ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 39,No. 6,Nov. 2024,pp. 1455－1469
DOI:10. 16337/j. 1004-9037. 2024. 06. 014

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

MSDAB-DETR:一种多尺度遥感目标检测算法

李　烨，周生翠，张　驰

(上海理工大学光电信息与计算机工程学院,上海 200093)

摘　要: 由于遥感图像中的目标尺寸差异大,且捕获不同尺度目标的信息非常困难,因此难以有效识别不同尺度目标。同时,传统Transformer在处理高分辨率图像时会出现计算资源不足的问题;单一的损失计算方式和匈牙利算法结合会增大代价损失的波动性,影响算法的收敛速度和精度。基于上述问题,本文提出一种基于改进DAB-DETR的多尺度遥感目标检测算法(Multi-scale dynamic anchor boxes for DETR, MSDAB-DETR)。首先,该算法通过创建一种新型的多尺度注意力融合模块,利用不同分辨率特征信息之间的差异,实现了对遥感图像的多尺度预测。其次,采用高效注意力机制对Transformer模型中的自注意力机制进行改进,降低原始模型的内存占用量。最后,利用SIoU损失函数作为边界框回归损失,与匈牙利算法相结合,削弱了二分图匹配的波动性,加快了收敛速度,并进一步改善了边界框的回归能力。实验结果表明,该方法在NWPU VHR-10和DIOR数据集上的检测精度分别高达95.3%和71.5%;在NWPU VHR-10数据集上,小、中、大3种尺度目标的平均检测精度相较于DAB-DETR模型分别提升了10.5%、1.8%和2.7%;内存占用量减少约9%。

关键词: 遥感图像检测;DAB-DETR模型;多尺度注意力融合;高效注意力Transformer;SIoU损失

中图分类号: TP391.41　　**文献标志码:** A

MSDAB-DETR: A Multi-scale Remote Sensing Target Detection Algorithm

LI Ye, ZHOU Shengcui, ZHANG Chi

(School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Due to the large differences of target size in remote sensing images and the difficulty in effectively capturing the effective features of targets at different scales, it is difficult to effectively identify targets at different scales. And, when dealing with high-resolution images, traditional Transformers may face the problem of insufficient computational resources. In addition, the combination of a single loss calculation method and the Hungarian algorithm can increase the fluctuation of cost loss and affect the convergence speed and accuracy of the algorithm. Therefore, a multi-scale remote sensing target detection algorithm, named as MSDAB-DETR, is proposed. Firstly, the algorithm creates a new multi-scale attention fusion module to leverage the differences between different resolution feature information to achieve multi-scale prediction of remote sensing images. Secondly, an efficient attention mechanism is adopted to improve the self-attention mechanism in the Transformer model, reducing the memory footprint of the original model. Finally, the SIoU loss function is used as the bounding box regression loss,

收稿日期:2023-12-04;**修订日期:**2024-03-04

combined with the Hungarian algorithm, to weaken the fluctuation of binary graph matching, accelerate the convergence speed, and further improve the regression ability of bounding boxes. Experimental results show that the detection accuracy of this method on the NWPU VHR-10 and DIOR datasets is as high as 95.3% and 71.5%, respectively. Among them, on the NWPU VHR-10 dataset, the average detection accuracy for small, medium, and large-scale targets is improved by 10.5%, 1.8%, and 2.7%, respectively compared to the DAB-DETR model. At the same time, the memory footprint is reduced by about 9%.

Key words: remote sensing image detection; DAB-DETR model; multi-scale attention fusion; efficient attention Transformer; SIoU loss

引　言

遥感目标检测是遥感领域的一个重要研究方向[1-2]，它利用遥感图像数据自动识别和定位特定目标或物体，如飞机、桥梁和汽车等。该技术在城市规划、军事监测及农业资源管理等多个领域得到了广泛应用，具有重要的实用价值。随着深度学习的迅速发展，神经网络的结构也越来越复杂，其提取出的高级特征蕴含的信息也愈加抽象，具有更强的语义表征能力和判别能力。因此，目前在遥感目标检测任务中已经普遍采用以卷积神经网络（Convolutional neural network，CNN）为代表的深度学习方法。

基于深度学习的目标检测算法主要可以分为两类。一类是基于无锚框的检测算法，此类算法不依赖于预定义的锚框，而是通过对输入图像采用多尺度分割或滑动窗口分割的方式，在图像的多个尺度和位置上，直接预测目标的位置和类别，典型的算法包括YOLOv1[3]、CornerNet[4]、CenterNet[5]、FCOS[6]和DETR[7]等。另一类是基于锚框的检测算法，此类算法结合了传统滑动窗口方法和生成一系列区域建议（如R-CNN[8]系列算法）的思想，在输入图像上预先生成一组固定大小和宽高比的锚框，并利用这些锚框来匹配目标对象，根据匹配结果对目标对象的位置进行精准定位，典型的算法包括Faster R-CNN[9-10]、SSD[11]、YOLOv2[12]、YOLOv3[13]、YOLOv4[14]、YOLOv5[15]、RetinaNet[16]和Cascade R-CNN[17]等。

近年来，目标检测算法开始广泛应用于遥感图像领域[18]。比如Zhang等[19]提出一种将FCOS与特征金字塔网络（Feature Pyramid network，FPN）[20]结合在一起的遥感目标检测算法，利用多尺度的特征金字塔提取特征。YANG等[21]在Faster R-CNN的基础上引入多维注意力网络和采样融合网络来减弱噪声的影响，提高对多尺度目标的检测精度。Hou等[22]在Cascade R-CNN的基础上，利用多个并行的RoIAlign模块进一步提升检测精度。Fu等[23]提出了一种基于FPN的Faster R-CNN目标检测算法，更好地解决不同尺度目标检测的问题。然而，以上所述算法主要基于CNN进行局部特征提取，对全局特征考虑得并不充分。当仅使用局部特征进行目标检测时，其容易受到背景干扰或相似物体的影响，使得目标与周围环境的区分变得困难，进而导致目标漏检和误检的问题。全局特征则能捕捉图像的整体结构，提供更全面的信息，有助于准确区分目标与周围环境，从而提高检测的准确性和鲁棒性。

当前目标检测算法提取全局特征的常用方法是使用Transformer[24-25]。Transformer可以为每个位置提供对其他位置的全局信息，从而捕获输入信息中的全局特征。将CNN与Transformer相结合，可以使算法获得更全面的语义信息，提升模型检测能力。Li等[26]提出的TRD算法，将Transformer与FPN相结合，增强算法的特征提取能力。Shen等[27]提出的MAME-YOLOX算法采用YOLOX作为基线网络，在其颈部模块中嵌入Swin Transformer，获取全局上下文信息以提取特征，提高检测精度。

Carion等[7]提出的DETR不仅成功地将CNN和Transformer相结合，实现了局部特征和全局特征的综合利用，而且还省去了传统目标检测算法中先验框筛选和后处理等步骤，实现了真正意义上的端

到端的目标检测。但是,DETR存在编码器计算复杂度高和收敛速度慢的问题。为此,Li等[28]提出了DN-DETR模型,将去噪方法引入DETR模型,以加速收敛。Zhang等[29]进一步提出了DINO模型,通过改进去噪训练、查询初始化和框预测方式进行优化。Liu等[30]提出了DAB-DETR(Dynamic anchor boxes for DETR)算法,该算法将锚框重新引入到DETR中,提供了良好的位置先验信息。通过这种方式,DAB-DETR能更快地关注到物体所在的位置,从而有效加速模型的收敛速度。然而,由于高分辨率遥感图像中同时存在大、中、小不同尺度的目标,将DAB-DETR直接应用到遥感图像中时会出现尺度多样性适应能力不足的问题,难以有效同时识别不同尺度目标,尤其是小尺度目标,容易出现漏检的情况。并且由于DAB-DETR使用传统Transformer,在处理高分辨率图像时,会出现计算资源不足的问题。此外,单一的损失计算方式和匈牙利算法结合会增大代价损失的波动性,从而影响算法的收敛速度和精度。

针对上述问题,本文提出了一种基于改进DAB-DETR的多尺度遥感目标检测算法(Multi-scale dynamic anchor boxes for DETR,MSDAB-DETR)。本文主要贡献如下:首先,创建一种新型的基于注意力的多尺度融合模块,通过分析不同分辨率特征之间的异同,使得高分辨率特征中的细节信息更显著。随后,利用基于注意力的特征融合模块将其与低分辨率特征融合,得到具有更丰富表征能力的特征,提升模型对多尺度目标的适应能力。其次,提出基于高效注意力机制(Efficient attention,EA)[31]的Transformer(EA Transformer,EAT),在处理高分辨率图像时,显著降低了计算资源消耗,展现出更好的可扩展性。最后,采用SIoU损失函数[32]作为边界框回归损失,与匈牙利算法相结合来进行二分类匹配。这种结合能够有效地减少总自由度,从而降低二分图匹配的波动性,加速模型的收敛过程。同时,这种方法还能减小边界框回归误差,提高模型的检测性能。本文方法不仅充分考虑了局部以及全局的特征提取,而且在降低计算成本的情况下,使多尺度遥感图像的检测精度达到了更好的效果。

1　DAB-DETR算法

1.1　整体结构

DAB-DETR在DETR的基础上,引入了以(x, y, w, h)表示的4维锚框作为Object Queries,并逐层更新。其中,锚框中心点(x, y)提供更好的位置先验,锚框的宽高信息(w, h)来调整位置注意力图,从而显著提高模型的收敛速度。

DAB-DETR的整体结构包括CNN骨干网络、Transformer编码器(Encoder)和解码器(Decoder),以及用于预测边界框和标签的预测头前馈网络(Feed forward network, FFN),如图1所示。输入1幅图片后,模型首先使用CNN骨干网络提取特征,随后将最后一层特征与其位置编码(Position encoder)相加,并传递到Transformer编码器以捕获全局上下文信息。然后,将一定数量的动态锚框嵌入到解码器输入中,以提取出更具针对性的特征。解码器的最后一层输出经过共享的FFN的处理,用于预测目标框的坐标和类别。最终,将预测框与真实值进行二分图匹配,并计算匹配的目标框的相应损失。

1.2　自注意力Transformer模型

Transformer是一种基于自注意力机制的神经网络模型,最初用于自然语言处理任务,如机器翻译。与卷积神经网络不同,Transformer能够在处理序列数据的同时关注所有位置,从而提高了处理长距离依赖关系的能力。DAB-DETR模型中采用了自注意力Transformer结构,自注意力机制(Self-attention,SA)是Transformer中最关键的内容,其计算方法为

$$\boldsymbol{D}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\rho(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}})\boldsymbol{V} \tag{1}$$

式中:ρ为归一化函数;$\boldsymbol{Q}$为查询向量,$\boldsymbol{K}$为键向量,$\boldsymbol{V}$为值向量,计算公式为

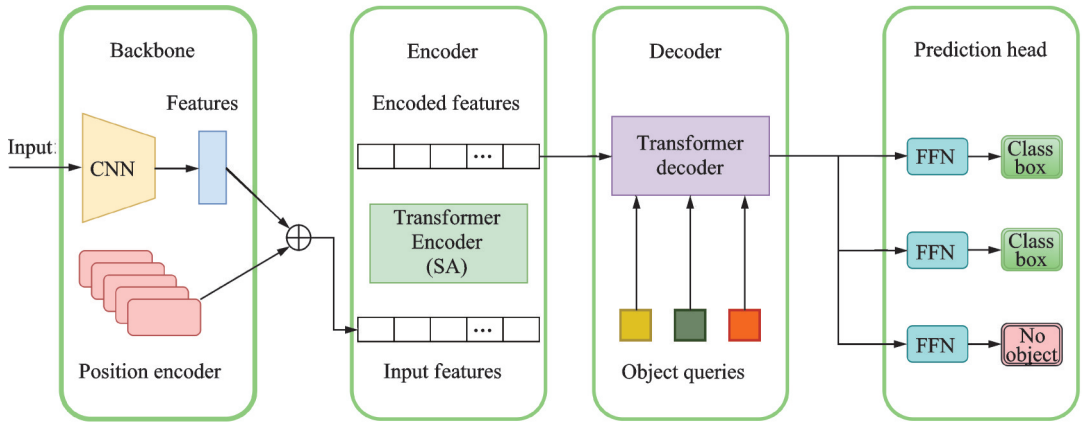


图1 DAB-DETR整体网络框架

Fig.1 Overall network structure of DAB-DETR

$$\boldsymbol{Q}=\boldsymbol{X}\boldsymbol{W}^{Q} \tag{2}$$

$$\boldsymbol{K}=\boldsymbol{X}\boldsymbol{W}^{K} \tag{3}$$

$$\boldsymbol{V}=\boldsymbol{X}\boldsymbol{W}^{V} \tag{4}$$

式中：$\boldsymbol{X}$为主干网络提取到的特征向量；$\boldsymbol{W}^{Q}$、$\boldsymbol{W}^{K}$、$\boldsymbol{W}^{V}$分别为$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$对应的权重矩阵。

从式(1)中可知SA包含了两个连续的矩阵乘法。第1个矩阵乘法($\boldsymbol{S}=\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}$)，计算不同位置之间的相似度，得到注意力矩阵$\boldsymbol{S}$。由于它计算每对位置之间的相似性，因此空间复杂度为$O(n^2)$。第2个矩阵乘法($\boldsymbol{D}=\boldsymbol{S}\boldsymbol{V}$)通过注意力矩阵乘法$\boldsymbol{S}$将值$\boldsymbol{V}$加权得到特征表示$\boldsymbol{D}$，其中矩阵$\boldsymbol{S}$表示不同像素之间的相似度信息。特征表示$\boldsymbol{D}$可以捕获不同位置之间的依赖关系，从而可以更好地表达输出。

1.3 损失函数

DAB-DETR通过综合计算分类损失和回归损失，构建代价矩阵。并在此基础上，使用匈牙利算法获取总代价最小的匹配方案，找到与真实框最匹配的预测框。最后，根据匹配结果计算损失值，对模型进行优化。

具体来说，首先将所有真实框y_i与每一个预测框$\hat{y}_i$两两匹配计算代价，得到真实框和预测框的代价矩阵，代价计算方法为

$$L_{\mathrm{match}}=-l_{\{c_i\neq\varnothing\}}\hat{p}_{\sigma(i)}(c_i)+l_{\{c_i\neq\varnothing\}}L_{\mathrm{box}}(b_i,\hat{b}_{\sigma(i)}) \tag{5}$$

式中：l为布尔函数；c_i和b_i分别为第i个真实框的类别和坐标信息；$\hat{p}_{\sigma(i)}(c_i)$为预测类别c_i的概率；$\sigma(i)$为匹配时某一种排列组合情况；$\hat{b}_{\sigma(i)}$为第i个预测框的坐标；L_{box}为回归损失，表达式为

$$L_{\mathrm{box}}=\lambda_{\mathrm{GIoU}}L_{\mathrm{GIoU}}(b_i,\hat{b}_{\sigma(i)})+\lambda_{\mathrm{L_1}}L_{\mathrm{L_1}} \tag{6}$$

式中：λ_{GIoU}和$\lambda_{\mathrm{L_1}}$分别为GIoU损失和$\mathrm{L_1}$损失的超参数；L_{GIoU}为GIoU损失函数；$L_{\mathrm{L_1}}$为$\mathrm{L_1}$损失函数。

在获取代价矩阵后，利用匈牙利算法得到每个真实框最优且唯一匹配的预测框$\hat{y}_{\hat{\sigma}(i)}$，并根据匹配结果计算分类损失和回归损失，得到最终的损失函数

$$\hat{\sigma}=\arg\min\sum_{i}^{N}L_{\mathrm{match}}(y_i,\hat{y}_{\sigma(i)}) \tag{7}$$

$$L_{\mathrm{Hungarian}}(y_i,\hat{y}_{\hat{\sigma}(i)})=\sum_{i=1}^{N}\left(-\log\hat{p}_{\hat{\sigma}(i)}(c_i)+l_{\{c_i\neq\varnothing\}}L_{\mathrm{box}}(b_i,\hat{b}_{\hat{\sigma}(i)})\right) \tag{8}$$

式中:$\hat{\sigma}(i)$为匹配代价最小的排列组合情况;N为预测结果数量。

2　MSDAB-DETR算法

MSDAB-DETR算法以DAB-DETR目标检测网络作为基线结构进行设计,如图2所示。相较于DAB-DETR算法,MSDAB-DETR算法在多尺度注意力融合模块、Transformer和损失函数3个方面进行了改进,改进部分在图中用红框标出,AFF(Attentional feature fusion)表示基于注意力的特征融合模块。

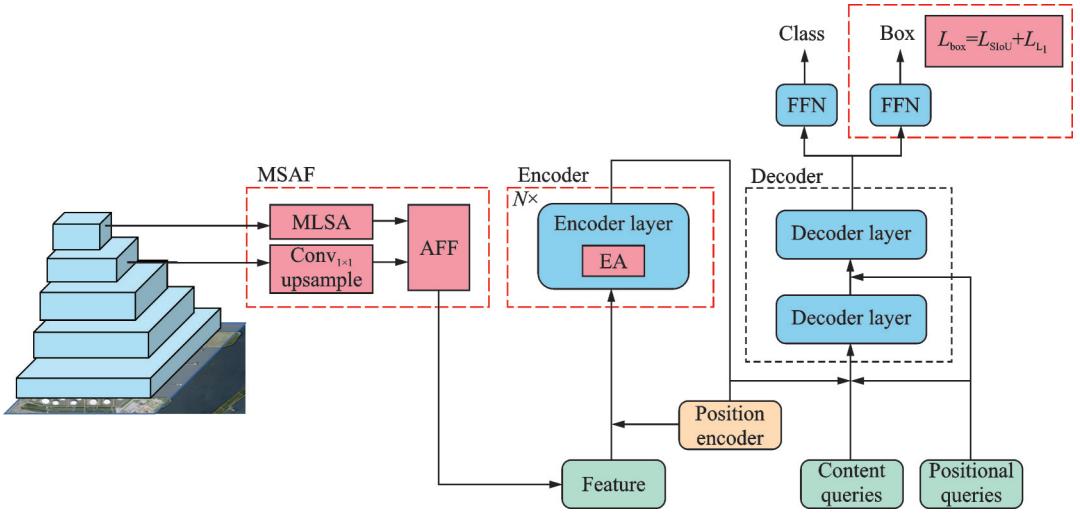


图2　MSDAB-DETR整体网络框架

Fig.2　Overall network structure of MSDAB-DETR

2.1　多尺度注意力融合模块

DAB-DETR的骨干网络采用逐渐减小特征空间大小的方式来提取最高级语义特征。对于大目标而言,其语义信息通常出现在低分辨率特征中;对于小目标,其细节信息则出现在高分辨率特征中。但随着网络的加深,小目标信息可能会丢失。为了充分挖掘不同分辨率特征所携带的特征信息,本文提出多尺度注意力融合模块(Multi-scale attention fusion, MSAF),以减少重要特征信息的损失,提高目标检测性能,其网络架构如图3所示。该模块通过针对多尺度问题设计的局部空间注意力模块(Multi-scale local spatial attention,MLSA)对比不同空间分辨率特征之间的差异,生成局部空间注意力系数图,以自适应强化高分辨率特征图中目标区域的显著性,从而检测出DAB-DETR所遗漏的小目标信息。图中MS-CAM(Multi-scale channel attention module)表示多尺度通道注意力模块。

具体来说,MSAF以ResNet50作为骨干网络,提取各个不同阶段输出的特征。由于前3层卷积的输出占用大量内存,因此提取Conv4和Conv5输出,并将它们表示为$\{C_4,C_5\}$。该模块首先使用1×1卷积将C_5的通道数量降低以便与C_4对齐,在执行逐元素相减之前,使用双线性插值操作对齐特征的空间大小,得到与C_4尺寸完全一致的新的特征图C_5',由此获取丢失的高分辨率特征信息。再经过Softmax激活函数进行非线性映射后生成局部空间注意力系数图,与C_4逐元素相乘,从而实现对C_4高分辨率特征信息的增强,生成具有局部空间区域显著特征的特征图C_6,表达式为

$$C_6=C_4\otimes \mathrm{soft\,max}\left(C_4-\mathrm{Conv}_{d_4}(\mathrm{Up}_{s_4}(C_5))\right) \tag{9}$$

式中:"$\otimes$"表示两个特征图对应元素相乘;Conv_{d_i}表示输出通道数为d_i的1×1卷积,d_i为特征C_i的通

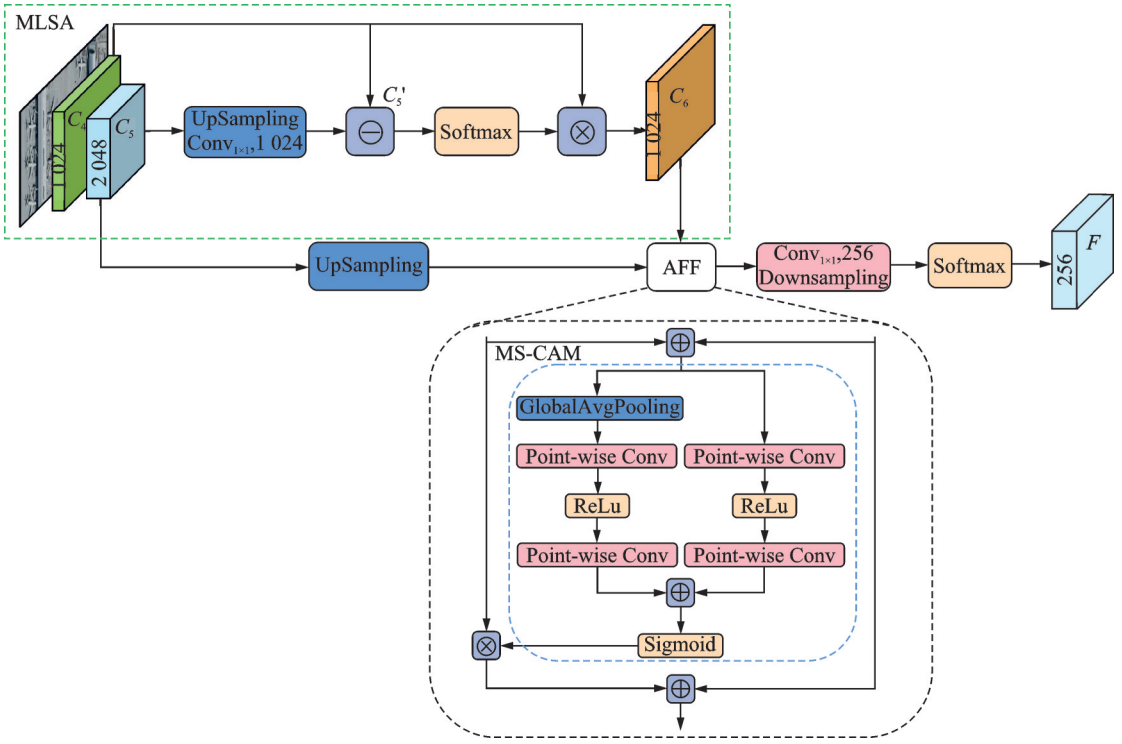


图3 多尺度增强注意特征融合模块

Fig.3 Multi-scale enhanced attention feature fusion module

道数;Up$_{s_i}$表示输出尺寸为s_j的上采样,s_j为特征C_i的分辨率。

为了在拥有低分辨率语义信息的同时也保有高分辨细节信息,将C_6和C_5进行融合得到具有更丰富表达能力的特征图F。在特征融合之前,对C_5进行上采样以便与C_6对齐,然后使用1×1卷积来减少通道数,防止引入更多的参数。相较于常用的特征融合处理方式(例如求和或拼接),这里采用AFF模块[33],该模块可以更好地融合语义和尺度不一致的特征,表达式为

$$\mathrm{AFF}(C_6, C_5)=M(C_6 \uplus C_5)\otimes C_6+(1-M(C_6 \uplus C_5))\otimes C_5 \tag{10}$$

式中:"$\uplus$"表示逐元素求和;M表示MS-CAM;符号"$\otimes$"表示逐元素乘法。

将$(C_6 \uplus C_5)$记作$\boldsymbol{X}$,$\boldsymbol{X}\in\mathbf{R}^{C\times H\times W}$,具有$C$个通道和大小为$H\times W$的特征映射,MS-CAM通过两种分支来分别提取出不同尺度的通道注意力,计算之后的权重值用来对输入特征$\boldsymbol{X}$逐元素相乘得到输出$\boldsymbol{M}$,$\boldsymbol{M}$与输入特征$\boldsymbol{X}$保持相同的分辨率,表达式为

$$\boldsymbol{M}(\boldsymbol{X})=\boldsymbol{X}\otimes\sigma(L(\boldsymbol{X})\oplus G(\boldsymbol{X})) \tag{11}$$

式中$L(\boldsymbol{X})$和$G(\boldsymbol{X})$分别为局部特征和全局特征的通道注意力,表达式为

$$L(\boldsymbol{X})=B\Big(\mathrm{PWConv}_2\big(\delta(B(\mathrm{PWConv}_1(\boldsymbol{X})))\big)\Big) \tag{12}$$

$$G(\boldsymbol{X})=B\Big(\mathrm{PWConv}_2\Big(\delta\Big(B(\mathrm{PWConv}_1(g(\boldsymbol{X})))\Big)\Big)\Big) \tag{13}$$

$$g(\boldsymbol{X})=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{X}_{[:,i,j]} \tag{14}$$

式中:PWConv$_1$和PWConv$_2$均为1×1卷积,PWConv$_1$用于将输入的特征$\boldsymbol{X}$的通道数减少,而

PWConv$_2$用于将通道数恢复到与原输入通道数相同;B表示BatchNorm层;δ表示ReLU激活函数。$G(\boldsymbol{X})$相较于$L(\boldsymbol{X})$多了一个全局平均池化操作$g(\boldsymbol{X})$。

2.2　高效注意力Transformer

遥感图像具有高分辨率特性[34-35],即具有更高的像素密度。在相同的图像尺寸下,遥感图像中包含的像素数量更多,每个像素都对应着一个相对较小的地理区域。例如,每个像素可能对应一个建筑物、车辆或船只等目标。由于DAB-DETR中的Transformer编码器使用的是SA,因此需要大量的内存和计算资源。在处理高分辨率图像时,会导致计算需求急剧增加,超出一般计算资源的承受范围,这限制了DAB-DETR在高分辨率遥感图像上的直接应用。因此,采用与SA等价的高效注意力机制EA,在实现相同性能的同时减少内存的占用。EA的定义为

$$E(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\rho_q(\boldsymbol{Q})\left(\rho_k(\boldsymbol{K}^{\mathrm{T}})\boldsymbol{V}\right) \tag{15}$$

式中ρ_q和ρ_k分别表示$\boldsymbol{Q}$和$\boldsymbol{K}$的归一化函数。

EA的主要思想是考虑到矩阵乘法的关联性,将$(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}})\boldsymbol{V}$的计算更改为$\boldsymbol{Q}(\boldsymbol{K}^{\mathrm{T}}\boldsymbol{V})$,并不影响其效果,还能将复杂度从$O(n^2)$降为$O(d_k,\ d_v)$。在实际情况下,$d_k\times d_v$的值明显小于$n^2$,从而使得基于EA提出的EAT能够在处理高分辨率图像输入时具有更好的可扩展性和较低的计算资源消耗。

2.3　损失函数设计

匈牙利算法利用代价矩阵计算寻找二分图匹配的最短路径,从而确定唯一的匹配结果。代价矩阵中细微的变化会导致结果产生显著差异。DAB-DETR中使用的GIoU[36]仅考虑了预测框和真实框的重叠率以及在最小闭包区域的比重。这种单一的损失计算方式无法全面反映预测框与真实框之间的差异,从而增大了代价损失的波动性。而这种波动性进一步加剧了预测框在训练过程中的不稳定性,进而影响算法的收敛速度和精度。为了提升模型检测性能,引入SIoU替代GIoU作为回归损失函数。

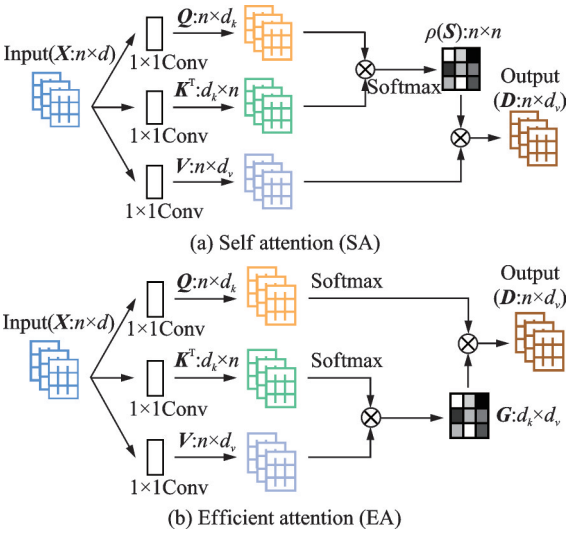


图4　SA与EA的对比

Fig.4　Comparison between SA and EA

SIoU损失主要包含4个部分,分别是IoU损失、形状损失、距离损失以及角度损失,其定义为

$$L_{\mathrm{SIoU}}=1-\mathrm{IoU}+\frac{\Omega+\Delta}{2} \tag{16}$$

式中:Ω为形状损失,通过计算真实框和预测框之间的宽高比来判断二者形状是否相似;Δ为距离损失,描述了真实框和预测框中心点之间的距离,损失与角度呈正相关,即

$$\Delta=\sum_{t=x,y}(1-\mathrm{e}^{\gamma\rho t}) \tag{17}$$

$$\rho_x=\left(\frac{b_{c_x}^{\mathrm{gt}}-b_{c_x}}{c_{\mathrm{W}}}\right)^2,\rho_y=\left(\frac{b_{c_y}^{\mathrm{gt}}-b_{c_y}}{c_{\mathrm{H}}}\right)^2,\gamma=2-\Lambda \tag{18}$$

式中:c_{H}和c_{W}为真实框和预测框最小外接矩形的宽和高;$(b_{c_x}^{\mathrm{gt}},b_{c_y}^{\mathrm{gt}})$和$(b_{c_x},b_{c_y})$分别为真实框和预测框的中心坐标;$\Lambda$为角度损失,表达式为

$$\Lambda = 1 - 2 \times \sin^2\left(\arcsin\left(\frac{c_h}{\sigma}\right) - \frac{\pi}{4}\right) \tag{19}$$

式中σ和c_h分别为真实框和预测框中心点的距离和高度差。

最终得到改进后的回归损失为

$$L_{box} = \lambda_{SIoU} L_{SIoU} + \lambda_{L_1} L_{L_1} \tag{20}$$

式中：λ_{SIoU}和λ_{L_1}分别为SIoU损失和L_1损失的超参数。

相较于GIoU，SIoU不仅考虑了真值框与预测框之间的重叠率，还进一步考虑了它们之间的方向、形状以及中心点之间的距离对匹配结果的影响。通过有效减少总自由度，即让预测框更快地接近或找到最近的轴，可以减弱二分图匹配的波动性，加速模型的收敛过程。二分图波动性减小后，预测框与真值框的匹配更为准确。这种准确的匹配会使得在计算最终损失函数时，回归精度得到提高，从而降低总损失。图5显示了DAB-DETR在分别使用GIoU和SIoU作为边界框损失函数时的损失曲线对比，共计训练50个轮次，在第25轮将学习率调低。从图5中可以看出，SIoU的损失值下降速度快于GIoU，这使得训练过程可以更早地达到收敛状态；此外，SIoU的损失值始终低于GIoU的损失值。这充分说明，使用SIoU作为损失函数不仅收敛速度更快，而且收敛效果也更好。

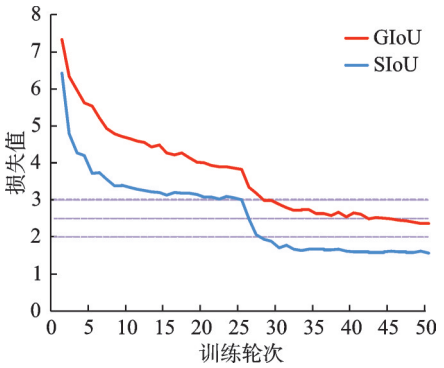


图5 不同损失函数的损失曲线

Fig.5 Loss curves for different loss functions

3 实验结果与分析

3.1 数据集

实验使用NWPU VHR-10数据集[37]和DIOR数据集[38]。NWPU VHR-10数据集广泛应用于遥感图像中的目标检测任务，共有10种不同类型的地理空间对象，包含从Google Earth获得的715幅彩色图像，以及从Vaihingen数据集中获得85幅锐化的彩色红外(Color infrared, CIR)图像，总共800幅高分辨率(Very high resolution, VHR)遥感图像。其中Google Earth图像的空间分辨率范围为0.5～2 m，CIR图像的空间分辨率为0.08 m。每一幅图像至少包含1个目标图像，一共含有3 775个实例。在原始数据集中，部分数据存在标注不完整的问题，导致这些数据无法直接用于实验。因此，对数据集进行数据清洗，最终获得了650幅图像，并将其用于实验。将NWPU VHR-10数据集按照9∶1的比例划分为训练集(585幅图像)和测试集(65幅图像)。数据集中各类别的图像和目标数如表1所示。

表1 NWPU VHR-10数据集中各类别的图像和目标数

Table 1 Numbers of images and targets for the categories in NWPU VHR-10 dataset

目标分类	图像数	目标数	原数据库标注
飞机	90	757	Airplane
舰船	57	302	Ship
油罐	26	655	Storage tank
棒球场	163	390	Baseball diamond
网球场	101	524	Tennis court
篮球场	76	159	Basketball court
田径场	151	163	Ground track field
港口	27	224	Harbor
桥梁	67	124	Bridge
汽车	86	477	Vehicle

DIOR数据集中有23 463幅航空图像，20个类别，共192 472个标注的实例目标。每幅图片的尺寸都是800像素×800像素，为了保证训练集和测试集中图像的一致性，DIOR数据集在使用中被随机划分到两个数据集中。其中，训练集包含11 725幅图像，测试集包含

11 738幅图像。数据集中各类别的图像如表2所示。

表2　DIOR数据集中各类别的图像

Table 2　Numbers of images for the categories in DIOR dataset

目标分类	图像数	原数据库标注
飞机	1 387	Airplane
机场	1 310	Airport
棒球场	2 440	Baseball field
篮球场	1 369	Basketball court
桥梁	2 176	Bridge
烟囱	854	Chimney
大坝	986	Dam
服务区	1 125	Expressway service area
收费站	1 218	Expressway toll station
高尔夫球场	946	Golf course
田径场	2 312	Ground track field
港口	1 474	Harbor
立交桥	2 019	Overpass
船舶	2 702	Ship
体育场	1 470	Stadium
储油罐	1 614	Storage tank
网球场	2 582	Tennis court
火车站	994	Train station
车辆	6 420	Vehicle
风车	1 616	Wind mill

3.2　参数设置

采用PyTorch开源深度学习框架搭建网络,在具有NVIDIA GeForce RTX 3090 GPU和24 GB显存的服务器上实现。由于从头训练DAB-DETR模型会耗费大量时间且对显存的要求比较高,因此使用预训练好的DAB-DETR模型对其训练参数进行微调,以节省时间和计算资源,提高效率。使用AdamW训练DAB-DETR,主干网络和其他模块的初始学习率分别设置为10^{-5}和10^{-4},权重衰减系数为10^{-4},batchsize设置为2。由于在实验中发现DAB-DETR和MSDAB-DETR模型收敛较快,因此训练的迭代次数设置为50轮,其他模型则设置为300轮。

3.3　评价指标

为了更好地评价检测方法的表现,实验采用了常用的平均精度(Average precision,AP)和平均召回率(Average recall,AR)作为检测性能的评价指标。检测结果可以分为3类,分别是正确预测的正样本(True positive,TP),错误预测的正样本(False positive,FP)以及错误预测的负样本(False negative,FN)。据此,精度P和召回率R定义为

$$\begin{cases} P=\dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \\ R=\dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \end{cases} \tag{21}$$

AP通过计算从召回率R的起点(即$R=0$,表示没有召回任何正例)到终点(即$R=1$,表示所有正例都被召回)的整个范围内精度P的平均值得到,更高的AP值意味着模型在该类别上的检测效果越好,反之亦然。在多类目标检测任务中,由于测试的数据集具有多个类别标签,而仅针对单个类别的AP值无法全面反映模型的整体性能。因此采用计算多个类别AP的平均值来评估模型性能,这个平均值常用均值平均精度(mean Average precision,mAP)表示。

3.4　对比实验

表3给出了MSDAB-DETR与其他目标检测模型在NWPU VHR-10数据集上的检测结果。与DETR等其他经典算法相比,DAB-DETR和MSDAB-DETR收敛所需训练周期更短。相比于其他模型,MSDAB-DETR模型拥有更好的目标检测性能,mAP值达到95.3%,在所有实验模型中最高。与Faster R-CNN模型以及SSD模型相比,MSDAB-DETR的mAP分别提高了6%和4.4%。除对轮船和港口的检测精度略低于上述两种模型外,其余目标的检测精度都有所提升,其中对于中小型目标的检

表3 NWPU VHR-10数据集中不同类别目标的检测精度对比

Table 3 Comparison of detection accuracy of different categories of targets in NWPU VHR-10 dataset

算法	RetinaNet	Faster R-CNN	SSD	Faster R-CNN with FPN	YOLOv5	TRD	DETR	DAB-DETR	本文算法
迭代次数	300	300	300	300	300	—	300	50	50
飞机	0.895	0.990	0.908	0.993	0.994	0.994	**1.000**	**1.000**	**1.000**
舰船	0.682	0.896	0.932	0.827	0.941	0.782	0.926	**0.943**	0.931
油罐	0.884	0.839	0.949	0.871	0.955	0.844	0.925	0.978	**0.980**
棒球场	0.959	0.993	0.987	0.979	0.943	0.942	0.965	**1.000**	**1.000**
网球场	0.842	0.892	0.940	0.940	0.937	0.820	0.946	0.902	**0.948**
篮球场	0.692	0.809	0.948	0.954	0.855	0.839	**1.000**	0.993	0.978
田径场	0.938	0.987	0.897	0.938	0.945	0.989	0.982	0.991	**0.995**
港口	0.856	0.963	0.797	**0.978**	0.846	0.784	0.646	0.838	0.864
桥梁	0.909	0.800	0.925	0.878	0.894	0.569	0.924	0.869	**0.988**
汽车	0.581	0.764	0.806	0.771	0.863	0.722	0.867	**0.874**	0.843
mAP	0.824	0.893	0.909	0.913	0.917	0.829	0.918	0.939	**0.953**

测精度提升尤为明显。例如,相比于SSD的飞机检测精度,MSDAB-DETR提升了9.2%。与YOLOv5相比,MSDAB-DETR的mAP也有3.6%的提升,除舰船和汽车之外的其他类别的检测精度值也都有所提高。相较于Faster R-CNN with FPN,MSDAB-DETR的mAP提高了4%,对港口以外的其他9种类别的检测精度均表现出更高的准确度。与DETR和DAB-DETR相比,MSDAB-DETR的mAP则分别提高了3.5%和1.4%,其中与DAB-DETR相比,对于油罐、网球场、田径场、港口和桥梁这些目标分别提高了0.2%、4.6%、0.4%、2.6%和11.9%。MSDAB-DETR的算法与采用相同Transformer架构的TRD[26]算法相比,mAP提升了12.4%,并且在各类别的检测精度上均表现出色,明显优于TRD算法。

为了进一步验证MSDAB-DETR的有效性,在DIOR数据集上与其他目标检测模型进行对比[38]。检测结果如表4所示,所提MSDAB-DETR的mAP值为71.5%。在所有实验模型中,MSDAB-DETR模型展现出了卓越的目标检测性能,达到了最高水平。与Faster R-CNN模型以及SSD模型相比,MSDAB-DETR的mAP分别提高了17.4%和12.9%。其中对于中小型目标的检测精度提升尤为明显。例如,相比于Faster R-CNN的飞机检测精度,MSDAB-DETR提升了25.1%。相较于Faster R-CNN with FPN,MSDAB-DETR的mAP提高了8.4%,对船舶和储油罐以外的其他18种类别的检测精度均表现出更高的准确度。与YOLOv3相比,MSDAB-DETR的mAP也有14.4%的提升,除舰船、储油罐和汽车之外的其他类别的检测精度值也都有提高。与DAB-DETR相比,MSDAB-DETR的mAP则提高了5.3%,除少数类型检测精度略低于上述其他模型外,如高尔夫球场和体育场,MSDAB-DETR在其余目标,尤其是小尺度目标上均有所提升。例如对飞机、桥梁、港口、船舶和车辆的检测精度分别提升了7.6%、8.2%、9.5%、10.8%和8.5%。

表5展示了在NWPU VHR-10数据集上,MSDAB-DETR与其他目标检测算法在处理不同尺度目标时的性能对比。评估算法性能的主要指标采用AP和AR,对于小、中、大3种尺度目标,具体分为AP_s、AP_m、AP_l、AR_s、AR_m和AR_l。其中,小目标是指尺寸小于32像素×32像素的目标,中目标是像素值范围在32像素×32像素到96像素×96像素之间的目标,而大于96像素×96像素的目标则被视为大目

表4　DIOR数据集中不同类别目标的检测精度对比

Table 4　Comparison of detection accuracy of different categories of targets in DIOR dataset

算法	Faster R-CNN	SSD	Faster R-CNN with FPN	YOLOv3	DAB-DETR	本文算法
飞机	0.536	0.595	0.541	0.722	0.711	**0.787**
机场	0.493	0.727	0.714	0.292	0.859	**0.885**
棒球场	0.788	0.724	0.633	0.740	0.755	**0.783**
篮球场	0.662	0.757	0.810	0.786	0.871	**0.887**
桥梁	0.280	0.297	0.426	0.312	0.384	**0.466**
烟囱	0.709	0.658	0.725	0.697	0.808	**0.830**
大坝	0.623	0.566	0.575	0.269	0.710	**0.755**
服务区	0.690	0.635	0.687	0.486	0.771	**0.836**
收费站	0.552	0.531	0.621	0.544	0.613	**0.696**
高尔夫球场	0.680	0.653	0.731	0.311	**0.827**	0.815
田径场	0.569	0.686	0.765	0.611	0.810	**0.842**
港口	**0.502**	0.494	0.428	0.449	0.395	0.490
立交桥	0.501	0.481	0.560	0.497	0.562	**0.610**
船舶	0.277	0.592	0.718	**0.874**	0.458	0.566
体育场	0.730	0.610	0.570	0.706	**0.735**	0.729
储油罐	0.398	0.466	0.535	**0.687**	0.372	0.457
网球场	0.752	0.763	0.812	0.873	0.844	**0.875**
火车站	0.386	0.551	0.530	0.294	0.630	**0.698**
车辆	0.236	0.274	0.431	**0.483**	0.349	0.434
风车	0.454	0.657	0.809	0.787	0.775	**0.863**
mAP	0.541	0.586	0.631	0.571	0.662	**0.715**

表5　不同尺度目标的精度和召回率对比

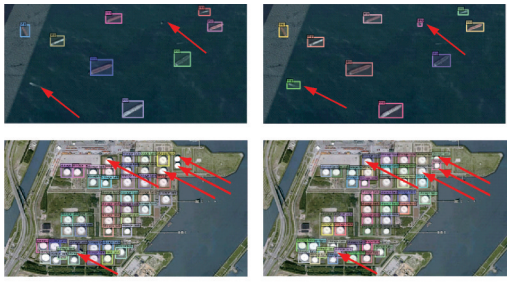
Table 5　Comparison of precision and recall rate of targets at different scales

算法	mAP	AP_s	AP_m	AP_l	AR_s	AR_m	AR_l
Faster R-CNN with FPN	0.913	0.432	0.543	0.573	0.456	0.579	0.621
DETR	0.918	0.406	0.549	0.592	0.438	0.583	0.644
DAB-DETR	0.939	0.485	0.584	0.616	0.487	0.656	0.673
本文算法	**0.953**	**0.590**	**0.602**	**0.643**	**0.621**	**0.669**	**0.715**

标。从表5中可以看出,相较于其他目标检测算法,MSDAB-DETR在各项评估指标上均取得了卓越的性能。具体来说,与Faster R-CNN with FPN相比,MSDAB-DETR在AP_s和AR_s上分别提升了15.8%和16.5%;在AP_m和AR_m上分别提升了5.9%和9%;在AP_l和AR_l上分别提升了7%和9.4%。与DETR相比,MSDAB-DETR在AP_s和AR_s上分别提升了18.4%和18.3%;在AP_m和AR_m上分别提升了5.3%和8.6%;在AP_l和AR_l上分别提升了5.1%和7.1%。与DAB-DETR相比,MSDAB-DETR在AP_s和AR_s上分别提升了10.5%和13.4%;在AP_m和AR_m上分别提升了1.8%和1.3%;在AP_l和AR_l上分别提升了2.7%和4.2%。这些结果表明,MSDAB-DETR在检测不同尺度目标时展现出更高的精度和召回率,尤其是在小目标检测方面,提升更为明显。说明该方法对不同尺度目标的适应性更强,能够更好

地应对尺度变化带来的挑战，而且还能有效降低误检和漏检的发生率。

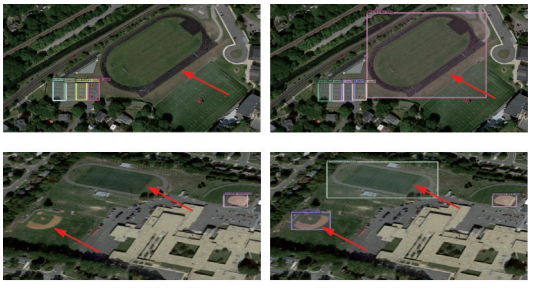
图6为DAB-DETR和MSDAB-DETR两种方法对小目标的可视化对比。从图6的2个图例可以看出，DAB-DETR对于较小的目标检测出现了漏检，尤其对于复杂环境，该现象更为明显；而MSDAB-DETR则较小发生漏检问题。此外，实验中发现，DAB-DETR在处理存在大量目标且目标尺度大小不均等的图片时会产生漏检。图7展示了一组针对多尺度目标检测的实验对比结果。可以看出，DAB-DETR无法检测到大尺度目标（如田径场）和部分小尺度目标（如棒球场）；而MSDAB-DETR能够准确检测出全部的大尺度和小尺度目标，并能够进行准确分类，对多尺度目标的检测能力更强。



(a) DAB-DETR (b) MSDAB-DETR

图6 DAB-DETR和MSDAB-DETR的小目标可视化对比

Fig.6 Small target visualization comparison between DAB-DETR and MSDAB-DETR



(a) DAB-DETR (b) MSDAB-DETR

图7 DAB-DETR和MSDAB-DETR的多尺度可视化对比

Fig.7 Multi-scale visualization comparison between DAB-DETR and MSDABDETR

3.5 消融实验

为了评估MSAF、EAT和SIoU三者对遥感图像目标检测性能的影响，在NWPU VHR-10数据集上从检测速度、参数量、计算量、GPU内存和mAP五个方面进行实验比较，实验结果如表6所示。从表6中可以看出，在DAB-DETR模型的基础上加入MSAF模块后，只需要增加2.4 MB的参数和2.07 GB的计算量，mAP就可以提高1.1%，这归功于MSAF可以获得更多的目标区分特征，从而提高目标检测性能。在DAB-DETR模型的基础上加入EAT后，其参数量和计算量与DAB-DETR相同，检测精度也基本保持一致。但检测速度提高了0.5帧/s，并且内存占用量减少了0.61 GB，约为原DAB-DETR内存占用量的9%。这表明，采用EAT可以在保持与采用自注意力Transformer的DAB-DETR模型相似性能的同时，减少检测时间和内存消耗，实现更高的效率。在损失函数中采用SIoU后，相对于DAB-DETR模型，其参数量和计算复杂度不变，检测速度略有提升，而mAP提高了0.9%，这得益于SIoU比GIoU更完善，降低了二分图匹配的波动性，并减小了边界框回归误差。综上所述，MSAF和

表6 单独加入不同模块的消融实验

Table 6 Ablation experiments of individually adding different modules

算法	MSAF	EAT	SIoU	检测速度/(帧·s^{-1})	参数量/MB	计算量/GB	GPU内存/GB	mAP
DAB-DETR				28.5	41.4	64.96	6.73	0.939
DAB-DETR+MSAF	√			27.9	43.8	67.03	6.66	0.950
DAB-DETR+EAT		√		29.0	41.4	64.96	6.12	0.938
DAB-DETR+SIoU			√	28.6	41.4	64.96	6.74	0.948
本文算法	√	√	√	28.6	43.8	67.03	6.12	**0.953**

SIoU在提升目标检测性能方面表现出色,而EAT则有效地降低了内存占用量,提高了检测速度。

为了进一步评估MSAF、EAT和SIoU对遥感图像目标检测性能的影响,设置了不同的模块组合。对比在不同IoU阈值和不同大小尺度下的平均检测精度,实验结果如表7所示。其中,AP_{50}和AP_{75}是评估模型在不同IoU阈值的召回率和误检率之间平衡的指标,ξ为MSDAB-DETR与DAB-DETR各个指标的提升比例。从表7可以看出,除了在AP_m指标上与DAB-DETR+MSAF+SIoU算法几乎持平之外,MSDAB-DETR的实验结果均优于其他所有方法。与MSDAB-DETR相比,DAB-DETR+MSAF+EAT算法的AP_{50}和AP_{75}分别降低了0.8%和1.7%,AB-DETR+MSAF+SIoU算法分别降低了0.4%和0.5%,而DAB-DETR+EAT+SIoU算法则分别降低了0.7%和0.3%,可见SIoU对于检测精度的影响更大。另一方面,与MSDAB-DETR相比,DAB-DETR+MSAF+EAT算法的AP_s、AP_m和AP_l分别降低了4.8%、1%和1%,DAB-DETR+MSAF+SIoU算法分别降低了2.3%、−0.1%和0.6%,而DAB-DETR+EAT+SIoU算法则降低了4.4%、0.5%和1.8%,这表明MSAF和SIoU对小目标的检测精度更为重要。

表7　不同模块组合的消融实验

Table 7　Ablation experiments of different module combinations

算法	MSAF	EAT	SIoU	$AP_{0.5}$	$AP_{0.75}$	AP_s	AP_m	AP_l
DAB-DETR				0.939	0.664	0.485	0.584	0.616
DAB-DETR+MSAF+EAT	√	√		0.945	0.683	0.542	0.592	0.633
DAB-DETR+MSAF+SIoU	√		√	0.949	0.695	0.567	**0.603**	0.637
DAB-DETR+EAT+SIoU		√	√	0.946	0.697	0.546	0.597	0.625
本文算法	√	√	√	**0.953**	**0.700**	**0.590**	0.602	**0.643**
ξ/%				+1.4	+3.6	+10.5	+1.8	+2.7

4　结束语

针对现有目标检测算法对于尺度多样的遥感图像检测精度不高的问题,提出了一种改进DAB-DETR的多尺度遥感目标检测算法MSDAB-DETR。首先,设计了一种多尺度注意力特征融合模块,利用不同尺度特征的差异来加强浅层特征,并通过基于注意力的特征融合模块与深层特征融合,实现遥感图像多尺度目标的预测。其次,采用高效注意力机制的Transformer,在拥有原模型相当的全局特征学习能力的同时降低计算复杂度,加快检测速度。最后,引入SIoU损失函数作为边界框回归损失函数,提高了检测准确率。实验结果表明,该算法具有较好的检测效果,实现了对多尺度遥感图像的高精度检测。

参考文献:

[1]　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.
NIE Guangtao, HUANG Hua. A survey of object detection in optical remote sensing images[J]. Acta Automatica Sinica, 2021, 47(8): 1749-1768.

[2]　廖育荣, 王海宁, 林存宝, 等. 基于深度学习的光学遥感图像目标检测研究进展[J].通信学报, 2022,43(5): 190-203.
LIAO Yurong, WANG Haining, LIN Cunbao, et al. Research progress of deep learning-based object detection of optical remote sensing image[J]. Journal of Communications,2022,43(5): 190-203.

[3]　REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016: 779-788.

[4] LAW H, DENG J. Cornernet: Detecting objects as paired keypoints[C]//Proceedings of the European Conference on Computer Vision (ECCV). [S.l.]: Springer International Publishing, 2018: 734-750.

[5] DUAN K, BAI S, XIE L, et al. CenterNet: Keypoint triplets for object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2019: 6569-6578.

[6] TIAN Z, SHEN C, CHEN H, et al. FCOS: Fully convolutional one-stage object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2019: 9627-9636.

[7] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with Transformers[C]//Proceedings of Computer Vision—ECCV 2020: 16th European Conference. Glasgow, UK: Springer International Publishing, 2020: 213-229.

[8] 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. [S.l.]: IEEE, 2014: 580-587.

[9] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. Advances in Neural Information Processing Systems, 2015. DOI: 10.1109/ICCV.2015.169.

[10] 韩松臣, 张比浩, 李炜, 等. 基于改进Faster-RCNN的机场场面小目标物体检测算法[J]. 南京航空航天大学学报, 2019, 51(6): 735-741.
HAN Songchen, ZHANG Bihao, LI Wei, et al. Small target detection in airport scene via modified Faster-RCNN[J]. Journal of Nanjing University of Aeronautics and Astronautics, 2019, 51(6): 735-741.

[11] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multi-box detector[C]//Proceedings of Computer Vision—ECCV 2016: 14th European Conference. Amsterdam, The Netherlands: Springer International Publishing, 2016: 21-37.

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

[13] REDMON J, FARHADI A. YOLOv3: An incremental improvement[EB/OL]. (2018-04-08)[2023-11-30]. http://arxiv.org/abs/1804.02767?context=cs.LG.html.

[14] BOCHKOYSKIY A, WANG C Y, LIAO H Y M. YOLOv4: Optimal speed and accuracy of object detection[EB/OL]. (2021-04-23)[2023-11-25]. https://arxiv.org/abs/2004.10934.

[15] JOCHER G. Yolov5[EB/OL]. (2020-08-10). https://github.com/ultralytics/yolo-v5.

[16] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE International Conference on Computer Vision. [S.l.]: IEEE, 2017: 2980-2988.

[17] CAI Z, VASCONCELOS N. CASCADE R-CNN: Delving into high quality object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018: 6154-6162.

[18] 阎菩提, 邱实, 岳程斐. 结合局部高清图像的遥感集群目标区域超分辨率重建[J]. 南京航空航天大学学报, 2023, 55(6): 956-965.
YAN Puti, QIU Shi, YUE Chengfei.Remote sensing image super-resolution reconstructure with local high-resolution clustered object images[J]. Journal of Nanjing University of Aeronautics and Astronautics, 2023, 55(6): 956-965.

[19] ZHANG F, WANG X, ZHOU S, et al. DARDet: A dense Anchor-free rotated object detector in aerial images[EB/OL]. (2021-10-26)[2023-11-25]. https://ieeexplore.ieee.org/abstract/document/9585487.

[20] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature Pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2017: 2117-2125.

[21] YANG X, YANG J, YAN J, et al. SCRDet: Towards more robust detection for small, cluttered and rotated objects[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2019: 8232-8241.

[22] HOU J, MA H, WANG S. Parallel cascade R-CNN for object detection in remote sensing imagery[J]. Journal of Physics, 2020, 1544(1): 012124.

[23] FU K, CHANG Z, ZHANG Y, et al. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 161: 294-308.

[24] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017. DOI: 10.48550/arXiv.1706.03762.

[25] 周丽娟, 毛嘉宁. 视觉 Transformer识别任务研究综述[J]. 中国图象图形学报, 2023, 28(10): 2969-3003.
ZHOU Lijuan, MAO Jianing. Vision Transformer-based recognition tasks: A critical review[J]. Journal of Image and Graphics, 2023, 28(10): 2969-3003.

[26] LI Q, CHEN Y, ZENG Y. Transformer with transfer CNN for remote-sensing-image object detection[J]. Remote Sensing, 2022, 14(14): 984.

[27] SHEN C, MA C, GAO W. Multiple attention mechanism enhanced YOLOX for remote sensing object detection[J]. Sensors, 2023, 23(3): 1261.

[28] LI F, ZHANG H, LIU S, et al. DN-DETR: Accelerate DETR training by introducing query denoising[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2022: 13619-13627.

[29] ZHANG H, LI F, LIU S, et al. DINO: DETR with improved denoising anchor boxes for end-to-end object detection[EB/OL]. (2022-04-27)[2023-11-25]. https://arxiv.org/abs/2203.03605.

[30] LIU S, LI F, ZHANG H, et al. DAB-DETR: Dynamic anchor boxes are better queries for DETR[EB/OL]. (2022-01-28)[2023-11-25]. https://arxiv.org/abs/2201.12329.

[31] SHEN Z, ZHANG M, ZHAO H, et al. Efficient attention: Attention with linear complexities[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. [S.l.]: IEEE, 2021: 3531-3539.

[32] GEVORGYAN Z. SIoU loss: More powerful learning for bounding box regression[EB/OL]. (2022-05-25)[2023-11-25]. https://arxiv.org/abs/2205.12740.

[33] DAI Y, GIESEKE F, OEHMCKE S, et al. Attentional feature fusion[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. [S.l.]: IEEE, 2021: 3560-3569.

[34] MA X. High-resolution image compression algorithms in remote sensing imaging[J]. Displays, 2023, 79(4): 102462.

[35] 冈萨雷斯. 数字图像处理[M]. 第3版. 北京: 机械工业出版社, 2021.
GONZALEZ R C. Digital image processing[M]. 3rd ed. Beijing: China Machine Press, 2021.

[36] REZATOFIGHI H, TSOI N, GWAK J Y, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019: 658-666.

[37] CHENG G, HAN J. A survey on object detection in optical remote sensing images[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2016, 117: 11-28.

[38] LI K, WAN G, CHENG G, et al. Object detection in optical remote sensing images: A survey and a new benchmark[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 159: 296-307.

作者简介:

李烨(1974-),男,博士,硕士生导师,研究方向:机器学习、图像处理、移动通信,E-mail: liye@usst.edu.cn。

周生翠(1998-),**通信作者**,女,硕士研究生,研究方向:深度学习与图像处理,E-mail: zhousc9805@163.com。

张驰(1998-),男,硕士研究生,研究方向:深度学习与图像处理。

(编辑:张黄群)