多分辨率特征融合的光学遥感图像目标检测

姚艳清[1,2]，程塨[1,2]，谢星星[1,2]，韩军伟[2]

1. 西北工业大学 深圳研究院, 深圳 518057;
2. 西北工业大学 自动化学院, 西安 710129

摘 要：高分辨率遥感图像目标检测是计算机视觉的一个重要研究领域，在民用与军事领域具有重要的应用价值。目前，基于深度学习的自然图像目标检测有了突破性进展。但是，由于遥感图像具有目标尺度差异大且类间相似度高的特点，使得处理自然图像的目标检测算法直接应用于遥感图像时仍面临着一些挑战。针对上述挑战，本文提出一种多分辨率特征融合的遥感图像目标检测方法。首先，通过特征金字塔提取多尺度特征图并在其后嵌入多分辨率特征提取网络，促使网络学习目标在不同分辨率下的特征，缩小不同特征层之间的语义差距。其次，为实现多分辨特征的有效融合，本文采用自适应特征融合模块挖掘更具判别性的多分辨特征表达。最后，将自适应特征融合模块的输出特征的相邻层进行深度融合。在公开的遥感图像目标检测数据集DIOR和DOTA上评估了本文方法的有效性，相比采用特征金字塔结构的Faster R-CNN，本文方法的准确率（mAP）分别提高2.5%和2.2%。

关键词：卷积神经网络，多分辨率特征融合，遥感图像，目标检测

引用格式：姚艳清,程塨,谢星星,韩军伟.2021.多分辨率特征融合的光学遥感图像目标检测.遥感学报,25(5): 1124-1137
Yao Y Q, Cheng G, Xie X X and Han J W. 2021. Optical remote sensing image object detection based on multi-resolution feature fusion. National Remote Sensing Bulletin, 25(5):1124-1137[DOI:10.11834/jrs.20210505]

1 引 言

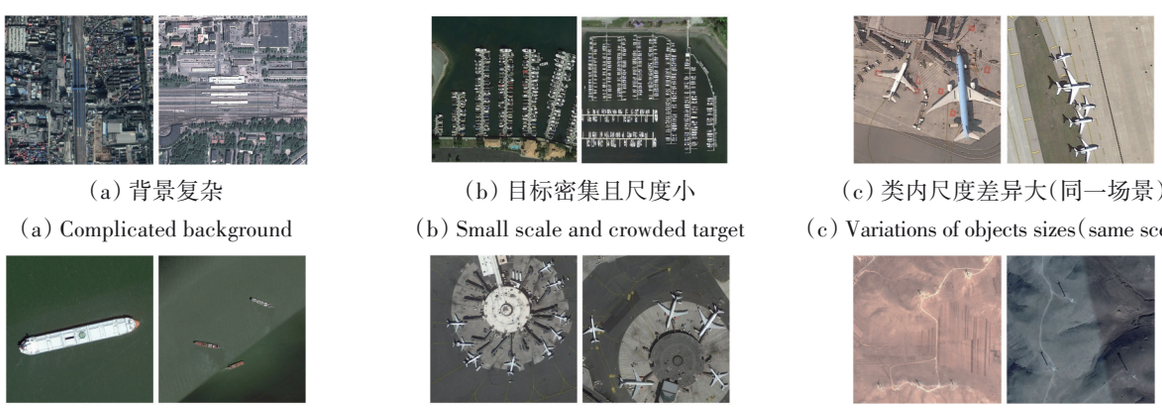
遥感图像目标检测是开展环境监测、城市规划、精准农业和土地测绘等对地观测应用的关键技术，具有重要的科研价值与广泛的应用前景（Cheng和Han，2016；Li等，2020，2018a；Xia等，2018；Zhou等，2019；Cheng等，2018，2019；孙显等，2020；龚健雅和钟燕飞，2016；周培诚等，2021）。在环境监测和资源探测方面，可以有效识别森林与植被覆盖率（曹琼等，2019），辅助城市有效规划等；在农业调查方面，可以快速获取农作物种植分布和种植面积等农情信息（陈凯强等，2020）；在海洋监测方面，可以识别海上船舰、港口等重要海况（姚红革等，2020），对军事侦察和民用监测具有重要意义。

传统的遥感图像目标检测算法，如方向梯度直方图HOG（Histogram of Oriented Gradient）（Dalal和Triggs，2005）、尺度不变特征变换SIFT（Scale-Invariant Feature Transform）（Lowe，1999）等，大多通过收集大量的先验知识来设计手工特征，然后以手工特征模板滑动整幅图像，选取激活值较高的区域，训练分类器对目标进行检测和识别。而随着计算机性能的提升，基于深度学习的目标检测算法在检测精度和速度上均远超传统检测算法，使遥感图像目标检测有了新的突破。R-CNN由Girshick等（2014）提出，提高了候选边界框的质量并可以提取深层特征。为减少R-CNN（Girshick等，2014）产生的大量冗余候选框，Fast R-CNN（Girshick，2015）、Faster R-CNN（Ren等，2017）相继被提出，Fast R-CNN（Girshick，2015）选择在特征图上进行感兴趣区域ROI（Region of Interesting）提取，从而减少计算量。而Faster R-CNN（Ren等，2017）摒弃了选择性搜索（Selective Search），采用区域生成网络

收稿日期：2020-11-12；**预印本**：2021-02-09
基金项目：深圳市科技创新委员会基金(编号：JCYJ20180306171131643)；国家自然科学基金(编号：61772425)
第一作者简介：姚艳清，1994年生，女，博士研究生，研究方向为高分辨率遥感图像理解。E-mail：yaoyanq@mail.nwpu.edu.cn
通信作者简介：程塨，1984年生，男，研究员，研究方向为高分辨率遥感图像理解。E-mail：gcheng@nwpu.edu.cn

RPN（Region Proposal Network）生成ROI，有效地提升了网络效率。为减少计算复杂度，R-FCN（Dai等，2016）将最后的全连接层替换卷积层。另一方面，为提高检测速度，以YOLO（Redmon等，2016）、SSD（Liu等，2016）为代表的单阶段目标检测算法相继提出。YOLO（Redmon等，2016）将目标检测作为回归问题，在输入图像上划分区域并进行目标的检测与定位。SSD（Liu等，2016）则很好地结合了Faster R-CNN（Ren等，2017）中的锚框机制和YOLO（Redmon等，2016）的回归思想，使得检测速度和准确率之间达到了一个较高水平的均衡。YOLO9000（Redmon和Farhadi，2017）在YOLO（Redmon等，2016）上增加了联合训练算法和批归一化BN（Batch Normalization，）层，进一步提升了检测速度。为解决目标检测中难、易样本不均衡的问题，RetinaNet（Lin等，2017b）中提出了Focal Loss损失函数。与此同时，基于深度学习的目标检测算法已推广至各个领域，不少学者开始致力于研究遥感图像目标检测。

然而，遥感图像与自然图像不同，由于成像平台与成像方式的不同，遥感图像具有背景复杂多样、目标尺度差异大、且方向任意等特点，导致检测效果不佳，如图1中的DIOR（Li等，2020）数据集样例所示。对此，不少研究学者提出了解决上述问题的遥感图像目标检测方法。



（a）背景复杂 （a）Complicated background
（b）目标密集且尺度小 （b）Small scale and crowded target
（c）类内尺度差异大（同一场景） （c）Variations of objects sizes（same scene）
（d）类内尺度差异大（不同场景） （d）Variations of object sizes（different scene）
（e）目标方向任意 （e）Arbitrary orientation
（f）俯瞰的视角 （f）Overylook perspective

图1 DIOR数据集样本

Fig. 1 Some examples of DIOR data set

（1）针对遥感图像类间相似性高的特点，文献（Cheng等，2020）提出一种跨尺度特征融合检测框架CSFF（Cross-Scale Feature Fusion），可以有效提升遥感图像目标检测精度。作者首先通过在顶层插入SE（Squeeze-and-Excitation Networks）模块来建模不同特征通道之间的联系，然后通过CSFF模块来融合不同尺度的特征，使检测器和分类器获得更为有效和更具有判别性的特征表达。Ma等（2019）提出一种多模型决策融合目标检测框架，分别建立了像素级上下文特征融合与目标级上下文特征融合子网络。上述两种方法都可以提高目标的检测精度，但网络计算量大幅增加。（2）针对遥感图像目标的多尺度特点，Hamaguch和Hikosaka（2018）提出一种多任务模型，通过学习多个检测器，每个检测器都专门用于检测固定大小的目标。FMSSD（Wang等，2020）利用由特征金字塔和多采样率构成的空间特征金字塔模块，将上下文信息融合到多尺度特征中。Li等（2018b）提出一种可以同时检测几十个像素到数千个像素的目标的端到端训练网络，利用分层选择滤波层，将不同尺度的特征映射到同一尺度空间。Chen等（2019）提出了基于上下文特征金字塔的多尺度目标检测框架，通过增强场景与目标之间的联系，提高多尺度目标检测的性能。这些方法大多通过特征金字塔结构来实现多尺度目标检测，但在检测精度与检测速度上难以实现平衡。（3）针对遥感图像小目标难以精准定位的问题，Long等（2017）提出一种边界框回归算法，并结合非极大值抑制算法对检测到的目标区域的边界盒进行优化。SCRDet（Yang等，2019）则提出新

的目标检测架构，设计了特征融合和锚点采样来提高小目标的检测精度，同时针对遥感目标的密集分布，提出了一种基于监督的注意力网络，以减少背景噪声的不利影响。但上述3种方法存在误分类问题，比如大型车辆与小型车辆，直升机与飞机。（4）针对于遥感图像目标的任意旋转角度问题，Cheng等（2016）提出一种学习旋转不变CNN模型RICNN（Rotation Invariant Convolutional Neural Networks），通过在现有CNN体系结构的基础上引入并学习新的旋转不变层来提高目标检测的性能。但是网络新增加的旋转不变层与原始CNN模型相比会引入额外的计算代价，本文将尝试在CNN模型的目标函数中嵌入旋转不变正则化器，而不引入额外的层来提高计算效率。

虽然在遥感图像目标检测领域已经取得了长足的进步，但遥感图像目标尺度差异大和类间相似性高仍然是两个亟待解决的问题。多尺度特征提取与融合是解决这类问题的有效方法。特征金字塔相较于图像金字塔可有效减少计算量，同时可嵌入到多种检测框架中。但由于特征金字塔分层检测不同尺度的目标，使得各层特征没有得到有效的融合。且每层特征学习到的特征单一。为获得更具判别性的多层特征表达，本文提出一种端到端训练的多分辨率特征融合模型。

2 本文方法

众所周知，无论是分类任务还是检测任务，特征提取的质量将直接影响最终的分类和检测精度。有效的特征表达可使目标更具判别性，并且定位更精准，易于检测。相反，不充分的特征表达将会导致目标定位不精确，在相似类间难以正确分类。因此，在遥感图像中选择一个有效的特征提取方法尤为重要。目前，由于神经网络通过训练可直接从原始图像像素中生成目标的特征表达，使得基于卷积神经网络CNN（Convolutional Neural Networks）的方法成为目标检测任务中提取特征的主要技术。因此，本文选择一种基于卷积神经网络的方法来提取光学遥感图像中目标检测的特征。

随着网络层数的加深，深层的特征图具有更丰富的语义信息，有益于目标的分类。而浅层的特征图则包含更准确的位置信息，有益于目标的定位。而特征金字塔首创性的提出将深层特征和浅层特征通过自上而下的支路进行有效的融合。作者选取每个阶段的最后一个残差块作为特征激活输出，对于Conv2、Conv3、Conv4、Conv5，作者将这些残差块的输出记为$\{C_2, C_3, C_4, C_5\}$，且相对于输入图像的步长为{4，8，16，32}。此处，作者没有选取Conv1参与预测的原因是Conv1特征层的尺度太大，会极大的增加网络运行时的空间复杂度。在自上而下的路径中，将每一层特征上采样2倍后与邻近的特征层逐像素合并，在合并前通过1×1的卷积将通道统一为256。最后将合并后的特征图经3×3的卷积生成最终的特征图$\{P_2, P_3, P_4, P_5, P_6\}$，其中$P_6$为$C_5$下采样所得。且在FPN（Lin等，2017a）中，作者在每个尺度的特征层上仅放置一个尺度的锚框，分别为$\{32^2, 64^2, 128^2, 256^2, 512^2\}$，长宽比为{1 ： 2，1 ： 1，2 ： 1}。这样的设置，使FPN（Lin等，2017a）存在以下几个缺点：

（1）由于每个尺度的特征层仅被分配检测单一尺寸的目标，导致每一层学到的目标的多尺度信息有限。

（2）特征金字塔的特征集的抽取方式，导致不同的特征层之间存在一定的语义差距。

（3）仅通过1×1的卷积将通道统一到256，使深层特征图丢失了大量信息。

因此，本文着重关注如何有效的结合深层特征与浅层特征各自的优点，生成更具判别性的特征，我们通过多分辨率特征融合网络实现。

本文算法在采用特征金字塔FPN（Feature Pyramid Network）结构（Lin等，2017a）的Faster R-CNN基础上进行改进。如图2所示，整个网络结构包括3个部分，分别为多分辨率特征提取网络、自适应特征融合模块和双尺度特征深度融合模块。为适应遥感图像中目标尺度大差异性，本文设计了多分辨率特征提取网络，目的是迫使网络学习多种分辨率目标的特征，从而提高网络对不同尺度的目标的学习能力。其次，自适应特征融合模块实质是通道和空间注意力机制，先通过通道注意力来建立通道间的联系，通过通道间的关联性来自适应的调整通道的特征响应，然后通过空间注意力来学习目标的空间位置信息，以此来建模不同通道、不同空间的信息，通过建立通道间和空间的关联性来自适应的调整通道、空间

的特征响应，进而起到抑制背景噪声，突出目标特征的作用。双尺度特征深度融合模块主要用于对相邻两个特征尺度进行特征融合和特征增强，利用双尺度特征深度融合模块，合并相邻两个尺度的特征上下文信息，更好地实现目标的特征提取，从而解决类间相似性问题。本文将ResNet（He 等， 2016） 的 conv1、 conv2_3、 conv3_4、conv4_6、conv5_3的输出作为多分辨特征提取网络的输入，并依次经过自适应特征融合模块和双尺度特征深度融合模块，使用Faster R-CNN检测器进行目标的分类与边界框回归。

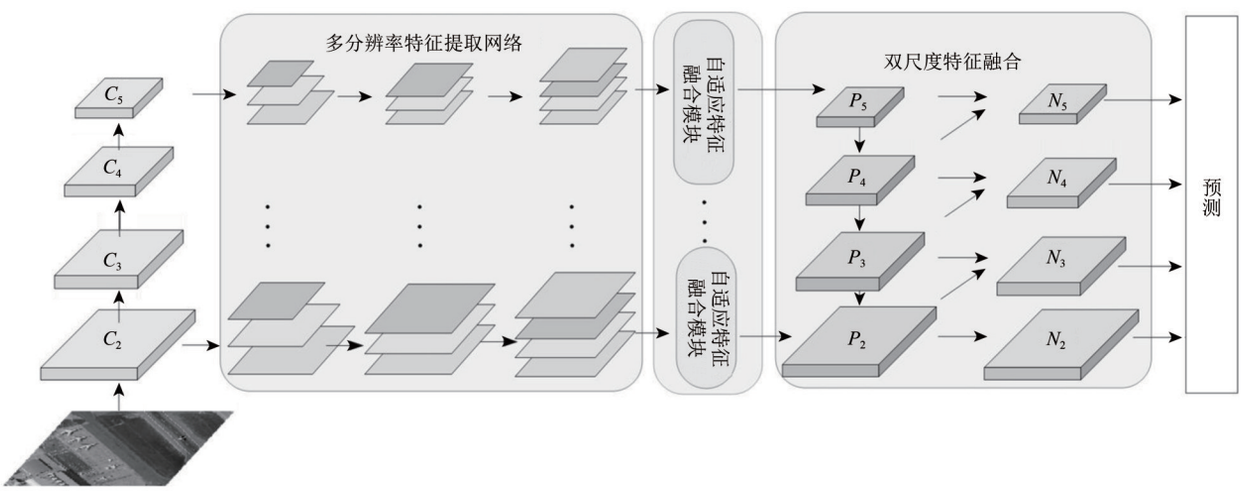


图2 多分辨率特征融合目标检测网络框架

Fig. 2 Overview of the proposed multi-resolution feature fusion object detection method

2.1 多分辨率特征提取网络MFE（Multi-resolution Feature Extract Network）

不同于特征金字塔结构，为缓解上述3个问题，本文提出一种多分辨特征提取网络。（1）将特征金字塔中的每个尺度的特征层分别下采样到不同分辨率，使每一层网络可以学习目标的不同分辨率下的特征。（2）多分辨率特征提取网络将特征金字塔的每个尺度的特征又细化为不同比例，由此缩短了网络的每一个尺度间的语义差距。（3）用多分辨率特征，相当于增加网络复杂度，可以缓解深层高通道特征层通过1×1卷积将通道骤减带来的信息损失问题。具体操作如图3所示，首先，本文抽取不同尺度的特征层C_i，i=2，3，4，5作为基础特征，通过平均池化生成多个不同分辨率的特征图$\{M_{i1}, M_{i2}, M_{i3}\}$，比例系数为$\{\mu_1, \mu_2, \mu_3\}$，通道数不变。其次，为了缓解深层网络通道骤减带来的信息损失，本文首先通过1×1的卷积将当前特征图的通道C降为$C/2$，然后再经过1×1的卷积将通道降为256，生成中间层特征$\{N_{i1}, N_{i2}, N_{i3}\}$。随后，为充分的利用不同分辨率特征的信息，通过上采样将特征图变为与C_i尺度一致，生成特征图$\{F_{i1}, F_{i2}, F_{i3}\}$。最后，将$C_i$用1×1的卷积降维后与$\{F_{i1}, F_{i2}, F_{i3}\}$拼接，得到多分辨率特征提取网络的输出$F_i$。多分辨率特征提取的计算为

$$Z_{\mathrm{M}}(C_{\mathrm{i}}) = Interp\left(f_{256}^{7\times7}\left(f_{C/2}^{7\times7}\left(AvgPool^{\mu_1,\ \mu_2,\ \mu_3}(C_{\mathrm{i}})\right)\right)\right) = Interp\left(f_{256}^{7\times7}\left(f_{C/2}^{7\times7}\left(M_{i1};\ M_{i2};\ M_{i3}\right)\right)\right) \quad (1)$$

式中，$Interp$表示插值上采样，$f_{256}^{7\times7}$表示卷积核的大小为7×7，通道数为256，$f_{C/2}^{7\times7}$表示卷积核大小为7×7，通道数为$C/2$，$AvgPool$表示平均池化操作。

如此，本文在迫使每一层网络学习目标在不同分辨率下的特征的同时，减少了传统特征金字塔不同尺度特征层之间的语义差距，并缓解了深层网络通道骤减所带来的信息丢失问题，加深了网络的语义表达。需要说明的是，由于ResNet（He等，2016）中Conv2_3的通道数为256，所以，以C_2为基础特征的多分辨率特征提取网络的整个过程中通道数均为256。

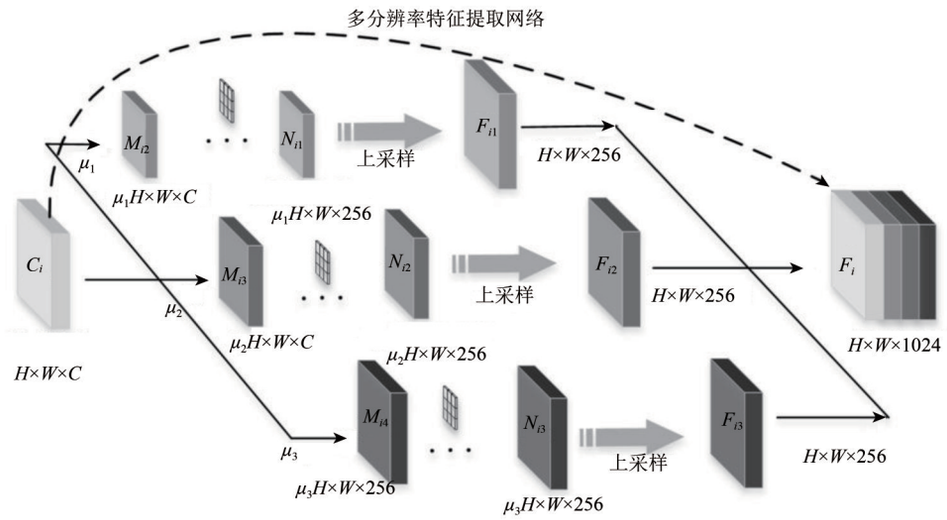


图3 多分辨率特征提取网络示意图

Fig. 3 Structure of multi-resolution feature extract network

2.2 自适应特征融合模块AFF（Adaptive Feature Fusion Module）

为了更好地利用多分辨率特征，本文采用注意力机制让网络更加关注目标的有效信息，抑制背景噪声。如图4所示，将多分辨特征提取网络的输出F作为自适应特征融合模块的输入，然后分别经过通道注意力和空间注意力模型，生成最终的融合特征F''。通道注意力利用特征的通道间的关系，使网络在优化的过程中逐渐学习出对目标检测结果有益的通道。首先，利用平均池化与最大池化操作来聚合特征图的空间信息，生成两个不同的空间向量：$\boldsymbol{F}^c_{\text{avg}}$与$\boldsymbol{F}^c_{\text{max}}$，分别表示平均池化特征与最大池化特征。然后，将两个向量送入多层感知器MLP（Multilayer Perceptron ）以生成通道注意力向量。随后将两个不同的通道注意力向量逐元素相加，得到最终的通道注意力向量，并与特征$\boldsymbol{F}$相乘生成通道注意力特征图$\boldsymbol{F}'$。通道注意力的计算为

$$\begin{aligned} M_c(\boldsymbol{F}) &= \left(\sigma\left(\text{MLP}\left(Pool_{\text{Avg}}(\boldsymbol{F})\right)+\right.\right. \\ &\left.\left.\text{MLP}\left(Pool_{\text{Avg}}(\boldsymbol{F})\right)\right)=\sigma\left(w_1\left(w_0\left(\boldsymbol{F}^c_{\text{avg}}\right)\right)+\right. \\ &\left. w_1\left(w_0\left(\boldsymbol{F}^c_{\text{max}}\right)\right)\right) \end{aligned} \tag{2}$$

式中，σ代表sigmoid函数，w_0、w_1为多层感知器MLP的网络参数，且对于两个不同的空间向量参数共享，$Pool_{\text{Avg}}$为平均池化操作，$Pool_{\text{max}}$为最大池化操作。空间注意力则主要关注目标的位置信息，即在网络优化的过程中，逐渐学习并突出对检测结果有益的局部特征。首先，同样利用平均池化和最大池化操作来聚合特征图的通道信息，得到两个不同的二维向量$\boldsymbol{F}^s_{\text{avg}} \in \boldsymbol{R}^{1\times \text{H}\times \text{W}}$与$\boldsymbol{F}^s_{\text{max}} \in \boldsymbol{R}^{1\times \text{H}\times \text{W}}$，分别表示跨通道的平均池化特征与最大池化特征。然后通过卷积生成的二维空间注意力特征向量。空间注意力的计算为：

$$\begin{aligned} M_c(F') &= \sigma\left(f^{7\times 7}\left(Pool_{\text{Avg}}(F')\,Pool_{\text{max}}(F')\right)\right)= \\ &\sigma\left(f^{7\times 7}\left(F^s_{\text{avg}};\ \ F^s_{\text{max}}\right)\right) \end{aligned} \tag{3}$$

式中，$f^{7\times 7}$代表7×7的卷积操作。利用通道注意力与空间注意力来融合多分辨率特征可以充分的挖掘不同分辨率的有效信息，增强特征表达，从而生成更具判别性的特征，解决遥感图像目标类间相似性大的问题。

2.3 双尺度特征深度融合模块DFDF（Dual-scale Feature Deep Fusion Module）

在特征金字塔中，同一目标可能会被分配到两个不同尺度的特征层中进行检测。为提取更具判别性的特征，本文融合两个相邻尺度的特征层。在算法整体概述中，提到随着网络层数的加深，深层特征图具有更丰富的语义信息，而浅层特征图则具有更多的位置信息。直接将5个不同尺度的特征层全部融合，会大大增加计算量，且在步长差距较大的特征层之间进行融合，需要将特征图进行高采样比的上采样或者下采样，容易丢失特征或使特征模糊，导致最终的融合结果出现混叠效应。因此，在均衡检测性能与计算量两个重要因素后，本文选取语义差距较小的相邻两个尺度

的特征进行融合。这样能够提取更具判别性的特征，也有利于检测尺度相近的目标。

双尺度特征深度融合模块结构图如图5（a）所示。本文将自适应特征融合模块的输出$\{P_2, P_3, P_4, P_5, P_6\}$作为双尺度特征深度融合模型的输入。首先将$P_i$，$i=2$，3，4，5，6下采样到与$P_{i+1}$相同尺度，然后$P_{i+1}$经过3×3卷积与下采样后的$P_i$像素求和，再经过3×3的卷积消除混叠效应，生成最终的预测特征层$\{N_2, N_3, N_4, N_5, N_6\}$。其中，$N_2$则由$P_2$通过3×3的卷积直接生成。具体计算为

$$N_{i+1}=f^{3\times3}\left(f^{3\times3}\left(P_{i+1}+Interp\left(P_i\right)\right)\right) \quad (4)$$

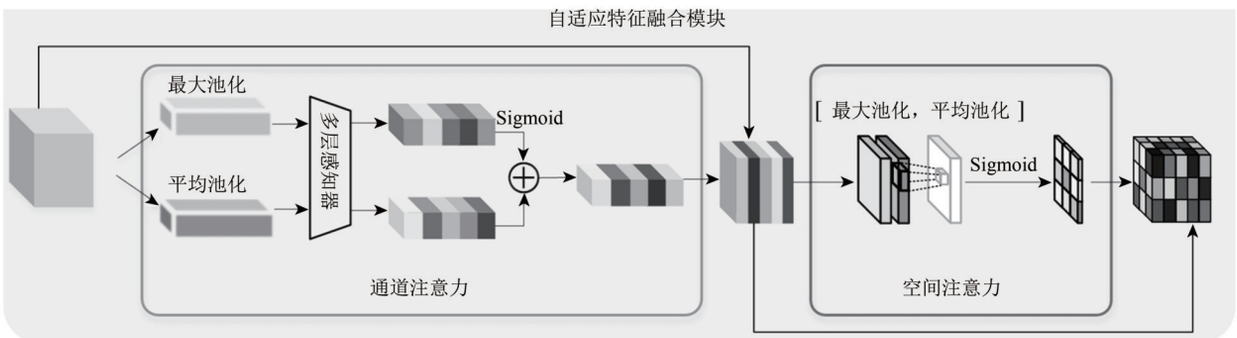
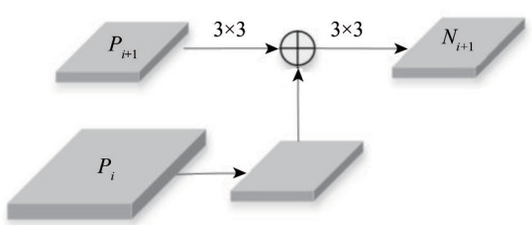


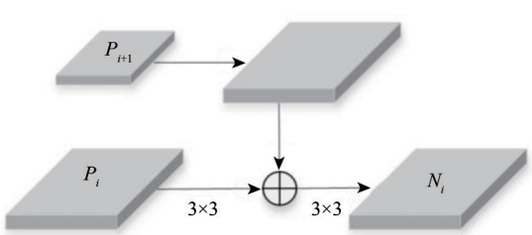
图4 自适应特征融合模块结构示意图

Fig. 4 Structure of adaptive feature fusion module



（a）结构1

（a）First structure



（b）结构2

（b）Second structure

图5 双尺度特征深度融合模块结构示意图

Fig. 5 Structure of dual-scale feature deep fusion module

在实验中，本文也尝试了图5（b）中的结构，即将P_{i+1}上采样到与P_i相同尺度，然后P_i经过3×3的卷积后与上采样后的P_{i+1}进行像素求和，再经过3×3生成最终的检测结果$\{N_2, N_3, N_4, N_5, N_6\}$。其中，$N_6$由$P_6$经过3×3的卷积得到。具体计算为

$$N_i=f^{3\times3}\left(f^{3\times3}\left(P_i+Interp\left(P_{i+1}\right)\right)\right) \quad (5)$$

3 实验与分析

3.1 实验数据

为评估本文提出的算法的有效性，本文在遥感图像公开检测数据集DIOR（Li等，2020）和DOTA（Xia等，2018）上进行实验验证。DIOR数据集共有23463张遥感图像，共包含192472个实例样本，有20个类别，分别为飞机、机场、棒球场、篮球场、桥梁、烟囱、水坝、高速公路服务区、高速公路收费站、港口、高尔夫球场、田径场、立交桥、轮船、体育场、储罐、网球场、火车站、车辆和风车，分别表示为C1—C20，图幅大小为800×800。其中训练集包括11725张图像，测试集包括11738张图像。DOTA数据集共有2806张图像，188282个实例，共15个类别：飞机、船、储罐、棒球场、网球场、游泳池、田径场、港口、桥梁、大型车辆、小型车辆、直升机、环岛、足球场和篮球场。图幅大小为800—4000。

在消融实验部分，采用FPN的Faster R-CNN（Ren等，2017）作为基准，并以ResNet50作为特征提取骨干网络。利用在ImageNet （Russakovsky等，2015）数据集预训练的模型参数对网络进行初始化。实验平台为mmdetection，版本号为1.2.0。选取SGD作为网络优化器，动量设为0.9，初始学

习率设为0.01，衰减系数为0.0001，batch-size为2，迭代12个epoch，分别在第8个和第11个epoch降学习率。模型在两张NVIDIA GeForce GTX 2080 Ti上训练及测试。

3.2 评价指标

本文采用已被广泛应用于目标检测评估的平均精度均值MAP（Mean Average Precision），MAP是指多个类别的平均精度AP（Average Precision）的平均值，每个类别都可以根据准确率（precision）和召回率（recall）绘制一条曲线，其在0到1区间范围内绘制的曲线与坐标轴所围成的面积即为平均精度，具体可表示为$AP = \int_0^1 p(r)\mathrm{d}r$。其中，准确率和召回率的定义如式（6）、（7）所示。

$$precision = \frac{TP}{TP + FP} \tag{6}$$

$$recall = \frac{TP}{TP + FN} \tag{7}$$

式中，TP代表真正例，FN代表假反例，FP代表假正例。实验中，统一采用VOC12的评价标准。

3.3 消融实验

3.3.1 多分辨率特征提取网络消融实验

为验证多分辨率特征提取网络的有效性，本文在基准上仅嵌入MFE（Multi-resolution Feature Extract Network）模型。所有的消融实验均以ResNet50作为骨干网络，并在DIOR（Li等，2020）数据集上训练并测试。实验结果如表2所示，嵌入MFE模块后，平均精度mAP为72.1，相较于基准模型提高了1.4%。由于mmdetection版本的不同，本文的基准相比Li等（2020）较高。同时，测试了不同的超参数对多分辨率特征提取网络的影响，为了使网络获得合适的融合特征，选取的比例系数控制在$(2, 10)$。实验结果如表1所示，最终比例系数$\{\mu_1, \mu_2, \mu_3\}$为{4，8，10}。这组的实验结果最优，平均精度mAP达到了73.2。需要说明的是，为了使整个网络学习到最好的超参数，超参数的选取实验是在基准模型上嵌入MFE、AFF和DFDF模块后的模型上进行的测试。

为了探究多分辨特征提取网络模块在遥感图像目标检测中起到的作用，将不同分支的分辨率特征热图进行了可视化，同时也将基准模型的特征图进行可视化用以对比。本文取基准特征图C_2与对应的3个不同分辨率特征图分支$\{F_{21}, F_{22}, F_{23}\}$进行可视化。本文取多个输入图像的特征图进行分析，结果如图6所示，每一行为一组输入图像的目标，每一行的5张图像分别为原始图像、基准特征热图以及3种分辨率下的特征热图。

表1 多分辨率特征提取网络超参数选取实验

Table 1 Experimental results of hyper-parameter selection of multi-resolution feature extraction network

μ_1, μ_2, μ_3	MAP
{2, 3, 4}	73.0
{2, 4, 6}	72.8
{2, 4, 8}	72.8
{4, 6, 8}	72.5
{4, 8, 10}	73.2

由图可以看出，在基准特征热图中，图像中的目标的关键特征点没有被有效的激活，导致网络的关注区域分散，不能有效提取有助于目标识别的关键特征。而在多尺度特征热图中，可以很明显的观察到不同的目标在不同分辨率的特征图中受到关注。如较窄的桥梁，密集的飞机在高分辨率的分支得到增强，而体育场等占地较大的建筑则在较低分辨率的分支得到增强，从而可以证明融合多分辨率分支对于增强目标的特征有效性。

此外，在实验中还发现，高分辨率的分支对物体的边缘信息较为敏感从而可以增强细小目标的边缘信息，而低分辨率的分支则对物体的主体信息更敏感。这也能从侧面说明多分辨率分支提取到的特征确实是对基准特征的有效补充。

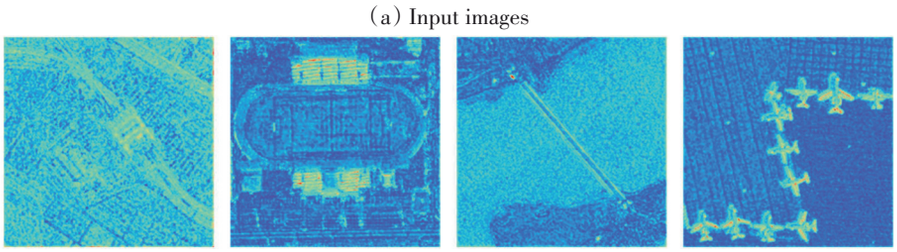
3.3.2 自适应特征融合模块消融实验

为验证自适应特征融合模块的有效性，我们在嵌入MFE的基准模型上加入AFF模块。且由于AFF（Adaptive Feature Fusion）模块的输入为MFE模块的输出，因此通过比较表2中的第3行与第5行，得出AFF模块可以在嵌入MFE模块的基准模型基础上将性能指标mAP提升0.5%。同样，可视化了AFF模块之后的特征热图，用来展示自适应特征融合模块的工作机制。如图7所示，6组图像分别为原始图像、基准特征、经自适应特征融合后的特征和类激活图。从图7中可以直观的看出，

较原始的特征图，融合后的特征图可以在抑制背景噪声的同时，将网络的注意力有效的聚焦在目标上，从而显著增强目标信息的表达，有助于提高网络的检测精度。

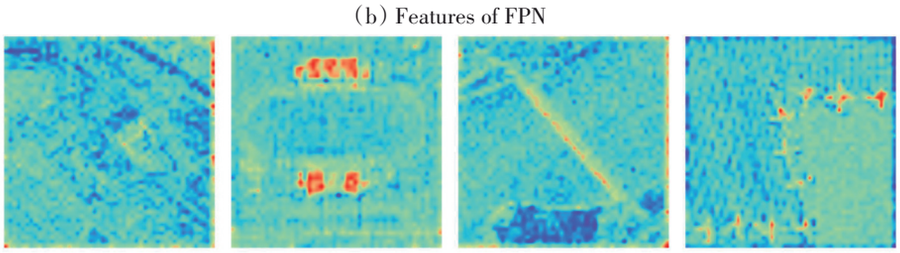
（a）原图

（a）Input images



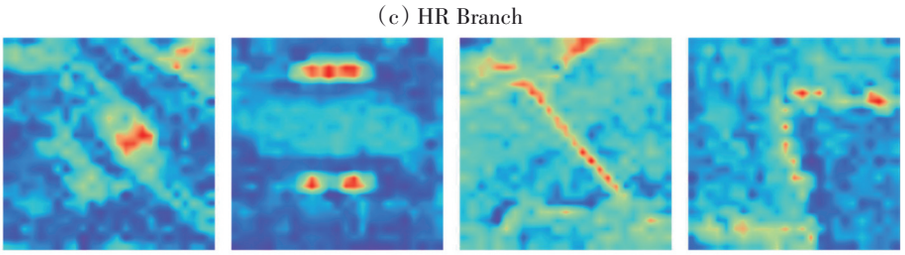
（b）基准特征

（b）Features of FPN



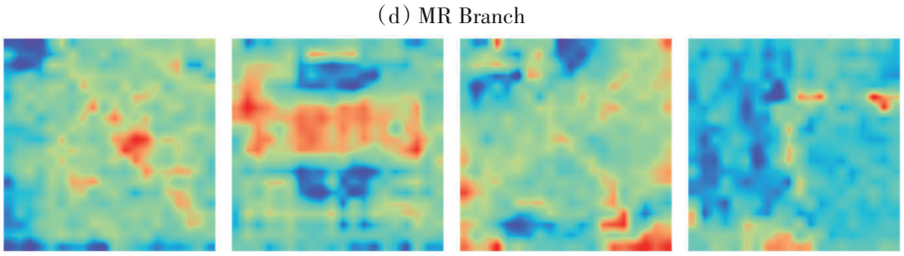
（c）高分辨率分支

（c）HR Branch



（d）中分辨率分支

（d）MR Branch



（e）低分辨率分支

（e）LR Branch

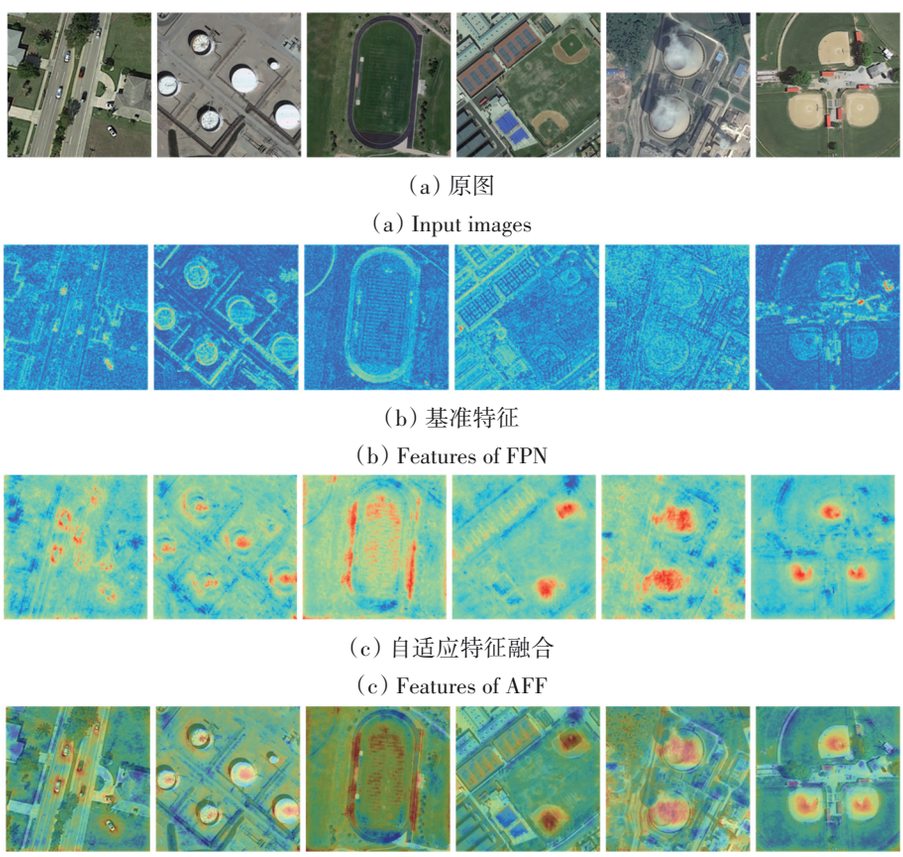
图6　多分辨率特征提取网络生成特征响应热图

Fig. 6　Feature heat maps of multi-resolution feature extraction network

3.3.3　双尺度特征深度融合模块消融实验

最后，也验证了DFDF（Dual-scale Feature Deep Fusion）模块的有效性，如表2所示，在基准模型上仅嵌入DFDF模块后，网络的检测性能指标

平均精度由70.7提升至72.0，提升了1.3%。同时，本文还尝试了图5（b）中的结构，与图5（a）中的结构不同的是，图5（b）中的结构将P_{i+1}上采样到与P_i相同尺度，然后P_i经过3×3的卷积后与上采样后的P_{i+1}进行像素求和，再经过3×3的卷积生成最终的检测结果$\{N_2, N_3, N_4, N_5, N_6\}$。在实验中，发现图5（b）结构不仅具有更高的网络复杂度，且在最终的检测精度较图5（a）中的结构也略低，分析可能的原因是使用插值进行上采样生成较大的特征图不利于网络提取有效的高维信息表达，从而降低网络的性能。由表2可见，双尺度特征深度融合模块可以通过融合两个相邻尺度的特征图，达到增强特征判别性，提高网络检测性能的目的。最终，在基准模型上同时嵌入本文提出的MFE、AFF与DFDF模块后，平均精度mAP可以达到73.2，相较于基准模型提高了2.5%。



（a）原图

（a）Input images

（b）基准特征

（b）Features of FPN

（c）自适应特征融合

（c）Features of AFF

（d）类激活图

（d）Class-specific activation maps

图7 自适应特征融合模块生成特征响应热图

Fig. 7 Feature heat maps of adaptive feature fusion module

表2 本文算法在DIOR数据集的消融实验结果比较

Table 2 Ablation studies on the DIOR data set

	MFE	AFF	DFDF	mAP/%
Baseline				70.7
√	√			72.1
√			√	72.0
√	√	√		72.6
√	√	√	√	73.2

注：“√”表示在基准网络的基础上添加相应的模块。

3.4 实验结果

为验证本文方法的有效性，本文与当前较为有效的遥感图目标检测算法的进行比较。在DIOR数据集上，比较了Faster R-CNN（Ren等，2017）、采用FPN（Lin等，2017a）结构的Faster R-CNN（Ren等，2017）和Mask R-CNN（He等，2017）、PANet（Liu等，2018）、以及CSFF（Cheng等，2020）。Faster R-CNN（Ren等，2017）、采用FPN结构的Faster R-CNN和Mask R-CNN（He等，

2017）和PANet（Liu等，2018）的实验结果均以Li等（2020）为准，CSFF则采用Cheng等（2020）中的结果。比较结果列为表3结果显示，本文算法与其他几种算法相比，检测结果最优，且相较于其他算法，在多数类别上均取得较大优势，尤其在飞机、棒球场、烟囱、体育场这几个类别上，以ResNet101为骨干的检测性能越超其他算法。而CSFF则在机场、篮球场、高速公路服务区、高速公路收费站几个类别上表现较为优异。

表3 几种算法在DIOR数据集的检测结果比较

Table 3 Comparisons of detection accuracy of six methods on the DIOR data set

/%

方法	骨干网络	地物类别																				mAP
		C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	
Faster RCNN	VGG16	53.6	49.3	78.8	66.2	28.0	70.9	62.3	69.0	55.2	68.0	56.9	50.2	50.1	27.7	73.0	39.8	75.2	38.6	23.6	45.4	54.1
Faster RCNN with FPN	ResNet50	54.1	71.4	63.3	81.0	42.6	72.5	57.5	68.7	62.1	73.1	76.5	42.8	56.0	71.8	57.0	53.5	81.2	53.0	43.1	80.9	63.1
	ResNet101	54.0	74.5	63.3	80.7	44.8	72.5	60.0	75.6	62.3	76.0	76.8	46.4	57.2	71.8	68.3	53.8	81.1	59.5	43.1	81.2	65.1
Mask RCNN with FPN	ResNet50	53.8	72.3	63.2	81.0	38.7	72.6	55.9	71.6	67.0	73.0	75.8	44.2	56.5	71.9	58.6	53.6	81.1	54.0	43.1	81.1	63.5
	ResNet101	53.9	76.6	63.2	80.9	40.2	72.5	60.4	76.3	62.5	76.0	75.9	46.5	57.4	71.8	68.3	53.7	81.0	62.3	43.0	81.0	65.2
PANet	ResNet50	61.9	70.4	71.0	80.4	38.9	72.5	56.6	68.4	60.0	69.0	74.6	41.6	55.8	71.7	72.9	62.3	81.2	54.6	48.2	86.7	63.8
	ResNet101	60.2	72.0	70.6	80.5	43.6	72.3	61.4	72.1	66.7	72.0	73.4	45.3	56.9	71.7	70.4	62.0	80.9	57.0	47.2	84.5	66.1
CSFF	ResNet101	57.2	79.6	70.1	87.4	46.1	76.6	62.7	82.6	73.2	78.2	81.6	50.7	59.5	73.3	63.4	58.5	85.9	61.9	42.9	86.9	68.0
Ours	ResNet50	66.7	**83.6**	74.8	**89.1**	**50.5**	80.6	**69.0**	**84.9**	**75.2**	**83.9**	**84.2**	53.8	65.2	75.6	74.6	62.7	88.1	**65.8**	46.4	**88.8**	73.2
	ResNet101	**91.0**	74.5	**93.3**	83.2	47.4	**91.9**	63.3	68.0	61.4	80.0	82.8	**57.4**	**65.8**	**80.0**	**92.5**	**81.1**	**88.7**	63.0	**73.0**	78.1	**75.8**

注：C1—C20表示飞机、机场、棒球场、篮球场、桥梁、烟囱、水坝、高速公路服务区、高速公路收费站、港口、高尔夫球场、田径场、立交桥、轮船、体育场、储罐、网球场、火车站、车辆和风车。黑体表示最优结果。

在DOTA数据集上，比较了R-FCN（Dai等，2016）、Faster R-CNN（Dai等，2016）、Deformable Faster R-CNN（记为Deformable FR-H）（Ren等，2018）、ICN（Azimi等，2019）以及FPN（Dai等，2016）。骨干网络及实验参数设置均以Xia等（2018）为准。比较结果如表4所示，本文方法相较于其他方法在多数类别上具有更高的检测精度，但在直升机这类目标上的检测效果一般。

表4 几种算法在DOTA数据集的检测结果比较

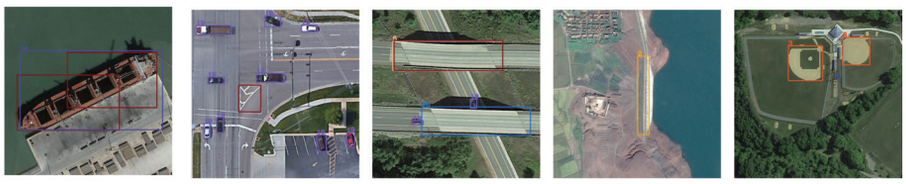
Table 4 Comparisons of detection accuracy of six methods on the DOTA data set

/%

方法	骨干网络	地物类别															mAP
		PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	
R-FCN	ResNet101	81.01	58.96	31.64	58.97	49.77	45.04	49.29	68.99	52.07	67.42	41.83	51.44	45.15	53.30	33.89	52.58
Faster R-CNN	ResNet101	80.32	77.55	32.86	68.13	53.66	52.49	50.04	90.41	75.05	59.59	57.00	49.81	61.69	56.46	41.85	60.46
Deformable FR-H	ResNet101	86.53	77.54	42.70	64.43	67.60	63.64	77.86	90.33	77.82	75.36	52.12	56.79	68.92	62.04	54.92	67.91
FPN	ResNet101	88.70	75.10	52.60	59.20	69.40	**78.80**	**84.50**	90.60	**81.30**	82.60	52.50	62.10	76.60	66.30	**60.10**	72.00
ICN	ResNet101	**90.00**	77.70	53.40	**73.30**	73.50	65.00	78.20	90.80	79.10	84.80	**57.20**	62.11	73.45	70.22	58.08	72.45
Ours	ResNet101	89.95	**83.79**	**55.91**	71.92	**79.74**	68.36	79.17	**90.87**	80.06	**85.00**	56.01	**62.74**	**77.53**	**75.05**	57.20	**74.20**

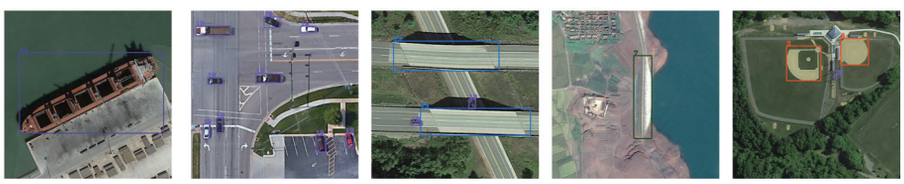
注：黑体表示最优结果。PL：飞机，BD：棒球场，BR：桥梁，GTF：田径场，SV：大型车辆，LV：小型车辆，SH：船，TC：网球场，BC：篮球场；ST：储罐，SBF：足球场，RA：环岛，HA：港口，SP：游泳池，AC：直升机。

图8为本文算法与采用FPN结构的Faster R-CNN的基准模型在DIOR数据集的检测结果可视化图比较，图8中红色的框表示虚警、深蓝色的框表示漏检。从图8中可以看出，本文算法的检测结果更准确，虚警、漏检等问题得到极大的改善，如图8所示，在第1列、第2列的检测结果中，基准模型检测的虚警明显比本文方法的检测结果高，对比第3列结果图发现，基准模型将图中的“立交桥”错检为“桥梁”，在第4列对比图中，基准模型将“水坝”误检为“机场”，而本文方法则可以准确识别。同样，对比第5列结果图发现，基准模型将图中的“小车”目标丢失（蓝色框），而本文方法可以有效地检测出来。对比5种图像，本文算法具有更好的鲁棒性，这是因为本文模型能够提取更具判别性的特征，且通过有效的自适应特征融合模块可以极大程度的抑制背景信息。



（a）基准模型

（a）Faster R-CNN



（b）本文方法

（b）Proposed method

图8 本文算法与基准模型在DIOR数据集的检测结果比较

Fig. 8 Comparison of detection results of our proposed method and Faster R-CNN with FPN on the DIOR data set

图9为本文算法的检测结果可视化图，从图9中可以看出本文算法可以有效的检测出多尺度目标，如飞机、储蓄罐等，并且对密集目标如储蓄罐、船只的检测结果表现也十分优异。此外，本文算法对复杂背景的抗干扰能力显著，比如由于光照导致的海水成镜像发射造成的光污染噪声和由于云雾天气导致的风电机组的遮挡。在这些复杂的背景下，本文算法依然能有效的检测出待测目标。

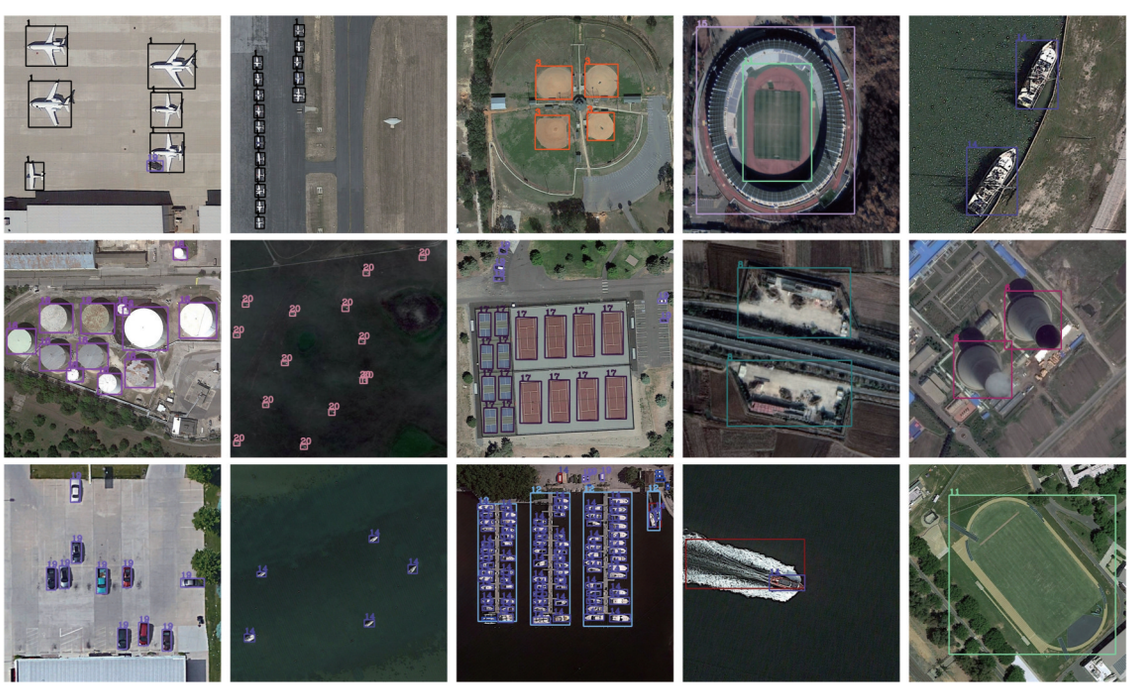


图9 本文方法在DIOR数据集的检测结果样例

Fig. 9 Some detection results of our proposed method on the DIOR data set

4 结 论

本文提出一种多分辨率特征融合的遥感图像目标检测算法，通过在Faster R-CNN结构上嵌入多分辨率特征提取网络（MFE）、自适应特征融合模块（AFF）和双尺度特征深度融合模块（DFDF）来提升网络检测性能。多分辨率特征提取网络旨在使网络学习目标在不同分辨率下的特征，同时减少了传统特征金字塔不同尺度特征层之间的语义差距，并缓解了深层网络通道骤减所带来的信

息丢失问题，增强了网络的语义表达。自适应特征融合模块则是为更好的融合多分辨率特征提取网络提取的多分辨率特征，通过注意力机制让网络更加关注目标的有效信息，抑制背景噪声。最后，利用双尺度特征深度融合模块融合两个相邻尺度的特征层，提取更具判别性的特征。为证明本文提出的3个模块的有效性，在遥感图像公开数据集DIOR上做了消融实验，实验结果显示，MFE、AFF和DFDF模块在基准模型的基础上分别可以将检测平均精度mAP提高1.4%、0.5%和1.3%。通过与目前较为有效的遥感图像目标检测算法进行比较，本文算法检测平均精度最优。在遥感图像公开数据集DIOR和DOTA也进行了测试，相较于基准模型，本文方法分别提高了2.5%和2.2%，mAP分别为73.2%和74.2%。

通过实验检测结果可视化图可以看出，本文的算法可以提取更具判别性的特征，可极大程度的改善错检、漏检等问题。且对密集目标、多尺度目标都有较好的检测性能，对复杂的背景噪声有较强的抑制作用。尽管，最终的效果有一定提升，但本文算法仍存在可以改进的部分。比如，我们的算法对快速移动的船只（快艇）的检测效果一般，可能的原因是游艇后常伴有较明显的浪花，导致游艇的特征被干扰。

参考文献（References）

Azimi S M, Vig E, Bahmanyar R, Körner M and Reinartz P. 2019. Towards multi-class object detection in unconstrained remote sensing imagery//14th Asian Conference on Computer Vision. Perth, Australia: Springer: 150-165 [DOI: 10.1007/978-3-030-20893-6_10]

Cao Q, Ma A L, Zhong Y F, Zhao J, Zhao B and Zhang L P. 2019. Urban classification by multi-feature fusion of hyperspectral image and LiDAR data. Journal of Remote Sensing, 23(5): 892-903 (曹琼, 马爱龙, 钟燕飞, 赵济, 赵贝, 张良培. 2019. 高光谱-LiDAR多级融合城区地表覆盖分类. 遥感学报, 23(5): 892-903) [DOI: 10.11834/jrs.20197512]

Chen C Y, Gong W G, Chen Y L and Li W H. 2019. Object detection in remote sensing images based on a scene-contextual feature pyramid network. Remote Sensing, 11(3): 339 [DOI: 10.3390/rs11030339]

Chen K Q, Gao X, Yan M L, Zhang Y and Sun X. 2020. Building extraction in pixel level from aerial imagery with a deep encoder-decoder network. Journal of Remote Sensing, 24(9): 1134-1142 (陈凯强, 高鑫, 闫梦龙, 张跃, 孙显. 2020. 基于编解码网络的航空影像像素级建筑物提取. 遥感学报, 24(9): 1134-1142) [DOI: 10.11834/jrs.20209056]

Cheng G and Han J W. 2016. A survey on object detection in optical remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing, 117: 11-28 [DOI: 10.1016/j.isprsjprs.2016.03.014]

Cheng G, Han J W, Zhou P C and Xu D. 2019. Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection. IEEE Transactions on Image Processing, 28(1): 265-278 [DOI: 10.1109/TIP.2018.2867198]

Cheng G, Li Z P, Han J W, Yao X W and Guo L. 2018. Exploring hierarchical convolutional features for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 56(11): 6712-6722 [DOI: 10.1109/TGRS.2018.2841823]

Cheng G, Si Y, Hong H L, Yao X W and Guo L. 2020. Cross-scale feature fusion for object detection in optical remote sensing images. IEEE Geoscience and Remote Sensing Letters, 18(3): 431-435 [DOI: 10.1109/LGRS.2020.2975541]

Cheng G, Zhou P C and Han J W. 2016. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 54(12): 7405-7415 [DOI: 10.1109/TGRS. 2016. 2601622]

Dai J F, Li Y, He K M and Sun J. 2016. R-FCN: object detection via region-based fully convolutional networks//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc.: 379-387 [DOI: 10.5555/3157096.3157139]

Dalal N and Triggs B. 2005. Histograms of oriented gradients for human detection//IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, CA, USA: IEEE, 1: 886-893 [DOI: 10.1109/CVPR.2005.177]

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

Girshick R, Donahue J, Darrell T and Malik J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 1: 580-587 [DOI: 10.1109/CVPR.2014.81]

Gong J Y and Zhong Y F. 2016. Survey of intelligent optical remote sensing image processing. Journal of Remote Sensing, 20(5): 733-747 (龚健雅, 钟燕飞. 2016. 光学遥感影像智能化处理研究进展. 遥感学报, 20(5): 733-747) [DOI: 10.11834/jrs.20166205]

Hamaguchi R and Hikosaka S. 2018. Building detection from satellite imagery using ensemble of size-specific detectors//IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, UT, USA: IEEE: 223-227 [DOI: 10.1109/CVPRW.2018.00041]

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

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Li K, Cheng G, Bu S H and You X. 2018a. Rotation-insensitive and context-augmented object detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 56(4): 2337-2348 [DOI: 10.1109/TGRS.2017.2778300]

Li K, Wan G, Cheng G, Meng L Q and Han J W. 2020. Object detection in optical remote sensing images: a survey and a new benchmark. ISPRS Journal of Photogrammetry and Remote Sensing, 159: 296-307 [DOI: 10.1016/j.isprsjprs.2019.11.023]

Li Q P, Mou L C, Liu Q J, Wang Y H and Zhu X X. 2018b. HSF-Net: multiscale deep feature embedding for ship detection in optical remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing, 56(12): 7147-7161 [DOI: 10.1109/TGRS. 2018. 2848901]

Lin T Y, Dollár P, Girshick R, He K M, Hariharan B and Belongie S. 2017a. Feature pyramid networks for object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE: 936-944. [DOI: 10.1109/CVPR.2017.106]

Lin T Y, Goyal P, Girshick R, He K M and Dollár P. 2017b. Focal loss for dense object detection//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2999-3007 [DOI: 10.1109/ICCV.2017.324]

Liu S, Qi L, Qin H F, Shi J P and Jia J Y. 2018. Path aggregation network for instance segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE: 8759-8768 [DOI: 10.1109/CVPR. 2018.00913]

Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y and Berg A C. 2016. SSD: single shot multibox detector//14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer: 21-37 [DOI: 10.1007/978-3-319-46448-0_2]

Long Y, Gong Y P, Xiao Z F and Liu Q. 2017. Accurate object localization in remote sensing images based on convolutional neural networks. IEEE Transactions on Geoscience and Remote Sensing, 55(5): 2486-2498 [DOI: 10.1109/TGRS.2016.2645610]

Lowe D G. 1999. Object recognition from local scale-invariant features//Proceedings of the Seventh IEEE International Conference on Computer Vision. Kerkyra, Greece: IEEE, 2: 1150-1157 [DOI: 10.1109/iccv.1999.790410]

Ma W P, Guo Q Q, Wu Y, Zhao W, Zhang X R and Jiao L C. 2019. A novel multi-model decision fusion network for object detection in remote sensing images. Remote Sensing, 11(7): 737 [DOI: 10. 3390/rs11070737]

Redmon J, Divvala S, Girshick R and Farhadi A. 2016. You only look once : unified, real-time object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE: 779-788 [DOI: 10.1109/CVPR.2016.91]

Redmon J and Farhadi A. 2017. YOLO9000: better, faster, stronger//Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE: 6517-6525 [DOI: 10.1109/CVPR.2017.690]

Ren S Q, He K M, Girshick R and Sun J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149 [DOI: 10.1109/TPAMI.2016.2577031]

Ren Y, Zhu C R and Xiao S P. 2018. Deformable faster R-CNN with aggregating multi-layer features for partially occluded object detection in optical remote sensing images. Remote Sensing, 10(9): 1470 [DOI: 10.3390/rs10091470]

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z H, Karpathy A, Khosla A, Bernstein M, Berg A C and Li F F. 2015. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 115(3): 211-252 [DOI: 10. 1007/s11263-015-0816-y]

Sun X, Liang W, Diao W H, Cao Z Y, Feng Y C, Wang B and Fu K. 2020. Progress and challenges of remote sensing edge intelligence technology. Journal of Image and Graphics, 25(9): 1719-1738 (孙显, 梁伟, 刁文辉, 曹志颖, 冯瑛超, 王冰, 付琨. 2020. 遥感边缘智能技术研究进展及挑战. 中国图象图形学报, 25(9): 1719-1738) [DOI: 10.11834/jig.200288]

Wang P J, Sun X, Diao W H and Fu K. 2020. FMSSD: feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing, 58(5): 3377-3390 [DOI: 10.1109/TGRS.2019.2954328]

Xia G S, Bai X, Ding J, Zhu Z, Belongie S, Luo J B, Datcu M, Pelillo M and Zhang L P. 2018. DOTA: a large-scale dataset for object detection in aerial images//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE: 3974-3983 [DOI: 10.1109/CVPR.2018.00418]

Yang X, Yang J R, Yan J C, Zhang Y, Zhang T F, Guo Z, Sun X and Fu K. 2019. SCRDet: towards more robust detection for small, cluttered and rotated objects//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 8231-8240 [DOI: 10.1109/ICCV.2019.00832]

Yao H G, Wang C, Yu J, Bai X J and Li W. 2020. Recognition of small-target ships in complex satellite images. Journal of Remote Sensing, 24(2): 116-125 (姚红革, 王诚, 喻钧, 白小军, 李蔚. 2020. 复杂卫星图像中的小目标船舶识别. 遥感学报, 24(2) 116-125) [DOI: 10.11834/jrs.20208238]

Zhou P C, Cheng G, Yao X W and Han J W. 2021. Machine learning paradigms in high-resolution remote sensing image interpretation. Journal of Remote Sensing, 25(1): 182-197 (周培诚, 程塨, 姚西文, 韩军伟. 2021. 高分辨率遥感影像解译中的机器学习范式. 遥感学报, 25(1): 182-197) [DOI: 10.11834/jrs.20210164]

Zhou P C, Han J W, Cheng G and Zhang B C. 2019. Learning compact and discriminative stacked autoencoder for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 57(7): 4823-4833 [DOI: 10.1109/TGRS.2019.2893180]

Optical remote sensing image object detection based on multi−resolution feature fusion

YAO Yanqing[1,2], CHENG Gong[1,2], XIE Xingxing[1,2], HAN Junwei[2]

1. Research & Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen 518057, China;
2. School of Automation, Northwestern Polytechnical University, Xi'an 710129, China

Abstract: In recent years, high-resolution remote sensing image object detection has attracted increasing interest and become an important research field of computer vision due to its wide applications in civil and military fields, such as environmental monitoring, urban planning, precision agriculture, and land mapping. The natural scene object detection frameworks based on deep learning have made a breakthrough progress. These algorithms have good detection performance on the open data sets of natural scenes. However, although these algorithms have greatly improved the accuracy and speed of remote sensing image object detection, they have not achieved the expected results. Given the large variations of object sizes and inter-class similarity, most of the conventional object detection algorithms designed for natural scene images still face some challenges when directly applied to remote sensing images. To address the above challenges, we propose an end-to-end multi-resolution feature fusion framework for object detection in remote sensing images, which can effectively improve the object detection accuracy. Specifically, we use a Feature Pyramid Network (FPN) to extract multi-scale feature maps. Then, a Multi-resolution Feature Extract (MFE) module, which can promote the network to learn the feature representations of the objects at different resolutions and narrow the semantic gap between different scales, is inserted into the feature layers of different scales. Next, to achieve an effective fusion of multi-resolution features, we use an Adaptive Feature Fusion (AFF) module to obtain more discriminative multi-resolution feature representations. Finally, we use a Dual-scale Feature Deep Fusion (DFDF) module to fuse two adjacent-scale features, which are the output of the adaptive feature fusion module. In the experiments, to demonstrate the effectiveness of each module of our proposed method, including the MFE, AFF, and DFDF modules, we first conducted extensive ablation studies on the large-scale remote sensing image data set DIOR, and the experimental results show that our proposed MFE, AFF, and DFDF modules could improve the average detection accuracy by 1.4%, 0.5%, and 1.3%, respectively, compared with the baseline method. Furthermore, we evaluate our method on two publicly available remote sensing image object detection data sets, namely, DIOR and DOTA, and obtain improvements of 2.5% and 2.2%, respectively, which are measured in terms of mAP comparison with Faster R-CNN with FPN. The detection results of the ablation studies and the comparison experiments indicate that our method can extract more discriminative and powerful feature representations than Faster R-CNN with FPN, which can significantly boost the detection accuracy. Moreover, our method works well for densely arranged and multi-scale objects. Although many improvements have been achieved in this work, some aspects still require improvement. For example, our method performs poorly in terms of detecting objects with big aspect-ratios, such as bridges, possibly because most anchor-based methods have difficulty ensuring a sufficiently high intersection over union rate with the ground-truth objects with big aspect-ratios. Our future work will focus on addressing these problems by exploring the advantages of anchor-free based methods.

Key words: convolutional neural networks, multi-resolution feature fusion, remote sensing images, object detection

Supported by Science, Technology and Innovation Commission of Shenzhen Municipality (No. JCYJ20180306171131643); National Natural Science Foundation of China (No. 61772425)