基于特征融合和自学习锚框的高分辨率图像小目标检测算法

李 超,黄新宇,王 凯
(湖北工业大学计算机学院,湖北武汉 430068)

摘 要: 为了提高高分辨率图像中小目标的检测精度,解决高分辨率图像在下采样和局部裁切时由于细节和背景信息丢失造成的漏检和误检问题,本文提出了一种基于特征融合和自学习锚框的小目标检测算法. 算法采用多路分支网络对高分辨率图像的全局语义和细节特征平滑后逐层融合,以同时增强特征图上小目标的细节和背景特征. 针对训练样本尺寸差异造成不同分支网络上特征表达不一致的问题,本文引入自学习锚框使融合后的特征图能够适应锚框的位置和形状. 使用本文算法与目前先进的目标检测算法对下采样图像和切块检测,大量实验结果验证了本文算法对高分辨率图像小目标检测的准确性和有效性.

关键词: 小目标检测; 特征融合; 自学习锚框; 高分辨率图像
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2022)07-1684-12
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20200917

Small Object Detection of High-Resolution Images Based on Feature Fusion and Learnable Anchor

LI Chao, HUANG Xin-yu, WANG Kai
(*School of Computer, Hubei University of Technology, Wuhan, Hubei* 430068, *China*)

Abstract: Small object detection of high-resolution images presents significant challenges. To solve the problem that downsampling and cropping of high-resolution images result in missed detections and false detections due to the loss of fine details and contextual information, an algorithm based on feature fusion and learnable anchor is proposed for small object detection of high-resolution images. Contextual and detailed features are extracted from downsampled images and cropped patches respectively, which are then fused layer-wise. The fused features are further combined with smoothed features to strengthen both fine details and contextual information. To mitigate the feature inconsistency, learnable anchor is applied to make the fused features accommodative to the location and shape of anchors. The proposed method is tested from the perspective of global inference and local inference compared to state-of-the-art detectors. The experimental results show the accuracy and effectiveness of the proposed method.

Key words: small object detection; feature fusion; learnable anchor; high resolution images

1 引言

小目标检测在自动驾驶、卫星图像和医学图像分析中发挥着重要的作用. 但是受到环境和测量手段限制,小目标本身携带的信息量较少,这给小目标的特征提取、识别和检测带来了极大挑战[1].

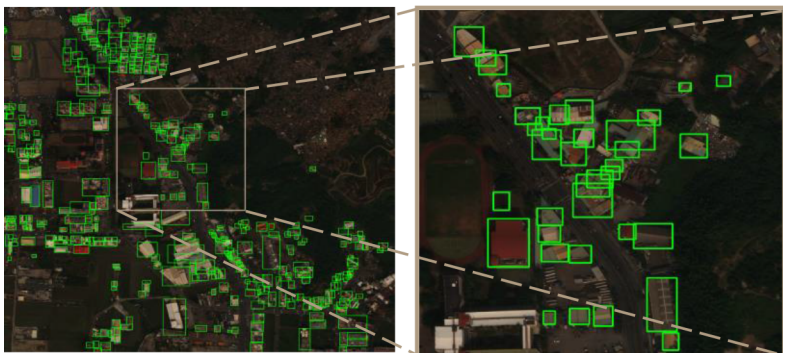
目前,小目标的定义主要有 2 种[2]. 第一种是绝对小目标. COCO 数据集指明当目标的像素尺寸小于 32×32 时,此类目标即可被看作绝对小目标[3]. 第二种是相对小目标. 这种目标的特点是图像尺寸较大但是目标相对原图的尺寸较小,如图 1 所示. 图 1(a)中红色框是普通图像中的绝对小目标,图 1(b)中绿色框是高分辨率图像中的相对小目标. 与普通图像相比,由于训练环境限制,尺寸较大且目标数量较多的高分辨率图像需要下采样或裁切到合适尺寸之后才能训练. 对高分辨率图像多次下采样后,相对小目标的细节信息损失更多,导致更难提取到有效特征,从而造成目标漏检[4,5].

收稿日期:2020-08-21;修回日期:2021-11-05;责任编辑:孙瑶
基金项目:湖北省自然科学基金(No.2017CFB326);国家级大学生创新创业训练计划项目(No.201810500016)

对高分辨率图像裁切后,目标细节信息虽然得以保留,但是目标周围的上下文语义信息缺失导致特征相似的目标被误检. 另外,高分辨率图像中小目标的分布极不均匀,由此造成的正负样本失衡进一步增加了小目标的检测难度.

针对高分辨率图像中小目标检测存在的问题,本文提出一种基于特征融合和自学习锚框的小目标检测算法,通过多路分支网络对高分辨率图像中的小目标提取语义和细节特征,并对提取的特征融合,利用融合后的特征指导锚框预测小目标的位置和形状. 为了评价算法的有效性,将本文算法与目前先进的检测算法进行比较. 实验结果表明,融合后的特征图能够有效弥补由下采样和局部裁切造成的细节特征和语义特征缺失,全局语义和细节特征在融合时出现的特征表达不一致问题得到了较大改善,本文算法可以有效提升高分辨率图像中小目标的检测精度.

(*a*)　普通图像中的绝对小目标



(*b*)　高分辨率图像中的相对小目标

图1　小目标样例

2　相关工作

针对高分辨率图像中小目标检测的难点,本文主要从特征融合和建议框改进等方面开展相关研究工作. 卷积神经网络提取的底层特征分辨率较高,因此细节信息较多;上层特征的分辨率较低但是具有更强的语义信息[6]. 为了让不同层次的特征图都具有较强的语义信息,特征金字塔网络(Feature Pyramid Network,FPN)[7]利用卷积神经网络的多尺度特性,通过自上向下的方式对上层特征图上采样后与其相邻的下层特征图融合.

受到特征金字塔网络的启发,Libra-RCNN检测模型[8]认为传统的特征金字塔仅关注特征图相邻层间的特征融合,非相邻层次中的语义信息在每次融合之后都会被弱化. 通过对不同层次的特征图进行缩放、整合与强化,Libra-RCNN让每个层次的特征图都能从其他层次的特征图上获得同等的信息. 裴伟等人[9]针对无人机拍摄的图像分辨率较低和目标尺度变化较大的问题,提出了一种基于残差网络的目标检测算法,该算法将低层特征和高层语义特征结合在一起,改进后的方法提升了无人机航拍的检测精度. 黄继鹏等人[10]认为专门为小目标设计的检测方法复杂度过高,不具有通用性,由此提出一种面向小目标的多尺度快速区域卷积神经网络,该网络可以同时使用低层和高层特征进行多尺度目标检测. Liu等人[11]认为低层特征包含的较强位置信息对小目标检测具有非常重要作用,因此提出将最低层的信息直接传到最高层特征中. Chen等人[12]通过2条分支网络提取图像的细节信息和全局信息,实现了对高分辨率图像的精准语义分割. Sun等人[13]在并行的多分辨率子网上反复交换信息,通过在不同分支之间进行多尺度的重复融合来减少信息损耗. Chen等人[14]设计了多任务、多阶段的混合级联结构,通过将语义分割分支的空间上下文信息融合到预测框和掩码分支中,使其在目标检测和实例分割任务上都取得了较好的效果.

除了提高特征融合的质量,建议框的优化对目标检测也发挥着重要的作用. Faster-RCNN[15]通过区域生成网络(Region Proposal Network,RPN)在特征图上以滑动窗口方式均匀生成大量密集的锚框,利用特征金字塔和建议框进行多尺度目标检测[16]. Cascade-RCNN[17]不断提高交并比(Intersection-over-Union,IoU)阈值优化建议框的质量,通过级联回归重采样使前一阶段重新采样过的建议框能够适应下一阶段更高阈值的检测器. Dynamic-RCNN[18]提出在训练过程中应随着建议框的分布变化动态调整IoU的阈值. Wang等人[19]关注目标的边缘特征以得到更高质量的边界信息来提升目标检测的精度. Wang等人[20]认为锚框分布与特征图的语义特征有关,并认为根据语义特征生成的锚框可以提升建议框的质量. 基于以上分析可以发现,特征融合质量和建议框的生成方式对目标检测非常重要,但是与

之前的工作相比，卷积神经网络对高分辨率图像提取特征时存在因下采样导致目标细节特征消失过快以及对原图裁切后上下文语义信息缺失的问题．针对高分辨率图像中小目标检测的难点，本文建立表征能力更强的多路分支特征融合网络，在融合的特征图上引入自学习锚框生成建议框，通过提高特征图和建议框的质量，提升高分辨率图像中小目标的检测精度．

3 基于特征融合和自学习锚框的目标检测算法

3.1 整体框架

高分辨率图像像素尺寸较大以及目标尺寸相对较小的特点，导致单一结构的特征金字塔网络难以同时提取到小目标的细节特征和周围的上下文语义特征．本文采用多路分支网络对高分辨率图像中小目标的细节特征和背景语义特征进行融合，整体结构如图2所示．

图2中网络整体由全局分支、局部分支和融合分支构成，特征融合分为从全局到局部的特征融合(Global-to-Local Feature Fusion)和局部到全局的特征融合(Local-to-Global Feature Fusion)．受到训练环境限制，首先在全局分支网络上对下采样图像训练，以提取整幅图像的全局特征(为突出特征融合过程，只描述其中一层)．之后，在局部分支网络上使用骨干网提取切块上目标的局部细节特征，并将切块对应的上下文特征从预训练的全局特征上裁切出来进行Global-to-Local特征融合．由于无法对高分辨率图像直接检测，接着在预训练的局部分支上对原图的每张切块遍历，对融合了上下文信息的局部特征图先下采样再依次拼接，将拼接的特征图与融合分支上的全局特征进行Local-to-Global特征融合．最后在融合了全局背景和局部细节的特征图上对下采样图像实现快速、准确的检测．

由于不同分支网络上的训练样本差异较大，相比单一结构的特征金字塔网络，多路分支网络对原图进行下采样和裁切时，目标的空间位置相对原图已经发生了较大变化．为了使目标的细节特征和上下文特征有效融合，本文通过空间映射对特征图进行裁切和拼接[21]．另外，全局分支和融合分支上的原始图像经过下采样之后目标尺寸变小，而局部分支上的原始图像经过裁切后目标的尺寸并未发生改变．当不同的分支网络采用预设的固定锚框在特征图上均匀滑动时，特征图上每个位置的特征表达是一致的，并且特征图的

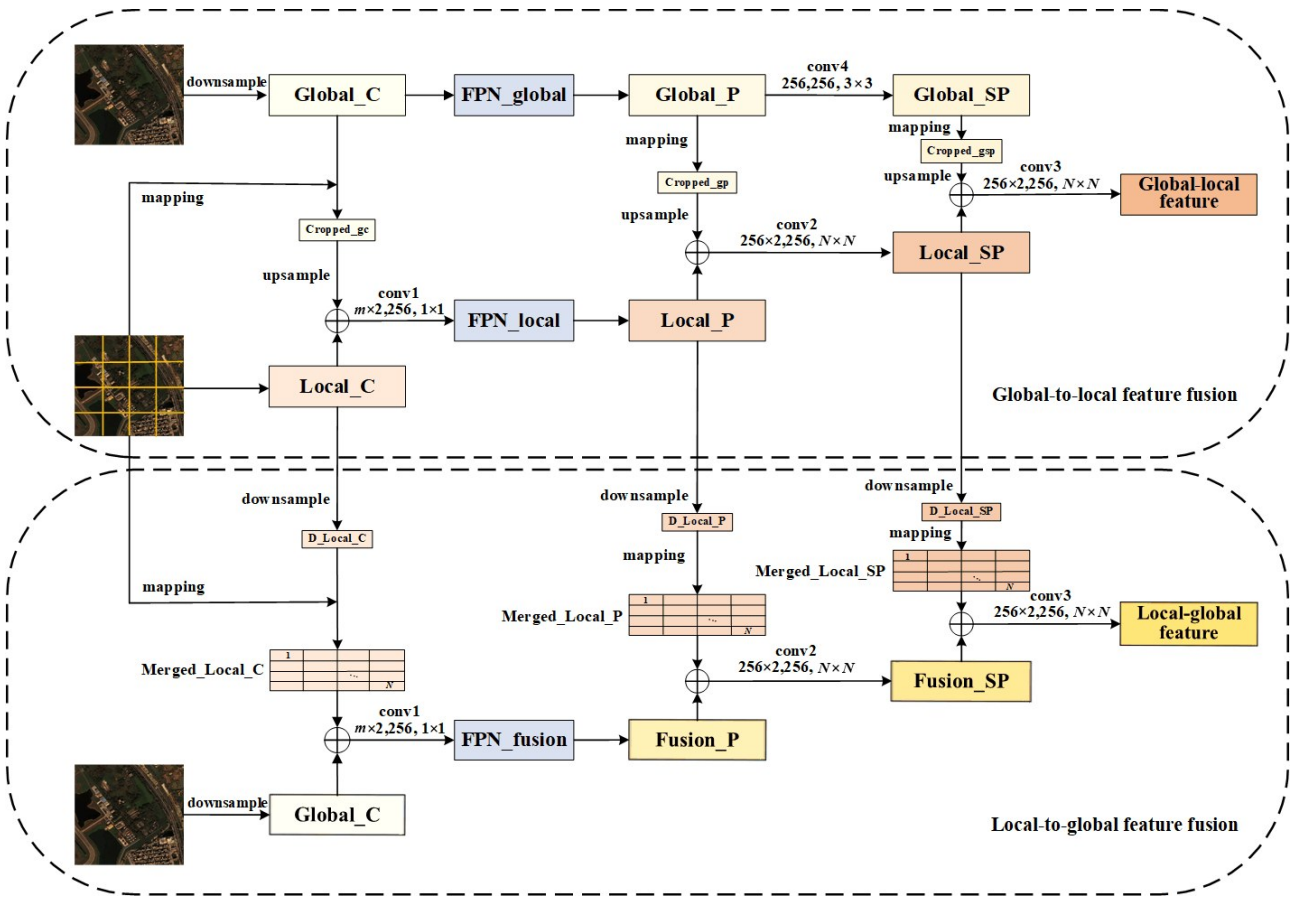


图2 全局和局部特征融合网络

每个像素点与锚框的中心点都能够保持对齐. 但是由于锚框对目标尺寸有不同的偏好,不同的分支网络需要根据目标大小使用不同尺寸的锚框,这将导致各分支网络上特征的感受野和语义范围不一致. 考虑到难以通过先验知识在融合的特征图上选择合适的锚框生成建议框,本文在多路分支网络上引入自学习锚框,使融合的全局局部特征与锚框能够更好地匹配.

3.2　全局和局部特征融合

本文对Global-to-Local特征融合做了详细描述,具体过程如图2所示. 首先,在预训练的全局分支网络上获取下采样图像的全局特征(Global_C),然后在局部分支网络上提取切块上目标的局部细节特征(Local_C). 由于高分辨率图像上的目标分布极不均匀,大量切块上没有目标样本,因此为提高正负训练样本比例,在Global-to-Local特征融合阶段只对含有目标的切块进行特征融合. 之后,利用空间映射将切块对应的预训练全局特征裁切后(Cropped_gc)进行上采样,使其与局部特征图的大小一致. 在此基础上,采用"级联"(concate)方法将裁切后的全局特征和局部特征在通道维度上拼接[22](m为输入通道数,从上至下依次为2 048,1 024,512,256),通过conv1卷积对全局和局部特征按照特征金字塔的方式从上至下逐层融合.

为了使不同尺度的特征图都具有较强的语义信息,特征金字塔FPN_Global对全局特征Global_C从上至下进行融合,输出特征融合结果Global_P. 另外,为了消除FPN_Global在自上向下过程中对上层特征Global_C上采样产生的混叠效应,通常还会采用3×3卷积核对特征融合结果Global_P进行再次卷积,以生成平滑的特征图Global_SP. 在Global-to-Local特征融合过程中,局部分支网络上的局部特征Local_P,Local_SP也要与全局特征Global_P,Global_SP做进一步融合. 在融合之前,仍然利用空间映射将切块对应的预训练全局特征裁切后(Cropped_gp,Cropped_gsp)进行上采样,然后将通道数均为256的全局和局部特征图"级联"为通道数为512的特征图,最后采用conv2,conv3卷积对"级联"的特征图进行卷积,输出通道数为256的特征图(Global-Local Feature),融合过程中只改变特征图的深度,不改变特征图的宽和高. 根据图2,Global-to-Local特征融合算法描述如算法1所示.

Local-to-Global特征融合过程与Global-to-Local特征融合类似,不同之处是Global-to-Local特征融合只对部分切块对应的预训练全局特征上采样,而Local-to-Global特征融合需要对每张切块的局部特征Local_C,Local_P,Local_SP下采样后再依次拼接,保证拼接的特征图Merged_Local_C,Merged_Local_P,Merged_Local_SP与融合分支网络上的特征图Global_C,Fusion_P,Fusion_SP大小相同. 在Local-to-Global特征融合过程中,使用conv1对拼接的局部特征和全局特征从上至下逐层融合,使用conv2和conv3卷积对"级联"的局部特征和融合的全局特征进行卷积,经过平滑后逐层融合,输出通道数为256的特征图(Local-Global Feature). Local-to-Global特征融合算法描述如算法2所示.

算法1　Global-to-Local特征融合算法

输入: image, patch

输出: Global-local feature

```
Global_C=Global_Backbone(downsample(image))
Local_C=Local_Backbone(patch)
Cropped_gc=crop(Global_C, mapping(patch))
Local_P=conv1(Local_C, upsample(Cropped_gc))
Global_P=FPN_global(Global_C)
Cropped_gp=crop(Global_P, mapping(patch))
Local_SP=conv2(Local_P, upsample(Cropped_gp))
Global_SP=conv4(Global_P)
Cropped_gsp=crop(Global_SP, mapping(patch))
Global-local feature=conv3(Local_SP, upsample(Cropped_gsp))
```

算法2　Local-to-Global特征融合算法

输入: image, Local_C,Local_P,Local_SP

输出: Local-global feature

```
Global_C=Fusion_Backbone(downsample(image))
Merged_Local_C=merge(downsample(Local_C), mapping(patch))
Fusion_P=conv1(Global_C, Merged_Local_C)
Merged_Local_P=merge(downsample(Local_P), mapping(patch))
Fusion_SP=conv2(Fusion_P, Merged_Local_P)
Merged_Local_SP=merge(downsample(Local_SP), mapping(patch))
Local-global feature=conv3(Fusion_SP, Merged_Local_SP)
```

3.3　基于特征融合的自学习锚框

由于不同分支间的训练样本尺寸差异较大,采用基准检测模型进行全局和局部特征融合时,如果锚框尺寸选择不恰当,将会影响检测结果. 为了让全局语义和局部细节特征更好地匹配,本文引入自学习锚框使融合的特征图以自适应方式匹配锚框的位置和形状,在特征表达一致的情况下对目标检测. 考虑到图像中目标的位置是非均匀分布的,并且目标形状与图像语义有着密切联系,Wang等人[20]认为锚框的位置和形状服从条件概率分布,通过图3中的位置预测分支使用1×1卷积将输入通道数为256的特征图变为输出通道数为1的特征图,并使用sigmoid函数将特征图上的值转换为概率值,在超过阈值的位置上生成锚框. 形状预测分支使用1×1卷积将特征图变为输出通道数为2的特征图,在特征图上对锚框形状w,h进行预测. 除此以外,自学习锚框模型还对目标进行分类和预测框回归,

因此模型的误差损失由4部分组成，即

$$\text{Loss}=L_{\text{loc}}+L_{\text{shape}}+L_{\text{cls}}+L_{\text{reg}} \tag{1}$$

通过对位置误差损失L_{loc}和形状误差损失L_{shape}最小化，自学习锚框能够尽量覆盖与其相邻目标的定位框，这种生成方式可以提高目标检测的泛化能力．但是在不同位置上生成形状任意的锚框将导致特征图上每个点的特征表达不一致．为解决该问题，自学习锚框模型使用3×3可变形卷积对锚框形状偏移量进行变换，以特征自适应方式使特征图的感受野在不同位置上保持一致，并且像素点和锚框中心点保持对齐．

虽然特征自适应模块解决了单条分支网络上的特征表达问题，但是自学习锚框在训练过程中需要将样本(x_g,y_g,w_g,h_g)映射到特征图上作为标签与预测值(x',y',w',h')一起计算位置误差和形状误差．由于全局分支和局部分支上目标样本的尺寸差异较大，全局和局部特征图上相应位置像素点的感受野和语义特征表达不一致．如果将预训练的特征图与待提取的特征图直接融合，仍然难以通过自学习锚框准确预测锚框的位置和形状．

为了使融合的特征图与自学习锚框自适应匹配，本文利用位置误差损失、形状误差损失和特征自适应模块对融合的特征图做适当修正．由于提取预训练特征的网络模型参数无法再作改变，本文在特征融合时创建新的特征金字塔，将该金字塔输出的特征与预训练特征重新融合，并将样本标签(x_g,y_g,w_g,h_g)重新映射到融合后的特征图上，根据特征图预测的锚框(x',y',w',h')重新计算位置误差和形状误差．对位置误差损失和形状误差损失最小化，使金字塔输出的特征与预训练特征在融合时每个特征点都能适应锚框的位置和形状．在预训练特征基础上，新建的特征金字塔参数只需适当微调，模型就能快速收敛[23]．

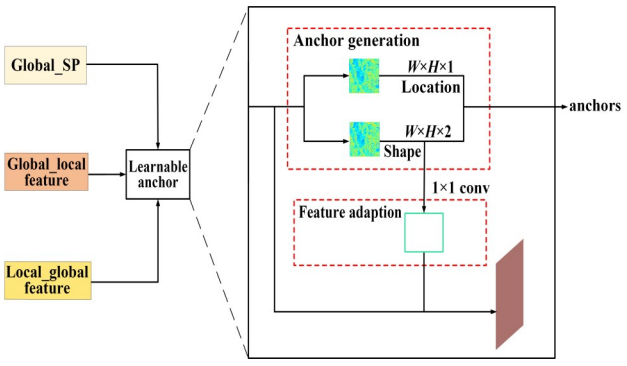


图3 通过自学习锚框对全局和局部特征进行自适应匹配

4 实验方案

4.1 实验数据集

根据高分率图像的特点，本文采用XVIEW数据集[24]作为训练和测试样本，该数据集中共有846张高分辨率图像．随机选取601张图像作为训练集，选取245张图像作为测试集．为了进一步说明高分辨率图像中小目标的特点，图4对XVIEW数据集的图像尺寸、目标尺寸、目标与图像的面积占比、每张图像中的目标数量进行了统计，并与COCO数据集进行了对比．

从图4(a)可以看到，XVIEW数据集的图像尺寸在2 500×2 500到5 000×3 500之间，COCO数据集的图像尺寸在100×100到800×800之间，前者图像的平均像素尺寸是后者的数十倍．另外，依据COCO数据集通过像素大小对目标尺寸的定义，图4(b)中XVIEW数据集的小目标占比为63.3%，而COCO数据集中小目标的比例为41.44%．图4(c)中，COCO数据集的目标与图像的面积归一化占比分布在$[10^{-3},10^{-2}]$区间，而XVIEW数据集中目标与图像的面积归一化占比分布在$[10^{-5},10^{-4}]$区间．由于目标相对原图的归一化比例较小，在对原图下采样后，XVIEW数据集中的目标尺寸将变得更小，导致特征提取网络更难提取到目标的特征．在图4(d)中，COCO数据集中目标数量少于10个的图像占比接近75%，而XVIEW数据集中目标数量少于10个的图像占比不到20%，目标数量超过100个的图像占比接近60%．以上分析表明，XVIEW数据集中的目标比COCO数据集的检测难度更大．

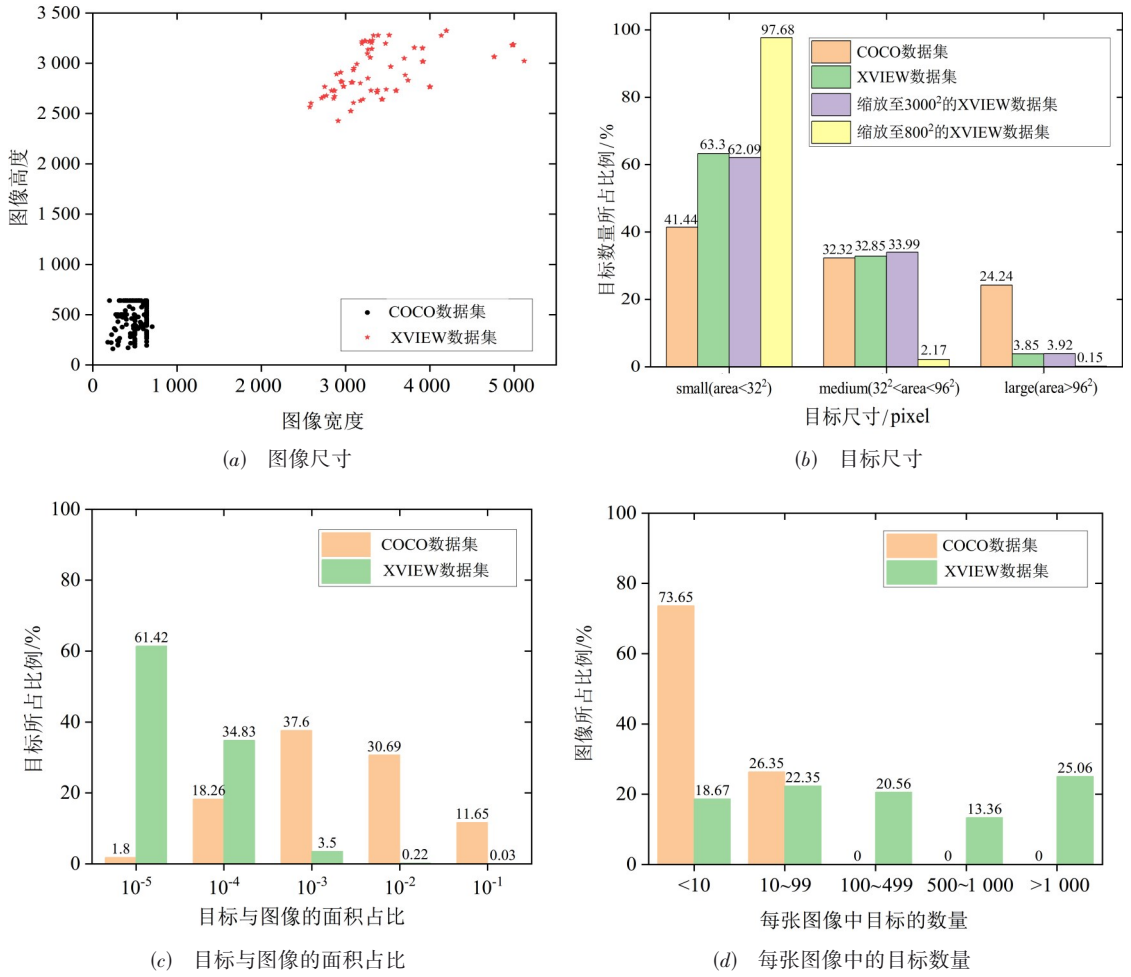
4.2 实验设计

本文算法在Ubuntu 18.04系统下运行，采用MMDetection 2.0目标检测框架[25]对模型进行训练和测试．CPU采用Intel Core i7-9700KF，内存为32 GB，GPU采用NVDIA RTX2080Ti，显存为11 GB，利用CUDA8.0和CUDNN5.0进行加速．

在对模型训练之前，通常需要对图像进行缩放、裁切、翻转、填充等预处理，一方面是为了进行数据增强，以提高网络的泛化性能，另一方面是为了让每个输入网络的batch拥有相同的尺寸，便于GPU做加速计算．由于XVIEW数据集的图像尺寸较大并且目标数量较多，除了对图像采取常用的预处理方法外，还要考虑GPU硬件环境限制和算法效率．根据图4(b)中XVIEW数据集的目标尺寸分布，当图像缩放到3 000×3 000后，目标尺寸的变化与原始数据集相比变化较小，因此将图像调整到3 000×3 000后均匀裁切．为了避免原图裁切后切块之间的边缘信息缺失，让切块之间保持一定程度的重叠．考虑到目标尺寸较小，设置切块之间的重叠区域为50．本文使用ResNet50和特征金字塔提取特征，对原图裁切之前根据式(2)分析特征提取的计算量[26]，即

$$\text{FLOPs}=2*\left(HWC_{\text{in}}K^2C_{\text{out}}+C_{\text{out}}\right) \tag{2}$$

其中,H和W是输出特征图的大小;C_{in}和C_{out}是输入和输出通道数;K是卷积核大小. 在卷积核大小和输入、输出通道数不变的情况下,提取特征的计算量主要由输出特征图的尺寸决定. 一般来说,图像尺寸越大,输出的特征图越大,提取特征的计算量也越大. 另外,对高分辨率图像提取特征时除了考虑切块尺寸的影响,还要考虑切块的个数,表1给出了ResNet50和特征金字塔对不同尺寸的切块和整副图像提取特征的计算量.



(a) 图像尺寸　(b) 目标尺寸

(c) 目标与图像的面积占比　(d) 每张图像中的目标数量

图4　COCO数据集和XVIEW数据集的特征对比

表1　ResNet50和特征金字塔对高分辨率图像提取特征的计算量

切块尺寸	切块个数	最底层特征图的大小	GFLOPs	
			每张切块	整幅图像
500×500	7×7	125×125	35.41	1 735.09
600×600	6×6	150×150	50.43	1 815.48
700×700	5×5	175×175	68.46	1 711.50
800×800	4×4	200×200	88.83	1 421.28
900×900	4×4	225×225	114.14	1 826.24
1000×1000	4×4	250×250	140.05	2 240.80

由表1可知,当切块尺寸为800×800时,对整副图像提取特征的计算量最少. 另外,在对整副图像下采样时,如果缩放的尺寸越小,提取特征的计算量也越少,但是损失的细节信息也将更多. 考虑多路分支网络对每张下采样图像只提取一次特征,并且为了便于切块与其对应的全局特征在空间上相互对应,本文将下采样图像的大小设置为800×800. 对图像下采样和裁切后,图2中Global_C和Local_C的大小自下而上依次为200×200,100×100,50×50,25×25.

4.3　训练参数和评价指标

本文采用随机梯度下降方式对全局分支和局部分支网络分别训练50个epoch,训练初始阶段学习率设为0.001,在前五百轮训练学习率线性增长并稳定在0.002 5,第45个epoch开始,学习率按epoch为单位指数衰减到0.000 3. 对融合分支网络训练时,只需对网络参数作适当微调,训练20个epoch后即能达到比较好的收敛状态.

在性能评价方面,将真实框与预测框的交并比(IoU)大于0.5视为正确的预测结果,在IoU阈值为0.5

下计算模型的检测精度(Average Precision,AP),准确率(Precision)和召回率(Recall). 为了比较的全面性,所有模型均在下采样图像和切块上对目标检测,并对模型的复杂度进行综合评价.

5 实验与结果讨论

本文首先在全局分支网络和局部分支网络上使用大小不同的基础锚框对基准检测模型Faster-RCNN进行训练,并给出基准模型在特征融合下的检测结果. 为了说明特征融合和自学习锚框对检测结果的影响,对特征融合前后基准模型和自学习锚框模型的检测性能进行对比. 之后,通过改变特征融合过程中卷积核的尺寸评价算法效率和检测性能之间的关系. 最后,将本文方法与目前一些先进的检测模型进行综合对比,并给出特征融合前后热度图的直观效果.

5.1 实验1:锚框对检测性能的影响

本实验使用基准模型和两组大小为4和8且长宽比为{1:2,1:1,2:1}的基础锚框(Base anchor)对特征融合前的全局分支网络和局部分支网络训练,表2给出了全局分支网络对下采样图像以及局部分支网络对切块的检测结果.

表2 特征融合前的检测结果

检测方式	Base anchor	AP	Recall	Precision
下采样图像检测	4	29.9	44.6	64.9
	8	14.3	23.4	56.0
切块检测	4	32.4	59.2	49.7
	8	33.3	64.9	45.2

表2表明锚框尺寸对检测结果影响较大,采用基础锚框为4的基准模型对下采样图像检测时,检测精度为29.9,召回率为44.6,采用基础锚框为8时,检测精度为14.3,召回率为23.4. 由此可见,高分辨率图像经过下采样后,由较小锚框生成的建议框可以更准确地命中小目标. 与此相反,局部分支网络上的样本保留了原始尺寸,使用基础锚框为8的检测结果优于基础锚框为4的检测结果.

5.2 实验2:特征融合实验

本实验在基准模型上使用特征融合算法对小目标检测,表3给出了特征融合后的检测结果.

从表3可以发现,在Global-to-Local特征融合下对切块上的小目标检测时,模型准确率相比特征融合前有了显著提升. 在4组全局和局部锚框组合中,在全局分支网络上使用基础锚框为8的基准模型提取的全局特征与局部分支网络上使用基础锚框为8的基准模型提取的细节特征融合后,检测精度最高. 在4→8锚框组合下检测精度有所下降,这说明特征融合前在全局分支网络上使用基础锚框为4的基准模型虽然能够取得较好的检测效果,但是固定锚框对训练样本尺寸较为敏感且具有一定偏好,导致特征图上感受野和特征表达不一致,使得提取的全局特征和局部特征在融合时不匹配.

表3 全局和局部特征融合的检测结果

融合方式	Base anchor	AP	Recall	Precision
Global-to-Local	4→4	37.8	54.4	64.5
	4→8	39.2	55.3	65.8
	8→4	37.7	53.3	65.6
	8→8	40.0	56.2	65.3
Local-to-Global	Base anchor	AP	Recall	Precision
	88→4	33.4	47.0	69.7
	88→8	13.4	21.7	36.0

本文进一步在融合分支上将拼接后的局部细节特征图与全局特征图融合,在Local-to-Global特征融合下对下采样图像检测. 实验结果表明,在Global-to-Local特征融合基础上,在融合分支网络上采用基础锚框4对下采样图像检测时,检测精度从特征融合前的29.9提升到33.4,召回率从44.6提升到47.0,说明经过Local-to-Global特征融合后,特征图上小目标的细节信息得到了显著增强.

实验2表明特征融合在一定程度上可以提高目标检测精度,但是在不同网络上采用固定锚框对预训练特征融合时会出现特征不匹配的问题. 为了避免这种情况,本文使用自学习锚框对小目标作进一步检测.

5.3 实验3:自学习锚框对检测结果的影响

为了体现自学习锚框的作用,本实验首先在特征融合前使用自学习锚框对下采样图像和切块检测,将检测结果与实验1中基准模型检测性能最好的结果进行比较. 为了体现自学习锚框对特征融合的影响,将特征融合算法用于自学习锚框,将检测结果与实验2中基准模型在特征融合下性能最好的结果进行比较,结果如表4所示.

从表4可以发现,在特征融合之前使用自学习锚框

表4 自学习锚框对检测结果的影响

方法		切块检测			下采样图像检测		
		AP	Recall	Precision	AP	Recall	Precision
特征融合前	基准模型	33.3	64.9	45.2	29.9	44.6	64.9
	自学习锚框	34.5	64.7	46.8	26.8	43.3	60.2
方法		切块检测			下采样图像检测		
		AP	Recall	Precision	AP	Recall	Precision
特征融合后	基准模型	40.0	56.2	65.3	33.8	47.0	69.7
	自学习锚框	41.9	57.6	66.4	33.9	52.2	62.7

对切块的检测精度为34.5,比基准模型的检测精度高了1.2个百分点. 相比对切块检测性能的提升,自学习锚框对下采样图像的检测精度、召回率和准确率都低于基准模型,这说明在细节信息损失较多的全局特征图上直接使用自学习锚框的效果并不理想.

经过Global-to-Local特征融合之后,自学习锚框对切块的检测精度为41.9,相比特征融合前检测精度提升了7.4个百分点,而基准模型相比特征融合前检测精度提升了6.7个百分点. 经过Local-to-Global特征融合后,自学习锚框对下采样图像的检测精度为33.9,相比特征融合前检测精度提升了7.1个百分点,而基准模型相比特征融合前检测精度提升了3.9个百分点,这说明在使用自学习锚框后,特征融合对检测性能的提升更为明显. 通过对比自学习锚框在特征融合前后的检测性能,可以发现自学习锚框对特征图的质量有着更高的要求. 在融合了全局语义和局部细节的特征图上,自学习锚框对小目标的检测有着更强的适应性.

5.4　实验4:算法效率和性能之间的关系

本实验通过模型的计算复杂度评价算法效率和检测性能之间的关系. 由于本文采用的骨干网和输出子网与基准模型相同,自学习锚框中的位置预测分支、形状预测分支和特征自适应模块带来的计算量很小,可以忽略不计,因此全局和局部特征融合的额外计算开销主要由图2中的conv1,conv2和conv3产生,其中conv1对每层特征图融合的计算量如表5所示.

表5　conv1对4层特征图进行特征融合的计算量

conv1	2 048×2, 256, 1×1	1 024×2, 256, 1×1	512×2, 256, 1×1	256×2, 256, 1×1
特征图尺寸	25×25	50×50	100×100	200×200
GFLOPs	0.66	1.31	2.62	5.52

从表5可以看到,conv1在全局和局部特征融合过程中产生的计算量为10.11GFLOPs. 根据式(2),在输入、输出通道数和特征图尺寸不变时,卷积操作产生的计算量与卷积核尺寸有关. 在实验2和实验3中,conv2,conv3卷积核尺寸为1×1. 为了体现算法效率和性能之间的关系,本实验使用尺寸为3×3的卷积核,表6给出conv2,conv3在不同尺寸下对4层特征图进行卷积的计算量.

表6　conv2,conv3分别对4层特征图进行特征平滑的计算量

conv2,conv3	256×2, 256, 1×1			
特征图尺寸	200×200	100×100	50×50	25×25
GFLOPs	5.25	1.31	0.32	0.08
conv2,conv3	256×2, 256, 3×3			
特征图尺寸	200×200	100×100	50×50	25×25
GFLOPs	47.17	11.79	2.95	0.74

从表6可以看出,使用1×1卷积核时,特征融合产生的计算量为13.92GFLOPs,使用3×3卷积核时,特征融合产生的计算量为125.3GFLOPs,这说明当卷积核尺寸增大后,模型的计算复杂度将显著增加. 由于增大卷积核尺寸将改变感受野,进而影响特征融合的效果,表7给出在全局和局部特征融合下基准模型和自学习锚框分别使用不同卷积核的检测性能.

表7　全局和局部特征融合卷积核尺寸对检测性能的影响

检测模型	卷积核大小	下采样图像检测			
		AP	Recall	Precision	GFLOPs
基准模型	1×1	33.4	47.0	69.7	158.53
	3×3	37.4	52.4	70.1	269.91
自学习锚框	1×1	33.9	52.2	62.7	158.72
	3×3	37.7	53.2	69.0	270.09
检测模型	卷积核大小	切块检测			
		AP	Recall	Precision	GFLOPs
基准模型	1×1	40.0	56.2	65.1	2536.48
	3×3	43.0	60.8	63.8	4318.56
自学习锚框	1×1	41.9	57.6	66.4	2539.52
	3×3	43.9	59.9	67.5	4321.44

根据图2,基准模型和自学习锚框模型对下采样图像检测时,全局特征图与拼接好的局部细节特征图进行一次特征融合,由于可以将下采样图像的检测结果直接映射回原图,因此可以直接得到对整副图像检测的计算量. 对切块检测时,原图被裁切成16块,每张局部特征图都要与其对应的预训练全局特征融合一次,并且每张切块检测的结果都要映射回原图,因此对整副图像检测的计算量是每张切块的16倍. 由表7可知,增大卷积核尺寸虽然能够提升模型对小目标检测的精度,但是不可避免地会增加计算量. 以自学习锚框为例,当增大卷积核尺寸后,对下采样图像的检测精度可以提升11.2%,但是计算量将增加70.16%,因此需要在算法效率和检测性能之间进行综合考虑.

5.5　实验5:与现有方法对比

为了更客观地评价算法的性能和效率,将本文方法与目前一些先进的目标检测模型进行对比,包括Libra-RCNN[8]、Cascade-RCNN[17]、Dynamic-RCNN[18]、Hybrid task cascade[14](只使用bbox分支)、SABL[19],以及使用HRNet[13]为特征提取网络的基准检测模型,其中本文使用表7中自学习锚框卷积核尺寸为1×1的检测结果,根据实验1对其他模型在下采样图像上使用大小为4的基础锚框,在局部切块上使用大小为8的基础锚框检测,表8统计了各检测模型在测试数据集上的检测精度、召回率、准确率和计算量.

由表8可知,在局部切块上对小目标检测时,本文算法具有更高的准确率和检测精度,其中检测精度比

表8 本文方法与其他模型目标检测性能对比

检测模型	切块检测				下采样图像检测			
	AP	Recall	Pricision	GFLOPs	AP	Recall	Precision	GFLOPs
Cascade-RCNN	35.4	64.1	50.2	2 595.68	23.3	35.9	63.0	162.23
Dynamic-RCNN	35.3	62.8	51.1	2 150.08	26.5	39.8	65.3	134.38
Libra-RCNN	31.8	65.7	41.3	2 160.64	25.3	50.7	47.3	135.04
SABL-RCNN	33.4	66.2	44.2	3 103.84	21.2	32.3	62.7	193.99
Hybrid task cascade	34.4	64.8	47.3	2 594.88	31.2	50.4	58.9	162.18
HRNetV2P_W32	36.3	67.1	47.9	2 933.28	33.7	51.7	63.8	183.33
本文方法	41.9	57.6	66.4	2 539.52	33.9	52.2	62.7	158.72

现有模型中检测精度最高的HRNetV2P_W32基准检测模型高5.6个百分点. 在下采样图像上对小目标检测时,本文算法具有更高的召回率和检测精度,HRNetV2P_W32基准检测模型的检测精度略低于本文算法,计算量比本文算法多24.61GFLOPs.

由表8可知,本文选取的6种具有代表性的目标检测模型分别从特征提取、建议框质量和正负例样本采样等方面对基准模型进行改进和优化. 除了HRNetV2P_W32基准检测模型外,本文选取的其他模型均采用ResNet50作为骨干网提取特征. Cascade-RCNN通过级联方式提升检测器的质量,在不改变特征金字塔的情况下对切块上的目标检测时取得了较好的检测效果. Dynamic-RCNN对切块的检测精度虽然略低于Cascade-RCNN,但是计算量远低于Cascade-RCNN. Libra-RCNN使不同层次的特征图在具体细节和抽象语义之间达到一种较为平衡的关系,但是在对局部切块上的目标检测时,仅在局部范围内对底层细节和高层语义进行了特征平衡,导致模型的准确率偏低. Hybrid task cascade混合任务级联检测模型通过对图像进行精细的像素级分类,将语义分割信息融合到预测框分支中,对下采样图像检测时取得了较高的召回率,而Cascade-RCNN和SBAL-RCNN对下采样图像检测时召回率都较低,说明两者不适用于细节特征消失太多的全局特征图上. 相比以上检测模型,HRNetV2P_W32通过重复跨并行卷积执行多尺度融合,使网络一直都保持高分辨率表征,以此达到同时增强语义信息和精准位置信息的目的,因此对小目标检测取得了较好的效果.

由于深度学习目标检测模型根据特征图生成对目标感兴趣的建议框,并且在建议框上对目标进行分类和位置回归,因此如何在特征图上区分出前景和背景,并且在前景区域上有针对性地生成建议框,对提高目标检测性能有着重要的影响[27,28]. 为了进一步说明特征图对高分辨率图像检测的重要性,本文对特征融合前后的热度图进行直观对比.

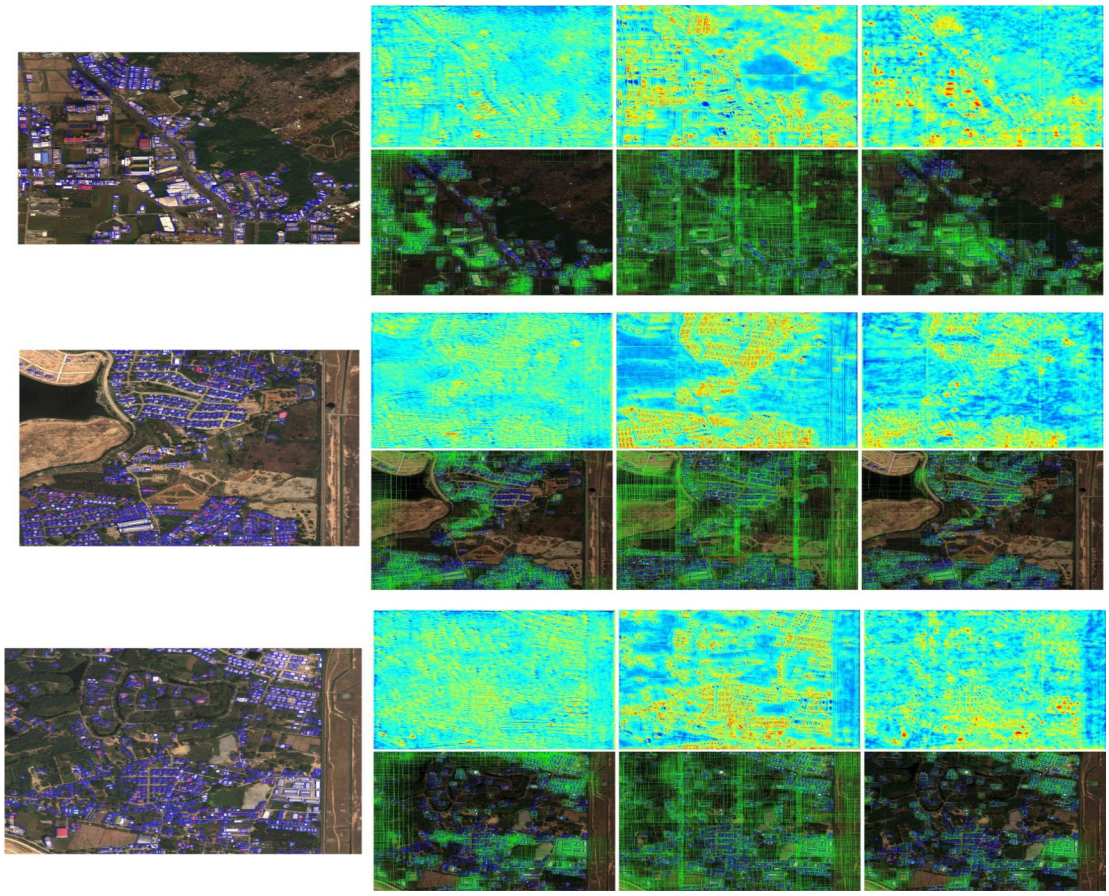
5.6 特征融合前后热度图比较

为了直观反映特征图对高分辨率图像中小目标检测的作用,将全局分支、局部分支和融合分支上的特征图转换成热度图,图5给出了特征融合前后的热度图以及在原图上生成的建议框.

由图5(*b*)可知,全局下采样后的热度图可以大致区分出前景和后景,大量像素较小的目标在热度图上颜色较浅,导致对这些小目标无法生成有效的建议框. 与此相反,图5(*c*)的热度图颜色较深,说明从切块上提取的局部特征图保留了丰富的细节特征,但是每块局部特征图的背景区域都生成了大量建议框,导致很多目标被误验. 经过全局和局部特征融合后,图5(*d*)在保留目标细节特征的同时,前景和背景区分也更为明显,因此能够在特征图上有针对性地生成高质量的建议框.

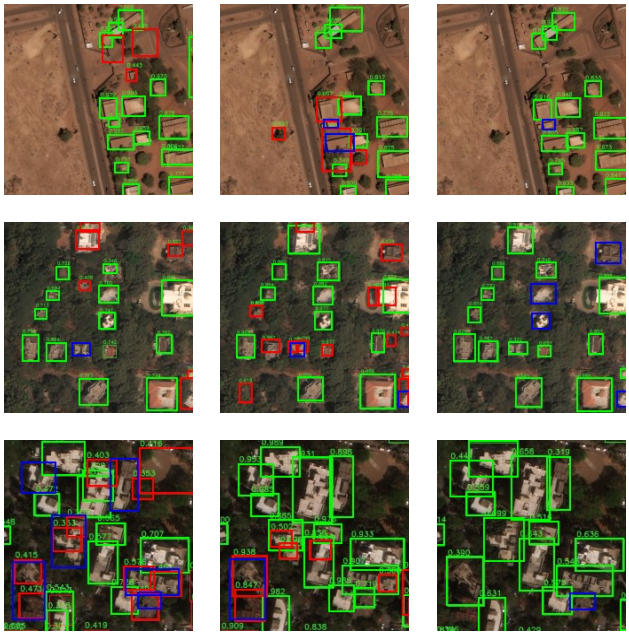
5.7 检测结果对比

基于特征融合和自学习锚框的检测模型使高分辨率图像中小目标检测的效果有所提升,但是由于高分辨率图像中的小目标多为稠密聚集性分布,并且个体之间的边界较为模糊,因此模型的检测效果有待进一步提升[29],这一点从图6中部分具有代表性的目视预测结果对比中可以看出. 图6中从左到右依次为使用全局分支网络、局部分支网络和融合分支网络给出的检测结果,绿色框为正确检测结果,蓝色框为漏检,红色框为误检. 在图6(*a*)背景较为复杂的地面场景中,部分尺寸较小且灰度与地面场景较为相似的小目标在下采样和卷积过程中,其特征很容易被周围背景特征掩盖,因此无法在特征图上提取该类目标的特征. 图6(*b*)中检测目标由于其形状和灰度与周围目标比较相似,因此很容易被误检. 经过特征融合之后,图6(*c*)中的漏检和误检情况相比图6(*a*)和图6(*b*)有了明显减少,但是仍然存在大量与检测目标联系比较紧密的个体对象. 为解决这类问题,一方面需要提高训练数据集中小目标标注的质量,另一方面也要从算法层面对这类稠密聚集性的目标做更精确的区分,这也是本研究团队下一步研究工作的重点[30~32].



(*a*) 原始图像　(*b*) 全局下采样的热度图　(*c*) 局部拼接的热度图　(*d*) 融合的热度图

图5　特征融合前后的热度图对比



(*a*) 全局分支网络　(*b*) 局部分支网络　(*c*) 融合分支网络

图6　测试集上小目标的检测结果

6　结语

高分辨率图像的下采样和裁切处理造成目标细节和上下文语义特征缺失,使得一些先进的目标检测算法对高分率图像中的目标检测时难以取得理想的效果. 本文通过多路分支网络对高分辨率图像提取全局语义和局部细节特征,并将两者有效地融合起来. 实验结果验证了在融合的特征图上采用自学习锚框能够有效检测高分辨率图像中的小目标. 在未来工作中,本文将进一步探索并优化网络结构以提升特征融合能力,从而提高小目标的检测精度.

参考文献

[1] KISANTAL M, WOJNA Z, MURAWSKI J, et al. Augmentation for small object detection[C]//9th International Conference on Advances in Computing and Information Technology(ACITY 2019). Sydney: Aircc Publishing Corporation, 2019: 119-133.

[2] 刘颖, 刘红燕, 范九伦, 等. 基于深度学习的小目标检测

研究与应用综述[J]. 电子学报, 2020, 48(3): 590-601.
LIU Y, LIU H Y, FAN J L, et al. A survey of research and application of small object detection based on deep learning[J]. Acta Electronica Sinica, 2020, 48(3): 590-601. (in Chinese)

[3] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common Objects in Context[C]//European Conference on Computer Vision. Cham: Springer, 2014: 740-755.

[4] BODLA N, SINGH B, CHELLAPPA R, et al. Soft-NMS–Improving object detection with one line of code[C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 5562-5570.

[5] 李宝奇, 贺昱曜, 强伟, 等. 基于并行附加特征提取网络的SSD地面小目标检测模型[J]. 电子学报, 2020, 48(1): 84-91.
LI B Q, HE Y Y, QIANG W, et al. SSD with parallel additional feature extraction network for ground small target detection[J]. Acta Electronica Sinica, 2020, 48(1): 84-91. (in Chinese)

[6] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770-778.

[7] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Hawaii: IEEE, 2017: 936-944.

[8] PANG J M, CHEN K, SHI J P, et al. Libra R-CNN: Towards balanced learning for object detection[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Long beach: IEEE, 2019: 821-830.

[9] 裴伟, 许晏铭, 朱永英, 等. 改进的SSD航拍目标检测方法[J]. 软件学报, 2019, 30(3): 738-758.
PEI W, XU Y M, ZHU Y Y, et al. The target detection method of aerial photography images with improved SSD [J]. Journal of Software, 2019, 30(3): 738-758. (in Chinese)

[10] 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. (in Chinese)

[11] LIU S, QI L, QIN H F, et al. Path aggregation network for instance segmentation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 8759-8768.

[12] CHEN W Y, JIANG Z Y, WANG Z Y, et al. Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long beach: IEEE, 2019: 8916-8925.

[13] SUN K, XIAO B, LIU D, et al. Deep high-resolution representation learning for human pose estimation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Long beach: IEEE, 2019: 5686-5696.

[14] CHEN K, PANG J M, WANG J Q, et al. Hybrid task cascade for instance segmentation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long beach: IEEE, 2019: 4969-4978.

[15] 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.

[16] OKSUZ K, CAM B C, KALKAN S, et al. Imbalance problems in object detection: A review[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(10): 3388-3415.

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

[18] ZHANG H K, CHANG H, MA B P, et al. Dynamic R-CNN: Towards High Quality Object Detection via Dynamic Training[M]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 260-275.

[19] WANG J Q, ZHANG W W, CAO Y H, et al. Side-aware boundary localization for more precise object detection [C]//European Conference on Computer Vision. Cham: Springer, 2020: 403-419.

[20] WANG J Q, CHEN K, YANG S, et al. Region proposal by guided anchoring[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Long beach: IEEE, 2019: 2960-2969.

[21] YANG F, FAN H, CHU P, et al. Clustered object detection in aerial images[C]//2019 IEEE/CVF International Conference on Computer Vision(ICCV). Seoul: IEEE, 2019: 8310-8319.

[22] LIN M, CHEN Q, YAN S. Network In Network[EB/OL].

(2014)[2020]. https://arxiv.org/abs/1312.4400.

[23] OUYANG W L, WANG X G, ZHANG C, et al. Factors in finetuning deep model for object detection with long-tail distribution[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 864-873.

[24] DARIUS L, KUZMA R, MCGEE K, et al. xView: Objects in Context in Overhead Imagery[EB/OL]. (2018)[2020]. https://arxiv.org/abs/1802.07856.

[25] CHEN K, WANG J, PANG J, et al. MMDetection: Open MMLab Detection Toolbox and Benchmark[EB/OL]. (2019)[2020]. https://arxiv.org/abs/1906.07155v1.

[26] MOLCHANOV P, TYREE S, KARRAS T, et al. Pruning Convolutional Neural Networks for Resource Efficient Inference[C]//Proceedings of the 5th International Conference on Learning Representations(ICLR2017). Toulon: ICLR, 2017: 1-17.

[27] OKSUZ K, CAM B C, AKBAS E, et al. Generating positive bounding boxes for balanced training of object detectors[C]//2020 IEEE Winter Conference on Applications of Computer Vision. Snowmass Village: IEEE, 2020: 883-892.

[28] FU Z H, CHEN Y W, YONG H W, et al. Foreground gating and background refining network for surveillance object detection[J]. IEEE Transactions on Image Processing, 2019, 28(12): 6077-6090.

[29] HE Y H, ZHU C C, WANG J R, et al. Bounding box regression with uncertainty for accurate object detection [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Long beach: IEEE, 2019: 2883-2892.

[30] CHEN K A, LI J G, LIN W Y, et al. Towards accurate one-stage object detection with AP-loss[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Long beach: IEEE, 2019: 5114-5122.

[31] HUANG X, GE Z, JIE Z Q, et al. NMS by representative region: Towards crowded pedestrian detection by proposal pairing[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Seattle: IEEE, 2020: 10747-10756.

[32] REZATOFIGHI H, TSOI N, GWAK J, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Long beach: IEEE, 2019: 658-666.

作者简介

李　超　男,1982年出生于湖北省洪湖市. 湖北工业大学计算机学院硕士生导师. 研究方向为计算机视觉、机器学习.
E-mail: lich.mail@163.com

黄新宇　男,1998年出生于湖北省黄石市. 湖北工业大学计算机学院本科生. 研究方向为目标检测与语义分割.
E-mail: 864546664@qq.com