© 山东大学科技期刊社版权所有

文章编号:1671-9352(2025)01-0001-13 DOI:10.6040/j.issn.1671-9352.7.2023.3979

双注意力引导特征融合的半弱监督目标检测

陈俊芬,李娜娜,谢博鋆*,张杰
(河北大学数学与信息科学学院, 河北 保定 071002)

摘要:为了降低标注成本,解决目标定位不准、细节信息遗漏等问题,提出双注意力引导特征融合的半弱监督目标检测算法,利用全标记和弱标记数据来平衡检测性能和标注成本,使用空间注意力将低层特征图与高层特征图进行像素级加权融合,使高层特征图具有丰富的低层信息,对融合后的特征图进行通道加权运算,得到细节、位置信息丰富的高层特征图。为了得到更准确的伪标注框,提出更具鲁棒性的候选框筛选策略。实验表明,本文提出的算法具有较优的检测性能,减少了全标记图像的数据量和额外的图像级标注。

关键词:弱监督目标检测;特征融合;注意力机制;半监督学习
中图分类号:TP391 **文献标志码**:A
引用格式:陈俊芬,李娜娜,谢博鋆,等.双注意力引导特征融合的半弱监督目标检测[J]. 山东大学学报(理学版),2025,60(1):1-13.

Semi-weakly supervised object detection using bi-attention-guided feature fusion

CHEN Junfen, LI Nana, XIE Bojun*, ZHANG Jie
(College of Mathematics and Information Science of Hebei University, Baoding 071002, Hebei, China)

Abstract: In order to reduce the cost of annotation and solve the problems of inaccurate target localization and omission of detail information, a semi-weakly supervised object detection method with bi-attention-guided feature fusion is proposed. Based on the method which fully labelled and weakly labelled data, the detection performance and annotation cost are balanced, and the spatial attention the low-level feature maps with the high-level feature maps with pixel-level weighting are fused, so that the high-level feature maps have rich low-level information, and performs channel-weighting operations on the fused feature maps to obtain high-level feature maps having rich details and location information. In order to get more accurate pseudo-labelled boxes, a more robust candidate box selection strategy is proposed. The proposed algorithm has better detection performance and reduce the amount of full-labeled image data and additional image-level labeling.

Key words: weakly supervised object detection; feature fusion; attention mechanism; semi-supervised learning

0 引言

目标检测是计算机视觉中的一项基本任务。在既有类别信息又有位置信息(称为细粒度边界框标注)的数据集上,全监督目标检测表现出良好的性能[1-3],然而,细粒度边界框的标注工作更耗时、费力。弱监督目标检测(weakly supervised object detection, WSOD)[4]仅依赖图像类别标签进行学习,减少了标注的工作量,因此,研究弱监督目标检测具有重要的应用价值。

计算机视觉算法通常将弱监督目标检测视为多示例学习(multi-instance learning, MIL)[5]问题。这类算法在输入图像上生成目标候选区域,将生成的候选区域作为实例,每一幅图像视为一个包含候选目标的示例包,在多示例学习约束下训练目标检测器。这些算法在高层特征图上进行预测,高层特征图分辨率低,导

收稿日期:2023-05-04; **网络出版时间**:2024-02-28 13:36:03
基金项目:河北省引进留学人员资助项目(C20200302); 河北省教育教学改革研究与实践项目(2020GJJG007)
第一作者:陈俊芬(1976—),女,副教授,博士,研究方向为机器学习及计算机视觉. E-mail: chenjunfen2010@ 126.com
*通信作者:谢博鋆(1981—),男,副教授,硕士生导师,博士,研究方向为机器学习及计算机视觉. E-mail: xiebojun@ 126.com

致像素占比小的物体的信息丢失,容易出现漏检现象。因为缺少物体整体信息的引导,检测器经常在物体最有区别性的部分产生较小边界框,所以,在高层特征图中同时保留细节信息和全局信息是一项非常有意义的任务。

本文提出了一种用双注意力引导特征融合的半弱监督目标检测算法,利用空间注意力将包含更多位置、细节的低层特征图信息融合到预测特征层,跨层融合后对不同的通道加权,提升了图像有效区域的特征提取效果,自下而上的跨层融合更有效地实现了特征层之间的信息交互。在候选框筛选模块中提出了更具鲁棒性的伪标注框采样策略:训练前期使用置信度较高的候选框作为伪标注框以减少噪声累积,后期随机采样候选框作为伪标注框,有效提升了模型检测精度。在检测部分用软式非极大值抑制(soft non-maximum suppression, Soft-NMS)[6]代替非极大值抑制(non maximum suppression, NMS)来解决漏检问题,提升了检测的召回率。

1 相关工作

1.1 弱监督目标检测

根据快速的基于区域的卷积神经网络(fast region-based convolutional networks, Fast RCNN)[7],Bilen等[8]提出了一种弱监督深度检测(weakly supervised deep detection network, WSDDN)算法,2个分支网络同时执行区域选择和分类任务,将区域得分进行加权求和得到图像类别的置信度。Tang等[9]提出了在线示例分类器更新(online instance classifier refinement, OICR)算法,通过多阶段示例分类器迭代细化预测;Yang等[10]在主干中添加了一个分类损失引导的注意力模块,用来提取特征图中潜在的位置信息;Wan等[11]在损失函数中添加2个熵来最小化目标定位的随机性。此后,Wan等[12]又提出了连续多示例学习(continuation multiple instance learning, C-MIL)算法,使用一系列平滑损失函数解决非凸的多示例学习池化问题。这些算法都使用多示例学习将目标检测问题转移为图像分类问题,但是有助于分类的信息并不一定有助于定位。为此,Tang等[13]提出了建议聚类学习(proposal cluster learning, PCL)算法,将生成的候选区域聚类成不同的类簇,通过迭代改进示例分类器的性能。Ren等[14]引入了示例关联的空间多样化约束,并建立了参数化的空间丢弃块来解决示例模糊和不完全定位的问题。很多弱监督检测算法都通过改进生成的伪标注框来提高检测效率。Li等[15]提出了一种两步策略,通过掩模分类收集较准确的候选框,将最优候选框作为伪标注框对检测器进行微调。Zhu等[16]使用弱监督构建了一个候选框生成器,将特征图转换成分数图,将其输入到分类层。Arun等[17]训练2个协同网络,其中一个是带有额外通道的条件网络,目标是共同最小化预测分布和条件分布之间的差异。

尽管上述弱监督检测算法取得了一定的成功,但是缺少位置信息引导,这些算法检测性能仍然远远落后于全监督检测器。由于训练集中的所有图像都是弱标记图像,只有一个子集带有边界框标注,因此利用带边界框标注的图像训练检测器可纠正弱监督检测器的错误预测[18]或生成更可靠的伪标注框[19]。Chen等[20]联合利用全标记数据和弱标记数据,提出了一个两阶段的弱监督和半监督相结合的目标检测框架。为了减少检测器对多阶段训练的依赖,Meethal等[21]提出了一种新的半弱监督训练算法,利用弱标记图像和一小部分全标记图像,以在线方式产生伪标注框,提高了检测性能。与这些算法类似,本文使用少量的全标记数据训练模型,并加入新的特征跨层融合方式,以缩小全监督与弱监督检测器间的性能差距。

1.2 特征融合

特征融合算法在目标检测[22]、微表情识别[23]等领域发挥了极其重要的作用。融合不同层的特征可以提高目标检测的性能,对解决漏检问题有很大的帮助。特征金字塔网络(feature pyramid networks, FPN)算法[24]利用自上而下和水平连接的方式融合高层和低层特征,改进了特征金字塔结构,成为多尺度特征融合的经典算法。在此之后,路径聚合网络(path aggregation network, PANet)[25]、加权双向特征金字塔网络(bidirectional feature pyramid network, BiFPN)[26]等多尺度特征融合算法(如图1)相继提出。图1中P3为第三层级的特征图,同理P4—P7为对应层级的特征图,每一层级包含多个特征图。

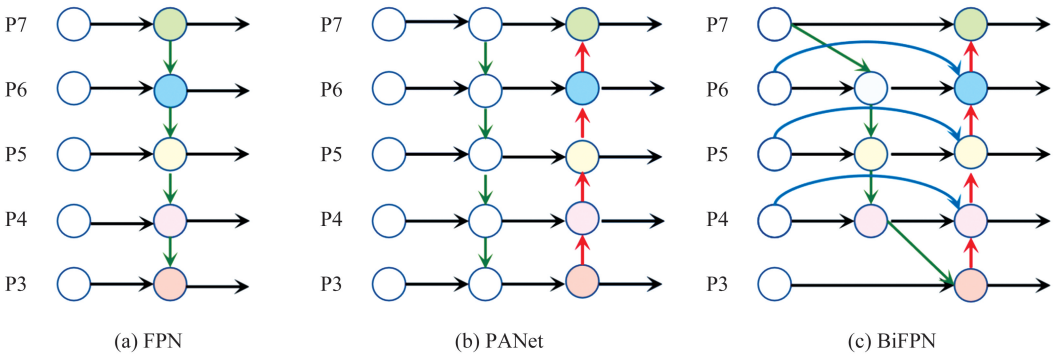


图 1　不同的特征融合方式

Fig.1　Different feature fusion methods

BiFPN 算法是由一系列的特征层级组成的一种自下而上和自上而下的双向特征金字塔网络,该算法融合不同层级和不同尺度的特征图,以获得更全面、更具有语义信息和更准确的特征表示。具体来说,BiFPN 算法使用反向连接自下而上和自上而下,将 2 个相邻层级的特征图进行融合。这种反向连接的特点是将低分辨率和高分辨率特征图的信息相互结合,获取更具有语义信息的特征表示。PANet 算法的结构包括 1 个主干网络和 1 个特征金字塔网络。特征金字塔网络接收主干网络的输出,将其传递给多个尺度的特征金字塔网络分支。每个分支对特征进行卷积和上采样,以生成相应尺度的特征图。PANet 算法引入了自下而上和自上而下的路径聚合机制,实现多尺度特征的信息交换和融合。自下而上的路径聚合将较低分辨率的特征图与较高分辨率的特征图级联,自上而下的路径聚合将较高分辨率的特征图与较低分辨率的特征图进行融合。PANet 算法可以在多个尺度上进行有效的目标检测,并具有更高的检测精度。

这些算法都可以有效地融合多尺度和多层次的特征图,提高目标检测的精度和效率,但这些特征融合方式较复杂,需要大量的全标记数据才能训练出较好的性能。相较于以上特征融合算法,本文提出的特征融合模块“即插即用”,仅使用弱标记数据就能训练出较好性能,在较小规模的模型上更能展现出其优势。

1.3　注意力机制

注意力机制可以促使网络专注于与目标相关的特征,大大提高了信息处理的效率与准确度。Seq2Seq 模型[27]首次提出注意力模块,之后出现了全局注意力模块和局部注意力模块[28]、卷积块注意力模块(convolutional block attention module, CBAM)[29]等不同变体。注意力模块加强特征提取的算法[30]在深度学习领域受到更多关注。CBAM 模块由通道注意力模块(channel attention module, CAM)和空间注意力模块(spatial attention module, SAM)组成。通道注意力模块通过 1 个全局平均池化层和 2 个全连接层来实现,学习不同通道之间的相互关系,更好地捕捉图像中的重要特征。空间注意力模块通过 1 个卷积层和 1 个全连接层来实现,用于调整原始特征图,使重要位置得到更大的响应,提高模型的性能。与 CBAM 不同的是,本文使用空间注意力模块和通道注意力模块进行自下向上式跨层的特征融合和最大池化。实验证明,空间注意力模块和通道注意力模块的结合可以更好地捕捉图像的关键信息。

2　改进的半弱监督目标检测框架

本文提出了双注意力引导特征融合的半弱监督目标检测算法,其框架如图 2 所示,选用 Faster RCNN[31]作为基础目标检测框架。输入图像是弱标记图像,使用选择性搜索(selective search, SS)算法[32]生成大量候选框,然后进入候选框筛选模块,此筛选模块将置信度较低的候选框过滤掉,筛选出置信度较高的候选框并进行重要性采样,将采样得到的候选框作为伪标注框。全标记图像和带有伪标注框的弱标记图像作为输入图像,残差网络(residual network, ResNet)[33]作为骨干网络提取特征,conv2_x 指 ResNet 网络结构中的第 2 部分,由多个卷积块组成; conv4_x 指 ResNet 网络结构中的第 4 部分,该部分也是由多个卷积块组成。将 conv2_x 的输出作为特征融合模块的输入特征层 layer1;conv4_x 的输出作为特征融合模块的输入特征层 layer3。接着进行感兴趣区域池化(region of interest pooling, ROI Pooling),经全连接层(fully connected layers, FC)后完成预测。最后,用输出的预测框类别置信度更新候选框筛选模块中候选框的置信度。

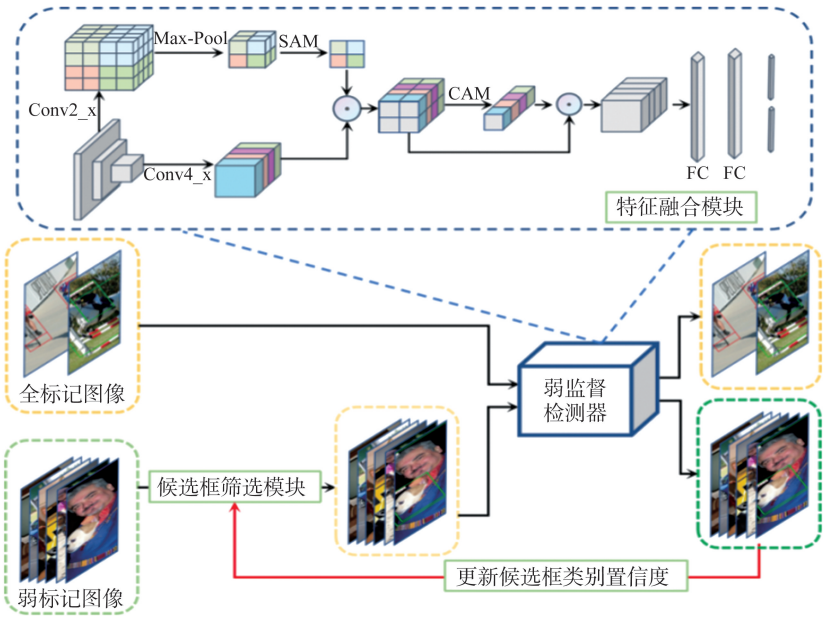


图 2 本文算法框架图

Fig.2 The framework of our method

2.1 候选框筛选模块

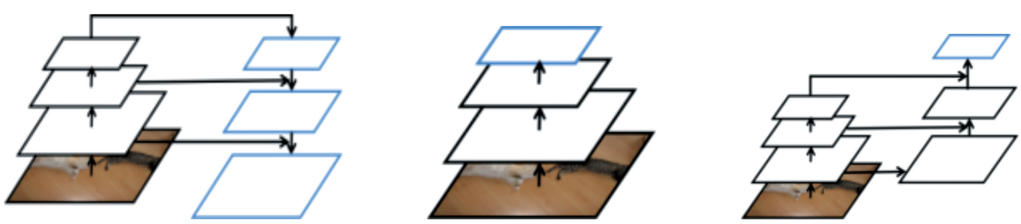
弱监督目标检测算法[21]生成伪标注框时,先过滤置信度较低的候选框后随机采样。若生成的伪标注框错误,随着迭代次数的增加,错误会逐渐积累,造成模型崩塌。如果只选择置信度最高的一个候选框作为伪标注框,由于过度自信产生带噪伪标注框,也会造成噪声累积并降低模型性能。为了克服这些缺陷,本文提出一个更具鲁棒性的伪标注框采样算法。

具体来说,使用选择性搜索算法得到的每个候选框都有其相应的类别置信度。筛选候选框时,将类别置信度低(置信度小于 0.3)的候选框过滤掉。训练前期,对一张图像中的每个类别分别选出最高置信度 $k_{\max}$ ($k_{\max}=5$)个候选框参与训练。训练后期,这些候选框的置信度作为权重进行重要性采样,并用此置信度对损失进行重加权,减小错误预测发生的概率,使模型得到较好的初始化,训练后期也可以修正前期过度自信带来的影响,使得伪标注框更具鲁棒性。

2.2 基于双注意力的特征融合模块

2.2.1 自下向上式的特征融合

最经典的特征融合方式是 FPN 算法[24],如图 3(a)。先进行自下而上的特征卷积,然后自上而下融合相邻的特征图,这种方式的高层特征图缺少细节信息。弱半监督检测(weakly-semi-supervised object detection, WSSOD)[20]、半弱监督检测(semi-weakly supervised object detection, Semi-WSOD)[21]等弱监督检测模型在单一尺度的特征图上进行检测,如图 3(b)所示,此特征图可能丢失很多细节信息。本文提出的特征融合算法克服了图 3(a)和图 3(b)存在的缺点,如图 3(c)所示,以自下而上的方式进行特征融合,将低层细节信息直接补充到语义信息较丰富的高层特征图中,在该层特征图上预测,使检测效果更准确。此算法适用于所有的弱监督目标检测模型。在图 3 中,蓝色的轮廓表示将要进行预测的特征图。



(a) 自上而下式融合 (b) 无融合 (c) 自下而上式融合

图 3 用于目标检测的特征图

Fig.3 Feature maps for object detection

2.2.2　双注意力引导的特征跨层融合

卷积操作混合了通道信息和空间信息提取特征,不一定能更理想地提取通道信息和空间信息特征。为此,本文分别使用空间注意力和通道注意力融合跨层特征图。利用空间注意力对低层特征图加权,将低层特征图信息融合到具有更强语义信息、对细节感知能力较差的高层特征图中。使用通道注意力对融合后特征图加权,使网络模型更多地关注包含物体的关键通道,在空间维度和通道维度上学习物体细节信息和整体信息。将 layer 1 的输出作为特征图 $F^{(1)}$,layer 3 的输出作为特征图 $F^{(2)}$。下面将对此过程进行详细的描述。

此模块如图 4 所示。假设特征图 $F^{(1)}$ 的维度为 $F^{(1)} \in \mathbf{R}^{C\times H\times W}$($C$ 表示特征图的通道数,H 表示特征图的高,W 表示特征图的宽),特征图 $F^{(2)}$ 的维度为 $F^{(2)} \in \mathbf{R}^{4C\times \frac{H}{4}\times \frac{W}{4}}$。图中 Conv3×3 表示卷积核大小为 3×3、Conv7×7为卷积核大小为 7×7, Avg-Pool 为平均池化操作,Max-Pool 为最大池化操作,多层感知机(multi-layer perceptron, MLP)表示为 MLP。

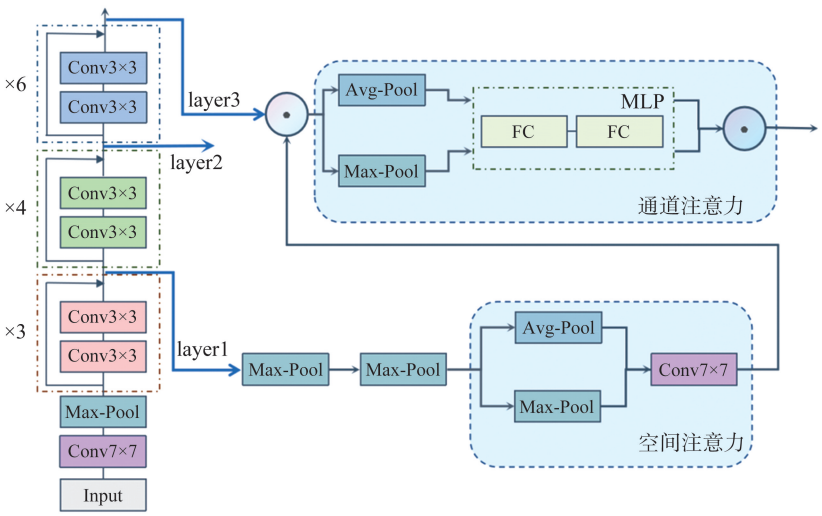


图 4　双注意力引导的特征跨层融合模块

Fig.4　Feature fusion of different layers guided by bi-attention

空间注意力:对特征图 $F^{(1)}$ 进行最大池化操作后得到特征图 $F^{(1')}$,其中 $F^{(1')} \in \mathbf{R}^{C\times \frac{H}{4}\times \frac{W}{4}}$,对于特征图 $F^{(1')}$,先对每个空间位置上的所有通道应用平均池化和最大池化操作,生成 2 个 2D 特征图,$F^{s}_{\text{avg}} \in \mathbf{R}^{1\times \frac{H}{4}\times \frac{W}{4}}$ 和 $F^{s}_{\text{max}} \in \mathbf{R}^{1\times \frac{H}{4}\times \frac{W}{4}}$,$S$ 是空间;按通道连接起来,送入一个卷积核大小为 7×7 的标准卷积层,经过 ReLU 得到空间注意力图 M^s,其中 $M^s \in \mathbf{R}^{1\times \frac{H}{4}\times \frac{W}{4}}$,$M^s$ 与 $F^{(2)}$ 经过哈达玛积运算后得到融合特征图 $F^{(2')}$,其中 $F^{(2')} \in \mathbf{R}^{4C\times \frac{H}{4}\times \frac{W}{4}}$,即

$$F^{(1')} = M_{\text{a}}(F^{(1)}), \tag{1}$$

$$\begin{aligned} M^s(F^{(1')}) &= \sigma(f^{7\times 7}([A(F^{(1')}); M_{\text{a}}(F^{(1')})])) \\ &= \sigma(f^{7\times 7}([F^{s}_{\text{avg}}; F^{s}_{\text{max}}])), \end{aligned} \tag{2}$$

$$F^{(2')} = M^s(F^{(1')}) \otimes F^{(2)}。 \tag{3}$$

式中,M_a 为最大池化操作,A 为平均池化操作,σ 为 sigmoid 函数,$f^{7\times 7}$ 表示卷积核大小为 7×7 的卷积运算。

通道注意力:对于 $F^{(2')}$,首先对每个通道上的所有空间位置应用平均池化和最大池化操作,得到 2 个 1D 特征图 F^{c}_{avg} 和 F^{c}_{max},其中 $F^{c}_{\text{avg}} \in \mathbf{R}^{4C\times 1\times 1}$,$c$ 为类别数, $F^{c}_{\text{max}} \in \mathbf{R}^{4C\times 1\times 1}$。然后,$F^{c}_{\text{avg}}$ 和 F^{c}_{max} 分别经过一个 MLP 操作。F^{c}_{avg} 和 F^{c}_{max} 相加并经过 ReLU 后生成通道注意图 M^c,其中 $M^c \in \mathbf{R}^{4C\times 1\times 1}$,$M^c$ 与 $F^{(2')}$ 经过哈达玛积运算后得到最终特征图 $F^{(2'')}$,其中 $F^{(2'')} \in \mathbf{R}^{4C\times \frac{H}{4}\times \frac{W}{4}}$,即

$$\begin{aligned} M^c(F^{(2')}) &= \sigma(M_1(A(F^{(2')})) + M_1(M_{\text{a}}(F^{(2')}))) \\ &= \sigma(W_1(W_0(F^{c}_{\text{avg}})) + W_1(W_0(F^{c}_{\text{max}}))), \end{aligned} \tag{4}$$

$$F^{(2'')} = M^c(F^{(2')}) \otimes F^{(2')}。 \tag{5}$$

式中,M_l 为一个 MLP 操作,$W_0 \in \mathbf{R}^{C\times 4C}$, $W_1 \in \mathbf{R}^{4C\times C}$。

2.3　损失函数

对于全标记图像 I^{la},损失函数为

$$L^{\mathrm{la}}(x,c,l,g,d)=\frac{1}{M}(L_{\mathrm{cls}}(c,l)+\lambda L_{\mathrm{loc}}^{\mathrm{la}}(x,g,d)), \tag{6}$$

监督信息为真值框 g,$g=(x_0,y_0,x_1,y_1)$是一个边界框的左上角坐标和右下角坐标。图像 I^{la}经过检测器后得到预测框 d 和对应的类别概率 l,M 为预测框个数。损失函数包含分类损失和定位损失,分类损失函数和定位损失函数分别为

$$L_{\mathrm{cls}}(c,l)=-\sum_{i=0}^{M}\log_2^{l_i^c}, \tag{7}$$

$$L_{\mathrm{loc}}^{\mathrm{la}}(x,g,d)=\sum_{i=0}^{M}\sum_{j=0}^{N}\sum_{m\in(x_0,y_0,x_1,y_1)}x_{ij}^{c}s_{L1}(d_i^m-g_j^m), \tag{8}$$

式中,λ 是一个控制分类损失和定位损失相对重要性的超参数,l_i^c 表示第 i 个预测框对类别 c 的概率值,定位损失 $L_{\mathrm{loc}}^{\mathrm{la}}$中,第 i 个预测框的坐标用 d_i^m 表示,第 j 个真值框的坐标用 g_j^m 表示,其中 $m\in(x_0,y_0,x_1,y_1)$。当第 i 个预测框和第 j 个真实框的交并比值超过一定阈值时,$x_{ij}^c=1$,否则为 0。式中 s_{L1}为:

$$s_{L1}(x)=\begin{cases}0.5x^2, & |x|<1,\\ |x|-0.5, & \text{其他。}\end{cases} \tag{9}$$

对于弱标记图像 I^{w},分类损失 L_{cls}与全标记图像相同,在定位损失 $L_{\mathrm{loc}}^{\mathrm{w}}$中将其伪标注框 b 和对应的类别 s 作为监督信息。定位损失函数为

$$L_{\mathrm{loc}}^{\mathrm{w}}(x,s,b,d)=\sum_{i=0}^{M}\sum_{j=0}^{N}\sum_{m\in(x_0,y_0,x_1,y_1)}x_{ij}^{s}s_j^s s_{L1}(d_i^m-b_j^m), \tag{10}$$

$$L^{\mathrm{w}}(x,c,s,l,b,d)=\frac{1}{M}(L_{\mathrm{cls}}(c,l)+\lambda L_{\mathrm{loc}}^{\mathrm{w}}(x,s,b,d)), \tag{11}$$

式中 s_j^c 表示第 j 个伪标注框对类别 c 的概率值。

一个图像总损失为

$$L=L^{\mathrm{la}}+\alpha L^{\mathrm{w}}, \tag{12}$$

式中 α 是一个控制全标记图像损失和弱标记图像损失相对重要性的超参数。一次训练中所有图像的损失求和取平均后,反向传播更新网络参数。

2.4 Soft-NMS

目标检测中常用的后处理算法是 NMS,用于过滤掉预测较差的候选框,只保留响应率较高的候选框。原始的 NMS 将交并比值大的相邻检测框的检测分数直接置为 0,未考虑对象的遮挡问题造成漏检。本文用线性 Soft-NMS[6] 代替 NMS 缓解漏检问题,提升召回率。在不增加任何额外训练的情况下,通过降低检测分数的方式抑制其余检测框。由于与置信度最高的检测框交并比值较大的检测框是假阳性的概率较大,需要更多的抑制,通过线性 Soft-NMS 可实现交并比越大分数越低的结果:

$$l_i=\begin{cases}0, & I(\tilde{d},d_i)<N_2,\\ l_1, & N_1\leqslant I(\tilde{d},d_i)<N_2,\\ l_i(1-I(\tilde{d},d_i)), & I(\tilde{d},d_i)\geqslant N_1,\end{cases} \tag{13}$$

式中,I 为交并比,l_i 为检测框概率值,$\tilde{d}$ 为类别概率值最高的检测框;d_i 为与 $\tilde{d}$ 相邻的其余检测框,N_1、N_2为设定的阈值。

3 实验及结果分析

3.1 实验细节

本文在 the PASCAL visual object classes (VOC)2007 和 VOC 2012 数据集上进行实验,VOC 2007 数据集有 9 963 幅图像,包含 20 个种类,其中训练集和验证集共 5 011 幅图像,测试集包含 4 952 幅图像。VOC 2012 数据集有 22 531 幅图像,其中训练集和验证集包含 11 540 幅图像,测试集包含 10 991 幅图像。训练集和验证集用于检测模型的训练,测试集用于检测模型的性能评估。

ResNet34 作为主干网络提取特征,采用动量为 0.9、权重衰减为 0.000 5 的随机梯度下降算法进行端到

端训练。初始学习率为 0.01,并在训练次数为 5、10 时,批量为 2,训练 20 次。在训练过程中,只使用水平翻转来进行数据增强,并对图像进行均值方差归一化。

实验中模型训练阶段使用的服务器配置为:内存为 12 GB,NVIDIA 3060 GPU,运行环境为 Python3.7,Pytorch1.8.0。

3.2　评估指标

目标检测的主要评价指标是平均检测精度。查准率 P 是衡量模型判断正确的比例,查全率 R 是衡量模型找出所有正样本的能力。计算公式分别为

$$P=\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{P}}},\tag{14}$$

$$R=\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{N}}},\tag{15}$$

式中,T_{P} 为真实标注是正例但被分为正例的样本数,F_{P} 为真实标注是负例但被分为正例的样本数,F_{N} 为真实标注是正例但被分为负例的样本数。平均准确率 A^{P} 是 P-R 曲线下的面积,即

$$A^{\mathrm{P}}=\int_0^1 P(R)\,\mathrm{d}R。\tag{16}$$

模型输出的检测框是 T_{P} 还是 F_{P},根据其与真值框的交并比(intersection over union, IoU)是否大于阈值来判断。模型输出的预测框使用 NMS 算法过滤后,计算其与真值框的交并比并进行匹配,存在 2 种情况:(1) 一个预测框可能与多个真值框的交并比大于阈值,选取交并比最大的一个预测框作为 T_{P};(2) 多个预测框匹配到一个真值框,取置信度最大的一个预测框作为 T_{P},剩下的都是 F_{P}。

3.3　实验结果对比

本次实验的对比算法有 WSSOD、Semi-WSOD、consistency-based semi-supervised learning for object detection (CSD)[34]、self-training and the augmentation driven consistency regularization (STAC)[35]、interpolation-based semi-supervised learning for object detection (ISD)[36]、budget-aware object detection (BAOD)[3]和 multiple instance self-training (MIST)[14]。当交并比阈值为 0.5 时,WSSOD、Semi-WSOD、CSD、STAC、ISD、BAOD、MIST 和本文算法的平均准确率为 $A^{\mathrm{P}}_{\mathrm{WSSOD}}$、$A^{\mathrm{P}}_{\mathrm{Semi\text{-}WSOD}}$、$A^{\mathrm{P}}_{\mathrm{CSD}}$、$A^{\mathrm{P}}_{\mathrm{STAC}}$、$A^{\mathrm{P}}_{\mathrm{ISD}}$、$A^{\mathrm{P}}_{\mathrm{BAOD}}$、$A^{\mathrm{P}}_{\mathrm{MIST}}$、$A^{\mathrm{P}}_{\mathrm{Ours}}$,如表 1 所示。当交并比阈值分别为 0.5、0.55、0.6、0.65、0.7、0.75、0.8、0.85、0.9、0.95 时,WSSOD、Semi-WSOD、CSD、STAC、ISD、BAOD、MIST 和本文算法的平均准确率均值为 $M^{\mathrm{AP}}_{\mathrm{WSSOD}}$、$M^{\mathrm{AP}}_{\mathrm{Semi\text{-}WSOD}}$、$M^{\mathrm{AP}}_{\mathrm{CSD}}$、$M^{\mathrm{AP}}_{\mathrm{STAC}}$、$M^{\mathrm{AP}}_{\mathrm{ISD}}$、$M^{\mathrm{AP}}_{\mathrm{BAOD}}$、$M^{\mathrm{AP}}_{\mathrm{MIST}}$、$M^{\mathrm{AP}}_{\mathrm{Ours}}$,如表 2 所示。

表 1　不同算法在 VOC 2007 测试集上的平均准确率

Table 1　Average precision of different methods on the VOC 2007 test set

算法	训练集	$A^{\mathrm{P}}_{\mathrm{WSSOD}}$	$A^{\mathrm{P}}_{\mathrm{Semi\text{-}WSOD}}$	$A^{\mathrm{P}}_{\mathrm{CSD}}$	$A^{\mathrm{P}}_{\mathrm{STAC}}$	$A^{\mathrm{P}}_{\mathrm{ISD}}$	$A^{\mathrm{P}}_{\mathrm{BAOD}}$	$A^{\mathrm{P}}_{\mathrm{MIST}}$	$A^{\mathrm{P}}_{\mathrm{Ours}}$
弱监督	弱标记数据 VOC07		54.5				52.7	50.9	**59.8**
半监督	全标记数据 VOC07+无标记数据 VOC12	78.0	77.8	74.7	77.4	74.4			**78.4**
半弱监督	全标记数据 VOC07+弱标记数据 VOC12	78.9	79.4						**79.6**

表 2　不同算法在 VOC 2007 测试集上的平均准确率均值

Table 2　Average precision means for Different Methods on the VOC 2007 Test Set

算法	训练集	$M^{\mathrm{AP}}_{\mathrm{WSSOD}}$	$M^{\mathrm{AP}}_{\mathrm{Semi\text{-}WSOD}}$	$M^{\mathrm{AP}}_{\mathrm{CSD}}$	$M^{\mathrm{AP}}_{\mathrm{STAC}}$	$M^{\mathrm{AP}}_{\mathrm{ISD}}$	$M^{\mathrm{AP}}_{\mathrm{BAOD}}$	$M^{\mathrm{AP}}_{\mathrm{MIST}}$	$M^{\mathrm{AP}}_{\mathrm{Ours}}$
弱监督	弱标记数据 VOC07								
半监督	全标记数据 VOC07+无标记数据 VOC12		44.2	42.7	44.6				**44.9**
半弱监督	全标记数据 VOC07+弱标记数据 VOC12		47.3						**47.5**

首先评估本文算法经过半弱监督训练后的性能,再与 WSSOD 和 Semi-WSOD 算法进行比较。在半监督设置下,本文算法在全标记图像上训练一个分类器,训练出的分类器可输出未标记图像的图像级标签,然后进行半监督训练。弱监督检测算法只使用带弱标记的 VOC 2007 数据集训练模型,半弱监督检测算法用带全标记的 VOC 2007 数据集和带弱标记的 VOC 2012 数据集训练模型。表 1、2 的结果表明,无论是在弱监

督、半监督还是半弱监督的情况下,本文算法的检测精度都优于其他算法。

3.4 消融实验结果

在 VOC 2007 数据集上进行消融实验,分别验证新的特征融合方式、伪标注框采样策略以及 Soft-NMS 的作用,结果如表 3 所示。VOC 2007 训练集中,10%是带有细粒度边界框的全标记图像,其余 90%都是弱标记图像。

表 3 本文算法的消融实验结果

Table 3 Ablation studies on each component of our method

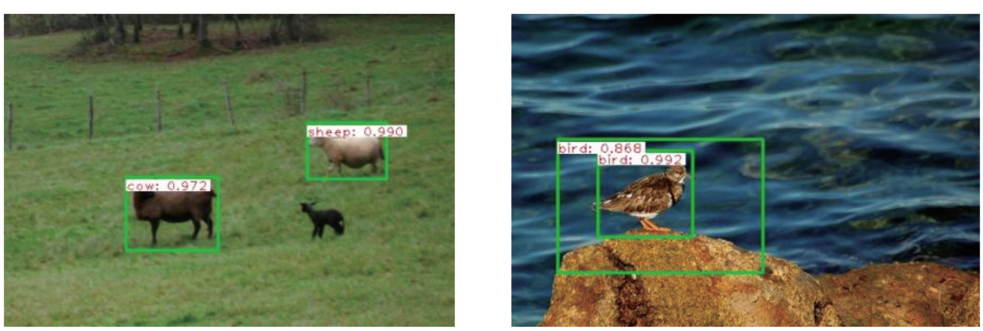
特征融合模块	伪标注框采样策略	Soft-NMS	交并比为 0.5 时的平均准确率/%
			60.57
√			62.74
	√		61.98
√	√		63.55
√		√	62.95
√	√	√	**64.86**

注:“√”表示评估本文算法的该组成部分。

3.4.1 双注意力引导的特征融合模块

为了验证特征融合模块的有效性,本节分别从是否加入特征融合模块、特征融合算法的选择、特征融合模块使用的注意力机制以及参与跨层融合的特征层这 4 个方面进行讨论。

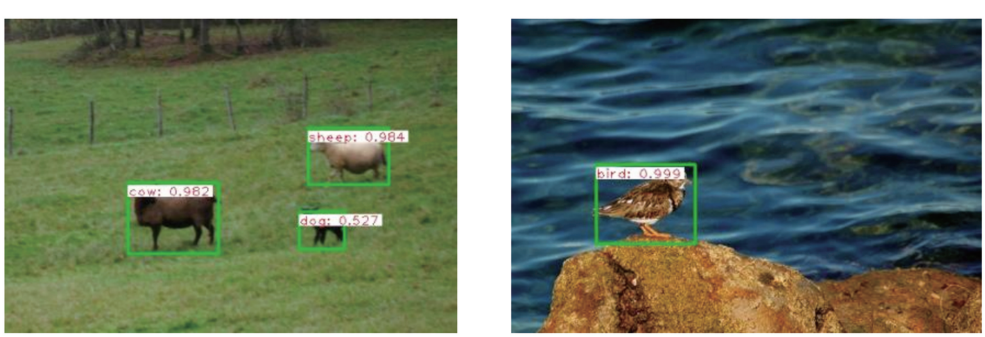
首先仅使用单一层的特征图进行检测,并与本文提出算法在精度(见表 3)和检测结果图上进行比较,结果如图 5、6 所示。由图 5(a)可看出,根据单一层的特征图检测出大部分物体,但存在小物体漏检现象。另外,如图 5(b)所示,在同一个物体上出现了多余的不准确的检测框,且这些检测框只框选出物体的关键部位。在特征融合后的特征图进行检测时,如图 6 所示,检测器对小物体的敏感度增加,且不再关注大物体的局部关键部位,而是将整个物体作为关注对象。



(a) 小物体漏检 (b) 多余的检测框

图 5 使用单层特征图的检测结果

Fig.5 Detection results using a single feature map



(a) 成功检测全部物体 (b) 精准的检测框

图 6 使用融合特征图的检测结果

Fig.6 Detection results on the fused feature map

其次,针对不同的特征融合算法进行消融实验。此次消融实验使用特征层 layer1+layer2+layer3 进行特

征融合,并且只使用融合后的最高层进行目标检测。实验结果如表 4 所示,None 表示在特征融合部分不加任何注意力机制,仅运行 faster-RCNN 的结果。由表 4 可知,使用 PANet 算法和 BiFPN 算法进行特征融合效果并不理想。相较于本文算法,PANet 算法和 BiFPN 算法特征融合方式更复杂,在加强特征提取的同时,有更多的特征信息在最高层汇聚,弱化了背景信息与前景信息间的差异。在检测模型较简单的情况下,较复杂的特征融合方式无法发挥出其最优效果,甚至会影响模型最终性能(采用 PANet 算法后,平均准确率均值由 60.57%下降至 60.26%)。实验证明本文的特征融合模块"即插即用",特别是在较小规模的模型上更能展示出其优势。

表 4　不同特征融合算法实验效果

Table 4　Results of different feature fusion algorithm

算法	交并比为 0.5 时的平均准确率/%
None	60.57
PANet	60.26(−0.31)
BiFPN	61.51(+0.94)
CBAM	62.26(+1.69)
Ours	**62.74**(**+2.17**)

针对在特征融合中使用的注意力机制进行消融实验。分 4 种情况进行对比实验:(1) None 算法不加任何注意力机制;(2) CAM 算法仅使用通道注意力融合特征。layer1 经过 2 次下采样,再用一个卷积核为1×1的卷积层扩展通道数,然后与 layer3 融合;(3) SAM 算法仅使用空间注意力引导 layer1 与 layer3 融合。(4) 本文使用的特征融合算法使用双注意力(SAM、CAM 算法)。

由图 7 可知,使用双注意力引导的特征跨层融合模块较大提升了模型性能,且训练过程较稳定。在特征融合过程中,空间注意力和通道注意力的先后顺序对模型有影响,先使用空间注意力再使用通道注意力效果更好。此后的所有消融实验及本文中的双注意力机制的使用顺序都是先空间注意力后通道注意力。

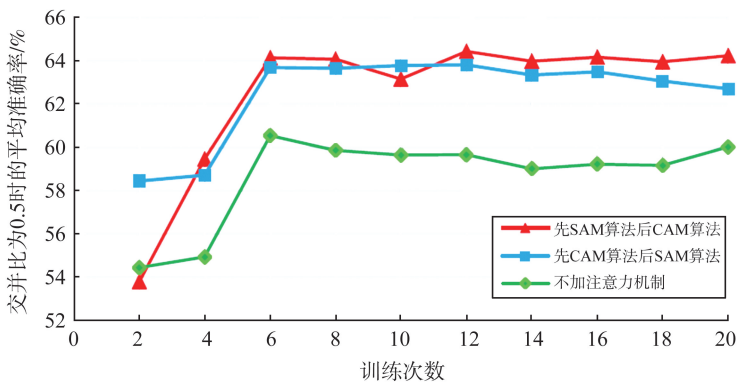


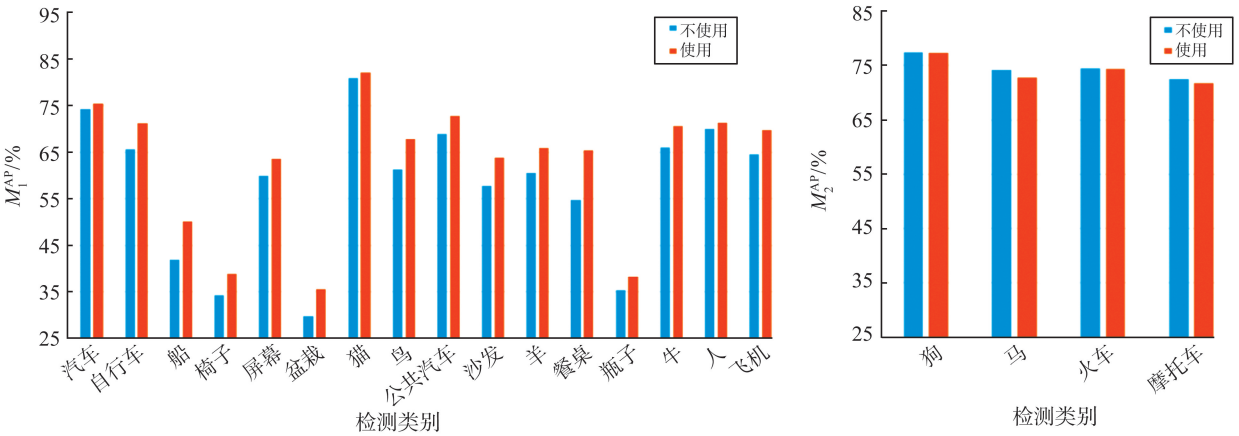
图 7　使用注意力机制的过程对比

Fig.7　The comparison of processes with/without attention mechanism

在 VOC 2007 的 16 个类别上,使用注意力机制的模型检测平均准确率均值都高于未使用注意力机制的模型检测平均准确率均值。相差最大的是餐桌类,平均准确率均值差达到 10.62%;相差最小的汽车类上也有 1.19%。这 16 个类别上平均准确率均值的平均差距约为 4.8%。如图 8(b)所示,在其余 4 个类别上,未使用注意力机制的模型检测精度略高于使用注意力机制的模型的检测平均准确率均值,两者的平均差距仅为 0.6%。这些结果验证了注意力机制在特征融合模块的必要性。交并比阈值分别为 0.5、0.55、0.6、0.65、0.7、0.75、0.8、0.85、0.9、0.95 时,M_1^{AP} 为使用注意力机制的平均准确率均值,M_2^{AP} 为不使用注意力机制 r 平均准确率均值,如图 8 所示。

由图 9(a)可知,双注意力机制在 10 个类别上都取得了最高平均准确率均值,特别是在船、椅子、盆栽等类别上,双注意力机制表现出良好的性能,验证了双注意力机制进行特征融合的有效性。对比图 9(b)、(c)可看出, SAM 在 7 个类别上都得到了最高精度,说明最高特征层更渴望补充细节信息。当交并比为 0.5 时,图 9(a)中 A_1^P 是使用 CAM+SAM 机制的平均准确率,图 9(b)中 A_2^P 是仅使用 SAM 机制的平均准确率,图 9(c)中 A_3^P 是仅使用 CAM 机制的平均准确率。

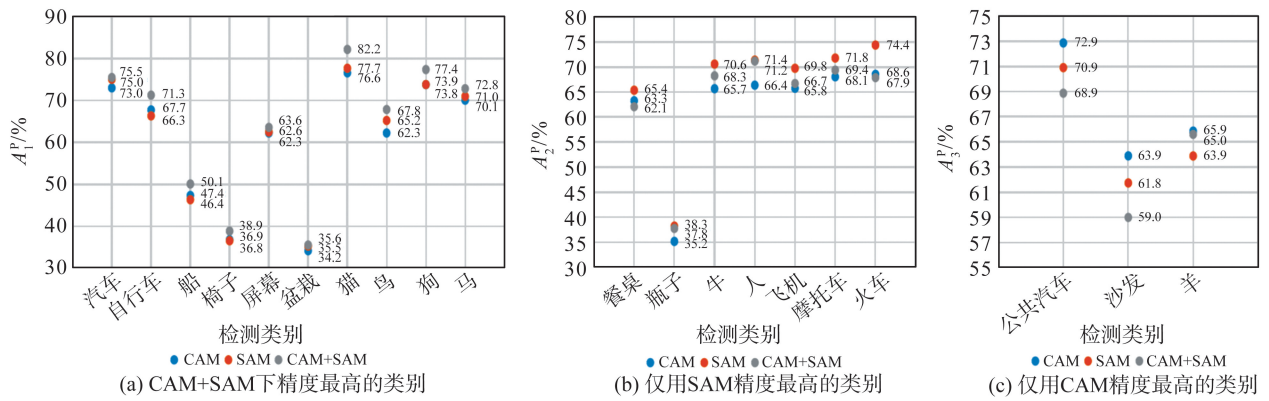
最后,对特征融合选取的特征层进行分析。特征层组合方式有 4 种:(1) layer3 只在最高特征层执行通道注意力与空间注意力;(2) (layer1+layer3)融合最低特征层和最高特征层;(3) (layer2+layer3)融合中间特征层信息和最高特征层;(4) (layer1+layer2+layer3)融合最低特征层、中间特征层和最高特征层。表 5 列出了本实验的结果。



(a) 使用注意力机制精度较高的类别 (b) 不使用注意力机制精度较高的类别

图 8 是否使用注意力机制的结果对比

Fig.8 The comparison with/without attention mechanism



(a) CAM+SAM下精度最高的类别 (b) 仅用SAM精度最高的类别 (c) 仅用CAM精度最高的类别

图 9 不同注意力机制的结果对比

Fig.9 The comparison for the different attention mechanism

由表 5 可知,若直接用最高层进行预测,平均准确率最低,layer1 和 layer3 的组合得到最优结果,剩余 2 种情况的效果相差不多,介于中间。只用 layer1 和 layer3 时,由于最低层和最高层的信息差异最大,因此能显著拉大背景与前景信息的差异。若 layer2 加入特征融合,反而会导致差异减小,不利于物体间区分。

表 5 不同特征层组合方式的实验效果

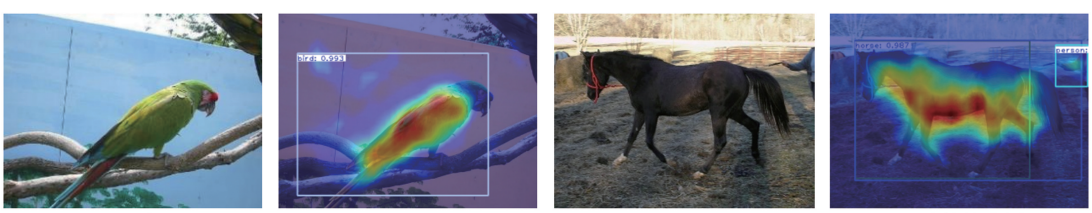
Table 5 Results of different combinations of the feature layers

特征层组合方式	交并比为 0.5 时的平均准确率/%
layer3	61.24
layer1+layer3	**63.55**
layer2+layer3	62.23
layer1+layer2+layer3	62.74

另外,热力图可视化更直观地展示了特征融合模块带来的效果。如图 10 所示,在单个物体上颜色均匀。如图 11 所示,多个物体之间有明显的边界,说明模型关注目标整体而不只是重点部位,不同物体间的区别更明显,更加验证了双注意力引导的特征融合模块的有效性。

3.4.2 伪标注框采样策略

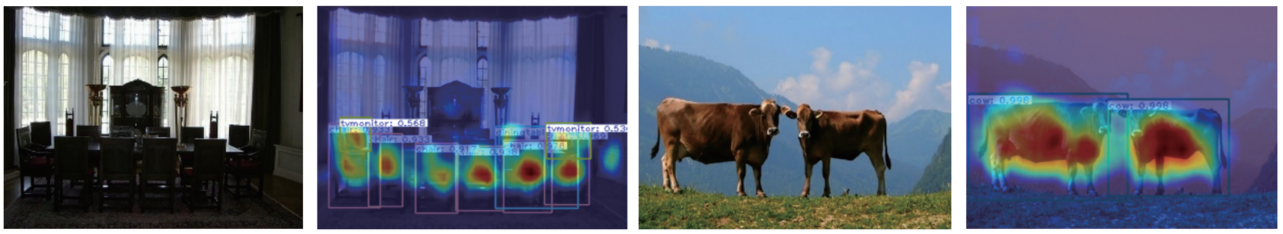
为了验证伪标注框采样策略的有效性,分别可视化不同训练阶段(训练次数为 5、10、15、20)生成的伪标注框,见图 12—14。刚开始存在伪标注框数量冗余(图 12)、分类错误(图 13)和定位不精准(图 14)等问题,但随着迭代次数的增加,算法逐渐筛选出更精确的候选框作为伪标注框,即逐渐收敛到更有意义的位置。



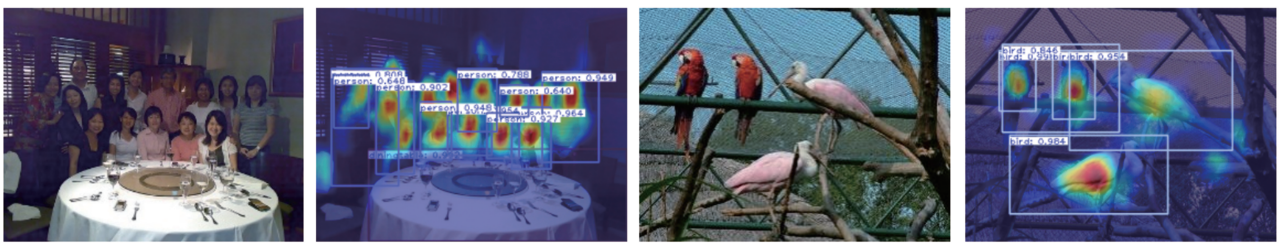
(a) 鸟 (b) 鸟的热力图 (c) 马 (d) 马的热力图

图 10 单个物体及其热力图可视化

Fig.10 Single object and the visualization of heat map



(a) 多个椅子　(b) 多个椅子的热力图　(c) 多头牛　(d) 多头牛的热力图



(e) 众人　(f) 众人的热力图　(g) 多只鸟　(h) 多只鸟的热力图

图 11　多个物体及其热力图可视化

Fig.11　Multiple objects and the visualization of heat map

(a) 训练5次　(b) 训练10次　(c) 训练15次　(d) 训练20次

图 12　存在伪标注框数量冗余时不同训练阶段的可视化

Fig.12　Visualization of different training stages in redundant box

(a) 训练5次　(b) 训练10次　(c) 训练15次　(d) 训练20次

图 13　存在分类错误时不同训练阶段的可视化

Fig.13　Visualization of different training stages in the presence of classification errors

(a) 训练5次　(b) 训练10次　(c) 训练15次　(d) 训练20次

图 14　存在定位不精准时不同训练阶段的可视化

Fig.14　Visualization of different training stages in the presence of imprecise positioning

3.4.3　Soft-NMS

对 Soft-NMS 中的超参数 N_1、N_2的取值进行消融实验。由表 6 可知,线性 Soft-NMS 的 2 个超参数对模型敏感性较小,检测性能基本稳定。

表 6 参数 N_1、N_2的不同取值对性能的影响

Table 6 Performance varying for the different values of N_1 and N_2

N_1	交并比为 0.5 的平均准确率/%	N_2	交并比为 0.5 的平均准确率/%
0.01	**64.86**	0.1	64.83
0.02	64.82	0.2	64.82
0.03	64.83	0.3	**64.86**
0.05	64.16	0.4	64.84

4 结语

本文介绍了一种双注意力引导特征融合的单阶段半弱监督目标检测算法,利用空间注意力和通道注意力加强了高层特征层对多尺度目标的位置、细节信息提取能力。在伪标注框生成过程中,通过候选框筛选和损失重加权方式,有效地利用了弱标记图像。训练是一个单阶段的过程,在 VOC 2007 数据集上的实验表明,本文算法能够提供较优的检测性能,减少了全标记图像的数据量和额外的图像级标注。

由于本文提出的算法仍依赖选择性搜索算法生成候选框,因此比较耗时且无法保证生成的候选框质量,若生成的候选框质量较差,性能就会受到很大影响。接下来的工作将继续探索更优的候选框生成算法以及进一步降低标注成本,并将算法框架应用到目标跟踪和姿态估计等任务中。

参考文献:

[1] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017:7263-7271.

[2] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//Computer Vision-ECCV 2016. Amsterdam: Springer, 2016:21-37.

[3] PARDO A, XU M, THABET A, et al. BAOD: budget-aware object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021:1247-1256.

[4] 任冬伟,王旗龙,魏云超,等. 视觉弱监督学习研究进展[J]. 中国图象图形学报,2022,27(6):1768-1798.
REN Dongwei, WANG Qilong, WEI Yunchao, et al. Progress in weakly supervised learning for visual understanding[J]. Journal of Image and Graphics, 2022, 27(6):1768-1798.

[5] DIETTERICH T G, LATHROP R H, LOZANO-PÉREZ T. Solving the multiple instance problem with axis-parallel rectangles [J]. Artificial Intelligence, 1997, 89(1/2):31-71.

[6] BODLA N, SINGH B, CHELLAPPA R, et al. Soft-NMS-improving object detection with one line of code[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017:5561-5569.

[7] GIRSHICK R. Fast R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015:1440-1448.

[8] BILEN H, VEDALDI A. Weakly supervised deep detection networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016:2846-2854.

[9] TANG Peng, WANG Xinggang, BAI Xiang, et al. Multiple instance detection network with online instance classifier refinement[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 2843-2851.

[10] YANG Ke, LI Dongsheng, DOU Yong. Towards precise end-to-end weakly supervised object detection network[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019:8372-8381.

[11] WAN Fang, WEI Pengxu, JIAO Jianbin, et al. Min-entropy latent model for weakly supervised object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018:1297-1306.

[12] WAN Fang, LIU Chang, KE Wei, et al. C-MIL: continuation multiple instance learning for weakly supervised object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019:2199-2208.

[13] TANG Peng, WANG Xinggang, WANG Angtian, et al. Weakly supervised region proposal network and object detection[C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich: Springer, 2018:352-368.

[14] REN Zhongzheng, YU Zhiding, YANG Xiaodong, et al. Instance-aware, context-focused, and memory-efficient weakly supervised object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020:10598-10607.

[15] LI Dong, HUANG Jianbing, LI Yali, et al. Weakly supervised object localization with progressive domain adaptation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016:3512-3520.
[16] ZHU Yi, ZHOU Yanzhao, YE Qixiang, et al. Soft proposal networks for weakly supervised object localization[C]// Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017:1841-1850.
[17] ARUN A, JAWAHAR C V, KUMAR M P. Dissimilarity coefficient based weakly supervised object detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019:9432-9441.
[18] PAN Tianxiang, WANG Bin, DING Guiguang, et al. Low shot box correction for weakly supervised object detection[C]// Proceedings of the 28th International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann, 2019: 890-896.
[19] BIFFI C, MCDONAGH S, TORR P, et al. Many-shot from low-shot: learning to annotate using mixed supervision for object detection[C]//European Conference on Computer Vision. Glasgow: Springer, 2020:35-50.
[20] CHEN Liangyu, YANG Tong, ZHANG Xiangyu, et al. Points as queries: weakly semi-supervised object detection by points [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021:8823-8832.
[21] MEETHAL A, PEDERSOLI M, ZHU Z, et al. Semi-weakly supervised object detection by sampling pseudo groundtruth boxes[C]//2022 International Joint Conference on Neural Networks (IJCNN). Padua: IEEE, 2022:1-8.
[22] 谢星星,程塄,姚艳清,等. 动态特征融合的遥感图像目标检测[J]. 计算机学报, 2022, 45(4):735-747.
XIE Xingxing, CHEN Gong, YAO Yanqing, et al. Dynamic feature fusion for object detection in remote sensing images[J]. Chinese Journal of Computers, 2022, 45(4):735-747.
[23] 钱泽锋,钱梦莹. 基于改进特征融合的微表情识别方法[J]. 软件工程,2021,24(4):26-29.
QIAN Zefeng, QIAN Mengying. Micro-expression recognition method based on improved feature fusion[J]. Software Engineering, 2021, 24(4):26-29.
[24] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017:2117-2125.
[25] LIU Shu, QI Lu, QIN Haifang, et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018:8759-8768.
[26] TAN Mingxing, PANG Ruoming, LE Quoc V. EfficientDet: scalable and efficient object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020:10781-10790.
[27] SUTSKEVER I, VINYALS O, LE Quoc V. Sequence to sequence learning with neural networks[C]//Advances in Neural Information Processing Systems. Montréal: MIT Press, 2014:3104-3112.
[28] LUONG M T, PHAM H, MANNING C D. Effective approaches to attention-based neural machine translation[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: ACL, 2015:1412-1421.
[29] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich: Springer, 2018:3-19.
[30] 赵珊,郑爱玲. 判别相关分析双注意力机制的目标检测算法[J]. 计算机工程与应用, 2022, 58(17):120-129.
ZHAO Shan, ZHENG Ailing. Object detection based on dual attention mechanism combined with discriminant correlation analysis[J]. Computer Engineering and Applications, 2022, 58(17):120-129.
[31] 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.
[32] UIJLINGS J R, SANDE K E, GEVERS T, et al. Selective search for object recognition[J]. International Journal of Computer Vision, 2013, 104(2):154-171.
[33] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016:770-778.
[34] JEONG J, LEE S, KIM J, et al. Consistency-based semi-supervised learning for object detection[C]//Advances in Neural Information Processing Systems, Vancouver: MIT Press, 2019:10759-10768.
[35] ZHOU Qiang, YU Chaohui, WANG Zhibin, et al. Instant-teaching: an end-to-end semi-supervised object detection framework[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 4081-4090.
[36] JEONG J, VERMA V, HYUN M, et al. Interpolation-based semi-supervised learning for object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021:11602-11611.

(编辑:陈丽萍)