引用格式：FU Hongjian, BAI Hongyang, GUO Hongwei, et al. Object Detection Method of Optical Remote Sensing Image with Multi-attention Mechanism[J]. Acta Photonica Sinica, 2022, 51(12):1210003
付宏建,白宏阳,郭宏伟,等.融合多注意力机制的光学遥感图像目标检测方法[J].光子学报,2022,51(12):1210003

融合多注意力机制的光学遥感图像目标检测方法

付宏建[1],白宏阳[1],郭宏伟[1],原瑜蔓[1],秦伟伟[2]
(1 南京理工大学 能源与动力工程学院,南京 210094)
(2 火箭军工程大学 核工程学院,西安 710025)

摘　要:针对光学遥感图像因目标尺度差异大、目标分布密集和背景复杂所导致的检测效果不佳的问题,提出了一种融合多注意力机制的遥感目标检测方法。设计了一种自适应感受野大小的坐标注意力模块,以加强网络对多尺度目标特征的提取能力,提升网络对复杂背景下目标的定位效果。基于Swin Transformer自注意力机制模块改进了YOLOv5网络的预测头,增强了网络对密集分布目标的识别能力。在DOTA公开遥感图像数据集上进行训练与测试,实验对比结果表明,所提方法在检测精度上比YOLOv5网络提高了3.6%,且优于多类典型对比方法;在Nvidia GTX 1080Ti平台上检测速度达49帧/s,证明该方法具有较好的实时检测能力。

关键词:光学遥感图像;目标检测;深度学习;注意力机制;感受野;多尺度;卷积神经网络
中图分类号:TP751　　**文献标识码**:A　　**doi**:10.3788/gzxb20225112.1210003

0　引言

光学遥感图像目标检测技术是指利用算法对感兴趣的遥感图像目标自动分类与定位的技术[1],在军事侦察、精准制导、交通管制、灾情预测等领域有着广泛的应用[2]。从发展历程看,光学遥感图像目标检测技术主要可分为传统目标检测算法和基于深度学习的目标检测算法。传统目标检测算法是指基于手工设计的特征描述子来提取候选目标并进行验证的方法[3],手工设计的特征一般为目标纹理、颜色、边缘等视觉信息[4]。传统目标检测算法主要使用支持向量机(Support Vector Machine, SVM)[5]、Adaboost[6]和K-means[7]等方法作为分类器。

相比传统目标检测算法,基于深度学习的目标检测技术可以自动提取目标特征,特征表达更具鲁棒性和泛化性[8]。根据有无候选框生成阶段作为区分[9],基于深度学习的目标检测技术主要分为以R-CNN(Region-CNN)系列(R-CNN[10]、Fast R-CNN[11]、Faster R-CNN[12]、Mask R-CNN[13])为代表的双阶段模型和以YOLO(You Only Look Once)系列(YOLOv2[14]、YOLOv3[15])、SSD系列(SSD[16]、DSSD[17])为代表的单阶段模型。现阶段,在遥感图像目标检测领域应用深度学习目标检测技术可以达到较好的检测效果,然而,遥感图像目标检测仍有几类难题亟待解决,如目标尺度差异大、目标分布密集、背景复杂等[18]。

针对上述问题,国内外学者在已有深度神经网络的基础上做了大量改进。HOU J Y等[19]在R-CNN的基础上利用多分支的感兴趣区域池化层(Regions Of Interest Pooling, ROI Pooling),将特征映射成不同尺度,并采用级联方式检测,提高了多尺度遥感目标的检测精度,但对于密集分布的遥感目标检测效果不理想。LONG H等[20]融合了传统方法与深度学习方法,提出一种特征融合的深度神经网络,有效提高了密集

基金项目:国家自然科学基金(No. U2031138)
第一作者:付宏建(1998—),男,硕士研究生,主要研究方向为深度学习、目标检测。Email: scarlllet@163.com
导师(通讯作者):白宏阳(1985—),男,教授,博士,主要研究方向为人工智能与计算机视觉。Email:hongyang@mail.njust.edu.cn
收稿日期:2022-05-18;**录用日期**:2022-07-14

分布目标和多尺度目标的检测效果，但该网络流程复杂，不具备工程应用价值。ZHANG Y K等[21]采用语义分割的方式将各类别目标先进行特征掩膜，再采用像素注意力机制对各类别目标加权计算，提升各类别目标的区分度，有利于复杂背景下的目标识别，但这种采用先验知识的方法不具备普适性。张永福等[22]基于Faster R-CNN目标检测框架，提出了一种融合特征的目标检测模型，检测精度得到提升，但模型的每秒传输帧数（Frames Per Second，FPS）仅为6.5左右，无法满足卫星在轨实时处理的需求。

因此，本文对YOLOv5检测网络做出改进，提出了一种融合多注意力机制的YOLOv5检测网络（Multi Attention-YOLOv5，MA-YOLOv5）。在网络中添加一种自适应感受野大小的坐标注意力模块，以加强网络对多尺度目标特征的提取能力，提升网络对复杂背景下目标的定位效果，并基于Swin Transformer自注意力机制模块改进了YOLOv5网络的预测头，增强了网络对密集分布目标的识别能力。

1 模型及改进

1.1 YOLOv5检测网络及改进

根据检测网络深度与宽度系数的不同，YOLOv5模型分为YOLOv5s、YOLOv5m、YOLOv5l、YOLOv5x四个版本，系数越大，模型越复杂，检测精度通常越高，但同时也会牺牲检测速度。考虑到遥感图像在轨实时处理的需求，需保证检测网络具备实时检测能力，因此选用网络深度与宽度系数均为1的YOLOv5l网络作为基础网络。YOLOv5l在结构上主要分为主干（Backbone）、颈部（Neck）和检测器（Prediction）三个部分。Backbone部分主要采用CSPDarknet的主干结构进行特征提取；Neck部分采用FPN（Feature Pyramid Network）+PAN（Path Aggregation Network）的特征金字塔结构进行特征融合；Prediction部分采用CIOU_loss作为损失函数进行计算。

针对遥感图像中目标尺度差异大、背景复杂的特点，在YOLOv5l网络的Neck部分添加一种自适应感受野大小的坐标注意力模块（Adaptive Receptive Field Coordinate Attention，ARFCA），以提升网络对不同尺度目标特征的提取能力，加强网络在复杂背景下对目标的定位能力。针对遥感目标分布密集的特点，向YOLOv5网络的预测头中添加滑动窗口变形器（Swin Transformer，STR）自注意力机制模块，增强网络捕获目标环境信息的能力，形成了MA-YOLOv5网络。MA-YOLOv5网络整体结构如图1所示。

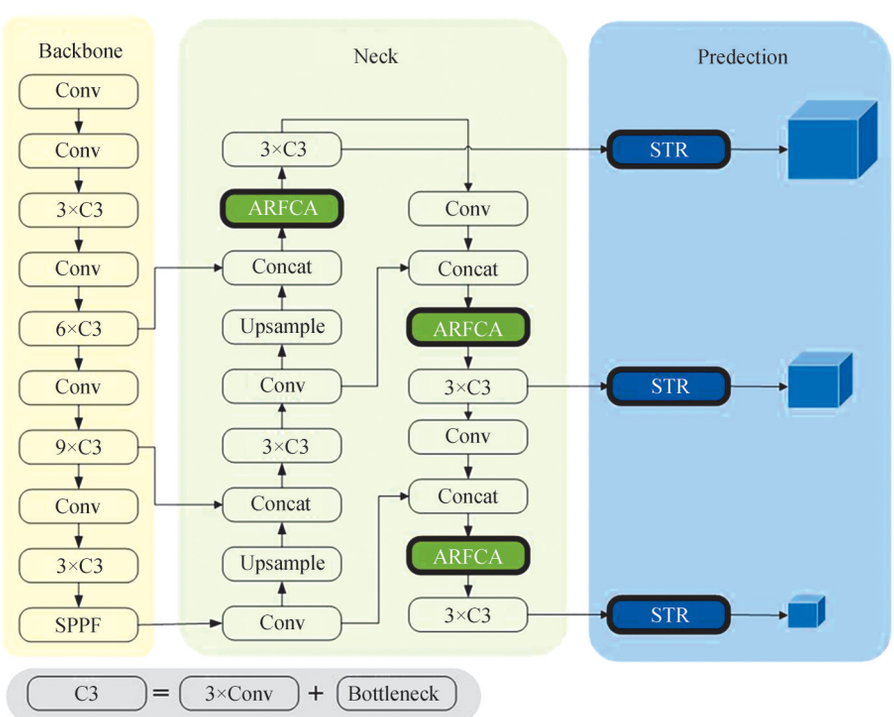


图1 MA-YOLOv5网络结构示意图

Fig.1 Schematic diagram of MA-YOLOv5 network structure

1.2 ARFCA模块

针对遥感图像中目标尺度差异大、背景复杂的特点，提出了一种自适应感受野大小的坐标注意力模块，模块结构如图2所示。该模块主要分为分离、坐标注意力和选择三个部分。

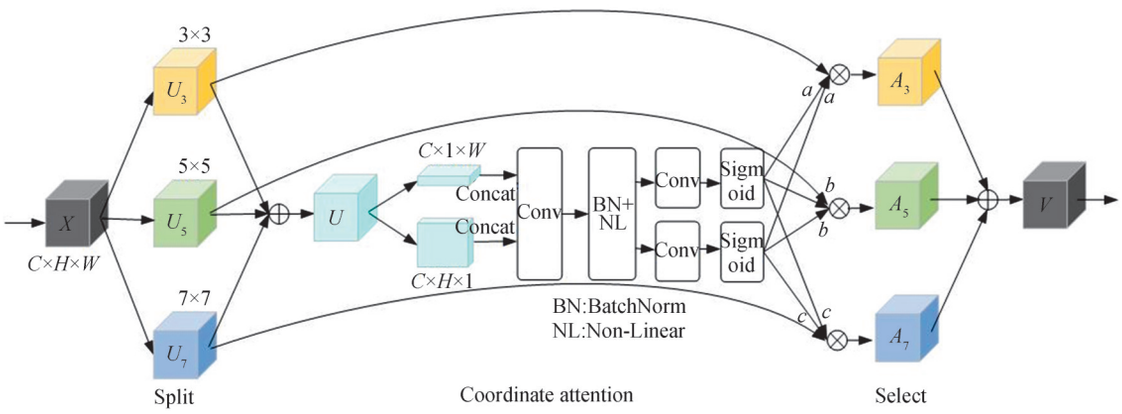


图2 ARFCA模块示意图(3通道)

Fig.2 Schematic diagram of ARFCA module (3 channels)

分离部分引入了不同卷积核大小的卷积作为并行分支进行特征处理，该模块的分支数可根据数据集的不同进行改变，图中展示了分支数为3时的情况(按图示易推得其他分支数时的情况)。三分支分别采用3×3、5×5和7×7的卷积核。为保证模型效率，5×5和7×7卷积实际为空洞率大小为2和4的3×3空洞卷积。空洞卷积可在不增加参数量的情况下增大卷积的感受野。

在坐标注意力部分，首先将不同尺寸卷积输出的信息相加，得到融合元素U，随后分别使用$(H,1)$和$(1,W)$的空间通道池化核对U元素进行高度和宽度通道方向上的池化，得到维度分别为$C\times1\times W$和$C\times H\times1$的特征。对于空间通道池化后的特征来说，第c个通道在高度h上输出可表示为

$$z_c^h(h)=\frac{1}{W}x_c(h,i) \tag{1}$$

第c个通道在宽度w上的输出可表示为

$$z_c^w(w)=\frac{1}{H}x_c(j,w) \tag{2}$$

上述变换完成后，对两方向特征进行拼接，随后进行卷积变换，并使用非线性激活函数进行激活，即

$$\boldsymbol{f}=\delta\left(F_1\left(\left[\boldsymbol{z}^h,\boldsymbol{z}^w\right]\right)\right) \tag{3}$$

式中，$\left[\boldsymbol{z}^h,\boldsymbol{z}^w\right]$代表拼接处理，$F_1$步骤为卷积变换，$\delta$为非线性激活函数。随后对$\boldsymbol{f}$分别进行高度和宽度方向的卷积变换和Sigmoid函数激活，得到坐标注意力的两个输出。公式为

$$\begin{cases}\boldsymbol{g}^h=\sigma\left(F_h\left(\boldsymbol{f}^h\right)\right)\\ \boldsymbol{g}^w=\sigma\left(F_w\left(\boldsymbol{f}^w\right)\right)\end{cases} \tag{4}$$

为实现ARFCA模块自适应选择不同感受野大小的卷积所输出的信息，在坐标注意力的中引入三个softmax注意力系数a、b、c，第c个通道中a_c、b_c和c_c的计算公式为

$$\begin{cases}a_c=\dfrac{\mathrm{e}^{\boldsymbol{A}_c\boldsymbol{g}^h\boldsymbol{g}^w}}{\mathrm{e}^{\boldsymbol{A}_c\boldsymbol{g}^h\boldsymbol{g}^w}+\mathrm{e}^{\boldsymbol{B}_c\boldsymbol{g}^h\boldsymbol{g}^w}+\mathrm{e}^{\boldsymbol{C}_c\boldsymbol{g}^h\boldsymbol{g}^w}}\\ b_c=\dfrac{\mathrm{e}^{\boldsymbol{B}_c\boldsymbol{g}^h\boldsymbol{g}^w}}{\mathrm{e}^{\boldsymbol{A}_c\boldsymbol{g}^h\boldsymbol{g}^w}+\mathrm{e}^{\boldsymbol{B}_c\boldsymbol{g}^h\boldsymbol{g}^w}+\mathrm{e}^{\boldsymbol{C}_c\boldsymbol{g}^h\boldsymbol{g}^w}}\\ c_c=\dfrac{\mathrm{e}^{\boldsymbol{C}_c\boldsymbol{g}^h\boldsymbol{g}^w}}{\mathrm{e}^{\boldsymbol{A}_c\boldsymbol{g}^h\boldsymbol{g}^w}+\mathrm{e}^{\boldsymbol{B}_c\boldsymbol{g}^h\boldsymbol{g}^w}+\mathrm{e}^{\boldsymbol{C}_c\boldsymbol{g}^h\boldsymbol{g}^w}}\end{cases} \tag{5}$$

式中，$a_c+b_c+c_c=1$，$\boldsymbol{g}^h$和$\boldsymbol{g}^w$是上一步H和W空间方向上的输出，$\boldsymbol{A}$、$\boldsymbol{B}$、$\boldsymbol{C}\in\mathrm{R}^{N\times I}$，分别代表三个通道的

softmax注意力权重，其中N是特征的通道数，A_c、B_c、$C_c\in \mathrm{R}^{1\times I}$代表第$c$个通道$A$、$B$、$C$对应的矩阵，$I$表示为

$$I=\max(N/r, L) \tag{6}$$

式中，r是为控制输出值而自定义的压缩比例系数，$L=32$是实验设置的常量。

在选择部分，将坐标注意力输出的三个注意力系数a、b、c分别与原特征输出U_3、U_5和U_7相乘，得到A_3、A_5、A_7三个不同尺度的特征分量，将这三个分量相加，获得特征输出。第c个通道上的特征输出V_c的计算公式为

$$V_c=A_{3c}+A_{5c}+A_{7c}=a_c\cdot U_{3c}+b_c\cdot U_{5c}+\mathrm{c}_c\cdot U_{7c} \tag{7}$$

则$V=[V_1, V_2, \cdots, V_c]$。

对比传统的注意力模块，ARFCA模块通过分离和选择机制实现了根据输入目标尺寸大小动态调整模块感受野大小的效果，从而提升了多尺度特征提取能力。在坐标注意力部分，分别沿两个方向进行空间特征的聚集，生成一对具有方向感知力的特征图。在保存一个方向上位置信息的同时捕捉到另一个空间方向的长期依赖关系，有助于网络更准确地定位感兴趣的目标。

1.3 Swin Transformer模块

受Swin Transformer网络的启发[23]，在YOLOv5的检测头中添加Swin Transformer模块，模块结构如图3所示。

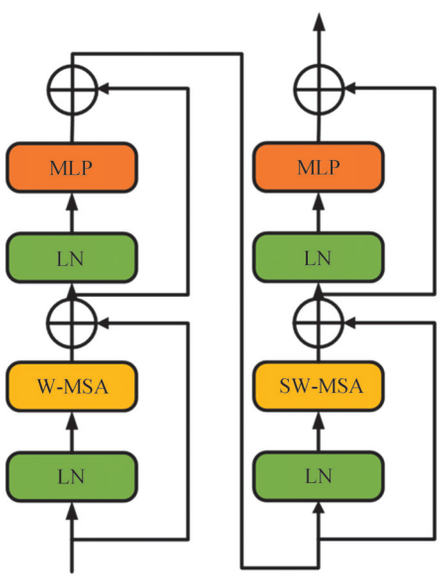


图3 Swin transformer模块示意图

Fig.3 Schematic of the Swin transformer module

该模块包含两个子层，主要由窗口多头自注意力层(Window Multi-head Self-Attention, W-MSA)和移位窗口多头自注意力层(Shifted-Window Multi-head Self-Attention, SW-MSA)组成。由于该结构限制，Swin Transformer模块的层数通常为2的整数倍。窗口多头自注意力层和移位窗口多头自注意力层均把自注意力的计算限制在窗口中，相比于传统Transformer中的自注意力MSA模块，STR模块大大降低了计算复杂度，式(7)～(8)展示了MSA模块与W-MSA、SW-MSA模块计算复杂度的对比，可以看出MSA模块计算量与hw(h、w分别表示特征高和宽的数值)呈二次关系，而W-MSA和SW-MSA模块与hw呈线性关系。

$$\Omega_{\mathrm{MSA}}=4hwC^2+2(hw)^2C \tag{8}$$

$$\Omega_{\mathrm{W-MSA/SW-MSA}}=4hwC^2+2\mathrm{M}^2hwC \tag{9}$$

式中，Ω表示计算复杂度，M为常量(一般设置为7)。

特征进入STR模块首先经过LN(Layer Normalization)层进行归一化，随后进入W-MSA层，在窗口中进行自注意力的计算，之后经过MLP层得到第一模块的输出。自注意力的计算公式为

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}+\boldsymbol{B}\right)\boldsymbol{V} \tag{10}$$

式中，Attention($\boldsymbol{Q}$,$\boldsymbol{K}$,$\boldsymbol{V}$)的计算是根据关注程度对Value乘以相应权重，权重由$\boldsymbol{Q}$和$\boldsymbol{K}$计算获得，计算结果为Value的加权和；Softmax函数是一种归一化指数函数；$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$分别是Query、Key、Value对应的矩阵，Query、Key是计算Attention权重的特征向量，Value表示输入特征的向量；d是$\boldsymbol{Q}$和$\boldsymbol{K}$的向量维度；$\boldsymbol{B}$是一个偏置矩阵。

在STR的第二模块中运用SW-MSA层，基于移动窗口的分割方法进行自注意力的计算，随后经过MLP层进行全局平均池化，得到最终的预测结果。Swin Transformer模块中的LN层能帮助模型更好地收敛，防止模型过度拟合，W-MSA和SW-MSA层中多头自注意力机制的引入不仅能帮助模型关注当前像素，还能增强模型捕获当前像素环境信息的能力，从而提升模型对密集分布目标的检测性能。

2 实验

2.1 数据集

采用DOTA[24]遥感图像公开数据集进行实验，验证改进网络的有效性，具体为DOTAv1.5版本。该版本包含16个类别，40万余个带有注释的目标，其中最小注释目标仅为10像素左右。具体类别分别为轮船、储罐、飞机、棒球场、网球场、篮球场、小型车辆、大型车辆、直升机、田径场、港口、桥梁、环岛、足球场、游泳池和集装箱起重机。数据集的注释文件分为两个版本，分别是旋转目标标注框(Oriented Bounding Box，OBB)和水平目标标注框(Horizontal Bounding Box，HBB)，本文采用水平目标标注框作为注释标签。数据集中原始图像分辨率最大达20 000×20 000左右，为避免大像素图像输入网络时在调整图片大小步骤造成图像信息损失，采用分割脚本将大分辨率图像切割成每张像素大小为600×600的图像，便于网络进行目标检测。数据集分为训练集和验证集两部分，训练集共71 254张图片，验证集共7 917张图片。

2.2 实验环境及参数设置

实验环境如下：操作系统为Ubuntu18.04，内存为16G，CPU使用Intel(R)Core(TM) i7-7700K@4.2GHz，GPU使用Nvidia GeForce GTX 1080Ti(显存为11G)，深度学习框架采用Pytorch1.10.2版本。实验设置训练最大迭代轮数为12个轮次，初始学习率为0.005，循环学习率为0.1，学习率动量为0.937，交并比损失系数为0.05，分类损失系数为0.5，有无物体系数为1.0。分别在第8和第10个轮次下降学习率。对于SSD和YOLOv5系列的网络，采用K-means聚类方法计算生成遥感图像目标对应尺度的锚点，计算结果为[11，12]，[21，21]，[30，44]，[47，32]，[51，75]，[96，56]，[106，120]，[176，209]，[356，365]。

2.3 ARFCA模块分支数消融实验

为验证ARFCA模块分支数的不同对模型所造成的影响，同时确定ARFCA模块最佳分支数量，设置了一组消融实验，以MA-YOLOv5网络为基础，将SKCA分支数分别设置为1、2、3、4，其他训练参数设置均保持一致进行实验。表1展示了不同分支数对应的平均检测精度和检测速度。

表1 ARFCA不同分支数的检测精度与检测速度

Table 1 Detection accuracy and speed of ARFCA with different branch numbers

Branch number of ARFCA	Convolution kernel size of each branch	mAP	FPS
1	5×5	68.0	50.9
2	3×3，5×5	68.4	50.2
3	3×3，5×5，7×7	68.5	49.4
4	3×3，5×5，7×7，9×9	68.5	48.3

观察表1数据可得，在分支数为1、2、3时，随着分支数的增加，模型的平均检测精度(mean Average Precision，mAP)有微弱提升，而当分支数为4时，模型mAP值与分支数为3时相比不再提升。从检测速度角度看，随着分支数的增加，模型的FPS始终呈下降趋势。因此，为实现模型的最高检测精度，同时保证模型检测速度，将ARFCA模块分支数确定为3，并进行后续实验。

2.4 结果与分析

对MA-YOLOv5、添加ARFCA模块的YOLOv5、添加STR模块的YOLOv5、YOLOv5原始网络以及三个以ResNet-50为骨干网络的一阶段网络：SSD、RetinaNet和FCOS，共7个网络进行实验对比，并采用全类平均精度mAP（IOU=0.5:0.95）、平均精度AP（IOU=0.5）、AP（IOU=0.75），coco数据集定义的小中大目标对应的mAP：APS（目标面积$<32^2$像素）、APM（$32^2<$目标面积$<96^2$像素）、APL（目标面积$>96^2$像素），共6种指标作为模型精度的评价标准，采用检测速度FPS作为模型速度的评价标准。各模型的详细参数结果如表2所示。

表2 不同网络在测试集下的性能

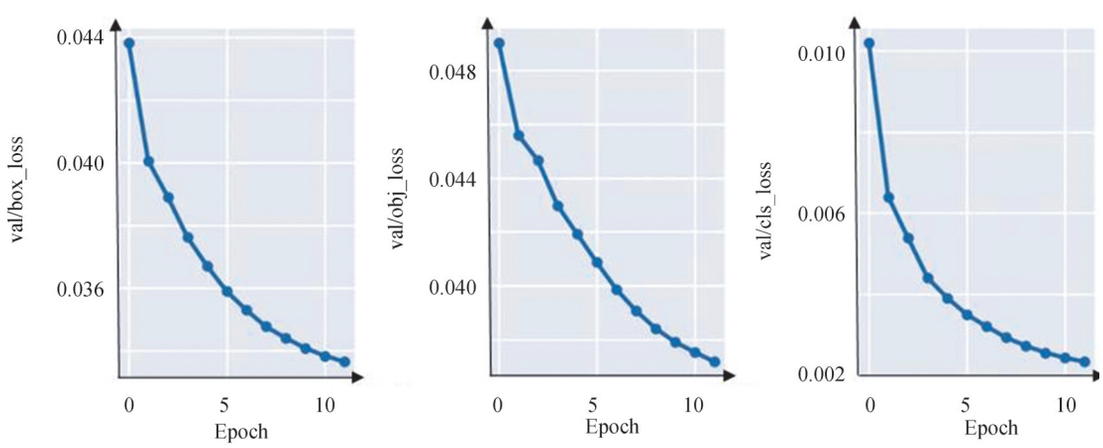
Table 2 The performance of different networks on the test set

Method	Backbone	mAP	AP50	AP75	APS	APM	APL	FPS
SSD	ResNet-50	48.4	80.9	51.5	29.0	57.5	51.3	**77**
RetinaNet	ResNet-50	60.7	88.8	70.5	54.2	66.2	54.8	28
FCOS	ResNet-50	63.5	89.8	74.3	56.2	69.3	58.2	38
YOLOv5	CSPdarknet	64.9	91.7	78.4	58.0	72.4	61.2	59
YOLOV5-STR	CSPdarknet	**66.3**	92.6	80.2	58.6	73.3	62.9	52
YOLOv5-ARFCA	CSPdarknet	**67.2**	93.0	80.6	59.2	73.5	63.4	54
MA-YOLOv5	CSPdarknet	**68.5**	93.4	82.8	60.3	76.5	65.3	49

观察表2结果可以看出，在检测精度方面，MA-YOLOv5网络在实验对比的7个网络中最高，mAP值达到了68.5%，对比原始YOLOv5网络，实现了3.6%的精度提升。添加ARFCA模块和STR模块的YOLOv5网络在平均检测精度上分别实现了2.3%和1.4%的精度提升。而SSD、RetinaNet和FCOS在检测精度上的表现均与YOLOv5系列网络有一定差距。在检测速度方面，SSD网络表现最佳，达到了77FPS，原始YOLOv5网络FPS达到59，在对比的YOLOv5系列网络中表现最好，而YOLOv5改进后的三个网络在FPS上略有下降，但仍具有实时检测的能力。

图4展示了MA-YOLOv5训练和验证数据集时对应的定位损失、置信度损失和分类损失曲线的变化情况，其中各图横坐标表示模型训练批次。

图5展示了MA-YOLOv5网络在DOTA数据集上测试时各类别的平均精度mAP（IOU=0.5:0.95）。可以看出，有6类目标检测精度大于0.7，包括尺寸较小的舰船类别和尺寸较大的篮球场、网球场类别，说明本文所提出方法对多尺度遥感目标具有较好检测性能。



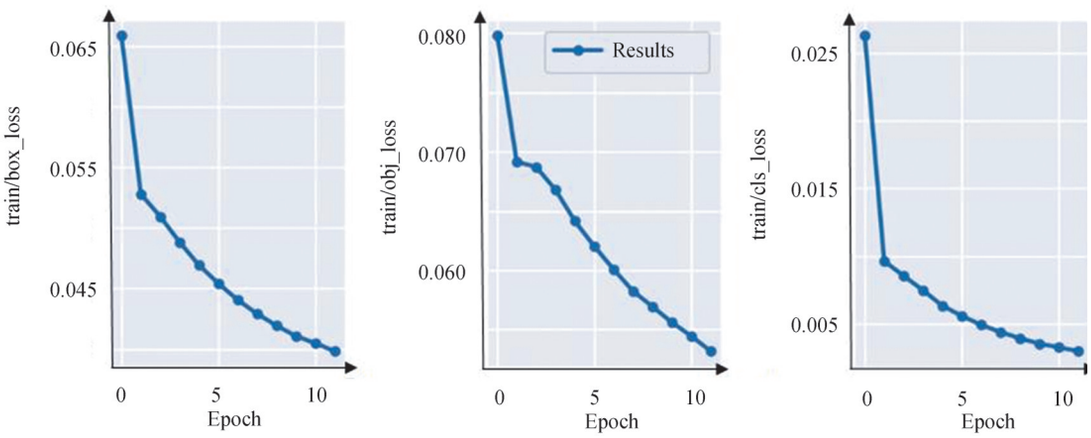


图4　MA-YOLOV5网络在训练与验证时的损失值

Fig.4　The loss value of MA-YOLOV5 network during training and validation

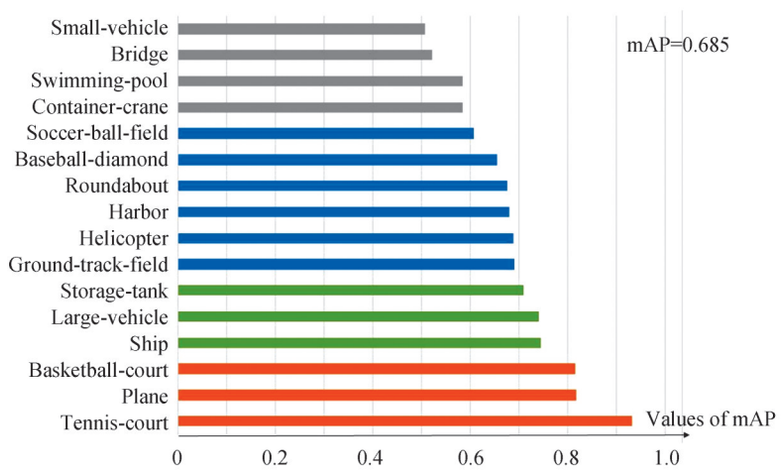


图5　各类别目标的mAP值

Fig.5　mAP values for each category of targets

3　结论

本文以YOLOv5算法为基础，提出了一种融合多注意力机制的光学遥感图像目标检测算法。针对遥感图像中目标尺度差异大、背景复杂的特点，设计了一种自适应感受野大小的坐标注意力（ARFCA）模块，通过模块中的分离和选择机制，根据输入目标大小，自适应地选择不同感受野大小的卷积所输出的信息，从而提升模型对于多尺度遥感目标的特征提取能力。同时，通过ARFCA模块中的坐标注意力机制，捕捉一个空间方向的长期依赖关系，并保存另一个空间方向的位置信息，有助于网络更准确地定位目标。此外，通过在YOLOv5预测头中加入Swin Transformer自注意力机制模块，增强了模型捕获目标环境信息的能力，提升模型对密集分布目标的检测性能。实验结果证明了MA-YOLOv5模型中ARFCA模块和Swin Transformer模块对于遥感图像目标检测效果提升的有效性，以及模型具有较好的实时性与一定的工程应用价值。

参考文献

[1] NIE Guangtao, HUANG Hua. A survey on object detection technology in optical remote sensing images[J]. Acta Automatica Sinica, 2021, 47(8): 1749-1768.
聂光涛, 黄华. 光学遥感图像目标检测算法综述[J]. 自动化学报, 2021, 47(8): 1749-1768.

[2] WANG Jianan, GAO Yue, SHI Jun, et al. Scene classification of optical high-resolution remote sensing images using vision transformer and graph convolutional network[J]. Acta Photonica Sinica, 2021, 50(11): 1128002.
王嘉楠, 高越, 史骏, 等. 基于视觉转换器和图卷积网络的光学遥感场景分类[J]. 光子学报, 2021, 50(11): 1128002.

[3] NI Kang, ZHAO Yuqing, CHEN Zhi. Multi-scale convolutional neural network driven by sparse second-order attention mechanism for remote sensing scene classification[J]. Acta Photonica Sinica, 2022, 51(6): 0610004.

倪康，赵雨晴，陈志. 稀疏二阶注意力机制驱动的多尺度卷积遥感图像场景分类网络[J]. 光子学报，2022，51(6)：0610004.

[4] WANG Zijian. Multi-scale remote sensing object detection based on attention mechanism[D]. Beijing：University of Chinese Acdaemy of Sciences，2021.
王子健. 基于注意力机制的多尺度遥感目标检测[D]. 北京：中国科学院大学，2021.

[5] OSUNA E，FREUND R，GIROSIT F. Training support vector machines：an application to face detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，1997：130-136.

[6] VIOLA P，JONES M. Rapid object detection using a boosted cascade of simple features[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2001：511-518.

[7] KAYASAL U. Magnetometer aided inerital navigation system：modeling and simulation of a navigation system with an IMU and a magnetometer[M]. Turkey：National Defense Industry Press，2009：74-77.

[8] NI Kang，LIU Pengfei，WANG Peng. Compact global-local convolutional network with multifeature fusion and learning for scene classification in synthetic aperture radar imagery[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing，2021，14：7284-7296.

[9] ZHAO Yongqiang，RAO Yuan. A survey of deep learning object detection methods[J]. Journal of Image and Graphics，2020，25(4)：629-654.
赵永强，饶元. 深度学习目标检测方法综述[J]. 中国图像图形学报，2020，25(4)：629-654.

[10] GIRSHICK R，DONAHUE J，DARRELL T，et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C].Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2014：580-587.

[11] GIRSHICK R. Fast R-CNN[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2015：1440-1448.

[12] REN S，HE K，GIRSHICK R，et al. Faster R-CNN：towards real-time object detection with region proposal networks[J]. IEEE Trans Pattern Analysis and Machine Intelligence，2017，39(6)：1137-1149.

[13] HE K，GKIOXARI G，DOLLAR P，et al. Mask R-CNN[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2017：2961-2969.

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

[15] REDMON J，FARHADI A. YOLOv3：an incremental improvement[J/OL].[2022-05-18]. http://arxiv.org/abs/1804.02767.

[16] LIU W，ANGUELOV D，ERHAN D，et al. SSD：single shot MultiBox detector[C]. European Conference on Computer Vision，Springer，Cham，2016：21-37.

[17] FU C Y，LIU W，RANGA A，et al. DSSD：deconvolutional single shot detector[J/OL].[2022-05-18]. https://arxiv.org/abs/1701.06659.

[18] GUO H W，BAI H Y，YUAN Y M，et al. Fully deformable convolutional network for ship detection in remote sensing imagery[J].Remote Sensing，2022，14(8)：1850-1869.

[19] HOU J Y，MA H B，WANG S J. Parallel cascade R-CNN for object detection in remote sensing imagery[J]. Journal of Physics：Conference Series，2020，1544：012124.

[20] LONG H，CHUNG Y，LIU Z B，et al. Object detection in aerial images using feature fusion deep networks[J]. IEEE Access，2019，7:30980-30990.

[21] ZHANG Y K，YOU Y N，WANG R，et al. Nearshore vessel detection based on Scene-mask R-CNN in remote sensing image[C]. 2018 International Conference on Network Infrastructure and Digital Content. Guiyang，China：IEEE Access，2018：76-80.

[22] ZHANG Yongfu，SONG Hailin. Deep learning remote sensing image target detection model based on fusion features[J]. Computer Technology and Development，2021，31(9)：48-54.
张永福，宋海林. 融合特征的深度学习遥感图像目标检测模型[J]. 计算机技术与发展，2021，31(9)：48-54.

[23] LIU Z，LIN Y T. Swin transformer：hierarchical vision transformer using shifted windows[J/OL].[2022-05-18]. https://arxiv.org/abs/2103.14030v2.

[24] XIA G S，BAI X，DING J，et al. DOTA：a large-scale dataset for object detection in aerial images[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2018：3974-3983.

Object Detection Method of Optical Remote Sensing Image with Multi-attention Mechanism

FU Hongjian[1], BAI Hongyang[1], GUO Hongwei[1], YUAN Yuman[1], QIN Weiwei[2]
(1 *School of Energy and Power Engineering, Nanjing University of Science and Technology, Nanjing* 210094, *China*)
(2 *School of Nuclear Engineering, Rocket Force University of Engineering, Xi'an* 710025, *China*)

Abstract: Optical remote sensing image target detection technology refers to the technology that uses algorithms to automatically classify and locate objects of interest. It has a wide range of applications in military reconnaissance, precision guidance and urban construction. From the perspective of development history, optical remote sensing image target detection technology can be mainly divided into traditional target detection algorithms and deep learning-based target detection algorithms. Compared with traditional target detection algorithms, deep learning-based target detection algorithms can automatically extract target features, and the feature expression is more robust and generalisable. In the field of remote sensing image target detection, the application of deep learning target detection technology can achieve better detection results. However, several problems still exist in remote sensing image target detection, such as large differences in target scales, dense target distribution and complex backgrounds. In response to the above problems, this paper makes improvements based on the YOLOv5 network, and proposes the MA-YOLOv5 (Multi Attention-YOLOv5) network, which improves the remote sensing target detection effect, and the experiments verify the effectiveness of the improvement. Considering the requirement of on-orbit real-time processing of remote sensing images, ensuring a certain detection speed is necessary. Therefore, this paper selects the YOLOv5l network whose network depth and width coefficients are oneas the basic network. YOLOv5 is mainly divided into three parts: Backbone, Neck and Prediction. The Backbone part mainly uses the backbone structure of CSP (Cross Stage Partial) Darknet for feature extraction; the Neck part uses the FPN (Feature Pyramid Network)+PAN (Path Aggregation Network) feature pyramid structure for feature fusion; the Prediction part uses CIOU_loss (C Intersection over Union_loss) as the loss function for calculation. To improve the detection effect of remote sensing images with multiple scales and complex backgrounds, this paper proposes a coordinate attention module with adaptive receptive field size. Through the separation and selection mechanism in the module, the network can adaptively select the information output by convolutions with different receptive field sizes according to the size of the target, thereby improving the feature extraction ability of the model for multi-scale remote sensing targets. At the same time, through the coordinate attention mechanism in the module, the long-term dependency of one spatial direction is captured, and the position information of another spatial direction is saved, which helps the network to locate the target more accurately. In addition, in view of the dense distribution of remote sensing targets, the Swin Transformer self-attention mechanism module is added to the protection head of the YOLOv5 network to enhance the network's ability to capture the target environment information. To verify the influence of the different number of branches of the ARFCA (Adaptive Receptive Field Coordinate Attention) module on the model, and to determine the optimal number of branches of the ARFCA module, a set of ablation experiments are set up in this paper. The experimental results show that the best effect is when the number of ARFCA branches is 3. Finally, this paper sets up a set of experiments to compare the following seven networks: The MA-YOLOv5, YOLOv5 with ARFCA module added, YOLOV5 with STR (Swin Transformer) module added, YOLOv5 original network, SSD, RetinaNet and FCOS. Seven categories of indicators are used for evaluation. The experimental results show that compared with the original YOLOv5 network, the MA-YOLOv5 network achieves a 3.6% improvement in accuracy and has a certain ability of real-time detection.

Key words: Optical remote sensing image; Target detection; Deep learning; Attention mechanism; Receptive field; Multiscale; Convolutional neural networks
OCIS Codes: 100.2000; 280.4788; 120.0280; 110.2970

Foundation item: National Natural Science Foundation of China (No. U2031138)