摘 要

遥感技术是通过各类遥感平台（如卫星或机载平台）的传感器探测和识别地物的一类综合技术。随着无人机技术和高分辨率卫星的不断发展，空间分辨率不断提高，遥感图像和光谱信息也日益丰富。如何高效、准确地对高分辨率遥感图像数据进行分析和处理，是当前遥感图像目标检测的研究重点。鉴于遥感图像在俯视视角下通常呈现任意方向排布，故常用的检测算法往往在遥感图像下检测效果不理想，对于复杂背景目标、小目标和密集目标的检测，还有较大的提升空间。

为了更好地实现遥感图像目标检测，本文在单阶段检测网络 YOLOv5 的基础上，提出了一种基于旋转框、注意力机制和加权双向特征金字塔网络结构的遥感图像目标检测算法（YOLOv5_CB）。在 DOTA 数据集上实验结果表明，改进的 YOLOv5_CB 模型比 YOLOv5 模型的 mAP@0.5、mAP@0.5:0.95 分别提升了 3.0% 和 1.1%。

关键词：遥感图像 目标检测 旋转框 注意力机制 加权双向特征金字塔

ABSTRACT

Remote sensing technology is a type of comprehensive technology that detects and identifies features through the sensors of various remote sensing platforms (such as satellites or airborne platforms). With the continuous development of UAV technology and high-resolution satellites, the spatial resolution is increasing and remote sensing images and spectral information are becoming more and more abundant. How to analyze and process high-resolution remote sensing image data efficiently and accurately is the focus of current research on remote sensing image object detection. In remote sensing image object detection, remote sensing images usually present arbitrary directional arrangements under the overhead view. So the commonly used detection algorithms often have unsatisfactory detection effects under remote sensing images. And there is still a large room for improvement in the detection of complex background objects, small objects and dense objects.

To better achieve remote sensing image object detection, I improve the single-stage detection network YOLOv5 in this thesis. I propose a remote sensing image object detection algorithm (YOLOv5_CB) based on rotating object frame, attention mechanism and weighted bi-directional feature pyramid network structure. Experimental results on the DOTA dataset show that the improved YOLOv5_CB model of this thesis outperforms the YOLOv5 model by 3.0% and 1.1% for mAP@0.5 and mAP@0.5:0.95, respectively.

Keywords: remote sensing image object detection rotating frame attention mechanism weighted bidirectional feature pyramid

目 录

第一章 绪论

1.1 研究背景及意义

遥感技术是一种以航空或卫星平台为基础的远程遥感技术，它能采集到地面目标的电磁波辐射、反射和散射特性，并将其成像为遥感影像。由于技术、经济等因素的影响，在我国开展遥感技术的研究相对滞后。而如今随着光学遥感器件的发展与研制，该领域也迎来快速发展。我国研制发射新一代遥感卫星。资源三号卫星搭载有四台地面分辨率分别为 3.6m、5.8m、2.1m 和 3.6m 的光学相机。高分二号卫星所拍摄的遥感图像分辨率可达到 0.8m，处于世界先进水平[1]。

随着高分辨率卫星的研发与使用，高质量的遥感图像数据包含着更加丰富的地物信息，遥感技术先起始于军事领域，后逐渐应用至民用，也在交通监控、城市规划、资源调查等方面起到广泛应用。因此，遥感图像目标检测具有重要的研究价值和应用前景[2]。

由于大范围的遥感影像具有复杂的前景和目标，因此如何准确、迅速地进行目标识别、分类和语义分割已是许多学者关注的焦点。利用该方法对多个特征进行提取是遥感影像中的一项关键技术。在传统的遥感影像中，一般采用手工破译的方法来进行特征的抽取，其中的几何特性包括形状、颜色和纹理等，但单靠人工进行视觉破译费时费力。另外，传统算法通常采取特定条件下的模版匹配，也就是对某一具体问题进行特征提取，故目标检测的良好效果只局限于某一特定的场景，故传统方法缺乏鲁棒性，明显不适用于地物环境复杂的遥感图像。传统目标检测算法处理遥感图像存在着明显缺陷。

作为 CV 领域有较大挑战性的课题之一，遥感目标检测取得的巨大进步也得益于深度学习技术的飞速发展。基于深度学习的目标检测技术是通过卷积神经网络来完成对被测物体的特征的自动学习与辨识。核心是建立一个多层次的机器学习系统，它能够获得海量的数据，并利用这些信息对其进行推理分析和分类预测。深度学习已有许多出色的结果，如双阶段的 Faster R-CNN 网络[3]、单阶段的 YOLO 网络[14]及 SSD 网络[15]。这些网络得到的模型具有良好的鲁棒性，与常规方法相比具有明显的优越性。

虽然有一定进展，但是当今技术距离高精度、强适应性和良好实时性的现实需求依然还有较为明显的差距。鉴于遥感图像检测有着图像尺寸大、背景复杂、小目标数量大等特征，因此存在着诸多挑战：（1）尺寸过大的航空图像导致常规目标检测网络显存超载；（2）小目标所占像素少，与周围背景难以进行区分；（3）目标分布不均匀、稀疏，导致检测效率低；（4）密集目标存在易被遮挡，导致检测困难；（5）数据集中类别不均衡，导致长尾效应。（6）俯拍的卫星遥感图像角度不固定且方向多变，导致检测难度大。

如果把上述问题放在首位，可以推动深度学习的深入发展，从而为实现我国遥感技术的智能化、自动化奠定基础。因此，无论是在理论上还是实践中，对已有的目标探测算法进行深入的研究都有很大的实用价值。

1.2 国内外研究现状

1.2.1 自然图像目标检测现状

回顾过去二十年的历史，自从 Hinton 等于 2006 年提出，若一个神经网络包含多个隐藏层，可以更好地获取特征，深度学习就逐渐引起了研究者们的重视。新的网络结构不断出现，例如 AlexNet[4]、Overfeat[5]、VGG[6]、GoogleNet[7]、ResNet[8]。目标检测领域的发展主要分为两个时期，前一个时期为以可变性组件模型（Deformable Part-based model, DPM）[9]、VJ 检测器、方向梯度直方图检测器为代表的传统目标检测。在提取特征方面，卷积神经网络相较于人工提取更加丰富全面。

Krizhevsky 等人在 2012 计算机视觉竞赛 ILSVRC 上推出了改良的 AlexNet，该网络架构在当时图像分类中具有最好的识别精度。此后，深度学习得到了越来越多的应用，R.Girshick 等在 2014 年提出典型的 R-CNN 目标检测算法[10]，该方法利用卷积神经网络对目标进行特征抽取，并将其分为两个子工作：获取对象的位置并分类。后续又以此为基础进行了两次改进：Fast R-CNN[11]、Faster R-CNN 算法。另外，R-CNN 的缺点也很明显，在生成目标候选区域时，由于使用了选择性搜索方法，产生了大量的冗余候选框，而且重复运算会使卷积神经网络的运算速度下降。此外，采用 AlexNet 导致其输入图像的大小必须是固定的，鲁棒性不强。针对以上问题，何恺明等于 2015 年发表 SPP-Net 算法[12]，提出 SPP 结构加于卷积层和全连接层中，这种新的方法既能有效地克服原有的图像失真和比例失调，提高检测效

率。并于 2 年后推出 Mask R-CNN 算法[13]，这种算法有效克服了特征图和图像的 RoI 不容易对齐的问题。

上述都是双阶段（Two-stage）方案，皆需候选区域生成和区域分类两个步骤。2016 年 YOLO 算法的提出开创了单阶段(One-stage)的先河 SSD。YOLO 算法彻底放弃了双阶段目标检测器“候选建议框+预测框”这一机制，将输入的图片按比例缩小到相同的尺寸，分割为若干个单元，然后根据该单元的位置对该对象进行预测。YOLO 算法仅仅只用一个卷积神经网络完成特征提取、候选框回归和分类，其中有 24 个卷积层，2 个全连接层。实时检测的速度可达 45 帧每秒（frames per second, fps），显著提升了目标检测器的速度，这意味着深度学习目标检测算法开始能真正着手应用于实时检测任务并有效。随后，LiuW 等人提出的 SSD 算法对 YOLO 进行了改进，从多个特征图上进行预测以检测不同尺度大小目标，使得 One-stage 算法保证了精度和较高的检测速度。在 2017 年，Lin T Y 等提出了基于 top-down 与 bottom-up 结构的特征金字塔（FPN）算法[16]，同年 R. Joseph 等采取一套完善 YOLO 的策略，提出了 YOLO9000[17]。YOLO9000 增加主干网络 Darknet-19，除了提高分辨率和多维聚类产生锚框等改善外，增加了全新的联合训练模式。

2018 年，R.Joseph 等又提出了 YOLOv3[18]，该算法使用了多尺度特征图来进行边界框预测，设计了 Darknet-53 网络用于特征提取。次年，Mingxing Tan 等提出了基于 EfficientNet[19]的可扩展模型架构 EfficientDet[20]。该架构的主要贡献有两点：一是新型的基于 BiFPN 的特征抽取方法；第二个特点是将网络宽度深度等进行统一缩放，使得网络在受到各种资源约束的情况下可以很容易地进行改动。2020 年，Alexey Bochkovskiy 等提出 YOLOv4[21]，该算法主要创新三点：首先在现有算法基础上继续进行简化和高效化；第二点是优化了空间注意力模块（Spatial Attention Module，SAM）、路径聚合网络（PathAggregation Network，PAN）等算法，使得模型训练便捷化，只需 1 个 GPU 即可对目标检测器进行有效的训练。YOLOv5 在短短两个多月后发布，输入端增加自适应图片缩放和 Mosaic 数据增强，Mosaic 数据增强的优点是使目标的背景丰富化，且 Mosaic 方法通过随机调整、缩放和拼接小目标数据量增大，通过丰富数据集的方式使网络更加鲁棒。同时引入自适应锚框计算、CSP 结构和 Focus 结构等，在保证精度的前提下，实现了模型的轻量化，并增强了 CNN 的学习性能。YOLOv5 可以减低模型的储存空间，从而减少总的配置成本。

1.2.2 遥感图像目标检测现状

虽然在2015年发布的UCAS-AOD[22]数据集和第二年公布的NWPU VHR-10[23]数据集规模很小，但是在遥感图像小目标检测与研究领域有着不可忽视的贡献。2018 年发布的 YOLT[24]算法作为阶段性突破应用在遥感目标检测领域，基于YOLOv2 进行改进，对遥感图像的像素阈值的选定和切割进行阐述，并分析了对冗余预测框的删除及后续的合并操作。Zhang G 等于 2019 年提出了 CAD-NetP4[25]，分别在全局场景级和局部目标级提取上下文信息，对 R-CNN 和 FPN 网络进行优化，空间感知注意模块的设计引导网络倾向于信息量更多的部分和对图像特征尺度进行适应性调整。

在国内，很多大学都在积极地进行遥感影像的目标检测研究，2017 年，北京大学 Zhang L.为提高对油罐的检测精度，提出综合了 CNN 和支持向量机的目标检测算法。同年，武汉大学和华中科技大学联合制作的 DOTA[26]数据集包含 2806 张遥感图像，共有 15 个类别，标注方法分为目标检测有向边界框（OBB）及水平边界框（HBB）。西北工业大学程塨教授等人于 2019 年开源了 DIOR[27]光学遥感图像数据集。近年来，公开数据集逐渐增多，适用于遥感图像的目标检测算法正在得到广泛应用。

为了解决遥感影像中的小物体探测问题，现有的主流算法是采用特征金字塔网络来提高检测精度。例如，Guo 等提出多尺度 RPN 对小目标检测的精度进行提升，Wang 等提出多尺度视觉注意力网络 MS-VANs 来缓解小目标与背景噪声难以区分的问题。

遥感图像中的目标在俯视下呈现任意方向排列，常规的基于水平框的检测算法无法满足此类场景的应用需求，近年来，基于旋转框的精细定位技术成为遥感领域研究的热点。Jiang 等根据 Faster RCNN 的结构，提出了 R2CNN 算法，在第二阶段对旋转框进行预测。Ma 等人提出双阶段的 RRPN 算法，增设旋转锚框并在其中加入方向参数，使其能检测任意方向目标。上述两种算法最初应用在倾斜文本检测，之后逐渐在遥感目标检测领域取得良好效果。Ding 等提出 RoI Transformer 算法，增设 RoI 空间变换结构，在 OBB 标注的监督下，空间变换的参数学习 RRoI，更具判别性的 RRoI 能对旋转目标检测性能作出明显提升。Yang 等提出 SCRDet 网络解决密集目标检测，增加特征融合结构 SF-Net，通过有监督的多维注意力网络

降低背景噪声的负面效应，在损失函数的改进方面，加入 IoU 常数因子缓解旋转框回归过程的边界问题。

从上述算法的发展历程不难看出，目前目标检测的主流算法是基于深度学习的，如何遥感图像信息的有效提取，实现特定目标快速准确地识别定位，并且提升特定对象尤其是小目标检测的精确度，是当今遥感图像解析处理的一个重要研究方向。

1.3 本文的研究内容与工作安排

第一章 绪论。首先介绍了遥感图像目标检测的背景和意义，然后阐述自然及遥感图像目标检测技术的发展。紧接着介绍了本文及各章节的主要工作内容。

第二章 目标检测相关理论基础。首先对 ANN 和 RNN 进行了介绍，并对当前的先进的目标识别算法作较为详尽的分析，接着阐述遥感图像目标检测的流程及评判指标。

第三章 YOLOv5 目标检测算法。针对 YOLOv5 目标检测算法，对其数据增强、网络结构及损失函数三个模块的相关理论进行详细阐述，最后选定 YOLOv5 算法为本文研究和改进的基础。

第四章 基于 YOLOv5 的遥感目标检测算法。首先给出了实验环境、数据集和性能评价指标的相应描述，然后基于 YOLOv5 算法提出了一种基于旋转框、注意力机制和加权双向特征金字塔网络结构的改进方案，通过试验验证了该改进方案，进行了相应的消融试验，并对其进行了理论上的分析和总结。

第五章 总结与展望。对本文的工作进行了简单的总结，并结合论文的框架，对今后的工作进行了展望。

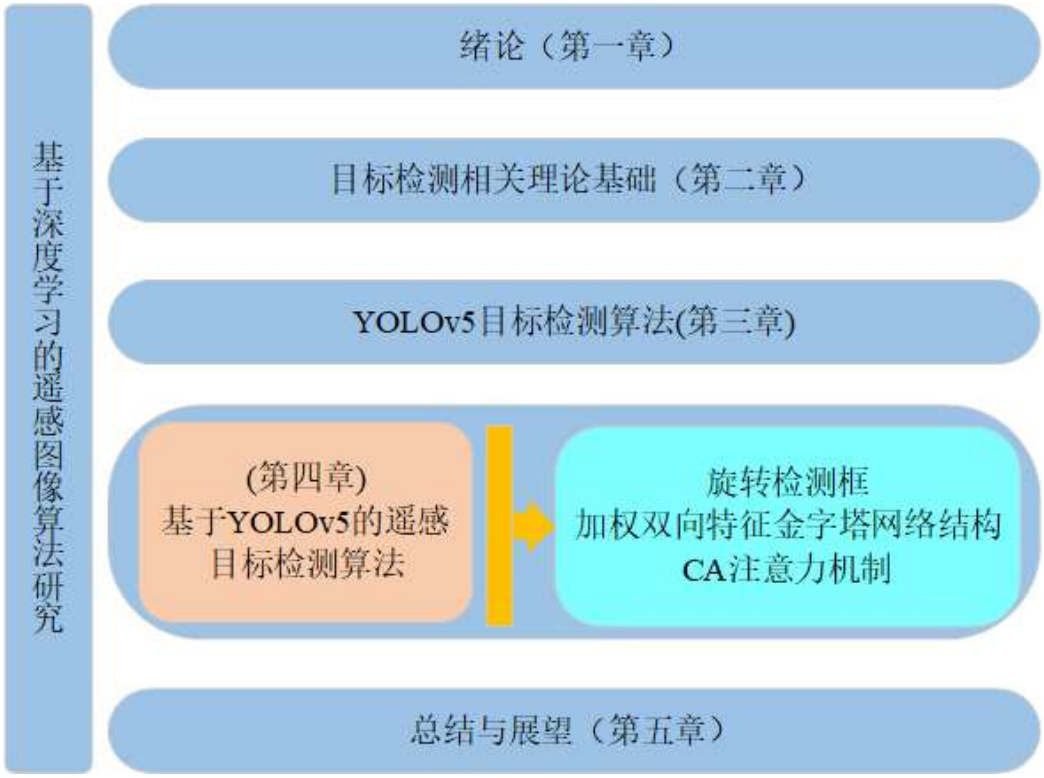


图 1.1 论文整体构架示意图

第二章 目标检测相关理论基础

计算机视觉通常可以分为分类、定位、检测、分割四大基本任务。本章的内容主要分为两个部分：目标检测算法综述和特征融合结构综述。

双阶段算法根据提议的候选区，先对象的检测框，然后根据检测框二次修正得到正确的标签和回归的结果；单阶段算法则是在产生边界框的时候进行分类和回归。本章将首先对人工神经网络和卷积神经网络的基础原理进行简要介绍，其次将依次介绍几种主要的双阶段算法，概述每种算法的改进以及不足之处；然后介绍单阶段算法，简要阐述几种经典的单阶段算法，并着重介绍 YOLOv5 的基本原理和网络结构，最后对本次实验中使用的目标检测评价指标进行分析和介绍，为后续第四章提出对该算法的改进做铺垫。

2.1 人工神经网络

人工神经网络主要分为三层：输入层、隐藏层和输出层。这是对人类大脑神经系统的抽象概括，包括非常多的神经元，彼此之间互相连接。从单个神经元进行分析，可将其视为一个特殊函数，可同时作为输入和输出。海量的神经元通过多级互联来达到信息的传递、分析和处理的功能，同时输出结果会被应用于不同领域。

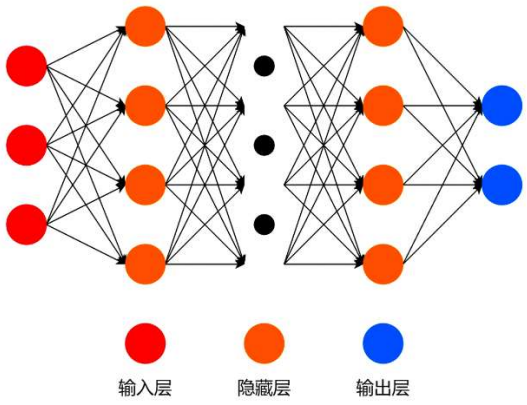


图 2.1 人工神经网络结构图

人工神经元是人工神经网络的基础组成单元，也叫感知器，上个世纪中期才被提出来。模型如下图所示，通常情况下，不同外部输入 $\mathbf{x}_1$、x_2、x_3、…分别与相

应权重w_1、w_2、w_3、…对应，将全部外部输入进行加权求和操作，再与内置偏移量相加得到感知器的输入。

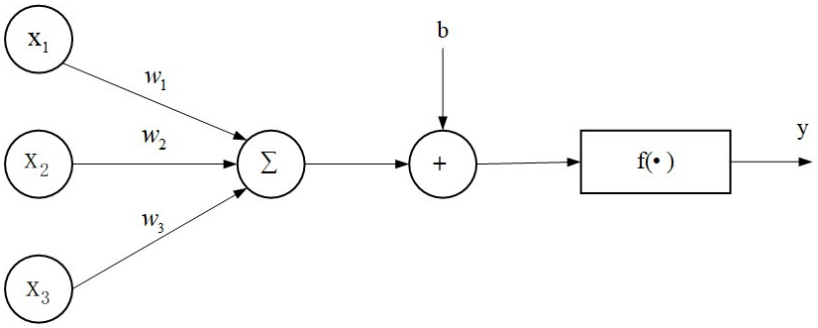


图 2.2 感知模型器

若感知器模型的输入有m个，则整个感知器的输出为公式（2-1）。由大量神经元构成的神经网络将原始的数据进行一层一层的处理，最终生成所需的输出。

$$y = f\left(\sum_{i=1}^{m} x_i w_i + b\right). \tag{2-1}$$

2.2 卷积神经网络

通过多年研究发现，对人工神经网络来说，ANN 中的隐藏层数量越多，其特征表达的性能就越好，获取的信息也更加体现图像的本质。所以，若能构造深层网络，以获取目标的特征图，进而表示图像的高级语义信息，特征鲁棒性就能得到提高。1998 年，卷积神经网络被提出，包含有局部感受野、池化和共享权重三个部分。该网络有两大优势，一是相对而言易于网络优化，二是针对过拟合现象能做到有效减缓。卷积神经网络用于处理图像数据的表现较佳，省去了大量非必需的预处理操作，该算法简单有效，直接输入 CNN，然后进行卷积特征的自动抽取。通常只对二维图像进行处理，提取拓扑、位置、颜色等特征，并且有着较佳的鲁棒性和较高的运算效率。

CNN 经过不同阶段的运算形成三个主要结构：卷积层(Convolution Layer, CONV)、池化层(Pooling)和全连接层(Full Connection, FC)，通采用适当的堆叠来构造一个完整的卷积神经网络，三种结构层通过交替使用构建的卷积神经网络也具有不同功能。

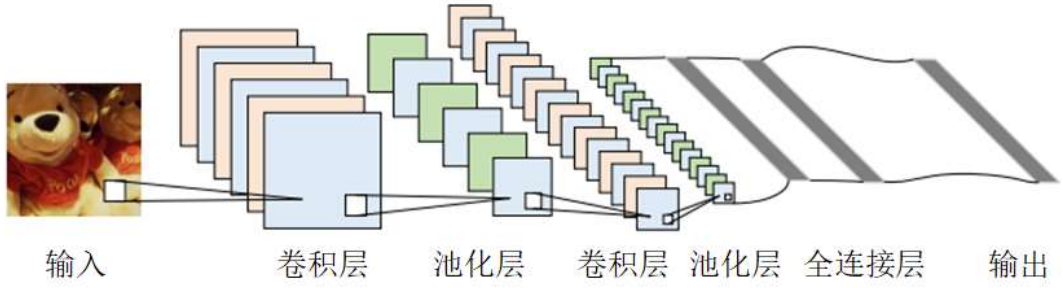


图 2.3 卷积神经网络结构图

2.2.1 卷积层

（1）卷积层：该层利用不同大小的卷积核分步进行局部卷积，得到了相应的特征图。卷积核尺寸（Filter size）指的是卷积核的宽高以及个数（图片通道数）。卷积核每次滑动的距离称为卷积步长（Stride）。

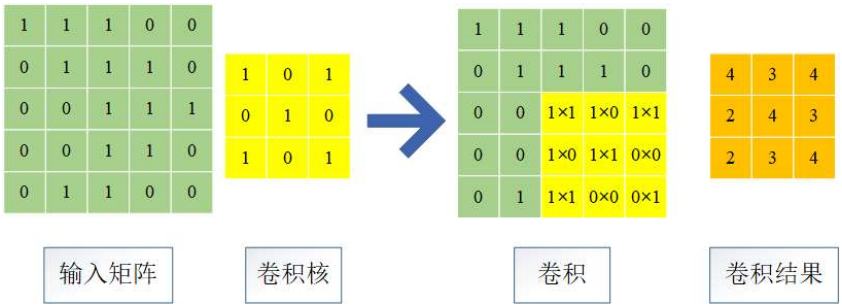


图 2.4 二维卷积操作

在上图的卷积操作中，5×5矩阵作为输入，卷积核是黄色3×3的矩阵，卷积步长的值为 1，最后获得的特征图为3×3的矩阵。如果输入的图像含有色彩信息，那就会有 *RGB* 这 3 个颜色通道，分别对应三个输入矩阵，卷积核个数也为 3。若输入图像通道数为n，也就是输入矩阵的数目是n，第k个输入矩阵是X_k，卷积核的第k个子卷积核矩阵为W_k，b是偏移值，那么可用公式（2-2）表示卷积结果元素值：

$$s(i,j)=\sum_{k=i}^{n}(X_k * W_k)(i,j)+b. \qquad (2-2)$$

由于权值分享的性质，使得在任何地方，特征都会呈现出同样的表现，所以权重向量在卷积核沿着输入矩阵的宽和高滑动的过程中维持不变。权值共享显著减少了卷积层参数的数量，从而达到了节约时间、降低网络复杂性等目标。

2.2.2 池化层

池化层这一重要结构层通常加在卷积层之后，为了减少特征图尺寸，简化参数，减少计算量，减少存储量，对特征图进行下采样。另外，在某种意义上，池化会降低过度拟合的危险。由于特征图特征不变这一特点，通过池化来缩小特征图尺寸，当这一特征图进行下一步卷积时，计算量大幅减少，显著提高计算效率。

常用的池化方法包括最大值池化、平均池化和 L2 范数池化。网络重点进行特征提取，将卷积层最大值池化相连，平均池化通常应用于检测器部分。与卷积层相似，池化层也按固定窗口滑动进行下采样。图 2.5 为最大值池化计算示意图。池化窗口的尺寸为2×2，池化层该位置的输出为特征图像素点最大的值，滑动步长是 2。这种池化方式获得了数值最大的特征，与此同时，也抑制了干扰噪声。

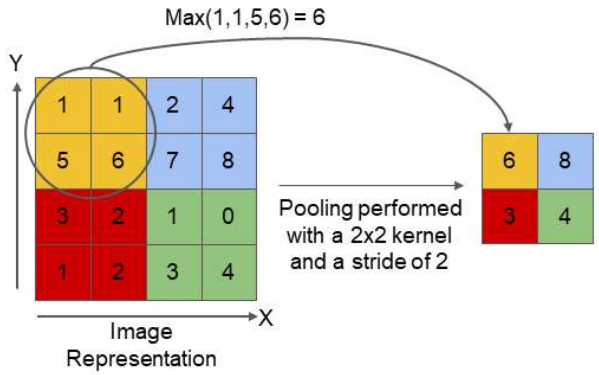


图 2.5 最大值池化计算示意图

2.2.3 全连接层

全连接层：综合在卷积层和池化层提取的特征，有助于进行类别区分的局部信息进行整合。该层顾名思义，上一层全部神经元都分别与该层的每个神经元相连接。如果在该层后再连接数个隐藏层，根据任务类型选取不同的分类器，常见的有 Softmax 层，这样可更深一步获取图像特征，最终再输出得到结果。

但由于所有神经元之间的完全连通，往往会导致过度拟合，故在卷积神经网络中，首先将前两个层次的特征图集进行整体平均池处理，然后将其与全连接层相连，然后在全连接层中引入弃权，从而减少过拟合。

2.3 基于卷积神经网络的目标检测算法

基于深度学习的目标检测算法按照检测方式的差异可以划分为双阶段（Two-Stage）算法和单阶段（One-Stage）算法。两阶段算法根据提议的候选区，先提取目标的检测框，然后根据检测框二次修正得到正确的标签和回归的结果。而单阶段算法则是在产生边界框的时候进行分类和回归。本文将依据时间顺序依次介绍二阶段算法和单阶段算法，并详细介绍二阶段算法中每种算法的优点和需要修改的地方，以及单阶段算法的相关原理和网络结构。单阶段算法是典型的端到端，仅仅只用一个完整的卷积神经网络即可进行实时的目标检测。

两类目标检测算法在精度和速度方面各有优缺点。最近几年，学者们逐渐将双阶段的一些优秀算法思维和模块应用至单阶段检测中，不仅在结构上统一了目标检测算法，使其训练便捷化，还大幅提升检测精度。以下将对上述提及的目标检测算法进行介绍。

2.3.1 双阶段目标检测算法

2.3.1.1 R-CNN 算法

在 AlexNet 问世之后，Ross Girshick 等提出的 R-CNN 是第一个将卷积神经网络用于目标检测的算法，其流程如图 2.6 所示。

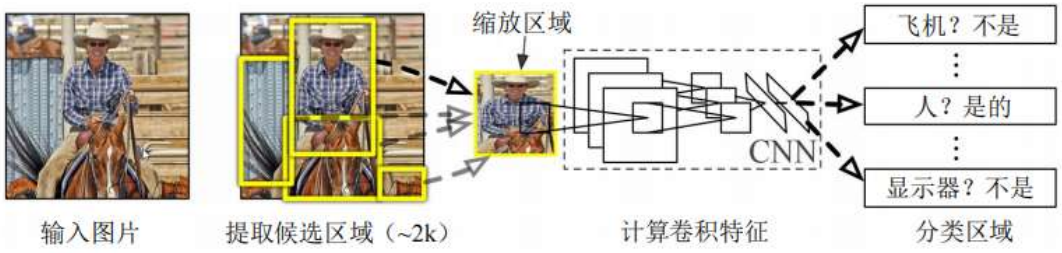


图 2.6 R-CNN 目标检测算法流程图[33]

R-CNN 是一个具有划时代意义的卷积神经网络。首先，该方法利用“选择性搜索”算法，产生多个包含目标的区域，并将其调整到一个统一尺寸；再利用卷积神经网络在这些区域上进行特征提取；之后通过支持向量机（SVM）分类器对其进行分类得到置信度的值；在最后使用边界框回归来做调整以及准确定位。

R-CNN 算法虽然是一个奠基之作，但仍有着一些需要进行改进的地方。比如：模型训练是分成多个步骤并且在多个平台上运行，而不是采用端到端的训练方式。这样的方法使 R-CNN 在进行模型训练和推理的过程中耗费很多的时间和空间，而不适合实际使用。

2.3.1.2 SPP-Net 算法

针对 R-CNN 算法在训练时需要对图像进行缩放处理，使图像以固定的尺寸作为网络的输入，从而导致图像失真这个问题。何凯明等将空间金字塔池化（SPP）层加入到网络结构的全连接层前面，提出了空间金字塔池化网络（SPP-Net）模型。该网络着重解决了 R-CNN 算法中需将候选区域的固定尺寸的图像作为网络的输入的问题，从而有效地克服了图像剪裁或者缩放过程中引起的图像失真。即使图像发生变形，依然能够稳定维持目标检测网络的性能。

但 SPP-Net 仍有一些缺陷：SPP 结构仅用于最后的全连接层之前，并没有改善卷积层和池化层的性能。此外，与 R-CNN 一样，其模型训练过程分成多阶段且在不同平台，占用大量的存储空间。

2.3.1.3 Fast R-CNN 算法

为了解决上述两个算法的不足，Fast R-CNN 算法应运而生，它可以被视为 R-CNN 的快速版本。Fast R-CNN 并没有像前面的算法一样对图像进行分割，而是采用了一种新的方法：即使用整幅图像作为输入来进行处理，并且引入新的概念：感兴趣区域池化层（ROI）。ROI 层通过从不同大小的提议区域中提取出固定大小的特征，再将这些特征用于后面的分类和回归的全连接层的输入。就本质上而言，Fast R-CNN 与 R-CNN 的分别对全部提议区进行卷积操作不同，改为将视野放在了整幅图像上，此外还使用 ROI 池化及映射关系得到固定的特征，因此可以显著的降低运算量。根据 VOC 2007 数据集的目标检测实验表明：和 R-CNN 相比较，Fast R-CNN 的 mAP 高出足足 19.7%，而且计算速率更是快了将近两百倍。尽管这个实验已经证实了 Fast R-CNN 有了很大的提高，而且也实现了由 R-CNN 传统的训练方式到多任务端到端的转变，但是该方法也有一定的缺陷：在实际训练过程中，仍然需要花费大量时间用于选择性算法来生成候选区域，故仍需找到一种更高效的候选框提议算法取代现有的算法。

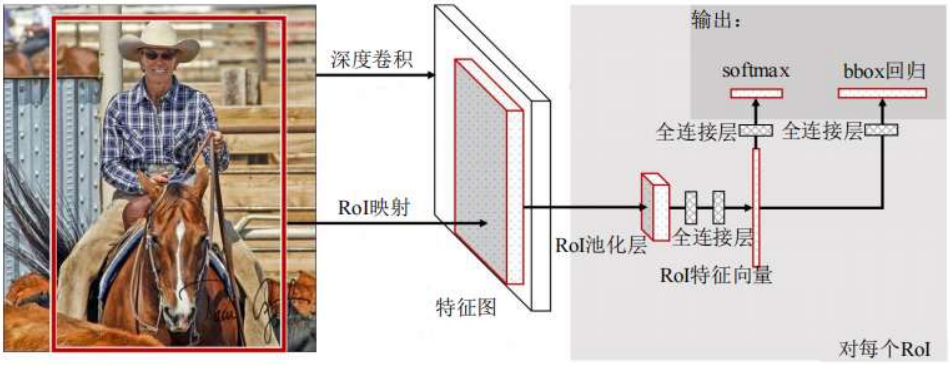


图 2.7 Fast R-CNN 目标检测流程图[33]

2.3.1.4 Faster R-CNN 算法

针对以上 Fast R-CNN 中所遇到的问题，Ren 等人将区域建议网络（RPN）引入到卷积层后方，来替代原先的选择性搜索算法。RPN 是一种全卷积网络，通过对输出结果进行映射处理，使其成为四个坐标值和一个概率值，还将边界框回归损失，二分类损失统一起来作为模型训练中的目标损失函数。通过对 VOC 2007 数据集进行的的目标检测实验表明：在相同 VGG 网络下，Faster R-CNN 的 mAP 提升幅度很少，但大大减少了模型的学习和推理过程时间。由此可以看到 Faster R-CNN 的模型训练速度得到了大幅提升，FasterR-CNN 的学习效率明显提高，但由于算法局限性导致还存在着大量冗余运算，有待进一步的完善。其算法流程图如图 2.8 所示。R-CNN 系列算法性能对比如表 2.1 所示。

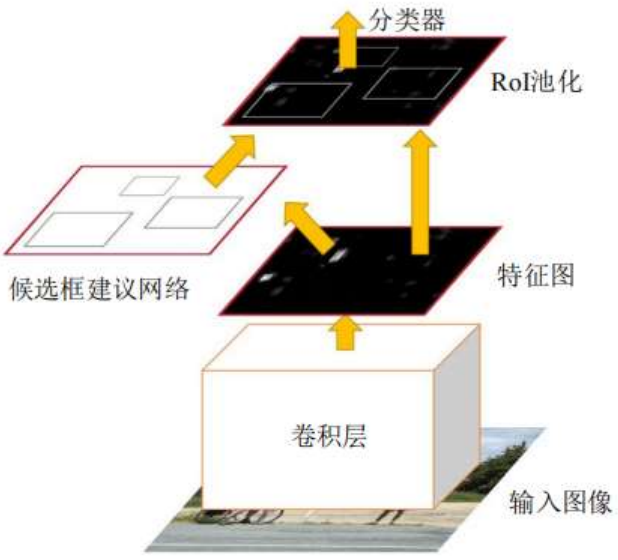


图 2.8 Faster R-CNN 目标检测流程图

表 2.1 R-CNN 系列算法性能对比

模型 / 指标	R-CNN	Fast R-CNN	Faster R-CNN
预测时间	50 秒	2 秒	0.2 秒
加速倍数	1x	25x	250x
mAP(VOC 2007)	66.0%	66.9%	66.9%

2.3.2 单阶段目标检测算法

2.3.2.1 SSD 算法

SSD（Single Shot MultiBox Detector）算法是现在目标检测领域主要的一阶段算法。它的想法是用回归问题来代替检测问题，可以只用一次即可对物体进行定位和分类。SSD 算法的主干特征提取网络使用的是 VGG16，使用卷积层替换了 VGG16 原有的最后的最后两个全连接层，并且在这个基础上还在后方添加了 4 个卷积层。基础神经网络不但可以使用 VGG16 网络结构，还可以使用 ResNet，MobileNet 等更为轻型的网络结构，从而节省训练中的时间和训练量。与其他一阶段算法比较，SSD 目标检测算法在检测小尺寸目标时，能较好实现准确定位。

SSD 算法具体流程如下：先将输入图片加载到预训练好的卷积神经网络中。其次产生默认框（Default box）。提取 6 个不同卷积层中的特征映射，在每个映射上构建 6 个不同尺寸的默认框，再对其进行检测与分类，生成满足需求的默认框。接着对上一步处理所得的若干候选框做非极大值抑制处理，删除重叠或错误候选框，最终确定了只有正确的默认框被保留下来。

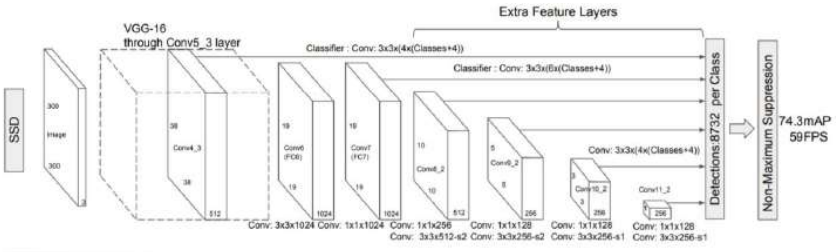


图 2.9 SSD 模型结构

2.3.2.2 YOLOv3 算法

YOLO 系列的算法一直凭借其实时检测性而受到学者的研究和欢迎。YOLOv3 相比于 YOLOv2、YOLOv1，检测速度和检测精度都有提高，尤其是对于小物体检测，故应用于各行各业。改进的 YOLOv3 能在检测速度比 SSD 算法快了将近 3 倍的同时做到精度不下降，YOLOv3 算法主要从以下几点进行改进。

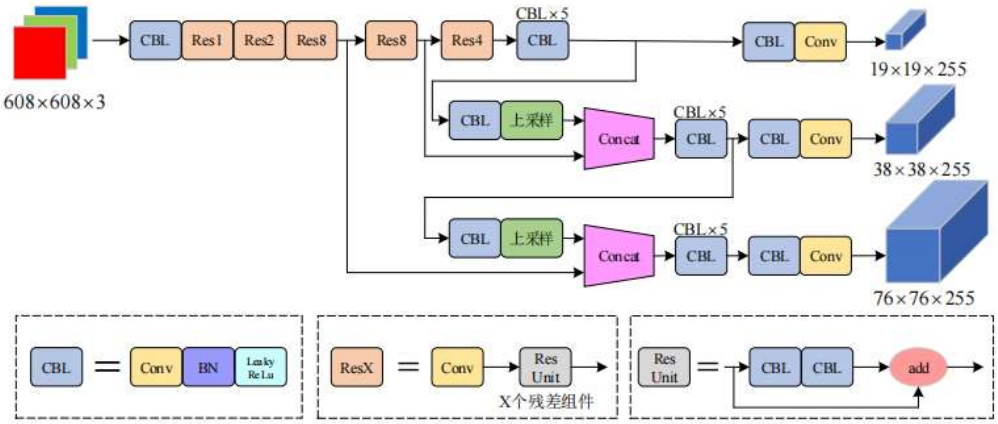


图 2.10 YOLOv3 结构图

YOLOv3 采用了新的网络 DarkNet-53 并加入残差跳跃连接层。同时，网络采用全卷积层，并且下采样通过调节卷积的步长进行实现，以此达到降低池化带来的梯度负面影响的目的。

YOLOv3 采用特征金字塔网络(FPN)进行多尺度特征融合，针对性提升小目标检测效果。它将网络中大小为 416×416×3 的图像处理后输出特征图尺度分别为 13×13×255、26×26×255 和 52×52×255 三条预测支路，并通过多尺度的模式检测各种尺度的目标。从特征获取预测结果后对其解码，对预测出的边界框得分排序与非极大抑制筛选。

2.4 遥感图像目标检测

2.4.1 遥感图像目标检测流程

与自然影像相比，遥感影像的背景复杂，尺度多样，对象密度大，但是其目标检测任务流程有较大相似度。具体检测流程如下：

第一步：选取遥感图像数据集，搭建目标检测模型，用所得数据集中的图像和对应类别标签训练模型以及后续的验证操作；

第二步：模型训练完成后，用其进行测试，得到预测目标的位置、分类和置信度，滤除重复的预测结果后将其进行可视化。

2.4.2 目标检测评判指标

在目标检测中，一般采用精确度、平均精度、类均值平均精度等作为评价标准，交并比是两个框交叉面积与合并面积之比，也就是重叠程度的指标，通常用于评价定位任务的准确性，*IoU* 公式所示如下：

$$IoU = \frac{|A \cap B|}{|A \cup B|} = \frac{Inter\,sec\,tionArea}{UnionArea}. \tag{2-3}$$

表 2.2 表示评判指标混淆矩阵，精确率是指在全部正例样本中，预测无误的部分在判定为正例中所占比例，也叫虚警比例。召回率的含义是预测正确的正例数在所有正样本数中的比例，也称作漏报比例。用上述两个指标作为横纵轴值，将精确率进行积分运算，就能计算出平均精度（公式 2-6）。

表 2.2 评判指标混淆矩阵

真实样本	预测为正例	预测为负例	总计
正样本数	True Positive(TP)	False Negative(FN)	P
负样本数	False Positive(FP)	True Negative (TN)	N
总计	P'	N'	S=P+N

$$Precision = \frac{TP}{P'}, \tag{2-4}$$

$$Recall = \frac{TP}{P}. \tag{2-5}$$

对于 IoU 的阈值取值，Pascal VOC metric 标准中 IoU 阈值为 0.5，表示为 mAP@0.5。而在 COCO 数据集标准下，阈值为 0.5 到 0.95 内以 0.05 为步长所得所有结果的均值。本方案选定 mAP 为主要性能评价指标，由所有检测类别的目标的 AP 求均值计算得出。

$$AP = \int_0^1 p(r)dr \tag{2-6}$$

第三章 YOLOv5 目标检测算法

3.1 引言

近年来 YOLO、 YOLOv2 和 YOLOv3 相继被提出并成为了那个时期主流和强大的目标检测方法。2020 年 6 月，Glenn Jocher 在 YOLOv3 的基础上作出改进，提出了 YOLOv5 模型。YOLOv5 具有快速、高性能且易上手的优点。其性能与 YOLOv4 不相上下，但是模型与 Darknet 比较，只占其 10%，大约为 27M。YOLOv5 网络模型包括 YOLOv5m、YOLv5l、YOLOv5x 和 YOLOv5s，YOLOv5s 网络规模较小，精度较低，速度较快。其余网络都是在 YOLOv5s 的基础上对网络的深度和宽度进行提升，在精度提升的同时也减慢了速度。

YOLOv5 和 YOLOv4 实质上是对 YOLOv3 的网络结构和训练方法的改进，以提高检测性能。YOLO 系列目标检测框架通常可以分为以下四部分：输入端、骨干网络、Neck 网络和输出端。本章将具体分析 YOLOv5 算法的改进之处：（1）输入端增加自适应图片缩放和 Mosaic 数据增强；（2）在 backbone 中应用 CSPDarknet53 结构和 Mish 激活函数等；（3）Neck 中增加了 SPP、FPN+PAN 模块；（4）输出端应用 CIOU_Loss、DIOU_nms。

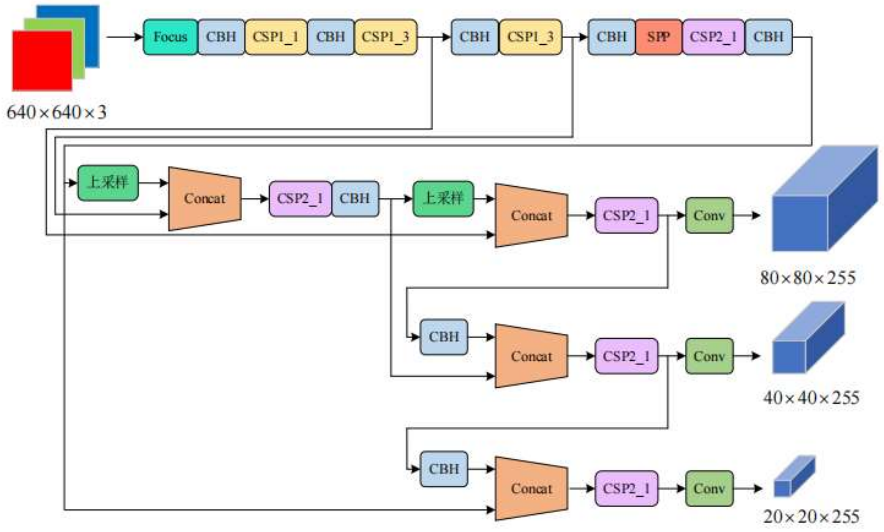


图 3.1 YOLOv5 网络结构

3.2 数据增强

YOLOv5 算法实验部分采用马赛克(Mosaic)数据增强方法，该方法原理上与CutMix[28]方法相近。区别是 Mosaic 方法需要 4 张图片而 CutMix 方法仅需 2 张。

Mosaic 数据强化首先是输入四幅图片，然后对全部图片进行角度调整、自适应缩放、饱和度和亮度调节等，用矩阵将每张图像固定区域截取后对处理完的 4 张图像进行拼接，就形成了新图像。示意图如图 3.2 所示。

图 3.2 Mosaic 数据增强示意图

Mosaic 数据增强的优点是使目标的背景丰富化，因为批量标准化计算会统一对四张图数据进行处理，Mini-batch 可以设得较小，不需要占用多少 GPU 便可取得较佳效果。另外，MS COCO 数据集各种尺寸的目标分布不均，而 Mosaic 方法通过随机调整、缩放和拼接小目标数据量增大，通过丰富数据集的方式使网络更加鲁棒。

3.3 网络结构

3.3.1 Focus 结构

Focus 结构是 YOLOv5 在主干网络的第一层提出的独有的结构，示意图如图 3.3 所示。其中的切片操作尤为关键，对特定尺寸的输入图像进行切片及后续的拼接，即将 RGB 三通道模式改为 12 通道，所得结果经过卷积后得到二倍下采样的特征图。

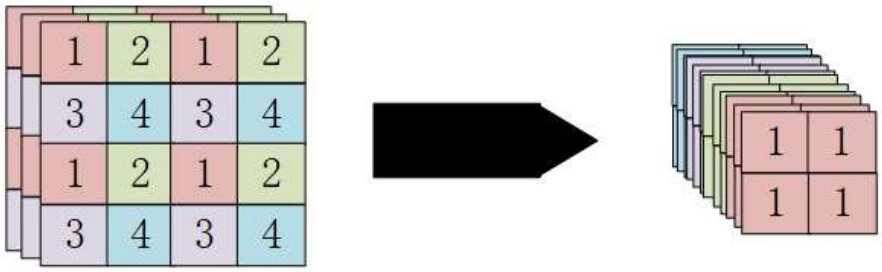


图 3.3 Focus 结构中的切片操作过程

Focus 结构最直观的作用是下采样，但其计算量比普通卷积下采样多 3 倍，该结构将高分辨率的图像信息自空间维度转移到通道维度上再用卷积层进行特征提取，该方法可以有效地降低输入图片大小和输入信息的损失，从而加快网络的训练和推理。

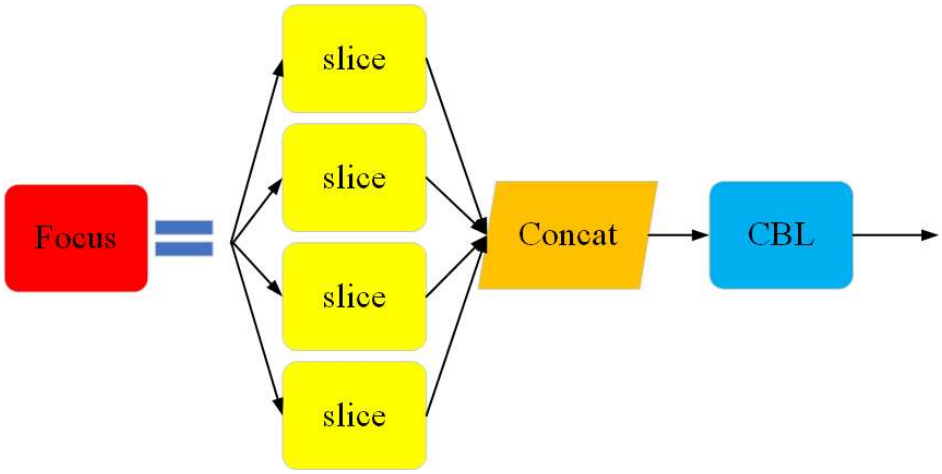


图 3.4 Focus 结构示意图

3.3.2 CSP 结构

YOLOv5 与 YOLOv4 将 CSPNet 结构与 DarkNet53 骨干网络相结合不同，而是以跨阶段局部网络为基础设计了 CSP1_X 与 CSP2_X 两种结构。如图 3.5 和 3.6 所示，两种结构分别应用于骨干网络和 Neck 部分，前者使用带有残差组件的 CSP 结构，后者进行卷积来替换。总的来说 CSPNet 提升了检测性能，增强了 CNN 的学习能力的同时计算量减少，推理速度加快，降低了内存成本并维持了准确率。

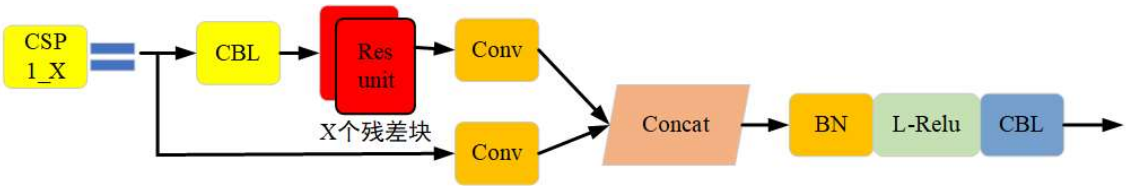


图 3.5 CSP1_X 结构

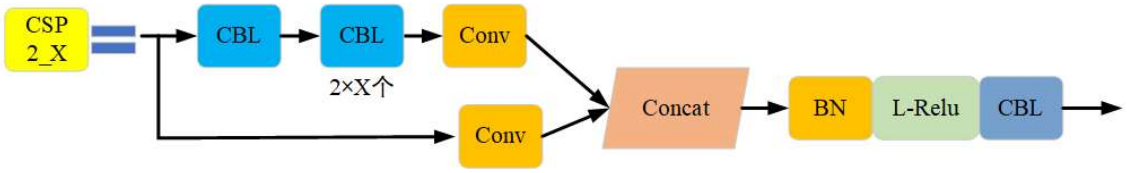


图 3.6 CSP2_X 结构

YOLOv5 在主干网络尾部加入了空间金字塔池化层。SPP 层加入的作用是解决因输入图像尺寸差异性导致重复的变尺度操作。具体做法为：预先将特征图划分为不同尺寸网格（如 1×1、2×2、4×4）后依次进行最大值池化，池化核大小和步长随图像尺寸而进行调整，SPP 使得任何尺寸的特征图都有 1+4+16 维度的输出。SPP 层的优点是解除了输入到主干网络的特征图尺寸限制和有助于分离上下文特征。

YOLOv5 在 Neck 模块中采用特征金字塔网络和路径聚合网络结构。FPN 主要在多尺度特征图上分别进行预测，是当下比较受欢迎的特征融合方案之一。FPN 算法使用横向连接且分为自底向上和自顶向下的两条路线。使用这种结构能使特征图在融合后的具有高语义信息和高几何信息的特点。在此之后，YOLOv5 借鉴又 PANet，加入自底向上的特征金字塔结构，不同的是 YOLOv5 采用的融合特征的方式不是短连接而是拼接。3 个不同尺寸的特征图通过 3 次拼接操作得到，再通过 CSP 结构和卷积得以在最终的三个特征图上分别推理。综上，这种将卷积神经网络特征图融合的方法能对强语义信息和强特征信息进行有效提取。

3.4 损失函数

目标检测任务的损失函数主要是边界框回归损失和分类损失两类，损失函数的优劣性将对检测器的性能与训练速度造成直接影响。当前主流的边框损失函数有：Smooth L1 损失、loU 损失、GIoU 损失、DIoU 损失和 CloU 损失。其中 CIoU

损失函数应用于 YOLOv5 中进行模型训练。本节将对用于模型训练和检测的边界框回归损失和分类损失进行阐述。

对于目标检测器，为了对其性能进行综合评估，常选定交并比（IoU）作为指标，其物理含义是预测框与真实框的重叠程度。

若存在真实框和预测框无任何相交部分的情况，根据上述公式可得 IoU 始终为 0，但是仅凭 IoU 为 0 这一条件不能体现两个框的位置和相互之间的距离，另一种情况是 loU 值相等时，无法正确区分重叠比例相同但重叠部分不同的情况。为解决以上问题，研究者提出 GloU 损失函数[29]，计算方法如公式(3-1)。其中和 U 表示两框最小闭包区域面积和并集面积。

$$GIoU = IoU - \frac{|A_C - U|}{A_C}. \tag{3-1}$$

DIoU 损失在基于 loU 损失的情况下，将实际框与预测框尺寸、两框中心间距、重叠度等因素结合起来，进一步改善了边框回归的稳定性。计算过程为公式(3-2)。

$$DIoU = IoU - \frac{\rho^2\left(b, b^{gt}\right)}{c^2}, \tag{3-2}$$

其中，c 和 ρ 分别代表着预测框和真实框最小闭包区域的对角线距离和欧式距离函数，b、b^{gt} 则是两框的中心点。模型在两框水平或垂直同向的前提下能实现快速回归。DIoU 损失函数达到加快收敛速度的效果的途径是将两个框之间的距离最小化。

研究者在 DIoU 的基础上将边界框的长宽比纳入考量，进而又提出了 CIoU 损失，CIoU 损失计算过程为公式(3-3)和公式(3-4)。

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2, \tag{3-3}$$

$$L_{CIoU} = 1 - IoU + \frac{\rho^2\left(b, b^{st}\right)}{c^2} + \alpha v, \tag{3-4}$$

其中，v 和 α 分别表示长宽比的相似性及权重系数。

第四章 基于 YOLOv5 的遥感目标检测算法

YOLOv5 算法作为当下最优秀的目标检测算法之一，不但具有很高的学术价值，其使用的领域也从军事领域延伸到了人们的生活中。但该检测算法在遥感图像下检测效果不理想，对于复杂背景目标、小目标和密集目标的检测，还有较大的提升空间。故本章基于 YOLOv5 目标检测算法进行改进，首先进行水平检测框至旋转检测框的改进，其次对多尺度特征检测进行研究，基于双向特征金字塔网络的结构优势（BiFPN），将 BiFPN 的思想加入至多尺度特征融合部分。最后研究注意力机制对目标检测的影响，在模型的主干网络部分引入 CA 注意力机制，提升 YOLOv5 算法的表现力。

4.1 模型的改进与优化

4.1.1 旋转检测框

随着目标检测算法被检测物体本身的形状特征发生变化，所采取的边框标注方式也要随之改变。因为原始 YOLOv5 项目在大多数情况下应用于自然场景下的目标，再加上一般情况下的视角为水平视角，故使用水平矩形框（Horizontal Bounding Box, HBB）作为目标检测边框，但是如上文所提及的，俯拍的卫星遥感图像角度不固定且方向多变，物体呈现在二维图像中的形状特征也会随着视角发生变化而改变，研究者们提出了多种边框的标记方法来与图像特征达到更好的匹配效果，例如将椭圆边框标注应用于交通监控（鸟瞰）视角下的目标。恰当的边框标注方式有以下两个优点：（1）通过精确的标注方式降低提供给网络训练时的冗余信息，充分的先验有助于使网络训练更具方向性和目的性，减少了网络的收敛时间；（2）对于密集的目标而言，精准的标注方式防止已经检出的目标被 NMS 清除。

对原始的遥感数据集进行剪裁后，得到分割后的图像数据集和 YOLO 格式注释文件，DOTA 数据集中标注方式采用的是任意四边形四点坐标，本文选取的旋转矩形框的表示方法为长边表示法。其形式为表 4.1。

表 4.1 长边表示法形式

Classid	x_c	y_c	longside	shortside	θ	$\theta \in [0,180)$

其中x_c与y_c分别表示旋转矩形框中心的横纵坐标；*longside*和*shortside*分别表示矩形框的长边和短边；θ为x轴顺时针旋转遇到最长边所经过的角度。

4.1.2 主干网络的改进

大多数现有的注意力机制通常采用最大池化或平均池化方法来处理通道，会对目标的空间信息造成损失。模型容量受限的轻量级网络往往难以负担注意力机制造成的计算开销，使得注意力的应用十分滞后。此外，由于遥感影像中的小物体所占据的像素数量较少，而且容易受到周围环境的干扰，YOLOv5 在进行卷积取样时容易失去其特征。故本文加入坐标注意力机制，有效地对小目标和密集目标进行特征提取，实现对检测准确率的提升。

坐标注意力机制(Coordinate Attention, CA)[30]在通道注意力中加入位置信息，达到减少计算量和以及使网络得以在更大区域上注意的目的。CA 机制将通道注意拆分为两个一维的特征编码，分别在两个方向上对特征进行聚合来得到精确的位置信息和远程依赖关系。再编码所得特征图得到一对方向感知和位置敏感的特征。通过以上处理，可以对 SENet[31]、CBAM[32]等机制在池化操作中造成的遗失位置信息进行有效改善。

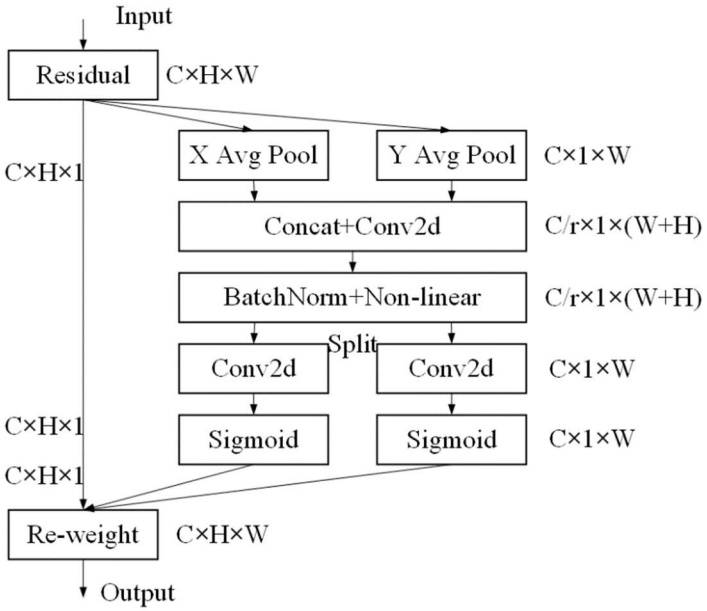


图 4.1 CA 模块结构

如图 4.1 所示，CA 机制凭借位置信息对长程依赖和通道关系进行编码，主要分为坐标信息嵌入和坐标注意力生成两步。CA 模块可被具象为对特征表示能力进行强化的计算单元，输入中间张量 $X=[x_1,x_2,...x_c]\in R^{C\times H\times W}$，输出 $Y=[y_1,y_2,...y_c]$ 与中间张量 X 具有相同尺寸，且有提高特征表达能力的作用。其中 C 、H 、W 分别表示通道数以及输入图片的高和宽。

Coordinate Attention 作为新颖简单且即插即用的模块，能在不产生额外计算量和开销的基础上对网络的精度进行提高。本文将 CA 模块添加至 Backbone 中，将原有的十层特征提取网络改为十三层的结构，提升网络模型对小目标的检测效果。

4.1.3 特征金字塔结构的改进

BiFPN 加权双向特征金字塔网络结构由 Google 首次提出，是一种全新的特征融合方式。BiFPN 是基于路径增强高效的双向跨尺度连接和加权特征融合的思路。依次进行自顶向下和自底向上的特征融合。

FPN、PANet 和 BiFPN 结构如图 4.2 所示，FPN 通过创建一条自上而下的通路来实现特征融合，但是所得的有着更高语义信息的特征层进行预测会受阻于单向信息流。Shu Liu 等人为了对上述问题进行优化，在 FPN 的基础上加入一条由下而上的通路，提出 PAN 结构，使得预测特征层既有底层的位置信息，又有上层的语义信息，从而极大地提高了目标的识别准确率。

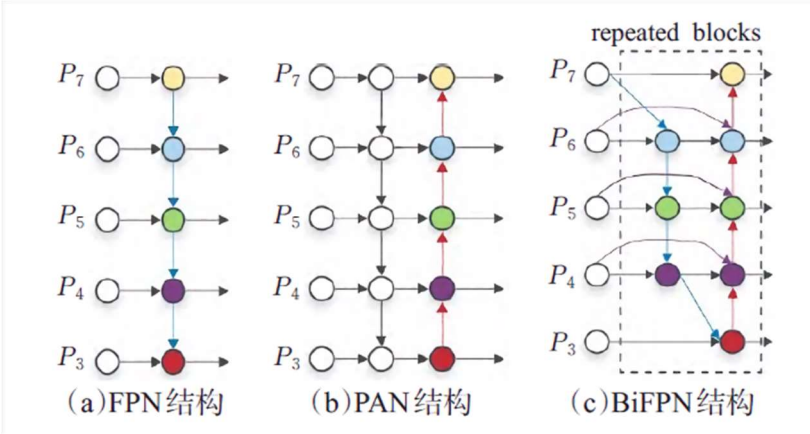


图 4.2 FPN、PANet 和 BiFPN 结构

BiFPN 以 PAN 为基础又做改进，双向跨尺度连接先对贡献度小的单输入节点进行删除，以此简化网络且几乎不造成影响，接着在输入输出节点间增设一条边来

融合更多的特征；最后，将两条路径（向上和向下）融合至同一模块中便于反复堆积，以达到高级特征融合。此外，BiFPN 将权值与所有权值相比来缩至$[0,1]$之间，这种快速归一化融合的模式提高了计算速度。公式为式（4-1）。通过激活函数 ReLU 来确保权重$w_i \geq 0$，I_i 表示输入的特征。

$$O = \sum_i \frac{wi}{\varepsilon + \sum_j w_j} \cdot I_i . \tag{4-1}$$

标量权重的无界性会引起训练过程的不稳定，故通过 softmax 来归一化。BiFPN 的输入为主干网络中提取出的三种不同尺度的特征 P_3、P_4、P_7，在加权特征融合和跨尺度连接后，最终确定了 20×20、40×40 和 80×80 三个不同大小的预测分支。

基于上述优点，本文用 BiFPN 模块取代 YOLOv5 结构中的 PAN 模块，以此优化特征融合，提升检测的精度和速度。

4.2 实验结果与分析

4.2.1 数据集介绍

为了对改进后的模型进行检测，本章使用 DOTA-v1.0 数据集对模型进行训练，并上传至服务器进行测试。DOTA-v1.0 是武大遥感国重实验室夏桂松和华科电信学院白翔等合作完成的航拍图像数据集，目前最大的光学遥感图像数据集 (图像来源于 GoogleEarth 和两颗中国的卫星 GF-2，JL-1)。该数据集共包含 2806 张遥感图像（图片尺寸从 800*800 到 4000*4000），188,282 个实例中涵盖了 15 个类别：飞机、船只、储蓄罐、棒球内场、网球场、篮球场、田径场、海港、桥、大型车辆、小型车辆、直升机、英式足球场、环形路线、游泳池。每个实例都由一个四边形边界框标注，顶点按顺时针顺序排列。官方使用 1/2 的图像作为训练集，1/6 的图像为验证集，1/3 的图像为测试集。

4.2.2 实验环境

表 4.2 实验环境配置表

操作系统	Linux
内存	48GB
CPU	Intel(R) Xeon(R) Gold 6330
GPU	RTX 3090 * 2
CUDA	11.3
Python	3.8
PyTorch	1.10.0

4.2.3 数据预处理

DOTA 数据集需要对数据进行格式转化方能用于模型训练，首先是 DOTA 水平边界框的转换，格式如表 4.3 所示。具体步骤为逐行读取 labels 文本文件，保留边界框位置与类别信息，对水平边界框四点坐标计算求得中心坐标和宽高，进行归一化后再逐行写入另一个文本文件。

表 4.3 DOTA 水平边界框格式转换

DOTA format	poly		class name		difficult
			To		
YOLO format	class name id	x	y	width	hight

第二个步骤是将所得标签格式转为 YOLO 长边表示法标签格式，得到新的注释文件。鉴于遥感图像检测有着图像尺寸大、背景复杂、小目标数量大等特征，而尺寸过大的图像易导致常规目标检测网络显存超载，故选取 1024 为分辨率阈值，对 train、val、test 文件夹中的图像进行剪裁，对于剪裁后的小图片，若其中没有 15 类检测目标之一，对其进行去除，达到减少冗余计算量，提高训练速度的效果。

4.2.4 实验结果

4.2.4.1 模型检测结果对比

对 YOLOv5、YOLOv5_CB 两个模型分别进行从头训练和检测，对所得的进行了剪裁的遥感图像标签进行合并，由于 DOTAv1.0 测试集未提供真实标签，因此需要将测试结果上传至 DOTA 官网上的评估服务来进行模型性能评估。本文选用 mAP@0.5、mAP@0.5:0.95 两个指标进行评估，mAP@0.5 指的是 IoU 阈值为 0.5 时的 mAP 值，mAP@0.5:0.95 则使用.50:.05:.95 之间的 10 个 IoU 阈值，对 loU 进行平均可以使检测器更好地定位。

表 4.4 YOLOv5 及改进算法的模型检测结果对比

Class	YOLOv5	YOLOv5_CB
Plane(AP@0.5)	0.895	0.895
baseball-diamond(AP@0.5)	0.662	0.763
bridge(AP@0.5)	0.449	0.440
ground-track-field(AP@0.5)	0.567	0.561
small-vehicle(AP@0.5)	0.754	0.748
large-vehicle(AP@0.5)	0.783	0.781
ship(AP@0.5)	0.880	0.880
tennis-court(AP@0.5)	0.908	0.908
basketball-court(AP@0.5)	0.627	0.798
storage-tank(AP@0.5)	0.803	0.803
soccer-ball-field(AP@0.5)	0.401	0.453
roundabout(AP@0.5)	0.563	0.555
harbor(AP@0.5)	0.605	0.687
swimming-pool(AP@0.5)	0.750	0.748
helicopter(AP@0.5)	0.527	0.592
all(mAP@0.5)	0.678	0.708
all(mAP@0.5:0.95)	0.403	0.414

4.2.4.2 消融实验

消融实验结果如表 4.5 所示，根据表中数据可得，加入 CA 注意力机制后，mAP@0.5 提高了 1.3%，mAP@0.5:0.95 提高了 0.9%，修改金字塔结构 mAP@0.5 提高了 1.7%，mAP@0.5:0.95 提高了 0.6%。将这两个改进共同加入模型之中，mAP@0.5 提高了 3.0%，mAP@0.5:0.95 提高了 1.1%，通过逐个对改进模块进行添加，验证单一模块的优化效果。优化后的 YOLOv5_CB 模型对密集目标和小目标的检测有较大提升。

表 4.5 消融实验结果

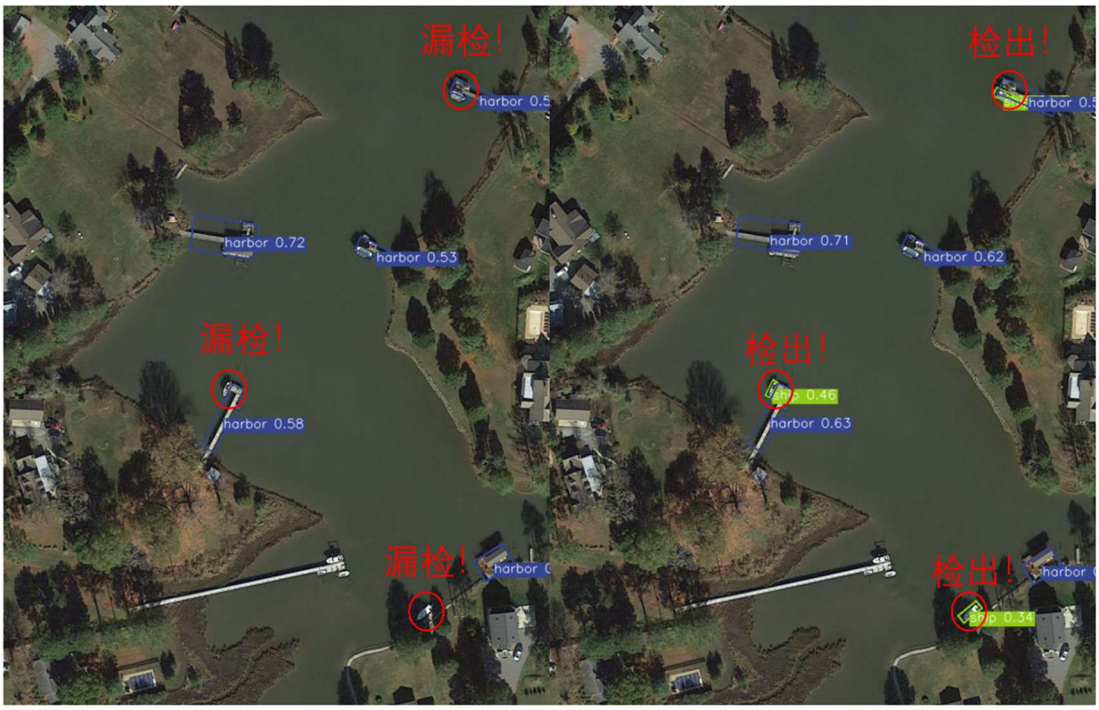
CA	BiFPN	mAP@0.5	mAP@0.5:.95
×	×	0.678	0.403
√	×	0.691	0.412
×	√	0.695	0.409
√	√	0.708	0.414

4.2.4.3 检测结果分析

为了进一步检验 YOLOv5_CB 模型的可信度，随机选取测试集图片进行检测。图 4.3 至图 4.7 为基础模型和改进模型的检测结果对比图，图 4.3 和图 4.4 是对密集目标的检测，左图和右图分别为 YOLOv5 模型和改进模型的检测可视化结果，可见左图对密集排布的船只出现了漏检，而右图能进行正常检测。图 4.5 是对遮挡目标的检测，与左图的漏检不同，右图则能检测出被遮挡的小车辆目标。图 4.6 和图 4.7 上图是对小车辆目标的错检和漏检，而下图则准确的检测到每个目标。

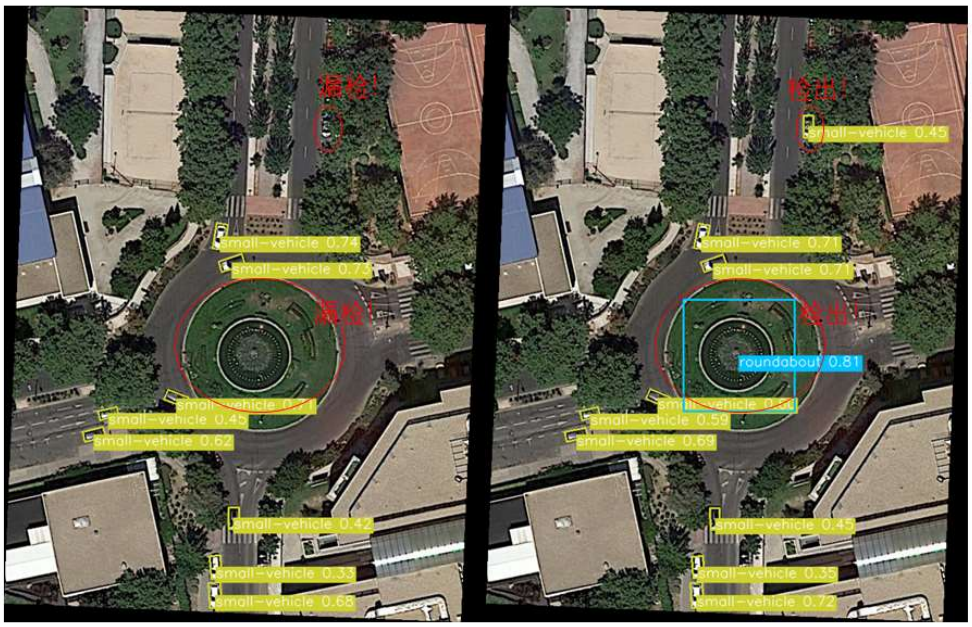
(a)YOLOv5 (b)YOLOv5_CB

图 4.3 密集目标的检测结果



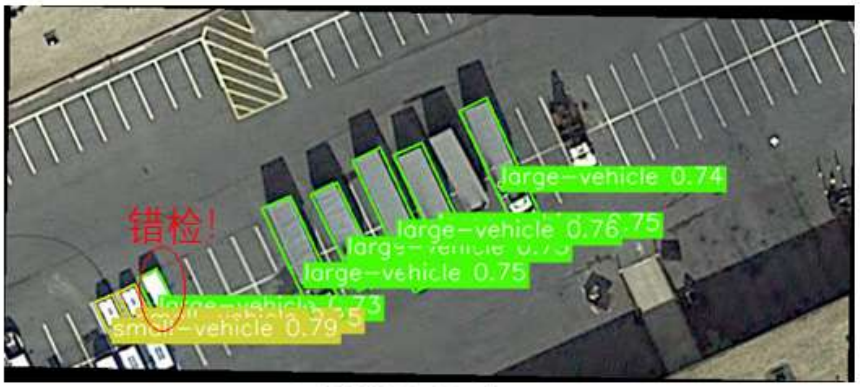
(a)YOLOv5 (b)YOLOv5_CB

图 4.4 密集目标的检测结果

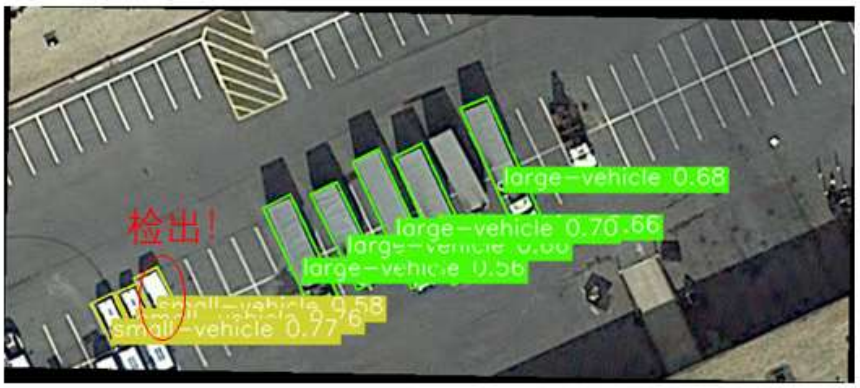


(a)YOLOv5 (b)YOLOv5_CB

图 4.5 被遮挡目标的检测结果



(a)YOLOv5



(b)YOLOv5_CB

图 4.6 错检的检测结果

(a)YOLOv5

(b)YOLOv5_CB

图 4.7 漏检的检测结果

4.3 本章小结

此章先对实验环境、数据集以及性能评估指标等方面给出了相应的描述，接着对改进的算法进行实验验证和检测结果可视化，并做了消融实验作为参照。改进的 YOLOv5_CB 模型在 DOTA 数据集上 mAP@0.5 和 mAP@0.5:0.95 分别提升了 3.0% 和 1.1%。

综上所述，YOLOv5_CB 模型相比于 YOLOv5 模型有着更优越的性能。YOLOv5 模型在复杂和大尺度的场景下检测效果不佳，尤其是对被遮挡目标、小目标及密集目标易发生漏检或误检的情况。YOLOv5_CB 模型对被遮挡目标、小目标和密集目标的检测效果要优于 YOLOv5 模型，具有更好的鲁棒性，表现出更优越的性能和更准确的定位精度。

第五章 总结与展望

5.1 论文工作总结

本文在对 YOLOv5 目标检测算法的检测原理深入分析后，针对小物体、密集物体和被遮挡物体检测漏检误检等问题，在 YOLOv5 模型的基础上提出了改进算法 YOLOv5_CB 模型。该改进主要是通过将边框标注方式由水平改为旋转检测框、修改特征金字塔结构、加入 CA 注意力机制来对模型进行优化。通过实验以及测试结果证明了本文的改进方案具有可行性，对于小物体检测和复杂背景检测性能都有提高，具有很好的泛化能力。

本文的主要工作如下：

（1）将 YOLOv5 的水平检测框改为旋转检测框，通过精确的标注方式降低提供给网络训练时的冗余信息，充分的先验有助于使网络训练更具方向性和目的性，减少了网络的收敛时间；另外对于密集的目标而言，精准的标注方式防止已经检出的目标被 NMS 清除。

（2）对 YOLOv5 中的特征金字塔结构作出了改进，通过 BiFPN 加权双向特征金字塔网络结构代替 FPN。

（3）对注意力机制的作用和原理作详细的分析，在 YOLOv5 模型的 backbone 模块引入 CA 注意力机制。检测器对小目标和复杂背景目标的检测精度均有提升。

5.2 论文工作展望

本文对 YOLOv5 目标检测算法的改进取得了一定的效果，但受本人现阶段知识水平和能力限制，算法的改进仍存在一定不足之处，有大量进一步优化的空间。一方面可以从理论上进行深入的分析，另一方面也可以通过大量的实验来积累经验。综合全文，未来的研究还可以从两点着手：

（1）就网络本身而言，找寻更佳的结构。未来可将注意力放至对自身结构的优化，比如卷积层使用膨胀卷积或空洞卷积及对卷积神经网络的参数计算量进行优化等。

（2）随着实例工程的应用，模型需要更好地适配硬件，由于嵌入式端装备和移动端的大量需求，模型的简化对其的应用和部署至关重要，需进一步考虑如何在

保证模型良好的检测性能的同时尽可能降低算法复杂度，未来研究可着重该方向，得到更适合实际应用的新型网络。

致谢

时光荏苒，转眼间 4 年本科生活已然接近尾声。回望过去四年的点点滴滴，这段校园时光里充满了温暖和快乐，我也在老师们和同学们的帮助下，知识水平得到加强，能力得到提升，视野也随之开阔。在完成本篇论文之际，我谨向所有关心、鼓励和帮助过我的人表示衷心的谢意。

感谢指导老师杨曦教授，感谢她对本篇论文从选题到算法研究和改进方面的悉心指导。她认真负责，对待学术一丝不苟的专业态度值得我学习。

感谢张鑫学长，认真负责，待人和善，在毕设方面给了我很多知识上的铺垫和文献上的导引，让我少走了很多弯路。

感谢我的朋友们和我的舍友们，与他们共同学习的四年生活是我一生的财富，感谢大家的陪伴与帮助，他们各方面的闪光点都是我为之赞叹和敬佩，使我学会了自律，学会了谦虚与踏实，受益良多。

最后，感谢父母和其他亲人对我一直以来无条件地鼓励支持和默默付出，成就了我今日的成绩，祝愿身边的每一位尊长平安顺遂。

参考文献

[1] 王素敏,李向英,何劲.资源三号卫星影像测绘性能分析[J].影像技术,2013,25(03):48-49+45.

[2] 曹连雨. 基于深度卷积神经网络的遥感影像目标检测技术研究及应用[D].北京科技大学,2022.DOI:10.26945/d.cnki.gbjku.2022.000125.

[3] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks [J]. Advances in Neural Information Processing Systems, 2015, 28.

[4] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25.

[5] Sermanet P, Eigen D, Zhang X, et al. Overfeat: Integrated recognition, localization and detection using convolutional networks [J]. arXiv preprint arXiv:1312.6229, 2013.

[6] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1-9.

[7] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[8] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[9] Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model[C]//2008 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, 2008: 1-8.

[10] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 580-587.

[11] Girshick R. Fast R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1440-1448.

[12] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[13] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection [C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 2980-2988.

[14] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 779-788.

[15] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector[C]//European Conference on Computer Vision. Springer, Cham, 2016: 21-37.

[16] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2117-2125.

[17] Redmon J, Farhadi A. YOLO9000: Better, faster, stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 7263-7271.

[18] Redmon J, Farhadi A. YOLOv3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[19] Tan M, Le Q. EfficientNet: Rethinking model scaling for convolutional neural networks[C]//International Conference on Machine Learning. PMLR, 2019: 6105-6114.

[20] Tan M, Pang R, Le Q V. EfficientDet: Scalable and efficient object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2020: 10781-10790.

[21] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[22] Zhu H, Chen X, Dai W, et al. Orientation robust object detection in aerial images using deep convolutional neural network[C]//2015 IEEE International Conference on Image Processing (ICIP). IEEE, 2015: 3735-3739.

[23] Cheng G, Zhou P, Han J. 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.

[24] Van Etten A. You only look twice: Rapid multi-scale object detection in satellite imagery [J]. arXiv preprint arXiv:1805.09512, 2018.

[25] Zhang G, Lu S, 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.

[26] Xia G S, Bai X, Ding J, et al. DOTA: A large-scale dataset for object detection in aerial images [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3974-3983.

[27] Li K, Wan G, Cheng G, et al. Object detection in optical remote sensing images: A survey and a new benchmark[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 159: 296-307.

[28] Yun S, Han D, Oh S J, et al. CutMix: Regularization strategy to train strong classifiers with localizable features [C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 6023-6032.

[29] Rahman M A, Wang Y. Optimizing intersection-over-union in deep neural networks for image segmentation [C]//International Symposium on Visual Computing. Springer, Cham, 2016: 234-244.

[30] Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2021: 13713-13722.

[31] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7132-7141.

[32] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 3-19.

[33] 赵婉月. 基于 YOLOv5 的目标检测算法研究[D]. 西安电子科技大学, 2021. DOI:10.27389/d.cnki.gxadu.2021.002918.

附录