面向遥感目标检测的无锚框Transformer算法

喻九阳,胡天豪,戴耀南,张德安,夏文凤
(武汉工程大学机电工程学院湖北省绿色化工装备工程技术研究中心,湖北武汉 430205)

摘 要: 遥感图像目标具有多方向排布、小且密集等特性,使基于深度学习的旋转目标检测算法存在检测精度不佳的问题. 针对这一问题,本文提出了一种面向遥感目标检测的无锚框Transformer算法. 首先,采用层次化Transformer采集不同分辨率的特征信息以扩大特征信息的采集范围. 其次,构建一种新的前馈网络(Spacial-FeedForward Neural network,SFFN). SFFN将3×3深度可分离卷积的局部空间特性和多层感知机(MultiLayer Perceptron,MLP)的全局通道特性融合在一起,以解决前馈网络(Feed Forward Neural network,FFN)在局部空间建模上的不足. 最后,基于SFFN架构搭建了无锚框检测器,将预测框回归问题分为水平框与旋转框,缓解了旋转框的损失不连续性问题. 在DOTA数据集上的测试结果表明,此方法的平均精度达到了75.83%,同时在NWPU VHR-10数据集上5类小目标检测结果达到了92.47%,在遥感目标检测精度上更具竞争力.

关键词: 遥感图像;目标检测;Transformer算法;无锚框检测器
基金项目: 湖北省重点研发计划(No.2020BAB030);湖北省自然科学基金(No.2023AFC010)
中图分类号: TP391;TP183 **文献标识码:** A **文章编号:** 0372-2112(2023)11-3238-10
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220612

Anchor-Free Transformer Algorithm for Aerial Remote Sensing Target Detection

YU Jiu-yang, HU Tian-hao, DAI Yao-nan, ZHANG De-an, XIA Wen-feng
(*Hubei Provincial Engineering Technology Research Center of Green Chemical Equipment*, *School of Mechanical and Electrical Engineering*, *Wuhan Institute of Technology*, *Wuhan*, *Hubei* 430205, *China*)

Abstract: Aerial remote sensing image targets have the characteristics of multi-directional arrangement, small, and dense. The rotating target detection algorithm based on deep learning has the problem of poor detection accuracy. To solve this problem, the article proposes a novel anchor-free Transformer algorithm for aerial remote sensing target detection. Firstly, hierarchical Transformer is used to collect feature information of different resolutions to improve the range of feature information collection. Secondly, a new feedforward network (Spacial-FeedForward Neural network, SFFN) is constructed. SFFN combines the local space characteristics of 3×3 depth separable convolution with the global channel characteristics of multi-layer perceptron (MLP) to solve the shortcomings of feed forward neural network (FFN) in local space modeling. Finally, an anchor-free detector is built based on SFFN architecture, and the regression problem of prediction frame is divided into horizontal frame and rotating frame, which alleviates the loss discontinuity problem of rotating frame. The test results on DOTA dataset show that the average accuracy of this method has reached 75.83%, respectively, while achieving 92.47% of 5 small targets on NWPU VHR-10 dataset, which is more competitive in remote sensing target detection accuracy.

Key words: remote sensing image; target detection; Transformer algorithm; anchor-free detector
Foundation Item(s): Hubei Provincial Key R&D Project (No.2020BAB030); Natural Science Foundation of Hubei Province (No.2023AFC010)

1 引 言

随着计算机视觉与遥感监测技术的发展,基于深度学习的目标检测方法在航空图像上的应用越来越广泛. 目前,针对遥感图像的目标检测主要借助深度卷积神经网络(Deep Convolution Neural Network,DCNN)来完成,主体上遵循“特征提取+边框回归”的设计思路,

收稿日期:2022-05-27;修回日期:2022-09-02;责任编辑:朱梅玉

取得了非常不错的检测效果[1]. 然而,目前典型的DCNN目标检测方法,无论是单阶段的方法(如SSD(Single Shot multibox Detector)、YOLO(You Only Look Once)系列、RetinaNet(ResNet+FPN+FCN网络)),还是Faster-RCNN(Faster Regions with CNN features)以及在其基础上改进的Cascade RCNN等两阶段方法,均未在遥感目标检测上达到精度与速度的平衡[2,3]. 原因主要包括以下三点:(1)一般卷积是从图像的固定位置采集特征,而遥感图像的排列分布不均,难以建立几何模型[4];(2)常规的多尺度特征金字塔网络(Feature Pyramid Network,FPN)[5]在面对目标尺度变化范围较大的遥感图像时,难以充分提取上下层间的语义信息[6];(3)遥感图像中密集排布大量小目标,在经过多次下采样后特征难以提取.

为了解决这一问题,研究人员相继提出了RRPN(Rotation Region Proposal Network)[7], R^{2}CNN(Rotational Region CNN)[8], RoI-Transformer[9], SCRDet[10], R^{3}Det(Refined Rotation RetinaNet)[11]等算法去实现旋转目标的检测. 然而,这些算法主要基于五参数回归的任意定向方法来预测角度,忽略了边界的损失不连续性,导致物体角度预测不准确. 肖进胜等[12]以角度预测为分类方式,采用圆形平滑标签处理分布函数,缓解了五参数法边界不连续的问题.

O^{2}DETR[13]是首个将Transformer应用到遥感图像旋转目标检测上的模型. 该模型提出了基于方向自适应的锚框检测器,同时用深度可分离卷积代替原始Transformer中的注意力机制[14],降低了Transformer对多尺度特征的计算复杂度,提高了目标检测效率. 何林远等人[15]提出一种基于稀疏Transformer的遥感旋转目标检测方法,利用K-means算法更好地提取稀疏域下的目标特征. 祝星馗等人[16]提出在YOLOv5网络基础上结合Transformer与CNN结构,同时使用加权双向特征金字塔网络,提高对小目标的检测能力.

Transformer及其改进算法在遥感航空目标检测上虽然取得了长足的进步,但仍存在以下不足:(1)Transformer的注意力机制存在位置不变的特性,若采用Transformer作为主干网络,则目标物需重新编码位置信息,导致模型的复杂度有所增加;(2)锚框检测器存在角度边界值的不连续性,需要一种消除边界不连续性的方法来解决.

为解决上述问题,本文提出一种面向遥感目标检测的无锚框Transformer算法. 首先,采用以Transformer为模型主干,避免了DCNN的复杂性. 此外,针对目标多尺度特性,采用层次化模块以分辨图像的不同特征信息,从而扩大特征信息的采集范围. 其次,设计一种新的前馈神经网络(Spacial-FeedForward Neural network,SFFN),将3×3深度可分离卷积(Depthwise Separable Convolution,DSC)和多层感知器(MultiLayer Perceptron,MLP)相结合,在不需要位置编码,并保证模型参量基本不变的前提下,解决前馈网络(Feed Forward Neural network,FFN)在局部空间建模不足的问题. 最后,基于SFFN架构搭建无锚框检测器(anchor-free),在旋转框接近水平时回归水平框,以缓解旋转框边界不连续的问题.

2　本文方法

2.1　整体网络框架

本文的整体网络框架如图1所示. 该网络架构主要是由Backbone,Neck和Heads三部分组成. 首先,Backbone模块采用层次化Transformer对图像进行特征提取,以降低注意力机制的计算复杂度,避免了DCNN复杂的结构. 其次,Neck结构通过所设计的SFFN结构增强了图像的多级特征信息,提高了FFN空间局部建模能力. 最后,Heads结构采用无锚框八参数回归方法对图像进行预测和检测,以降低常用锚框检测器结构的复杂性,并缓解旋转框的损失不连续性问题[17].

2.2　层次化主干网络

Transformer目标检测网络大多都没有考虑图像的内容信息[18],这是由于在遥感图像中,图像的分辨率大,目标小而密集,而Transformer目标检测模型仍采用固定大小(16×16)图像分割法(patch partition),这使模型在经历下采样之后,图像块中目标物的占比很小,非常容易丢失待检测的目标.

不同于Transformer全程使用大小固定的图像分割法,针对密集的小目标可使用较小的图像块,为此,本文将输入大小为(H×W×3)的图片分割成(4×4)大小的图像块,并将其嵌入图像块编码(patch embedding)中.

针对密集目标检测,本文通过层次化分层设计将Transformer分成4个不同分辨率的层级,其目的是将特征金字塔引入Transformer中,从而生成多尺度感受野用于密集目标的检测任务. 图像特征分辨率从第1层到第4层,逐次下降,依次为$\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C_i$, $i\in\{1,2,3,4\}$, C_{i+1}大于C_i. 每个分层结构是由结构相同但数量不同的注意力单元组成,不同层之间采用图像块合并(patch merging)进行下采样操作以缩小分辨率,通过通道数的调整实现结构的层次化,层次化主干网络如图2所示.

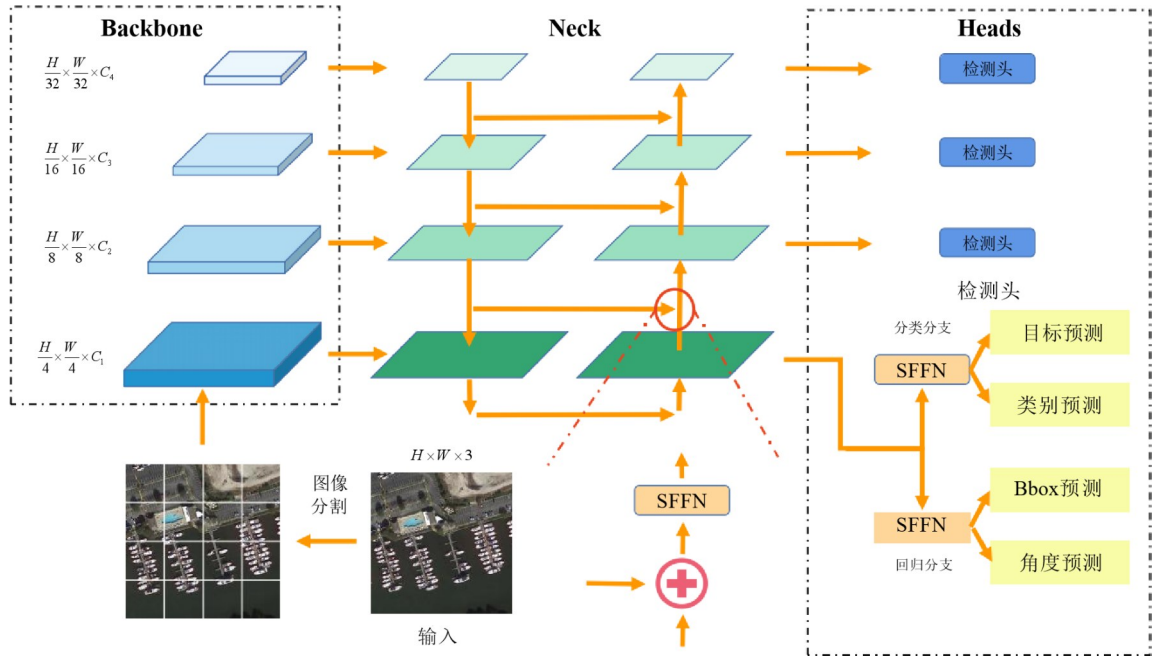

图1 无锚框Transformer算法结构

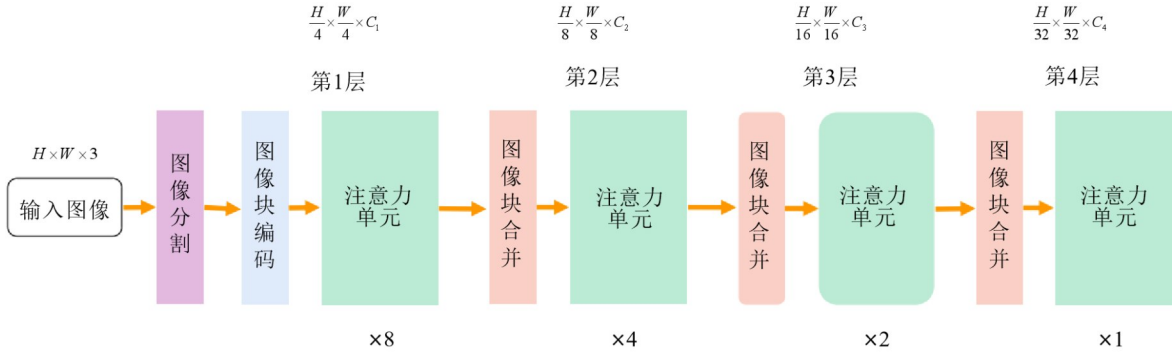

图2 层次化主干网络

2.3 注意力单元

传统的Transformer采用多头自注意力机制模型，包括scaled dot-product attention和multi-head attention[19]，就长距离捕捉信息能力而言，多头自注意力机制模型（图3）和递归神经网络（Recursive Neural Network，RNN）相当. 其表达式如下：

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}\left(\frac{\boldsymbol{K}^{\mathrm{T}}\boldsymbol{Q}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{1}$$

$$\text{MuitiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Concat}(\text{head}_1,\cdots,\text{head}_h)\boldsymbol{W}^o \tag{2}$$

其中，$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$分别是查询向量序列，键向量序列和值向量序列（Query，Key，Value）；$\text{head}_i=\text{Attention}(\boldsymbol{QW}_i^Q,\boldsymbol{KW}_i^K,\boldsymbol{VW}_i^V)$；$\sqrt{d_k}$对内积进行缩放，避免softmax的结果非0即1.

图3中，位置编码（position embedding）、多头注意力、FFN组成为一个Transformer注意力单元块. 多头自注意力机制模型在应用中需提供位置编码来确定输入的位置信息，同时需要结合残差连接上下文信息，并且存在头部冗余，从而加大了模型的计算复杂度. 为解决这一问题，受深度可分离卷积（Depthwise Separable Convolution，DSC）与Twins-SVT注意力机制的启发[20]，本文的注意力机制将注意力单元分为两部分改进，注意力单元结构如图4所示.

2.3.1 全局子采样注意力

对每个局部注意力单元添加全局注意力层可降低自注意机制的复杂度，但该方法会使时间的复杂度增加到$O((k_1\times k_2\times m\times n)^2\times d))$，故采用全局子采样注意（Global Sub-Sampled Attention，GSSA）方法，通过sub-sampling函数来将整个过程（局部+全局注意力的过程）的时间复杂度降低到$O\left(\left(\frac{H^2W^2d}{k_1k_2}\right)^2+k_1k_2HWd\right)$[21]. 其中，$m$和$n$表示为感受野被分成了$(m\times n)$个子窗口，$k_1=H/m$，$k_2=W/n$.

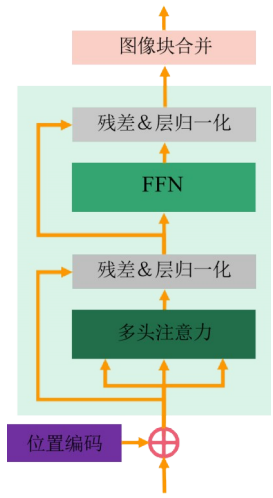

图3　多头注意力单元结构

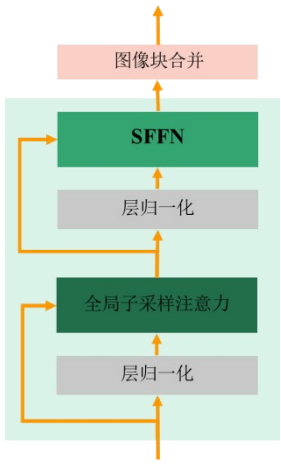

图4　本文注意力单元结构

GSSA 模块的表达式如下[22]:

$$\boldsymbol{F}^c=\boldsymbol{Q}(\rho(\boldsymbol{K}^{\mathrm{T}}))\boldsymbol{V}) \tag{3}$$

$$f_{ab}^c=(q_{ab}\boldsymbol{R}^{c\mathrm{T}})\boldsymbol{V}_{ab}^c \tag{4}$$

其中,ρ 表示对每行分别应用 softmax 归一化的操作;$\boldsymbol{V}$ 表示在 Value 特征图的 (a,b) 像素点的同一列上取 L 长度的矩阵($-L/2$ 到 $L/2$),其尺寸为 $(L\times \mathrm{dout})$;$\boldsymbol{R}$ 表示沿着列的 L 个空间偏移可学习的相对位置嵌入的矩阵;q 为 Query 特征图 (a,b) 像素点的值.

本文使用如下所示的约简比 R_r 对序列的长度进行约简[23].

$$\boldsymbol{K}'=\mathrm{Reshape}(N/R_r^{\,2},C\cdot R_r)(\boldsymbol{K}) \tag{5}$$

$$\boldsymbol{K}=\mathrm{Linear}(C\cdot R_r^{\,2},C)(\boldsymbol{K}') \tag{6}$$

其中,$\boldsymbol{K}$ 的序列降低;$\boldsymbol{K}$' 的维数为 $N/(R_r\cdot C)$;从第 1~4 层,注意机制的复杂性由 $O(N^2)$ 降低到 $O(N^2/R_r^{\,2})$;R_r 设置为[8,4,2,1].

2.3.2　SFFN

由于图像块编码限制图像块的分辨率大小,在分层分辨率时,需要对位置编码进行插值,但这会导致精度的下降[24]. 传统 Transformer 中编码器和解码器中每一层都包含一个 FFN,并由 ReLU 函数激活[25]:

$$\mathrm{FFN}(\boldsymbol{x})=\max(0,\boldsymbol{x}\boldsymbol{W}_1+\boldsymbol{b}_1)\boldsymbol{W}_2+\boldsymbol{b}_2 \tag{7}$$

其中,$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 为权重矩阵,$\boldsymbol{b}_1$ 和 $\boldsymbol{b}_2$ 为偏置向量.

为解决位置编码所带来的复杂性,以及 FFN 缺乏局部空间建模能力,本文设计了 SFFN,通过层归一化(Layer Normalization,LN)[26],采用上采样的方式将 FFN 和 3×3DSC 相结合,以融合二者的全局、局部能力. 深度可分离卷积具有逐通道卷积和逐像素卷积两部分. 其中,逐通道卷积组数与输入特征通道数一致,故可进行空间上的建模,用于替换 Transformer 的自注意机制;而逐像素卷积的卷积核大小是 1×1,可进行通道上的建模,故 SFFN 内置属性可在不影响模型性能的情况下删除网络中的位置信息,同时利用深度可分离卷积可以一次融合 patch embedding 的空间和通道信息[28]. 两者相结合既加强了网络的局部建模效果,又避免了 DCNN 的结构复杂性[29]. SFFN 的结构如图 5 所示.

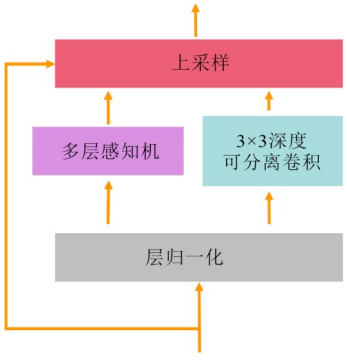

图5　SFFN 结构

其中,上采样采用 PixelShuffle 方法,层归一化的作用是对同一个样本的不同通道进行归一化处理. 层归一化是一个独立于 batch size 的算法,样本数量不会影响其计算数据量,层归一化的表达式为

$$h=f\left(\frac{g}{\sqrt{\sigma^2+\varepsilon}}\cdot(a-\mu)+b\right) \tag{8}$$

其中,σ 和 μ 为 LN 的归一化统计量;a 为 LN 归一化后的值;g 为增益(gain);b 为偏置(bias);ε 为非零项以确保等式的成立.

2.4　无锚框检测器

旋转目标锚框预测中,典型的 Transformer 主要预测出在中心点下的边框横纵坐标值,其产生的水平边界框明显不适用于遥感图像的斜框回归.

本文基于自适应训练样本选择(Adaptive Training Sample Selection,ATSS)的正样本筛选机制,采用八参数回归法设计检测头,将旋转框回归问题以角度分类

为旋转框与水平框. 当旋转框面积和最初水平框面积的重合比例大于0.95时,网络回归水平框. 检测头结构如图6所示,其中,分类分支主要用于目标物和目标类别的分类,回归分支用于预测框大小与角度计算.

本文旋转框八参数定义方法除了网格中心点与预测框的高度和宽度外,还包含Bbox水平框的4个回归点,这8个参数设为$[X,Y,H,W,a,b,c,d]$,定义的左上角为起点,其余点按逆时针顺序排列[22],如图7所示.

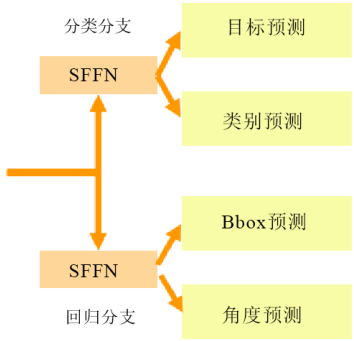

图6 检测头结构

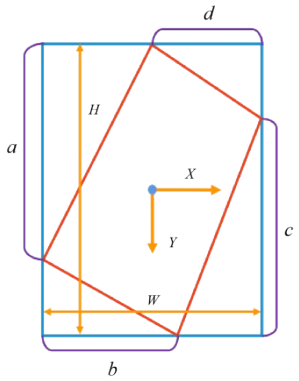

图7 旋转锚框八参数回归法

2.5 损失函数设计

本文是在Transformer的框架下对遥感旋转图像进行目标检测,目标检测结果是一个无序的集合. 为了限制梯度值,本文设计的模型的水平框将采用CIOU函数来作为损失函数(系数为2). CIOU损失函数在DIoU的基础上增加了长与宽的损失计算,能够进一步地加快收敛速度和提升性能[30,31],如图8所示,其表达式为

$$L_{\mathrm{CIOU}}=1-\mathrm{IOU}(A,B)+\rho^2(A_{\mathrm{ctr}},B_{\mathrm{ctr}})/c^2+\alpha v \tag{9}$$

$$v=\frac{4}{\pi^2}\left(\arctan\frac{w^{\mathrm{gt}}}{h^{\mathrm{gt}}}-\arctan\frac{w}{h}\right)^2 \tag{10}$$

$$\alpha=\frac{v}{(1-\mathrm{IOU})+v} \tag{11}$$

其中,式(10)和式(11)分别对应式(9)中长宽相似参数v与权重系数α的计算公式;w^{gt}和h^{gt}分别表示真实框的宽和高;w和h分别表示预测框的宽和高.

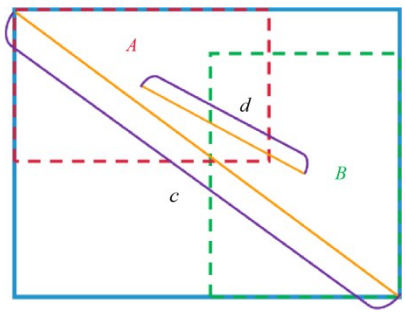

图8 CIOU损失函数计算方法

旋转框则采用L_{mr}作为损失函数[32],L_{mr}针对8参数回归特性,将损失分为三种情况计算:(1)将预测框的4个顶点顺时针移动一个位置;(2)保持预测框顶点的顺序不变;(3)将预测框的4个顶点逆时针移动一个位置. 取其中最小值作为损失值,具体过程为

$$L_{\mathrm{mr}}=\min\begin{cases}\sum_{i=0}^{3}\left(\left|x_{(i+3)\%4}-x_i^*\right|+\left|y_{(i+3)\%4}-y_i^*\right|\right)\\ \sum_{i=0}^{3}\left(\left|x_i-x_i^*\right|+\left|y_i-y_i^*\right|\right)\\ \sum_{i=0}^{3}\left(\left|x_{(i+1)\%4}-x_i^*\right|+\left|y_{(i+1)\%4}-y_i^*\right|\right)\end{cases} \tag{12}$$

其中,x_i和y_i分别表示预选框的第i个顶点和参考框的第i个顶点之间的坐标偏移.

本文总损失函数为

$$\begin{aligned}L=\frac{1}{N}\sum_{n=1}^{N}t_n\Bigg\{&t_\theta\sum_{j\in\{X,\,Y,\,H,\,W,\,a,\,b,\,c,\,d\}}L_{\mathrm{reg}\theta}(\boldsymbol{V}_{nj}^1,\boldsymbol{V}_{nj}^{\mathrm{T}})+(1-t_\theta)\\&\times\sum_{j\in\{x,y,h,w\}}L_{\mathrm{reg}}(\boldsymbol{V}_{nj}^2,\boldsymbol{V}_{nj}^{\mathrm{T}})\Bigg\}\\&+\frac{\lambda_1}{N}\sum_{n=1}^{N}L_{\mathrm{cls}}(p_n,t_n)+\frac{\lambda_2}{N}\sum_{n=1}^{N}L_{\mathrm{reg}O}(\boldsymbol{O}_n^1,\boldsymbol{O}_n)\end{aligned} \tag{13}$$

其中,N代表网络产生的锚框的数量;t_n有两个取值,分别为0和1,当t_n等于0时,代表背景,当t_n等于1时,代表前景;t_θ也有两个取值,分别为0和1,当t_θ为0时,代表此时预测锚框与最初水平框重合比大于0.95,计算水平框损失,当t_θ为1时,计算旋转框损失;$\boldsymbol{V}_{nj}^1$代表预测的旋转框8参数偏移量;$\boldsymbol{V}_{nj}^2$代表预测的水平框4参数偏移量;$\boldsymbol{V}_{nj}^{\mathrm{T}}$代表实际的偏移量;$t_n$代表目标的类别;$p_n$是该目标属于不同类别的概率;$\boldsymbol{O}_n^1$为预测方向;$\boldsymbol{O}_n$为实际朝向. λ_1与λ_2是权重因子,控制损失函数的权重,默认值为1;$L_{\mathrm{reg}\theta}$代表L_{mr}损失函数;L_{reg}代表CIOU损失函数;L_{cls}代表目标类别分类损失函数,使用Focal Loss;$L_{\mathrm{reg}O}$代表方向回归损失函数,使用Balanced L1 Loss.

3　实验与结果分析

3.1　数据集及实验设置

3.1.1　DOTA

DOTA数据集中有2 806幅航空图像,总共包含188 282个用水平包围框及旋转包围框标注的实例目标. DOTA数据集中的对象类别参考文献[22]. 在DOTA数据集中,训练集、验证集和测试集的图像数量分别为总数据集的1/2,1/6和1/3. 其中,每张图片的尺寸都在800 pixel×800 pixel至4 000 pixel×4 000 pixel之间,一般定义10~50像素的目标为小目标,DOTA数据集中小目标占比达57%. 因此,针对DOTA数据集进行测试更具有实际应用价值.

3.1.2　NWPU VHR-10

NWPU VHR-10是一个来自Google Earth和Vaihingen数据集的公共目标检测数据集. 这个数据集包括650张带有标记的图像. 每个类的样本数都小于DOTA数据集的样本数. NWPU VHR-10数据集几乎没有面积小于1 000像素的目标. 这些目标被分为10种类型:飞机(PL)、船舶(SH)、储罐(ST)、棒球场(BD)、网球场(TC)、篮球场(BC)、地面跑道(GT)、港口(HA)、桥梁(BR)和车辆(VH).

3.1.3　实验环境与参数

实验采用的软硬件配置如下:操作系统为Ubuntu20.04;CPU为Intel(R)Core(TM)i9-12000K;GPU为Nvidia RTX 3090 Ti;学习框架为Pytorch 1.7.0,cuda 11.0.

本次DOTA数据集实验中,输入图片大小为800×800,总批量大小为16,对应于每次训练16张图像. 训练轮次(Epochs)总数为500次;初始学习率为0.000 1;权重衰减率为0.000 1,IoU为0.1.

NWPU VHR-10数据集实验中,为保证实验环境的一致性,训练参数调整如下:总批量大小为4;训练轮次(Epochs)总数为300次;初始学习率为0.000 1;权重衰减为0.9,IoU为0.1.

3.2　实验结果分析

3.2.1　评价指标

为了准确评估算法的检测效果,本文采用平均精度(Average Precision,AP)、均值平均精度(mean Average Precision,mAP)来评价模型的检测精度.

针对DOTA数据集,计算公式为

$$\mathrm{AP}=\int_0^1 \mathrm{Pre}(\mathrm{Re})\mathrm{d}(\mathrm{Re})\times 100\% \tag{14}$$

$$\mathrm{mAP}=\frac{\sum_{i=1}^{15}\mathrm{AP}}{15} \tag{15}$$

其中,Pre为准确率,Re为召回率.

3.2.2　消融实验

为了验证算法中各改进模块的有效性,对层次化Transformer Stage(STS)模块、SFFN模块,以及CIOU-L_{mr}损失函数(C-L_{mr})模块进行了消融实验. 实验结果如表1所示,其中,加粗数据为最大值.

由表1可知,在添加STS模块后,算法的均值平均精度从69.20%提高到70.26%,STS模块提取多层特征以提取丰富特征信息,有效增强对遥感图像目标检测精度. 在此基础上添加SFFN后,均值平均精度进一步提高了3.16%,SFFN模块弥补了传统FFN的局部建模能力,增强了算法的局部信息感知. 最后,添加C-L_{mr}模块后,算法的均值平均精度达到75.83%. C-L_{mr}模块拆分了锚框的回归问题,旋转框与水平框的分类回归缓解了传统算法存在的损失不连续性问题.

表1　加入不同模块的目标检测精度

STS	SFFN	C-L_{mr}	mAP/%
			69.20
√			70.26
√	√		73.39
√	√	√	**75.83**

3.2.3　实验对比

将本文算法与目前5种先进的旋转目标检测方法即(RoI-Transformer(2019)、O^2DETR(2021)、稀疏Transformer(2021)、PolarDet(2021)和RS-Resnet(2021))进行比较,以验证本文算法的检测精度. 表2总结了不同算法在DOTA数据集中15个类别的检测结果,表3总结了不同算法对DOTA数据集的均值平均精度与检测速度,表4是进一步比较本文算法与RS-Resnet的模型大小及耗时量,其中,加粗数据为最大值.

由表2可知,本文算法在桥梁、大型车辆、船舶以及篮球场这四类目标中获得了最佳的AP值. 另外,在小型车辆上的AP值仅次于最高. 桥梁、篮球场均为大型目标,特征明显,检测较为容易,而DOTA数据集中的大型车辆、小型车辆以及船舶均布置为密集目标,相邻目标会有遮挡问题,但本文仍取得了最好的检测结果.

由表3可知,本文方法检测速度达到了15.16 frame/s,相比RoI-Transformer这类依赖复杂的NMS后处理的方法以及另一种基于Transformer的遥感旋转目标检测模型而言,检测速度有着明显的提高. 由表3可知,本文算法的均值平均精度(mAP)优于除RS-Resnet外的其他检测算法,RS-Resnet算法采用NAS-FPN模块对随机选取两层不同特征进行融合,提高了算法的综合检测性能,但对密集小目标的检测仍有不足.

由表4可知,RS-Resnet的模型大小为本文算法的1.45倍,测试单张图片的速度稍快于本文算法.

表2 不同算法在DOTA数据集上的AP值对比 单位:%

模型	RoI-Transformer	O^2DETR	稀疏Transformer	PolarDet	RS-Resnet	本文方法
PL	88.64	86.01	**89.91**	89.73	88.70	89.56
BD	78.52	75.92	85.78	**87.05**	82.46	79.34
BR	43.44	46.02	50.65	45.30	52.81	**56.91**
GTF	75.92	66.65	**78.16**	63.32	68.75	74.67
SV	68.81	**79.70**	64.34	78.44	78.51	78.53
LV	73.68	79.93	75.43	76.65	81.45	**84.29**
SH	83.59	89.17	75.78	87.13	86.41	**89.93**
TC	90.74	90.44	**90.88**	90.79	90.02	90.33
BC	77.27	81.19	78.67	80.58	85.37	**86.56**
ST	81.46	76.00	84.45	85.89	**86.31**	83.60
SBF	58.39	56.91	57.91	60.97	**65.10**	61.12
RA	53.54	62.45	63.56	**67.94**	65.20	62.14
HA	62.83	64.22	64.56	**68.20**	67.80	65.89
SP	58.93	65.80	66.74	**74.63**	69.29	69.63
HC	47.67	58.96	66.33	**68.67**	64.83	64.95

表3 不同算法在DOTA数据集上的mAP值及检测速度对比

模型	Backbone	mAP/%	检测速度/(frame/s)
RoI-Transformer	ResNet101	69.56	5.76
稀疏Transformer	ResNet101	72.15	6.78
PolarDet	ResNet50	72.87	—
RS-Resnet	ResNet152	75.53	15.76
本文方法	VIT	75.83	15.16

表4 RS-Resnet和本文算法模型的耗时和模型大小对比

模型	模型大小/MB	t/s
RS-Resnet	426.40	0.063
本文方法	293.28	0.067

表5 不同算法在NWPU VHR-10数据集上的检测精度与速度对比

指标	类别	Faster R-CNN	YOLOv3	YOLOv4	SSD	RetinaNet	本文方法
AP/%	PL	88.51	88.59	91.31	84.71	72.41	94.37
	VH	87.48	82.36	89.47	84.08	78.12	91.97
	SH	85.74	88.18	95.46	78.32	73.24	95.29
	BC	93.22	86.32	88.43	90.63	89.28	90.32
	ST	81.25	87.45	94.27	76.07	79.59	90.43
mAP/%		87.24	86.58	91.79	82.76	78.53	92.47
Time/s		0.227	0.053	0.034	0.085	0.124	0.031

我们进一步使用NWPU VHR-10数据集进行了实验,实验结果如表5所示. 除了使用Faster R-CNN,SSD,RetinaNet这类传统遥感图像目标检测算法,本文增加了YOLO系列[33,34]这类目前最先进的单阶段高性能遥感图像目标检测算法进行5类小目标检测对比实验. 观察可以发现,本文方法的mAP达到了92.47%. 与其他算法相比,所提算法的精度更高、鲁棒性更强. 可以发现,本文方法在飞机和车辆这两类方向多变、广邻域稀疏、多邻域聚集的目标上取得了较好的检测效果.

3.2.4 可视化检测结果

为了直观显示目标检测效果,将本文算法与RoI-Transformer,RS-Resnet在DOTA数据集中部分图像检测效果进行可视化[35],结果如图9和图10所示.

通过观察可以发现,本文方法在检测水平目标时,预测锚框更加切合,证明了本文方法缓解边界损失不连续的可靠性. 同时,对于密集目标及模糊图像,本算

(*a*) RoI-Transformer

(*b*) RS-Resnet

(*c*) 本文方法

图 9　不同模型检测效果可视化

图 10　本文方法在 DOTA 数据集上检测效果可视化

法均展示了不错的性能.

4 结论

本文针对遥感图像目标特性导致检测效果不佳的问题,提出了一种面向遥感目标检测的无锚框Transformer算法. 该算法设计了层次化Transformer主干网络、基于深度可分离卷积的局部空间特性强化的前馈网络,以及搭建了无锚框检测器,将预测框回归问题分为水平框与旋转框,缓解了旋转框的损失不连续性问题. 在DOTA数据集上的对比结果表明,该算法在检测精度上,相较稀疏Transformer提升了2.77%,相较RoI-Transformer提升了6.07%. 在以Transformer为主干架构的算法中,取得了更好的检测效果,能很好地完成遥感图像中的目标检测任务. 同时,在NWPU VHR-10数据集上的实验进一步验证了本文方法的泛用性.

未来将进一步优化算法,提升算法的检测性能,充分结合Transformer与深度卷积神经网络特性,以期设计高性能、高实时性的遥感目标探测器. 同时,目前采用的数据集均为大样本数据集,如何在较少样本情况下获得足够的训练精度,也是下一步要研究的重点.

参考文献

[1] DUAN K W, BAI S, XIE L X, et al. Centernet: Keypoint triplets for object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2020: 6568-6577.

[2] 刘颖, 刘红燕, 范九伦, 等. 基于深度学习的小目标检测研究与应用综述[J]. 电子学报, 2020, 48(3): 590-601.
LIU Y, LIU H Y, FAN J L, et al. A survey of research and application of small object detection based on deep learning[J]. Acta Electronica Sinica,2020, 48(3): 590-601. (in Chinese)

[3] 方青云, 王兆魁. 基于改进YOLOv3网络的遥感目标快速检测方法[J]. 上海航天, 2019, 36(5): 21-27, 34.
FANG Q Y, WANG Z K. Remote sensing target rapid detection method based on improved YOLOv3 network[J]. Aerospace Shanghai, 2019, 36(5): 21-27, 34. (in Chinese)

[4] DAI J F, QI H Z, XIONG Y W, et al. Deformable convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 764-773.

[5] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]//European Conference on Computer Vision. Cham: Springer, 2020: 213-229.

[6] CAO J, CHEN Q, GUO J, et al. Attention-guided context feature pyramid network for object detection[EB/OL]. (2020-05-23)[2022-05-27]. https://arxiv.org/abs/2005.11475v1.

[7] MA J Q, SHAO W Y, YE H, et al. Arbitrary-oriented scene text detection via rotation proposals[J]. IEEE Transactions on Multimedia, 2018, 20(11): 3111-3122.

[8] JIANG Y, ZHU X, WANG X, et al. R2CNN: Rotational region CNN for orientation robust scene text detection[EB/OL]. (2017-06-29)[2022-05-27]. https://arxiv.org/abs/1706.09579.

[9] DING J, XUE N, LONG Y, et al. Learning RoI Transformer for oriented object detection in aerial images[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 2849-2858.

[10] 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. Piscataway: IEEE, 2019: 8232-8241.

[11] YANG X, YAN J C, FENG Z M, et al. R3Det: Refined single-stage detector with feature refinement for rotating object[EB/OL]. (2019-08-15)[2022-05-27]. https://arxiv.org/abs/1908.05612.

[12] 肖进胜, 张舒豪, 陈云华, 等. 双向特征融合与特征选择的遥感影像目标检测[J]. 电子学报, 2022, 50(2): 267-272.
XIAO J S, ZHANG S H, CHEN Y H, et al.Remote sensing image object detection based on bidirectional feature fusion and feature selection[J].Acta Electronica Sinica, 2022, 50(2): 267-272. (in Chinese)

[13] LI W T, CHEN Y J, HU K X, et al. Oriented reppoints for aerial object detection[EB/OL]. (2021-05-24)[2022-05-27]. https://arxiv.org/abs/2105.11111.

[14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000-6010.

[15] 何林远, 白俊强, 贺旭, 等. 基于稀疏Transformer的遥感旋转目标检测[J]. 激光与光电子学进展, 2022, 59(18): 1810003.
HE L, BAI J Q, HE X, et al. Remote sensing rotation object detection based on sparse Transformer[J]. Laser & Optoelectronics Progress, 2022, 59(18): 1810003. (in Chinese)

[16] 祝星旭, 蒋球伟. 基于CNN与Transformer的无人机图像目标检测研究[J]. 武汉理工大学学报(信息与管理工程版), 2022, 44(2): 323-331.
ZHU X K, JIANG Q W. Research on object detection for UAV images based on CNN and transformer[J].Journal of Wuhan University of Technology (Information & Management Engineering), 2022, 44(2): 323-331. (in Chinese)

[17] 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. Piscataway: IEEE, 2017: 2999-3007.

[18] CHEN Z Y, ZHU Y S, ZHAO C Y, et al. DPT: Deformable patch-based transformer for visual recognition[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 2899-2907.

[19] WU P, GONG S Q, PAN K K, et al. Reduced order model using convolutional auto-encoder with self-attention[J]. Physics of Fluids, 2021, 33(7): 077107.
[20] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 1251-1258.
[21] CHU X X, TIAN Z, WANG Y Q, et al. Twins: Revisiting the design of spatial attention in vision transformers[C]//35th Conference on Neural Information Processing Systems. Virtual Event: NeurIPS, 2021: 1-12.
[22] 戴耀南. 基于障碍物图像感知的油气管道机器人路径规划策略研究[D]. 武汉: 武汉工程大学, 2022.
DAI Y N. A Dissertation Submitted in Partial Fulfillment of the Requirements forthe Degree of Doctor of Philosophy in Engineering[D]. Wuhan: Wuhan Institute of Technology, 2022. (in Chinese)
[23] WANG W H, XIE E Z, LI X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 568-578.
[24] PAN X J, REN Y Q, SHENG K K, et al. Dynamic refinement network for oriented and densely packed object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 11204-11213.
[25] AZIMI S M, VIG E, BAHMANYAR R, et al. Towards multi-class object detection in unconstrained remote sensing imagery[C]//Asian Conference on Computer Vision. Cham: Springer, 2019: 150-165.
[26] RAFFEL C, SHAZEER N, ROBERTS A, et al.Exploring the limits of transfer learning with a unified text-to-text transformer[EB/OL]. (2019-10-23)[2022-05-27]. https://arxiv.org/abs/1910.10683.
[27] SHANG R H, HE J H, WANG J M, et al. Dense connection and depthwise separable convolution based CNN for polarimetric SAR image classification[J]. Knowledge-Based Systems, 2020, 194: 105542.
[28] SHI P F, ZHAO Z X, FAN X N, et al. Remote sensing image object detection based on angle classification[J]. IEEE Access, 2021, 9: 118696-118707.
[29] SANDLER M, HOWARD A, ZHU M L, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 4510-4520.
[30] GAO J F, CHEN Y, WEI Y M, et al. Detection of specific building in remote sensing images using a novel YOLO-S-CIOU model. Case: Gas station identification[J]. Sensors, 2021, 21(4): 1375.
[31] BOCHKOVSKIY A, WANG C Y, LIAO H Y M.Yolov4: Optimal speed and accuracy of object detection[EB/OL]. (2020-04-23)[2022-05-27]. https://arxiv.org/abs/2004.10934.
[32] QIAN W, YANG X, PENG S L, et al. Learning modulated loss for rotated object detection[EB/OL]. (2019-11-19)[2022-05-27]. https://arxiv.org/abs/1911.08299.
[33] 赵玉卿, 贾金露, 公维军, 等. 基于pro-YOLOv4的多尺度航拍图像目标检测算法[J]. 计算机应用研究, 2021, 38(11): 3466-3471.
ZHAO Y Q, JIA J L, GONG W J, et al. Multi-scale aerial image target detection algorithm based on pro-YOLOv4[J]. Application Research of Computers, 2021,38 (11): 3466-3471. (in Chinese)
[34] 肖振久, 杨玥莹, 孔祥旭. 基于改进YOLOv4的遥感图像目标检测方法[J]. 激光与光电子学进展, 2023, 60(6): 0628009.
XIAO Z J, YANG Y Y, KONG X X. Remote sensing image target detection based on improved YOLOv4[J]. Laser & Optoelectronics Progress, 2023, 60(6): 0628009. (in Chinese)
[35] ZHAO P, QU Z, BU Y, et al. PolarDet: A fast, more precise detector for rotated target in aerial images[J]. International Journal of Remote Sensing, 2021, 42(15): 5831-5835.

作者简介

喻九阳　男, 1962年生, 湖北武汉人. 硕士. 武汉工程大学二级教授、博士生导师. 主要研究方向为过程装备与人工智能、机械视觉.
E-mail: yjy@wit.edu.cn

胡天豪　男, 1997年生, 湖北武汉人. 武汉工程大学硕士生. 主要研究方向为深度学习、图片处理与目标检测.
E-mail: hutianhao12@qq.com

戴耀南(通讯作者)　男, 1993年生, 湖北武汉人. 博士. 武汉工程大学讲师. 主要研究方向为计算机视觉、工业机器人.
E-mail: 22060302@wit.edu.cn