

中图法分类号: TP391 文献标识码: A 文章编号: 1006-8961(2025)03-0696-14

论文引用格式: Yu L X, Hao J and Zuo L. 2025. Supervised attention-based oriented object detection in remote sensing images. Journal of Image and Graphics, 30(3):0696-0709(余凌霄, 郝洁, 左量. 2025. 基于监督注意力的遥感图像定向目标检测. 中国图象图形学报, 30(3):0696-0709)
[DOI:10. 11834/jig. 240247]

基于监督注意力的遥感图像定向目标检测

余凌霄¹, 郝洁^{1,2*}, 左量¹

1. 南京航空航天大学计算机科学与技术学院, 南京 210016; 2. 软件新技术与产业化协同创新中心, 南京 210023

摘要: 目的 任意方向的多尺度目标和复杂的背景信息,使得在遥感图像上的目标检测相比一般目标检测任务具有更大的挑战。尽管一些现有的检测模型取得了令人满意的成果,但它们主要基于锚框实现,其检测性能依赖于预定义锚框的设计。本文在 Faster R-CNN(faster region-based convolutional neural network)的基础上,通过结合无锚检测思想与监督掩码注意力技术,提出了一种新的两阶段无锚检测模型。**方法** 首先,对于骨干网络提取的特征金字塔构建一个监督掩码注意力模块,通过注意力机制和掩码监督方法引导检测模型关注目标区域,减少背景噪声的干扰,从而提高目标特征的质量。其次,结合 FCOS(fully convolutional one-stage object detection)的回归思想以及中心点偏移技术设计了一个基于关键点预测的无锚定向区域提议网络,并且在训练阶段采用动态调整的软标签策略,实现标签的合理分配,提高模型的检测精度。**结果** 在两个公开的遥感数据集 DOTA(dataset for object detection in aerial images)和 HRSC(high-resolution ship collections)2016 上进行了大量实验,平均精确率均值(mean average precision, mAP)分别达到 76.36% 和 90.51%,超过大多数的定向检测模型,表明了本文方法的先进性和有效性。**结论** 本文提出的检测模型通过将区域提取网络的无锚框设计、监督方法以及注意力机制进行结合,能够有效适应复杂遥感图像中的定向目标检测场景。

关键词: 深度学习; 遥感图像; 定向目标检测; 无锚框; 注意力机制; 多尺度特征融合

Supervised attention-based oriented object detection in remote sensing images

Yu Lingxiao¹, Hao Jie^{1,2*}, Zuo Liang¹

1. School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China;

2. Center for Collaborative Innovation in Software Technology and Industrialization, Nanjing 210023, China

Abstract: Objective With the emergence of convolutional networks in recent years, deep learning-based object detection algorithms for remote sensing images have achieved considerable effectiveness. Compared to object detection in natural scenes, object detection in remote sensing images faces remarkable challenges, including arbitrary orientation, dense arrangements, and multiscale distributions. Traditional bounding box annotations use horizontal bounding boxes aligned with the coordinate axes to represent object locations. However, using horizontal boxes as anchor boxes or proposals for detecting oriented objects leads to notable drawbacks. Therefore, object detection in remote sensing images generally employs oriented bounding boxes to accurately describe the objects. However, the existing oriented object detection algorithms mostly identify potential regions using predefined anchor boxes. While these predefined anchor boxes can assist oriented object detection algorithms in recognizing objects of different scales and shapes, they also have notable limitations.

收稿日期: 2024-05-16; 修回日期: 2024-08-13; 预印本日期: 2024-08-20

* 通信作者: 郝洁 haojie@nuaa.edu.cn

First, the anchor boxes are predefined. Therefore, any deviation of object dimensions from these boxes can result in decreased detection accuracy, potentially leading to higher miss rates or false detection rates. In addition, the number, scale, and aspect ratio of anchor boxes must be determined based on experience or parameter tuning, which can hinder the model's capability to generalize across different scenarios or datasets. Second, anchor-based object detection models often define a large number of anchor boxes to cover objects of various sizes and aspect ratios to achieve a high recall rate. However, this phenomenon notably increases the computational complexity and further exacerbates the positive-negative sample imbalance problem. Additionally, remote sensing images often depict highly complex scenes, such as dense urban buildings, which introduce a substantial amount of disturbing information in the images. This complexity makes it challenging for traditional feature extraction networks based on classic backbone networks and feature pyramid networks to accurately extract and highlight the important features of the objects. This paper proposes a novel high-precision oriented object detection model for remote sensing images by combining supervised mask attention module (SMAM) and anchor-free oriented region proposal networks (AFORPN) to address these challenges. **Method** In this paper, a two-stage detector based on region-based convolutional neural network (Faster R-CNN) is introduced for oriented object detection in remote sensing images. This model comprises the following four components: the feature extraction backbone network, the SMAM, the AFORPN, and the R-CNN. The main contributions are as follows: An SMAM is constructed for the extracted feature pyramid to focus more on the object regions, suppress background noise, and achieve fine feature extraction. The SMAM comprises three parts: multiscale feature fusion, spatial attention, and supervised mask enhancement module. First, a sub-pixel convolution technique is adopted for the multiscale fusion of feature pyramids to achieve the unity of spatial scales by upsampling the feature maps through convolutional learning and channel rearrangement, retaining more object information compared with adjacent upsampling. Then, the channel attention mechanism is used to learn the weight coefficients of each channel and adaptively fuse different levels of feature maps, improving the representation capability of the model for input features. Subsequently, a self-attention mechanism is employed in the design of the spatial attention module for the processing of fused feature maps, allowing each pixel in the feature map to consider the information from all other pixels, thereby establishing a global wise dependency. This approach contributes to an improved understanding of the image background and semantic correlations within the model, thereby enhancing its capability to comprehend the surrounding environment and suppress the blending effects due to fusion. Finally, in the design of the supervised mask enhancement module, the model is guided to learn the semantic features and object contour information in the form of mask loss feedback, which drastically improves the accuracy of classification and localization. An AFORPN is proposed to avoid complex and redundant anchor box designs. The AFORPN comprises localization and classification branches. In the design of the localization branch, the keypoint regression method based on fully convolutional one-stage object detection (FCOS) is adopted, and the gliding vertex technique is introduced to achieve the generation of oriented bounding boxes. The prediction of midpoint offsets effectively alleviates the sensitivity issue to angle variations, allowing the bounding boxes to better conform to object shapes and improving keypoint regression performance. During the training stage, based on the spatial alignment between samples and objects, as well as the regression performance of the samples, a new label assignment criterion is proposed. This criterion gradually improves the weight assignment of regression performance through dynamic adjustment, enabling accurate label assignment and effectively mitigating the inconsistency between classification and regression that may arise from static heuristic assignment rules. **Result** Ablation analysis is performed on the dataset for object detection in aerial images (DOTA) dataset to demonstrate the effectiveness of the method. Experimental results showed that the proposed SMAM and AFORPN improved the mean accuracy precision (mAP) by 1.52% and 2.65%, respectively, compared to the baseline. The proposed approach is evaluated on the DOTA and HRSC2016 datasets, comparing it with other oriented detection algorithms to demonstrate its state-of-the-art performance. Without special processing, a mAP of 75.36%, which is higher than most oriented object detection models, is achieved on the DOTA dataset. On the HRSC dataset, the model also achieved a mAP up to 90.51%. **Conclusion** Extensive experimental results demonstrate that the proposed SMAM improves the quality of feature maps, while the proposed AFORPN generates high-quality region proposals, thereby further enhancing the detection performance of oriented objects. Overall, the proposed oriented detection model, which combines SMAM with AFORPN, exhibits promising detection capabilities and can effectively adapt to complex oriented object detec-

tion scenarios in remote sensing images.

Key words: deep learning; remote sensing image; oriented object detection; anchor-free; attention mechanism; multi-scale feature fusion

0 引言

随着遥感技术的进步,精准定位和分类遥感图像中的特定目标对城市规划、土地利用和资源调查等领域变得至关重要。这使得遥感图像目标检测方法受到了广泛关注。

在深度学习和卷积神经网络(convolutional neural network, CNN)的推动下,基于水平边界框(horizontal bounding box, HBB)的现代通用目标检测方法取得了巨大成功(聂光涛和黄华, 2021)。这些方法主要采用两种策略:两阶段检测和单阶段检测。其中两阶段检测策略首先使用RPN(region proposal network)提取区域提议(proposals),然后通过CNN对区域提议进行分类和边界框细化。代表性的算法有Faster R-CNN(faster region-based convolutional neural network)(Ren等, 2017)和Mask R-CNN(He等, 2017)等。单阶段检测策略则直接利用单个CNN估计类别概率和边界框偏移。代表性的算法有YOLO(you only look once)(Redmon等, 2016)、SSD(single shot multibox detector)(Liu等, 2016a)和RetinaNet(Lin等, 2017b)等。

然而,遥感图像中大部分目标呈现任意方向和密集排列的特点,因此HBB检测算法并不适合直接应用于遥感图像。为了解决上述挑战,相关研究引入方向信息扩展HBB算法,提出了定向边界框(oriented bounding box, OBB)检测算法。目前主流的定向检测算法大多基于密集放置的预定义锚框设计。Ma等人(2018)使用RRPN(rotated region proposal network)和RRoI(rotated region of interest)池化输出任意方向的OBB;RoI Transformer(Ding等, 2019)提出了监督旋转框机制,实现将RoI转换成RRoI;Gliding vertex(Xu等, 2021)提出在HBB的4条对应边上滑动每个顶点来表示OBB。

虽然锚框的设计提高了模型的精度,但也带来了新的问题。1)大量的预定义锚框增大了模型的计算复杂度,而且在训练过程中大部分的锚框被定义为负样本,使得出现严重的正负样本失衡问题。2)

由于锚框的尺寸和宽高比固定,模型在处理极端尺寸的检测目标时可能会遇到匹配困难的问题。

鉴于上述问题,无锚检测模型由于结构简单,并且无需预先定义锚框而受到越来越多的关注。无锚模型通过相关关键点(例如角点、极值点或中心)来表示正样本,并直接从特征中回归出目标的类别和位置。例如,基于HBB的检测算法中,CenterNet(Duan等, 2019)使用一个中心点来表示目标,并直接根据中心位置的特征回归其他属性,如目标大小、维度和姿态。FCOS(fully convolutional one-stage object detection)(Tian等, 2019)通过预测像素点到水平框4条边的距离来检测物体的边界框。而在OBB算法中,IENet(interactive embranchment network)(Lin等, 2024)基于FCOS(Tian等, 2019),使用了独立分支来回归方向参数。Axis learning(Xiao等, 2020)通过直接预测特征图每个像素的轴和宽度来检测目标。BBA-Vectors(box boundary-aware vectors)(Yi等, 2021)基于CenterNet(Duan等, 2019),通过设计一组框边界感知向量来定位OBB。这些无锚方法在处理计算复杂性和性能之间的权衡方面显示出极大的潜力。

此外,特殊的视角使得遥感图像相比一般图像包含着更多且复杂的背景信息,这给模型提取目标特征增加了困难。大量的研究已经表明,注意力机制在图像分类(Wang等, 2017; Zhong等, 2017)和一般物体检测(Liu等, 2016a)等视觉应用中取得不错的性能。因此,部分研究同样也在遥感图像的检测模型的构造中结合注意力模块的设计。例如, Yang等人(2019)使用由像素注意网络和通道注意网络组成的有监督多维注意网络来抑制噪声并突出前景; Li等人(2020)提出了一种多尺度空间和通道注意力机制,使网络像人类视觉一样更多地关注航空图像中的物体; Liu等人(2022)使用中心-边界双重注意模块来提取定向目标的中心和边界区域的特征; Zhang等人(2021)使用方向注意模块来增强方向敏感度,以实现精确的旋转目标回归。

综上,本文提出了一种新的无锚定向检测模型。首先,针对遥感图像中的背景干扰问题,通过结合掩码监督方法和注意力机制,突出遥感图像中感兴趣

区域,使得模型更准确地学习目标的轮廓信息和语义信息。其次,对于复杂冗余的预定义锚框,本文结合FCOS的无锚框思想设计RPN,通过关键点直接回归OBB,实现输出高质量区域提议。其中对于角度信息的预测,本文选择采取中心点偏移预测的方式,这有效缓解了大宽高比目标的角度敏感问题。此外,在RPN的训练阶段,本文采用一种软标签分配策略来替代启发式预定义规则的静态标签分配策略。这种软标签分配依据结合了样本的空间先验信息和生成高质量预测的能力,能够实现更精确的标签分配,从而提高模型性能。

本文的主要创新在于:1)通过结合注意力机制和掩码监督方法,设计了一个监督掩码注意力模块(supervised mask attention module, SMAM),从而提高了模型的特征提取能力;2)基于FCOS与中心点偏移技术,设计了一个基于关键点的无锚框定向区

域提取网络(anchor-free oriented region proposal network, AFORPN),实现输出高质量区域提议;3)结合样本的空间先验信息和生成高质量预测的能力,提出了一种适用于无锚检测的软标签分配策略,实现正样本的动态选取,从而优化了样本质量。

1 模型

如图1所示,本文提出的检测模型基于两阶段目标检测模型Faster R-CNN,包含以下4个部分:1)用于特征提取的骨干网络Backbone;2)用于融合多尺度特征,并通过注意力机制进行背景抑制的监督掩码注意力网络SMAM;3)用于多尺度目标检测并筛选高质量区域提议的无锚定向区域提议网络AFORPN;4)用于提议框的分类和回归调整,并生成最终的检测结果的R-CNN Head网络。

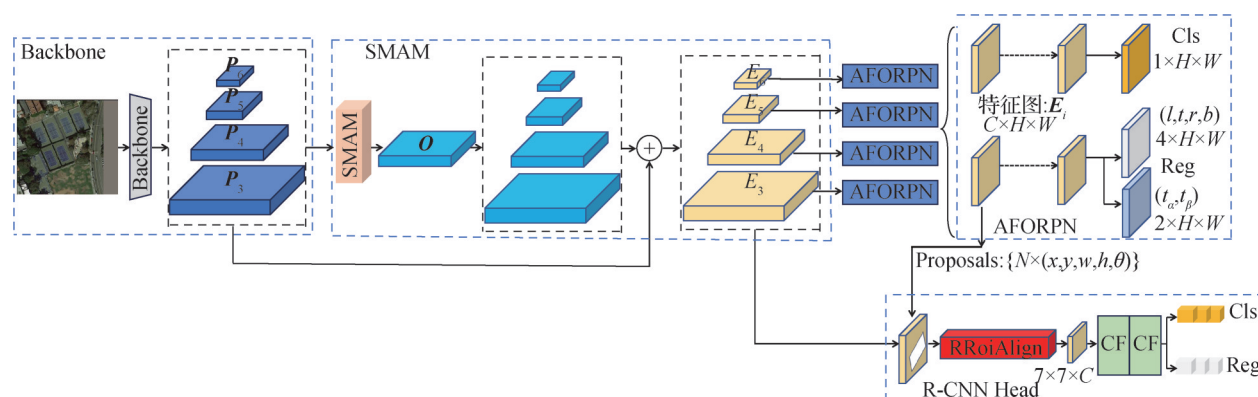


图1 模型框架

Fig. 1 Model framework

1.1 特征提取骨干网络

对于特征提取骨干网络,本文采用ResNet(residual network)(He等,2016)以及FPN(feature pyramid network)。其中ResNet利用若干个残差卷积获取不同粒度的特征图(feature map)。FPN通过自上而下的路径从顶层到低层逐步上采样低分辨率特征图,然后与邻近的高分辨率特征图进行融合,实现对多尺度特征的有效整合,进而得到最终的特征图集合 $\{P_3, P_4, P_5, P_6\}$,这些特征图相对于输入图像具有 $\{8, 16, 32, 64\}$ 像素的感受野。

1.2 监督掩码注意力模块

与自然图像检测相比,遥感图像的检测面临着更多的挑战。例如复杂背景和模糊区域的影响使得

目标与背景特征之间的边界变得模糊不清;以及同一目标在不同图像中表现出巨大的特征差异等。这些因素加剧了特征提取的难度,从而对后续的检测造成干扰。而如上文所述,注意力机制和掩码监督方法可有效应对这些挑战。受此启发,本文提出SMAM,通过结合注意力机制和掩码监督方法使得模型更加关注目标的轮廓以及语义特征,用以协助后续的检测网络产生更精确的预测框。如图2所示,SMAM由3部分构成,分别是多尺度特征融合、空间注意力模块以及监督掩码增强模块。

1.2.1 多尺度特征融合

特征融合是SMAM中至关重要的一环。然而,如果直接依赖于邻近插值上采样和简单逐像素相加

其次,采用通道注意力自适应融合多尺度特征。SENet (squeeze-and-excitation networks) (Zhong 等, 2019)提出不同特征图通道在检测任务的贡献并不相同,简单的相加方法可能无法充分利用每个通道的信息。因此,本文引入了通道注意力机制,实现对各个通道权重的动态调整,使得关键特征通道获得更多关注,从而提高融合特征的质量。具体如图2所示,在获取到变换后的4层特征图后,模型将其在通道维度进行拼接并输入到CAM(channel attention module)模块。CAM通过全局池化操作以及全连接层对其进行处理并输出一个 $1 \times 1 \times 4C$ 通道注意力图。考虑到相同通道,不同层级的重要性也是不同的,本文将拼接特征以及注意力图重构为 $H \times W \times 4 \times C$ 和 $1 \times 1 \times 4 \times C$ 的形式,然后在层级维度上对变换后的权重进行softmax操作,得到归一化的通道权重,再将其与拼接特征进行逐像素相乘,最后通过特征图之间的逐像素相加操作,得到最终融合特征 $F \in \mathbb{R}^{H \times W \times C}$ 。

1.2.2 空间注意力模块

为了缓解融合可能出现的混叠效应以及捕获特征之间的长距离依赖关系来增强融合特征的表达,本文设计了一个空间注意力模块(spatial attention module, SAM)。利用输入序列内部的关联信息来加权每个元素的重要性,自适应地整合局部特征和全局依赖。具体如图2所示,首先将融合特征图 F 输入到3个 1×1 卷积进行降维得到3个特征图: $Q \in \mathbb{R}^{H \times W \times (C/k_1)}$ 、 $K \in \mathbb{R}^{H \times W \times (C/k_1)}$ 和 $V \in \mathbb{R}^{H \times W \times C}$ 。式中, k_1 表示放缩倍数,在本文设置为8。然后,对重构后的 Q, K 执行矩阵乘法,并通过归一化操作,得到空间像素注意力权重图 $D \in \mathbb{R}^{H \times W}$ 。此后,将注意力图 D 与特征图 V 执行矩阵相乘操作并进行重构,得到细化后的特征图 $F' \in \mathbb{R}^{H \times W \times C}$ 。

1.2.3 监督掩码增强模块

最后,在获得细化后的特征图 F' 后,监督掩码增强模块通过卷积、激活函数对其进行处理,分别生成预测掩码 $M \in \mathbb{R}^{H \times W \times N}$ (N 表示类别数目)和目标权重 $A \in \mathbb{R}^{H \times W \times C}$,然后通过掩码损失反馈引导网络提高目标权重,从而实现过滤背景特征,增强目标特征。如图2所示,监督掩码增强模块的构造由两个分支组成,其中预测掩码分支由 3×3 卷积、ReLU激活函数以及 1×1 卷积构成;目标权重分支由 1×1 卷积和sigmoid激活函数构成。具体为

$$\begin{aligned} M &= \text{Conv}_1(\text{ReLU}(\text{Conv}_3(F'))) \\ A &= \text{sigmoid}(\text{Conv}_1(F')) \\ O_{ij,c} &= A_{ij,c} \times F'_{ij,c} + F'_{ij,c} \end{aligned} \quad (2)$$

式中, c 表示通道索引, (i,j) 表示上采样特征图的空间位置。

此外,由于遥感检测数据集中没有语义掩码标注,因此本文将目标真值框内的像素填充为其类别标识,其他则填充为背景标识0,通过伪掩码的形式实现监督。而对于掩码监督损失,采取Focal Loss (Lin 等, 2017b)用以计算。同时考虑到特征图 O 中具有更多的目标信息,因此本文将特征图 O 通过缩放操作使得与原始FPN特征图 $\{P_3, P_4, P_5, P_6\}$ 保持大小一致,然后执行逐像素相加,得到最终增强特征图 $\{E_3, E_4, E_5, E_6\}$ 。

1.3 无锚框定向区域提议网络

如图1所示,经过SMAM增强后的特征图 $\{E_3, E_4, E_5, E_6\}$ 被输入到AFORPN网络中。AFORPN网络通过卷积对输入的特征图进行处理,生成一组稀疏的定向提议框。在处理过程中,AFORPN网络针对每个尺度的特征图进行单独处理,即在每个特征图上运行相同的AFORPN网络,以产生密集的预测框。随后,利用非极大值抑制方法从生成的预测框中获取最终的稀疏提议框。AFORPN网络的设计基于FCOS,其输出包含两个分支:第1个分支用于判断候选区域是否包含目标;而第2个分支则用于预测候选区域的边界框。以下对AFORPN的设计进行详细描述。

1.3.1 中心点偏移预测

如图4所示,一个OBB通常使用 (X_c, Y_c, W, H, θ) 来表示,其中 (X_c, Y_c) 表示中心点坐标, W, H 表示OBB的宽、高, $\theta \in [-90^\circ, 90^\circ)$ 表示对应于OBB的长边 W 与 x 轴之间的夹角。FSDet(detector with feature enhancement and soft label assignment) (Yu 等, 2022)提出在基于FCOS的HBB预测的基础上增加角度预测分支,用于进行定向检测。具体来说,对于特征图上一样本点 $P(x_a, y_a)$ 所预测的OBB, FSDet采用一个5维向量 $d = (l, t, r, b, \theta)$ 来解码获得,其中, l, t, r 和 b 表示样本 P 在原图坐标下到该OBB的左、上、右和下侧的垂直距离, θ 表示预测角度。

但是,在遥感图像中存在着大量的目标,它们的形状呈细长状,这使得其OBB对于角度变化十分敏

感,即使角度的轻微误差也会导致检测重叠度(intersection over union, IoU)的快速下降,从而影响检测精度。因此为了缓解角度的敏感问题,相对于角度的直接预测,本文采用了一种中心点偏移预测方法。如图4所示,中心点偏移预测方法通过一个HBB(X_c, Y_c, w, h)以及相对于其宽、高的中心点偏移距离(α, β)来表示OBB。通过回归偏移距离与对应边的长度比,有效表征HBB每侧的相对滑动偏移量。与FSDet提出的角度直接预测相比,这种偏移预测方式使得模型对于OBB方向的敏感性更小。因此,对于样本 $P(x_a, y_a)$,对应输入图像上的位置(x, y)可以表示为

$$x = \left\lfloor \frac{s_i}{2} \right\rfloor + x_a \times s_i, y = \left\lfloor \frac{s_i}{2} \right\rfloor + y_a \times s_i \quad (3)$$

式中, s_i 表示特征图 E_i 的感受野, $\lfloor \cdot \rfloor$ 表示向下取整。那么对应于该点的预测框的回归参数 $d = (l, t, r, b, t_\alpha, t_\beta)$ 可以表示为

$$\begin{cases} l = \frac{x - \left(x_c - \frac{w}{2}\right)}{s_i}, t = \frac{y - \left(y_c - \frac{h}{2}\right)}{s_i} \\ r = \frac{\left(x_c + \frac{w}{2}\right) - x}{s_i}, b = \frac{\left(y_c + \frac{h}{2}\right) - y}{s_i} \\ t_\alpha = \frac{\alpha}{w}, t_\beta = \frac{\beta}{h} \end{cases} \quad (4)$$

式中, (l, t, r, b) 表示该点到对应HBB四边的垂直距离与感受野的比值, (t_α, t_β) 表示中心点偏移量与HBB对应边的长度比。在该点为正样本时,其目标值计算同样如上所述。

1.3.2 软标签分配策略

FCOS根据样本与真值框的位置关系进行标签分配,落在框内的点视为正样本,落在框外的点视为

负样本。FCOSR(Li等,2023)则提出使用高斯密度概率函数来计算样本感受野与真值框的重合度,从而进行样本筛选。然而每个样本的相关特征感受野与任意方向的定向框并未很好地对齐,特别是对于长宽比较大的目标。这使得不同位置的正样本在回归定向框时表现出不一致的定位能力。因此,仅依赖于启发式预定义规则的标签分配可能不够准确,从而在训练过程中引入标签噪声。受DAL(dynamic anchor learning)(Qi等,2021)提出的匹配度的启发,本文提出了一种适用于无锚框定向目标检测的软标签分配策略。

首先,对于真值框的分配,本文采用多层预测的思想,根据其面积大小分配给指定特征层。具体来说,对于一个给定的特征图 E_i ,其面积分配区间为 $[k_2 s_i^2/4, 4k_2 s_i^2]$,其中 k_2 作为超参数用以调控范围,在本文设置为16, s_i 表示特征图 E_i 的感受野。对于特征图 E_3 分配区间的最小值,本文设置为0,对于 E_6 的最大值,本文设置为1 000 000。

然后,针对特征图 E_i 中的某一个样本 $P(x, y)$ 的标签分配,在本文的设计中,综合考虑了其分配给 E_i 中任一真值框 $G(X_c, Y_c, W, H, \theta)$ 的空间先验信息以及所预测的定向框 B_p 与 G 的IoU值。其标签筛选标准定义为

$$f(P) = \lambda g(P) + (1 - \lambda) \text{IoU}(B_p, G) \quad (5)$$

式中, λ 表示权重因子, $g(P)$ 表示一个高斯密度函数值,反映了样本 P 与真值框 G 的空间先验信息。具体而言,本文通过一个未归一化的2D高斯密度函数来表示真值框 G 与样本 P 的感受野的重合度,函数定义为

$$g(P) = \exp\left(-\frac{1}{2} (P - \mu)^T \Sigma^{-1} (P - \mu)\right) \quad (6)$$

式中, $g(P) \in (0, 1)$, Σ 表示协方差矩阵, μ 表示均值。对于参数 Σ, μ 的设计为

$$\Sigma = R_\theta \times \Sigma_0 \times R_\theta^T, \mu = (X_c, Y_c) \\ R_\theta = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}, \Sigma_0 = \frac{1}{12} \begin{bmatrix} W^2 & 0 \\ 0 & H^2 \end{bmatrix} \quad (7)$$

式中, Σ_0 表示角度等于0时的协方差矩阵, R_θ 表示旋转变换矩阵。本文将 $g(P)$ 的阈值设置为0.23,以在训练过程中避免偏离目标区域的样本干扰,并确保图像中的小目标能得到一定的候选正样本覆盖。因此,对于小于阈值的样本,将其划分为负样本,而对于剩下的候选正样本,采用软标签分配策略进行再

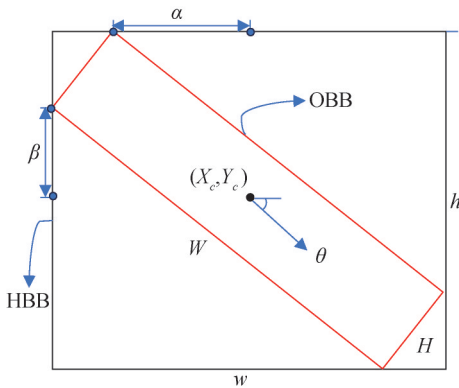


图4 边界框参数说明

Fig. 4 Explanation of bounding box parameters

次筛选。其中对于落在重合区域的候选正样本, 将其分配给 $g(\mathbf{P})$ 取值较大的真值框。

为了实现更稳定的训练并且提高正样本的质量, 本文迭代调整训练过程中样本的空间先验信息的权重。其设置为

$$\lambda(t) = \begin{cases} 1 & t < 0.1 \\ 5(\lambda_0 - 1) \times t + 1.5 - 0.5 \times \lambda_0 & 0.1 \leq t < 0.3 \\ \lambda_0 & t \geq 0.3 \end{cases} \quad (8)$$

式中, t 表示当前训练迭代次数与总训练迭代次数的比值。在本文中, 将 $f(\mathbf{P})$ 的阈值设置为 0.5, λ_0 设置为 0.25。同时对于每一个真值框 G 的所有候选正样本, 本文将其中 $f(\mathbf{P})$ 值最大的样本也划分为正样本, 以保证每一个真值框都有正样本。

1.3.3 损失函数

AFORPN 的损失由分类损失和回归损失构成。沿用 Focal Loss (Lin 等, 2017b) 计算分类损失虽然很好地解决了正负样本的不平衡问题, 但是却忽略了分类与回归的不一致问题导致在非极大值抑制阶段会出现高质量提议框由于低分类置信度而被过滤的现象。这主要是由于分类与回归任务在训练过程中是独立优化的, 两者之间缺乏交互。为了缓解上述问题, 本文使用一致性损失 (quality focal loss) (Li 等, 2023) 作为分类损失函数, 将目标的定位质量直接融合到分类损失中。这不仅能很好地解决正负样本的不平衡问题, 还能引导分类与回归联合优化。分类损失的定义为

$$L_{\text{QFL}}(\sigma) = -|I - \sigma|^{\eta} ((1 - I) \log(1 - \sigma)) + I \log(\sigma) \quad (9)$$

式中, I 表示预测框与真值框的 IoU 值, σ 表示相应的类别分数, η 表示权重因子。对于回归损失, 为了缓解距离回归与角度回归的量纲不一致性问题, 本文采用 Yu 等人 (2022) 提供的回归损失设计, 将其划分为两部分, 分别是水平边界框的回归损失以及角度的中心点偏移回归损失。

$$L_{\text{reg}}(\mathbf{d}, \mathbf{d}^*) = L_{\text{HBB}}((l, t, r, b), (l^*, t^*, r^*, b^*)) + \lambda_2 L_{\text{angle}}((t_{\alpha}, t_{\beta}), (t_{\alpha}^*, t_{\beta}^*)) \quad (10)$$

式中, $\mathbf{d} = (l, t, r, b, t_{\alpha}, t_{\beta})$ 表示回归预测参数, 同理 $\mathbf{d}^*$ 表示相应的目标值。 L_{HBB} 采用的是 GIoU (generalized intersection over union) 损失 (Rezatofighi 等, 2019), L_{angle} 采用的是 Smooth L1 损失。对于角度损失权重, 本文保留了 Yu 等人 (2022) 提供的设计, 其

值为 0.2。

1.4 R-CNN Head

如图 1 所示, R-CNN 负责对 AFORPN 生成的提议框进行处理。首先依据 Lin 等人 (2017a) 提出的映射规则, R-CNN 将提议框映射到指定的特征图层; 接着通过 RoiAlign 操作将区域特征调整为 7×7 的特征图块; 然后将其输入到后续的全连接层进行分类和提议框回归调整; 最终输出检测结果。

对于提议框的回归, 采用上述的中心点偏移预测方法。如图 4 所示, 对于一个提议框 $(x_p, y_p, W_p, H_p, \theta_p)$, 同样可以使用 $(x_p, y_p, w_p, h_p, \alpha_p, \beta_p)$ 表示, 则其预测参数 $\mathbf{t} = (t_x, t_y, t_w, t_h, t_{\alpha}, t_{\beta})$ 表示为

$$\begin{cases} t_x = \frac{(x - x_p)}{w_p}, t_y = \frac{(y - y_p)}{h_p} \\ t_w = \log\left(\frac{w}{w_p}\right), t_h = \log\left(\frac{h}{h_p}\right) \\ t_{\alpha} = \frac{\alpha - \alpha_p}{w_p}, t_{\beta} = \frac{\beta - \beta_p}{h_p} \end{cases} \quad (11)$$

式中, $(x, y, w, h, \alpha, \beta)$ 表示经过 R-CNN 调整后的提议框。同理, 其目标值 $\mathbf{t}^* = (t_x^*, t_y^*, t_w^*, t_h^*, t_{\alpha}^*, t_{\beta}^*)$ 同样如上述计算。

对于 R-CNN 的训练, 本文沿用原始 Faster R-CNN 的样本分配策略和损失函数。

2 实验

2.1 数据集和实验环境

DOTA1.0 (dataset for object detection in aerial images) (Xia 等, 2018) 是一个用于遥感图像目标检测的大规模数据集, 具有 OBB 注释, 包含 2 806 幅图像和 188 282 个目标。这些目标主要分为 15 个类别, 包括飞机 (PL)、棒球场 (BD)、桥梁 (BR)、田径场 (GTF)、小型车辆 (SV)、大型车辆 (LV)、船 (SH)、网球场 (TC)、篮球场 (BC)、储罐 (ST)、足球场 (SBF)、环形交叉路口 (RA)、港口 (HA)、游泳池 (SP) 和直升机 (HC)。本文使用标准的训练、验证和测试数据划分, 分别包含 1 411、458 和 937 幅图像。由于 DOTA 数据集中的图像尺寸范围从 800×800 像素到 $4\,000 \times 4\,000$ 像素不等, 本文使用官方提供的工具包将原始图像裁剪成重叠度为 200 像素, 尺寸为 $1\,024 \times 1\,024$ 像素的子图像。通过这些处理, 本文提取了 9 736 个训

练数据,3 066个验证数据,以及1 0833个测试数据。

HRSC2016 (high-resolution ship collections) (Liu 等, 2016b) 是一个高分辨率船舶检测数据集, 包含了 1 061 幅图像, 其中训练集、验证集和测试集分别包含 436 幅图像、181 幅图像和 444 幅图像, 尺寸范围从 300×300 像素到 $1\,500 \times 900$ 像素不等, 并且涵盖了 20 多个类别的船舶, 具有多样的外观特征。本文将所有图像的短边调整为 800 像素, 长边根据原始图像的长宽比自适应调整大小。

本文模型是在 MMRotated 框架上实现。除非另有说明, 所有实验的主干网络都采用在 ImageNet (Krizhevsky 等, 2009) 上预训练的 ResNet 网络。模型使用随机梯度下降优化器进行优化, 权重衰减和动量分别设置为 0.000 1 和 0.9, 并且在训练过程中使用随机翻转和随机旋转进行数据增强。在 DOTA 数据集上, 模型共训练 12 个 epoch, 初始学习率为 0.005, 在第 8 和第 11 个 epoch 时学习率减小 10 倍。在 HRSC 数据集上, 模型共训练 24 个 epoch, 学习率与上述相同, 并在第 16 和第 22 个 epoch 时进行衰减。本文在搭载了一个 NVIDIA GTX 2080ti 的设备上训练模型, 批量大小设置为 2。

本文采用平均准确率均值 (mean average precision, mAP) 评估所有实验的检测精度。

2.2 检测结果分析

为了详细分析每个提出的模块, 在 DOTA 数据集上进行消融实验。为了公平比较, 除了对相应模块进行指定更改之外, 所有实验使用相同的参数设置。消融实验结果如表 1 所示, FCOS-based 表示在 FCOS 的基础上添加角度分支来实现定向检测。

1) SMAM 模块的消融分析。通过表 1 可以看出, 相比基线网络 72.97% 的 mAP 值, 使用本文提出的 SMAM 进行特征增强后, 模型的 mAP 值达到了 74.49%, 实现了 1.52% 的提升, 表明了 SMAM 的优越性。为了进一步理解 SMAM 的效果, 图 5 展示了不同级别特征图的热图。第 1 行展示了原图以及 FPN 的输出特征图 $\{P_3, P_4, P_5, P_6\}$ 的热图, 第 2 行则展示了相应的掩码标注以及经过 SMAM 增强后的特征图 $\{E_3, E_4, E_5, E_6\}$ 的热图。可以观察到, 通过 SMAM 增强后, 有效减少了 FPN 自顶向下过程中的冗余信息, 加强了每个级别的目标特征, 同时提高了目标区域的显著性, 使得网络输出特征图的边界更加清晰, 从而改善了最终的检测结果。

表 1 所提出模块在 DOTA 数据集上的消融分析

Table 1 Ablation analysis of proposed modules on DOTA dataset

SMAM	FCOS-based RPN	AFORPN	mAP/%
—	—	—	72.97
√	—	—	74.49 (+1.52)
—	√	—	73.42 (+0.45)
√	√	—	74.96 (+1.99)
—	—	√	75.62 (+2.65)
√	—	√	76.36 (+3.39)

注: “√”表示采用, “—”表示未采用。

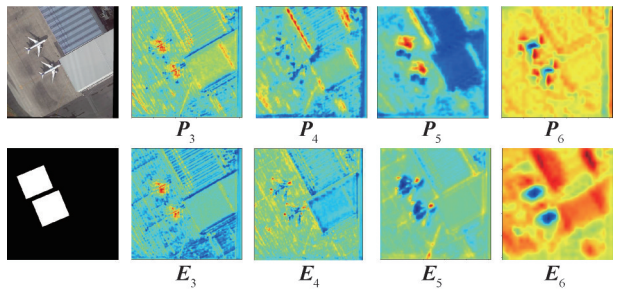


图 5 SMAM 增强前后的不同级别特征图热图对比

Fig. 5 Comparison of heatmaps of different levels of feature maps before and after SMAM enhancement

此外, 本文对 SMAM 模块的设计进行了消融分析。表 2 展示了以 Faster R-CNN 为基线, 结合 SMAM 各项设计后的模型检测结果。通过分析表中数据可以看出, 与 CBAM 实现的 0.34% 性能提升相比, 本文提出的空间注意力机制实现了 0.65% 的提升。这表明相对于增大卷积核来扩大特征感受野, 通过自注意力机制来建立特征之间的长距离依赖更能够增强对目标特征的学习, 进而提高检测性能。另外, 在将自适应多尺度融合模块用于特征融合时, 性能提升了 1.19%。这表明了亚像素卷积上采样能够更好地保留目标的细节信息, 而通道注意力模块的自适应融合使得网络能够以动态选择的方式融合不同层级特征图, 进而提升模型性能。另外, 通过增加掩码监督同样给模型的检测性能带来了提升。这些实验结果显示了所提出模块的有效性和先进性, 同时也证明了这些设计的结合也是有效的。

2) AFORPN 模块的消融分析。正如表 1 的第 4 行和第 6 行所示, 采用无锚框思想设计的 RPN 能够生成高质量的提议框, 并显著提升模型的检测精度。

表 2 SMAM在DOTA数据集上的消融分析
Table 2 Ablation analysis of SMAM on the DOTA dataset

CBAM	空间注意力	自适应多尺度融合	掩码监督	mAP/%
-	-	-	-	72.97
√	-	-	-	73.31(+0.34)
-	√	-	-	73.62(+0.65)
-	√	√	-	74.16(+1.19)
-	√	√	√	74.49(+1.52)

注:“√”表示采用,“-”表示未采用。

不过从表中可以观察到,相较于直接使用FCOS作为RPN设计,本文提出的AFORPN模块的提升效果更为显著。图6给出了RPN网络、FCOS-based RPN网络以及AFORPN网络对于大尺度目标、密集目标、大宽高比目标以及复杂背景情况下小目标的检测结果。其中,蓝色框表示检测结果,红色框标注了各种RPN的明显差异。从图中可以对比看出,本文的AFORPN网络所得到的检测结果具有更高的准确性。而且相比于RPN和FCOS-based RPN,AFORPN输出的检测框与目标更加贴合。

此外,本文对AFORPN模块的设计进行了消融分析。表3给出了以FCOS-based RPN为基线,结合AFORPN中的各个设计的模型的部分检测结果。从表中可以看出,结合AFORPN设计的模型在检测结果中,无论是类别性能还是总体性能方面都有不同程度的提升。具体来说,中心点偏移带来了1.51%的性能提升,值得注意的是,大宽高比类别HA实现了2.81%的提升;软标签策略带来了0.52%的性能提升;一致性损失带来了0.69%的提升。这些结果表明了AFORPN中各个设计的有效性。

2.3 检测结果比较

为了进一步验证本文方法的有效性,与DOTA数据集上的几种先进方法进行比较。其中,本文所有的结果都以固定格式在DOTA评估服务器上进行评估。表4显示了本文方法与目前先进方法的定量比较。在DOTA数据集上,模型的特征提取骨干网络与在HRSC数据集上保持一致(详见表5)。从表4可以观察到,使用ResNet50为骨干网络时,本文方法实现了最佳效果,mAP达到了76.36%,高于大多先进方法。值得注意的是,本文的模型在BR、LV、HA、SBF和TC这5个类别中表现最优。对于其他类

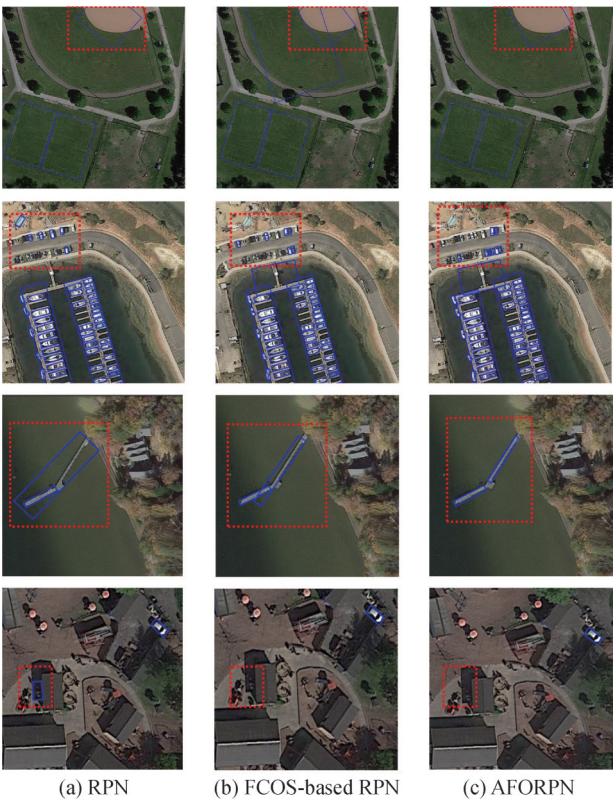


图 6 不同RPN网络在DOTA数据集上的可视化结果
Fig. 6 The visualization results of different RPN networks on the DOTA dataset ((a) RPN; (b) FCOS-based RPN; (c) AFORPN)

表 3 AFORPN在DOTA数据集上的消融分析
Table 3 Ablation analysis of AFORPN on DOTA dataset /%

方法	SV	LV	SH	HA	GTF	mAP
基线	77.44	80.71	87.53	72.95	68.56	73.42
+中心点偏移	78.68	82.28	88.01	75.76	73.76	74.93
+软标签策略	78.28	81.27	87.95	73.62	71.37	73.94
+一致性损失	77.94	81.54	87.65	73.58	72.47	74.11
AFORPN	78.61	82.30	88.04	75.60	74.88	75.62

注:“+”表示在基线的基础上增加相应设计。

别,虽然AP值不是最优,但是大多与最优结果相近,如PL、BC、SV等。这表明了本文方法具有较大的潜力。此外,图7给出了本文方法在DOTA各个类别的可视化结果。从图中观察到,本文方法在各种场景中对于不同尺度和外观多样性的目标都表现良好。对于图像中密集排列的船舶和车辆等目标,能够实现准确的区分和精准的定位。此外,对于例如港口和桥梁等具有大宽高比的目标,本文方法也可以精确地检测它们。这表明本文方法可以处理所有

表 4 在DOTA数据集上与当前先进方法的对比
Table 4 Comparison with state-of-the-art methods on the DOTA dataset

/%																	
方法		PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
单阶段	PIoU	80.90	69.70	24.10	60.20	38.30	64.40	64.80	90.90	77.20	70.40	46.50	37.10	57.10	61.90	64.00	60.50
	DAL	88.68	76.55	45.08	66.80	67.00	76.76	79.74	90.84	79.54	78.45	57.71	62.27	69.05	73.14	60.11	71.44
	FCOS-R	88.57	69.54	48.18	62.42	79.08	76.14	85.81	90.33	79.25	84.92	55.62	64.59	61.81	68.67	37.96	70.19
	R3Det	88.76	83.09	50.91	67.27	76.23	80.39	86.72	90.78	84.68	83.24	61.98	61.35	66.91	70.63	53.94	73.79
	RSDet	89.80	82.90	48.60	65.20	69.50	70.10	70.20	90.50	85.60	83.40	62.50	63.90	65.60	67.20	68.00	72.20
	S2ANet	89.11	82.84	48.37	71.11	78.11	78.39	87.25	90.83	84.90	85.64	60.36	62.60	65.26	69.13	57.94	74.12
双阶段	RRPN	80.94	65.75	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	60.01
	SCRDet	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
	Faster R-CNN	89.19	82.00	50.62	67.84	78.05	74.14	86.20	90.75	85.10	85.03	56.53	61.94	66.97	68.87	51.57	72.97
	Glid vertex	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
	RoI Trans	88.65	82.60	52.53	70.87	77.93	76.67	86.87	90.71	83.83	82.51	53.95	67.61	74.67	68.75	61.03	74.61
	Oriented R-CNN	89.46	82.12	54.78	70.86	78.93	83.00	88.20	90.90	87.50	84.68	63.97	67.69	74.94	68.84	52.28	75.87
本文(R50)		89.53	81.73	54.80	73.86	78.55	83.10	87.93	90.90	87.12	84.59	68.02	64.40	75.90	69.81	57.22	76.36
本文(R101)		88.64	81.54	54.49	74.02	78.69	83.30	88.12	90.90	87.24	85.22	67.64	64.20	75.21	69.26	56.87	76.31

注:加粗字体表示各列最优结果。

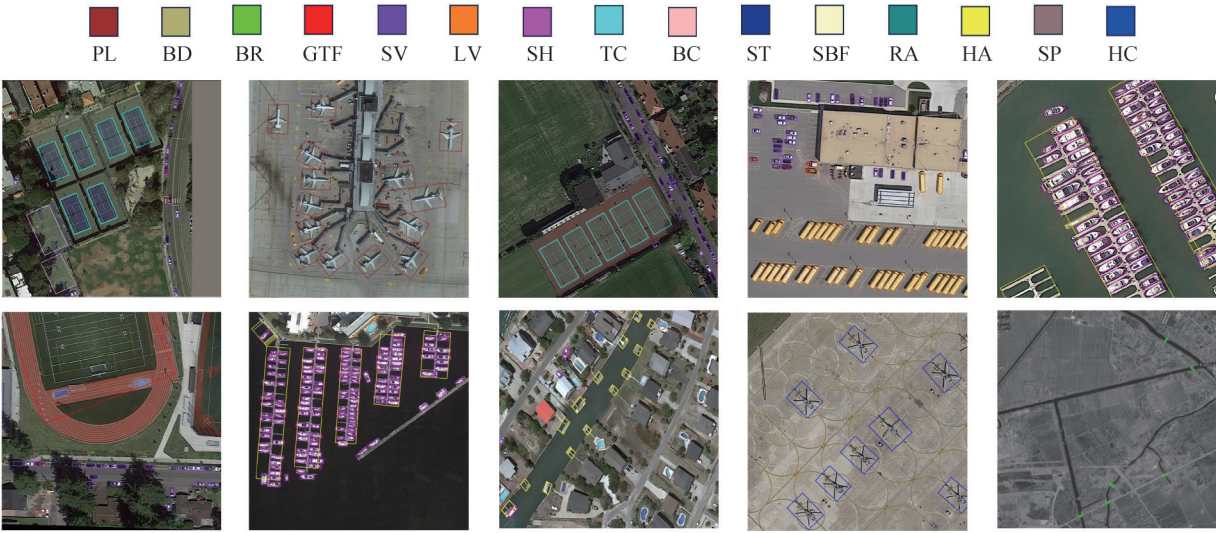


图 7 DOTA数据集上的可视化结果
Fig. 7 Visualization results on the DOTA dataset

这些具有挑战性的情况并显示出理想的结果,证明了所提出方法的稳健性。

为了进一步验证本文方法在特殊场景的检测性能,选用大宽高比船舰数据集 HRSC 进行实验,对比本文方法与其他算法的检测性能,结果如表 5 所示。从表 5 可以观察到,在使用 ResNet101 时,本文方法

实现了最佳效果,mAP 达到 90.51%,相比 Glid vertex 提升了 2.31%。

图 8 给出了 HRSC 的一些可视化检测结果。可以看到,本文方法能够准确地检测具有任意方向的船舶。此外,它在检测大宽高比和不同尺寸的船舶时表现了出色的性能。这些结果表明,本文方法在

表 5 在 HRSC 数据集上与当前先进方法的比较
Table 5 Comparison with state-of-the-art methods
on the HRSC dataset

	方法	特征提取网络	mAP/%
单 阶 段	PIoU	DLA-34	89.20
	DAL	R-50-FPN	89.77
	R3Det	R-101-FPN	89.26
	RSDet	R-101-FPN	86.50
	S2ANet	R-50-FPN	90.17
双 阶 段	RRPN	R-101	79.08
	Glid vertex	R-101-FPN	88.20
	Roi trans	R-50-FPN	86.20
	Oriented R-CNN	R-50-FPN	90.40
	本文	R-50-FPN	90.49
	本文	R-101-FPN	90.51

注:加粗字体表示最优结果。R (ResNet) 和 DLA (deep layer aggregation)表示骨干网络。

检测具有任意方向、不同尺度和大宽高比的目标时表现出很好的鲁棒性。

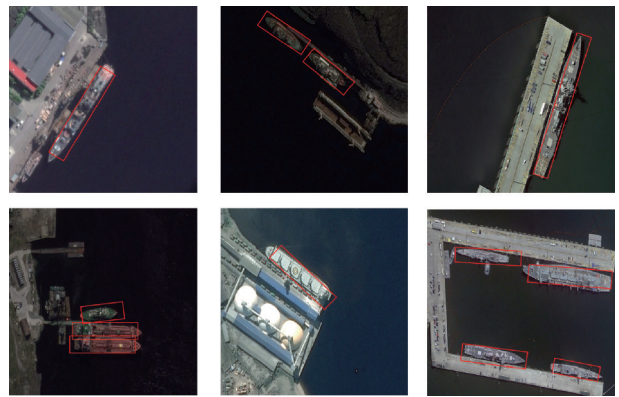


图 8 HRSC 数据集上的可视化结果
Fig. 8 Visualization results on the HRSC dataset

最后,本文对方法的不足之处进行了详细分析。从表 4 的数据可以看出,尽管本文方法在 BR 和 RA 等小目标检测上超越了大多数现有检测模型,但其检测精度分别仅有 54. 80% 和 64. 40%,小目标仍存在大量漏检现象。尽管本文引入了多尺度特征融合增强的方法,但由于特征采样过程中分辨率的下降,模型难以捕捉到检测小目标所需的足够细节信息,从而导致其定位和识别精度的不足。此外,正样本数量不足也限制了模型的学习效果,这进一步削弱

了其对小目标的识别能力,并且小目标容易被复杂背景所淹没,尤其是在场景中存在大量干扰物的情况下,模型对小目标检测的鲁棒性受到干扰,导致误检和漏检的情况增加。图 9 展示了原图以及在模拟现实云层影响和图像失真情况下的小目标检测结果,其中绿色标注表示与原图相比的误检和漏检结果。检测结果验证了上述分析的准确性。综上,尽管本文的方法在一定程度上提升了小目标检测的精度,但仍存在明显不足。在未来的研究中,将继续优化模型结构和特征提取策略,以进一步提升小目标检测性能。

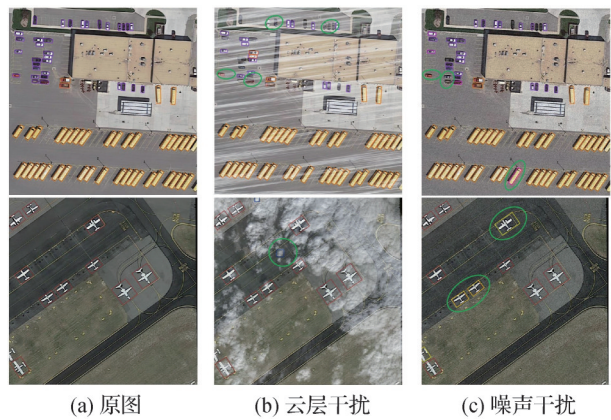


图 9 小目标在不同干扰下的检测结果对比
Fig. 9 Comparison of small object detection results under different interference conditions ((a) original images; (b) with cloud interference; (c) with noise interference)

3 结 论

针对复杂遥感图像的目标检测任务,提出了一种基于监督注意力的两阶段无锚定向目标检测模型。所设计的 SMAM 通过注意力机制和掩码监督方法引导网络提高对待检测区域的关注,从而更好地实现了对于任意方向目标对象特征的学习;所设计的 AFORPN 结合 FCOS 的关键点回归思想和中心点偏移技术既实现了无需预定义锚框也可生成高质量的定向区域提议,同时缓解了大量锚框带来的正负样本失衡问题;所设计的软标签策略综合多种信息作为样本筛选标准,有效改善了正样本的质量,缓解了无锚框的样本分配不准确问题。通过消融实验,验证了所设计模块的优越性。在两个公共数据集上的对比实验结果表明,本文方法显著优于其他先进

方法:在DOTA和HRSC数据集上的mAP分别达到76.36%和90.51%。然而,尽管本文方法表现优异,但由于检测过程需要通过SMAM模块进一步增强特征,这不可避免地增加了参数和计算成本。此外,对于在小目标特征不足的情况下,本文方法的检测效果未能实现显著改善。因此,在未来的工作中,将在如下几个方面继续开展研究:1)优化特征增强模块的设计,在减少计算负担的情况下,提高模型对于目标特征的捕捉能力;2)改进样本选择策略,实现在小目标样本不足的情况下,提高其检测精度。

参考文献(References)

- Cai Q, Qian Y, Li J, Lyu J, Yang Y H, Wu F and Zhang D. 2023. HIPA: hierarchical patch Transformer for single image super resolution. *IEEE Transactions on Image Processing*, 32: 3226-3237. DOI:10.1109/TIP.2023.3279977.
- Ding J, Xue N, Long Y, Xia G S and Lu Q K. 2019. Learning RoI Transformer for oriented object detection in aerial images//*Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, USA: IEEE: 2849-2858 [DOI: 10.1109/CVPR.2019.00294]
- Duan K W, Bai S, Xie L X, Qi H G, Huang Q M and Tian Q. 2019. Centernet: keypoint triplets for object detection//*Proceedings of 2019 IEEE/CVF International Conference on Computer Vision*. Seoul, Korea (South): IEEE: 6569-6578 [DOI: 10.1109/ICCV.2019.00669]
- He K M, Gkioxari G, Dollár P and Girshick R. 2017. Mask R-CNN. *Proceedings of 2017 IEEE International Conference on Computer Vision*, 45(10): 2961-2969 [DOI: 10.1109/ICCV.2017.322]
- He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]
- Krizhevsky A, Sutskever I and Hinton G E. 2012. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6): 84-90 [DOI: 10.1145/3065386]
- Li X, Lv C Q, Wang W H, Li G, Yang L F and Yang J. 2023. Generalized focal loss: towards efficient representation learning for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3139-3153 [DOI: 10.1109/TPAMI.2022.3180392]
- Li Y Y, Qin H, Xuan P, Jiao L C and Shang R H. 2020. RADet: refine feature pyramid network and multi-layer attention network for arbitrary-oriented object detection of remote sensing images. *Remote Sensing*, 12(3): #389 [DOI: 10.3390/rs12030389]
- Li Z, Hou B, Wu Z T, Ren B and Chen Y. 2023. FCOSR: a simple anchor-free rotated detector for aerial object detection. *Remote Sens*, 15(23): #5499 [DOI: 10.3390/rs15235499].
- Lin T Y, Dollár P, Girshick R, He K M, Hariharan B and Belongie S. 2017a. Feature pyramid networks for object detection [EB/OL]. [2024-05-16]. <https://arxiv.org/pdf/1612.03144.pdf>
- Lin T Y, Goyal P, Girshick R, He K M and Dollár P. 2017b. Focal loss for dense object detection//*Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy: IEEE: 2999-3007 [DOI: 10.1109/ICCV.2017.324]
- Liu S, Zhang L, Lu H C and He Y. 2022. Center-boundary dual attention for oriented object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1-14 [DOI: 10.1109/TGRS.2021.3069056]
- Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y and Berg A C. 2016a. SSD: single shot MultiBox detector//*Proceedings of the 14th European Conference on Computer Vision*. Amsterdam, the Netherlands: Springer: 21-37 [DOI: 10.1007/978-3-319-46448-0_2]
- Liu Z K, Wang H Z, Weng L B and Yang Y P. 2016b. Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds. *IEEE Geoscience and Remote Sensing Letters*, 13(8): 1074-1078 [DOI: 10.1109/LGRS.2016.2565705]
- Ma J Q, Shao W Y, Ye H, Wang L, Wang H, Zheng Y B and Xue X Y. 2018. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia*, 20(11): 3111-3122 [DOI: 10.1109/tmm.2018.2818020]
- Nie G T and Huang H. 2021. A survey of object detection in optical remote sensing images. *Acta Automatica Sinica*, 47(8): 1749-1768 (聂光涛, 黄华. 2021. 光学遥感图像目标检测算法综述. *自动化学报*, 47(8): 1749-1768) [DOI: 10.16383/j. aas. c200596]
- Qi M, Zhou Z Q, Miao L J, Zhang H W and Li L H. 2021. Dynamic anchor learning for arbitrary-oriented object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3): 2355-2363 [DOI: 10.1609/aaai.v35i3.16365]
- Redmon J, Divvala S, Girshick R and Farhadi A. 2016. You only look once: unified, real-time object detection//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, USA: IEEE: 779-788 [DOI: 10.1109/CVPR.2016.91]
- Ren S, He K, Girshick R and Sun J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6): 1137-1149 [DOI: 10.1109/TPAMI.2016.2577031]
- Rezatofighi H, Tsoi N, Gwak J Y, Sadeghian A, Reid I and Savarese S. 2019. Generalized intersection over union: a metric and a loss for bounding box regression. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 42(7): 658-666 [DOI: 10.1109/CVPR.2019.00078]

- Tian Z, Shen C H, Chen H and He T. 2019. FCOS: fully convolutional one-stage object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 9626-9635 [DOI: 10.1109/ICCV.2019.00972]
- Wang F, Jiang M Q, Chen Q Yang S, Li C, Zhang H G, Wang X G and Tang X O. 2017. Residual attention network for image classification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 41 (6): 3156-3164 [DOI: 10.1109/CVPR.2017.033]
- Xia G S, Bai X, Ding J, Zhu Z, Belongie S, Luo J B, Dateu M, Pelillo M and Zhang L P. 2018. DOTA: a large-scale dataset for object detection in aerial images//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 3974-3983 [DOI: 10.1109/CVPR.2018.00418]
- Xiao Z F, Qian L J, Shao W P, Tan X W and Wang K. 2020. Axis learning for orientated objects detection in aerial images. Remote Sensing, 12(6): #908 [DOI: 10.3390/rs12060908]
- Xie X X, Cheng G, Wang J B, Yao X W and Han J W. 2021. Oriented R-CNN for object detection//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 3500-3509 [DOI: 10.1109/iccv48922.2021.00350]
- Xu Y C, Fu M T, Wang Q M, Wang Y K, Chen K, Xia G S and Bai X. 2021. Gliding vertex on the horizontal bounding box for multi-oriented object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43 (4): 1452-1459 [DOI: 10.1109/TPAMI.2020.2974745]
- Yang X, Yan J C, Feng Z M and He T. 2021. R3Det: refined single-stage detector with feature refinement for rotating object//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtually: AAAI: 3163-3171 [DOI: 10.1609/aaai.v35i4.16426]
- Yang X, Yang J R, Yan J C, Zhang Y, Zhang T F, Guo Z, Sun X and Fu K. 2019. SCRDet: towards more robust detection for small, cluttered and rotated objects//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 8231-8240 [DOI: 10.1109/ICCV.2019.00832]
- Yi J R, Wu P X, Liu B, Huang Q Y, Qu H and Metaxas D. 2021. Oriented object detection in aerial images with box boundary-aware vectors//Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE: 2149-2158 [DOI: 10.1109/WACV48630.2021.00220]
- Yu Y, Yang X, Li J and Gao X. 2022. Object detection for aerial images with feature enhancement and soft label assignment. IEEE Transactions on Geoscience and Remote Sensing, 60: 1-16 [DOI: 10.1109/TGRS.2022.3177255]
- Zhang K, Zeng Q Q and Yu X U. 2021. ROSD: refined oriented staged detector for object detection in aerial image. IEEE Access, 9: 66560-66569 [DOI: 10.1109/ACCESS.2021.3076596].
- Zhong X, Gong O B, Huang W X, Li L and Xia H X. 2019. Squeeze-and-excitation wide residual networks in image classification//Proceedings of 2019 IEEE International Conference on Image Processing. Taipei, China: IEEE: 395-399 [DOI: 10.1109/ICIP.2019.8803000]

作者简介

余凌霄,男,硕士研究生,主要研究方向为遥感图像目标检测。E-mail:ylx2022@nuaa.edu.cn

郝洁,通信作者,女,硕士生导师,主要研究方向为目标检测和多模态数据融合。E-mail:haojie@nuaa.edu.cn

左量,男,硕士研究生,主要研究方向为小目标检测。E-mail:1147126030@qq.com