

基于角度敏感的空间注意力机制的轻量型旋转目标检测器

尹开石¹, 杨萌², 顾曦³, 王志成³

(1. 东华大学计算机科学与技术学院, 上海 201620;

2. 中国舰船研究设计中心, 湖北 武汉 430064;

3. 同济大学电子与信息工程学院, 上海 201804)

摘要: 随着深度学习的快速发展, 近年来越来越多的基于锚框的目标检测算法被应用于遥感图像上, 然而提高算法精度的代价是牺牲了检测速度。因此, 选择无锚框的目标检测网络框架, 针对遥感场景的特点, 提出了一种旋转框的遥感检测算法。根据旋转框与其外接矩形框的空间位置关系, 提出了一种简单有效的旋转框表示方式。此外, 设计了一种用于辅助检测旋转目标的角度敏感的空间注意力机制, 通过引入角度信息提升模型对旋转目标的检测能力。提出的算法在公开遥感数据集 DOTA 上进行了实验, 旋转框的目标检测网络的平均精度均值达到了 68.5%, 检测速度达到了每秒 17.4 帧图像。

关键词: 旋转检测; 遥感图像; 无锚框; 注意力机制

中图分类号: TP181

文献标识码: A

doi: 10.11959/j.issn.2096-6652.202133

Light weight rotating object detector based on angle sensitive spatial attention mechanism

YIN Kaishi¹, YANG Meng², GU Xi³, WANG Zhicheng³

1. School of Computer Science and Technology, Donghua University, Shanghai 201620, China

2. China Ship Development and Design Center, Wuhan 430064, China

3. College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China

Abstract: With the rapid development of deep learning, more and more target detection algorithms based on anchor frame are applied to remote sensing images in recent years. However, the cost of improving the accuracy of the algorithm is to sacrifice the detection speed. Therefore, the target detection network framework of anchor free was chosen, and a remote sensing detection algorithm of rotating frame was proposed according to the characteristics of remote sensing scene. A simple and effective representation of rotating frame was proposed according to the spatial position relationship between rotating frame and its external rectangular frame. In addition, an angle sensitive attention mechanism was designed to assist the detection of rotating targets. By introducing angle information, the detection ability of the model for rotating targets was improved. The proposed algorithm was tested on the open remote sensing dataset DOTA. The mean average precision of target detection network of rotating frame is 68.5% and the detection speed is 17.4 frames per second.

Key words: rotation detection, remote sensing image, anchor free, attention mechanism

1 引言

在遥感图像的处理技术中, 目标检测是一个重

点研究对象。将目标检测与遥感图像结合, 可以在地球表面资源评估、自然灾害预警、军事领域等方面提供巨大的帮助。因此, 研究目标检测在遥感图

收稿日期: 2021-05-17; 修回日期: 2021-08-06

通信作者: 尹开石, 1291020864@qq.com

基金项目: 国防基础科研计划资助项目 (No.JCKY2020206B037)

Foundation Item: The National Defense Basic Research Program of China (No.JCKY2020206B037)

像上的应用是非常有必要的。

在目标检测中,不同于常规图像,遥感图像的检测难度更大。在大尺寸遥感图像中的目标识别(例如船舶、飞机和车辆检测)是一项具有挑战性的任务,因为目标的尺寸小且数量大,并且周围环境复杂,这可能导致检测模型将无关的地面物体误认为目标。而自然图像中的物体相对较大,并且与遥感图像相比,局部区域中的环境并不复杂,因此更易于识别目标。这是检测遥感目标和自然目标的主要区别之一。

目标检测中使用的深度学习方法可以分为两大类:无监督方法和有监督方法。无监督方法从输入数据中学习特征,无须了解相关标签或其他监管信息,而有监督方法则使用输入数据以及附加到输入的监管信息来有区别地学习特征。但是,这两种深度学习方法都对对象图像中学习特征,并且学习过程可以统一到同一框架中。

近年来许多基于深度学习的目标检测方法被陆续提出,显著提高了目标检测的性能。通常可以基于有无锚框(anchor)将用于目标检测的现有深度学习方法分为两个流派,分别是基于 anchor 的区域建议方法(以下简称区域建议方法)和基于无锚框(anchor free)的回归方法。

在过去的几年中,区域建议方法在自然场景图像中取得了巨大的成功。这种方法将目标检测的框架分为两个阶段。第一阶段着重于生成可能包含目标的一系列候选区域。第二阶段将第一阶段获得的候选区域分类为目标前景或背景,并进一步微调边界框的坐标。其中 R-CNN^[1]是基于区域建议网络(region proposal network, RPN)的目标检测方法的典型算法之一。R-CNN 先通过选择性搜索获取可能存在目标的候选区域,然后在候选区域利用卷积神经网络(convolutional neural network, CNN)提取特征向量,再将特征向量传入支持向量机进行分类。尽管它在图像特征提取过程中引入了神经网络,避免了人工提取特征的弊端,但对每个候选区域进行卷积运算难以满足速度要求。Fast R-CNN 算法^[2]、Faster R-CNN 算法^[3]、Mask R-CNN 算法^[4]等对 R-CNN 做了不同的改进,使得区域建议方法获得了卓越的性能。

尽管如此,目前的移动设备的存储空间和计算能力非常有限,区域建议方法为了获得较高的召回率,需要在图像中预先设定数以万计的锚框,这对

小目标检测效果提升明显,但需要先验知识设计尺度、长宽比、交并比等超参数,且会造成正负样本严重不平衡,导致它的检测速度难以进一步提高。

因此,研究人员开始尝试优化基于 anchor free 的回归方法,使用单阶段目标检测模型进行预测,从而简化了目标检测的两个步骤,优点是具有更大更灵活的解空间,并大大降低了目标检测的计算复杂度。与基于 anchor 的区域建议方法相比,基于 anchor free 的回归方法更加简单高效,因为无须生成候选区域建议和随后的特征重采样阶段。2015 年年初, Huang L C 等人提出^[5]的 DenseBox 是基于 anchor free 的回归方法的早期作品。其使用全卷积网络,实现了端到端的训练和识别,在人脸检测任务中有不俗的效果。但由于论文正式发表的时间较晚,这种基于 anchor free 的回归方法比区域建议方法更晚进入人们的视野,因此区域建议方法在目标检测领域占据主导地位。

Redmon J 等人^[6]提出了 YOLO (you only look once),将目标检测看作一个单一的回归问题,直接从图像生成目标矩形框坐标和分类概率。由于没有区域提案生成步骤,YOLO 使用一组候选区域直接预测检测结果。与基于区域的方法使用局部区域的特征来预测不同,YOLO 使用的是全局图像的特征。由于完全抛弃了区域提案生成步骤,YOLO 速度很快。但与 Fast R-CNN 相比,YOLO 的定位误差更大,这是因为边界框位置、比例的划分比较粗略。

Law H 等人^[7]质疑了 anchor 在目标检测框架中的主导作用,提出了基于 anchor free 的目标检测方法 CornerNet,指出了 anchor 存在的缺陷,如造成正负样本之间的巨大不平衡、减慢训练速度和引入额外的超参数,并借鉴了在多人位姿估计中关联嵌入(associative embedding)的思想,将边界框目标检测视为检测成对的左上角和右下角关键点。在 CornerNet 中,基础网络由两个堆叠的沙漏(hourglass)网络组成,并使用简单的角池化方法来更好地定位角点。在 COCO 数据集^[8]上,CornerNet 实现了 42.1%的平均精度,超过了之前所有的单阶段检测器,然而在 Titan X GPU 上明显慢于 YOLO。CornerNet 很难决定哪些关键点对应应该被分到同一个目标中,这会导致检测错误。

为了进一步改进 CornerNet 方法,Zhou X Y 等人^[9]提出了 CenterNet,在目标的中心引入一个额外的关键点,使得角点匹配更加精确,降低了

CornerNet 中错误匹配角点的概率, 将 COCO 数据集上的平均精度提高到 47.0%, 但这导致了 CenterNet 的预测速度比 CornerNet 慢。

针对目标矩形框表示会产生干扰背景信息, 且与目标本身形状和姿态信息存在较大差距的问题, Zhou X Y 等人^[10]提出了 ExtremeNet, 利用关键点预测网络预测目标矩形框的 4 个极值点(最顶部、最左侧、最底部、最右侧)和 1 个中心点, 取得了不错的检测效果。

Ma T S 等人^[11]基于 CornetNet, 提出了一种新的角点匹配的 anchor free 目标检测方法。Li Q M 等人^[12]设计了一种新颖有效的无锚网络 MPAF-Net 用于多尺度行人检测, 将行人检测简化为一个简单的中心和尺度预测任务。

尽管卷积神经网络特征在自然场景图像中已经取得了很大的成功, 但是由于难以有效地处理物体旋转变化的问题, 无法将其直接用于光学遥感图像中的物体检测。本质上, 此问题对于自然场景图像而言并不是关键问题, 由于地球的重力作用, 目标通常处于垂直方向, 因此图像上的方向变化通常很小。相反, 由于从上空拍摄了遥感图像, 遥感图像中的飞机场、建筑物和车辆等物体通常具有不同的方向。另外, 遥感图像存在小目标占比多、分布密集, 图像拍摄条件不稳定等问题, 还容易受到光照、云层、伪装等因素的影响。因此, 目标检测在遥感图像中的准确率往往低于自然图像。

检测速度在目标检测领域也非常重要。例如, 欧洲哨兵 1 号 (Sentinel-1) 地球观测卫星每 6 天对整个地球进行一次成像, 要求算法必须足够快且具有足够的可转移性才能应用于整个地球表面^[13]。

然而, 目前的遥感目标检测算法大多基于 anchor, 有大量的冗余和较高的计算复杂度, 检测速度相对较慢。基于 anchor free 的目标检测模型虽然损失了部分精度, 但往往能拥有更快的检测速度。因此, 进一步探究基于 anchor free 的目标检测方法对于提升遥感图像的应用能力具有重要意义。

综上, 本文选择 anchor free 的目标检测网络框架, 针对遥感场景的特点, 提出了一种基于角度敏感的空间注意力机制的遥感图像旋转目标检测算法。本文根据旋转框与其外接矩形框的空间位置关系, 提出了一种简单有效的旋转框表示方法。此外, 本文设计了一种用于辅助检测旋转目标的角度敏感的空间注意力机制, 通过引入角度信息, 提升模

型对旋转目标的检测能力。本文提出的算法在公开遥感数据集 DOTA 上进行了实验, 旋转框的目标检测网络平均精度均值 (mAP) 达到了 68.5%, 检测速度达到每秒 17.4 帧图像。

2 基于角度敏感的空间注意力机制的遥感图像旋转目标检测算法

2.1 整体网络结构

本文提出的 anchor free 的基于角度敏感的空间注意力机制的遥感图像旋转目标检测算法基于 CenterNet^[7], 其网络结构如图 1 所示。该网络结构简单, 编码网络负责提取输入图像的特征, 解码网络负责从特征图中提取有用信息并输出。首先, 编码网络主要通过多尺度网络分别采用自下而上、自上而下和横向连接这 3 种路径提取底层、高层特征进行融合, 使得模型能更好地捕获全局和局部信息。接着, 多尺度网络提取的每一层特征图都将通过通道注意力机制, 使得重要的通道能够被分配到更高的权重。针对旋转框目标具有旋转角度这一特点, 本文设计了一个角度敏感的空间注意力机制模块, 将角度信息融入特征图, 使得模型更加关注旋转目标, 提升对旋转目标的检测精度。本文将解码网络改造成适用于表达旋转框的形式。最后, 最终的特征图传入解码网络, 负责预测热力图中每个网格是检测目标中心点的概率、中心点的偏移量以及对应的长和宽、旋转偏移量。

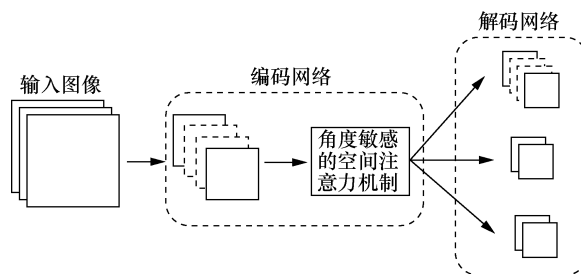


图1 基于角度敏感的空间注意力机制的遥感图像旋转目标检测算法的网络结构

2.2 特征提取网络

遥感图像尺寸大、场景复杂, 缩小尺寸后特征不太明显, 因此需要深层的特征提取网络才能有效地提取物体特征。为了避免深度网络可能存在的梯度消失问题, 本文将 Resnet^[14]作为检测网络的特征提取网络。

Resnet 由于结构简单、有效而被广泛应用在目标检测任务上。Resnet 引入的深层残差块将底层特

征跳跃连接到高层特征，有效解决了由于网络深度的加深而产生的梯度消失、准确率无法有效提升的问题。

Resnet 由头部的卷积层、中间的多个卷积块和尾部的输出层构成，根据中间堆叠的卷积块数量又能细分成 Resnet18、Resnet50、Resnet101。显然，卷积块堆叠得越多，网络越深，能提取到的特征模式越复杂，但是网络参数量也会相应地增加。本文通过详细的实验对比得出，Resnet101 的检测效果更好，带来的时间损耗也在可接受范围内，因此将 Resnet101 作为主干网络。

2.3 多尺度特征融合

由于遥感图像的检测目标的尺寸跨度很大，小至几十像素，大至几百甚至上千像素，仅仅使用高层特征检测目标很容易造成小目标物体漏检，因此需要利用底层特征辅助。本文提出的目标检测网络借鉴特征金字塔网络（feature pyramid network, FPN）^[15]的思想，引入了横向连接方法进行多尺度融合。通常底层特征的语义信息比较少，但是空间位置信息相对准确；高层特征的语义信息比较丰富，但是空间位置信息比较粗略。FPN 能将底层特征和高层特征融合起来，从而更好地利用全局特征。

Resnet 能够自下而上地提取出 5 层不同深度的特征图，如何充分利用这些特征图尤为重要。FPN 的思路是对提取出的每一层特征图分别进行预测，然后汇总每一层的预测结果。大多数应用多尺度结构的基于 anchor 的目标检测算法也沿用这样的方法，在 RPN 阶段分别预测每一层特征图的每个锚框的回归系数和置信度，将回归系数应用在初始锚框的坐标上以获得所有边界框的预测坐标。然后使用非极大值抑制（non-maximum suppression, NMS）算法从所有预测结果中筛选出最优质的边界框，并将其作为最终的输出。这样能从多层预测的大量边界框中筛选出足够少的合适的边界框，减少时间开销。

在基于 anchor 的方法中应用多尺度特征网络通常采用的手段是对每一层特征图进行候选框预测，然后对所有层的候选框进行统一的筛选操作。与基于 anchor 的方法不同，基于 anchor free 的方法并不是直接预测目标矩形框的坐标，而是预测中心点特征图、偏移量和长宽，这就导致每一层的边界框坐标的计算更复杂。因此，本文采用的多尺度特征融合模块如图 2 所示，其中实线箭头指卷积操作，虚线箭头指上采样操作， $\oplus$ 是两个卷积在通道维度

上的连接操作，最终得到的卷积特征图 C'_2 会传到解码网络进行后续的操作。该方法将每一层不同尺寸的特征图上采样至同一尺寸并拼接在一起。这样不仅能够使得高层特征和底层特征都得到充分的利用，而且检测速度也没有太大的损失。

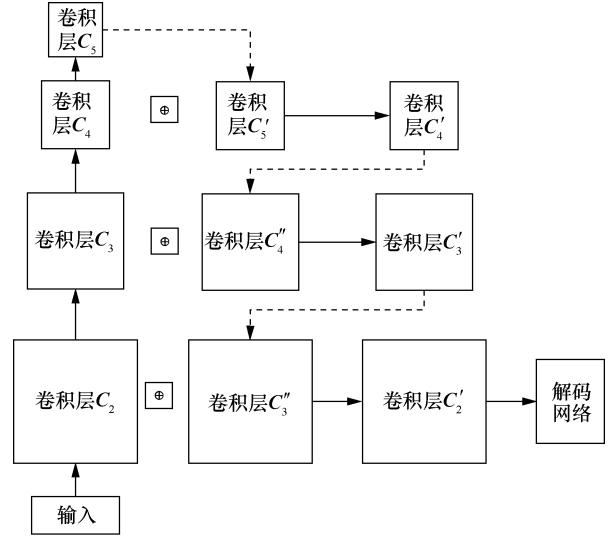


图2 多尺度特征融合模块

2.4 通道注意力机制

虽然卷积神经网络有较强的非线性表达能力，但是当信息量过于复杂时，需要构建更深的网络模型才能获得更强的表达能力。为了降低模型的复杂程度，可以借助人脑处理过载信息的方式，使用注意力机制提高神经网络处理信息的能力。注意力机制能够让模型关注更加重要的信息，忽视不重要的信息。实际上，注意力机制的实现分为 3 个步骤：信息输入、计算注意力分布、根据注意力分布计算输入信息的加权平均。

第 2.3 节介绍了本文提出的目标检测网络使用的多尺度特征融合机制，这个机制需要统一特征提取网络产生的多层特征图的尺寸并在通道维度上进行连接，整个多尺度融合过程使用了多通道的特征。因此，本文希望注意力机制能够为多尺度融合过程涉及的通道分配合适的权重，使得模型可以根据输入图像的特征，动态调整高层特征和底层特征的权重关系。

本文提出的目标检测网络引入了 SENet^[16]中的通道注意力模块 SE block。SE block 主要由两部分构成，分别是压缩和激活，这两部分对原始特征图的非线性计算可看作原始特征图与 SE block 相乘，然后通过 Softmax 激活函数将权重矩阵归一化到

[0,1], 再与原始特征图相乘, 得到权重改变后的新特征图。此时, 原始特征图的通道权重已经得到了合理的分配, 越重要的通道权重越高, 模型越“重视”这个通道的特征。

为了增大感受野, SE Block 将全局空间信息压缩成一个聚合特征, 压缩操作由全局平均池化完成。接下来的激活操作用于激活聚合特征中的信息。这种机制能够学习到通道与通道之间的非线性关系, 而且能够学习到非互斥的关系, 这样才能保证多个通道能够同时作用。

SE block 能够灵活嵌入各种模型中, 因此本文有多种使用 SE block 的方式。经过充分的实验, 最终将 SE block 嵌入特征提取器的每一层特征图, 这使得每一层的特征通道都会得到合理的分配, 最大限度地提高模型的精确度。

融合 SENet^[16]中 SE block 的多尺度特征提取网络结构如图 3 所示。其中左图是 SENet^[16]中的通道注意力模块 SE block, 右图是 SE block 嵌入主干网络形成的多尺度下的通道注意力机制。

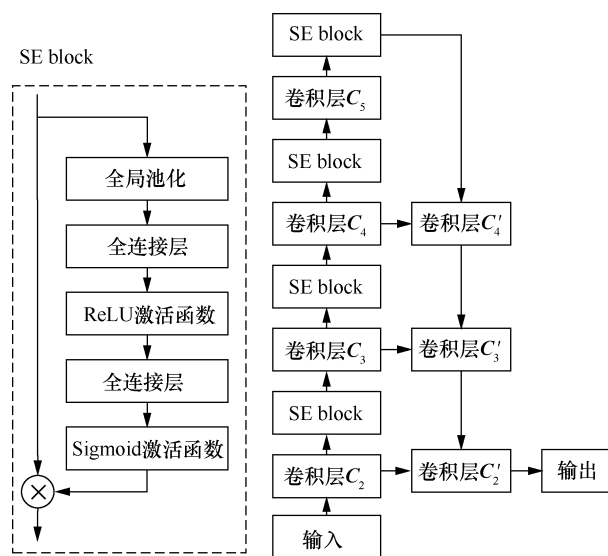


图3 融合 SENet^[16]中 SE block 的多尺度特征提取网络结构

2.5 角度敏感的空间注意力机制

本文利用特征图的空间关系生成了一个空间注意力图。与通道注意力机制不同, 空间注意力机制聚焦图像中与目标检测任务相关的重要的局部信息, 并抑制不相关的信息, 这对通道注意力机制起到了补充作用。

本文设计的角度敏感的空间注意力机制结构如图 4 所示。首先输入图像经过编码网络生成特征图, 经过一个卷积网络后生成了尺寸为(batch,1,w,h)

的角度特征图, 它表示长度为 w 、宽度为 h 的输入图像中对应空间位置上潜在目标的旋转角度, batch 为批次大小。接着计算角度特征图的空间注意力, 沿通道轴分别应用最大池化和平均池化, 分别生成尺寸为(batch,1,w,h)的特征图 **Max** 和 **Avg**。沿通道轴进行连接操作, 再经过一个卷积网络增加非线性表达能力, 然后使用 Sigmoid 激活函数将特征值分布到[0,1], 得到角度敏感的空间注意力权重图, 并与原始的特征图相乘, 给特征图在空间角度分配不同权重的注意力, 生成目标中心热力图。

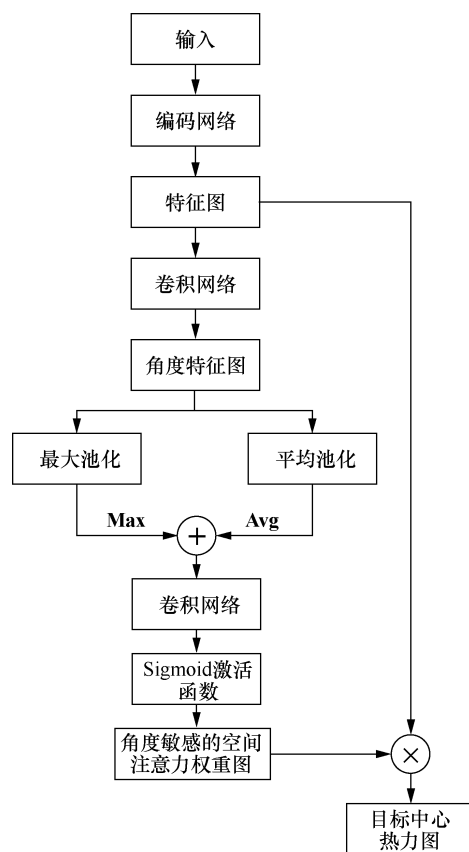


图4 角度敏感的空间注意力机制结构

角度敏感的空间注意力机制的原理是在具有一定旋转角度的空间位置上存在目标的可能性更大。因此, 既然旋转标注比水平框标注增加了旋转角度的信息, 就应该充分利用这一信息。但是实验证明, 直接利用预测出来的旋转角度预测旋转框的效果并不理想, 旋转角度的轻微误差都会导致检测效果剧烈下降。因此, 本文并不直接使用预测出来的旋转角度, 而是间接利用这一个特征对其生成一个角度敏感的空间注意力机制, 即这个注意力机制更加关注具备旋转角度的空间位置。理论上, 用角度敏感的空间注意力生成的特征图预测目标中心

热力图会更加精确，漏检会更少，后续的实验也证明了角度敏感的空间注意力机制的有效性。

3 训练与推理

3.1 数据标签格式

旋转框研究的难点在于合理地表示旋转框，因此本文提出了多种旋转框的表示方式。经过实验，选择出最适用于本文提出的目标检测网络的表示方式。

本文提出的数据标签格式是 $(c_x, c_y, w, h, r_x, r_y, w_1, h_1, w_2, h_2)$ ，利用旋转框的外接矩形来表示旋转框的坐标。本文提出旋转框参数表示方法（方法1）如图5所示，虚线的水平边界框（以下简称边界框）是旋转框的外接矩形。假设模型的输出为 $c_x, c_y, w, h, r_x, r_y, w_1, h_1, w_2, h_2$ ，其中 w, h 分别表示边界框的长与宽， c_x, c_y, r_x, r_y 由以下计算式获得：

$$c_x = \text{int}(p_x) \quad (1)$$

$$c_y = \text{int}(p_y) \quad (2)$$

$$r_x = p_x - c_x \quad (3)$$

$$r_y = p_y - c_y \quad (4)$$

其中， (p_x, p_y) 是中心点的浮点数坐标， (c_x, c_y) 是中心点取整后的坐标， r_x, r_y 为偏移量。

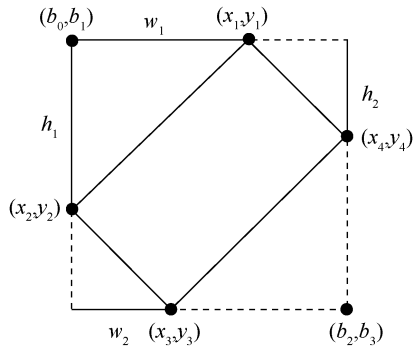


图5 本文提出的旋转框参数表示方法（方法1）

旋转框的4个点坐标 $(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)$ 可用以下计算式获得：

$$b_0 = c_x + r_x - w/2 \quad (5)$$

$$b_1 = c_y + r_y - h/2 \quad (6)$$

$$b_2 = c_x + r_x + w/2 \quad (7)$$

$$b_3 = c_y + r_y + h/2 \quad (8)$$

$$x_1 = b_0 + w_1, y_1 = b_1 \quad (9)$$

$$x_2 = b_0, y_2 = b_1 + h_1 \quad (10)$$

$$x_3 = b_2 + w_2, y_3 = b_3 \quad (11)$$

$$x_4 = b_2, y_4 = b_1 + h_2 \quad (12)$$

3.2 损失函数

本文提出的预测分支共4个，总损失函数的计算式如下：

$$L_{\text{total}} = L_{\text{hm}} + \lambda_1 L_{\text{wh}} + \lambda_2 L_{\text{off}} + \lambda_3 L_{\text{theta}} \quad (13)$$

其中， L_{hm} 是目标中心热力图分支的损失函数， L_{wh} 是边界框宽高预测分支的损失函数， L_{off} 是中心点偏移量预测分支的损失函数， L_{theta} 是角度预测分支的损失函数。 λ_1, λ_2 和 λ_3 是用于调整这几个分支权重的参数。

下面分别介绍这4个分支的损失函数。

（1）目标中心热力图分支的损失函数

该分支的输出实际含义是每个坐标位置是边界框中心点的概率。由于一张遥感图像中的边界框中心点数量占整张图像素的比例很小，即负样本的比例远远高于正样本，因此正负样本数量极度不均匀。为了解决这个问题，本文提出的目标检测网络将改进后的焦点损失（focal loss）作为损失函数。

改进后的损失函数如下所示：

$$L_{\text{hm}} = -\frac{1}{N} (l_{\text{pos}} + l_{\text{neg}}) \quad (14)$$

$$l_{\text{pos}} = \sum (1 - p_{\text{pos}})^\alpha \log p_{\text{pos}} \quad (15)$$

$$l_{\text{neg}} = \sum p_{\text{neg}}^\alpha (1 - \text{gt})^\beta \log (1 - p_{\text{neg}}) \quad (16)$$

其中， N 是热力图像素点的总数。 p_{pos} 和 p_{neg} 分别是目标中心热力图上所有正样本的像素值和所有负样本的预测值，当某个点的真实值 gt 大于一定阈值时，这个点被视为正样本，其余视为负样本。 l_{pos} 、 l_{neg} 分别表示目标中心热力图上所有正样本的像素值的焦点损失函数和所有负样本的像素值的焦点损失函数。 α 和 β 在 focal loss 中被称为调制系数，修改这些系数能够调整易分类样本的权重，使得模型在训练时更专注于难分类的样本。

对于距离目标中心点较远，即真实值较小且模型输出的预测值较大的像素点，通过式（14）~式（16）加大此类负样本的损失，让模型更加关注这些难分类的像素点，更利于模型学习。

划分正负样本的阈值设置得越小，正样本数量越多，但预测精度越不准确。本文提出的旋转目标

检测方法考虑到目标中心热力图是原图下采样 1/4 得到的,本身就有一定的偏差,若阈值设置得较小,模型难以提取出比较精确的中心点,因此这里取阈值为 1,即将热力图中真实值为 1 的点看作正样本,其余均看作负样本。

(2) 边界框宽高预测分支的损失函数

该分支的损失函数是 L_1 损失,其计算式如下:

$$L_{wh} = \frac{1}{N} \sum |\widehat{wh} - p_{pos}| \quad (17)$$

其中, $\widehat{wh} \in R^{w \times h \times 2}$ 是预测的边界框的宽和高。

(3) 中心点偏移量预测分支的损失函数

为了恢复由原图下采样引起的离散误差,本文提出的目标检测网络还预测了每个中心点的局部偏移量 $\hat{o} \in R^{w \times h \times 2}$ 。每一个通道共享相同的预测偏移量。其损失函数为 L_1 损失,计算式如下:

$$L_{off} = \frac{1}{N} \sum |\hat{o} - p_{pos}| \quad (18)$$

注意,这里只计算所有正样本的偏移量,负样本忽略不计。

(4) 角度预测分支的损失函数

L_{theta} 是角度预测分支的损失函数,其损失函数仍然采用 L_1 损失。这里的角度定义如下:

$$\theta = \begin{cases} \text{atan} \frac{w_i}{h_i}, h_i \neq 0 \\ 0, h_i = 0 \end{cases} \quad (19)$$

4 实验验证

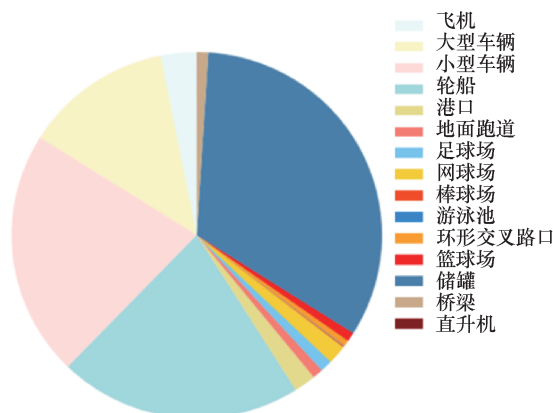
4.1 DOTA 数据集

DOTA 数据集是用于航空图像中目标检测的大规模数据集。它的图像是从 Google Earth 中手动收集的,其中一些是由吉林一号卫星拍摄的,其他则是由中国资源卫星数据与应用中心的高分二号卫星拍摄的。它包含 2 806 张航拍图像,每张图像的大小在 800×800 像素到 4 000×4 000 像素之间,并且包含具有不同的比例、方向和形状的样本。DOTA 数据集包含 15 个类别:飞机、轮船、储罐、棒球场、网球场、篮球场、地面跑道、港口、桥梁、大型车辆、小型车辆、直升机、环形交叉路口、足球场和游泳池。

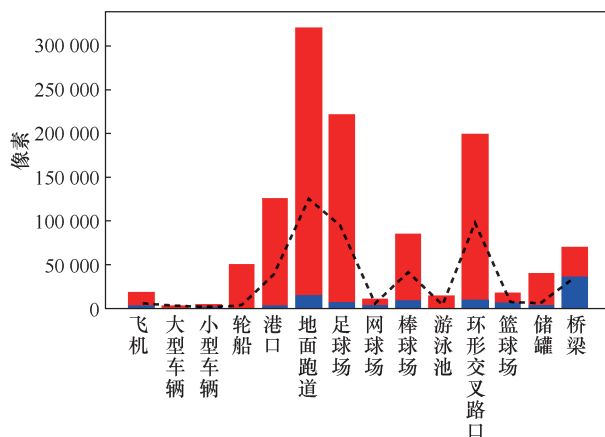
DOTA 数据集提供了 1 411 张训练集图像与 458 张测试集图像。本文对 DOTA 数据集中各类别数量进行了统计,不同类别数量占比如图 6(a)

所示,可以看出在所有样本中,80%以上是大型车辆、小型车辆、轮船和储罐,其余 11 个类别样本数量很少。样本类别不均匀造成了训练和检测的困难。

由于 DOTA 数据集中直升机的样本数量偏少,本文只对其余 14 个目标的尺寸进行了统计,不同类别的样本最大像素和最小像素总和如图 6(b) 所示。不同类别之间的样本尺寸相差很大,甚至同一个类别的不同样本之间的尺寸相差也很大,因此很难选择一个合适的尺寸分割原始图像。经过实验,本文将原始图像切割成 600×600 像素的图像,并将其作为输入图像,得到输出结果后重新合成原始图像的尺寸。采用 RandAugment 算法^[17]对输入图像进行图像增强。



(a) 不同类别目标数量占比



(b) 不同类别的样本的最大像素和最小像素总和

图 6 DOTA 数据集数据统计

4.2 不同旋转框参数表示方法的效果验证

本节针对多种不同的旋转框参数表示方法进行效果测试,证明本文提出的旋转框参数表示方法更加简单有效。

(1) 旋转框参数表示方法 2 (方法 2): $(c_x, c_y, w, h, r_x, r_y, w_1, h_1)$

该方法同样利用旋转框的外接矩形来表示旋转框的坐标, 但没有充分利用旋转框本身也是一个矩形的特征, 因此存在信息冗余。实际上, 只需要两个额外参数 w_1 和 h_1 , 根据两个三角形的全等关系就可以得出对边长度相等, 如图 7 所示。

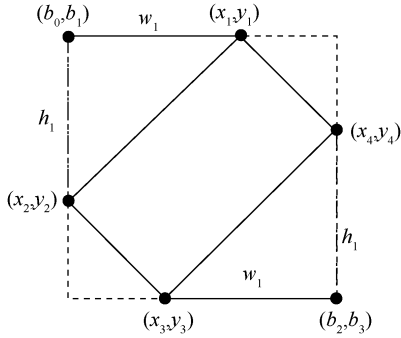


图 7 旋转框参数表示方法 2

旋转框的 4 个点坐标 (x_1, y_1) 、 (x_2, y_2) 、 (x_3, y_3) 、 (x_4, y_4) 可以用以下计算式得到:

$$x_1 = b_0 + w_1, y_1 = b_1 \quad (20)$$

$$x_2 = b_0, y_2 = b_1 + h_1 \quad (21)$$

$$x_3 = b_2 - w_1, y_3 = b_3 \quad (22)$$

$$x_4 = b_2, y_4 = b_3 - h_1 \quad (23)$$

(2) 旋转框参数表示方法 3 (方法 3): $(c_x, c_y, w, h, r_x, r_y, \theta)$

旋转框一般用旋转框中心点坐标、长宽与旋转角度来表示, 实际上 $(c_x, c_y, w, h, r_x, r_y, w_1, h_1)$ 中的 w_1 、 h_1 蕴含了角度关系。可以直接预测角度 θ , 再算出 w_1 、 h_1 即可表示旋转框, 如图 8 所示。

旋转框的 w_1 、 h_1 可以用下式计算得到:

$$t = \tan \theta \quad (24)$$

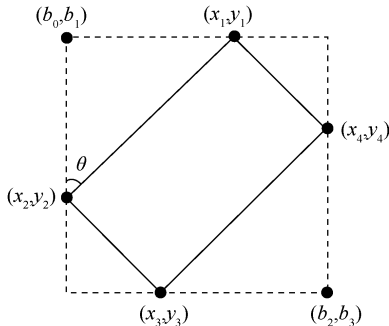


图 8 旋转框参数表示方法 3

$$h_1 = \frac{t(tw - h)}{t^2 - 1} \quad (25)$$

$$w_1 = th_1 \quad (26)$$

(3) 旋转框参数表示方法 4 (方法 4): $(c_x, c_y, w_0, h_0, r_x, r_y, \theta_0)$

除了借助外接矩形的方法表示旋转框, 还可以直接表述旋转框, 如图 9 所示, w_0 、 h_0 分别表示旋转框的宽和高, θ_0 表示该旋转框与水平轴的旋转角度, 限定 θ_0 的范围为 $0^\circ \sim 90^\circ$ 。根据已知变量可以求出 $w_1 = w_0 \cos \theta$, $h_1 = w_0 \sin \theta$ 。

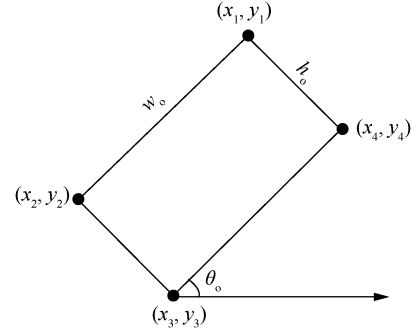


图 9 旋转框参数表示方法 4

本节根据提出的 4 种旋转框参数表示方法分别进行测试, 实验将 CenterNet 作为基础网络, 主干网络选择 Resnet101。不同表示方法的旋转框实验结果见表 1。从实验结果看, 虽然这些方法基于同一个网络框架, 但是检测效果却比较悬殊。其中, 方法 1 的 mAP 最高, 达到了 63.7%, 比方法 2 高 4.4%。方法 1 和方法 2 的原理都是先预测出一个边界框, 再基于此预测旋转框的 4 条边的偏差值, 而方法 3 和方法 4 则是预测物体的旋转角度。从实验结果看, 方法 1 和方法 2 的整体效果远远好于方法 3 和方法 4。预测旋转角度是极为困难的, 角度细微的偏差都会使最终预测的边界框的差距极大。而预测边的偏差则对误差的承受能力更强, 一条边的误差并不会导致最终预测的边界框整体变形。方法 2 比方法 1 的 mAP 稍低, 这表明虽然方法 2 的结构相对简单, 但是不可避免地会带来精度损失。

4.3 角度敏感的空间注意力机制的有效性验证

本节对提出的角度敏感的空间注意力机制 (用 atten 表示) 的有效性进行验证。将不添加角度敏感的空间注意力机制的本文提出的目标检测网络作为参考网络, 与添加该注意力机制的网络进行对比。实验中分别将 Resnet50 和 Resnet101 作为特征主干网络, 实验结果见表 2。

表1 不同表示方法的旋转框实验结果

类别	方法1	方法2	方法3	方法4
飞机	93.0%	92.2%	71.8%	57.1%
轮船	80.6%	80.1%	14.5%	51.9%
储罐	89.3%	89.7%	83.9%	60.4%
棒球场	67.7%	54.3%	57.4%	37.7%
网球场	94.6%	94.8%	55.7%	51.8%
篮球场	62.3%	50.7%	46.6%	40.6%
地面跑道	53.7%	44.4%	42.3%	41.5%
港口	55.5%	46.5%	19.3%	34.5%
桥梁	29.7%	24.2%	9.0%	23.5%
大型车辆	78.7%	76.5%	9.3%	43.4%
小型车辆	53.4%	57.1%	17.0%	38.8%
直升机	13.8%	25.7%	41.5%	29.6%
环形交叉路口	69.6%	65.3%	64.3%	47.2%
足球场	62.6%	39.4%	47.9%	38.8%
游泳池	52.2%	49.2%	28.1%	35.1%
mAP	63.7%	59.3%	40.5%	42.1%

表2 本文提出的目标检测网络的角度敏感注意力机制效果测试

类别	Resnet50	Resnet50+atten	Resnet101	Resnet101+atten
飞机	88.6%	88.7%	89.0%	88.8%
轮船	79.0%	82.7%	79.2%	84.0%
储罐	86.2%	86.6%	87.1%	85.0%
棒球场	68.6%	72.6%	73.8%	72.8%
网球场	90.7%	90.4%	90.8%	90.7%
篮球场	65.2%	52.2%	63.0%	63.5%
地面跑道	57.7%	55.0%	57.6%	57.1%
港口	63.1%	61.8%	63.0%	61.6%
桥梁	45.5%	40.2%	42.9%	41.9%
大型车辆	80.7%	77.9%	79.4%	80.7%
小型车辆	60.3%	60.5%	61.8%	62.8%
直升机	25.2%	57.7%	54.9%	53.1%
环形交叉路口	66.0%	60.8%	63.7%	66.8%
足球场	66.6%	63.3%	54.3%	56.2%
游泳池	59.8%	59.8%	62.2%	62.8%
mAP	66.8%	67.3%	68.1%	68.5%

从两个对比试验结果来看,增加了角度敏感的空间注意力机制以后,mAP提升了0.4%~0.5%,提升的效果接近。使用Resnet101比使用Resnet50的mAP更高。

4.4 通道注意力机制的有效性验证

本文为每一层特征都加上提出的通道注意力机制,使得特征通道的分配更合理。通道注意力机制实验结果见表3,实验发现mAP提高了2.3%,检测速度达到了每秒18.6帧图像,效果十分显著。

4.5 多尺度特征融合算法实验效果对比

多尺度特征融合实验结果见表4,可以看出,多尺度特征融合对Resnet101的提升较大,达到了1.2%,然而对Resnet50没有提升,甚至有轻微的降低。分析原因发现,一方面可能是训练不稳定导致效果轻微降低,另一方面可能是Resnet101提取出的高层特征的语义信息比Resnet50更多更全面,和底层特征的融合效果更好。总之,多尺度特征融合在合适的主干网络深度的情况下能够起到较好的效果。

表3 通道注意力机制实验结果

算法	特征提取网络	mAP
无通道注意力机制	Resnet101	62.8%
有通道注意力机制	Resnet101	65.1%

表4 多尺度特征融合实验结果

算法	特征提取网络	mAP
+多尺度融合	Resnet50	62.4%
无多尺度融合	Resnet50	62.7%
+多尺度融合	Resnet101	64.0%
无多尺度融合	Resnet101	62.8%

4.6 不同算法在DOTA数据集上的实验效果

将本文提出的基于角度敏感的空间注意力机制的遥感图像旋转目标检测算法与不同的预测旋转框的遥感算法RRPN^[18]、R3det^[19]、R2CNN^[20],以及经典的Faster R-CNN^[3]的旋转框版本进行对比分析。其中mAP结果采用了参考文献[19]论文中的实验结果,速度测试均在相同环境下进行,实验结果见表5。

表5 不同算法在DOTA数据集上的实验结果

算法	mAP	每秒处理帧数/帧	模型大小/MB
Faster R-CNN ^[3]	52.9%	12.5	237
R2CNN ^[18]	60.6%	0.05	359
RRPN ^[20]	61.0%	0.09	369
R3det ^[19]	73.7%	1.5	787
本文提出的算法	68.5%	17.4	218

从检测的精确度来看，R3det 的 mAP 最高，本文提出的算法次之，Faster R-CNN 的效果最差，只有 52.9%。但是从检测速度上看，R2CNN、RRPN、R3det 虽然同样是基于 anchor 的二阶段网络，但是比原版 Faster R-CNN 慢很多，这是因为它们针对旋转框加入了过多的针对性的改进，导致检测速度下降。本文提出的算法速度远超其他算法，这是基于 anchor free 网络的优势。在模型的大小上，本文网络的模型也是最小的。R3det 虽然精确度最高，但是模型太大，不够轻量化。因此虽然旋转框比水平框的检测精确度稍低，但是同样达到了精确度与检测速度的平衡。

4.7 水平框与旋转框对比分析

本文提出的基于角度敏感的空间注意力机制的遥感图像旋转目标检测算法可以同时检测水平框与旋转框，这是因为水平框是一种特殊角度的旋转框。因此，本节对比分析了基于 4 种旋转框参数表示方法的遥感图像目标检测效果，见表 6。

表 6 4 种旋转框参数表示方法的水平框与旋转框的 mAP

旋转框参数表示方法	水平框	旋转框
方法 1	64.9%	63.7%
方法 2	64.3%	59.3%
方法 3	64.4%	40.5%
方法 4	64.7%	42.1%

从表 6 可以看出，其实 4 种方法的水平框的检测效果几乎一致，都在 64% 左右。而旋转框的标注更细致，对检测的精确度要求更高。方法 3 和方法 4 虽然检测出了旋转目标的大致范围（即水平框），却无法对旋转框的检测起到很大的帮助，这说明旋转框检测比水平框检测更难，旋转框检测不仅需要检测出目标物体的位置，还需要精准地检测出旋转角度。此外，不同的旋转框的表示方法也对模型效果造成了非常重要的影响。

本文对旋转框与水平框的部分检测图像进行了可视化，以便于分析与对比。图 10 的每张图像代表了每个类别的部分检测结果（矩形框内部的黄色框为旋转框）。从旋转框与水平框的整体标注效果来看，旋转框的效果明显优于水平框。例如图 10 (e)、图 10 (j)，水平框的标注存在大量重叠，导致肉眼无法很好地分辨不同的目标物体；旋转框的标注则分界线十分清晰，完美贴合了目标物体。

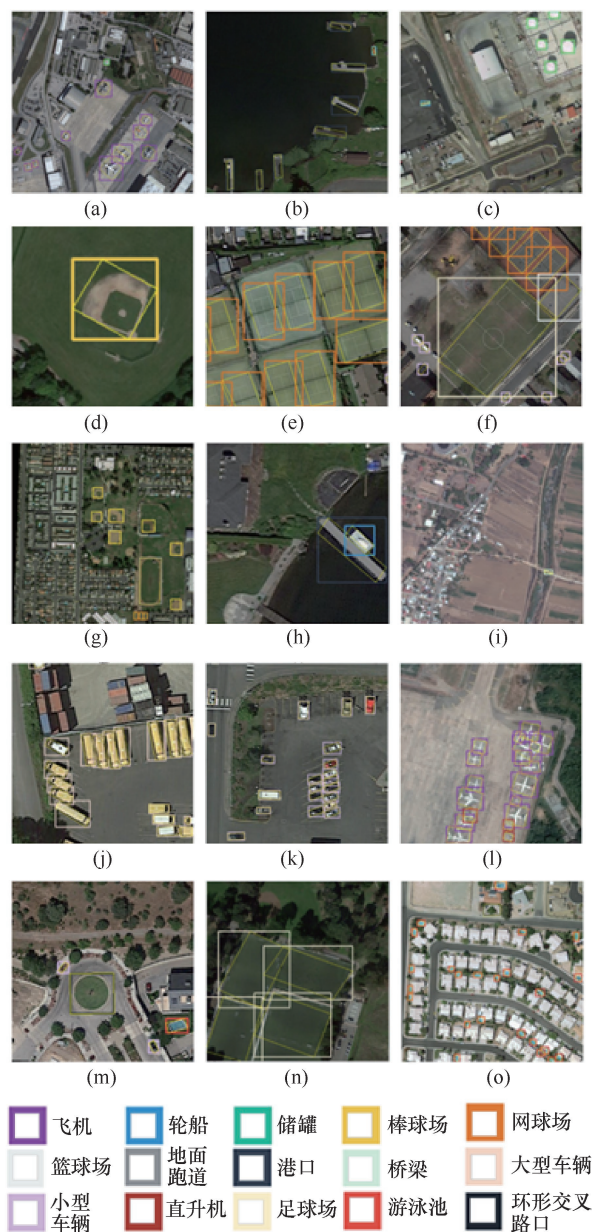


图 10 DOTA 数据集上水平框与旋转框的检测结果

从旋转框的检测效果来看，本文提出的算法对旋转角度的预测效果很好，不管是哪一类别，只要找准了正确的目标中心与正确的水平框的长宽，基本上能够正确地找出旋转的角度。因此，本文提出的算法中旋转框的准确度实际上是非常依赖水平框的准确度的，只有水平框检测准确，旋转框才有可能检测准确。

总体上，本文提出的目标检测网络能够较好地克服遥感场景的困难之处。遥感目标物体的分布密集、部分重叠，如图 10 (k)、图 10 (l) 的目标十分密集，但是本文提出的目标检测网络能够较好地检测出所有目标，并且旋转框不会重叠。甚至本文

提出的目标检测网络可以解决两个目标物体重叠的问题,如图 10 (h),轮船和港口紧靠在一起,会出现不同类别的标注重叠在一起的情况,本文提出的目标检测网络能够通过多通道各司其职、独立检测来克服这个困难,这对于基于 anchor 的方法来说是一个难题。遥感目标物体的尺寸跨度很大,如图 10 (i) 中的桥梁,在整张图中显得极为渺小,但是本文提出的目标检测网络仍然能够检测出来。又如图 10 (f) 中小型车辆和篮球场的尺寸差距很大,但检测效果也很好。

5 结束语

遥感图像的目标检测技术在农业土壤管理与灌溉、自然灾害预警、地球表面资源预估、军事管理等方面占据着非常重要的地位。

一方面,随着卫星技术的提高,每天从空中收集到数以万计的遥感数据,并传输到地球表面,这对遥感检测的速度有了更高的要求;另一方面,利用基于 anchor free 的目标检测网络检测遥感图像的研究甚少。虽然基于 anchor free 的目标检测网络的检测速度较快,但是在检测遥感目标上仍有很多的困难与挑战:基于 anchor free 的目标检测网络自身的精度不高,难以满足精度的要求;遥感目标可以任意角度旋转且分布密集,水平框的标注效果很差。针对这两个难点,本文分别对水平框的目标检测算法与旋转框的目标检测算法进行了研究,并提出了两种用于遥感图像的目标检测算法,具体成果如下。

相比水平框的目标物体检测,旋转框的目标物体检测在遥感场景中更具优势,但也更有挑战性。本文根据旋转框与其外接矩形框的空间位置关系,提出了一种简单高效的旋转框位置的表示方法。针对旋转框增加的旋转角度的信息,设计了一个角度敏感的空间注意力机制用于辅助检测旋转框。在 DOTA 数据集上,本文提出的算法的 mAP 达到了 68.5%,检测速度达到了每秒 17.4 帧图像。与其他网络相比,本文设计的目标检测网络具有突出的速度优势,在检测精度上也取得了不错的效果。

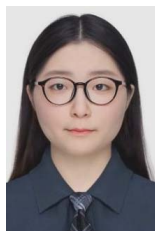
参考文献:

- [1] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 580-587.
- [2] GIRSHICK R. Fast R-CNN[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2015: 1440-1448.
- [3] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence. Piscataway: IEEE Press, 2015: 1137-1149.
- [4] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 2980-2988.
- [5] HUANG L C, YANG Y, DENG Y F, et al. DenseBox: unifying landmark localization with end to end object detection[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2015.
- [6] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 779-788.
- [7] LAW H, DENG J. CornerNet: detecting objects as paired keypoints[J]. International Journal of Computer Vision, 2020, 128(3): 642-656.
- [8] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[M]//Computer Vision-ECCV 2014. Cham: Springer International Publishing, 2014: 740-755.
- [9] ZHOU X Y, WANG D Q, KRÄHENBÜHL P. Objects as points[J]. arXiv preprint, 2019, arXiv:1904.07850.
- [10] ZHOU X Y, ZHUO J C, KRAHENBUHL P. Bottom-up object detection by grouping extreme and center points[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019.
- [11] MA T S, TIAN W H, KUANG P, et al. An anchor-free object detector with novel corner matching method[J]. Knowledge-Based Systems, 2021, 224: 107083.
- [12] LI Q M, QIANG H, LI J. Conditional random fields as message passing mechanism in anchor-free network for multi-scale pedestrian detection[J]. Information Sciences, 2021, 550: 1-12.
- [13] CHENG G, ZHOU P C, HAN J W. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2016, 54(12): 7405-7415.
- [14] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.
- [15] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 936-944.
- [16] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7132-7141.

- [17] CUBUK E D, ZOPH B, SHLENS J, et al. RandAugment: practical automated data augmentation with a reduced search space[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2020: 3008-3017.
- [18] 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.
- [19] YANG X, LIU Q Q, YAN J C, et al. R3det: refined single-stage detector with feature refinement for rotating object[J]. arXiv preprint, 2019, arXiv:1908.05612.
- [20] JIANG Y Y, ZHU X Y, WANG X B, et al. R2CNN: rotational region CNN for orientation robust scene text detection[J]. arXiv preprint, 2017, arXiv:1706.09579.



杨萌（1986—），男，博士，中国舰船研究设计中心高级工程师，主要研究方向为舰船总体技术、环境智能感知。



顾曦（1996—），女，同济大学电子与信息工程学院硕士生，主要研究方向为机器学习、深度学习。

[作者简介]



尹开石（2001—），男，东华大学计算机科学与技术学院在读，主要研究方向为图像分析、机器学习。



王志成（1975—），男，博士，同济大学电子与信息工程学院副研究员，主要研究方向为机器学习、深度学习。