第 44 卷 第 6 期/2024 年 3 月/光学学报 研究论文

光学学报

增强小目标特征的多尺度光学遥感图像目标检测

单慧琳[1,2]，王硕洋[1]，童俊毅[1]，胡宇翔[2]，张雁皓[2]，张银胜[1,2*]

[1]南京信息工程大学电子与信息工程学院，江苏 南京 210044；

[2]无锡学院电子信息工程学院，江苏 无锡 214105

摘要 针对光学遥感图像目标分布密集、尺度变化范围较大及小目标特征信息过少等造成目标检测精度不高、泛化能力差等问题，本文提出了一种增强小目标特征的多尺度神经网络(ESF-MNet)。首先在骨干网络中引入注意力模块构建出高效层注意力聚合结构，以增强特征提取能力；此外，在浅层特征图与颈部网络融合之前加入感受野增强模块，以捕获不同尺度的上下文信息。其次，使用GSConv构成颈部网络，减少网络层参数量，保持网络的特征提取能力，并通过基于内容感知的特征重组模块提高识别精度。最后，采用下采样率分别为4、8和16倍的三个下采样模块作为头部网络输入，来提高小目标的检测效果。为了证明该方法的有效性，在DOTA数据集和NWPU NHR-10数据集上进行实验，平均检测精度分别达78.6%和94.3%，计算复杂度为94.7 G，整体模型大小为26.2 M。该方法具备检测精度高、计算复杂度低、模型权重小等特点，能有效提高小目标的检测精度，进一步改善光学遥感图像小目标检测性能。

关键词 光学遥感图像；目标检测；感受野增强；特征融合；注意力机制

中图分类号 TP753 **文献标志码** A **DOI**：10.3788/AOS231676

1 引 言

近年来，由于人造卫星和航空拍摄技术的进步，使得遥感技术的应用日益广泛，光学遥感图像在日常生活中有着广泛的运用[1-4]。然而，由于光学遥感图像存在尺度跨度大、目标尺寸小、分布不平衡和分布密集等问题，导致常常出现错检和漏检的情况，提高光学遥感图像的检测效果是当前亟待解决的问题。

随着计算机运算速度的不断提高，基于卷积神经网络(CNN)的目标检测算法得到了迅猛发展。基于深度学习的目标检测算法可以分为双阶段和单阶段检测。双阶段检测算法首先生成大量候选框区域，然后通过分类和回归的方式精确定位和识别目标。以R-CNN[5]、Fast R-CNN[6]、Faster R-CNN[7]和Mask R-CNN[8]等算法为典型代表，其检测精度高，但是效率相对较低。单阶段目标检测算法能够直接利用主干网络提取目标的特征，从而对目标的类型和位置进行预测。以SSD[9]、RetinaNet[10]和YOLO系列[11-14]等算法为典型代表，其大大提高了检测速度。然而，在处理复杂背景下的小目标检测时，上述算法仍存在一定的局限性。

近年来，国内外学者对光学遥感图像中的小目标检测进行了大量的研究，其中，Qu等[15]提出了一种新的基于空洞卷积的特征融合方法，有效提高了小目标的检测效果。闫钧华等[16]提出了一种基于多层次信息融合的目标特征分层提取方法，并在此基础上引入了多层次信息，从而提高了对光学遥感图像中小目标的探测能力。张寅等[17]提出了一种新型的“级联注意”方法，通过融合底层特征图中的多个感受野信息，增强了对小目标的捕捉能力。张廓等[18]提出了一种基于感知野与特征增强的轻型神经网络，利用深度可分卷积降低参数数量、加快检测速度，并引入感知野增强和注意力机制等模块，以提高检测准确性。薛俊达等[19]提出了一种基于目标框分组聚类与高效特征融合的FFC-SSD模型，利用反池化策略减少参数与计算量，增强浅层特征并与高层语义融合，提高了特征图输出准确性。吴洛冰等[20]提出了一种基于多尺度特征提取的旋转遥感目标检测算法，通过结合空洞卷积设计感受野扩展模块，并嵌入自适应特征融合结构，提升了模型对复杂环境中多尺度目标的检测能力。Jiang等[21]提出了一种基于深度学习的最优深度神经网络模型，将图像中的物体特征与原始数据的构建相结合，有效地改善了小尺寸物体的检测性能。Teng等[22]提出了一种基于全局到局部的目标检测网络模型，通过多尺度感

收稿日期：2023-10-20；**修回日期**：2023-11-15；**录用日期**：2023-12-15；**网络首发日期**：2023-12-23

基金项目：国家自然科学基金(62071240，62106111)、江苏省一流本科课程项目(2021YLKC005)、江苏省产教融合型一流课程项目(2022-133)

通信作者：*yorkzhang@nuist.edu.cn

知模块提取的全局上下文线索，来消除复杂背景的影响，并设计自适应锚模块缓解语义尺度差异。Zhao等[23]设计了一种轻量级的注意力网络，利用多尺度特征转换和标签注册方法提高模型学习能力，进而提高对光学遥感图像中小目标的检测精度。胡杰等[24]提出了一种高性能的基于深度语义和位置信息融合的双阶段三维目标检测算法，在俯视图中提取目标深层次纹理和语义特征的同时，增强了网络的自适应特征提取能力及中心点的聚合能力。王思启等[25]提出了一种基于MVSNet深度学习网络实现空间目标三维重建的方法，通过使用多尺度卷积提取深度特征，编码解码结构融合上下文信息进行立体匹配，残差网络解决边界平滑问题，有效提升了卫星图像重建效果。以上几种方法都有不同的改进，但是对于小目标的检测仍然有局限性。首先，以上方法没有考虑到感受野随深度的增加而呈缓慢线性增长，有限感受野不能与其特征尺度匹配，从而难以实现对小目标特征的有效提取。其次，特征金字塔融合解决了尺度差异的检测问题，但高层语义信息和低层空间信息融合仍有改进空间。

本文提出了一种增强小目标特征的多尺度神经网络（ESF-MNet），针对复杂背景下遥感小目标检测的创新有：1）在骨干网络中构建高效层注意力聚合结构作为其主要特征提取模块，更好地提取各种类别的目标特征；2）加入感受野增强模块（RFE）增强特征图的感受野，提高多尺度目标检测和识别的精度；3）使用GSConv构成Neck层，减少网络层参数量，采用一次聚合的方法设计跨阶段部分网络（GSCSP）模块VoV-GSCSP，保持网络的特征提取能力，同时使用CARAFE上采样模块获取丰富语义信息，改善上采样语义信息丢失的问题；4）在检测头结构中使用下采样率分别为4、8和16倍的特征输出作为输入，有效提高了对小目标的检测能力。与现有的检测算法相比，本文提出的ESF-MNet模型是一种更优越的检测算法，可以实现不同环境条件下对小目标的精准检测。

2　网络结构

ESF-MNet模型框架如图1所示，该模型主要由主干网络（Backbone）、颈部（Neck）和头部（Head）三部分组成，其主干网络由若干CBH（Conv+BN+MISH）层、高效层注意力聚合模块和若干MPConv层组成。高效层注意力聚合模块采用多残差级联结构，在提高检测精度的同时，能更好地捕捉到小目标的细节信息。MPConv通过Maxpool和CBH降低图像尺寸和通道数，并通过拼接操作融合特征，以增强特征提取能力。在特征提取阶段，通过CBH层获得基础的特征信息，然后通过四个高效层注意力聚合模块提升特征表达能力，每模块间经过MPConv层实现特征图的下采样，最后输出四个尺度的特征图（C2/C3/C4/

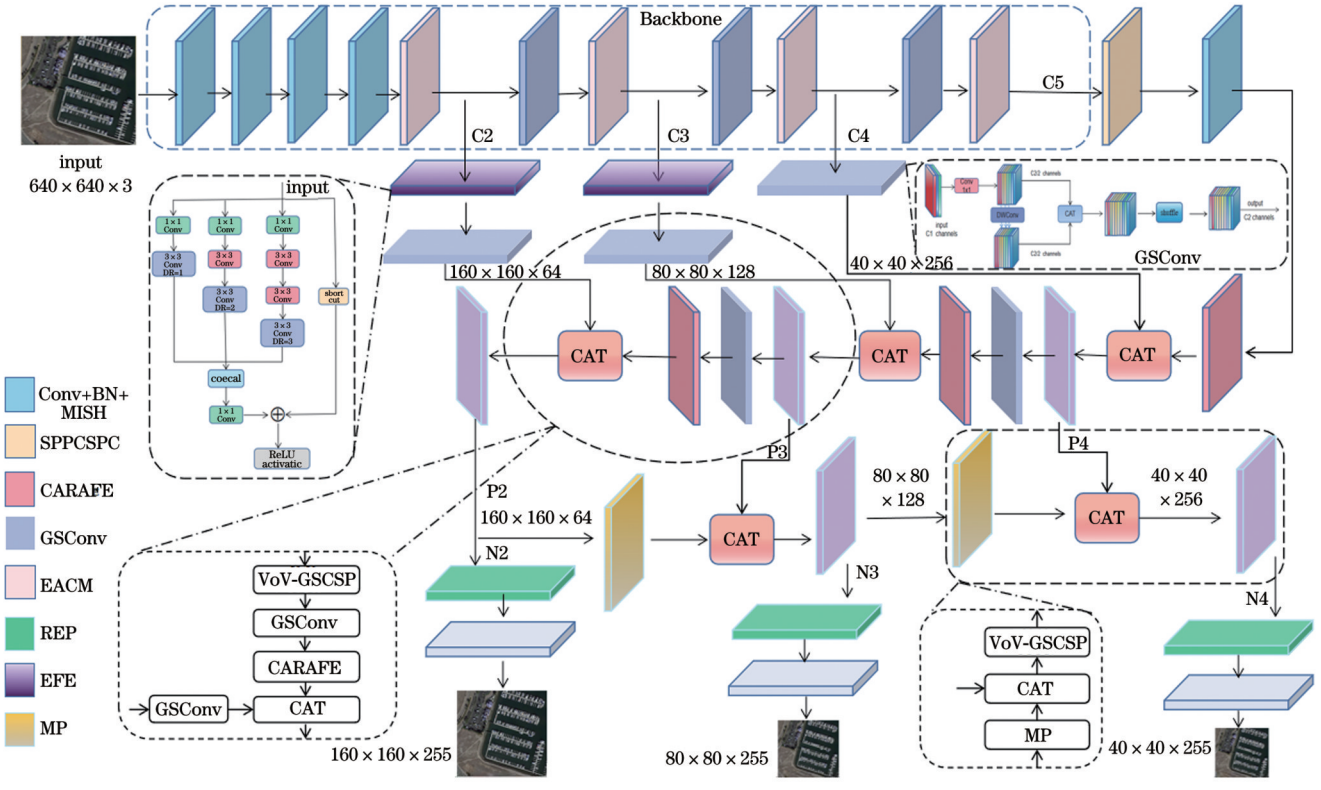

图1　ESF-MNet模型框架图

Fig. 1　Framework diagram of ESF-MNet model

C5)。Neck层是由SPPCSPC层、若干GSConv层、若干MPConv层、CARAFE模块组成。首先对主干输出的两层浅特征输出(C2/C3)进行感受野增强操作,再对输出的四个尺度的特征图(C2/C3/C4/C5)进行通道调整,然后自顶向下地使用CARAFE模块进行上采样和融合操作,生成三个尺度的融合特征图(P2/P3/P4)。接着,对这三个尺度的融合特征图自底向上地利用MPConv进行下采样和融合操作,得到三个尺度的融合特征图(N2/N3/N4)。最后将这三个尺度的融合特征图输入检测头,经过RepConv和Conv层后,得到了三层不同尺寸的特征图,最终用于生成预测结果。

2.1　EACM网络结构

高效层聚合网络(ELAN)具有特征提取、增强泛化能力的作用。在遥感图像中,小目标容易与其他地物混淆,尤其是外观相似的情况下,识别变得更加困难。因此在Backbone的ELAN模块中引入一种坐标注意力(CA)模块[26],这种结构既能兼顾到空间和通道信息,又能兼顾到长距离相关性,在轻量化的同时,还能提高精度,其结构如图2所示。其中,Avgpool为平均池化,F1为共享卷积核,F_h和F_w分别为两个方向的特征图。

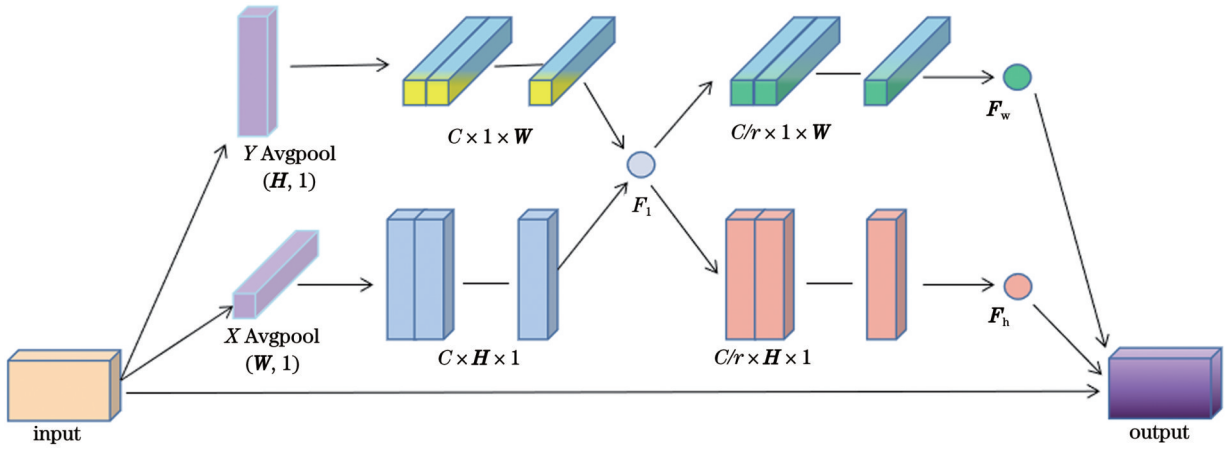

图2　CA注意力机制结构图

Fig. 2　Structure diagram of CA attention mechanism

高效层注意力聚合模块(EACM)结构如图3所示,EACM是由多个CBH和CA模块组成的多残差级联层。其中,CBH使用卷积、归一化和Mish激活函数,可以更好地保留梯度信息,提高模型的学习能力。通过对两条分支中不同层级的输出进行特征融合,充分挖掘多尺度特征的信息,可以更好地捕捉到小目标的细节信息,提高目标检测的精度。同时,引入的CA模块能够更好地增强目标特征表达,减少背景干扰,加速模型的训练和推理。

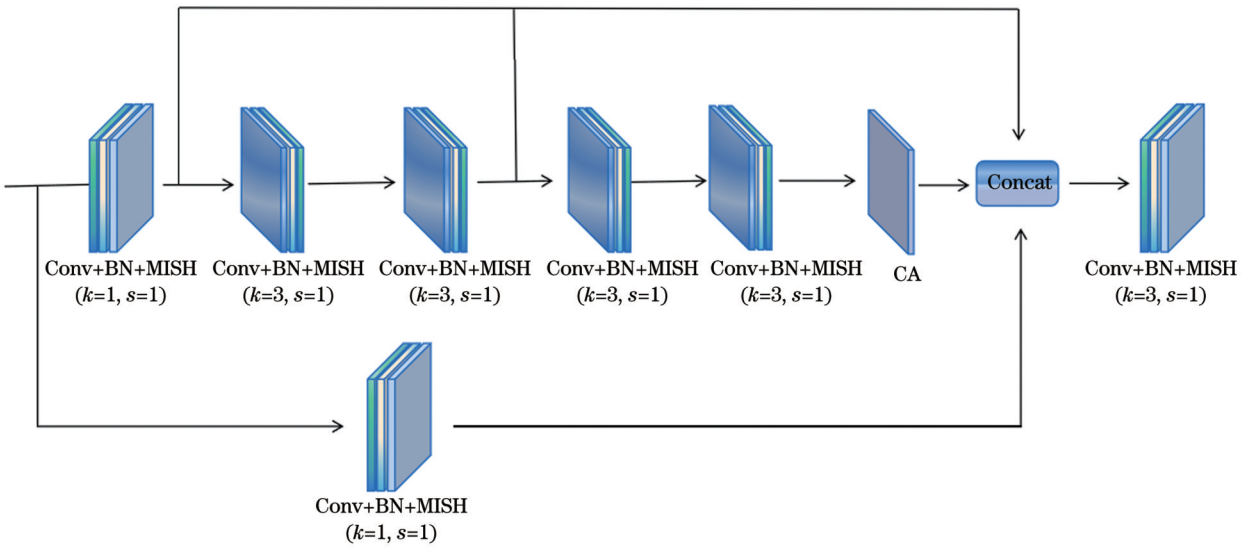

图3　EACM模块示意图

Fig. 3　Schematic diagram of EACM module

2.2 感受野增强模块

在使用特征提取网络中，为了应对光学遥感图像分辨率较大且背景复杂的问题，需要通过增强特征图的感受野来提高准确性。为此，使用RFENet模块[27]对本文模型中的两层浅层特征图的感受野进行处理。该模块由三个分支和一个旁路残差连接组成，每一分支中都有不同尺寸的卷积核及不同的空洞卷积。该结构扩大了卷积层的感受野，并融合了不同深度的特征图，提升了网络的信息表达能力。每个分支通过卷积操作获取不同层次的语义信息，然后级联输出特征图，以更好地融合深层语义信息和空间信息来增强语义特征和对小目标的检测能力。此外，模块利用空洞卷积在不降低特征图分辨率的情况下增大感受野，通过使用不同的空洞率(1、2、3)来获得不同尺度的上下文信息，并提高了对特征的感知和撷取能力。因此，该模块可以帮助网络更好地响应不同尺度的遥感目标。

2.3 GSConv

本文将常规卷积与深度可分离卷积(DW)相结合，构建一个轻量化网络，在保证网络准确率的前提下，提升学习与推理速度。传统的标准卷积方法虽然能有效地提取出图像中的各种特征，但其计算效率较低且耗时较长。因此，一些网络模型如Xception[28]、MobileNet等[29]使用了深度可分离卷积模块，极大地提高了检测速度，但却牺牲了模型的检测精度。为了解决这个问题，本文引入了一种名为GSConv的卷积操作[30]，它结合了普通卷积、深度可分离卷积和Shuffle混合策略。当输入通道数为C1的特征图进入网络中时，首先，采用1×1逐点卷积(PW)方法，将特征图中的通道数缩减至输出通道数的1/2；然后，利用深度可分离卷积法对图象进行处理；最后，将两个特征图结合起来，得到一个具有C2通道数的特征图。用Shuffle操作将数据的通道混排，将逐点卷积产生的信息渗透到深度可分离卷积生成的信息的每个部分中，增加随机性，提高网络泛化能力。本文以GSConv为基础，进一步设计了GSbottleneck模块，其结构如图4(a)所示，同样，采用一次聚合的方法设计跨阶段部分网络(GSCSP)模块VoV-GSCSP。其网络结构如图4(b)所示，在Neck端使用GSConv及VoV-GSCSP模块，既可以保证网络的精度，又可以减少网络结构的复杂程度和计算成本。

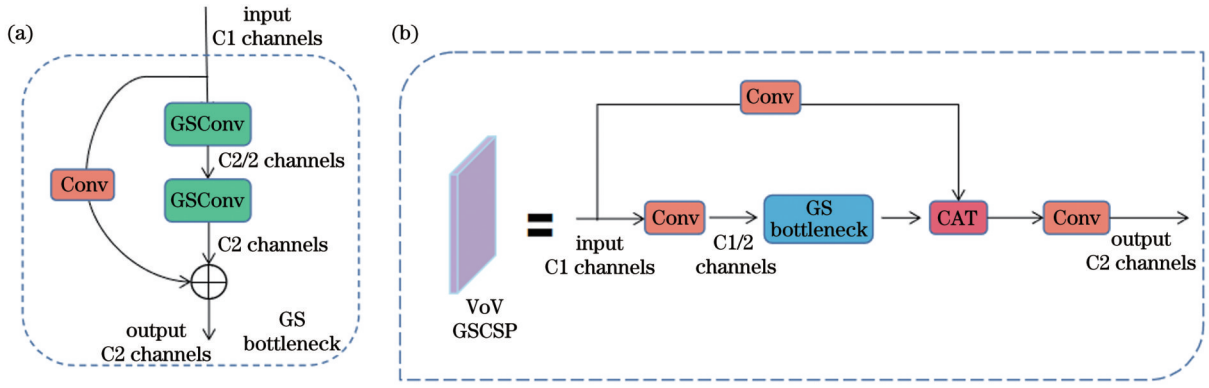

图4 GSbottleneck模块和VoV-GSCSP网络结构图。(a) GSbottleneck；(b) VoV-GSCSP

Fig. 4 Structures of Gsbottleneck module and VoV-GSCSP network. (a) Gsbottleneck; (b) VoV-GSCSP

2.4 CARAFE模块

双线性插值限制了图像质量和细节捕捉能力，特别是边缘区域的处理。而CARAFE[31]可以更好地利用上下文信息和高分辨率特征图，生成更准确且丰富的上采样结果，提升网络的语义提取能力。CARAFE分为两个主要模块，分别是上采样核预测模块和特征重组模块，其结构如图5所示，图中，C、H、W为特征图尺寸参数，σ为倍率，K_{up}为重组内核尺寸。具体步骤如下：

1) 利用1×1卷积，上采样核预测模块将原始C个通道压缩到C_m；

2) 使用一个卷积核大小为$K_{encoder}\times K_{encoder}$的卷积层作为内容编码器，用于预测上采样核，输入通道数为C_m，输出通道数为$\sigma\times\sigma\times K_{up}\times K_{up}$，并将通道维度展开到空间维度，生成尺寸为$\sigma H\times\sigma W\times K_{up}\times K_{up}$的上采样核，其中$K_{encoder}$为编码器参数；

3) 对步骤2)中所获得的上取样核进行归一化，使其权值和等于1；

4)使用上采样核对输入特征图中的$K_{up}\times K_{up}$区域进行点积操作，生成重组后的特征。

2.5 检测层

由于遥感数据集中小目标数量较多，因此本文对头部输入网络进行了如下构造：将下采样率4倍的浅特征输出作为头部网络的输入特征(即N2)，并与8倍和16倍输出特征进行联合融合。颈部网络将PANet的输出层N2、N3和N4作为头部分支的输入，以提高对小目标的检测精度。本文模型检测尺度为40×40×255、80×80×255、160×160×255。其结构如图6所示。

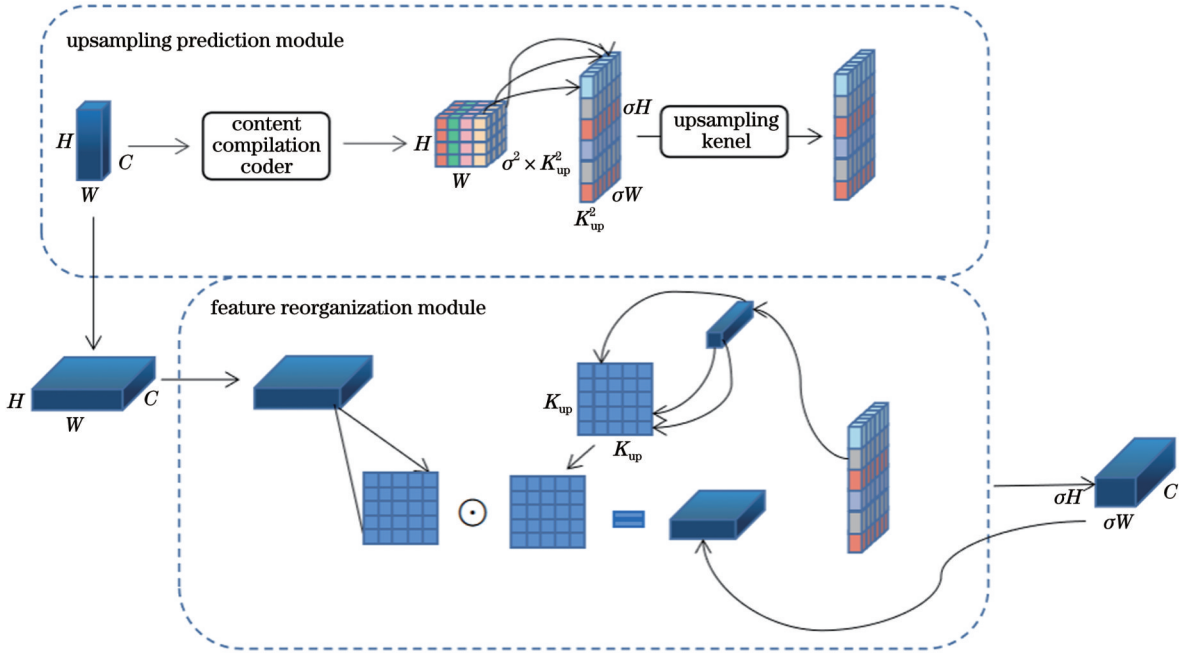

图 5 CARAFE结构图

Fig. 5 Structure diagram of CARAFE

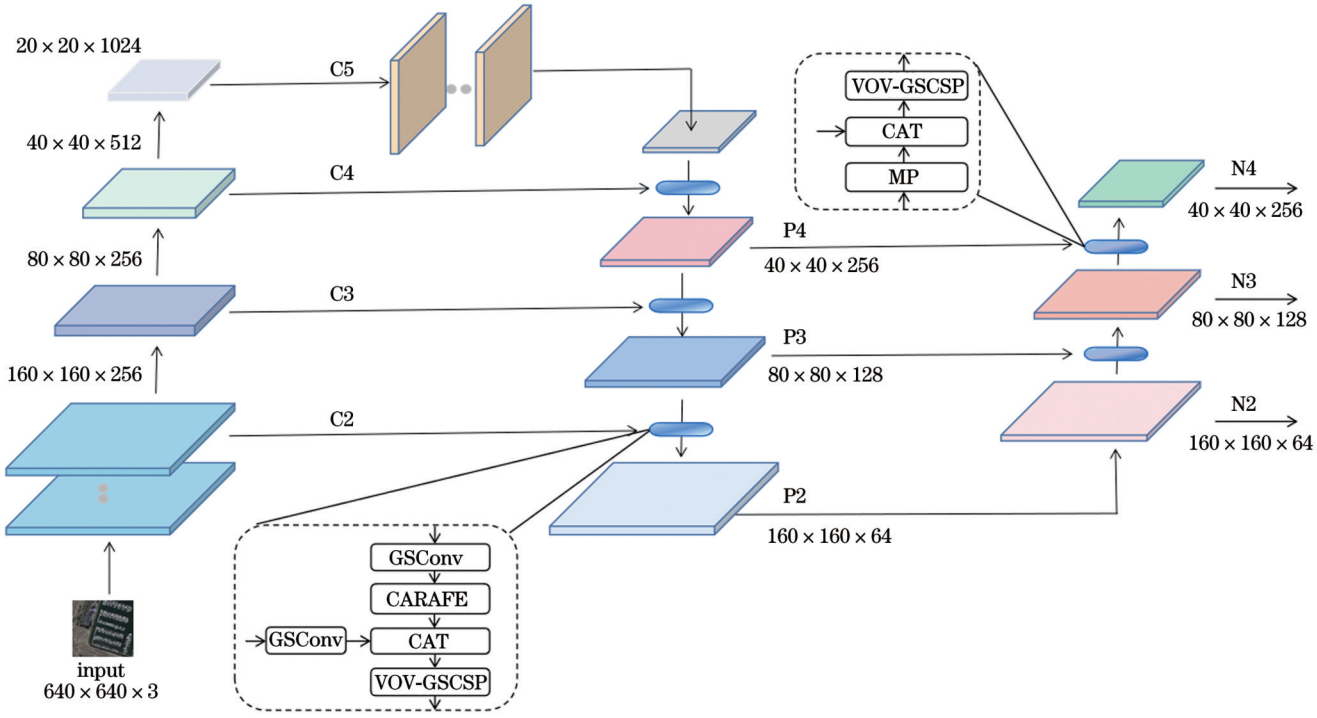

图 6 检测头结构

Fig. 6 Structure of detection head

3 实验结果与分析

3.1 实验数据集

在遥感数据集中，小目标是指在图像上占据较小区域、尺寸相对较小的物体。这些目标包括飞机、车辆、船舰等。本文实验采用数据集DOTA和NWPU VHR-10来训练和评估模型，DOTA数据集[32]包含2806张800 pixel×800 pixel~4000 pixel×4000 pixel不等的遥感图像，共计15个类别，分别是港口(HA)、飞机(PL)、轮船(SH)、环形交叉路口(RA)、储罐(ST)、游泳池(SW)、小型车辆(SV)、网球场(TC)、桥梁(BD)、篮球场(BC)、直升机(HC)、田径场(GTF)、大型车辆(LV)、足球场(SBF)和棒球场(BF)。NWPU VHR-10数据集[33]由西北工业大学发布，其拥

有650张包含目标的图像和150张无目标的背景图像，共计800张。图像中总共有3896个目标，目标种类包括飞机(PL)、船舶(SH)、储罐(ST)、棒球场(BF)、网球场(TC)、篮球场(BC)、田径场(GTF)、港口(HA)、桥梁(BD)和车辆(VH)，共计10个类别。

3.2 评价指标

光学遥感图像目标检测的精度综合考虑了预测结果的定位精度和分类精度。为了衡量模型对目标检测任务的准确性，文中主要采用了平均精确率(mAP，公式中用m_{AP}表示)指标。mAP用于评估目标检测模型在类别识别和位置定位方面的准确性，由精度(P)和召回率(R)共同计算得出。其公式为

$$P_{recision}=S_{TP}/(S_{TP}+S_{FP}), \tag{1}$$

$$R_{ecall}=S_{TP}/(S_{TP}+S_{FN}), \tag{2}$$

式中：S_{TP}为真正例(true positive)，表示预测值为1，真实值为1；S_{FP}为假正例(false positive)，表示预测值为1，真实值为0；S_{FN}为假反例(false negative)，表示预测值为0，真实值为1。P和R是一对互相影响的值，P值较高时R值就会偏低，而R值较高时P值往往偏低。为了综合考虑精度和召回率对检测模型的评估，引入平均精度(AP，公式中用A_P表示)：

$$A_P=\int_0^1 P(r)\mathrm{d}r。 \tag{3}$$

对所得到的各对象的AP值进行相加，并对其进行平均，从而得到分类平均精度，计算方法如下式所示：

$$m_{AP}=\frac{1}{k}\sum_{i=1}^{k}A_{Pi}。 \tag{4}$$

为验证ESF-MNeT模型对遥感小目标检测的可行性，除了通过mAP、P和R作为评价指标以外，还统算了参数量(Param.)、计算量(FLOPs)作为评估指标，通过对照实验、消融实验等方法来验证该方法的适用性。

3.3 实验环境及配置

本文实验基于Windows 10操作系统，深度学习环境搭载于CUDA11.7及Pytorch 2.0.1框架，使用NVIDIA GeForce RTX4080 GPU加速模型进行训练。训练配置参数设置如下：训练周期(epoch)为300，批次大小(batch_size)为12，图像尺寸为640×640。

3.4 消融实验

为了验证本文所构建网络模型的有效性，分别在DOTA数据集和NWPU NHR-10数据集上进行了消融实验，并以YOLOv7作为基准模型，作为后面六组对比标准，实验结果如表1所示。从消融实验的结果可以看出，在Backbone中构建EACM网络作为主干提取模块后，mAP分别提高了1.3%和3.3%，说明该模块可以有效增强网络进行特征提取，从而提高网络的准确率。通过增加RFE后，mAP分别提高了0.8%和2.6%，增强特征图的感受野使得遥感图像中的复杂语义得到充分提取。此外，使用GSConv构成的VoV-GSCSP模块相较于Baseline参数量减少了14.9%，FLOPs降低了大约13.8%，mAP分别下降了0.2%和0.3%，说明VoV-GSCSP模块在保持精准度的同时，能有效降低网络结构的复杂性和计算成本。此外，通过加入CARAFE上采样算子后，mAP在两个数据集上分别提高了0.2%和0.9%，说明CARAFE算子能够更好地捕获特征之间的空间关系，提升识别精度。此外，使用下采样率分别为4、8和16倍的特征输出作为头部网络输入，使得mAP在两个数据集上分别提高了0.7%和1.2%，说明构建的检测头更适合小目标检测，减少了随着网络深度增加而丢失的小目标信息。

表1 消融实验结果

Table 1 Ablation experimental results

Method	DOTA			NWPU NHR-10			Param. /M	FLOPs /G
	P /%	R /%	mAP0.5 /%	P /%	R /%	mAP0.5 /%		
Baseline	78.3	71.2	74.9	93.9	84.7	89.8	37.2	105.4
+EACM	80.1	71.8	76.2	97.2	87.2	93.1	37.3	105.5
+RFE	79.7	71.8	75.7	95.7	87.3	92.4	38.5	105.4
+VoV-GSCSP	77.1	71.4	74.7	92.6	84.9	89.5	31.9	91.6
+CARAFE	78.7	71.3	75.1	93.6	87.4	90.7	37.3	105.5
+HEAD	79.3	70.8	75.6	92.6	85.3	91.0	28.5	124.2
+ALL	80.8	74.3	78.6	95.8	89.9	94.3	26.2	94.7

为了进一步验证ESF-MNet模型中各模块的有效性，本文通过替换不同模块的网络模型进行可视化效果图对比，如图7所示，通过VGCE(VoV-GSCSP+CARAFE)和基准模型的可视化效果图可以看出，VGCE提取的特征信息优于基础网络，在降低网络结构复杂性和计算量的同时，保证了算法的准确性，并在CARAFE模型的加持下，使得更多的特征信息被提取出来。此外，通过RFE和VGCE的可视化效果图对比可以看出，RFE提取的特征信息不亚于VGCE，但与VGCE不同的是，RFE通过增强特征图的感受野，帮

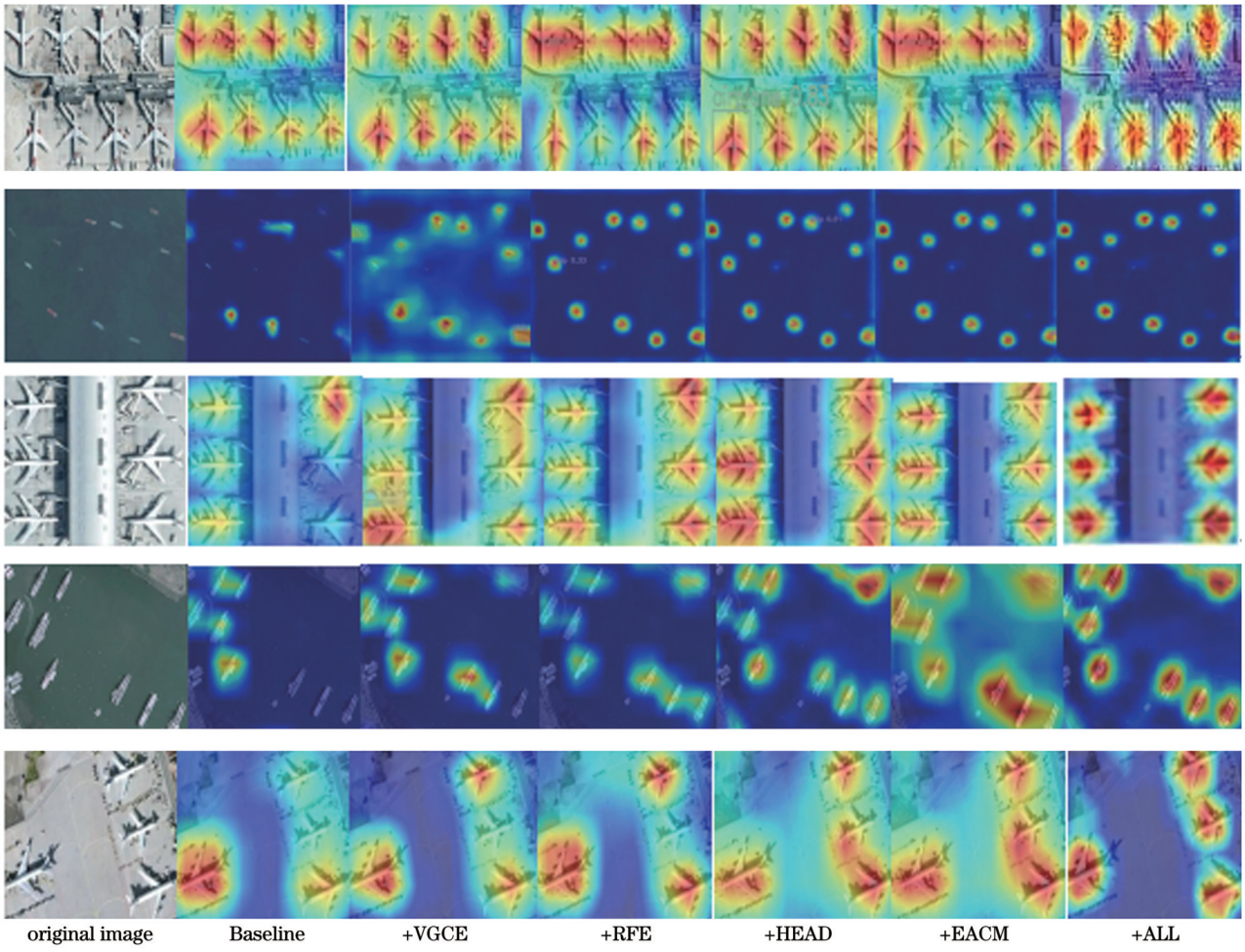

图 7　各模块可视化效果图对比

Fig. 7　Comparison of visual renderings of each module

助网络更好地响应不同尺度的遥感目标。此外，通过HEAD、EACM与RFE的可视化效果图对比可以看出，HEAD和EACM可以提取更多的目标邻域特征信息，从而获取更丰富的特征图。最后，对比ESF-MNet和其他模块的可视化效果图发现，ESF-MNet能够有效降低背景干扰，使其提取的特征信息能够更好地集中于目标周围，这也说明ESF-MNet的特征提取和分析能力相较于其他模块的网络模型有了较大的提升。

3.5　对比实验

为了进一步展示本文算法在小目标检测模型方面的优势，将使用相同的训练配置参数进行实验。同时，还会对比其他主流算法，包括FMSSD[34]、Faster-RCNN、YOLOv5s、YOLOv7和YOLOv8s，并评估它们在整体和局部上的性能表现。因此，可以更加准确地评估各个算法之间的差异，以及本文算法的优势所在。表2的实验结果显示，相比经典算法，本文算法在平均检测精度上表现最好。该算法在DOTA数据集上的平均检测精度达到了78.6%。相比于经典算法，本文算法在9类目标检测中表现更出色。此外，针对小目标检测方面，通常对飞机、直升机、舰船和大、小型车辆这五类目标进行评估。实验结果表明，本文算法在这五类目标的平均检测精度(mAP)达到了83.6%。与FMSSD、文献[22]以及文献[23]的算法相比，本文算法的小目标(SG)mAP分别高出了9.8百分点、10.9百分点和5.3百分点。因此，可见本文算法在小目标检测方面取得了极为显著的效果。然而，其在一些类别中的检测精度并未达到最优，尤其在检测一些大型目标地物(例如田径场、篮球场)时，精度还没有达到最佳水平，存在的主要原因是模型轻量化后的网络深度较浅，下采样倍数较小。然而，如果增加网络的深度和下采样倍数，虽然可以提高大型目标的检测效果，但会导致对小型目标检测效果较差。因此，本文的研究重点是在确保对大目标具有更高的检测准确率的前提下，提升中小目标的检测准确率。

在NWPU NHR-10数据集上，将其中60%的图像作为训练集，20%的图像作为验证集，剩下20%作为测试集进行实验。结果如表3所示，本文模型在10类目标的平均检测精度达到了94.3%，相比经典算法，本

表 2 不同算法在DOTA数据集上的目标检测结果对比

Table 2 Comparison of target detection results of different algorithms on DOTA dataset unit: %

Class	Model							
	FMSSD	Faster-RCNN	YOLOv5s	Ref. [22]	Ref. [23]	YOLOv7	YOLOv8s	ESF-MNet
SV	69.2	48.6	66.9	71.0	66.9	74.7	74.7	**78.1**
LV	73.6	73.4	85.6	70.8	85.6	88.7	88.3	**91.5**
PL	89.1	82.8	92.0	88.3	92.0	94.0	94.2	**95.5**
ST	73.3	52.5	70.9	85.2	70.9	78.6	77.1	**79.8**
SH	76.9	77.6	87.5	76.5	87.5	89.7	89.9	**92.3**
HA	72.4	62.6	85.1	64.5	85.1	85.8	86.5	**88.6**
GTF	67.9	**78.2**	67.9	64.7	67.9	71.8	70.4	76.2
SBF	52.7	66.6	53.0	53.4	53.0	71.8	70.5	**72.5**
TC	90.7	62.7	94.1	89.4	94.1	95.3	95.1	**96.2**
SP	**80.6**	57.7	62.7	67.0	62.7	63.9	63.7	71.0
BF	81.5	**86.4**	78.4	79.5	78.4	77.1	78.8	82.8
RA	**67.5**	62.7	63.0	66.0	63.0	56.6	60.2	62.9
BC	82.7	69.4	64.6	**83.5**	64.6	70.7	71.3	76.2
BD	48.2	48.9	49.6	44.2	49.6	49.5	50.8	**56.3**
HC	**60.2**	42.8	59.7	57.1	59.7	55.4	53.6	59.8
mAP	72.4	64.9	72.1	70.8	72.1	74.9	75.0	**78.6**
SG mAP	73.8	65.0	78.3	72.7	78.3	80.5	80.1	**83.6**

表 3 不同算法在NWPU NHR-10数据集上的目标检测结果对比

Table 3 Comparison of target detection results of different algorithms on NWPU NHR-10 dataset unit: %

Class	Model							
	FMSSD	Faster-RCNN	YOLOv5s	Ref. [21]	Ref. [23]	YOLOv7	YOLOv8s	ESF-MNet
PL	**99.7**	94.6	99.4	93.0	**99.7**	99.5	99.6	**99.7**
SH	89.9	82.3	94.9	84.5	91.0	95.1	95.0	**95.9**
ST	90.3	65.3	**99.3**	87.1	91.6	94.2	95.2	90.5
BF	98.2	95.5	98.4	92.8	90.8	98.5	98.8	**99.0**
TC	86.0	81.9	94.1	82.0	90.7	99.4	97.7	**99.5**
BC	**96.8**	89.7	76.8	89.0	98.7	86.8	87.0	86.9
GTF	**99.6**	92.4	95.8	78.0	96.5	99.5	99.4	**99.6**
HA	75.6	72.4	83.0	76.0	90.4	91.9	91.8	**96.1**
BD	80.1	57.5	60.8	81.0	**90.9**	61.1	73.4	78.4
VH	88.2	77.8	58.8	84.5	**97.9**	71.5	66.3	97.3
mAP	90.4	80.9	86.1	84.8	93.8	89.8	90.4	**94.3**
SG mAP	92.6	84.9	84.4	87.3	96.2	88.7	86.7	**97.6**

文算法在6类目标检测中表现更出色。此外，针对小目标检测方面，通常对飞机、舰船和车辆这三类目标进行评估。实验结果表明，本文算法在这三类目标的平均检测精度(mAP)达到了97.6%。与FMSSD、文献[21]以及文献[23]的算法相比，本文算法的小目标mAP分别高出了5.0百分点、10.3百分点和1.4百分点。因此，实验结果验证了本文模型对小目标检测的有效性。

3.6 检测结果分析

由图8中目标检测结果对比可知，从各类别目标检测精度来看，本文所提出的ESF-MNet模型与YOLOv5s、YOLOv7和YOLOv8s算法相比，能够更好地检测到小目标，同时又能保证较高的精度，在提高小目标检测精度上有很大的优势，并且在复杂背景下，小目标检测具有显著的优越性，准确性更高。

图8 目标检测效果对比图

Fig. 8 Comparison of target detection effects

由图9的可视化结果对比可知，YOLOv7和YOLOv8s算法相比于YOLOv5s，提取到更多的目标邻域特征信息，获取到更丰富的特征图。然而，YOLOv7和YOLOv8s的效果图内，颜色较深的地方却是目标周围，并非目标本身。因此，对于目标的提取仍有所欠缺。与其他三种算法对比发现，本文提出的ESF-MNet模型可以更清楚地标识出对应的改变区域的特性信息，这说明本文模型可以从原来的语义不明显的物体中取出更多的特征，并使其提取的特征信息能够更好地集中于目标周围。

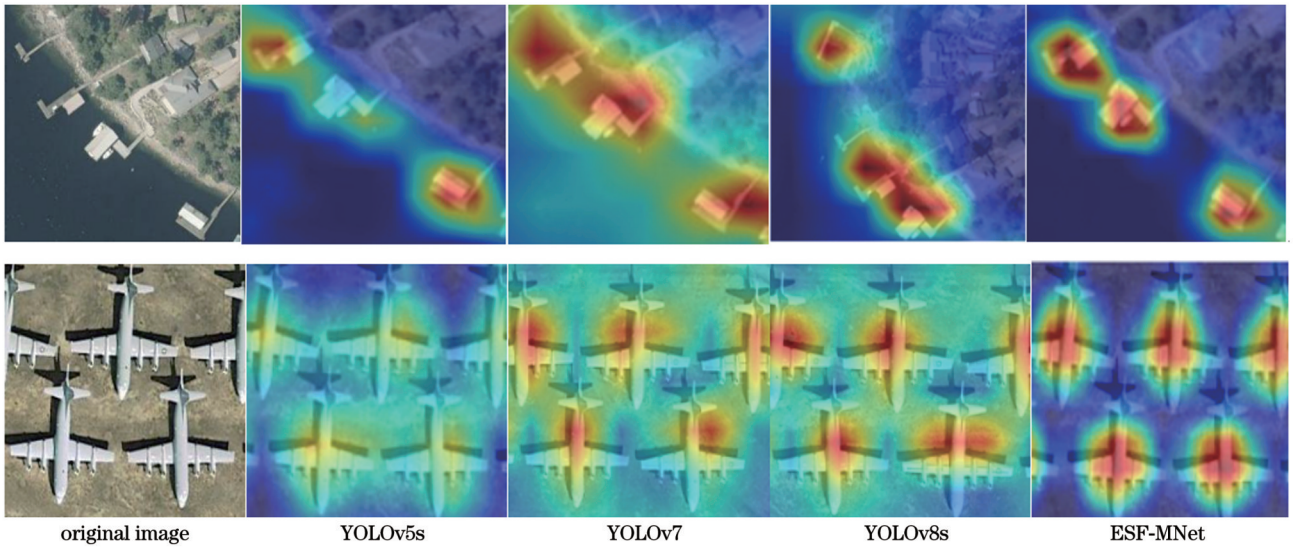

图9 可视化效果对比图

Fig. 9 Comparison of visual effects

通过图10中漏检目标检测结果的对比，可以发现本文模型相较于YOLOv5s、YOLOv7和YOLOv8s算法，在解决前景和背景类别不平衡导致的漏检问题上表现出了显著改进。此外，该算法对物体的定位、识

别、分类等也具有显著的优越性。

通过图 11 中误检目标检测结果的对比，可以看出尽管本文模型在提高检测精度方面取得了显著进展，但也不可避免地出现了一些轻微的误检情况。然而，与 YOLOv5s、YOLOv7 和 YOLOv8s 算法相比，本文模型表现出更低的误检率，表明在提取小目标的几何特征和感知能力方面，本文算法比 YOLOv5s、YOLOv7 和 YOLOv8s 更具优势。

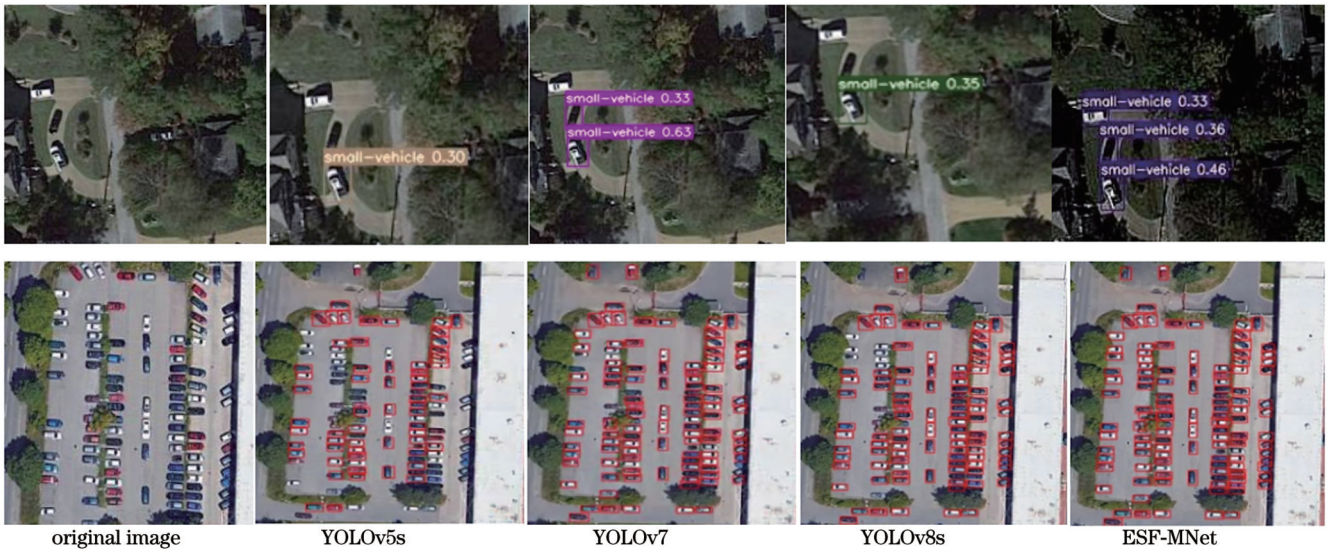

图 10　小目标检测漏检效果对比图

Fig. 10　Comparison of missed detection effects of small target detection

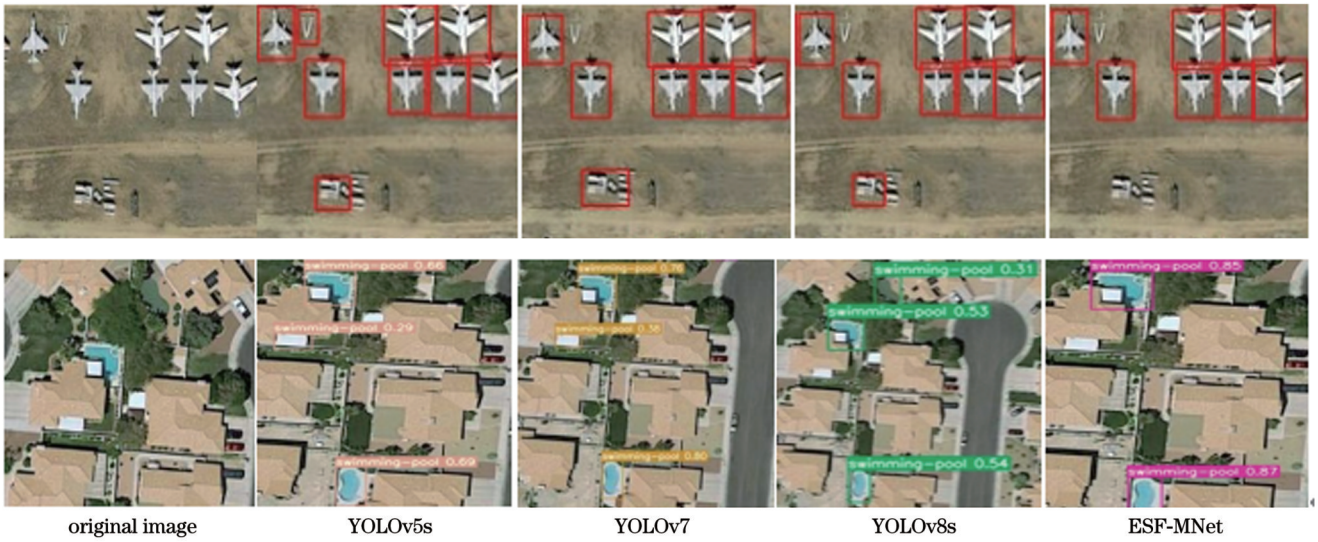

图 11　小目标检测误检效果对比图

Fig. 11　Comparison of false detection effects of small target detection

4　结　　论

对光学遥感图像中的目标进行探测和识别，在军用和民用领域都有着非常重要的意义。但是，在复杂的背景、密集的小目标以及缺乏特征信息的情况下，小目标的识别非常困难。本文通过在主干网络中构建一种高效层注意力聚合模块，来提取各种类别的目标特征，并使用感受野增强模块对不同深度的特征图进行融合，以提高网络的信息表达能力；此外，通过使用 GSConv 及 CARAFE 模块构成 Neck 层，采用通道数减半的压缩方法，对颈部进行精细处理，采用一次聚合的方法设计跨阶段部分网络模块 VoV-GSCSP 模块，能有效降低网络运算量，提升检测速度，并在加入 CARAFE 模块下提高了检测准确度；此外，通过在检测头结构中使用下采样率分别为 4、8 和 16 倍的特征输出层来构造多尺度网络，有效提高了对小目标的检测能力。实验结果表明，该模型对于复杂背景中的小目标检测的实时性好、鲁棒性强。虽然本文模型在漏检和误检方面有了很大的改进，但仍然可能存在一些遗漏和错误检测的情况。另外，尽管当前遥感图像中的物

体检测方法已日趋成熟，但是，由于CNN的计算量巨大，再加上遥感图像本身的复杂性，因而寻求一种既准确又高效的目标检测方法仍存在许多难点。在未来工作中，将继续针对上述问题及难点展开进一步的研究。

参 考 文 献

[1] Cheng G, Xie X X, Han J W, et al. Remote sensing image scene classification meets deep learning: challenges, methods, benchmarks, and opportunities[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 13: 3735-3756.

[2] 王浩雪, 曹杰, 邱诚, 等. 基于改进YOLOv4的航拍图像多目标检测方法[J]. 电光与控制, 2022, 29(5): 23-27.
Wang H X, Cao J, Qiu C, et al. Multi-target detection method of aerial images based on improved YOLOv4 algorithm[J]. Electronics Optics & Control, 2022, 29(5): 23-27.

[3] Cho S, Shin W, Kim N, et al. Priority determination to apply artificial intelligence technology in military intelligence areas[J]. Electronics, 2020, 9(12): 2187.

[4] Fukuda G, Hatta D, Guo X L, et al. Performance evaluation of IMU and DVL integration in marine navigation[J]. Sensors, 2021, 21(4): 1056.

[5] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE Press, 2014: 580-587.

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

[7] 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.

[8] He K M, Gkioxari G, Dollár P, et al. Mask R-CNN[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2980-2988.

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

[10] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 318-327.

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

[12] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6517-6525.

[13] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: optimal speed and accuracy of object detection[EB/OL]. (2020-04-23)[2023-11-09]. http://arxiv.org/abs/2004.10934.

[14] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[EB/OL]. (2022-07-06)[2023-11-09]. http://arxiv.org/abs/2207.02696.

[15] Qu J S, Su C, Zhang Z W, et al. Dilated convolution and feature fusion SSD network for small object detection in remote sensing images[J]. IEEE Access, 2020, 8: 82832-82843.

[16] 闫钧华, 张琨, 施天俊, 等. 融合多层级特征的遥感图像地面弱小目标检测[J]. 仪器仪表学报, 2022, 43(3): 221-229.
Yan J H, Zhang K, Shi T J, et al. Multi-level feature fusion based dim small ground target detection in remote sensing images [J]. Chinese Journal of Scientific Instrument, 2022, 43(3): 221-229.

[17] 张寅, 朱桂熠, 施天俊, 等. 基于特征融合与注意力的遥感图像小目标检测[J]. 光学学报, 2022, 42(24): 2415001.
Zhang Y, Zhu G Y, Shi T J, et al. Small object detection in remote sensing images based on feature fusion and attention[J]. Acta Optica Sinica, 2022, 42(24): 2415001.

[18] 张廓, 陈章进, 乔栋, 等. 基于感受野和特征增强的遥感图像实时检测[J]. 激光与光电子学进展, 2023, 60(2): 0228001.
Zhang K, Chen Z J, Qiao D, et al. Real-time image detection via remote sensing based on receptive field and feature enhancement[J]. Laser & Optoelectronics Progress, 2023, 60 (2): 0228001.

[19] 薛俊达, 朱家佳, 张静, 等. 基于FFC-SSD模型的光学遥感图像目标检测[J]. 光学学报, 2022, 42(12): 1210002.
Xue J D, Zhu J J, Zhang J, et al. Object detection in optical remote sensing images based on FFC-SSD model[J]. Acta Optica Sinica, 2022, 42(12): 1210002.

[20] 吴洛冰, 谷玉海, 吴文昊, 等. 基于多尺度特征提取的遥感旋转目标检测[J]. 激光与光电子学进展, 2023, 60(12): 1228010.
Wu L B, Gu Y H, Wu W H, et al. Remote sensing rotating object detection based on multi-scale feature extraction[J]. Laser & Optoelectronics Progress, 2023, 60(12): 1228010.

[21] Jiang S L, Yao W, Wong M S, et al. An optimized deep neural network detecting small and narrow rectangular objects in google earth images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 13: 1068-1081.

[22] Teng Z, Duan Y N, Liu Y, et al. Global to local: clip-LSTM-based object detection from remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 5603113.

[23] Zhao B Y, Wang Q, Wu Y F, et al. Target detection model distillation using feature transition and label registration for remote sensing imagery[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15: 5416-5426.

[24] 胡杰, 安永鹏, 徐文才, 等. 基于激光点云的深度语义和位置信息融合的三维目标检测[J]. 中国激光, 2023, 50(10): 1010003.
Hu J, An Y P, Xu W C, et al. 3D object detection based on deep semantics and position information fusion of laser point cloud[J]. Chinese Journal of Lasers, 2023, 50(10): 1010003.

[25] 王思启, 张家强, 李丽圆, 等. MVSNet在空间目标三维重建中的应用[J]. 中国激光, 2022, 49(23): 2310003.
Wang S Q, Zhang J Q, Li L Y, et al. Application of MVSNet in 3D reconstruction of space objects[J]. Chinese Journal of Lasers, 2022, 49(23): 2310003.

[26] Hou Q B, Zhou D Q, Feng J S. Coordinate attention for efficient mobile network design[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 13708-13717.

[27] Liu S T, Huang D, Wang Y H. Receptive field block net for accurate and fast object detection[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11215: 404-419.

[28] Chollet F. Xception: deep learning with depthwise separable convolutions[C]//2017 IEEE Conference on Computer Vision

and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 1800-1807.

[29] Howard A, Sandler M, Chen B, et al. Searching for MobileNetV3[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Republic of Korea. New York: IEEE Press, 2019: 1314-1324.

[30] Li H, Li J, Wei H, et al. Slim-Neck by GSConv: a better design paradigm of detector architectures for autonomous vehicles[EB/OL]. (2022-06-06)[2023-11-09]. https://arxiv.org/abs/2206.02424.

[31] Wang J Q, Chen K, Xu R, et al. CARAFE: content-aware reassembly of features[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Republic of Korea. New York: IEEE Press, 2019: 3007-3016.

[32] Ding J, Xue N, Xia G S, et al. Object detection in aerial images: a large-scale benchmark and challenges[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(11): 7778-7796.

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

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

Multi-Scale Optical Remote Sensing Image Target Detection Based On Enhanced Small Target Features

Shan Huilin[1,2], Wang Shuoyang[1], Tong Junyi[1], Hu Yuxiang[2], Zhang Yanhao[2], Zhang Yinsheng[1,2*]

[1]*School of Electronics & Information Engineering, Nanjing University of Information Science & Technology, Nanjing* 210044, *Jiangsu, China*;

[2]*School of Electronic & Information Engineering, Wuxi University, Wuxi* 214105, *Jiangsu, China*

Abstract

Objective Remote sensing technology is a method to observe and obtain information about objects and phenomena on the Earth's surface by satellites and aircraft. It allows us to obtain large-scale, multi-spectral, and high-resolution data from remote locations on Earth. The global and real-time technology features multi-spectral observation, high resolution, and multi-source data fusion without contact. Remote sensing target detection is a process of target recognition and extraction using remote sensing data. It aims to automatically detect, locate, and identify specific target types from remote sensing images, which is of significance for disaster warning and response, environmental monitoring, and ecological protection.

Methods The traditional remote sensing image target detection algorithms include valley threshold and Sobel operator and convolutional neural network (CNN) algorithm, of which the most widely employed is the CNN. The algorithm has sound feature extraction and pattern recognition capabilities, but it is sensitive to locations and scale and may still perform poorly when small targets or large-scale changes are involved. Therefore, for the detection of remote sensing targets, it is necessary to consider many factors such as complex background, unbalanced target distribution, dense target, false detection, and missed detection. Therefore, we propose a multi-scale neural network for enhancing small target features (ESF-MNet) to deal with the low detection accuracy and poor generalization of current remote sensing targets. The core idea is to combine multiple CBH modules and CA attention mechanism to form a multi-residual cascade layer and perform efficient aggregation to enhance target feature expression. The RFE module is introduced to help the network better respond to remote sensing targets of different scales. GSConv and CARAFE modules are utilized to form the main structure of the Neck end. While reducing the amount of parameters and maintaining accuracy, the CARAFE module is adopted to improve the semantic extraction ability of the network. Meanwhile, a detection head that is more suitable for small targets is constructed to reduce the lost small target information as the network depth increases.

Results and Discussions Qualitative and quantitative experiments are carried out on mainstream remote sensing detection models such as ESF-MNet, with ablation experiments analyzed. To verify the effectiveness of each improvement point, we conduct seven experiments on DOTA and NWPU NHR-10 datasets under the same environment and parameters based on the YOLOv7 network model. The detected image targets have complex backgrounds, as shown in Table 1. If the attention effect is not employed alone, the mentioned EACM module can significantly improve the effect. The proposed receptive field enhancement module effectively captures context information at different scales. The constructed Neck layer simplifies the network structure and improves the semantic extraction ability, and the proposed detection layer is suitable

for small targets and enhances the fusion of shallow features. The mAP0.5 is improved by 3.7% and 4.5% on the two datasets respectively, which proves the effectiveness of each module. The proposed algorithm is compared with other algorithms to further compare the model performance. The experimental environment is the same, with the same training set and test set adopted. Faster R-CNN, FMSSD, YOLOv5s, YOLOv7, YOLOv8s, algorithms in Refs. [21-23], and the proposed algorithm are shown in Tables 2 and 3. In terms of average accuracy value, the ESF-MNet model performs best. Especially in the aspect of custom small targets, the performance is more prominent. The mAP reaches 83.6% and 97.6% respectively. However, the algorithm accuracy does not reach the best level when detecting some large target objects (such as track and field, basketball court). The main reason is that the network depth after model lightweight is shallow and the downsampling multiple is small. If the network depth and the downsampling multiple increase, although the detection effect of large targets can be improved, poor detection of small targets will be caused. Therefore, our research focus is to improve the detection accuracy of small and medium-sized targets on the premise of ensuring higher detection accuracy for large targets. Generally, compared with other algorithms, the proposed algorithm still has obvious advantages in mAP, greatly reduces the false detection rate, and also meets the basic needs of real-time detection.

Conclusions The detection and recognition of targets in optical remote sensing images is of significance for civilian applications. However, in the case of complex background, dense small targets, and lack of feature information, the identification of small targets is very difficult. Meanwhile, we construct an efficient layer attention aggregation module in the backbone network to extract the target features of various categories and employ the receptive field enhancement module to fuse the feature maps of different depths and thus improve the information expression ability of the network. Additionally, by utilizing GSConv and CARAFE modules to form the Neck layer, and adopting the compression method of halving the number of channels, the neck is finely processed, and the cross-stage partial network (GSCSP) module VoV-GSCSP module is designed by one-time aggregation method, which can reduce the network computation and improve the detection speed. With the addition of the CARAFE module, the detection accuracy is improved. In addition, a multi-scale network is constructed by leveraging a feature output layer with a lower sampling rate of 4, 8, and 16 times in the detection head structure, which effectively improves the detection of small targets. Experimental results show that the model has sound real-time performance and strong robustness for small target detection in complex background. Although the model has been improved, it may still has missed detection and error detection. Although the remote sensing image target detection method is mature, it is still difficult to calculate the large and complex, accurate, and efficient method. However, we will continue to study and solve these problems in the future.

Key words optical remote sensing image; target detection; receptive field enhancement; feature fusion; attention mechanism