多元信息监督的遥感图像有向目标检测

王家宝，程塨，谢星星，姚艳清，韩军伟

西北工业大学 自动化学院, 西安 710129

摘 要：遥感图像有向目标检测是遥感图像解译中的一项基础任务，在许多领域有着广泛的应用。由于遥感图像目标尺度差异性大、方向任意且紧密排列，传统目标检测所使用的水平框无法准确的定位目标。因此，遥感图像有向目标检测成为目前遥感领域的研究热点。受益于深度学习的发展，遥感图像有向目标检测取得了突破性进展，但是大多数方法仅在检测头部加入角度预测参数，在训练过程中没有充分利用角度信息和语义信息。本文提出了一种多元信息监督的遥感图像有向目标检测方法。首先，在感兴趣区域提取阶段利用角度信息监督网络学习目标方向，从而使网络第一阶段生成更加贴近遥感图像目标的有向候选区域。其次，为了充分利用图像语义信息，本文在网络第二阶段增加语义分支，并使用图像语义标签进行监督学习。本文以Faster R-CNN OBB为基准，在DOTA数据集上验证所提方法的有效性。本文方法相比基准，平均精度（mAP）提升了2.8%，最终的检测精度（mAP）达到74.6%。

关键词：目标检测，有向目标检测，区域建议提取，多元信息，遥感图像

中图分类号：TP701/P2

引用格式：王家宝，程塨，谢星星，姚艳清，韩军伟.2023.多元信息监督的遥感图像有向目标检测.遥感学报，27(12)：2726-2735

Wang J B，Cheng G，Xie X X，Yao Y Q and Han J W. 2023. Multi-information supervision in optical remote sensing images. National Remote Sensing Bulletin，27(12)：2726-2735[DOI：10.11834/jrs.20211564]

1 引 言

高分辨率遥感图像解译有很高的实用价值，在森林资源管理（曹琼 等，2019）、农业调查（陈凯强 等，2020）、海洋检测（姚红革 等，2020）等诸多领域有着广泛的应用。在遥感图像解译的过程中，获取图像中高价值目标信息，精确定位目标是必不可少的工作。如图1所示，与自然图像不同，遥感图像目标的尺度差异性大、方向任意性且密集排布（Cheng等，2016），这使的通用目标检测算法无法精确定位遥感图像中的目标。因此，在遥感图像目标检测任务中，通常使用有向框来定位目标（Xia等，2018）。相较于通用目标检测，有向目标检测引入了目标角度的概念，要求模型在定位目标时输出目标所在位置、长宽和朝向。这种目标定位方式增加了目标定位输出的自由度，使任务变地更加困难。因此，高分辨率遥感图像有向目标检测具有很高的研究价值、并且富有挑战性。

近年来，随着深度学习算法飞速发展，各类计算机视觉任务在精度和速度方面都得到了极大的提升。最初，深度学习算法被应用于图像分类任务中，AlexNet（Krizhevsky等，2017）首次提出卷积层、激活层和池化层的结构，为之后的研究工作打下基础。VGG（Simonyan和Zisserman，2015）尝试通过加深网络层数来获得更好的分类效果。之后，ResNet（He等，2016）通过构建残差连接，解决了深层网络梯度消失问题，成功训练了上百层卷积神经网络。另一方面GoogLeNet系列（Ioffe和Szegedy，2015；Szegedy等，2017，2015，2016）尝试设计更加精巧的网络结构来提高精度，在每一层结构中将不同大小卷积核得到的特征图进行级联，用于获取语义更为丰富的特征。

在目标检测任务中，R-CNN（Girshick等，2014）首次使用了深度卷积神经网络。相比于传统方法，R-CNN获得了很高的检测精度。R-CNN首先通过区域生成算法获得感兴趣区域，之后通

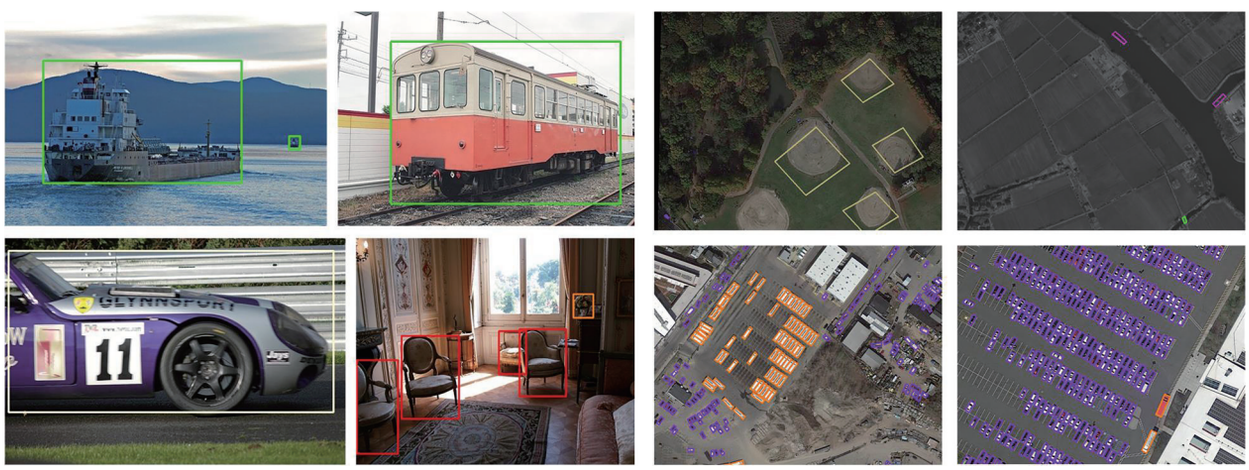
收稿日期：2021-08-21；**预印本**：2021-10-22

基金项目：国家自然科学基金(编号：61772425)；陕西省杰出青年科学基金(编号：2021JC-16)

第一作者简介：王家宝，研究方向为高分辨率遥感图像理解。E-mail：jbwang@mail.nwpu.edu.cn

通信作者简介：程塨，研究方向为高分辨率遥感图像理解。E-mail：gcheng@nwpu.edu.cn

过卷积神经网络对感兴趣区域进行分类和回归，得到最终的检测结果。在此之后，研究人员不断改进R-CNN并相继提出了Fast R-CNN（Girshick，2015）和Faster R-CNN（Ren等，2017），逐步将图像特征提取、感兴趣区域生成与区域分类回归融为一体，形成端到端目标检测框架。Faster R-CNN在目标检测上取得了显著的效果，为之后的许多模型提供了基准框架（Pang等，2019；Song等，2020；Wang等，2019；Wu等，2020），同时也被应用于下游任务（He等，2020）当中。这一类先获取感兴趣区域及其特征，而后对其进行分类和回归的方法被称为两阶段目标检测。



（a）自然图像 （b）遥感图像

（a）Natural images （b）Remotesensing images

图1 自然图像目标检测与遥感图像目标检测对比

Fig. 1 The comparison of natural image object detection and remote sensing image object detection

与两阶段目标检测不同，单阶段目标检测方法去除了感兴趣区域特征提取与感兴趣区域分类回归阶段，仅通过卷积操作得到最终的检测结果。YOLO（Redmon等，2016）是经典的单阶段目标检测方法。它将目标检测任务当作回归任务，在每一个位置上分别回归物体存在置信度、类别置信度与目标长宽，从而得到检测结果。相比于Faster R-CNN，YOLO模型在速度方面有着很大的提高。在此基础上，YOLO又进行了一系列改进（Redmon和Farhadi，2017，2018）。虽然单阶段目标检测方法取得了一定的效果，但是在检测精度方面，单阶段目标检测方法还是与双阶段目标检测方法有一定差距。Lin等（2020）指出单阶段目标检测面临着严重的正负样例不平衡问题，导致单阶段目标检测在精度方面远落后于双阶段目标检测方法。为了解决这个问题，Lin等（2020）提出了一种新的分类损失Focal Loss，为每一个样本分配不同的权重，减少易分负例样本在损失中的比重，从而减轻正负样本不平衡的影响。另外，Lin等（2020）构造了单阶段网络RetinaNet，并使用Focal Loss作为损失函数训练网络。RetinaNet在COCO数据集上达到了与Faster R-CNN相当的结果（Lin等，2014）。

在深度学习的推动下，高分辨率遥感图像解译也开始使用深度学习方法（陈凯强 等，2020；龚健雅和钟燕飞，2016；孙显 等，2020；姚红革等，2020；姚艳清 等，2021；周培诚 等，2021）。周培诚 等（2021）详细介绍了深度学习方法在各个领域的应用。在有向目标检测中，SCRDet（Yang等，2019）通过融合不同尺寸的特征来增强特征图的判别性，从而提升有向目标检测的精度，同时SCRDet还设计IoU-smooth L1 Loss来解决有向目标检测中的角度周期性与边界交换问题。R^3Det（Yang等，2021）在检测的过程中融合中点特征与角点特征，从而提高模型检测精度。Huang等（2022）构建了Nonlocal-aware金字塔注意力模块，提高模型的分类准确度。S^2ANet（Han等，2022）中使用AlignConv解决了有向检测中的特征不对齐问题，使得单阶段有向目标检测精度得到了很大的提升。在有向感兴趣区域生成方法中，较早的算法RRPN（Ma等，2018）直接在特征图的每个位置上放置不同尺寸与角度的锚框来生成有向候

选框。这种方法虽然取得了一定的成绩，但是过多的锚框增加了检测器的计算量和内存占用。在之后的工作中，RoI Transformer（Ding等，2019）通过训练小型全连接分支RoI Learner将水平的区域建议转化为有向区域建议。RoI Learner输出的有向区域建议能很好的匹配遥感图像中的有向目标，减少背景冗余信息，从而极大的提升检测效果。在LO-Det（Huang等，2022）中，作者设计轻量化遥感图像目标检测网络，使深度学习算法离实际应用更进一步。

虽然遥感图像有向目标检测已经取得了显著的效果，但在数据的利用方面还存在改进的空间。现阶段有向目标检测方法大都是以水平框作为基准，第一个阶段依然生成的是水平的候选区域，仅在最后一个阶段训练生成有向目标检测框。本文在第一阶段使用目标的角度信息监督候选区域的生成，从而使网络在第一阶段就能够生成有向的候选区域。这样可以获得与目标匹配更好的候选区域。同时如图2所示，图像语义信息可以清晰且直观的表示目标位置与大小，对网络特征的学习有一定的指导作用。现阶段有向目标检测方法中，很少利用到图像语义信息监督检测网络。本文将图像语义信息利用起来，使用图像语义标签监督不同尺度图像特征学习。同时本文将输出作为目标位置的先验信息，以此过滤假正例目标框。本文在DOTA（Xia等，2018）上验证所提算法的有效性。相较于基准Faster R-CNN OBB，第一阶段角度回归可以增长2.2% mAP，语义信息监督可以使检测精度增长0.8% mAP，最终本文方法的检测精度（mAP）可达74.64%。

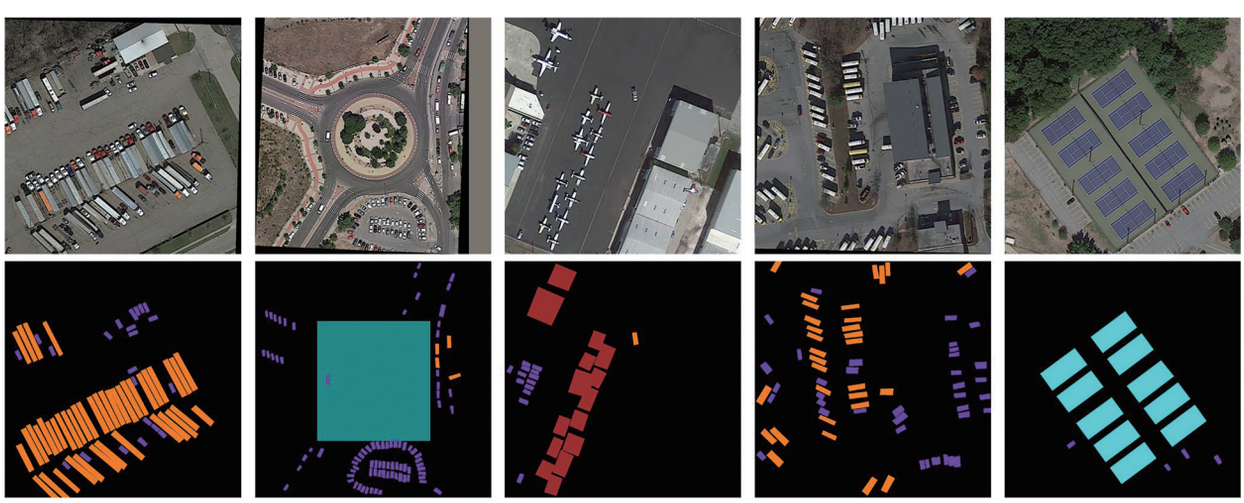


图2 有向目标语义标签

Fig. 2 The semantic labels of oriented objects

2 本文方法

本文提出了一种多元信息监督的有向目标检测方法，充分利用角度信息与语义信息。图3为本文方法的整体流程图。如图3所示，待检测的遥感图像先通过主干网络与特征金字塔网络生成不同尺度的特征图。在候选区域生成阶段，不同于基准方法Faster R-CNN OBB，本文直接通过角度信息来监督区域建议网络RPN（Region Proposal Network）生成有向的区域建议。与水平的区域建议相比，有向的区域建议更加贴近遥感图像中的目标，这不仅减少了冗余背景信息的干扰，还使得网络需要回归的目标值更小，提升有向目标检测效果。在第二阶段，本文在每一层特征上都增加了语义分支，通过图像的语义标签监督语义分支学习，从而让网络直接学习局部区域的语义特征，提高网络特征的平移可变性。同时，本文将语义分支的输出作为目标框位置的先验信息，从而过滤掉一些处于背景区域的假正例，提高有向区域建议的质量。在这一节中，本文分别详细介绍有向区域建议生成和语义信息分支。

2.1 带角度的区域建议网络ARPN（Angle-based Region Proposal Network）

原始的RPN在特征图的每一个位置放置形状大小不同的锚框作为区域建议回归的基准。默认

情况下，同一位置上会放置3种不同长宽比（1∶1，2∶1，1∶2）的锚框，并且随着每一个特征图尺寸的变化，锚框的面积（32^2，64^2，128^2，256^2，512^2）也发生变化。在推理过程中，特征图经过3×3的卷积后，再经过两个并行的1×1卷积得到分类置信度与回归偏差。分类置信度的通道数C与每一个位置上放置的锚点框个数相同，它表示相应锚框属于前景的置信度。回归偏差的通道数为$4\times C$，分别代表了相应锚点框的中心点坐标偏差t_x，t_y和长宽缩放偏差t_w，t_h。通过网络输出偏差值的大小，可以将形状规则的锚框回归为任性形状。设锚框的中心点坐标与长宽为x_a，y_a，w_a，h_a，回归后的区域建议的中心点坐标与长宽为x_p，y_p，w_p，h_p。它们之间的计算方式如式（1）所示：

$$\begin{aligned} x_p &= w_a \times t_x + x_a \quad w_p = w_a \times \exp(t_w) \\ y_p &= h_a \times t_y + y_a \quad h_p = h_a \times \exp(t_h) \end{aligned} \tag{1}$$

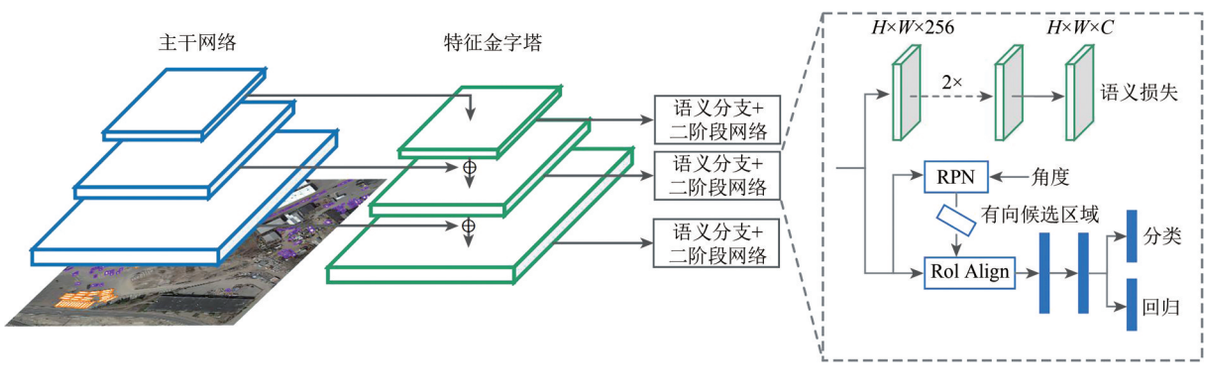


图3 网络整体流程图

Fig. 3 The main pipeline of the network

在原始的有向目标检测中，RPN通过目标的水平外接矩形进行监督，输出水平区域建议。在第二阶段，先提取区域建议特征，并以水平的区域建议为基准回归有向的目标框。这种有向目标回归方法存在以下几点问题：（1）水平区域建议无法和有向的目标贴合，造成区域建议特征提取过程中大量的无关噪声被引入，使得分类效果下降。（2）这种有向目标检测方法没有充分地利用二阶段的结构优势，只进行了一次角度回归。

为了解决上述问题，本文将角度信息引入RPN中，并提出带角度的区域建议网络ARPN（Angle-based Region Proposal Network）。在ARPN中，网络的回归分支通道数为$5\times C$，分别代表了C个锚框的中心点坐标偏差t_x，t_y、长宽缩放偏差t_w，t_h和角度t_t。在训练过程中，本文根据锚框与有向目标水平外接矩形之间的交并比IoU（Intersection of Union）将所有的锚框分为正例与负例。一个锚框与目标之间的IoU大于0.7，或者是这个目标IoU最大的锚框，这个锚框会被当作正例。如果一个锚框与所有的目标框之间的IoU都小于0.3，则这个锚框被认为是负例。分配完正负例之后，本文随机选取256个正例与256个负例进行训练，损失函数如式（2）所示：

$$\begin{aligned} L(\{p_i\}, \{t_i\}) = &\frac{1}{N_{\text{cls}}}\sum_i L_{\text{cls}}(p_i, p_i^*) + \\ &\frac{1}{N_{\text{reg}}}\sum_i p_i^* L_{\text{reg}}(t_i, t_i^*) \end{aligned} \tag{2}$$

式中，N_{cls}与N_{reg}分别表示进行分类与回归的实例数量，L_{cls}与L_{reg}表示分类和回归的损失函数，p_i表示网络输出的锚框的置信度。p_i^*代表锚框的正负例分类结果，当锚框为正例时，p_i^*为1，否则为0。本文使用交叉熵损失作为分类损失。网络的回归偏差输出为$t_i=(t_x, t_y, t_w, t_h, t_t)$。在训练过程中，本文只针对正例训练回归偏差。回归的目标值t_i^*通过相应的锚框（x_a，y_a，w_a，h_a）与其相匹配的有向目标（x^*，y^*，w^*，h^*，t^*）通过式（3）求得，回归使用L_1范数作为损失函数。

$$\begin{aligned} t_x^* &= (x^* - x_a)/w_a \quad t_w^* = \log(w^*/w_a) \\ t_y^* &= (y^* - y_a)/h_a \quad t_h^* = \log(h^*/h_a) \quad t_t^* = t^* \end{aligned} \tag{3}$$

通过角度信息监督，ARPN能够直接生成有向区域建议。在区域建议生成阶段，本文根据回归分支的输出，将水平锚框转化为有向的区域建议。同时，本文按照分类置信度，将所有的区域建议排序，并选择前2000个区域建议作为ARPN的输出。这些区域建议被输入到第二阶段，进行进一步的分类与回归。在第二阶段中，本文通过有向

区域建议与有向目标之间的IoU将区域建议与目标匹配。如果区域建议与目标之间的IoU大于0.5，将这个区域建议划分为正例，且类别与相匹配的目标一致。之后，随机选择512个样本进行训练，正例和负例的比例为1∶3。因为第一阶段已经获得初始的角度，第二阶段的回归将在第一阶段的基础上预测角度差值，从而进一步修正检测框。设有一组相互匹配的有向候选区域（x_p，y_p，w_p，h_p，t_p）与目标框（x^*，y^*，w^*，h^*，t^*），则从有向区域建议回归到目标框的偏差可由式（4）求得：

$$t_x^* = (x^* - x_p)/w_p \quad t_w^* = \log(w^*/w_p)$$
$$t_y^* = (y^* - y_p)/h_p \quad t_h^* = \log(h^*/h_p) \quad t_t^* = t^* - t_p \tag{4}$$

通过使用角度信息监督，ARPN能够在第一阶段生成有向的候选区域。这不仅减少了背景噪声对感兴趣区域的影响，同时还充分利用第二阶段的特性，对有向目标框进行微调。通过实验证明，ARPN能够有效提升有向目标检测精度。

2.2 图像语义信息分支（Semantic Branch）

除了目标的角度信息以外，图像的语义信息也对有向目标检测任务有所帮助。图像语义标签可以直观的反应出目标所在位置与相应类别，能够监督检测网络学习平移可变特征，从而提高检测效果。同时，图像语义预测可以作为目标位置的先验信息，将一些出现在背景语义上的假正例过滤。本文在第二阶段增加了图像语义预测分支，并将图像语义信息作为监督。如图3所示，新增的语义分支由2个3×3和1个1×1的卷积组成，最终语义信息分支的通道数C为数据集的类别数。本文将特征金字塔输出的5个不同尺度的图像特征图（P_2，P_3，P_4，P_5，P_6）分别输入语义分支，生成不同尺度的预测图（S_2，S_3，S_4，S_5，S_6）。在训练过程中，本文根据目标尺寸的大小，将目标分配到不同尺度的语义预测特征图上进行训练。设有向目标的参数为（x，y，w，h，t），则目标被分到的预测特征图的编号可由式（5）计算得到：

$$k = \left\lfloor k_0 + \log_2\left(\sqrt{wh}/224\right)\right\rfloor \tag{5}$$

式中，k_0表示面积为224^2的目标被分配到的预测特征图的编号。这里k_0为4。将所有目标分配到不同的特征图后，本文通过判断预测点是否位于目标内部来确定相应的语义标签。如果预测点位于相应尺度下的目标内部，则这个预测点的语义标签为这个目标的类别，否则为背景类。如果预测点在多个目标内部，则语义标签与面积较小的目标相同。在训练过程中，使用Focal Loss（Lin等，2020）计算预测值与语义标签的损失。

通过图像语义信息监督，可以增强网络特征的平移可变性，同时语义分支的预测值也可以作为目标所在位置的先验信息。如图4所示，语义分支的预测值可以清晰的反映出前景区域与背景区域，从而可以用于过滤背景区域中出现的假正例。在检测网络预测过程中，将所有类别语义预测的最大值作为这一区域是否有目标的先验。得到目标位置先验后，本文将位置先验与ARPN的类别置信度相乘，降低背景区域目标的置信度，从而过滤掉一些在背景区域中产生的虚警。

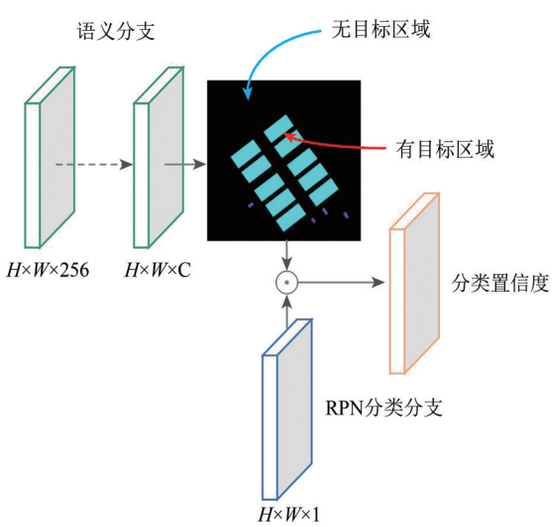


图4 通过语义预测值来过滤虚警

Fig. 4 Using semantic prediction to filter out false positives

3 实验与分析

3.1 实验条件

本文在DOTA数据集上验证所提方法的有效性。DOTA数据集是遥感图像有向目标检测的常用数据集，共有2806张图像，188282个实例，15个类别。类别包含：飞机（PL）、船（SH）、储罐（ST）、棒球场（BD）、网球场（TC）、游泳池（SP）、田径场（GTF）、港口（HA）、桥梁（BR）、小型车辆（SV）、大型车辆（LV）、直升机（HC）、环岛（RA）、足球场（SBF）和篮球场（BC）。图幅大小为800—4000。因为DOTA的图像面积较大，在训练之前需要进行裁图预处理。训练过程

中，将图像以大小为1024×1024、步长为824的滑窗进行裁图。测试过程中以大小为1024×1024、步长为824的滑窗进行裁图。所有的实验在单张Nvidia 1080Ti上进行。训练时仅使用随机翻转作为数据增广策略。训练的batch size为2，权重衰减为0.00001。总共训练12个epoch，初始学习率为0.005，并在epoch 8与epoch 11之后学习率变为原来的1/10。

3.2 消融实验

为了验证的有效性，本文使用ResNet50为主干网络，在DOTA数据集上进行消融实验。结果如表1所示，原始的Faster R-CNN OBB在DOTA数据集的检测精度为71.8% mAP的效果。本文方法以Faster R-CNN OBB为基准。相比于基准Faster RCNN OBB，ARPN可以使检测准确率上升2.2% mAP，达到74.0% mAP。ARPN的区域建议可视化结果如图6所示，从图6中可以看出，ARPN能够生成紧密包围有向目标的区域建议，这表明使用角度信息监督有向区域建议生成对有向目标检测很有必要。有向区域建议不仅更加贴近遥感图像目标，使得背景噪声减少，还能充分利用第二阶段结构的特性，进一步对目标角度微调。

仅增加语义分支可以使检测模型准确率提高0.8% mAP，达到72.6% mAP。这验证了使用语义信息监督的有效性。它可以提高网络特征的平移可变性，同时还能够作为目标位置的先验，用于过滤虚警。相较于原始的Faster RCNN OBB，本文方法将有向目标检测精度提升了2.8% mAP，最终达到了74.6% mAP，从而证明了方法的有效性。

表1 本文算法在DOTA数据集的消融实验

Table 1 Ablation studies on the DOTA dataset

	ARPN	Semantic Branch	mAP/%
基线			71.8
本文方法	√		74.0
		√	72.6
	√	√	74.6

3.3 对比实验

将本文方法与当前其他遥感图像有向目标检测方法进行了对比。在DOTA数据集上，对比了Faster RCNN OBB（Ren等，2017）、RetinaNet（Lin等，2020）、DAL（Ming等，2020）、RRPN（Ma等，2018）、SCRDet（Yang等，2019）、RoI Transformer（Ding等，2019）等遥感图像有向目标检测方法。其中RetinaNet、DAL为单阶段方法，Faster RCNN OBB、RRPN、SCRDet和RoI Transformer为两阶段方法。各个方法的检测精度在表2中给出，其中黑体代表每一类别的最高检测效果。

表2 不同算法在DOTA数据集上检测精度的对比

Table 2 Comparison of detection accuracy of different methods on the DOTA dataset

方法	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
Faster RCNN OBB	R-50-FPN	89.40	82.59	46.43	70.57	78.17	67.54	83.71	90.90	84.78	84.64	60.78	61.52	54.25	68.84	53.98	71.87
RetinaNet OBB	R-50-FPN	88.67	77.62	41.81	58.17	74.58	71.64	79.11	90.29	82.18	74.32	54.75	60.60	62.57	69.67	60.64	68.43
DAL	R-50-FPN	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
RRPN	R-101	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	R-101-FPN	**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
RoI Transformer	R-101-FPN	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	47.67	69.56
本文	R-50-FPN	89.59	**82.85**	**52.37**	71.68	**79.08**	**77.19**	**87.23**	**90.88**	85.31	84.88	62.87	62.17	66.05	70.41	57.05	**74.64**

注：加粗数字为该类别目标检测效果最好的结果。

结果显示，本文方法在棒球场（BD）、桥梁（BR）、小型车辆（SV）、大型车辆（LV）、船（SH）、网球场（TC）等类别目标上都取得了最好的检测效果。这些类别中包含大量小目标与紧密排列目标，如小车和船。如何提高小目标与紧密排列目标是遥感图像有向目标检测的难点问题。从检测精度来看，本文方法能够在一定程度上解决这些问题。同时，本文方法在DOTA数据集上的平均精度也是7种有向目标检测器中最高的。这进一步验证了本文方法的有效性。同时，我们将DOTA数据集上的部分检测结果进行可视化，可视化结果如图5所示。从可视化结果上来看，本文方法能够很好的处理紧密排列与尺度较小的目标，如紧密排列的车辆与船只。我们分析这是因为ARPN提取的有向区域建议能够更精准地与紧密排列的目标进行匹配，减少背景信息带来的干扰，从而提高这些目标的检测效果。

图5 DOTA数据集检测结果

Fig. 5 The detection results on DOTA dataset

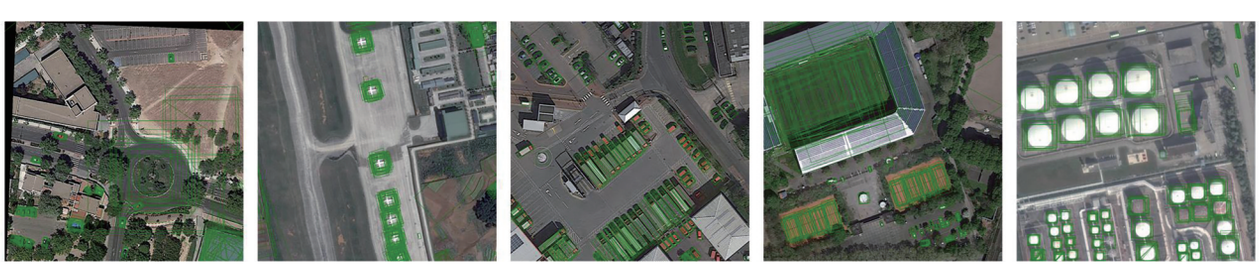


图6 ARPN区域建议可视化结果

Fig. 6 The visualization of proposals of ARPN

4 结 论

本文提出了一种多元信息监督的遥感图像有向目标检测方法。为了充分利用目标角度信息，本文提出了带角度的区域建议网络（ARPN），通过角度信息监督网络生成有向区域建议。同时，本文在检测网络中加入了语义分支（Semantic Branch）。图像语义信息的引入有利于检测网络学习平移可变的特征，从而提升检测模型的效果。本文在DOTA数据集上验证了方法的有效性。相比于基准，引入ARPN检测精度可以提升2.2% mAP，使得检测结果达到74.0% mAP；增加语义分支后检测网络精度能增加0.8% mAP，检测结果可达72.6% mAP。同时使用两种方法，最终的检测精度能达到了74.64% mAP。与其他遥感图像有向目标检测方法对比，本文方法也有一定优势，在棒球场（BD）、田径场（GTF）、小型车辆（SV）、大型车辆（LV）、船（SH）、网球场（TC）等类别上都

达到了最好的效果。从可视化结果来看，本文方法能够很好的处理密集排列的有向目标，并且对小目标也有很好的检测效果。

通过实验验证，在第一阶段输出有向的区域建议可以极大的提高遥感图像中有向目标的检测精度。但有向的区域建议在提高模型检测精度的同时，也会使模型计算复杂度上升。下一阶段将会重点研究如何降低有向区域建议网络的模型复杂度，在确保检测精度的条件下使模型推理速度增加，从而提高模型的实际应用价值。

参考文献（References）

Cao Q, Ma A L, Zhong Y F, Zhao J, Zhao B and Zhang L P. 2019. Urban classification by multi-feature fusion of hyperspectral image and LiDAR data. Journal of Remote Sensing, 23(5): 892-903 (曹琼, 马爱龙, 钟燕飞, 赵济, 赵贝, 张良培. 2019. 高光谱-LiDAR多级融合城区地表覆盖分类. 遥感学报, 23(5): 892-903) [DOI: 10.11834/jrs.20197512]

Chen K Q, Gao X, Yan M L, Zhang Y and Sun X. 2020. Building extraction in pixel level from aerial imagery with a deep encoder-decoder network. Journal of Remote Sensing (Chinese), 24(9): 1134-1142 (陈凯强, 高鑫, 闫梦龙, 张跃, 孙显. 2020. 基于编解码网络的航空影像像素级建筑物提取. 遥感学报, 24(9): 1134-1142) [DOI: 10.11834/jrs.20209056]

Cheng G, Zhou P C and Han J W. 2016. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 54(12): 7405-7415 [DOI: 10.1109/tgrs.2016.2601622]

Ding J, Xue N, Long Y, Xia G S and Lu Q K. 2019. Learning RoI transformer for oriented object detection in aerial images//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE: 2844-2853 [DOI: 10.1109/cvpr.2019.00296]

Girshick R. 2015. Fast R-CNN//2015 IEEE International Conference on Computer Vision (ICCV). Santiago: IEEE: 1440-1448 [DOI: 10.1109/iccv.2015.169]

Girshick R, Donahue J, Darrell T and Malik J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation//2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE: 580-587 [DOI: 10.1109/cvpr.2014.81]

Gong J Y and Zhong Y F. 2016. Survey of intelligent optical remote sensing image processing. Journal of Remote Sensing, 20(5): 733-747 (龚健雅, 钟燕飞. 2016. 光学遥感影像智能化处理研究进展. 遥感学报, 20(5): 733-747) [DOI: 10.11834/jrs.20166205]

Han J M, Ding J, Li J and Xia G S. 2022. Align deep features for oriented object detection. IEEE Transactions on Geoscience and Remote Sensing, 60: 5602511 [DOI: 10.1109/tgrs.2021.3062048]

He K M, Gkioxari G, Dollar P and Girshick R. 2020. Mask R-CNN. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2): 386-397 [DOI: 10.1109/tpami.2018.2844175]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE: 770-778 [DOI: 10.1109/cvpr.2016.90]

Huang Z C, Li W, Xia X G, Wang H, Jie F R and Tao R. 2022a. LO-Det: lightweight oriented object detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 60: 5603515 [DOI: 10.1109/tgrs.2021.3067470]

Huang Z C, Li W, Xia X G, Wu X, Cai Z Q and Tao R. 2022b. A novel nonlocal-aware pyramid and multiscale multitask refinement detector for object detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 60: 5601920 [DOI: 10.1109/tgrs.2021.3059450]

Ioffe S and Szegedy C. 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift//Proceedings of the 32nd International Conference on Machine Learning. Lille: JMLR.org: 448-456

Krizhevsky A, Sutskever I and Hinton G E. 2017. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6): 84-90 [DOI: 10.1145/3065386]

Lin T Y, Goyal P, Girshick R, He K M and Dollar P. 2020. Focal loss for dense object detection. IEEE Trans actionson Pattern Analysis and Machine Intelligence, 42(2): 318-327 [DOI: 10.1109/TPAMI.2018.2858826]

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L. 2014. Microsoft COCO: common objects in context//13th European Conference on Computer Vision. Zurich: Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48]

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]

Ming Q, Zhou Z Q, Miao L J, Zhang H W and Li L H. 2020. Dynamic anchor learning for arbitrary-oriented object detection. arXiv preprint arXiv: 2012.04150

Pang J M, Chen K, Shi J P, Feng H J, Ouyang W L and Lin D H. 2019. Libra R-CNN: towards balanced learning for object detection//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE: 821-830 [DOI: 10.1109/cvpr.2019.00091]

Redmon J, Divvala S, Girshick R and Farhadi A. 2016. You only look once: unified, real-time object detection//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE: 779-788 [DOI: 10.1109/cvpr.2016.91]

Redmon J and Farhadi A. 2017. YOLO9000: better, faster, stronger//2017 IEEE Conference on Computer Vision and Pattern Recognition

(CVPR). Honolulu: IEEE: 6517-6525 [DOI: 10.1109/cvpr.2017.690]

Redmon J and Farhadi A. 2018. YOLOv3: an incremental improvement. arXiv preprint arXiv: 1804.02767

Ren S Q, He K M, Girshick R and Sun J. 2017. 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]

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556

Song G L, Liu Y and Wang X G. 2020. Revisiting the sibling head in object detector//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE: 11560-11569 [DOI: 10.1109/cvpr42600.2020.01158]

Sun X, Liang W, Diao W H, Cao Z Y, Feng Y C, Wang B and Fu K. 2020. Progress and challenges of remote sensing edge intelligence technology. Journal of Image and Graphics, 25(9): 1719-1738 (孙显, 梁伟, 刁文辉, 曹志颖, 冯瑛超, 王冰, 付琨. 2020. 遥感边缘智能技术研究进展及挑战. 中国图象图形学报, 25(9): 1719-1738) [DOI: 10.11834/jig.200288]

Szegedy C, Ioffe S, Vanhoucke V and Alemi A A. 2017. Inception-v4, inception-ResNet and the impact of residual connections on learning//Proceedings of the Thirty-first AAAI Conference on Artificial Intelligence. San Francisco: AAAI Press: 4278-4284

Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A. 2015. Going deeper with convolutions//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE: 1-9 [DOI: 10.1109/cvpr.2015.7298594]

Szegedy C, Vanhoucke V, Ioffe S, Shlens J and Wojna Z. 2016. Rethinking the inception architecture for computer vision//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE: 2818-2826 [DOI: 10.1109/cvpr.2016.308]

Wang J Q, Chen K, Yang S, Loy C C and Lin D H. 2019. Region proposal by guided anchoring//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE: 2960-2969 [DOI: 10.1109/cvpr.2019.00308]

Wu Y, Chen Y P, Yuan L, Liu Z C, Wang L J, Li H Z and Fu Y. 2020. Rethinking classification and localization for object detection//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE: 10183-10192 [DOI: 10.1109/cvpr42600.2020.01020]

Xia G S, Bai X, Ding J, Zhu Z, Belongie S, Luo J B, Datcu M, Pelillo M and Zhang L P. 2018. DOTA: a large-scale dataset for object detection in aerial images//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE: 3974-3983 [DOI: 10.1109/cvpr.2018.00418]

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 AAAI Conference on Artificial Intelligence, 35(4): 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//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE: 8231-8240 [DOI: 10.1109/iccv.2019.00832]

Yao H G, Wang C, Yu J, Bai X J and Li W. 2020. Recognition of small-target ships in complex satellite images. Journal of Remote Sensing (Chinese), 24(2): 116-125 (姚红革, 王诚, 喻钧, 白小军, 李蔚. 2020. 复杂卫星图像中的小目标船舶识别. 遥感学报, 24(2): 116-125) [DOI: 10.11834/jrs.20208238]

Yao Y Q, Cheng G, Xie X X and Han J W. 2021. Optical remote sensing image object detection based on multi-resolution feature fusion. National Remote Sensing Bulletin, 25(5): 1124-1137 (姚艳清, 程塨, 谢星星, 韩军伟. 2021. 多分辨率特征融合的光学遥感图像目标检测. 遥感学报, 25(5): 1124-1137) [DOI: 10.11834/jrs.20210505]

Zhou P C, Cheng G, Yao X W and Han J W. 2021. Machine learning paradigms in high-resolution remote sensing image interpretation. National Remote Sensing Bulletin, 25(1): 182-197 (周培诚, 程塨, 姚西文, 韩军伟. 2021. 高分辨率遥感影像解译中的机器学习范式. 遥感学报, 25(1): 182-197) [DOI: 10.11834/jrs.20210164]

Multi-information supervision in optical remote sensing images

WANG Jiabao, CHENG Gong, XIE Xingxing, YAO Yanqing, HAN Junwei

School of Automation, Northwestern Polytechnical University, Xi'an 710129, China

Abstract: Oriented object detection is a basic task in the interpretation of high-resolution remote sensing images. Compared with general detectors, oriented detectors can locate instances with oriented bounding boxes, which are consistent with arbitrary-oriented ground truths in remote sensing images. Currently, oriented object detection has greatly progressed with the development of the convolutional neural network. However, this task is still challenging because of the extreme variation in object scales and arbitrary orientations. Most oriented

detectors are evolved from horizontal detectors. They first generate horizontal proposals using the Region Proposal Network (RPN). Then, they classify these proposals into different categories and transform them into oriented bounding boxes. Despite their success, these detectors exploit only the annotations at the end of the network and do not fully utilize the angle and semantic information.

This work proposes an Angle-based Region Proposal Network (ARPN), which learns the angle of objects and generates oriented proposals. The structure of ARPN is the same as that of RPN. However, for each proposal, instead of outputting four parameters for regression, ARPN generates five parameters, which are the center (x, y), shape (w, h), and angle (t). In the training, we first assign anchors with ground truths by the Intersection of Unions. Then, we directly supervise the ARPN with the shape and angle information of ground truths. We also propose a semantic branch to output image semantic results for utilizing the advantage of the semantic information. The semantic branch consists of two convolutional layers and is parallel with the detection head. We first assign objects to different scale levels according to their areas. Then, we create semantic labels in each scale and use them to supervise the semantic branch. With the semantic information supervision, the model will learn translation-variant features and improve accuracy. Moreover, the outputs of the semantic branch indicate the objectness in each place, which can filter out false positives of final predictions.

We conduct comprehensive experiments on the DOTA dataset to validate the effectiveness of the proposed methods. In the data preparation, we first crop original images into 1024×1024 patches with the stride of 824. Compared with the baseline, the ARPN achieves a 2.2% increase in mAP, while the semantic branch contributes an additional 0.8% improvement in mAP. Finally, we combine both methods and achieve a 74.64% mAP, which is competitive with those obtained by other oriented object detectors. We visualize some results on the DOTA dataset. The results show that our method is highly effective for small objects and densely packed objects.

We proposed ARPN and the semantic branch to utilize the multi-information in remote sensing images. The ARPN can directly generate oriented proposals, which can lead to better recall of oriented objects. The semantic branch increases the translation-variant property of the features. Experiments demonstrate the effectiveness of our method, which achieves a 74.64% mAP on the DOTA dataset. In the future works, we will focus on the model efficiency and the inference speed.

Key words：object detection, oriented object detection, region proposal generation, multi-information, remote sensing images

Supported by National Natural Science Foundation of China (No. 61772425); Shaanxi Science Foundation for Distinguished Young Scholars (No. 2021JC-16)