激光与光电子学进展 先进成像

基于稀疏Transformer的遥感旋转目标检测

何林远[1,2*]，白俊强[1]，贺旭[2]，王晨[2]，刘旭伦[2]

[1]西北工业大学无人系统技术研究院，陕西 西安 710072；

[2]空军工程大学航空工程学院，陕西 西安 710038

摘要 针对遥感图像目标广邻域稀疏、多邻域聚集、方向多样等特性导致检测难度大的问题，提出了一种基于稀疏Transformer的遥感旋转目标检测方法。首先，所提方法在典型端到端Transformer网络的基础上，根据遥感图像的特性，利用K-means算法实现多域聚集，从而更好提取稀疏域下的目标特征；其次，为适配旋转目标的基本属性，在边框生成阶段，利用目标包围框的中心点及边框特征学习的策略高效获取目标回归斜边框；最后，为提升网络对遥感目标的检测率，对网络的损失函数进行了优化。在DOTA和UCAS-AOD遥感数据集上的实验结果表明，所提方法的平均精度分别为72.87%和90.4%，能很好地适应遥感图像中各类旋转目标的形状与分布特性。

关键词 图像处理；遥感图像；旋转目标检测；稀疏Transformer；K-means

中图分类号 V221^{+}.3；TB553 **文献标志码** A **DOI**：10.3788/LOP202259.1810003

Sparse Transformer Based Remote Sensing Rotated Object Detection

He Linyuan[1,2*], Bai Junqiang[1], He Xu[2], Wang Chen[2], Liu Xulun[2]

[1]*Unbanned System Research Institute, Northwestern Polytechnical University, Xi'an, Shaanxi* 710072, *China*;

[2]*School of Aeronautical Engineering, Air Force Engineering University, Xi'an, Shaanxi* 710038, *China*

Abstract A remote sensing rotating target detection approach based on a sparse Transformer is proposed to address the problem of remote sensing image target detection, which is challenging due to the wide neighborhood sparse, multi-neighborhood aggregation, and multiple orientations characteristics. First, this method uses the K-means clustering algorithm to produce multi-domain aggregation, to better extract the target features in the sparse domain, based on the typical end-to-end Transformer network, and the characteristics of a remote sensing image. Second, to adapt to the basic characteristics of the rotating target, a learning technique based on the target bounding box's center point and the frame features is proposed in the frame generation stage, to efficiently obtain the target regression oblique frame. Finally, the network's loss function is further optimized to improve the detection rate of the remote sensing target. The experimental results on DOTA and UCAS-AOD remote sensing datasets show that the average accuracy of this technique is 72.87% and 90.4%, respectively; thus indicating that it can adapt effectively to the shape and distribution characteristics of various rotating targets in remote sensing images.

Key words image processing; remote sensing image; rotated object detection; sparse transformer; K-means

1 引 言

光学遥感成像下的目标具有直观、准确、抗电子干扰能力强的特点，因此光学遥感目标检测与识别一直是航空航天侦查的重要手段。当前高效的目标检测方法主要借助深度卷积神经网络(DCNN)来完成定位和分类任务，主体上遵循“特征提取+边框回归”的研究思路，取得了非常不错的检测效果[1-4]。然而，当上述适用于自然图像的检测方法遇到背景复杂、目标大小和分布不均匀及方向多变的遥感图像时，检测性能便会急剧下降。主要原因可以归结为以下两个：1)检测器中的特征提取算子没有捕获到稳定鲁棒的特征集；2)检测器对遥感图像下目标分布呈多方向的特点估计不足。目前典型的DCNN目标检测方法，无论是基于锚点估计的方法，如Faster-RCNN[1]、RetinaNet[2]，还是基于关键点的方法，如CornerNet[3]、CenterNet[4]，都

收稿日期：2021-06-07；**修回日期**：2021-06-29；**录用日期**：2021-07-20

基金项目：国家自然科学基金(61701524，62006245)、中国博士后基金(2019M653742)

通信作者：*hal1983@163.com

无法有效解决遥感旋转目标角度偏移及漏检较多等问题。虽然后续的改进方法[5-6]将包含角度的旋转锚框纳入考量范围，但由于目标的旋转角度较多，这无疑增大了后端非极大抑制算法(NMS)的复杂性，极大地增加了计算量，导致检测速度大大下降。可见，从遥感图像的属性出发，以目标的旋转特性为准绳，找到一种有效的检测识别方法，是遥感图像目标检测的前进方向。

End-to-end object detection with transformers (DETR)[7]是首个将自然语言处理中常用的Transformer方法迁移到目标检测领域的模型，然而面对更具挑战性的遥感图像旋转目标检测问题，DETR的潜力仍有待进一步挖掘。O^{2}DETR[8]是首个将DETR应用到遥感图像旋转目标检测上的模型，该模型应用Transformer直接精准定位目标，免去了繁琐的旋转锚框设计，同时用深度可分离卷积代替原始Transformer中的注意力机制[9]，大大降低了Transformer中使用多尺度特征的计算复杂度与内存，提高了检测效率，然而该模型没有充分考虑遥感图像下的感兴趣目标(舰船、飞机等)的广邻域稀疏、多邻域聚集、方向多样的特点，性能依然受限。

为了解决以上问题，本文提出了一种基于稀疏Transformer的遥感旋转目标检测方法。首先，利用“切块+嵌入”策略，在卷积神经网络(CNN)骨干网络的作用下提取图像基本特征，并转换为一维向量；其次，对二维图像每个点的位置进行编码，并与之前的一维向量相加，形成张量，送入编码器进行编码；然后，在编码后的输出与目标查询向量的联合作用下，利用解码网络结构解析旋转目标基本属性；最后，通过前馈网络进行梳理、筛选，找到目标及其精准位置。与一些基于中心点[10-11]和基于锚框[12-16]的方法相比，所提方法预测中心点时不受网格分布及各网格预测目标数量的限制，且可以避免复杂的旋转锚框计算，能够更灵活地对遥感图像的旋转目标实施精确检测。

2 DETR模型

Transformer是一种在自然语言处理(NLP)[17]领域应用广泛的方法，近年来逐渐被挖掘出迁移到其他任务上的潜力。DETR已经尝试将Transformer机制应用到图像检测任务上，按照二维图像的“思考”方式，提升了模型的归纳偏置能力，使其强大的广域注意力效能很好地被利用。DTER算法的步骤如下：

1) 预处理阶段。首先将经过骨干网络的二维数据在“切块+嵌入”策略下变成一维序列，并输入Transformer模块中。此时每个子块就相当于NLP中的一个字，这个过程也可以表示为

$$\boldsymbol{X}\in\mathbf{R}^{H\times W\times C}\rightarrow\boldsymbol{X}_p\in\mathbf{R}^{N\times(P^2\cdot C)},\tag{1}$$

式中：$\boldsymbol{X}$是输入图片；$\boldsymbol{X}_p$则是处理后的子图序列；P^2则是子图的分辨率；N则是切块后的子图数量(即序列长度)，显然有$N=HW/P^2$。由于Transformer只接受一维序列作为输入，还需要对每个二维图像块进行重整，变成嵌入的一维向量，一般利用线性变换层将二维图像块嵌入表示为一维向量。

2) 编码阶段。将空间的维度(高和宽)压缩为一个维度，即把步骤1)得到的$\boldsymbol{X}_{\text{input}}\in\mathbf{R}^{B\times d\times H\times W}$reshape成($HW$, B, 256)维的feature map，其中B为batch size的大小。此外，为了适配图像的二维位置属性，对横纵两个方向的位置采用sin/cos模式进行编码，每个方向各编码128维向量，这种编码方式更符合图像特点。位置编码的输出张量维度为(B, d, H, W)，$d=256$，其中d代表位置编码的长度，H、W代表张量的位置。即特征图上的任意一个点(H_1, W_1)均有对应的位置编码，且这个编码长度为256，其中前128维代表H_1的位置编码，后128维代表W_1的位置编码。

$$\begin{cases}E_{(p_x,2i)}=\sin\left(p_x/10000^{2i/128}\right)\\E_{(p_x,2i+1)}=\cos\left(p_x/10000^{2i/128}\right)\\E_{(p_y,2i)}=\sin\left(p_y/10000^{2i/128}\right)\\E_{(p_y,2i+1)}=\cos\left(p_y/10000^{2i/128}\right)\end{cases},\tag{2}$$

式中：$i\in[0,1,2,\cdots,128/2]$；(p_x, p_y)表示图像中任意位置；$p_x\in[1,HW]$，$p_y\in[1,HW]$。将p_x代入式(2)的前两个公式可得到两个128维向量$E_{(p_x,2i)}$和$E_{(p_x,2i+1)}$，它代表p_x的位置编码。将p_y代入式(2)的后两个公式可得到两个128维向量$E_{(p_y,2i)}$和$E_{(p_y,2i+1)}$，它代表p_y的位置编码。将这两个128维向量拼接起来，得到256维的向量，它代表了图(p_x, p_y)处的位置编码。

通过计算可得到整个batch的位置编码，编码矩阵维度为(B, 256, H, W)，将其序列化成维度为(HW, B, 256)维的张量，准备与(HW, B, 256)维的feature map相加后输入编码器(Encoder)。

3) 解码阶段。与传统Transformer不同，DETR中的解码器(Decoder)一次性处理全部的目标队列信息，即一次性输出全部的预测信息；而不像原始Transformer从左到右一个词一个词地逐步给出。这里Decoder主要包含两个输入：包括编码的输入及预测目标属性队列，其中编码输入就是步骤2)中的(HW, B, 256)的编码矩阵，目标属性队列则为一个维度为(100, B, 256)维的张量，用以学习预测具体类别和目标边框。

DETR模型通过适配二维数据，对传统的Transformer模型进行了改进，实现了典型的目标检测。然而，这种检测方式却不能直接应用在面向航空航天领域的遥感图像上，其主要原因在于遥感图像是俯视视角拍摄的，数据量大，但有效目标较少；其次，其目标在影像上的方向不是正向，而是随机无序的。因此，这既要求目标在特征提取上尽可能稀疏计算，又要在框回归上是方向相关的。DETR模型利用常规矩形框作为预测对象，显然没有契合遥感图像的特点，因此

降低了目标检测精度。

3　基于稀疏 Transformer 遥感图像旋转目标检测算法

通过对 DETR 模型的细致分析，本实验组围绕 DETR 模型开展旋转特性的相关研究，以此更好适配遥感图像的目标检测。众所周知，无论是有锚框还是无锚框的检测方法，主要遵循的都是“特征＋旋转框”的思路，尤其是基于中心点的方法，正成为近年来破解旋转目标的主要利器。受由中心点到旋转框估计方法[18-20]的启发，本实验组在 DETR 模型的基础上，有针对性地对其进行了相关改进。

3.1　整体框架

所提模型主要由三部分组成，包括 CNN 骨干结构、稀疏的编码器-解码器网络结构、适配旋转参数的前馈网络结构，具体结构如图 1 所示。

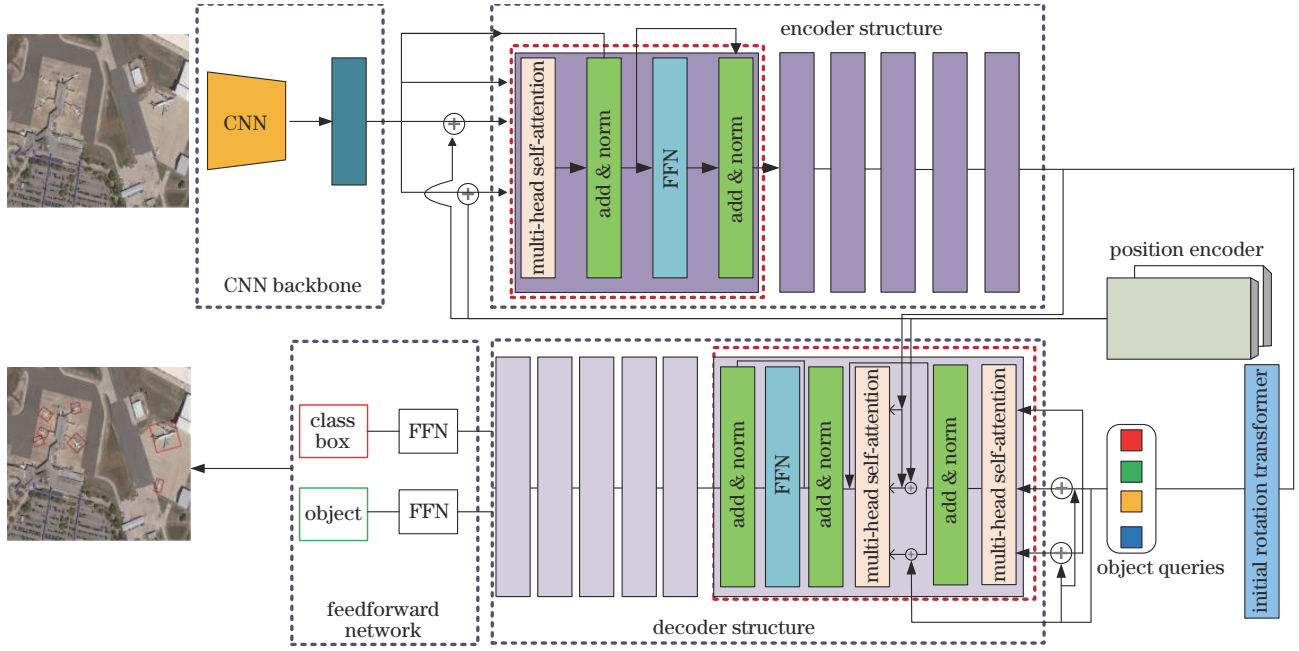

图 1　所提模型结构示意图

Fig. 1　Structure diagram of proposed model

CNN 骨干结构的主要作用在于对一幅图像实施切块，划分成很多的小块 $\boldsymbol{X}\in\mathbf{R}^{B\times 3\times H_0\times W_0}$，并将其转换为相对应的高维度 $\boldsymbol{X}_p\in\mathbf{R}^{B\times C\times H\times W}$。稀疏的编码器-解码器网络结构先用 1×1 卷积将通道数由 C 降维为 d，将输入 $\boldsymbol{X}_p\in\mathbf{R}^{B\times C\times H\times W}$ 转化为 $\boldsymbol{X}_{\text{input}}\in\mathbf{R}^{B\times d\times H\times W}$，并通过形变 reshape 成 $(HW, B, 256)$ 维，再与之后的位置索引相加一起送入 Encoder 处。编码模块主要由多头自注意单元、归一化单元、前馈单元组成。输入分别与矩阵 (W_q, W_k, W_v) 相乘，并在 K-means[21] 指导下进行稀疏不变特征提取与相关位置的记忆。在解码模块中，引入目标属性队列，并融合编码输出特征嵌入向量与位置编码向量之和。前馈网络结构将稀疏的编码器-解码器网络结构传过来的输入信息进行线性变换，经过一个激活函数后输出，最终实现场景中的目标属性和边框的回归预测。边框为适配遥感图像的旋转框，其参数为目标中心的位置、宽度、高度和方向。

3.2　稀疏不变的特征提取

目标特征提取，其核心在于找到不依赖于各种变化，具备鲁棒特性的特征点。传统的 DETR 模型，依靠自注意力机制对序列化的张量求解点积。若两个张量的点积与它们之间夹角的余弦成正比，则上述张量在方向上越接近，点积就越大。因此也就越相关，由此可以找到最为鲁棒的特征点。然而，DETR 并没有考虑像素距离远近。因此，分配给所有特征像素的注意力权重几乎是均等的，这就造成了模型需要长时间去学习关注真正有意义的特征。而在遥感图像这里，这些特征往往具有稀疏特性。因此，首先要解决的问题就是如何利用有限的稀疏特征来重新学习相关的注意力权重[22]。考虑到遥感图像的特性，这里的稀疏，一方面，与图像局部邻域归属一个物体有关；另一方面，与图像在广域范围内的相同物体有关。简而言之，这里的稀疏，就是要构建一个在短程上与绝对位置相关，在长程上与特征属性相关的自注意力新策略[23]，以较好契合遥感图像的特点。

综上所述，本实验组构建了一种稀疏表征模型，以期学会特征选择的稀疏聚类，从而提升 Transformer 的稀疏表征能力，从而更好提升运行效率。这里的聚类簇，就是构建关于每个键和查询的内容的函数。具体来说，首先将图像送入骨干网络，然后对其输出特征图进行维度转换，得到 $\boldsymbol{a}^i=\boldsymbol{W}\boldsymbol{X}^i$，并且构造位置编码张量。接着将这个嵌入向量 $\boldsymbol{a}^i$ 送入自注意力层，并将每个 $\boldsymbol{a}^i$ 分别乘以 3 个不同的 Transformation 矩阵。

$$\begin{cases} \boldsymbol{Q}=\boldsymbol{A}\boldsymbol{W}_{\boldsymbol{Q}} \\ \boldsymbol{K}=\boldsymbol{A}\boldsymbol{W}_{\boldsymbol{K}} \\ \boldsymbol{V}=\boldsymbol{A}\boldsymbol{W}_{\boldsymbol{V}} \end{cases}, \tag{3}$$

式中：$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$分别表示可拆解的查询向量、关键向量、价值向量；$\boldsymbol{W}_{\boldsymbol{Q}}$、$\boldsymbol{W}_{\boldsymbol{K}}$、$\boldsymbol{W}_{\boldsymbol{V}}$分别表示可以学习到的映射矩阵。由于遥感图像目标具有稀疏表征特性，本实验组假设每个查询向量都有一组可以与之相匹配的关键向量，而并不是跟DTER一样，所有的查询向量要遍历图像上所有的关键向量。因此，首先对查询和关键向量进行聚类，仅考虑来自同一组下的关键向量实施注意力机制。所提模型将k个关键向量和Q个查询向量在小批量样本下实现K-means聚类，其均值中心$\boldsymbol{\mu}=\{\mu_1,\mu_2,\cdots,\mu_k\}$，一旦聚类成立，令$\boldsymbol{\mu}\{\boldsymbol{Q}_i\}\in\boldsymbol{\mu}$的最近邻$\boldsymbol{Q}_i$和$\boldsymbol{\mu}\{\boldsymbol{K}_i\}\in\boldsymbol{\mu}$的最近邻$\boldsymbol{K}_i$进行相关匹配，具体可描述为

$$\boldsymbol{B}_i'=\sum_{j:\boldsymbol{K}_j\in\mu(\boldsymbol{Q}_i),j<i}^{n}\boldsymbol{A}_{ij}\boldsymbol{V}_j。\tag{4}$$

按照此种方式，此时的稀疏自注意力就如[图2(c)]所示。图2为列举的三种自注意力机制，其中对角线上每个点的注意力权重与其对应列的除白色以外的浅色点相关，其中[图2(a)]代表了邻域相关的注意力，这种注意力与相邻查询向量相关，[图2(b)]为经过跨步长处理后的注意力机制，这种注意力与跨步长相隔的查询向量相关，[图2(c)]表示为通过聚类的方式找到的基于稀疏的K个相关查询向量的自注意力。

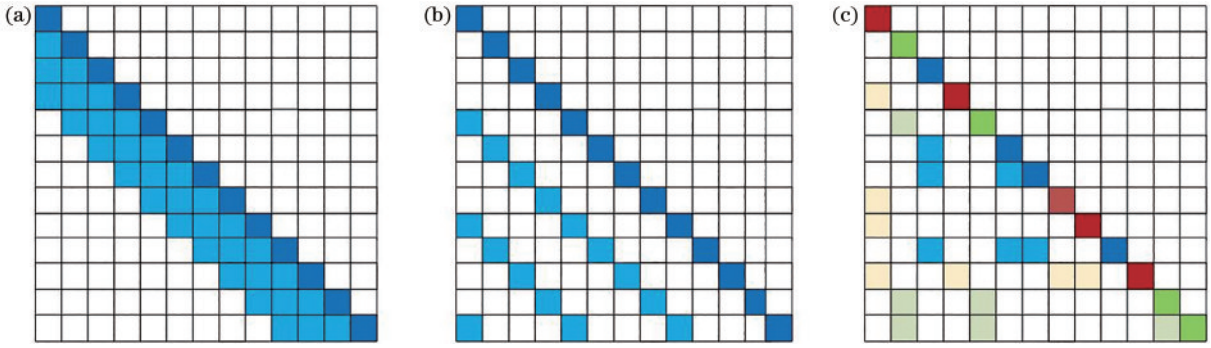

图2　多样注意力示意图。(a)邻域相关注意力；(b)跨步长注意力；(c)稀疏注意力

Fig. 2　Schematic diagrams of multiple self attention. (a) Neighborhood related self attention; (b) cross step self attention; (c) sparse self attention

此时，这里的查询向量就被最近邻的K个所关联。由于在单位向量上执行K-means算法等价于在球面实施K-means算法，查询和关键向量投影到单元球面的表达式为

$$\left\|\boldsymbol{Q}_i-\boldsymbol{K}_j\right\|^2=\left\|\boldsymbol{Q}_i\right\|^2+\left\|\boldsymbol{K}_i\right\|^2-2\boldsymbol{Q}_i^{\mathrm{T}}\boldsymbol{K}_j=2-2\boldsymbol{Q}_i^{\mathrm{T}}\boldsymbol{K}_j。\tag{5}$$

此时，若$\boldsymbol{Q}_i$和$\boldsymbol{K}_i$具有相同的聚类中心，即$\boldsymbol{\mu}(\boldsymbol{Q}_i)=\boldsymbol{\mu}(\boldsymbol{K}_i)=\boldsymbol{\mu}$时，对于任意的$\varepsilon$，有$\left|\boldsymbol{\mu}(\boldsymbol{Q}_i)-\boldsymbol{\mu}\right|=\left|\boldsymbol{\mu}(\boldsymbol{K}_i)-\boldsymbol{\mu}\right|<\varepsilon$。按照三角不等式，有

$$\left\|\boldsymbol{Q}_i-\boldsymbol{K}_j\right\|\leqslant\left|\boldsymbol{\mu}(\boldsymbol{Q}_i)-\boldsymbol{\mu}\right|+\left|\boldsymbol{\mu}(\boldsymbol{K}_i)-\boldsymbol{\mu}\right|<2\varepsilon。\tag{6}$$

将其代入式(5)中，有

$$\boldsymbol{Q}_i^{\mathrm{T}}\boldsymbol{K}_j>1-2\varepsilon^2。\tag{7}$$

此时，当具有相同的均值时，$\left\|\boldsymbol{Q}_i-\boldsymbol{K}_j\right\|$较小，而此时式(7)中的$\boldsymbol{Q}_i^{\mathrm{T}}\boldsymbol{K}_j$较高，代表关联特性更加紧密，由此可以判定相关程度的高低，而无需遍历整个向量。

3.3　旋转目标框的回归预测

旋转目标边框预测，DETR解码模型中主要依赖目标查询向量。预先设定N个目标查询向量，这里的N值远比训练/测试图像中的目标种类多。将编码器的输出送入Transformer的解码器后，便可以得到N个解码输出嵌入向量，经过前馈神经网络处理后就得到了N个预测的边框和这些边框的类别。假设真值的边框个数为m，生成的预测边框的数量N远大于真值的边框数量m。因此前面多出来的$N-m$个预测向量便会和背景类别相配对。这样就可以将边框预测和目标类别的配对看作两个等容量的集合的二分图匹配，其中主要采用的方法就是利用匈牙利算法实施相关优化预测，从而预估出每个预测目标归一化的中心点横坐标、中心点纵坐标、边框横距离、边框纵距离(c_x,c_y,w,h)。典型的Transformer中的目标查询向量的主要作用在于预估出在中心点下的边框横纵坐标值。很明显，上述方式产生的水平边界框不适用于遥感图像的斜框回归。

受中心点网络(CenterNet)[4]启发，只要得到边框的点和斜框的尺寸及倾斜角度信息，就可以唯一确定地表示旋转边框。利用全连接层从每个特征图中预测旋转框的5个几何参数，这几个参数的几何表示如图3所示，其中(x,y,w,h,θ)为用于分别表示旋转框的中心点的横坐标、纵坐标、宽度、高度和方向(长边与x轴的夹角)。在网络结构上，全连接层的作用是为了给每个特征图输出一个可学习的量。

由于有6层解码层，每层预测的结果均来自于对上一层预测结果的细化修复。假设第$d-1$层解码输

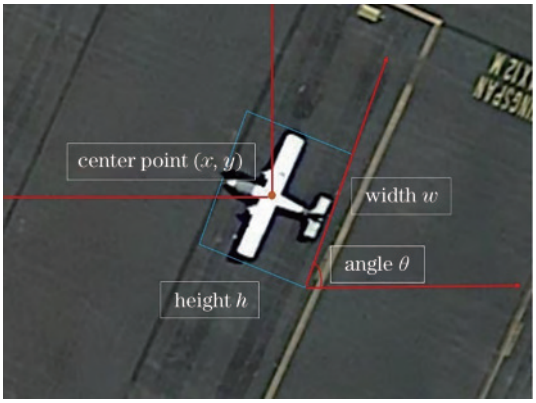

图 3 旋转包围框几何表示示意图

Fig. 3 Geometric representation of rotated bounding box

出为 $b^{d-1}_{q\{x,y,w,h,\theta\}}$，要求第 d 层的结果为 $\Delta b^{d}_{q\{x,y,w,h,\theta\}}$，即

$$\hat{b}^{d}_{q}=\left[\sigma\left(\Delta b^{d}_{q(x,y,w,h,\theta)}+\sigma^{-1}\left(\hat{b}^{d-1}_{q(x,y,w,h,\theta)}\right)\right)\right]。\tag{8}$$

为了加快网络训练速度，与传统的 DETR 网络不同，而是借鉴 Faster-RCNN[24]、Mask-RCNN[25]等网络结构，将编码后的网络直接通过三层前馈神经网络进行连接，得到粗粒度的回归坐标和前/背景分类结果。

3.4 损失函数设计

因为本实验是在稀疏 Transformer 的框架下对遥感旋转图像进行预测的，所以预测结果不像传统目标检测结果一样，是个有序集合，而是一个无序的集合。这就需要在双边匹配算法下进行合理优化，具体为

$$\hat{\sigma}=\arg\min_{\sigma\in\sum_{N}} L_{\text{match}}\left[y_i,\hat{y}_{\sigma(i)}\right],\tag{9}$$

式中：$\hat{y}_{\sigma(i)}$ 为真值所对应的预测值；$\sum_{N}$ 为从真值索引到预测值索引的所有的映射的可能排列；L_{match} 为真值 y_i 与预测值 $\hat{y}_{\sigma(i)}$ 之间的距离：

$$-1_{\{c_i\neq\varnothing\}}\hat{p}_{\sigma(i)}\left(c_i\right)+1_{\{c_i\neq\varnothing\}}L_{\text{box}}\left[\boldsymbol{b}_i,\hat{\boldsymbol{b}}_{\sigma(i)}\right]。\tag{10}$$

σ 为真值索引到预测值索引的所有的映射，对于图片上的每一个预测结果 i，找到对应的预测值 σ_i，再看分类网络的结果 $\hat{p}_{\sigma_i}\left(c_i\right)$，并取反作为式(10)的第 1 部分。对于回归的结果，作为式(11)的第 2 部分。接着，利用匈牙利函数找到对应的匹配值，即

$$L(y,\hat{y})=\sum_{i=1}^{N}\left[-\log\hat{p}_{\hat{\sigma}(i)}(c_i)+1_{\{c_i\neq\varnothing\}}L_{\text{box}}\left(\boldsymbol{b}_i,\hat{\boldsymbol{b}}_{\hat{\sigma}(i)}\right)\right]。\tag{11}$$

这里的 L_{boss} 即为检测中交并比(IOU)与 L1 范数的线性组合。同时 L_{IOU} 就是模型产生的目标 box 与正确 box 的交并比，考虑到旋转框带有角度，用倾斜交并比[26]来衡量两个 box 之间的重叠程度。

$$L_{\text{box}}\left(\boldsymbol{b}_i,\hat{\boldsymbol{b}}_{\hat{\sigma}(i)}\right)=\lambda_{\text{IOU}}L_{\text{IOU}}\left(\boldsymbol{b}_i,\hat{\boldsymbol{b}}_{\hat{\sigma}(i)}\right)+\lambda_{\text{L1}}\left\|\boldsymbol{b}_i,\hat{\boldsymbol{b}}_{\hat{\sigma}(i)}\right\|_1\tag{12}$$

$$L_{\text{IOU}}\left(\boldsymbol{b}_i,\hat{\boldsymbol{b}}_{\hat{\sigma}(i)}\right)=\frac{\left|\hat{\boldsymbol{b}}_{\hat{\sigma}(i)}\cap\boldsymbol{b}_i\right|}{\left|\hat{\boldsymbol{b}}_{\hat{\sigma}(i)}\cup\boldsymbol{b}_i\right|}。\tag{13}$$

4 实验结果与分析

4.1 数据集

4.1.1 DOTA 数据集

DOTA 数据集[27]由 2806 幅航空图像组成，总共包含 188282 个用水平包围框及旋转包围框标注的实例目标，在实验中，主要采用旋转包围框的标注形式。DOTA 数据集的类别主要包括飞机(Pl)、轮船(SH)、储油罐(ST)、棒球场(BD)、网球球场(TC)、游泳池(SP)、田径场(GTF)、港口(HA)、桥梁(BR)、大型车辆(LV)、小型车辆(SV)、直升机(HC)、环形交叉路口(RA)、足球场(SBF)及篮球场(BC)这 15 个类别，其中小型车辆和大型车辆是车辆类别的子类别。这个数据集中训练集、验证集和测试集的图像数量分别占总图像数的 1/2、1/6 和 1/3。每张图片的尺寸都在 800 pixel×800 pixel～4000 pixel×4000 pixel 之间。该数据集中的图像类别多样、方向分布均匀、目标尺度变化大，是最具有挑战性的遥感数据集之一。

4.1.2 UCAS-AOD 数据集

UCAS-AOD 数据集[28]包含飞机和汽车两种类型的目标，所有目标采用旋转边界框的标注。此数据集主要包含 1000 张彩色飞机图像和 510 个彩色汽车图像，共标注了 14626 个待测目标，包括 7482 个飞机目标和 7144 个汽车目标。UCAS-AOD 的每张图像的尺寸均为 1280 pixel×659 pixel。在实验中，随机地将其按照 7∶3 的比例划分为训练集和测试集。

4.2 实验细节及评价指标

实验环境基于深度学习框架 Pytorch 1.2 和 Ubuntu 16.04，CPU 为 Intel(R) E52603v4@2.20 GHz，同时采用 12 GB 的 Nvidia RTX 2080 Ti GPU 进行加速计算。在训练时，采用 AdamW 算法[29]对网络参数进行优化，并将 Transformer 的初始学习率设为 10^{-4}，骨干网络的初始学习率设为 10^{-5}，权值衰减率设为 0.0001，Batch size 设为 16，实验中的 Transformer 的初始权重采用 Xavier init[30]进行设置，并且采用事先在 ImageNet 中训练好的 ResNet[31]网络进行骨干网络预训练，来初始化模型，采用 ResNet-101 骨干网络作为特征提取网络，实验中的其他细节参数根据文献[7]进行设置。

采用平均精度(AP)、均值平均精度(mAP)来评价模型的检测精度，同时采用帧率来评估模型的检测速度。对于某种类别的测试结果，其准确率(precision)和召回率(recall)可表示为

$$R_{\text{precision}}=\frac{N_{\text{TP}}}{N_{\text{TP}}+N_{\text{FP}}},\tag{14}$$

$$R_{\text{recall}}=\frac{N_{\text{TP}}}{N_{\text{TP}}+N_{\text{FN}}},\tag{15}$$

式中：N_{TP} 代表真正例的数量；N_{FP} 代表假正例的数量；N_{FN} 代表假反例的数量。由准确率及召回率即可以得

到用来评估某种类别测试精度的重要指标AP值，表达式为

$$P_{\mathrm{AP}}=\int_0^1 R_{precision}(R_{\mathrm{recall}})d(R_{\mathrm{recall}})\times 100\%。\quad(16)$$

遍历各类目标的平均精度并取其均值，即可以得到衡量数据集中所有类别的重要指标均值mAP值。

4.3 实验结果与分析

为了验证所提方法在遥感图像旋转目标检测上的有效性，在DOTA数据集及UCAS-AOD数据集上进行了对比实验。实验中的所有模型均在Pytorch深度学习框架上运行。

4.3.1 在DOTA数据集上的实验

表1为不同的模型对于DOTA数据集中15个类别的检测结果，选取了5种典型的遥感旋转目标检测器，其中RRPN[13]基于Faster-RCNN基础框架并针对旋转框提出了任意方向区域提取网络及旋转感兴趣区域池化，有效地解决了斜框的检测问题。R2CNN[12]采用多尺度池化来提取长宽比信息同时提出了针对旋转框的倾斜非极大值抑制(R-NMS)算法，有效提高了检测的精度。CAD-Net[14]引入了注意力模块来提取全局及局部信息，使得网络更加关注有效的信息。RoI-Trans[15]将空间变换应用在感兴趣区域的提取过程中，有效避免了大量用于定向物体检测的旋转锚框的设计。O^2DETR是一种将Transformer应用到遥感旋转目标上检测模型，它用深度可分离卷积代替原始Transformer中的注意机制，同时借助于Transformer来预测目标的形态及位置，可实现遥感旋转目标的精准定位。以上几种模型均在遥感旋转目标检测任务上取得了一定的效果，从表1可以发现，所提模型取得的AP值在飞机、桥梁、网球场等多个类别的目标上均有所提升。不同模型在DOTA数据集上的检测精度及速度定量比较如表2所示，与其他4种模型相比，所提基于稀疏变形网络在mAP值上均有一定程度的提高，与基于Faster-RCNN的网络RRPN及R2CNN相比分别提升了2.4个百分点，3.8个百分点，与采用了特征金字塔网络(FPN)的CAD-Net相比提升了3.9个百分点，同时所提模型的精度也超过了高性能的RoI-Trans检测器与基于DETR的O^2DETR检测器，有效验证了所提模型在遥感目标检测任务上的可靠性及有效性。通过对比各模型的检测速度可以发现，所提模型的检测速度为6.78 frame/s，相比于

表1 不同模型在DOTA数据集上的AP值

Table 1 AP values of different models on DOTA dataset unit: %

Model	R2CNN[12]	RRPN[13]	CAD-Net[14]	RoI-Trans[15]	O^2DETR[8]	Proposed model
Pl	80.89	88.52	87.80	88.64	86.01	89.91
BD	65.75	71.20	82.40	78.52	75.92	85.78
BR	35.34	31.66	49.40	43.44	46.02	50.65
GTF	67.44	59.30	73.50	75.92	66.65	78.16
SV	59.93	51.85	71.10	68.81	79.70	64.34
LV	50.91	56.19	63.50	73.68	79.93	75.43
SH	55.81	57.25	76.70	83.59	89.17	75.78
TC	90.67	90.81	90.90	90.74	90.44	90.88
BC	66.92	72.84	79.20	77.27	81.19	78.67
ST	72.39	67.38	73.30	81.46	76.00	84.45
SBF	55.06	56.69	48.40	58.39	56.91	57.91
RA	52.23	52.84	60.90	53.54	62.45	63.56
HA	55.14	53.08	62.00	62.83	64.22	64.56
SP	53.35	51.94	67.00	58.93	65.80	66.74
HC	48.22	53.58	62.20	47.67	58.96	66.33

表2 不同算法在DOTA数据集上的mAP值及检测速度对比

Table 2 mAP values and detection frame rate of different detection algorithms on DOTA dataset

Model	mAP /%	Backbone	Frame rate /(frames·s^{-1})
RRPN[13]	61.01	VGG-16	5.25
R2CNN[12]	60.67	VGG-16	3.81
CAD-Net[14]	69.90	ResNet101	5.82
RoI-Trans[15]	69.56	ResNet101	5.76
O^2DETR[8]	72.15	ResNet101	6.58
Proposed model	72.87	ResNet101	6.78

RRPN、R2CNN、CAD-Net、RoI-Trans这类依赖于复杂的NMS后处理的模型及另一种基于Transformer的遥感旋转目标检测模型而言，检测速度更快。综合来看，所提模型兼顾高精度与高实时性两个特点，性能更好。所提模型在DOTA数据集上的检测结果如[图4(a)]所示。从图中可以看出，所提模型在DOTA数据集的多类别目标上达到了较好的检测效果。

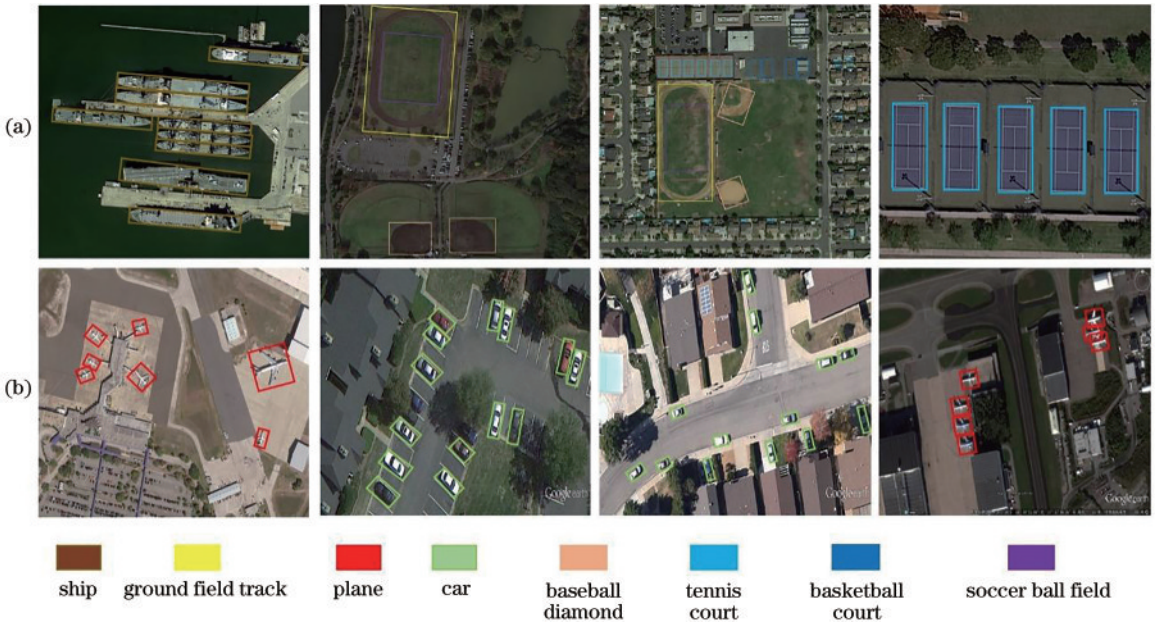

图4　所提模型在不同数据集上的检测结果。(a) DOTA；(b) UCAS-AOD

Fig. 4　Detection results of proposed model on different datasets. (a) DOTA; (b) UCAS-AOD

4.3.2　在UCAS-AOD数据集上的实验

进一步使用UCAS-AOD数据集进行了实验，实验结果如表3所示。除了使用RRPN、RDFPN、R2CNN这类基于旋转锚框范式的目标检测器，还增加了P-RSDet这类基于无锚框机制的高性能遥感图像目标探测器进行对比实验。从表3可以看出，所提模型的mAP达到了90.40%，与其他模型相比，所提模型的精度更高，鲁棒性更强。所提模型在UCAS-AOD数据集上的结果如图[4(b)]所示。从图中可以看出，所提模型在飞机和小汽车这两类方向多变、广邻域稀疏、多邻域聚集的目标上取得了较好的检测效果。

表3　不同模型在UCAS-AOD数据集上的AP值对比

Table 3　AP values of different models on UCAS-AOD dataset　　unit: %

Model	Plane	Car	mAP
RRPN[13]	88.04	74.36	81.20
R2CNN[12]	89.76	78.89	84.32
R-DFPN[16]	88.91	81.27	85.09
P-RSDet[11]	92.69	87.38	90.03
Proposed model	91.22	89.58	90.40

4.3.3　实验结果讨论

从表2的检测速度对比可以发现，相比于基于NMS的模型，这种无NMS的检测框架更直接，检测速度也更快。同时将所提模型与高性能的检测器RoI-Trans、O^{2}DETR在多类目标上的检测结果进行了可视化分析并进行了对比，如图5所示。从图中可以看出，RoI-Trans与O^{2}DETR模型的检测结果均出现了一定的定位误差及漏检现象，检测效果均没有所提模型好，进一步验证了所提模型更能适应遥感目标的形态特征，性能更优异。但是从表1的DOTA数据集上检测到的多类目标的性能指标可以发现，所提模型在桥梁上及足球场这两类目标上的检测效果依然不佳，其主要原因在于桥梁的长宽比较大，微小的角度误差都会引起较大的交并比变化，从而急剧降低检测的精度，而足球场常常位于田径场的内部，环境背景更加复杂，容易引起混淆，在检测时容易被漏检，这也大大降低足球场的检测精度。这两个问题也将会是我们下一步工作的方向，一方面要设计充分考虑目标长宽比的特性的损失函数，另一方面也要提高检测器对于具有强类间相似度的目标的辨别能力。

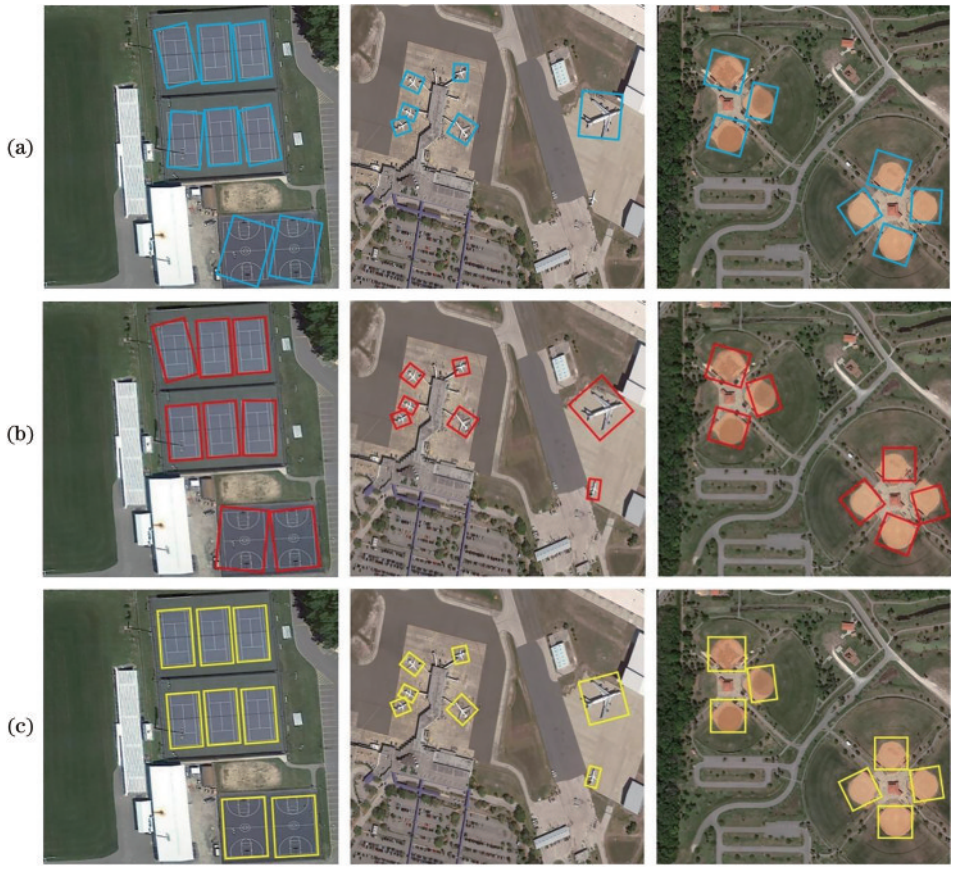

图 5　不同模型的检测结果可视化。(a) RoI-Trans 模型；(b) O²DETR 模型；(c) 所提模型

Fig. 5　Visualization of test results of different models. (a) RoI-Trans model; (b) O²DETR model; (c) proposed model

5　结　　论

提出了一种基于稀疏 Transformer 网络的遥感旋转目标检测方法。首先在典型端到端 Transformer 网络的基础上，针对遥感图像的特性，利用 K-means 算法实现目标的多域聚集，从而更好提取稀疏域下的目标特征；然后为了更好地匹配遥感旋转目标的特点，在边框生成阶段，使用基于中心点及边框特征学习的策略高效地获取目标斜边框。最后，为了提升网络对遥感目标的检测率，针对网络的特点对损失函数进行了优化。在 UCAS-AOD 和 DOTA 数据集上的实验结果表明，所提模型在检测精度与检测速度方面保持着一定的优势，能很好地完成遥感图像中的目标检测任务，具有一定的应用价值。在下一步的工作中，将继续对网络进行优化，进一步提升模型的检测性能，设计出基于 Transformer 的高效率、高性能、高实时性的遥感目标探测器。

参　考　文　献

[1] 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.

[2] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2999-3007.

[3] Law H, Deng J. CornerNet: detecting objects as paired keypoints[J]. International Journal of Computer Vision, 2020, 128(2): 642-656.

[4] Duan K W, Bai S, Xie L X, et al. CenterNet: keypoint triplets for object detection[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 6568-6577.

[5] 徐志京，丁莹. 自适应旋转区域生成网络的遥感图像舰船目标检测[J]. 激光与光电子学进展，2020，57(24)：242805.
Xu Z J, Ding Y. Ship object detection of remote sensing images based on adaptive rotation region proposal network[J]. Laser & Optoelectronics Progress, 2020, 57(24): 242805.

[6] 朱煜，方观寿，郑兵兵，等. 基于旋转框精细定位的遥感目标检测方法研究[J]. 自动化学报，2020，45(x)：1-10.
Zhu Y, Fang G S, Zheng B B, et al. Research on detection method of refined rotated boxes in remote sensing[J]. Acta Automatica Sinica, 2020, 45(x): 1-10.

[7] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020.

Lecture notes in computer science. Cham: Springer, 2020, 12346: 213-229.

[8] Teli M, Ma M Y, Mao M Y, et al. Oriented object detection with Transformer [EB/OL]. (2021-06-06) [2021-06-06]. https://arxiv.org/abs/ 2106.03146.

[9] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[EB/OL]. 2017-06-12] [2021-02-12]. https://arxiv.org/abs/1706.03762.

[10] Wei H R, Zhang Y, Chang Z H, et al. Oriented objects as pairs of middle lines[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 169: 268-279.

[11] Zhou L, Wei H R, Li H, et al. Arbitrary-oriented object detection in remote sensing images based on polar coordinates[J]. IEEE Access, 2020, 8: 223373-223384.

[12] Jiang Y Y, Zhu X Y, Wang X B, et al. R2CNN: rotational region cnn for orientation robust scene text detection[EB/OL]. (2017-06-29) [2021-05-06]. https://arxiv.org/abs/1706.09579.

[13] 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.

[14] Zhang G J, Lu S J, Zhang W. CAD-net: a context-aware detection network for objects in remote sensing imagery [J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(12): 10015-10024.

[15] Ding J, Xue N, Long Y, et al. Learning RoI transformer for oriented object detection in aerial images[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 2844-2853.

[16] Yang X, Sun H, Fu K, et al. Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks[J]. Remote Sensing, 2018, 10(1): 132.

[17] Lin T Y, Maire M, Belongie S J, et al. Microsoft COCO: common objects in context[EB/OL]. (2014-04-01)[2021-02-05]. https://arxiv.org/abs/1405.0312.

[18] Yang X, Yang J R, Yan J C, et al. SCRDet: towards more robust detection for small, cluttered and rotated objects[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 8231-8240.

[19] Zhang Z H, Guo W W, Zhu S N, et al. Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks[J]. IEEE Geoscience and Remote Sensing Letters, 2018, 15(11): 1745-1749.

[20] Li Y Y, Huang Q, Pei X, et al. RADet: refine feature pyramid network and multi-layer attention network for arbitrary-oriented object detection of remote sensing images[J]. Remote Sensing, 2020, 12(3): 389.

[21] 刘叶, 吴晟, 周海河, 等. 基于K-means聚类算法优化方法的研究[J]. 信息技术, 2019, 43(1): 66-70.
Liu Y, Wu S, Zhou H H, et al. Research on optimization method based on K-means clustering algorithm[J]. Information Technology, 2019, 43(1): 66-70.

[22] Sun P Z, Zhang R F, Jiang Y, et al. Sparse R-CNN: end-to-end object detection with learnable proposals[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 14449-14458.

[23] 汪亚妮, 汪西莉. 基于注意力和特征融合的遥感图像目标检测模型[J]. 激光与光电子学进展, 2021, 58(2): 0228003.
Wang Y N, Wang X L. Remote sensing image target detection model based on attention and feature fusion[J]. Laser & Optoelectronics Progress, 2021, 58(2): 0228003.

[24] 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.

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

[26] 张筱晗, 姚力波, 吕亚飞, 等. 基于中心点的遥感图像多方向舰船目标检测[J]. 光子学报, 2020, 49(4): 0410005.
Zhang X H, Yao L B, Lü Y F, et al. Center based model for arbitrary-oriented ship detection in remote sensing images[J]. Acta Photonica Sinica, 2020, 49(4): 0410005.

[27] Xia G S, Bai X, Ding J, et al. DOTA: a large-scale dataset for object detection in aerial images[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 3974-3983.

[28] Zhu H G, Chen X G, Dai W Q, et al. Orientation robust object detection in aerial images using deep convolutional neural network[C]//2015 IEEE International Conference on Image Processing, September 27-30, 2015, Quebec City, QC, Canada. New York: IEEE Press, 2015: 3735-3739.

[29] Loshchilov I, Hutter F. Decoupled weight decay regularization[EB/OL]. (2017-11-14)[2021-05-04]. https://arxiv.org/abs/1711.05101.

[30] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks[C]//Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, May 13-15, 2010, Chia Laguna Resort, Sardinia. Cambridge: JMLR, 2010: 249-256.

[31] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.