激光与光电子学进展 先进成像

基于稀疏掩模Transformer的遥感图像目标检测方法

刘旭伦[1]，马时平[1]，何林远[1,2*]，王晨[1]，贺旭[1]，陈哲[3]

[1]空军工程大学航空科学与工程学院，陕西 西安 710038；

[2]西北工业大学无人系统技术研究院，陕西 西安 710072；

[3]西安邮电大学网络空间安全学院，陕西 西安 710121

摘要 针对遥感图像中目标尺度差异较大和方向分布随机等导致检测精度较低的问题，提出一种基于稀疏掩模Transformer的遥感目标检测方法。该方法以Transformer网络为基础，首先引入角度参量，使其适应遥感目标的旋转特性；其次在特征提取部分以多层级特征金字塔为输入，以应对遥感图像目标尺寸变化大的特点，提高对不同尺度目标的检测效果，尤其对小目标的检测效果提升明显；最后以稀疏-插值注意力模块代替自注意力模块，有效缓解了Transformer网络检测高分辨遥感图像时计算量大的缺陷，并且加快了网络的收敛速度。在大型遥感数据集DOTA上的实验结果表明，所提方法的平均检测精度为78.43%，检测速度为12.5 frame/s，与基准方法相比，平均精度均值(mAP)提高了3.07个百分点，证明了所提方法的有效性。

关键词 Transformer；旋转目标检测；自注意力；稀疏掩模

中图分类号 V221+.3；TB553 **文献标志码** A **DOI**：10.3788/LOP202259.2228005

Target Detection Method for Remote Sensing Images Based on Sparse Mask Transformer

Liu Xulun[1], Ma Shiping[1], He Linyuan[1,2*], Wang Chen[1], He Xu[1], Chen Zhe[3]

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

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

[3]*School of Cyberspace Security, Xi'an University of Posts & Telecommunications, Xi'an* 710121, *Shaanxi, China*

Abstract Addressing the challenge of low detection accuracy due to large differences in target scale and random direction distribution in remote sensing images, this study proposes a remote sensing object detection method based on a sparse mask Transformer. This approach is based on a Transformer network. First, the angle parameter is added to the Transformer network for realizing appropriate rotational characteristics of remote sensing targets. Then, in the feature extraction section, the multi-level feature pyramid is employed as an input to deal with the large variations of the remote sensing image targets' size and enhance the detection impact for targets with various scales, particularly for small targets. Finally, the self-attention module is replaced with a sparse-interpolation attention module, which efficiently reduces the error due to the large computation amount of Transformer network detecting high-resolution images, and accelerates the network convergence speed during the training phase. The detection findings on the large-scale remote sensing dataset DOTA reveal that the proposed method's average detection accuracy is 78.43% and the detection speed is 12.5 frame/s. Compared to the traditional methods, the proposed method's mean average precision (mAP) is improved by 3.07 percentage points, which shows the proposed method's effectiveness.

Key words Transformer; rotating object detection; self-attention; sparse mask

收稿日期：2021-09-03；**修回日期**：2021-09-26；**录用日期**：2021-10-13

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

1 引 言

随着遥感技术的发展，高分辨率遥感图像日益增多，为遥感技术的应用奠定了基础。遥感图像目标检测在军事侦察、智能检测、智慧城市等多个领域起着至关重要的作用。相比于自然场景图像，遥感图像具有目标尺度变化大、目标排列方向任意、背景复杂等特点，这使得遥感图像目标检测成为当前目标检测领域的难点和热点。

近年来，随着深度学习技术的不断发展以及对遥感图像目标检测的深入研究，涌现了许多性能良好的旋转框检测算法。Cheng 等[1]通过在训练过程中引入旋转不变正则化和 Fisher 判别正则化，提高检测精度。Zhang 等[2]通过捕获全局场景和局部特征的相关性增强特征。Xu 等[3]通过受长宽比约束的非最大值抑制来提高候选区域的质量，并利用可变形的卷积神经网络来对物体的几何变化进行建模，有效改善了目标检测性能。Yang 等[4]利用像素注意力机制抑制图像噪声，突出目标特征，并在 Smooth L1 损失中引入交并比（IoU）常数因子解决旋转框边界问题，使旋转框预测更加精确。Liu 等[5]将传统的边界框替换为可旋转边框，并嵌入 SSD 中，使得算法可以预测目标的方向角，具有旋转不变性。这些算法是基于传统卷积神经网络针对遥感图像的改进算法，在一定程度上改善了遥感图像目标检测性能。但遥感图像中目标检测角度偏移、漏检较多、召回率较低等问题仍在，如何针对这些问题来提高对遥感图像的目标检测精度仍需进一步研究。

目前，一种基于自注意力机制的 Transformer[6]深度学习模型在自然语言处理、计算机视觉和音频处理等各个领域展现不错的竞争效果。在目标检测方面，Carion 等[7]提出了一种新的检测思路，即 End to End Object Detection with Transformer（DETR），通过结合卷积神经网络和 Transformer 的自注意力机制，摒弃了传统检测方法中的手工部件[8]，如非极大值抑制操作、锚框生成，大大简化了物体检测的设计流程，做到真正的端到端检测。在 COCO 数据集[9]上，DETR 的准确率和运行效率与高度优化的 Faster-RCNN[10]基本持平，在大目标上的检测效果优于 Faster-RCNN，成为目标检测领域的新范式。

尽管 DETR[7]在自然场景图像目标检测任务上有着优秀的表现，但在遥感图像目标检测领域还未有相关应用。本文以 Transformer 和 DETR 网络为基础，提出一种基于稀疏掩模 Transformer 的遥感目标检测方法，将 Transformer 模型应用于遥感图像目标检测。具体贡献总结如下：所提方法将 Transformer 模型应用于遥感图像目标检测，引入角度参量，使得检测旋转框与角度任意的目标端到端匹配，避免了经典检测方法中复杂的旋转锚框设计；遥感图像中目标尺度变化大，因此在特征提取阶段以多层级特征金字塔为输入，为遥感图像不同尺度的目标提供丰富的纹理信息和语义信息，以提高对不同尺度目标的检测精度，但与此同时，由于 Transformer 自注意力机制，多尺度特征图输入会造成巨额的计算量以及复杂的存储问题，成为亟需解决的难点；以稀疏-插值自注意力模块替代 DETR 中的注意力模块，可以快速将注意力集中在稀疏有意义的位置，加快了网络训练收敛速度，且本文提出的稀疏掩模使得 Query 和 Key 相关计算变得简单快速，解决了多尺度特征图输入带来的计算量大、存储复杂的问题。在 DOTA 数据集[11]上对模型进行评估，与基准方法相比，所提方法的平均精度均值（mAP）提升了 3.07 个百分点，证明了所提方法的有效性。

2 相关工作

2.1 基于自注意力的 Transformer

2017 年，Vaswani 等[6]提出了一种具有编码-解码架构的 Transformer 网络模型，此网络基于多头自注意力机制和前馈神经网络，以获得全局信息来完成自然语言处理任务，取得很好的效果。相比于 RNN[12-14]，Transformer 优秀的全局和记忆能力，使得它能更有效地对长序列信息进行建模，从而替代 RNN 应用于自然语言处理、语音处理和计算机视觉任务中。

2.2 DETR 目标检测

DETR[7]将 Transformer 模型应用于目标检测领域，该模型把目标检测看成一个集合预测的问题。给定一幅图像，模型采用 Transformer 编码器-解码器架构，预测所有目标的无序集合（每个目标基于类别表示，周围各有一个锚框），结合匈牙利算法进行双边匹配，实现预测值与真值的一一对应，最后利用损失函数进行优化。鉴于所提网络模型主要以 DETR 为基础，下面介绍 DETR 网络模型的基本架构。

1）骨干网络。输入图像 $\boldsymbol{x}_{\mathrm{img}} \in \mathbb{R}^{3\times H\times W}$，通过 CNN 骨干网络（ResNet[15]）提取特征图 $\boldsymbol{f} \in \mathbb{R}^{C\times H\times W}$，并进行通道压缩和序列化数据转换，将结果融入位置编码信息，得到输入 $\boldsymbol{z}_0 \in \mathbb{R}^{HW\times 256}$。

2）标准 Transformer 编码器-解码器架构。在编码器中对 $\boldsymbol{z}_0 \in \mathbb{R}^{HW\times 256}$ 进行编码，得到候选目标的增强特征，将其送入解码器中进行并行解码。

3）检测头。将解码的信息分别送入 2 个 FFN 进行类别和检测框的预测。

3 基于稀疏掩模 Transformer 的遥感目标检测方法

所提方法的整体网络结构如图 1 所示，主要工作如下。

1）以 ResNet 为骨干网络，提取 C_3 至 C_5 特征图，用 1×1 的卷积将特征图的维度降为 256，作为输入特征

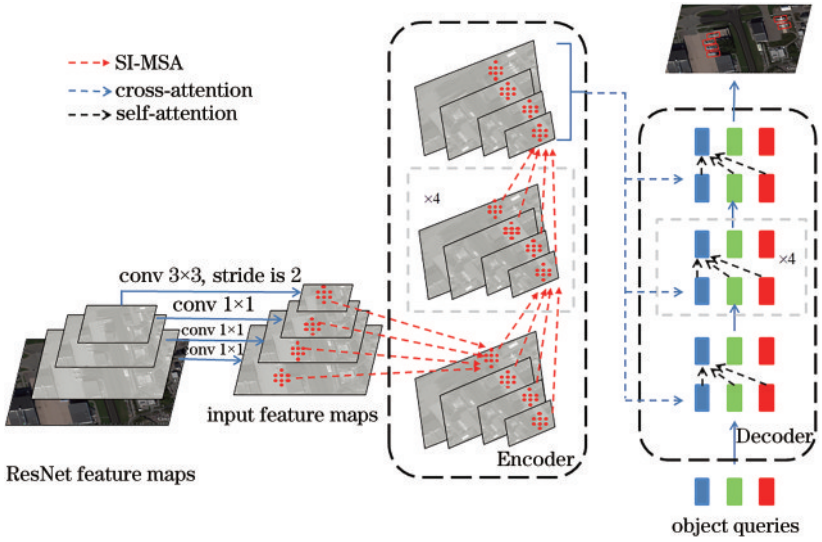

图 1　所提网络模型的结构示意图

Fig. 1　Structure diagram of the proposed network

图 $\boldsymbol{X}_1$ 至 $\boldsymbol{X}_3$；用 3×3 的卷积将 C_5 下采样，得到尺度减半、维度为 256 的输入特征图 $\boldsymbol{X}_4$；$\boldsymbol{X}_1 \sim \boldsymbol{X}_4$ 为本网络的输入特征。

2）以稀疏-插值注意力模块替代 Encoder 中的自注意力模块（如图 2 所示）。

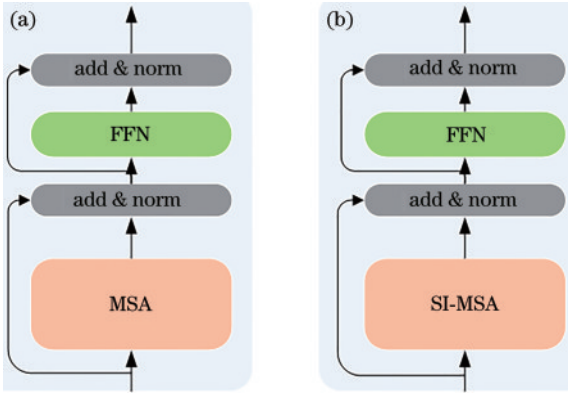

图 2　注意力模块示意图，MSA 为多头注意力，SI-MSA 为稀疏-插值多头注意力。(a) 标准注意力模块；(b) 稀疏-插值注意力模块

Fig. 2　Schematic of attention block. MSA is multi-head attention, SI-MSA is sparse-interpolation multi-head attention. (a) Standard attention block; (b) sparse-interpolation attention block

3）在网络加入角度参量 θ，在学习的过程中，object queries 将学习到一个包含角度的位置编码。Decoder 输出经两个前馈网络的张量维度为 $(b, 100, c_{\text{class}}+1)$ 和 $(b, 100, 5)$，前者是 100 个预测框的类型，后者为 100 个预测框，5 代表预测目标归一化的位置参量 (C_x, C_y, w, h, θ)。最终经匈牙利双边匹配算法完成预测值与真值的一一对应。

3.1　稀疏-插值注意力

由于图片中目标实例只占少数，存在着大量的信息冗余，这些冗余的信息增加了计算量和存储复杂度。在自注意力计算中所有信息两两交互，所以表现更为明显，因此设计了稀疏-插值注意力模块，通过采样，减少信息冗余，从而降低模型计算注意力时的计算量。

稀疏-插值注意力模块的网络结构如图 3 所示，从以下 4 个方面进行说明。

1）模块有三条支路，其中两条支路和标准的多头自制注意力模块一样，通过对输入 $\boldsymbol{Z}_0 \in \mathbb{R}^{n \times d_m}$ 进行线性映射得到 $\boldsymbol{K}$ 和 $\boldsymbol{V}$，维度为 $k \times n \times d_k$，其中 k 为多头的个数，$d_k = d_m / k$。

2）剩余的分支通过 reshape 将 2D 的输入 $\boldsymbol{Z}_0 \in \mathbb{R}^{n \times d_m}$ 变形为 3D 的空间图形，尺寸为 $d_m \times H \times W$，通过采样模块对其采样，获得稀疏掩模

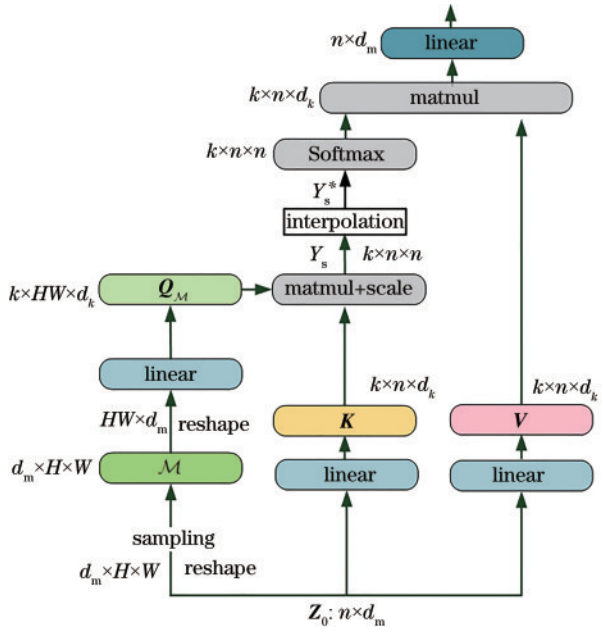

图 3　稀疏-插值多头自注意力

Fig. 3　Sparse-interlation multi-head self-attention

$\mathcal{M}\in\mathbb{R}^{d_m\times H\times W}$。掩模 $\mathcal{M}$ 通过 reshape 后进入线性层，映射得到 $\boldsymbol{Q}_{\mathcal{M}}\in\mathbb{R}^{k\times HW\times d_k}$。$\mathcal{M}(\boldsymbol{p})=1$ 代表图片中的采样点，只在采样点处进行相关性计算。由于 $\mathcal{M}$ 是稀疏的，且相比于自注意力模块需要通过巨量训练才能将注意力集中在有意义的点上，对 $\mathcal{M}$ 的训练要简单快速，因此通过变形和映射后，得到稀疏-注意力聚焦的 $\boldsymbol{Q}_{\mathcal{M}}$，可大大降低 $\boldsymbol{Q}_{\mathcal{M}}$ 和 $\boldsymbol{K}$ 相关计算量。

3）计算 $\boldsymbol{Q}_{\mathcal{M}}$ 与 $\boldsymbol{K}$ 之间的相似度 Y_s。在 DETR[7] 中，对于 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$，是通过点积对所有信息进行两两交互得到相似度的。而在所提方法中，只在采样点处对 $\boldsymbol{Q}^{i'}_{\mathcal{M}}$ 与 $\boldsymbol{K}$ 中的所有向量进行计算，在非采样点处，相似度记为 0，计算公式为

$$Y_s(\boldsymbol{p})=\begin{cases}\boldsymbol{Q}^{i'}_{\mathcal{M}}\boldsymbol{K}^{\mathrm{T}}, & \mathcal{M}(\boldsymbol{p})=1\\ 0, & \mathcal{M}(\boldsymbol{p})=0\end{cases}。\tag{1}$$

需要注意，二值化的 $\mathcal{M}$ 是不可微的，在训练阶段梯度不可被反向传播。因此，在所提方法中从训练阶段到推理阶段，$\mathcal{M}$ 是逐渐二值化的。在训练阶段，$\boldsymbol{Q}_{\mathcal{M}}$ 与 $\boldsymbol{K}$ 之间的相似度 Y_s 为

$$Y_s=\boldsymbol{Q}^{i'}_{\mathcal{M}}\boldsymbol{K}^{\mathrm{T}},\quad i'\leqslant MN。\tag{2}$$

4）对每个头的输出进行拼接，再将经过一次线性变换得到的值作为多头 Attention 的结果。计算单头 Attention 的公式为

$$\text{SI-MSA}(\boldsymbol{Q}_{\mathcal{M}},\boldsymbol{K},\boldsymbol{V})=\text{Softmax}\left[\text{Inter}\left(\frac{Y_s^*}{\sqrt{d_k}}\right)\right]\boldsymbol{V}。\tag{3}$$

3.2　稀疏采样模块

在确定性采样中，当置信度大于某一阈值，将被采样。如图 4(a)所示，置信度大于 0.5 的点均被采样，小于 0.5 的点不被采样。而在随机采样中，较高的置信度只代表被采样的几率相对较高，如图 4(b)所示，置信度高则被采样的几率大，置信度低则被采样的几率小。

从骨干网中提取的特征图存在空间冗余，而相邻点一般具有相似的特征和置信度，因此确定性抽样通常对相邻点进行抽样或不抽样。而随机采样可以更加均衡地采样，未采样点的相关性通过插值模块得到，使计算相对精确。受文献[16]启发，以文献[16]中的二分类 Gumbel-Softmax 分布为基础模拟随机采样，将掩模 $\mathcal{M}$ 定义为

$$\mathcal{M}(\boldsymbol{p})=\frac{\exp\left\{\left[\log\boldsymbol{\pi}_1+g_1(\boldsymbol{p})\right]/\tau\right\}}{\sum_{j\in\{0,1\}}\exp\left\{\left[\log\boldsymbol{\pi}_j+g_j(\boldsymbol{p})\right]/\tau\right\}},\tag{4}$$

式中：$\boldsymbol{\pi}$ 代表二分类 Softmax 激活函数产生的置信图；0 和 1 分别代表背景和实例；g 表示服从标准 Gumbel 分布的噪声；τ 是温度参数，当 τ 趋近于 0 时，$\mathcal{M}(\boldsymbol{p})$ 近似为二元化。在训练过程中，$\boldsymbol{\pi}$ 是通过 3×3 的卷积层和二分类 Softmax 激活函数产生的。$\tau=\alpha^{n_{\text{iter}}}\tau_0$，$\alpha$ 是衰减因子，n_{iter} 是迭代次数，τ_0 为初始值，在本文实验中设置 $\tau_0=1$。

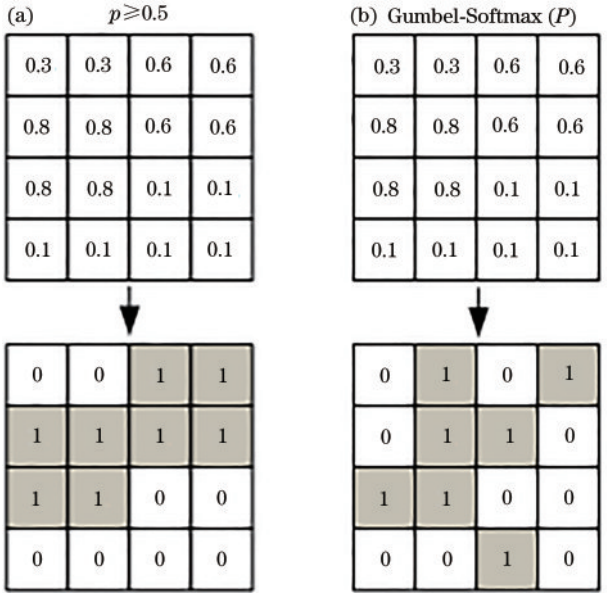

图 4　确定性采样和随机性采样。(a)确定性采样；(b)随机性采样

Fig. 4　Deterministic sampling and stochastic sampling. (a) Deterministic sampling; (b) stochastic sampling

因此，在训练初始阶段，掩模的取值是光滑连续的，梯度可被反向传播，采样模块可训练。训练结束，τ 趋近于零，掩模取值近似为二元化。

3.3　插值模块

通常，相邻点位置具有很强的相关性。为了提高 Y_s 的相对准确性，对于 Y_s 中为零点处的值由 Y_s 周围不为零点处的值插值得到，计算公式为

$$Y_s^*(\boldsymbol{p})=I(Y_s)(\boldsymbol{p})=\begin{cases}\boldsymbol{Q}^{i'}_{\mathcal{M}}\boldsymbol{K}^{\mathrm{T}}, & \mathcal{M}(\boldsymbol{p})=1\\ \dfrac{\sum_{s_q}W_I(\boldsymbol{p},\boldsymbol{s}_q)Y_s(\boldsymbol{s}_q)}{\sum_{s_q}W_I(\boldsymbol{p},\boldsymbol{s}_q)}, & \mathcal{M}(\boldsymbol{p})=0,\boldsymbol{s}_q\in\mathcal{R}_s^{\eta}(\boldsymbol{p})\end{cases},\tag{5}$$

式中：$\mathcal{R}_s^{\eta}(\boldsymbol{p})=\left\{\boldsymbol{s}_q\middle|\boldsymbol{s}_q\in\Omega,\left\|\boldsymbol{s}_q-\boldsymbol{p}\right\|_{\infty}\leqslant\eta\right\}$；$W_I(\boldsymbol{p},\boldsymbol{s}_q)\geqslant 0$ 表示插值权重，本文中 W_I 表示平均池化。

3.4　损失函数设计

在 DETR[7] 中，预测输出为一组无序集合，通过匈牙利算法双边匹配实现预测值与真值一一对应，具体为

$$\hat{\delta}=\arg\min_{\delta\in\sum_N}\sum_i^N\mathcal{L}_{\text{match}}(y_i,y_{\hat{\delta}(i)}),\tag{6}$$

式中：y_i 为某一个真值，$y_i=(c_i,b_i)$，c_i 表示类别，$b_i=(C_{x_i},C_{y_i},w_i,h_i)$ 表示真值框，加入角度参量 θ 后，$b_i=(C_{x_i},C_{y_i},w_i,h_i,\theta_i)$；$y_{\hat{\delta}(i)}$ 为与真值对应的预测值；$\sum_N$ 为真值索引 i 到预测值索引 $\delta(i)$ 映射的所有可能排列。用 $\mathcal{L}_{\text{match}}$ 最小化真值 y_i 和预测值 $y_{\hat{\delta}(i)}$ 之间的距离，

$\mathcal{L}_{\text{match}}$具体为

$$-1_{\{c_i\neq\varnothing\}}p_{\hat{\delta}(i)}(c_i)+1_{\{c_i\neq\varnothing\}}L_{\text{box}}(b_i,b_{\hat{\delta}(i)}),\tag{7}$$

可以使$\mathcal{L}_{\text{match}}$最小的排列$\hat{\delta}$就是所要找的排列，即对于第$i$真值，第$\delta(i)$预测值与之相配。根据找到的所有排列，计算损失函数，表达式为

$$L_{\text{Hungarian}}(y,\hat{y})=\sum_{i=1}^{N}\left[-\log\hat{p}_{\hat{\delta}(i)}(c_i)+1_{\{c_i\neq\varnothing\}}\mathcal{L}_{\text{box}}(b_i,\hat{b}_{\hat{\delta}(i)})\right]。\tag{8}$$

所提方法由于加入了稀疏采样模块，为了更好地让网络训练，产生稀疏-注意力聚焦的掩模，定义采样模块的损失函数为

$$L_{\text{sparse}}=\sum_{l}\left\|\boldsymbol{\pi}_1^l\right\|_1,\tag{9}$$

式中：$\boldsymbol{\pi}_1^l$表示第l层的置信度图；$\|.\|_1$表示L1-norm。

因此，可将所提基于稀疏掩模自注意力Transformer网络构架的损失函数定义为

$$\begin{aligned}\mathcal{L}&=\mathcal{L}_{\text{Hungarian}}(y,\hat{y})+\gamma L_{\text{sparse}}=\\&\sum_{i=1}^{N}\left[-\log\hat{p}_{\hat{\delta}(i)}(c_i)+1_{\{c_i\neq\varnothing\}}\mathcal{L}_{\text{box}}(b_i,\hat{b}_{\hat{\delta}(i)})\right]+\\&\gamma\sum_{l}\left\|\boldsymbol{\pi}_1^l\right\|_l,\end{aligned}\tag{10}$$

式中：γ为采样模块损失函数的权重。

4 实　　验

4.1 数据集

为了验证所提方法的有效性，在DOTA数据集上进行对比实验。DOTA数据集是由旋转框标注的大型公开数据集，主要用于遥感图像目标检测任务。该数据集由2806幅来自不同传感器和平台，大小从800×800像素到4000×4000像素不等的遥感图像组成，其中包含了188282个不同尺度、方向、形状的目标实例。DOTA数据集主要包括15种常见类别：飞机(PL)、直升机(HC)、游泳池(SP)、环形车道(RA)、港口(HA)、篮球场(BC)、足球场(SBF)、网球场(TC)、田径场(GTF)、棒球场(BD)、储油罐(ST)、桥梁(BR)、船舶(SH)、小型车辆(SV)、大型车辆(LV)。选取该数据集的1/2作为训练集，1/6作为验证集，1/3作为测试集，将所有图像统一裁剪为1024×1024大小的子图像。

4.2 评估标准

采用平均精度(AP)、平均精度均值(mAP)来评价模型的检测准确率，采用帧率(FPS)评估模型的检测速度。其中AP的计算表达式为

$$P_{\text{AP}}=\int_0^1 p(r)\mathrm{d}r,\tag{11}$$

其计算了准确率(p)和召回率(r)在[0,1]范围内绘制的曲线与坐标轴所围成的面积。准确率和召回率分别定义为

$$\begin{cases}p=\dfrac{N_{\text{TP}}}{N_{\text{TP}}+N_{\text{FP}}}\\[2ex]r=\dfrac{N_{\text{TP}}}{N_{\text{TP}}+N_{\text{FN}}}\end{cases},\tag{12}$$

式中：TP表示真正例；FP表示假正例；FN表示假反例。mAP的计算表达式为

$$P_{\text{mAP}}=\frac{\sum_{i=1}^{15}P_{\text{AP}i}}{15}。\tag{13}$$

FPS的计算表达式为

$$F=\frac{N_{\text{test}}}{T_{\text{time}}},\tag{14}$$

式中：N_{test}为测试集的样本数量；T_{time}为对测试集进行检测所消耗的时间。

4.3 实验配置

实验硬件采用Intel E5-2683 CPU，NVIDIA Tesla 3xV100GPU(批处理大小为6)，64G内存的服务器，实验环境为Ubuntu 16.04.4、Cuda 10.0、Cudnn 7.4.2、Pytorch 1.2。

本实验骨干网络采用在ImageNet[17]上训练好的ResNet网络。在训练阶段，采用AdamW算法[18]对网络参数进行优化，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}。模型训练50个epoch，从第40个epoch开始学习率下降1/10。

4.4 检测结果与分析

为了评估所提方法的性能，选取了7种典型的遥感目标检测算法进行对比，实验结果如表1所示。

其中R2CNN[19]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non-maximum suppression(NMS)算法，提高了检测的精度。RRPN[20]基于Faster-RCNN基础框架，提出了任意方向区域提取网络和旋转感兴趣区域池化(RRoI-pooling)，有效解决了斜框检测问题。RT[21]为RoI-Transformer方法，将水平感兴趣区域(HRoI)转换为旋转感兴趣区域(RRoI)，引入RPS-RoI-Align模块，从中提取旋转不变特征，以促进分类和回归。CAD-Net[2]基于RCNN和FPN，引入了空间感知注意模块，引导网络关注信息更加丰富的区域和更适合的图像特征尺度。SCRDet[22]设计了一种采样融合网络，将多层特征融合到有效的锚框中，提高了对小型目标的检测灵敏度。GV[3]是GV R-CNN方法的简称，通过学习角点在非旋转矩形上的偏移来定位出目标斜框。BBAVectors[23]以CenterNet方法[24]为基础检测物体中心点，回归一个包围框的边缘感知向量(BBAVectors)来得到有方向的包围框。

从表1可以看出，所提基于稀疏掩模的Transformer遥感目标检测方法优于其他方法，mAP值达78.43%。在飞机、小型车辆、大型车辆、船舶这些具有尺寸小且排列密集特点的目标上取得了很好的

表 1　不同方法在 DOTA 数据集的检测精度对比

Table 1　Comparison of detection accuracy of different methods in DOTA dataset

Method	AP /%															mAP /%
	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	
R2CNN[19]	80.94	65.67	35.34	67.44	59.92	50.91	55.81	90.67	66.92	72.39	55.06	52.23	55.14	53.35	48.22	60.67
RRPN[20]	88.52	71.20	31.66	59.30	51.85	56.19	57.25	90.81	72.84	67.38	56.69	52.84	53.08	51.94	53.58	61.01
RT[21]	88.64	78.52	43.44	75.92	68.81	73.68	83.59	90.74	77.27	81.46	58.39	53.54	62.83	58.93	47.64	69.56
CAD-Net[2]	87.80	82.40	49.40	73.50	71.10	64.50	76.60	90.90	79.20	73.30	48.40	60.90	62.00	67.00	62.20	69.90
SCRDet[22]	89.98	80.65	52.09	68.36	68.36	60.32	72.41	90.85	87.94	86.86	65.02	66.68	66.25	68.24	65.21	72.61
GV[3]	89.64	85.00	52.26	77.34	73.01	73.14	86.82	90.74	79.02	86.81	59.55	70.91	72.94	70.86	57.32	75.02
BBAVectors[23]	88.63	84.06	52.13	69.56	78.26	80.40	88.06	90.87	87.23	86.39	56.11	65.52	67.10	72.08	63.96	75.36
Proposed method	89.14	84.40	54.73	76.80	79.21	82.01	89.23	91.34	86.05	88.54	68.65	69.90	70.83	74.27	71.37	78.43

检测效果，说明所提方法对于这类场景的检测更具优势。图 5 展示了模型检测的部分结果，从检测结果可以看到，所提方法在遥感图像目标漏检、角度偏移等问题上也有明显改善，说明所提方法能够有效地用于遥感图像的目标检测。

图 5　部分检测结果展示

Fig. 5　Visualization of parts of the detection results

表 2 为不同方法在 DOTA 数据集上的检测速度对比。可以看到，由于所提方法避免了复杂的后端处理，检测速度相应得到提高，达 12.5 frame/s。

表 2　不同方法在 DOTA 数据集上的 mAP 值及检测速度对比

Table 2　mAP value and detection speed of different detection methods on DOTA dataset

Model	Backbone	mAP /%	Speed /(frame·s^{-1})
R2CNN	VGG-16	60.67	5.9
RRPN	VGG-16	61.01	7.2
RT	R101-FPN	69.56	7.8
CAD-Net	R101-FPN	69.90	7.9
SCRDet	R101-FPN	72.61	8.4
GV	R101-FPN	75.02	11.6
BBAVectors	ResNet-101	75.36	13.7
Proposed method	ResNet-101	78.43	12.5

4.5　消融实验

为了验证所提方法中不同模块的有效性，分别对多尺度特征金字塔模块、稀疏采样模块、插值模块进行消融实验，探究各个模块对模型效果的贡献。其中“Baseline”为实验的基准设置，采用 ResNet101 提取单尺度特征图，使用标准的注意力模块。消融实验结果如表 3 所示。

通过表 3 发现：使用标准注意力模块后需要 500 轮的迭代，模型才能达到最优解；与之相比，采用稀疏采样模块后模型有着更好的表现，可以更加快速将注意力集中在稀疏有意义的点上，加快了收敛速度，迭代次数降为其 1/10，且可以看出，稀疏采样模块大大缓解了多尺度输入带来的巨额计算量问题；此外，加入的多尺度金字塔特征输入对检测精度提升明显，相比单尺

表 3 消融实验

Table 3 Ablation study

Baseline	Multi-scale input	Sampling module	Interpolation module	Epoch	GFLOPs	mAP /%
√				50	152	65.33
√				500	152	76.41
√	√			50	1890	67.18
√	√			500	1890	78.23
√	√	√		50	138	77.86
√	√	√	√	50	140	78.43

度特征，mAP 值提升 1.85 个百分点。加入稀疏采样模块后，模型的精度有少许降低。但鉴于稀疏采样模块的主要作用是加快模型训练的收敛速度，降低多尺度特征图、高分辨率图片所带来的巨额计算量，0.37 个百分点的 mAP 损失可以忽略，且通过插值模块，检测精度得到提高。

5 结 论

提出了基于稀疏掩模的 Transformer 遥感图像目标检测方法，成功地将这一新方法运用于遥感图像目标检测，并以稀疏-插值自注意力模块代替原注意力模块，取得很好的效果。所提方法解决了 DETR 检测方法训练收敛慢、注意力计算量大且存储复杂的问题。另外，在遥感目标检测上，所提方法有效地改善了漏检、角度偏移的常见缺陷，提高了对不同尺度目标的检测精度。下一步工作中，将继续优化改进网络，在 COCO 数据集[9]上进行实验，对标 Deformable DETR 检测方法[25]。

参 考 文 献

[1] Cheng G, Han J W, Zhou P C, et al. Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection[J]. IEEE Transactions on Image Processing, 2019, 28(1): 265-278.

[2] 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.

[3] Xu Y C, Fu M T, Wang Q M, et al. Gliding vertex on the horizontal bounding box for multi-oriented object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(4): 1452-1459.

[4] 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.

[5] Liu L, Pan Z X, Lei B. Learning a rotation invariant detector with rotatable bounding box[EB/OL]. (2017-11-26)[2021-04-05]. https://arxiv.org/abs/1711.09405.

[6] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. New York: Curran Associates, 2017: 5998-6008.

[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] Liu L, Ouyang W L, Wang X G, et al. Deep learning for generic object detection: a survey[J]. International Journal of Computer Vision, 2020, 128(2): 261-318.

[9] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context[M]//Fleet D, Pajdla T, Schiele B, et al. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8693: 740-755.

[10] 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.

[11] 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.

[12] Goodfellow I, Bengio Y, Courville A. Deep learning [M]. Cambridge: The MIT Press, 2016.

[13] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[14] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[15] 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.

[16] Jang E, Gu S X, Poole B. Categorical reparameterization with Gumbel-Softmax[EB/OL]. (2016-11-03) [2021-05-04]. https://arxiv.org/abs/1611.01144v5.

[17] Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition, June 20-25, 2009, Miami, FL, USA. New York: IEEE Press, 2009: 248-255.

[18] Kingma D P, Ba J. Adam: a method for stochastic optimization[EB/OL]. (2014-12-22) [2021-04-05]. https://arxiv.org/abs/1412.6980.

[19] 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) [2021-01-04]. https://arxiv.org/abs/1706.09579.

[20] 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.

[21] Ding J, Xue N, Long Y, et al. Learning RoI transformer for detecting oriented objects in aerial images[EB/OL]. (2018-12-01)[2021-04-05].https://arxiv.org/abs/1812.00155.

[22] Yang X, Yang J R, Yan J C, et al. Towards more robust detection for small, cluttered and rotated objects [EB/OL]. (2018-11-17)[2021-02-04]. https://arxiv.org/abs/1811.07126.

[23] Yi J R, Wu P X, Liu B, et al. Oriented object detection in aerial images with box boundary-aware vectors[C]// 2021 IEEE Winter Conference on Applications of Computer Vision, January 3-8, 2021, Waikoloa, HI, USA. New York: IEEE Press, 2021: 2149-2158.

[24] Zhou X Y, Wang D Q, Krähenbühl P. Objects as points [EB/OL]. (2019-04-16)[2021-04-05]. https://arxiv.org/abs/1904.07850.

[25] Zhu X Z, Su W J, Lu L W, et al. Deformable DETR: deformable transformers for end-to-end object detection [EB/OL]. (2020-10-08)[2021-04-05]. https://arxiv.org/abs/2010.04159.