基于浅层定位的动态细化目标检测网络

郑棨元, 陈 莹†

(江南大学 轻工过程先进控制教育部重点实验室，江苏 无锡 214122)

摘 要: 现有的目标检测框架中,浅层弱分类能力是制约着网络精度进一步提高的关键. 对此,提出基于浅层定位信息的动态细化检测网络. 该网络在单阶段算法的基础上,通过增加多连接模块来增强浅层特征,同时去除浅层的分类操作以最大程度地保留浅层的定位结果,并将其作为候选框送入深层网络. 深层网络通过使用引入自适应因子的感受野模块构建特征金字塔,以获得丰富的语义信息用于对浅层的回归结果进行判别和微调. 最后设计基于自注意的可变形卷积头,通过对候选框的偏移来自发进行定位校准,使得网络获得精确的检测结果. 在PASCAL VOC和MS COCO数据集上的实验结果表明,所提出的网络结构可以实现优异的检测精度.

关键词: 目标检测；可变形卷积；感受野；特征金字塔；自适应因子；单阶段算法

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.0700 **开放科学(资源服务)标识码(OSID):**

引用格式: 郑棨元,陈莹. 基于浅层定位的动态细化目标检测网络[J]. 控制与决策, 2023, 38(1): 49-57.

Dynamic refinement networks for object detection based on shallow localization

ZHENG Qi-yuan, CHEN Ying†

(Key Laboratory of Advanced Process Control for Light Industry of Ministry of Education，Jiangnan University，Wuxi 214122，China)

Abstract: In the existing detection framework, the weak classification ability of the shallow layer is the key that restricts the further improvement of network accuracy. In order to solve the problem, a dynamic refinement detection network based on shallow positioning information is proposed. Based on single-stage algorithms, the network enhances the features of the shallow layer by adding multiple connection modules and removes the classification operations of the shallow layer to retain the location results of the shallow layer to the maximum. The location is used as the default boxes of the deep-level network. The deep level network is constructed by using a receptive field module with adaptive factors to obtain rich semantic information for the discrimination and fine-tuning of the regression results from the shallow layer. Finally, the designed deformable convolution head based on self-attention can automatically calibrate the position by shifting the detection box, which helps the network obtain accurate detection results. The experimental results on PASCAL VOC and MS COCO datasets show that the proposed network architecture achieves excellent detection accuracy.

Keywords: object detection；deformable convolution；receptive field；feature pyramid；adaptive factor；one-stage algorithms

0 引 言

在单阶段方法中,SSD(single shot multiBox detector)[1]通过构造多尺度层次结构来减轻单尺度特征的负担,在速度与检测精度之间取得了良好的平衡. SSD针对不同尺度的目标,通过综合多尺度特征的预测结果使得网络能够覆盖各尺度的变化. 然而,由于各层之间缺少交互联系,使得网络精度的进一步提升受制于其小目标的检测能力.

为了改善这种缺陷,现有的单阶段主流网络框架基于SSD的多层级金字塔结构开发了一系列网络变体. 如: FPN(feature pyramid networks)[2]通过自顶而下和横向连接结构将上下层信息相联系; FSSD(feature fusion single shot multiBox detector)[3]通过堆叠上下层特征,使得浅层特征聚合相邻层的通道信息; STDN(scale-transferrable object detection networks)[4]通过缩减通道数实现对尺度的扩增,从而获得不同尺度的语义信息.

上述方法的出发点都是通过向分辨率大但语义

收稿日期: 2021-04-22；**录用日期:** 2021-09-22.
基金项目: 国家自然科学基金项目(62173160).
†通讯作者. E-mail: chenying@jiangnan.edu.cn.

信息弱的浅层特征引入深层的语义信息来获得留有小物体信息的强表征特征. 然而,骨干网络采样因子较大,导致小目标残留在深层的语义信息几乎消失殆尽,致使只通过深浅层特征融合的方式带来的提升有限. 因此,诸如RefineDet[5]和Cascade RCNN[6]引入了级联的思想,通过组合多个线性弱回归器来逐渐逼近真实形状,在特征点定位任务上取得了重大突破. 其通过传递浅层的分类和回归结果至深层,实现了由粗到细的微调,使得定位结果更加精确,但也因此导致了深层的检测结果受制于浅层的筛选能力. 而这根本原因归结于分类和定位任务对同一特征图参数的共享[7]. 对于单阶段网络,浅层特征其内在的弱语义信息与强细节信息的矛盾性,致使其在强化定位能力的同时,又由于弱分类能力的误判而使所产生的回归结果被进一步“过滤”[8]. 作为其结果,深层分类和定位依据的候选框范围将会被进一步缩小,制约着网络精度的提高.

针对上述问题,本文提出一种基于浅层定位信息的动态细化目标检测网络(dynamic refinement networks, DRNet). 该方法改进了单阶段检测器多尺度定位分类结构,通过摒弃浅层的分类分支并融合级联的思想,使得深层特征能充分利用浅层的定位信息,而又不受制于浅层的弱分类能力,从而建立一个高效精准的检测网络结构. 所提出的结构相较于双阶段算法中区域候选网络(RPN),由于不再需要浅层的预先分类作用,而使得网络能像单阶段算法一样进行有效的一次训练,同时又结合单阶段算法中多尺度回归的优点,使得深层不局限于浅层的单一尺度回归结果.

1 基本原理

1.1 问题分析

以SSD[1]为首的一系列单阶段网络框架如图1所示. 这类网络一般以VGG (visual geometry group)[9]模型为基线,通过提取原始分类模型的stage 4和fc 7层,并在此基础上形成特征金字塔结构. 其中预测层由8倍下采样开始依序减小,通过利用大分辨率的浅层特征和小分辨率的深层特征实现对各尺度的覆盖.

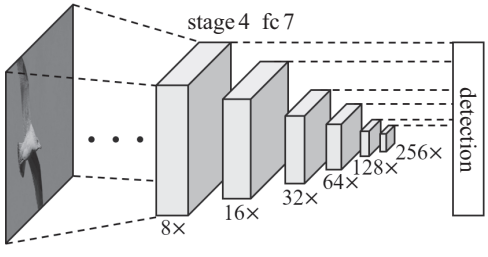
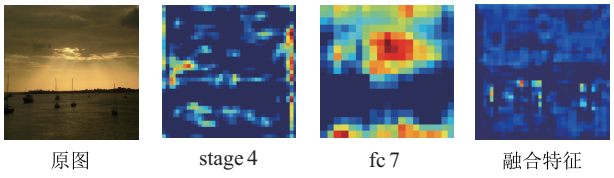
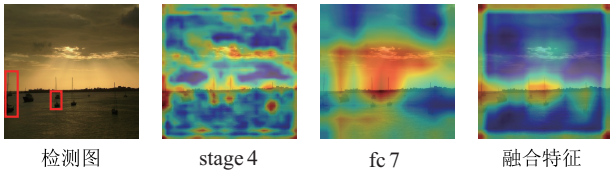


图 1 多层次金字塔结构

然而,实际中SSD本身对小目标的检测效果并不理想,其原因是特征图中语义信息与细节信息的矛盾性. 如图2(a)所示:分辨率较大的stage 4特征包含了更多的边缘轮廓信息,但与此同时也引入环境背景的噪声;而分辨率更小的fc 7特征更为抽象,也代表着语义信息更为丰富,但也因此丢失了对小物体的定位信息. 从图2(b)中可看出:stage 4虽然保留了细节信息,但是缺少足够的语义信息分辨物体使得对噪声过分关注,造成注意力的偏移和负样本的产生;与之相反,fc 7则由于缺少细节信息约束,导致注意力发散. 由于各层之间独立进行分类和回归任务,导致了最终综合的结果对小目标并不友好.



(a) 特征信息可视化



(b) 注意力可视化

图 2 网络可视化分析

特征融合操作通过向浅层特征引入深层的语义信息来提高浅层对大量噪声的判别能力,实现对小目标检测能力的改善[2]. 如图2最后一列所示,融合特征由于深层语义信息的引入,使背景噪声得到了有效的抑制,同时浅层定位信息提供约束作用,使得注意力限制在一定范围. 然而,也因为深层信息对小物体语义信息的缺失,致使小目标的定位信息也被过滤掉一部分,最终使注意力无法准确聚焦,从而制约了该方法的作用效果.

1.2 网络框架

针对上述问题,本文设计一种如图3所示的基于浅层信息的动态目标检测网络. 该网络结合单阶段算法的多尺度思想和双阶段算法的级联思想,在摒弃浅层弱分类作用的基础上,实现高效精准的网络架构. 相对于单阶段原始网络框架,该网络主要包括如下工作:

1) 在轻量级VGG的基础上弃用已经过多损失小目标信息的小分辨率层,通过将8倍下采样的stage 4层和16倍下采样的stage 5层、fc 7层送入所设计的浅层增强模块(shallow reinforcement module, SRM)进行信息的聚合,得到表征能力更强的浅层特征.

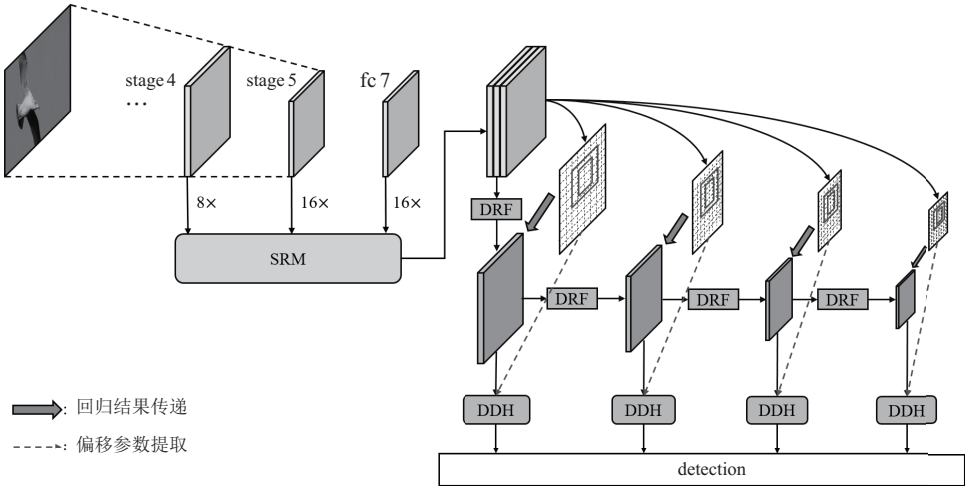


图 3　网络总体框架

2) 将增强的浅层特征通过所设计的动态感受野模块(dynamic receptive field module, DRF)构建出具有8倍、16倍、32倍、64倍下采样率的多尺度特征. 通过自适应地复合不同感受野上的信息,构建语义信息丰富的动态多尺度金字塔.

3) 基于增强的浅层信息进行多尺度回归操作,并将回归的结果传递给相应尺度语义增强特征,作为动态多尺度金字塔的候选框. 通过多尺度金字塔特征对源于浅层回归结果的候选框进行分类和微调,使网络摆脱浅层的弱分类能力.

4) 基于当前特征的空间自注意力并由低层回归结果所提取的信息作为偏移参数. 设计动态的可变形卷积头(dynamically deformable head, DDH)作为检测头,从而实现网络定位信息校准的目的,进而提升网络预测结果的准确性.

各模块的具体实现将会在接下来的章节中介绍.

1.3　浅层增强模块

如图4所示, SRM模块是一个轻量级的多连接模块,用于增强浅层特征表示,包括上采样连接、下采样连接和分辨率恒定连接的复用. 其通过相邻层之间的交互,获得多粒度信息. 为减轻上采样操作所造成的信息稀释,首先采用级联融合的方式,对stage 4层执行1 × 1、步长为2的卷积操作下采样至与stage 5特征同大小,其通道数统一调整为1 024,进行逐元素相加来补足stage 5的信息. 同理,补足后的stage 5特征也与fc 7层进行融合. 融合特征则采用双线性插值上采样与原始stage 4特征进行堆叠,并通过1 × 1卷积将通道数还原为stage 4的初始通道数512. 通过该操作使stage 4层整合了来自相邻层的多粒度信息,生成高质量的浅层特征. 整个过程可表示为

$$\begin{cases} \boldsymbol{T}_5 = f_{1\times1}(\boldsymbol{S}_4) \oplus f_{1\times1}(\boldsymbol{S}_5), \\ \boldsymbol{T}_7 = f_{1\times1}(\boldsymbol{T}_5) \oplus f_{1\times1}(\boldsymbol{S}_7), \\ \boldsymbol{y} = f_{1\times1}(C(\boldsymbol{S}_4, U(\boldsymbol{T}_5), U(\boldsymbol{T}_7))). \end{cases} \tag{1}$$

其中: $\boldsymbol{S}_4 \in \boldsymbol{R}^{H/8\times W/8\times 512}$ 代表stage 4特征; $\boldsymbol{S}_5, \boldsymbol{S}_7 \in \boldsymbol{R}^{H/16\times W/16\times 1\,024}$ 代表stage 5和fc 7层特征; $f_{k\times k}(\cdot)$ 为 $k\times k$ 卷积; $\oplus$ 为逐元素相加; $C(\cdot)$ 为通道堆叠; $U(\cdot)$ 为上采样操作; $\boldsymbol{y} \in \boldsymbol{R}^{H/8\times W/8\times 512}$ 为该模块的输出特征.

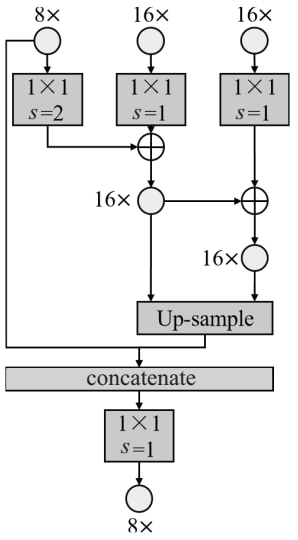


图 4　SRM模块结构

1.4　动态感受野模块

群体感受野大小随着视网膜离心率增加而增加[10]. 其主要实现是通过Inception[11]结构的多分支卷积来模拟离心率,而空洞卷积用于模拟感知尺度及离心率的关系. 这样使得距离卷积中心较近的权重使用小卷积核赋予更大的权重,从而获得更大的感受野,捕获更多的上下文信息. 然而,实际中视觉皮层的神经元感受野受激励调制,能够据输入的内容实现对

不同尺寸的调整[12]. 因此,为提高模型对不同尺度的泛化能力,本文提出了DRF模块,其结构如图5所示.

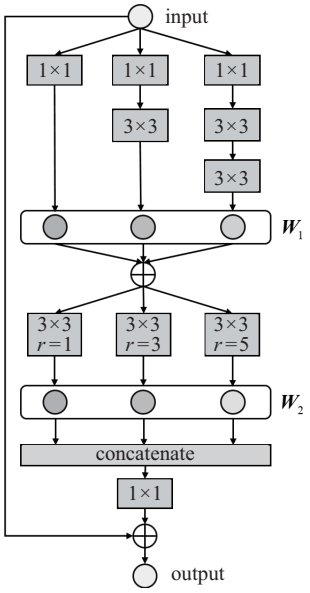


图 5 DRF模块结构

对于输入$\boldsymbol{x}$,首先通过1×1、3×3、5×5的多分支卷积来捕获不同尺度的信息,其中大卷积核分支使用1×1卷积以减轻参数量,而5×5卷积则用两个3×3卷积替代. 其次,引入长度为分支数的向量$\boldsymbol{W}_1$,通过参与网络训练自学习得到向量值. 各分支输出特征与向量所对应元素做乘积,以实现各尺度信息的权重分配,模拟出针对不同尺度的局部刺激,起到对冗余信息进行重加权聚合的软注意力作用. 同理,所得加权特征再次经过多分支结构,在维持分辨率、通道数不变的情况下,根据全局刺激对不同膨胀率的空洞卷积进行权重$\boldsymbol{W}_2$的学习以调整感受野,加深语义信息. 最后,融合特征经由1×1卷积调整通道数,对输入特征$\boldsymbol{x}$进行加和,用于增强原始特征. 整个过程可表示为

$$\begin{cases} e=\displaystyle\sum_{i=0}^{2}\boldsymbol{W}_1[i]*f_{(2i+1)\times(2i+1)}(\boldsymbol{x}),\\ \boldsymbol{T}[i]=\boldsymbol{W}_2[i]*f_{3\times3}^{2i+1}(e),\\ \boldsymbol{U}=f_{1\times1}([C(\boldsymbol{T}[i])]_{i=0}^{2}\oplus\boldsymbol{x}). \end{cases} \tag{2}$$

其中:$f_{k\times k}^{r}(\cdot)$为膨胀率r的$k\times k$卷积,i为第i条分支,$[C(\boldsymbol{T}[i])]_{i=0}^{N}$为$N+1$个特征图的堆叠,$\boldsymbol{U}$为该模块的输出特征.

1.5 浅层多尺度回归

为了防止浅层不充分的语义信息过滤掉大量的正样本,造成结果的损害,在级联回归的阶段,针对浅层特征只进行回归操作,考虑到回归框的个数应与深层尺度相对应,即低层每个像素点上都应匹配有深层的回归框. 这里在浅层特征上进行了多尺度的回归检测. 通过简单的最大池化缩放到相应的多个尺度,再接回归头输出对候选框中心及宽高位置的4个偏移参数. 多尺度定位过程如下所示:

$$\boldsymbol{D}_k=f_{3\times3}(M^k(\boldsymbol{y})),\ k=0,1,2,3. \tag{3}$$

其中:输入$\boldsymbol{y}$为浅层特征,即SRM模块的输出特征;M^k代表进行了k次最大池化操作,其下采样率为2^{3+k};$\boldsymbol{D}_k$为输出特征,通道数为$N_{\text{box}}\times4$,代表相对于每个点所配置的N_{box}个候选框中心和宽高的4个偏置量. 候选框的设置与基线SSD一致,其回归损失为

$$L_{\text{loc}}(\boldsymbol{l},\boldsymbol{g})\sum_{i}^{N}\sum_{m\in(cx,cy,w,h)}\text{smooth}_{L1}(\boldsymbol{l}_i^m-\hat{\boldsymbol{g}}^m).$$

$$\text{smooth}_{L1}(x)=\begin{cases}0.5x^2,\ |x|<1;\\ |x|-0.5,\ x\geqslant1.\end{cases} \tag{4}$$

其中:cx、cy、w、h为候选框的中心及宽高坐标;N为候选框的总数量;$\boldsymbol{l}$为所有$\boldsymbol{D}_k$预测结果的整合,表示对所有N个候选框的4个预测偏置;$\hat{\boldsymbol{g}}$为所对应的真实框相对于候选框的4个偏置.

1.6 回归结果细化

候选框的高生成质量有助于提升检测网络的总体精度,为此,本文进一步设计DDH检测头以取代原始的常规卷积检测头,从而利用可变形卷积的自偏移操作和空间注意力的自适应能力完成定位回归结果的细化. 该模块通过对引入浅层定位信息的深层输出特征施加自学习的偏移量,以使配置在各特征点的候选框发生不规则的位移,丰富了网络浅层所生成回归框的多样性. 其同样包含了分类分支和回归分支,分别用于深层对偏移后的候选框进一步地类别预测和边界框回归,实现“细化”浅层回归结果的功能. 模块结构如图6所示,其中锁头代表分类和回归分支共享偏移参数,以使得分类和回归是针对同样偏移后的候选框进行操作.

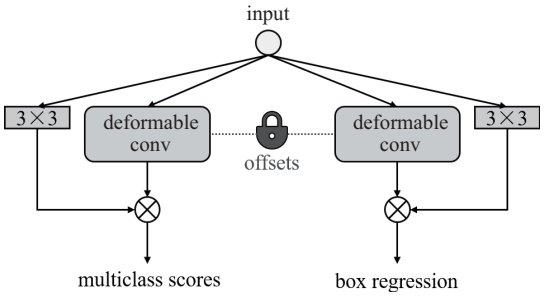


图 6 DDH检测头结构

1.6.1 动态可变形卷积头的偏移操作

可变形卷积源于解决卷积网络固定的几何结构而局限于模型几何变换的问题[13]. 基于对空间采样的位置信息作进一步位移调整的想法,使得采样区域

做自由调整[14]. 常规的卷积操作主要使用规则网格$\mathcal{R}$采样并进行加权运算,而在可变形卷积的操作中,采样位置通过增加一个偏移量$\Delta\boldsymbol{p}_n$,使之成为了不规则位置. 对于在输出的特征图上的每个位置,可变形卷积的计算如下:

$$\boldsymbol{O}(\boldsymbol{p}_0)=\sum_{\boldsymbol{p}_n\in\mathcal{R}} w(\boldsymbol{p}_n)\cdot\boldsymbol{I}(\boldsymbol{p}_0+\boldsymbol{p}_n+\Delta\boldsymbol{p}_n). \tag{5}$$

其中:$\mathcal{R}$定义了感受野的大小和扩张:$\mathcal{R}=\{(-1,-1),(-1,0),\dots,(0,1),(1,1)\}$;$\boldsymbol{p}_n$是对$\mathcal{R}$中所列位置的枚举;$w(\cdot)$为卷积核中所对应位置的权重值,$\boldsymbol{I}(\cdot)$为所对应位置的输入特征值,$\boldsymbol{O}(\cdot)$为所对应位置的输出特征值. 由于偏移量$\Delta\boldsymbol{p}_n$通常是小数,不规则位置的特征值计算采用双线性插值.

本模块进一步利用了可变形卷积对特征值偏移的性质,使得网络能自发学习深浅层特征的对齐. 为此,偏移量$\Delta\boldsymbol{p}_n$的提取,由浅层回归头特征$\boldsymbol{D}_k$进行3×3卷积得到,其输出通道数为$k\times k\times2$,代表对k大小的卷积核中每个位置的偏移参数. 模块结构如图7所示.

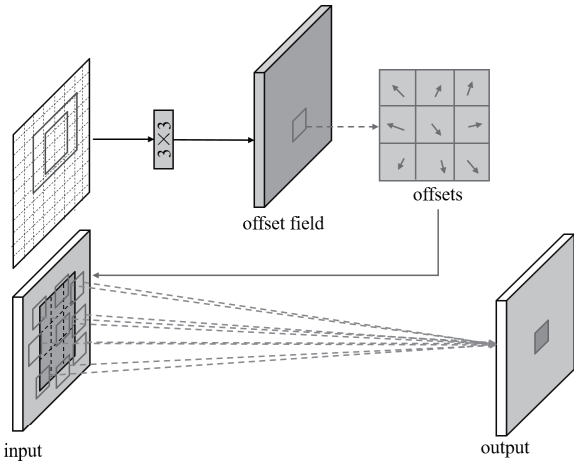


图 7　偏移参数提取

该操作使网络针对不同候选框生成不同的搜索范围,并进行微调和匹配. 此外,3×3的shortcut连接实现了全局空间自注意力,使得网络能基于当前特征的分布对各尺度对象动态地分配权重,让最后的检测结果能更加精准.

1.6.2　定位细化

浅层的偏置结果将会传递至深层特征并结合DDH检测头对候选框进行预调整,以使得由语义信息更强的金字塔特征做进一步的微调和筛选,而不依赖于浅层的弱分类能力,起到对回归结果的“细化”作用. 深层预调整的候选框中心及宽高坐标如下所示:

$$\begin{aligned} cx^*&=cx+\Delta\boldsymbol{p}|_x+l^{cx}\times w,\\ cy^*&=cy+\Delta\boldsymbol{p}|_y+l^{cy}\times h,\\ w^*&=e^{l^w}\times w,\\ h^*&=e^{l^h}\times h. \end{aligned} \tag{6}$$

其中$\Delta\boldsymbol{p}|_x$和$\Delta\boldsymbol{p}|_y$为DDH检测头偏移量$\Delta\boldsymbol{p}$关于x和y方向的分量.

深层金字塔网络的分类损失如下所示:

$$\begin{aligned} &L^*_{\text{conf}}(\boldsymbol{x},\boldsymbol{c})=-\sum_i^{N_{\text{pos}}} x_{ij}^t\log(\hat{c}_i^t)-\sum_i^{N_{\text{neg}}}\log(\hat{c}_i^0),\\ &\hat{c}_i^t=\frac{\exp(c_i^t)}{\sum_t\exp(c_i^t)}. \end{aligned} \tag{7}$$

其中:$x_{ij}^t\in0,1$表示第i个预测框与第j个真实框关于类别t是否匹配;$\hat{c}_i^t$为类别置信度的softmax损失;N_{pos}和N_{neg}分别为正负样本的数量.

回归损失如下所示:

$$\begin{aligned} &L^*_{\text{loc}}(\boldsymbol{x},\boldsymbol{l},\boldsymbol{g})=\\ &\sum_i^{N_{\text{pos}}}\sum_{m\in(cx^*,cy^*,w^*,h^*)} x_{ij}^t\text{smooth}_{L1}(\boldsymbol{l}_i^m-\hat{\boldsymbol{g}}_j^m). \end{aligned} \tag{8}$$

因此,网络的总损失为

$$\text{loss}=L_{\text{loc}}+L^*_{\text{loc}}+L^*_{\text{conf}}. \tag{9}$$

2　实验结果与分析

本文的实验环境基于Pytorch框架,硬件配置为两块NVIDIA 2080Ti. 所选用的数据集为PASCAL VOC[15]和MS COCO[16]. 其中PASCAL VOC有11 K的训练图片,4 952张的测试图片,共计20类,使用均值平均精度(mAP)作为评价指标. MSCOCO有80 K的训练图片,5 K的测试图片,共计80类,使用平均精度(AP)作为评价指标.

2.1　算法评估指标

本文使用的评估指标主要由精度、召回率两部分组成. 精度p也称正确率,定义为分类器认为是正类并且确实是正类的部分占所有分类器认为是正类的比例. 与正确率一同使用的是召回率r,定义为分类器认为是正类并且确实是正类的部分占所有确实是正类的比例. 即

$$p=\frac{\text{TP}}{\text{TP}+\text{FP}},\ r=\frac{\text{TP}}{\text{TP}+\text{FN}}. \tag{10}$$

其中:TP表示被分正确的正样本,FP表示被分错误的正样本,FN表示被分错误的负样本.

定义AP为不同召回率下所有精度的平均值,mAP为所有类AP的平均值. 即

$$\text{AP}=\frac{\sum_{r\in\{r_1,\dots,r_M\}}p(r)}{M},\ \text{mAP}=\frac{\sum_{i=0}^{\text{class}}\text{AP}_i}{\text{Class}}. \tag{11}$$

其中:M为召回率的数量,$p(r)$为该召回率下的精

度, class为类别数量.

2.2 算法实现细节

所有模型均使用ImageNet[17]预训练权重进行初始化. 优化器采用SGD, 动量因子设为0.9, 权重衰减因子设为5×10^{-3}. 训练的批量大小为16, 初始学习率为2×10^{-3}, 在前5个epoch中, 采用warmup预热阶段. 对于PASCAL VOC数据集, 总共训练250个epoch, 其中在150、200的epoch阶段学习率下降为原来的0.1. 对于MS COCO数据集, 总共训练160个epoch, 其中在90、120、140的epoch阶段将学习率下降为原来的0.1.

2.3 消融实验

本部分通过建立不同的模型, 在PASCAL VOC 2007测试数据集验证每个模块对检测性能的影响. 其结果如表1所示. 为了公平比较, 所有实验的参数设置与基线SSD一致, 基线的结果展现在第1组.

表1 不同模块对总体精度的影响

浅层定位	SRM	DRF	DDH	mAP/%
–	–	–	–	77.2
✓	–	–	–	80.4
✓	✓	–	–	81.0
✓	✓	✓	–	81.2
✓	✓	✓	✓	**81.8**

从表1中数据可以看出, 当利用浅层定位信息时, 网络的总体精度由原来的77.2 %提高到80.4 %. 在此基础上引入SRM模块, 进一步生成高质量的浅层特征, 网络的总体精度达到81.0 %. DRF模块对网络感受野的调整, 实现了0.2% 的微小提升. 而DDH模块实现了对级联回归阶段高低层传递结果的对齐, 使网络再次提升了0.6 %, 达到了最高精度81.8 %.

为了进一步验证所提出改进模块的有效性, 将对应模块替换成未改进模块进行对比, 见表2. 通过对浅层特征重新加入分类头进行多分类和二分类, 最终得到的精度为80.2 %和81.2 %, 分别下降了1.6 %和0.6 %, 说明浅层的弱语义信息导致的低分类能力, 将会过滤掉有用的正样本从而生成大量的负样本, 这在级联阶段中致使深层的微调结果受制于该层的错误分类. 而第3组和第4组的实验则将改进的模块分别替换为未改进的感受野模块(RFB)和可变形卷积头(DH), 可以发现精度均有不同程度的下降, 从而表明了改进模块的作用性.

表2 未改进模块对总体精度的影响

算法	mAP/%
DRNet	**81.8**
+浅层分类(多分类)	80.2
+浅层分类(二分类)	81.2
−DRF/+RFB	81.4
−DDH/+DH	81.1

2.4 PASCAL VOC上的整体性能

本节与现有的目标检测算法进行比较, 以展示所提出算法的优异性, 所有模型均采用VOC07和12的联合训练集训练, 结果如表3所示.

表3 基于PASCAL VOC2007测试集的精度比较

方法	输入尺寸	骨干网络	mAP/%
Faster R-CNN[18]	$\sim 600 \times 1\,000$	ResNet-101	73.2
OHEM[19]	$\sim 600 \times 1\,000$	VGG-16	74.6
ION[20]	$\sim 600 \times 1\,000$	VGG-16	76.5
R-FCN[21]	$\sim 600 \times 1\,000$	ResNet-101	79.5
DMFFM[14]	$\sim 600 \times 1\,000$	ResNet-101	82.0
SSD[1]	300×300	VGG-16	77.2
DSSD[22]	321×321	ResNet-101	78.6
DSOD[23]	300×300	DS/64-192-48-1	77.7
STDN[4]	321×321	DenseNet-169	79.3
RefineDet[5]	320×320	VGG-16	80.0
CADNet[24]	320×320	VGG-16	79.4
Shifted SSD[25]	300×300	VGG-16	78.3
MFRDet[26]	300×300	VGG-16	80.7
BPN[27]	320×320	VGG-16	80.3
DRNet	320×320	VGG-16	**81.8**
SSD[1]	512×512	VGG-16	79.5
DSSD[22]	512×512	ResNet-101	81.5
STDN[4]	512×512	DenseNet-169	80.9
RefineDet[5]	512×512	VGG-16	81.8
CADNet[24]	512×512	VGG-16	80.6
MFRDet[26]	512×512	VGG-16	82.0
BPN[27]	512×512	VGG-16	81.9
DRNet	512×512	VGG-16	**82.7**

从表3数据可知, 在小输入分辨率320×320下, 本文所提出的模型达到了81.8 %的精度, 高于基线SSD网络4.6 %, 不仅超过了Faster RCNN系列等双阶段算法, 而且优于一系列SSD变体, 如DSSD[22]、CADNet[24]、MFRDet[26]、BPN[27]等. 对比同样采取级联思想的RefineDet网络, 也提高了1.8 %, 说明了本文所提出模型架构的优越性. 当采取大输入分辨率512×512时, 本文网络的总体精度进一步提高到82.7 %, 高于基线网络3.2 %, 其结果同样也优于相同输入大小的SSD变体.

图8展示了所提出网络在PASCAL VOC上20个类别的检测精度. 从对比中可反映出, 该算法在各类别上相较于基线网络和RefineDet网络几乎都有提

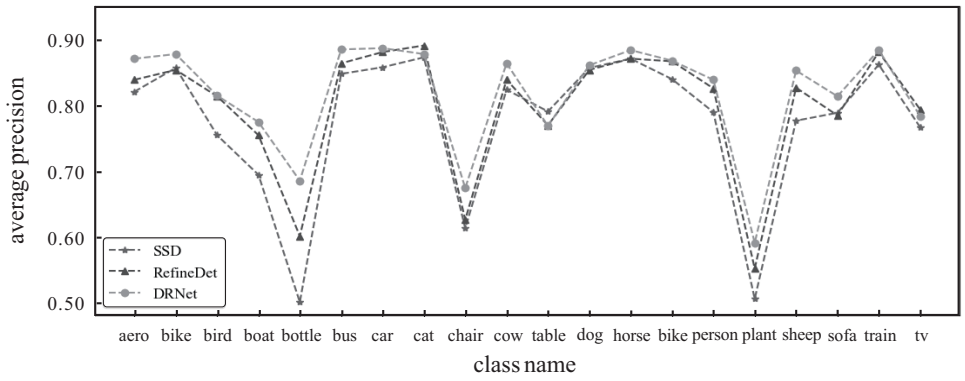


图 8　PASCAL VOC2007测试数据集中20类的AP结果

升. 其中DRNet网络在诸如bottle、chair、plant等典型小目标类别上提升显著,相对于基线分别提升了18.4 %、6.1 %和8.5 %,相对于RefineDet网络也提升了8.4 %、4.8 %和3.9 %. 可见,利用浅层定位信息的网络结构有助于捕捉小对象实例.

2.5　MS COCO上的整体性能

MS COCO也是当前目标检测的权威数据集之一,其包含了许多现实生活的复杂场景,从而使评估结果更具说服力. 依照COCO的评估指标,将目标实例分为3个等级. 其中依照像素将目标分为: 小目标($\text{area} < 32^2$),中目标($32^2 < \text{area} < 96^2$)和大目标($\text{area} > 96^2$),对应的检测指标为$\text{AP}_s$、$\text{AP}_m$、$\text{AP}_l$. 依照交并比(IOU)的不同取值分为3个指标AP、AP_{50}、AP_{75},分别代表在IOU值取0.5～0.95、0.5、0.75下的平均检测精度. 表4展示了本文模型与当前主流算法在COCO测试集上的评估结果.

表 4　基于MS COCO test-dev的精度比较

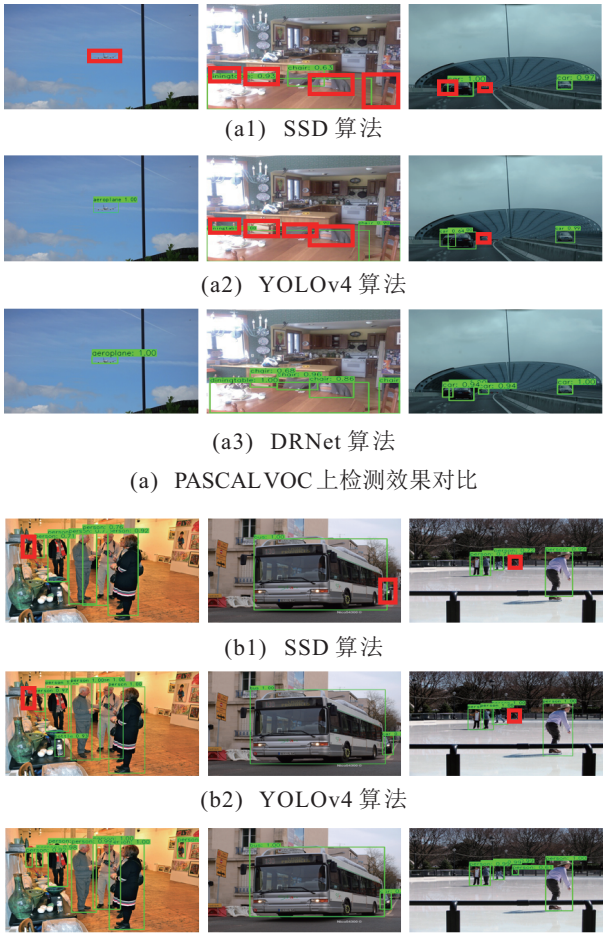
方法	年份	出处	输入尺寸	骨干网络	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Faster R-CNN[18]	2015	NIPS	～600 × 1 000	VGG-16	21.9	42.7	–	–	–	–
OHEM[19]	2016	CVPR	～600 × 1 000	VGG-16	22.6	42.5	22.3	5.0	23.7	37.9
OHEM++[19]	2016	CVPR	～600 × 1 000	VGG-16	25.5	45.9	26.1	7.4	27.7	40.3
ION[20]	2016	Arxiv	～600 × 1 000	VGG-16	23.6	43.2	23.6	6.4	24.1	38.3
R-FCN[21]	2016	NIPS	～600 × 1 000	ResNet-101	29.2	51.5	–	10.3	32.4	43.3
SSD[1]	2016	ECCV	300 × 300	VGG-16	23.2	41.2	23.4	5.3	23.2	39.6
DSSD[22]	2017	Arxiv	321 × 321	ResNet-101	28.0	46.1	29.2	7.4	28.1	47.6
STDN[4]	2018	CVPR	300 × 300	DenseNet-169	28.0	45.6	29.4	7.9	29.7	45.1
FAENet[28]	2019	ICIP	300 × 300	VGG-16	28.3	47.9	29.7	10.5	30.9	41.9
CADNet[24]	2019	TCSVT	320 × 320	VGG-16	27.8	47.1	29.0	8.5	30.1	43.8
DRNet	–	–	320 × 320	VGG-16	**33.2**	**52.9**	**35.6**	**13.3**	**38.7**	**49.2**
SSD[1]	2016	ECCV	512 × 512	VGG-16	26.8	46.5	27.8	9.0	28.9	41.9
DSSD[22]	2017	Arxiv	512 × 512	ResNet-101	33.2	53.3	35.2	13.0	35.4	51.1
STDN[4]	2018	CVPR	512 × 512	DenseNet-169	31.8	51.0	33.6	14.4	36.1	43.4
FAENet[28]	2019	ICIP	512 × 512	VGG-16	31.8	51.2	33.5	16.0	35.8	42.7
CADNet[24]	2019	TCSVT	512 × 512	VGG-16	30.5	50.8	32.1	11.4	35.0	44.8
EfficientDet-D0[29]	2020	CVPR	512 × 512	EfficientNet	34.6	53.0	37.1	12.4	39.0	**52.7**
YOLOv4[30]	2020	Arxiv	512 × 512	CSPDarknet53-PANet-SPP	36.6	55.5	39.6	21.2	41.1	47.0
FCOS[31]	2019	ICCV	～800 × 1 333	ResNet-50-FPN	36.3	54.8	38.7	20.5	39.8	47.8
DRNet	–	–	512 × 512	VGG-16	**37.8**	**58.7**	**40.6**	**22.2**	**42.8**	49.6

从表4中的数据可知,在低分辨率输入下,所提出的模型对比基线SSD, AP从原来的23.2 %提升至33.2 %. 在高分辨率输入下, AP从原来的26.8 %提升至37.8 %. 而在AP_{50}和AP_{75}的指标对比中,低分辨率版本分别提升了11.7 %和12.2 %,高分辨率版本分别提升了12.2 %和12.8 %. 可以看出,所设计网络在高阈值下的精度提升幅度较大,这也说明了引入浅层定位信息更有助于提升网络定位的准确性. 同样,由于该信息的引入,网络在中小目标的精度提升也较为明显,低分辨率版本下分别提升了15.5 %和8.0 %,

高分辨率版本下分别提升了13.9 %和13.2 %,这与本文的动机相符. 所提出的算法不仅在SSD及其变体中取得了最优的精度,也优于最新的SOTA算法,如EfficientDet、YOLOv4、FCOS等. 其中在AP_{50}指标上有着大幅提升,比起三者分别提升了5.7 %、3.2 %和3.9 %,且在中小目标的精度上达到了最优的42.8 %和22.2 %.

2.6 检测结果可视化分析

图9展示了基线网络SSD与本文网络在PASCAL VOC和MS COCO数据集上的检测视觉效果对比. 每组图片第1行为SSD的可视化结果,中间行为最新YOLOv4算法的可视化结果,最后一行为DRNet的可视化结果,红框为漏检目标. 从图9(a)和图9(b)第1列图可以看出,本文算法成功检测出了远处的飞机类和行人类小目标,YOLOv4对于远处的行人产生漏检,而SSD在两类上都产生了漏检. 从图9(a)和图9(b)第2列图可以看出,对于受遮挡的目标,如图中的椅子类和汽车类,SSD算法在两类上都产生了漏检,YOLOv4算法对椅子类产生了漏检,而本文算法则很好地框出了露出部位. 从图9(a)和图9(b)第3列图可以看出,对于暗光环境下的远处目标,本文算法也能准确地捕获.



(a1) SSD算法

(a2) YOLOv4算法

(a3) DRNet算法

(a) PASCAL VOC上检测效果对比

(b1) SSD算法

(b2) YOLOv4算法

(b3) DRNet算法

(b) MS COCO上检测效果对比

图9 SSD算法、YOLOv4算法、DRNet算法检测结果的可视化效果对比

3 结 论

本文针对现有目标检测框架中分类与定位任务的矛盾性,提出了基于浅层定位的动态细化检测网络. 该网络继承级联的思想,通过引入浅层特征的回归结果进行进一步的校准以提升网络对中小目标的捕获能力. 同时设计了多种自适应的动态融合模块以丰富特征的表示,使网络适应于各尺度的多变性. 在当前主流的大型数据集PASCAL VOC和MS COCO上的实验结果表明,该网络相比于其他算法在检测任务上获得了更为优异的检测效果.

参考文献(References)

[1] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector[C]. European Conference on Computer Vision. Amsterdam: Springer, 2016: 21-37.

[2] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]. IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 936-944.

[3] Li Z, Zhou F. FSSD: Feature fusion single shot multibox detector[J/OL]. 2017, arXiv: 1712.00960.

[4] Zhou P, Ni B B, Geng C, et al. Scale-transferrable object detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 528-537.

[5] Zhang S F, Wen L Y, Bian X, et al. Single-shot refinement neural network for object detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 4203-4212.

[6] Cai Z W, Vasconcelos N. Cascade R-CNN: Delving into high quality object detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 6154-6162.

[7] 唐乾坤, 胡瑜. 基于注意力机制的单阶段目标检测锚点框部件感知特征表达[J]. 计算机辅助设计与图形学学报, 2020, 32(8): 1293-1304.
(Tang Q K, Hu Y. Attention based part-aware features of anchor boxes for single-shot object detection[J]. Journal of Computer-Aided Design & Computer Graphics, 2020, 32(8): 1293-1304.)

[8] 黄继鹏, 史颖欢, 高阳. 面向小目标的多尺度Faster-RCNN检测算法[J]. 计算机研究与发展, 2019, 56(2): 319-327.
(Huang J P, Shi Y H, Gao Y. Multi-scale faster-RCNN algorithm for small object detection[J]. Journal of Computer Research and Development, 2019, 56(2): 319-327.)

[9] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J/OL]. 2014,

arXiv: 1409.1556.

[10] 王松, 纪鹏, 张云洲, 等. 自适应感受野网络的行人重识别[J]. 控制与决策, 2022, 37(1): 119-126.
(Wang S, Ji P, Zhang Y Z, et al. Adaptive receptive network for person re-identification[J]. Control and Decision, 2022, 37(1): 119-126.)

[11] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]. IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 2818-2826.

[12] 陈科, 杨棹夕, 廖柏涛, 等. 初级视皮层神经元非经典感受野研究的新进展[J]. 生物化学与生物物理进展, 2021, 48(4): 386-392.
(Chen K, Yang Z X, Liao B T, et al. The research progress of non-classical receptive field in primary visual cortex[J]. Progress in Biochemistry and Biophysics, 2021, 48(4): 386-392.)

[13] Dai J F, Qi H Z, Xiong Y W, et al. Deformable convolutional networks[C]. IEEE International Conference on Computer Vision. Venice, 2017: 764-773.

[14] 李雅倩, 盖成远, 肖存军, 等. 基于细化多尺度深度特征的目标检测网络[J]. 电子学报, 2020, 48(12): 2360-2366.
(Li Y Q, Gai C Y, Xiao C J, et al. Object detection networks based on refined multi-scale depth feature[J]. Acta Electronica Sinica, 2020, 48(12): 2360-2366.)

[15] Everingham M, Gool L, Williams C K I, et al. The pascal visual object classes (VOC) challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303-338.

[16] Chen X L, Fang H, Lin T Y, et al. Microsoft COCO captions: Data collection and evaluation server[J/OL]. 2015, arXiv: 1504.00325.

[17] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database[C]. IEEE Conference on Computer Vision and Pattern Recognition. Miami, 2009: 248-255.

[18] 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.

[19] Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining[C]. IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 761-769.

[20] Shrivastava A, Sukthankar R, Malik J, et al. Beyond skip connections: Top-down modulation for object detection[J/OL]. 2016, arXiv: 1612.06851.

[21] Dai J F, Li Y, He K M, et al. R-FCN: Object detection via region-based fully convolutional networks[J/OL]. 2016, arXiv: 1605.06409.

[22] Fu C Y, Liu W, Ranga A, et al. DSSD: Deconvolutional single shot detector[J/OL]. 2017, arXiv: 1701.06659.

[23] Shen Z Q, Liu Z, Li J G, et al. DSOD: Learning deeply supervised object detectors from scratch[C]. IEEE International Conference on Computer Vision. Venice, 2017: 1937-1945.

[24] Duan K W, Du D W, Qi H G, et al. Detecting small objects using a channel-aware deconvolutional network[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(6): 1639-1652.

[25] Fang L J, Zhao X, Zhang S Q. Small-objectness sensitive detection based on shifted single shot detector[J]. Multimedia Tools and Applications, 2019, 78(10): 13227-13245.

[26] Wei L X, Cui W, Hu Z Y, et al. A single-shot multi-level feature reused neural network for object detection[J]. The Visual Computer, 2021, 37(1): 133-142.

[27] Wu X W, Sahoo D, Zhang D X, et al. Single-shot bidirectional pyramid networks for high-quality object detection[J]. Neurocomputing, 2020, 401: 1-9.

[28] Li W Q, Liu G Z. A single-shot object detector with feature aggregation and enhancement[C]. IEEE International Conference on Image Processing. Taiwan, 2019: 3910-3914.

[29] Tan M X, Pang R M, Le Q V. EfficientDet: Scalable and efficient object detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, 2020: 10778-10787.

[30] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal speed and accuracy of object detection[J/OL]. 2020, arXiv: 2004.10934.

[31] Tian Z, Shen C H, Chen H, et al. FCOS: Fully convolutional one-stage object detection[C]. IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, 2019: 9626-9635.

作者简介

郑棨元 (1996−), 男, 硕士生, 从事目标检测的研究, E-mail: zhengqiyuan@stu.jiangnan.edu.cn;

陈莹 (1976−), 女, 教授, 博士生导师, 从事计算机视觉、模式识别、多媒体信息融合、深度模型压缩等研究, E-mail: chenying@jiangnan.edu.cn.

(责任编辑：李君玲)