Net 分别取得了 9 个类别和 4 个类别的最高和次高定位准确度，有效地证明了 FSD-Net 的优越性。

表 5 VOC 2007 测试集上每个类别的检测性能

Table 5 Detection performance of each category on VOC 2007 test set unit: %

Method	OICR[12]	WSCDN[22]	MGR[23]	CMIL[13]	WSOD2[24]	CMIDN[25]	B-OICR[15]	OIM+IR[14]	FDC[26]	FSD-Net
Aero	60.6	61.2	55.2	62.5	65.1	53.3	**68.6**	55.6	61.7	61.0
Bicycle	67.1	66.6	66.5	58.4	64.8	71.5	62.4	67.0	72.3	**76.6**
Bird	44.3	48.3	40.1	49.5	**57.2**	49.8	55.5	45.8	50.1	51.9
Boat	24.5	26.0	31.1	32.1	**39.2**	26.1	27.2	27.9	23.9	34.9
Bottle	19.2	15.8	16.9	19.8	24.3	20.3	21.4	21.1	9.1	**28.0**
Bus	68.9	66.5	69.8	70.5	69.8	70.3	71.1	69.0	70.9	**74.0**
Car	65.9	65.4	64.3	66.1	66.2	69.9	**71.6**	68.3	67.8	70.7
Cat	55.8	53.9	67.8	63.4	61.0	68.3	56.7	70.5	56.7	**74.6**
Chair	25.7	24.7	27.8	20.0	**29.8**	28.7	24.7	21.3	6.1	29.1
Cow	49.7	61.2	52.9	60.5	64.6	65.3	60.3	60.2	56.6	**70.7**
Table	43.7	46.2	47.0	**52.9**	42.5	45.1	47.4	40.3	40.8	49.5
Dog	47.4	53.5	33.0	53.5	60.1	64.6	56.1	54.5	61.8	**65.6**
Horse	33.8	48.5	60.8	57.4	**71.2**	58.0	46.4	56.5	55.8	60.5
Mbike	66.7	66.1	64.4	68.9	70.7	71.2	69.2	70.1	66.0	**74.8**
Person	10.6	12.1	13.8	8.4	**21.9**	20.0	2.7	12.5	3.3	16.0
Plant	24.8	22.0	26.0	24.6	**28.1**	27.5	22.9	25.0	20.6	25.3
Sheep	38.0	49.2	44.0	51.8	**58.6**	54.9	41.5	52.9	48.2	54.7
Sofa	52.5	53.2	55.7	58.7	59.7	54.9	47.7	55.2	60.5	**63.9**
Train	64.5	66.2	68.9	66.7	52.2	69.4	71.1	65.0	59.7	**72.0**
TV	66.2	59.4	65.5	63.5	64.8	63.5	**69.8**	63.7	58.0	52.2

此外，表 7 和表 8 分别展示了 FSD-Net 及 SOTAs 方法在 Pascal VOC 2007 和 VOC 2012 数据集上取得的 mAP 和 CorLoC 结果。FSD-Net 在 Pascal VOC 2007 和 Pascal VOC 2012 数据集上，分别取得了 55.3% 的 mAP、70.4% 的 CorLoc 和 48.7% 的 mAP、68.0% 的 CorLoc，明显优于 OICR、C-MIL、WSOD2、FDC、DPS 等近年主流的弱监督目标检测算法，证明了本文 FSD-Net 的卓越性能。

为了进一步证明本文 FSD-Net 的有效性和鲁棒性，本文在另一个大型公开数据集 MS-COCO 上开展了相关实验，训练数据选择 train 2014，测试数据选择 minival 2014，具体实验结果如表 9 所示。可以清楚地看到，本文所提出的 FSD-Net 在 MS-COCO 数据集上取得了 22.7% AP_{50}（IoU 取 0.5 时计算得到的平均准确度）的测试精度，相比 Baseline 有 8.4 个百分点的性能提升，优于 WSCDN、C-MIDN、CSC、PG-PS、Grading-Net 等近几年主流的弱监督目标检测算法。

表 6 VOC 2007 训练验证集上每个类别的定位性能

Table 6 Localization performance of each category on VOC 2007 trainval set unit: %

Method	OICR[12]	WSCDN[22]	PGE[27]	MGR[23]	WSOD2[24]	B-OICR[15]	FDC[26]	FSD-Net
Aero	81.7	85.8	85.5	81.7	**87.1**	86.7	83.6	80.8
Bicycle	80.4	80.4	79.6	81.2	80.0	73.3	86.8	**88.6**
Bird	48.7	73.0	68.1	58.9	**74.8**	72.4	64.6	64.3
Boat	49.5	42.6	55.1	54.3	**60.1**	55.3	35.9	54.8
Bottle	32.8	36.6	33.6	37.8	36.6	46.9	25.4	**48.9**
Bus	81.7	79.7	83.5	83.2	79.2	83.2	82.3	**84.8**
Car	85.4	82.8	83.1	86.2	83.8	**87.5**	**87.5**	85.0
Cat	40.1	66.0	78.5	77.0	70.6	64.5	65.6	**81.7**
Chair	40.6	34.1	42.7	42.1	43.5	44.6	18.4	**44.8**
Cow	79.5	78.1	79.8	83.6	**88.4**	76.7	78.0	86.3
Table	35.7	36.9	37.8	**51.3**	46.0	46.4	46.0	46.8
Dog	33.7	68.6	61.5	44.9	74.7	70.9	76.0	**79.5**
Horse	60.5	72.4	74.4	78.2	**87.4**	67.0	80.8	83.3
Mbike	88.8	91.6	88.6	90.8	90.8	88.0	90.6	**94.0**
Person	21.8	22.2	32.6	20.5	**44.2**	9.6	7.6	27.4
Plant	57.9	51.3	55.7	56.8	52.4	56.4	53.9	**59.7**
Sheep	76.3	79.4	77.9	74.2	**81.4**	69.1	70.8	76.3
Sofa	59.9	63.7	63.7	66.1	61.8	52.4	**75.1**	68.5
Train	75.3	74.5	78.4	81.0	67.7	79.8	77.4	**81.4**
TV	81.4	74.6	74.1	**86.0**	79.9	82.8	82.4	71.7

表 7 VOC 2007 数据集上与主流方法的比较

Table 7 Comparison with mainstream methods in Pascal VOC 2007 dataset

Method	mAP /%	CorLoc /%
WSCDN[22]	48.3	64.7
PGE[27]	47.6	66.7
MGR[23]	48.6	66.8
C-MIL[13]	50.5	65.0
WSOD2[24]	53.6	69.5
C-MIDN[25]	52.6	68.7
CSC[28]	43.0	62.2
B-OICR[15]	49.7	65.7
OIM+IR[14]	50.1	67.2
FDC[26]	47.5	64.4
ICM[29]	54.8	68.8
PG-PS[16]	51.1	69.2
MGML[30]	53.0	67.1
DPS[17]	50.9	66.5
GradingNet-MELM[31]	52.5	63.2
OICR[12](Baseline)	46.5	60.6
FSD-Net	**55.3**	**70.4**

表 8 VOC 2012 数据集上与主流方法的比较

Table 8 Comparison with mainstream methods in Pascal VOC 2012 dataset

Method	mAP /%	CorLoc /%
WSCDN[22]	43.3	65.2
PGE[27]	43.4	66.7
C-MIL[13]	46.6	67.4
CSC[28]	37.1	61.4
FDC[26]	44.2	65.1
B-OICR[15]	46.7	66.3
OIM+IR[14]	45.3	67.1
PG-PS[16]	48.3	**68.7**
MGML[30]	48.5	67.4
DPS[17]	43.8	-
GradingNet-MELM[31]	48.6	62.8
OICR[12](Baseline)	37.9	62.1
FSD-Net	**48.7**	68.0

3.5 可视化

为了更直观地证明本文方法的有效性，图 4 展示了 Baseline 及 FSD-Net 最后一层卷积层 conv5 提取到的特征，其中第一列图片为原始输入图像，第二列为 Baseline 提取到的特征，最后一列为 FSD-Net 提取到的特征。可以清楚地看到，FSD-Net 提取到的特征更加完整和准确，较好地解决了 Baseline 仅聚焦于高辨别性头部区域的问题。

表 9　MS-COCO 数据集上与主流方法的比较

Table 9　Comparison with mainstream methods in MS-COCO dataset

Method	AP_{50} /%
WSCDN[22]	11.5
MELM[32]	18.8
PCL[33]	19.4
WS-JDS[18]	20.3
C-MIDN[25]	21.4
CSC[28]	20.3
PG-PS[16]	20.7
GradingNet-MELM[31]	22.6
OICR[12](Baseline)	14.3
FSD-Net	**22.7**

此外，图 5 展示了 FSD-Net 和 Baseline 在 Pascal VOC 2007 数据集上的预测结果，其中密集点线框、实线框及虚线框分别表示真实边界框、预测正确的边界框（和真实边界框的 IoU 大于等于 0.5）和预测错误的边界框，预测框的左上方为预测类别及其置信度。可以清楚地看到，Baseline（图 5 第 1、3、5 列）容易陷入局部最优解，目标检测网络仅定位了狗、猫等目标高辨别性的头部却忽略了完整的目标，本文提出的 FSD-Net（图 5 第 2、4、6 列）显著地改善了这种问题。此外，FSD-Net 也有效地修正了 Baseline 中预测框过大等问题（图 5 第 2 行）。对于开放世界中“多个物体存在于同一张画面”这一应用场景（图 5 第 3 行），FSD-Net 也展示了其有效性和鲁棒性。

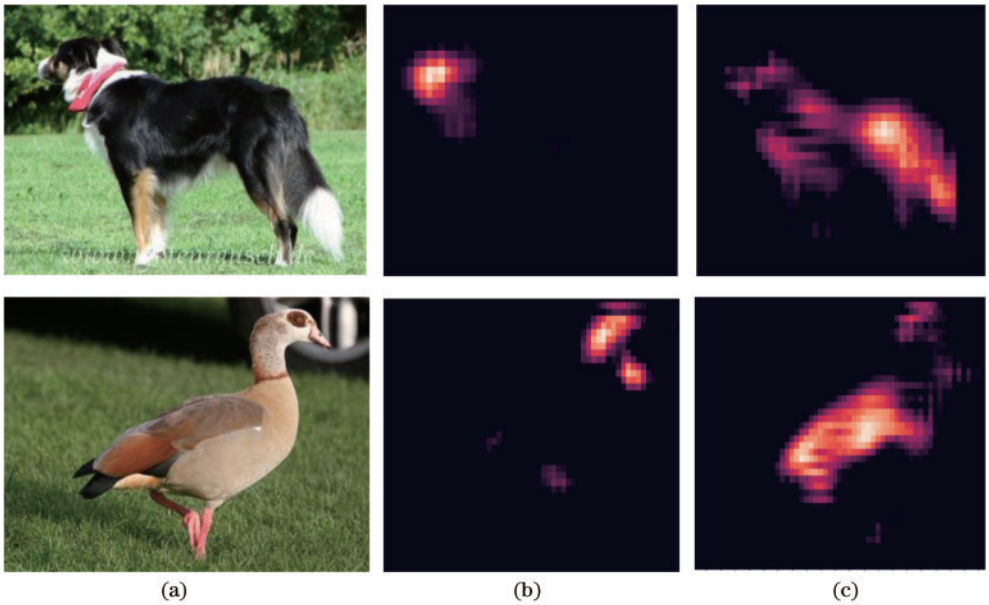

图 4　Baseline 及 FSD-Net conv5 层输出特征可视化。(a)原始输入图像；(b)Baseline 提取到的特征；(c)FSD-Net 提取到的特征

Fig. 4　Visualization of features extracted from Baseline and FSD-Net conv5 layer. (a) Original input image; (b) features extracted from Baseline; (c) features extracted from FSD-Net

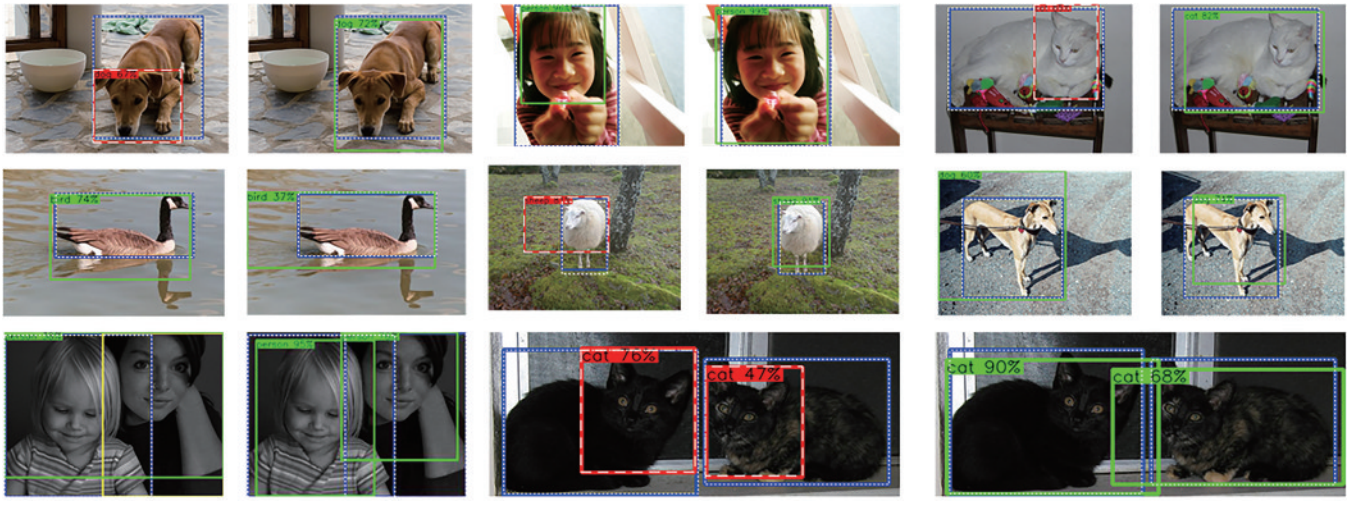

图 5　Baseline（第 1、3、5 列）和 FSD-Net（第 2、4、6 列）预测结果的可视化

Fig. 5　Visualization of prediction results of Baseline (columns 1, 3, and 5) and FSD-Net (columns 2, 4, and 6)

4 结 论

为了解决弱监督目标检测中局部定位问题，提出了一种FSD-Net，其充分利用不同层级特征表示中的细节信息和语义信息，并通过改进伪监督信息生成算法、引入回归分支等策略，在Pascal VOC 2007和VOC 2012数据集上mAP分别达到了55.3%和48.7%，在MS-COCO数据集上AP_{50}达到了22.7%，明显改善了弱监督目标检测器的局部定位问题及综合性能。

参 考 文 献

[1] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9905: 21-37.

[2] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 779-788.

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

[4] 罗会兰, 陈鸿坤. 基于深度学习的目标检测研究综述[J]. 电子学报, 2020, 48(6): 1230-1239.
Luo H L, Chen H K. Survey of object detection based on deep learning[J]. Acta Electronica Sinica, 2020, 48(6): 1230-1239.

[5] Bilen H, Vedaldi A. Weakly supervised deep detection networks[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 2846-2854.

[6] Qi D L, Zhang S A, Yang C S, et al. Single-shot compressed ultrafast photography: a review[J]. Advanced Photonics, 2020, 2(1): 014003.

[7] Park J, Brady D J, Zheng G A, et al. Review of bio-optical imaging systems with a high space-bandwidth product[J]. Advanced Photonics, 2021, 3(4): 044001.

[8] Zhou Z H, Liu W, He J J, et al. Far-field super-resolution imaging by nonlinearly excited evanescent waves[J]. Advanced Photonics, 2021, 3(2): 025001.

[9] Li D, Huang J B, Li Y L, et al. Weakly supervised object localization with progressive domain adaptation [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 3512-3520.

[10] Diba A L, Sharma V, Pazandeh A, et al. Weakly supervised cascaded convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 5131-5139.

[11] Jie Z Q, Wei Y C, Jin X J, et al. Deep self-taught learning for weakly supervised object localization[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 4294-4302.

[12] Tang P, Wang X G, Bai X, et al. Multiple instance detection network with online instance classifier refinement [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 3059-3067.

[13] Wan F, Liu C, Ke W, et al. C-MIL: continuation multiple instance learning for weakly supervised object detection[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 2194-2203.

[14] Lin C H, Wang S W, Xu D Q, et al. Object instance mining for weakly supervised object detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11482-11489.

[15] Zeni L F, Jung C R. Distilling knowledge from refinement in multiple instance detection networks[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 14-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 3324-3333.

[16] Cheng G, Yang J Y, Gao D C, et al. High-quality proposals for weakly supervised object detection[J]. IEEE Transactions on Image Processing, 2020, 29: 5794-5804.

[17] Jiang W H, Zhao Z C, Su F, et al. Dynamic proposal sampling for weakly supervised object detection[J]. Neurocomputing, 2021, 441: 248-259.

[18] Shen Y H, Ji R R, Wang Y, et al. Cyclic guidance for weakly supervised joint detection and segmentation[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 697-707.

[19] 李阳, 王璞, 刘扬, 等. 基于显著图的弱监督实时目标检测[J]. 自动化学报, 2020, 46(2): 242-255.
Li Y, Wang P, Liu Y, et al. Weakly supervised real-time object detection based on saliency map[J]. Acta Automatica Sinica, 2020, 46(2): 242-255.

[20] Yin Y F, Deng J J, Zhou W G, et al. Instance mining with class feature banks for weakly supervised object detection[C]//The 35rd AAAI Conference on Artificial Intelligence, February 2-9, 2021, Virtual Event. Virigina: AAAI Press, 2021: 3190-3198.

[21] Girshick R. Fast R-CNN[C]//2015 IEEE International Conference on Computer Vision, December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 1440-1448.

[22] Wang J J, Yao J C, Zhang Y, et al. Collaborative learning for weakly supervised object detection[C]// Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, July 13-19, 2018, Stockholm, Sweden. California: International Joint Conferences on Artificial Intelligence Organization, 2018:

971-977.

[23] Yang K, Li D S, Dou Y. Towards precise end-to-end weakly supervised object detection network[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 8371-8380.

[24] Zeng Z Y, Liu B, Fu J L, et al. WSOD2: learning bottom-up and top-down objectness distillation for weakly -supervised object detection[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 8291-8299.

[25] Yan G, Liu B X, Guo N, et al. C-MIDN: coupled multiple instance detection network with segmentation guidance for weakly supervised object detection[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 9833-9842.

[26] Zhang D W, Han J W, Zhao L, et al. From discriminant to complete: reinforcement searching-agent learning for weakly supervised object detection[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(12): 5549-5560.

[27] Kosugi S, Yamasaki T, Aizawa K. Object-aware instance labeling for weakly supervised object detection [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 6063-6071.

[28] Shen Y H, Ji R R, Yang K Y, et al. Category-aware spatial constraint for weakly supervised detection[J]. IEEE Transactions on Image Processing, 2020, 29: 843-858.

[29] Ren Z Z, Yu Z D, Yang X D, et al. Instance-aware, context-focused, and memory-efficient weakly supervised object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 10595-10604.

[30] Ji R Y, Liu Z Y, Zhang L B, et al. Multi-peak graph-based multi-instance learning for weakly supervised object detection[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2021, 17(2s): 70.

[31] Jia Q F, Wei S K, Ruan T, et al. GradingNet: towards providing reliable supervisions for weakly supervised object detection by grading the box candidates[C]//The 35rd AAAI Conference on Artificial Intelligence, February 2-9, 2021, Virtual Event. Virigina: AAAI Press, 2021: 1682-1690.

[32] Wan F, Wei P X, Jiao J B, et al. Min-entropy latent model for weakly supervised object detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 1297-1306.

[33] Tang P, Wang X G, Bai S, et al. PCL: proposal cluster learning for weakly supervised object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(1): 176-191.