分类号 _____ 密级 _____

UDC[注 1] _____

南京理工大学

NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY

硕 士 学 位 论 文

基于深度学习的视觉探测低慢小无人机技术研究

（题名和副题名）

余振滔

（作者姓名）

指导教师姓名 苏岩 教授

学 位 类 别 工学硕士

学 科 名 称 微系统与测控技术

研 究 方 向 模式识别与目标检测

论文提交日期 **2020 年 12 月**

注 1: 注明《国际十进分类法 UDC》的类号

基于深度学习的视觉探测低慢小无人机技术研究

南京理工大学

硕 士 学 位 论 文

基于深度学习的视觉探测低慢小无人机技术研究

作　　者：余振滔

指导教师：苏岩 教授

南京理工大学

2020 年 **12** 月

Ph.M. Dissertation

Research on Visual Detection of Low & Slow & Small Drones Based on Deep Learning

By

Zhentao Yu

Supervised by Prof. ***Yan Su***

Nanjing University of Science & Technology

December, 2020

声　明

本学位论文是我在导师的指导下取得的研究成果, 尽我所知, 在本学位论文中, 除了加以标注和致谢的部分外, 不包含其他人已经发表或公布过的研究成果, 也不包含我为获得任何教育机构的学位或学历而使用过的材料。与我一同工作的同事对本学位论文做出的贡献均已在论文中作了明确的说明。

研究生签名：_____　　年　　月　　日

学位论文使用授权说明

南京理工大学有权保存本学位论文的电子和纸质文档, 可以借阅或上网公布本学位论文的部分或全部内容, 可以向有关部门或机构送交并授权其保存、借阅或上网公布本学位论文的部分或全部内容。对于保密论文, 按保密的有关规定和程序处理。

研究生签名：_____　　年　　月　　日

摘　要

随着低慢小无人机广泛应用到各个领域，其造成的安全隐患问题也越来越突出，因此有效监控低慢小无人机成为当前一个亟需解决的问题。本文面向这一背景，开展基于深度学习的视觉探测低慢小无人机算法研究，具体成果包含以下几个方面：

(1) 分别制作了低慢小无人机可见光和红外数据集。收集了不同场景下，包含无人机，风筝和飞鸟三类物体的可见光图像 15659 张，红外图像 5546 张，再通过标注和划分，构建了本文的实验数据集，且指定了相关算法的评价指标。

(2) 分别设计并实现了面向可见光和红外图像的低慢小无人机目标检测模型。进行了二阶检测模型 Faster R-CNN 与一阶检测模型 SSD 基准实验，确立了检测的基础模型为 Faster R-CNN。针对可见光数据集多尺度和复杂背景问题，对应引入特征金字塔、可变性卷积和自注意力机制、GIoU 回归损失函数、在线困难样本挖掘技术进行改进；针对红外数据集存在 (极) 小目标的难点，除了使用特征金字塔和在线困难样本挖掘技术，还替换特征提取网络为 HRNet，保持特征图的高分辨率。最后的实验证明了两个改进模型的有效性，可以满足全天候监测低慢小无人机的需求。

(3) 提出了一种低慢小无人机弱监督语义分割系统。采用 GrabCut 算法和标注框生成无人机的“伪像素标签”，针对低慢小无人机在图像中的多尺度变化，正负像素样本不均衡的问题，引入语义分割模型 S-Net，并利用焦点损失函数，Dice 损失函数，图像混合方法进行训练和全连接条件随机场进行后处理，通过在可见光数据集上的实验证明了整个算法的可行性，加深了无人机识别定位效果。

关键词：低慢小无人机，深度学习，目标检测，语义分割，小目标，多尺度，复杂背景

Abstract

With the wide application of low & slow & small drones in various fields, the potential safety problem caused by them have become more and more prominent. Therefore, effective monitoring of low & slow & small drones has become an urgent matter to be solved. Under the circumstances, this dissertation conducts research on the algorithm of visual detection of low & slow & small drones based on deep learning. The specific contributions include the following aspects:

(1) Construct the visible light optical and infrared thermal datasets of low & slow & small drones respectively. Collect 15,659 visible light optical images and 5546 infrared thermal images of three classes of objects including drones, kites, and birds in different scenarios. Then, build the experimental datasets after annotation and partition. And offer the explicit evaluation metrics of related algorithms.

(2) Design and implement low & slow & small drones object detection models for visible light optical and infrared thermal images respectively. Determine basic detection algorithm to be Faster R-CNN after carried out benchmark experiments about two-stage model Faster R-CNN and one-stage model SSD. Aiming at multiscale and complex background difficulties of visible light optical datasets, correspondingly introduce feature pyramid network、deformable convolution and self-attention mechanisms、GIoU regression loss、online hard example mining techniques to improve Faster R-CNN. For infrared datasets that have the problem of (very) tiny targets, in addition to using feature pyramid network and online hard example mining, we also replace the feature extraction network with HRNet to maintain the high resolution of the feature maps. Experiments show that the two modified models are effective and can meet the demands of all-weather monitoring low & slow & small drones.

(3) Propose a weakly supervised semantic segmentation system for low & slow & small drones. Using GrabCut algorithm and labeled bounding box to generate the "pseudo pixel labels" of the drones. And next, in order to solve the problems of drones' scale changes in the image and the imbalance of positive and negative pixels, we introduce the semantic segmentation model S-Net and adopt focal loss, Dice loss, Mixup for training and fully connected conditional random fields for post-processing. Experiments on the visible light optical dataset show that the whole algorithm system is feasible, which can deepen the recognition and localization results of low & slow & small drones.

Key word: Low & Slow & Small Drone, Deep Learning, Object Detection, Semantic Segmentation, Small Target, Multiscale, Complex Background

目录

图表目录

1 绪论

1.1 选题的背景和意义

现阶段，全球民用无人机行业爆发式增长，与此同时，无人机“黑飞”给社会公共安全造成了巨大的威胁[1]。进入 21 世纪以来，智能芯片算力的提升以及制造成本的降低，无人机已经成为现阶段先进智能设备的代表，应用场景和使用需求不断扩大[2]。其中，以多旋翼为代表的小型无人机由于获取便利，操纵简便等特点，已经广泛应用于航空摄影，车流监管，遥感测绘，基础设施隐患防查，森林防护，气象勘探以及个人娱乐等领域[3]。根据中国民航科学技术研究院航空器适航研究所公布的数据，截至今年 5 月 26 日，我国无人机生产企业 1353 家，无人机登记数量 330034 架，注册用户数量为 310218 个[4]。然而，无人机管控系统以及相关追责法规的滞后，导致很多小型无人机带来的安全威胁和和经济损失无法得到及时消除和追究。此外，除了广大无人机消费者的“黑飞”，“禁飞”等情况，还可能出现不法分子利用小型无人机实施危害社会公共安全，甚至是国家安全等行为，因此反无人机技术和系统的需求愈发强烈。

低慢小无人机是对低空飞行、飞行速度慢、不易被侦测发现等特征的小型无人航空器目标的统称[5]，目前普遍以低空或超低空飞行 (飞行高度在 1km 以下)、移动速度较慢 (速度在 200km/h 以下)、体型较小 (雷达散射截面在 $2m^2$ 以下) 的标准来界定此类无人机[6]。此类无人机是目前消费级和部分军用级的主要使用产品，其威胁方式如图 1.1 所示，主要有以下几种[3]。

图 1.1 低慢小无人机带来的多种安全威胁

(1) 机场，高铁附近管制领域“黑飞”，“禁飞”低慢小无人机，造成旅客航班，车次延误；

(2) 广场，演唱会，火车站等集会区域或者其他人员密集区域，低慢小无人机出现故障或操作失误造成人员伤亡，或者被恐怖分子利用实施恶意袭击，制造恐慌等；

(3) 监狱外部人员通过低慢小无人机这类不易察觉的设备向罪犯提供枪支弹药，情报信息等，实施违法活动；

(4) 军事上，敌方利用低慢小无人机进行阵地侦察，军情刺探，甚至是作为武器实施打击等。

因此，针对低慢小无人机的监测，打击与反制技术已经成为各个领域备受关注的研究课题。

1.2 低慢小无人机探测方案研究现状

低慢小无人机的管控主要分为两个功能部分，一个是侦察探测，综合利用多种传感器和目标无人机的外观、物理属性，实现对其的识别和定位；另一个是反制，利用信号干扰，捕获接管，甚至采用摧毁的手段来消除目标无人机的威胁[7]。就低慢小无人机探测技术来说，典型的主要有雷达探测，无线电探测，声学探测和光电探测等[3]。

雷达通过发射电磁波，利用多普勒技术获取目标无人机的距离、尺寸、速度等信息，技术成熟度较高，但是识别性能差，存在鸟类这一相似对象的干扰，受背景杂波影响大，无法探测悬停的无人机，此外低慢小目标的制作材料中金属含量低，雷达散射截面积小，也不易被雷达设备发现与识别[8–12]。无线电探测是通过对低慢小无人机与外界的飞控通信和图像传输等信号的频谱、功率谱特征进行分析，根据到达时间法 (Time of Arrival, TOA)，接受信号强度法 (Receive Signal Strength Indicator, RSSI)，到达角度法 (Angle of Arrival, AOA) 等解算出无人机的位置信息[13–15]。此外，还可以根据无人机无线电的“指纹信息”，创建黑白名单，以实现低慢小无人机精准识别与预警[1]，但是城市环境中无线电信号复杂，探测易受到干扰，且随着低慢小无人机通信加密技术的提升，破解难度也会变大。声学探测是利用低慢小无人机飞行时产生的一定程度的噪声 (不同型号，不同飞行状态具有不同的声纹特征)，对其进行收集，分析和比对来实现无人机的检测和型号识别，成本较低且隐蔽性好，可以识别任意状态下的目标，但是易受环境噪声影响，无人机声纹库也需要不断进行更新维护，无法识别未知型号的无人机[16–19]。光电探测是指利用可见光摄像头或者红外摄像头采集指定空域的视频图像，并通过图像处理和模式识别算法来实现低慢小无人机的检测与识别，进而实现跟踪，该技术可以引导精确打击，实现全天候监控，并具备视频取证功能，此外利用云台和光学变焦等辅助设备和功能可以进一步提升系统监测的距离[20–23]。由此可见，基于光电的探测方案具有抗干扰性强，直观清晰，布站灵活的特点，可以在一定程度上避开前面几种方案的缺陷，便捷地实现对低慢小无人机的检测识别。

在光电探测系统中，视觉检测算法是其中的核心，也是决定整体识别和定位性能的

最关键因素。近年来，随着硬件算力的大幅提升以及深度学习理论的发展，计算机视觉领域的研究得到了突破性进展，相关成果也逐步应用到了人脸检测，安防监控，无人驾驶，自动零售等各个领域。基于深度学习的视觉模型具有自己根据目标任务提取相关特征的能力，可以根据训练的图片数据不断迭代更新，大大减少了图片的预处理以及手工特征设计工作，而且训练后的模型对物体的形变，位置变化以及遮挡等都具有不错的鲁棒性，其精度和速度都远超以往的传统算法，因此可以满足低慢小无人机检测识别任务的需求。

1.3 视觉检测算法国内外研究概况

目标检测算法是计算机视觉领域中核心研究方向之一，相关的研究成果已经应用到了生活中的各个场景，也逐渐催生出了多种智能化需求业务。图像目标检测任务主要分为两个步骤，首先是目标定位，即利用矩形框或者其他规则框 (多边形框，椭圆框) 等表示方法给出检测目标在图像中的具体位置，框越贴近物体的真实轮廓表明算法检测精度越高；其次是目标分类，对框出的目标进行物体分类，确定该物体属于哪一类目标类别，并给出置信度，正确类别的置信度越高表明算法的分类精度越高[24]。除了定位和分类这两个主要问题，实际应用目标检测算法时还需要考虑光照变化，运动模糊，相似动态物体干扰，尺度差异，遮挡形变，甚至是人为添加的噪声进行对抗攻击等问题。

1.3.1 传统的目标检测算法

传统的目标检测算法主要是基于模板匹配的方法，利用一系列不同尺寸的经验框，手工设计的图像特征和分类器来实现目标物体的定位和分类。首先利用经验框，以固定的步长在图像上进行滑动，或者利用启发式搜索，比如使用 Selective Search[25] 算法聚类并合并相似区域 (颜色，纹理，尺寸，交叠等)，以提取候选区域，作为目标的备选定位结果，然后对候选区域进行特征提取，抽象成特征向量，最后利用经典机器学习中的分类器对特征向量进行分类，筛选出目标类别，最终完成目标检测任务。

传统检测算法的重点在于图像特征和分类器的设计，不同类别的分类任务也会采用不同的图像特征。举例来说，Harr 特征常用于人脸检测任务，其反映了图像局部的灰度值变化情况，但容易受到光强，运动模糊的影响[26,27]；HOG(Histogram of Oriented Gragient) 特征常用于行人检测，其在局部区域统计梯度直方图，特征表达能力有所加强，在一定范围内对运动变化不敏感[28–30]；针对光强变化，LBP(Local Binary Patterns) 算子将图像的各个像素与其固定领域的像素值比较，并将比较结果编码为二进制数，实时性较快[31,32]；SIFT(Scale invariant Feature Transform) 算子作为一种稳定的局部特征，通过提取图像的关键点并对关键点进行描述来形成该图像的代表性特征，是传统手工设计特征中的代表性特征算子，具有容忍一定尺度变化，亮度变化等的鲁棒性[33,34]。分类器方面，代表性的算法有 SVM(Support Vector Machine) 二分类线性分类器，其可以通过添加高斯核函数将特征空间进行映射，从而实现非线性分类功能[35,36]；Adaboost 分

类方法通过增加困难错分样本的权重比例，改变训练数据的概率分布，并利用加权组合多个弱分类器得到最终的强分类器[37]；Decision Tree 通过建立树形结构，挖掘样本中与任务对应的特征属性，预测时使用层层推理的逻辑方式完成分类[38]。传统目标检测领域的集大成巅峰算法是由 Felzenszwalb 于 2008 年提出的 DPM(Deformable Part Model)，其使用滑动窗口，HOG 特征和 SVM 分类器，采取多部件模型策略，先对各个子部件进行检测然后对结果进行融合以获得最后的预测结果[39–42]。DPM 作为多任务学习的例子，并同时利用困难样本挖掘，上下文信息对应等辅助技巧来提高检测精度的思想，也启发了后续的基于深度学习的检测方法。

然而传统的目标检测算法性能和效率较低，存在诸多的局限性，应用场景受限。传统算法的主要瓶颈在于两点：(1) 手工设计的特征往往只是考虑到了某些因素，鲁棒性差，针对不同的物体可能还需要不同的图像特征，因此难以泛化；(2) 无论是利用滑动窗口还是启发式搜索获取目标的预选区域，都会带来较大的计算量，复杂度高，因此检测速度低，精度也得不到保证。自深度学习在计算机视觉领域大规模应用后，传统的目标检测算法已经基本被端到端的模型框架取代。

1.3.2 基于深度学习的目标检测算法

自基于深度卷积神经网络 (Deep Convolutional Neural Network, Deep CNN) 设计的 AlexNet[43] 在 ImageNet 大型公开数据集[44] 上一举取得物体分类精度的巨大飞跃后，越来越多的研究者开始利用 CNN 来完成复杂的图像特征提取工作，并使用坐标回归，区域提取，难例样本挖掘等模块组成多任务端到端目标检测学习框架。目前主流的基于深度卷积神经网络目标检测方法大致可以分为基于锚框 (Anchor-Based) 和无锚框 (Anchore-Free) 两种。基于锚框的检测方法主要是利用一系列不同的人工预设或者根据数据集聚类得到的锚框 (Anchor) 来进行目标建议区域预选工作，根据建议区域是否有二次对准回归的操作，还可以将该类方法分为二阶 (Two-Stage) 和一阶 (One-Stage) 方法。其中二阶方法精度较高，对小目标尺度也有比较好的适应性，但是速度较慢，典型的代表模型有 R-CNN[45], SPP-Net[46], Fast R-CNN[47], Faster R-CNN[48], R-FCN[49], Mask R-CNN[50] 等。一阶方法由于是直接根据预选框进行一次回归分类，所以预测速度快，可以满足大部分场景下的实时性要求，但是精度较差，尤其是在小目标上的效果不够理想，典型的代表模型有 YOLOv2[51], SSD[52] 等。随着研究工作的加深和业务需求的不断细化，二阶法和一阶法之间也在相互借鉴，界限也逐渐模糊。无锚框检测方法将物体的矩形描述框用左上角和右下角两个点来确定，将目标检测问题转换成了关键点 (Key-Point) 检测与匹配，更加符合人眼视觉特征，近年来研究热度不断提高，其代表模型有 CornerNet[53], CenterNet[54], FCOS[55] 等。

基于锚框的检测算法中，二阶模型的设计思路是先以某种方式得到一定数量的目标建议区域，然后根据 CNN 提取出的对应局部特征对其进行分类和坐标修正，得到修正备选框，接着再找出修正备选框对应的特征区域，进行第二次分类和坐标修正。2015 年，Ross Girshick 提出的 R-CNN 是二阶模型的开山之作，该检测系统首先使用 Selective

Search 算法在图片中提取出一系列可能目标物体的候选区域，然后将这些区域放大到固定大小 (227×227) 送入到 ImageNet 预训练模型 AlexNet 中进行微调，实现小数据集的迁移学习，最终得到候选区域的特征向量，接着使用这些特征向量训练 SVM 分类器，通过非极大值抑制算法 (Non-maximal Suppression, NMS) 并进行预选区域的筛选，保留局部最优的建议框，最后在这些建议框上训练每个目标类别的坐标回归器，合并分类结果后获得最后的检测结果。R-CNN 与传统的目标检测算法流程类似，只是借用 CNN 提取的特征来代替手工设计的特征，虽然整体流程繁杂，但是为后续的相关工作奠定了基础。Fast R-CNN 吸取了 SPP-Net 只前向提取图像特征一次的优点和特征金字塔池化合成区域判别向量的设计，对 R-CNN 进行了改进，仅保留了 Selective Search 生成预选区域的过程，利用 CNN 和感兴趣区域池化 (Region of Interest Pooling, RoI Pooling) 实现物体分类和边框回归的多任务学习，大大减少了 R-CNN 的推理时间和训练难度。Faster R-CNN 则进一步在 Fast R-CNN 上进行优化，设计了建议区域提取网络 (Region Proposal Network, RPN)，并提出了多尺度锚框机制，使用人为设定的 9 种不同大小和尺寸的锚框来自动提取可能的目标区域，以此替代 Selective Search 算法的工作，完全实现了端到端的训练和推理。由于其良好的检测性能和抗干扰性，Faster R-CNN 也成为了工业界比较成熟的检测方案之一。2016 年，代季峰提出了 R-FCN，针对 Faster R-CNN 中 RoI Pooling 阶段计算无法共享带来的速度问题进行了改进，使用位置敏感分布图 (Position Sensitive Score Map) 解决了“分类网络的位置不敏感性”和“检测网络的位置敏感性”之间的矛盾，在提高了检测速度的同时也提升了检测精度。2017 年，何恺明针对 RoI Pooling 中由于区域抠取的两次取整操作带来的区域位置精度损失提出了基于双线性插值的 RoI Align 方法，并加入了目标框内部的像素类别预测分支，设计出了 Mask R-CNN 模型，进一步提升了目标检测的效果。

与二阶模型相对应的一阶检测模型则是根据预设的锚框直接在 CNN 的特征图位置上进行预测，不需要复杂的候选区域生成和特征抠取操作，因此处理速度较快。YOLOv2 算法先把每张图按照设定的行列数分成一个个栅格，每个栅格上设立同等数量的锚框，网络需要预测这些预设锚框与目标位置的偏移量，存在物体的置信度和物体的类别概率。不同于 Faster R-CNN 的是，锚框的尺寸和大小是根据当前数据集的真值 (Ground Truth) 聚类而来，因此更加更具有针对性，便于模型训练优化。此外，YOLOv2 中移除了全连接层 (Fully Connected Layer)，使用全卷积替代，进一步加快了算法的推理速度。SSD 算法与 Faster R-CNN 中的 RPN 类似，其主要贡献在于使用多个特征图进行预测，考虑到浅层特征包含更多的物体位置细节，深层特征包含更抽象物体特征，每个特征图分配不同大小的锚框，分别检测小-中-大对象，以此加强算法检测多尺度目标的鲁棒性。从整体上来看，虽然一阶算法结构简洁，但是模型训练时存在正负样本不均衡，分类与回归任务特征不对齐等问题，因此检测精度和场景适应性不如二阶算法。

无锚框检测算法近两年来研究热度不断上升，由于去掉了复杂的图像区域选取和对齐，算法的灵活性变强，容易实现。2018 年，Law 等提出了 CornerNet 检测模型，该

方法基于配对左上角和右上角检测框的关键点，通过角点池化 (Corner Pooling) 和热图 (Heatmap) 集合来预测不同物体类别的角点位置，此外添加了嵌入向量 (Embedding) 预测分支来优化同类物体对应角点之间的距离至最小。该算法虽然结构新颖，检测精度也较优，但是处理时间长，缺乏对物体全局信息的考虑，在两个对应角点的组合上存在一些误判。Duan 等随后提出了 CenterNet 检测网络，在 CornerNet 的基础上加了一个目标物体中心点的偏置预测，以更加准确地定位目标，同时提出中心池化 (Center Pooling) 和级联角点池化 (Cascade Corner Pooling) 来提高算法的误检和漏检情况，最后的检测精度较 CornerNet 有很大的提高。Tian 等在 2019 年提出了基于全卷积神经网络的目标检测器 FOCS，其直接借鉴语义分割的处理思路对特征图上的每个像素点进行预测，检测精度与 CornerNet 相差不大，但是可以充分利用图像的上下文信息和全局信息，而且没有复杂的角点配对分组流程，因此更快更简洁[56]。目前来看，无锚框检测算法还不是很成熟，即使流程简洁，但是关键点的配对组合却带来了速度上和精度上的瓶颈，在面对一些复杂场 (背) 景或者 (极) 小目标物体情况下，检测的性能还不是很令人满意，因此还需要再进行相关的研究以获得突破。

1.4 视觉检测低慢小无人机研究现状与难点

近年来，随着反制无人机的需求增加，基于视觉的低慢小无人机探测的研究工作也陆续开展起来。张鹏飞[57] 利用 Canny 边缘检测算法和帧间差，光流差和背景差这三种运动目标检测算法实现了面向天空区域的无人机检测；岳子涵[58] 提出了一种自适应阈值的高斯混合模型来快速检测简单背景下无人机，并利用不动点迭代法和循环神经网络进行模糊状态下的无人机图像超分辨率重建，以缓解小目标识别问题；于鹏[59] 分别根据 HOG 特征，支持向量机和级联卷积神经网络设计了简单背景和复杂背景下的无人机识别算法；虞晓霞等[60] 通过改进经典的 LeNet-5 网络结构来实现禁飞区的监控预警系统中无人机的识别；何志祥等[61] 根据 LeNet-5 构建了一个深度为 6 的分类卷积神经网络来识别无人机；张辉等[62] 通过自建可见光无人机图像数据集，对 YOLO, SSD, Faster R-CNN 等检测算法进行了比较，探索测试了不同算法在无特殊优化条件下对所搭建数据集的检测能力；甘雨涛[63] 通过数据增强和多尺度训练的手段，利用 YOLO 算法实现了低空空域下的无人机检测；蒋兆军[64] 等采用多个摄像头组成视觉传感网络，进行白昼的图像捕捉和存储，并根据帧差法提取的无人机目标区域利用卷积神经网络进行分类；周光兵[65] 利用时空连续性算法，YOLO 以及 STC 运动目标跟踪算法构建了一套光电反无人机系统。此外，2020 年的 IEEE 国际计算机视觉与模式识别会议 (Conference on Computer Vision and Pattern Recognition, CVPR) 也举办了第一届反无人机挑战赛[66]，以吸引相关研究者开展对无人机检测和追踪算法设计。

由于低慢小无人机出现的场景多变，且其飞行距离，具体型号，飞行速度等各不相同，检测难度较大，就上述的无人机视觉探测现状而言，普遍存在算法的鲁棒性不够好，研究力度不够深入的情况，未能很好地解决探测任务中存在的难点。图 1.2 展示了视觉

检测低慢小无人机的部分难点，具体而言，任务难点主要包括以下几个方面。

(1) 深度学习模型的训练和优化需要大量的数据，而现有的低慢小无人机公开数据集很少，背景单一，质量参差不齐，因此需要手工收集相关的数据集并进行整理和标注；

(2) 低慢小无人机由于飞行距离不同，在图像上呈现出的大小也会不同。一般的大物体检测精度较高，算法也较为成熟，但是小目标检测目前没有一个完美的解决办法。对于在高空飞行的小无人机来说，其在图像中占据的像素很少，因此如何有效检测出小目标无人机，并容忍一定的无人机尺度变化是一个很大的挑战；

(3) 低慢小无人机的形态在视觉上与飞鸟，风筝等物体类似，因此检测算法需要具有一定的区分能力和抗干扰性；

(4) 低慢小无人机也是一个移动目标，因此要在保证检测精度的前提下，兼顾检测算法的实时性问题。

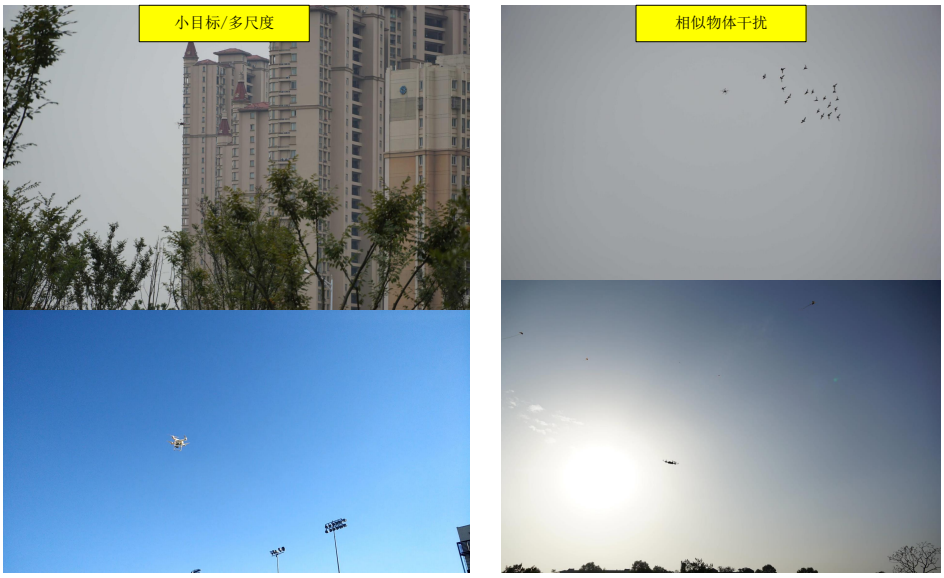


图 1.2 视觉检测低慢小无人机的部分难点示意

针对这些难点，传统的目标检测算法已经无法进行很好的处理，因此本文将采取基于深度学习的方法，通过神经网络主动提取一系列抽象程度不同的特征，设计合适的算法来实现低慢小无人机的检测识别。

1.5 论文的主要研究内容与结构安排

本文以视觉探测低慢小无人机为背景，开展基于深度学习的低慢小无人机图像检测与分割技术研究。考虑无人机的飞行姿态、位置、距离以及外部环境等因素，制作包含不同场景和相似干扰物体的低慢小无人机可见光和红外图像数据集，并在充分分析国内外相关目标检测算法的基础上，根据当前检测任务的难点，分别设计合适的检测模型，通过实验验证其性能，最终实现白天黑夜全天候监测，同时减少相似动态目标对象 (风筝和飞鸟) 的干扰。此外本文还提出了一种基于标注框的无人机弱监督语义分割系统，

尝试达到无人机的像素级定位，在不增加标注成本的情况下进一步加深检测定位无人机的能力。论文的整体研究框架如图 1.3 所示。

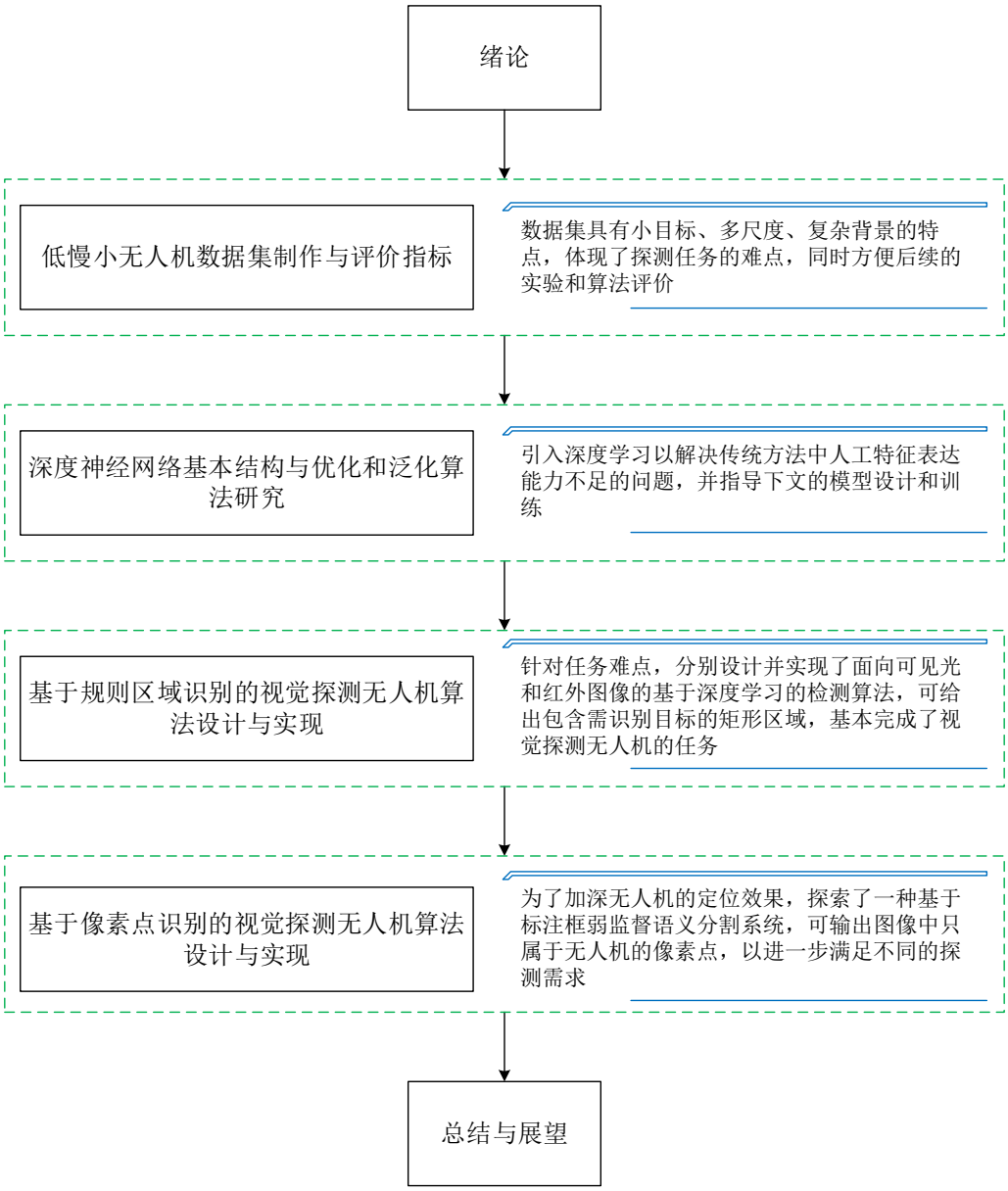


图 1.3 本文的结构安排

本文的各章节内容安排如下：

第一章，绪论。阐述本文研究的背景和意义，总结低慢小无人机检测现状与技术，梳理视觉目标检测技术的发展历程和先进算法，确立基于深度学习的低慢小无人机检测方案的有效性，最后介绍本文的主要研究内容。

第二章，低慢小无人机数据集制作与评价指标。首先引入 PASCAL VOC 和 COCO 这两个权威公开数据集，为本文数据集的组织、制作和标注提供了范例。然后梳理低慢小无人机可见光和红外热成像这两个模态的数据集制作过程，包括多方位、多形态、多

场景、多距离下的数据拍摄以及公开网络图片的扩充，和基于 PASCAL VOC 数据集格式的框标注，部分样本的像素级标注。最后研究目标检测和语义分割算法的相关评价指标。

第三章，深度神经网络基本结构与优化和泛化算法研究。介绍人工神经网络的基本内容，对神经元，非线性激活函数以及前馈神经网络做建模分析，通过引入梯度下降和反向传播算法，进一步深入研究适用于视觉探测网络模型的优化与泛化算法，并在之后引出本文使用的卷积神经网络基本架构，阐述其中的典型结构和模块。最后简单介绍几种简洁高效的代表性卷积神经网络模型，以作为后续无人机检测任务中特征提取阶段的参考。

第四章，基于规则区域识别的视觉探测无人机算法设计与实现。根据可见光数据集和红外数据集的难点分别设计基于矩形框形式的低慢小无人机的检测算法，实现全天候探测。首先承接第三章，开展对基于锚框的二阶检测算法 Faster R-CNN 和一阶检测算法 SSD 的结构分析，研究其工作流程，并通过它们在可见光数据集上的基准对比试验，分析各自的优缺点。然后针对可见光数据集中的物体存在多尺度分布，背景复杂的难点对 Faster R-CNN 进行改进，通过实验验证了改进模型的有效性。最后在可见光检测算法研究的基础上，根据红外数据集中待检物体主要为小目标甚至极小目标的特点，并考虑到红外图颜色、纹理特征单一，引入高分辨率特征提提取网络 HRNet，通过对比试验表明改进手段的可行性。综上，完成面向可见光和红外热成像探测设备的无人机检测任务。

第五章，基于像素点识别的视觉探测无人机算法设计与实现。提出一种弱监督语义分割系统实现了无人机的像素级定位，以加深探测识别力度。首先考虑到像素标注的成本消耗比框标注大，设计一种训练样本的“伪像素标签”生成流程，以供给模型学习。然后在分析基于深度学习的典型语义分割模型结构特点和缺陷的基础上，确立本文的分割模型，并针对低慢小无人机中多尺度，背景复杂等难点，引入新的损失函数和基于图像混合的抑制过拟合技术。最后对整个系统进行充分的实验，验证方案的有效性和可行性。

第六章，总结与展望。对全文的研究内容和工作进行总结，分析本文所作工作存在的不足之处，并提出后续的改进思路和方向。

2 低慢小无人机数据集制作与评价指标

利用深度神经网络来完成低慢小无人机的检测识别，需要大量的无人机图像数据，以便让模型可以在训练阶段进行充分地学习。同时为了检验相关算法的有效性，也需要合适的，多样的无人机数据来进行评价。考虑到现阶段相关的低慢小无人机公开数据集比较匮乏，本文根据实际需求，在不同的场景 (公园，城市建筑，校园等) 下，不同的距离 (几十米，几百米) 外拍摄了不同型号的多旋翼无人机图像。进一步地，考虑到全天候监测的必要性，本文采集了可见光彩色图片 (RGB) 和红外热成像图片 (Thermal) 两种数据，以加强检测系统的鲁棒性和针对性。此外，本章也详细讨论了目标检测算法的相关评价指标，以及后续用于加强定位程度的语义分割算法的评价指标，通过这些指标来衡量不同算法之间的精度优劣。

2.1 权威公开数据集概况

深度学习从本质上来说是一门数据科学，基于深度学习的目标检测需要大量的数据进行拟合学习。目前在目标检测领域中两个最常见的公开数据集是 PASCAL VOC(The PASCAL Visual Object Classes) 数据集[67] 和 MS COCO 数据集 (Microsoft Common Objects in Context)[68]，很多优秀的检测模型都是在这两个数据集上进行测试和比较，其数据组织方式也为后来其他的公开数据集沿用。

PASCAL VOC 自 2005 年开始举办视觉挑战赛，从最初的图像分类，到后面的目标检测，语义分割，动作识别等，数据集的体量和质量不断提高，也见证了很多优秀的基于深度学习的计算机视觉模型的诞生。目前仍被相关目标检测模型使用的数据集是 PASCAL VOC 2007 和 PASCAL VOC 2012，它们都总共有 4 个大类，分别是 Vehicles(车辆)、Household(家常)、Animals(动物)、Person(人)，其中前三类又被细分为更加具体的物体类别，总共形成 20 个类，这些数据主要用来关注检测和分割任务。PASCAL VOC 2007 包含图片 9663 张，物体 24649 个，PASCAL VOC 2012 包含图片 23080 张，物体 54900 个，每个训练集都被划分成训练集 (train)、验证集 (val) 和测试集 (test) 三个部分，其中训练集和验证集用以模型优化，测试集用以检验模型性能。

PASCAL VOC 数据集的标注较为谨慎，也制定了统一的标准[69]，目标检测标注采用 xml 文件格式，语义分割标注则是生成像素标签图。一份标准的目标检测标注文件内容如下所示：

```
1       <annotation>
2       <folder>VOC2007</folder>
3       <filename>000001.jpg</filename>       <!--文件名-->
4       <source>                              <!--图片来源-->
5       <database>The VOC2007 Database</database>
6       <annotation>PASCAL VOC2007</annotation>
7       <image>flickr</image>
```

```
        <flickrid>341012865</flickrid>
        </source>
        <owner>
        <flickrid>Fried Camels</flickrid>
        <name>Jinky the Fruit Bat</name>
        </owner>
        <size>                          <!--图片大小-->
        <width>353</width>
        <height>500</height>
        <depth>3</depth>
        </size>
        <segmented>0</segmented>        <!--是否分割-->
        <object>                        <!--目标信息-->
        <name>dog</name>                <!--类名-->
        <pose>Left</pose>               <!--拍摄角度-->
        <truncated>1</truncated>        <!--目标是否被截断或者被遮挡-->
        <difficult>0</difficult>        <!--物体检测难易程度(根据大小，光照等判断)-->
        <bndbox>                        <!--目标框的左上和右下两个角点坐标信息-->
        <xmin>48</xmin>
        <ymin>240</ymin>
        <xmax>195</xmax>
        <ymax>371</ymax>
        </bndbox>
        </object>
        </annotation>
```

PASCAL VOC 数据集定量评价检测模型的指标为 mAP(mean Average Percision)，其衡量了所有类别中检测的精确率 (Precision, P) 和召回率 (Recall, R) 的平均性能。对于二分类问题而言，可以根据样本的真实类别和分类器的预测类别进行比对划分为真正例 (True Positive, TP)、假正例 (False Positive, FP)、真反例 (True Negative, TN) 和假反例 (False Negative, FN) 四种情况，分类结果的混淆矩阵 (Confusion Matrix) 如表 2.1 所示[70]。精确率和召回率分别定义为：

$$P = \frac{TP}{TP + FP} \tag{2.1}$$

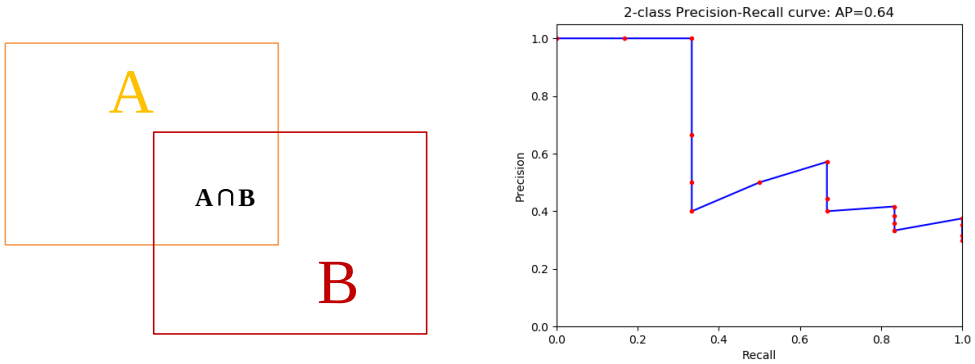
$$R = \frac{TP}{TP + FN} \tag{2.2}$$

表 2.1 分类结果混淆矩阵

真实情况	预测为正	预测为假
正例	TP(真正例)	FN(假反例)
反例	FP(假正例)	TN(真反例)

精确率越高，代表误检少，召回率越高，代表漏检少，这两个度量是相互矛盾的，往往不可能全部兼顾。在目标检测中，通过指定预测框与真实框的交并比 (Intersection

over Union, IoU) 阈值来判断结果真假 (PASCAL VOC 中指定该阈值为 0.5)，从而计算上述两个度量值，并根据置信度结果对样本排序，逐步计算并绘制出 P-R 曲线。该曲线与横轴围成的面积即为 AP 值，这样计算出每一类检测目标的 AP 值再取平均便得到了模型在该数据集上的检测结果 mAP。图 2.1 展示了 IoU 的定义和 P-R 曲线示例。而语义分割的评价指标则相对简单，可以直接使用 mIoU(mean IoU)，这两个指标具体的计算流程与近似公式将在2.3节详细论述。



(a) *IoU = area(A ∩B)/area(A ∪B)*　　(b) P-R曲线示例

图 2.1 IoU 计算过程与 P-R 曲线图样例

MS COCO 数据集是微软于 2014 年出资发布的大型公开数据集，也是当今目标检测与分割最权威的数据集之一。该数据集主要从日常复杂场景中截取，总共约有 33 万张图片，80 个物体类别，并提供了物体检测框的精确坐标和像素级别的位置标注，其精度均为小数点后两位。由于增加了像人类关键点检测，场景理解等复杂任务，所以标注上较 PASCAL VOC 繁琐一些。在目标检测任务上，COCO 也采用 AP 指标，但是其 IoU 的设定阈值分为十个档次 (0.5-0.95，中间以 0.05 递增)，因此检测性能的衡量会更加全面和严苛。COCO 的 AP@0.5 即与 PASCAL VOC 的 mAP 指标相同。

本文制作的数据集物体类别较少，体量上与 PASCAL VOC 接近，而且重点关注目标检测和语义分割任务，因此制作流程和标注格式将参照 PASCAL VOC 进行。

2.2 数据集制作

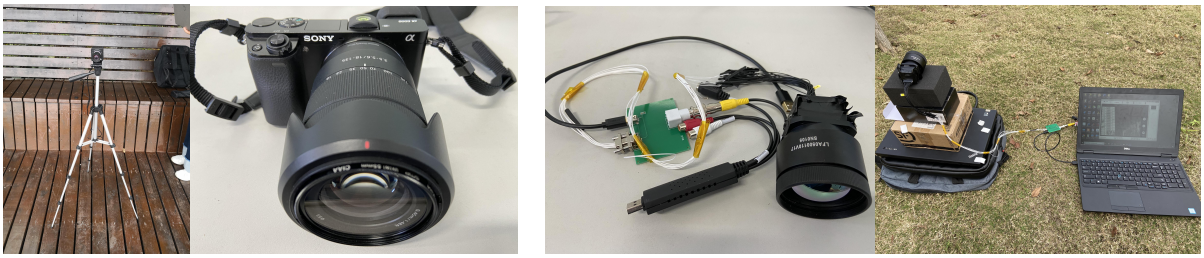
如第一章所述，本文主要根据 PASCAL VOC 公开数据集来组织自己的低慢小无人机数据集，本文的数据集以无人机为主要采集对象，同时也兼顾了风筝和鸟这两类干扰物体。数据集的制作主要分为数据采集与扩充，数据标注两个部分。

2.2.1 数据采集与扩充

本文主要是针对低慢小无人机视觉检测进行研究，因此采集了可见光彩色图片和红外热成像图片这两种模态的数据，以保证检测系统的容错性和工作时长。同时为了适应无人机多样性，移动性以及场景复杂性的特点，本文利用了多种型号的旋翼无人机，包

括六旋翼，四旋翼等，其无人机最长轴距为 120cm，最短轴距为 20cm。图 2.2 展示了这些用于实验的无人机设备。如图 2.3 所示，彩色图像的采集设备是锐尔威视的 USB 摄像头和一台 Sony α6000 微单，红外图像的采集设备是艾睿光电的 SN0105 热成像头。本文通过照片直拍和录制视频再截取的方式来获取无人机及相关物体图像数据。

图 2.2 用于拍摄的不同型号的无人机



(a) 可见光图像采集摄像头 (b) 红外热图采集摄像头

图 2.3 数据拍摄设备

由于低慢小无人机拍摄难度较大，为了获取更加充分的 RGB 数据，本文另从公开数据集 USC Drone[71] 和网络视频中选取了一些无人机图像作为补充，最终彩色图像总共收集了 15659 张，其中包含无人机对象 15227 架，风筝 1407 个，鸟 7437 只。图 2.4 显示了不同场景下的无人机样张，包括公园，建筑，强光，弱光，雾天等。图 2.5 给出了不同角度，不同距离外拍摄的无人机图像，体现了无人机移动目标带来的多尺度特点，其中无人机对象的最小像素为 10，最大像素为 193626，平均像素为 6612。图 2.6 给出了风筝和鸟这两类干扰物体的图像样张，其中风筝的最小像素为 42，最大像素为 164690，平均像素为 5040；鸟的最小像素为 4，最大像素为 10890，平均像素为 976，它们在视觉上也属于小物体，形态上与低慢小无人机类似。

图 2.4 不用场景下的低慢小无人机彩色图像样张

图 2.5 不同距离下的低慢小无人机彩色图像样张

图 2.6 风筝和鸟类彩色图像样张

红外图像采集方面，由于低慢小无人机的自发红外辐射主要来自于机身外壳和工作状态下发热的电池，红外辐射能量很低，因此需要在收集前进行无人机的红外特性分析，确立合适的探测距离，保证镜头透过率，以满足图像质量。此外，也要选择合适的电路结构对背景噪声进行抑制，加强目标信号。

本文主要在公园和建筑物旁对低慢小无人机及其他物体进行了拍摄，总共收集了 5546 张红外图片，其中包含无人机 4438 架，风筝 1707 个，鸟类 62 只。图 2.7 给出了不同场景下，不同形态的无人机、风筝等物体的红外图像数据样例。像素大小方面，无人机的最小像素为 15，最大像素为 3504，平均像素为 249；风筝的最小像素为 36，最大像素为 2278，平均像素为 366；鸟的最小像素为 28，最大像素为 760，平均像素为 236。

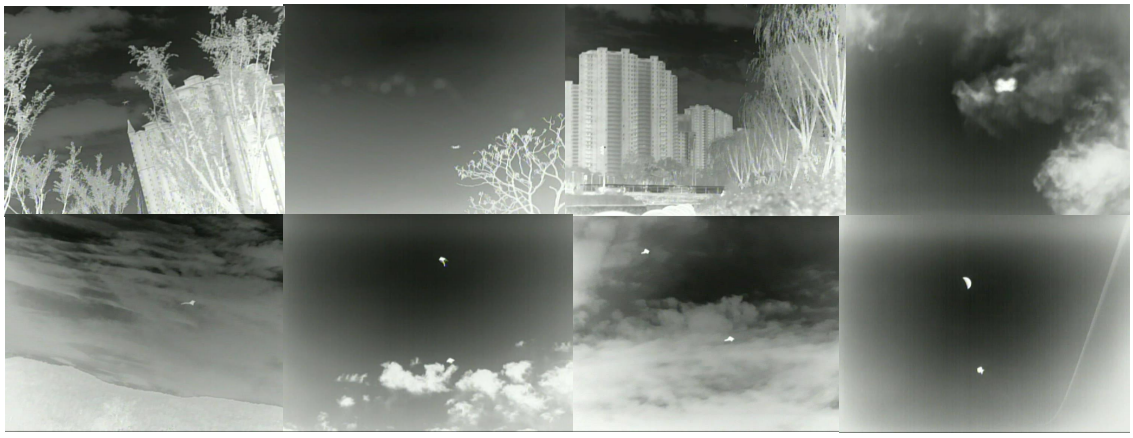


图 2.7 低慢小无人机，风筝等红外图像样张

2.2.2 数据标注

由于收集的数据体量并不是很大，因此 RGB 图像和红外图像都只分了训练集和测试集，不单独分出验证集，在训练的时候直接采用测试集进行过拟合的判断，以进行提

前停止。RGB 图像数据中，训练集为 12366 张，测试集为 3293 张；红外图像中，训练集为 4021 张，测试集为 1525 张，两类模态的数据集都是大致按照 4：1 的比例划分。本文对所有的图像都按照 PASCAL VOC 的格式进行了框标注，其中标注文件的最主要内容为图像中物体的类名以及该物体的坐标位置，该位置信息用矩形标注框的左上角和右下角坐标表示，即 $(x_{min}, x_{max}, y_{min}, y_{max})$。标注软件使用的是 LabelImg[72]，具体操作界面如图 2.8 所示。

图 2.8 使用 LabelImg 软件对图片进行矩形框标注

此外，考虑到第五章要进行仅使用标注框来实现低慢小无人机的弱监督语义分割(仅针对彩色图像)，为了评价算法的有效性，需要对 RGB 图像数据中的测试集进行像素级标注，以计量算法的精度。其中，像素标注的软件为 Labelme[73]，标注过程如图 2.9 所示，2.10 为生成的像素标签，以图像的形式存储，作为弱监督语义分割算法的评价真值。

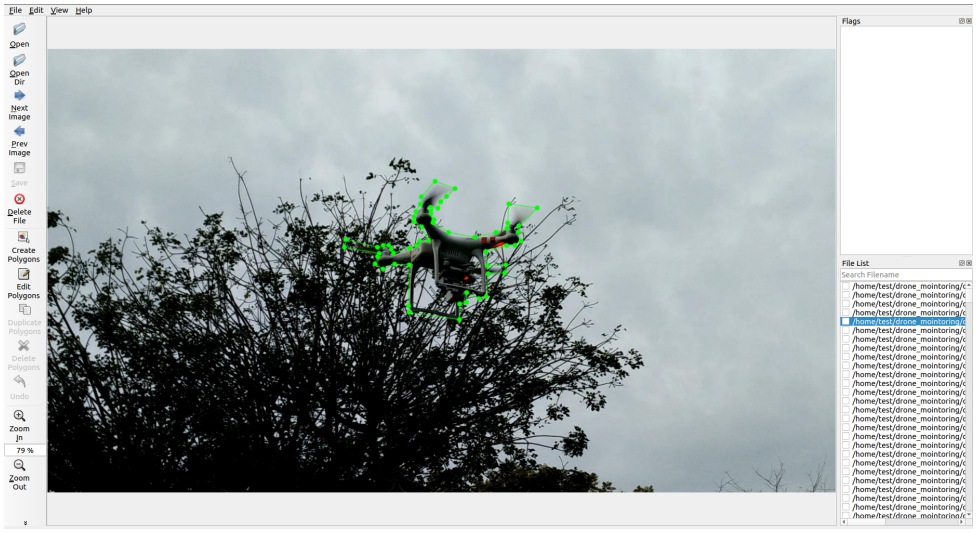


图 2.9 使用 Labelme 软件对图片进行多边形框标注，使后续可生成像素标签

图 2.10 利用标注的多边形框生成无人机像素标签 (红色部分为无人机，黑色部分为背景)

2.3 算法评价指标

在目标检测中，算法的评价指标主要是 mAP (mean Average Precision), 在语义分割中，算法的主要评价指标为 mIoU (mean Intersection of Union), 下面主要对这两个指标的计算流程进行介绍。

如第2.1节所述，对模型检测结果的成功与否判定是通过 IoU(交并比) 来进行的，本文也依据 PASCAL VOC 数据集的标准来设置 IoU 阈值，即 0.5。为了画出 P-R 曲线，首先将预测结果按照类别置信度降序排序，然后依次选取每个预测结果的置信度做为正负样本划分阈值，即大于该置信度的全部认定为正样本，小于该阈值的全部认定为负样本。接着将样本判定结果与真实的标注结果比较，统计出真正例 (True Positive, TP)、假正例 (False Positive, FP)、真反例 (True Negative, TN) 和假反例 (False Negative, FN) 这四种例别数量，并根据公式 (2.1) 和 (2.2) 计算出该阈值下的精确率 (Precision, P) 和召回率 (Recall, R)。最后以精确率为纵轴，召回率为横轴，把这些得到的一系列 (精确率，召回率) 坐标点连起来画出 P-R 曲线。每个物体类别都会有一个 P-R 曲线，其检测精度 AP 是通过计算该曲线与横轴围成的面积得到，即召回率从 0 到 1 范围内对精确率的积分值。

AP 的值在 [0,1] 范围之间，越大代表检测精度越高。在实际应用中，大多是通过近似计算而非准确积分来得到 AP 的值。以 PASCAL VOC 为例，首先记录那些在绘制 P-R 曲线时计算出的召回率值 (假设总共有 N 个)；然后找出每个召回率值在 P-R 曲线上对应点之后的最大精准率值；接着以该精准率值为高，相邻召回率点之间的距离为宽，相乘得到小矩形块的面积；最后重复以上步骤，将 N 个矩形面积加起来作为近似积分值，其表达式如 (2.3) 所示。

$$\mathrm{AP} = \int_0^1 P(R)dR \approx \sum_{k=1}^{N} P(k)\Delta R(k) \tag{2.3}$$

每类物体，都可以按照式 (2.3) 计算出对应的 AP 值，然后以它们的平均值 (即 mAP) 作为检测算法的定量精度值，以此衡量算法的优劣。

语义分割算法的常用精度评价指标有像素准确率 (Pixel Accuracy, PA)，平均像素准确率 (mean Pixel Accuracy, mPA) 和平均交并比 (mean Intersection over Union, mIoU)[74]，其中 mIoU 是目前语义分割的标准评价指标，本文第五章的弱监督语义分割算法也将采取 mIoU 作为精度衡量指标。mIoU 表示的是算法预测出的每类物体像素掩模 (Mask) 与真实标注掩模间的平均交并比，这里的交并比含义与目标检测一样，只不过语义分割中的交并比对象不是矩形框，而是不规则的图形。

假设 k 为语义分割的物体类别 (加上背景)，p_{ij} 代表真实类别 i 被错分成类别 j 的像素数量，p_{ii} 代表真实类别 i 被正确分为类别 i 的像素数量，则 mIoU 的计算公式为：

$$\mathrm{mIoU} = \frac{1}{k}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}} = \frac{1}{k}\sum_{i=0}^{k}\frac{TP}{TP+FP+FN} \tag{2.4}$$

2.4 本章小结

本章主要介绍了收集低慢小无人机数据集时所参考的公开数据集概况，以及整体制作过程。本章共收集了 15659 张可见光图片和 5546 张红外热成像图片，这两类模态数据涵盖的物体类别为无人机，风筝和飞鸟，背景包括公园，建筑，不同光强等，大小上具有多尺度分布和小尺度的特点，无人机的形态也不尽相同。之后按照 PASCAL VOC 数据集的格式对所有图片进行了框标注，对可见光图像的测试集进行了像素级标注，以保证算法的训练和评价。此外还讨论了目标检测和语义分割算法的定量精度衡量指标，作为后续算法比较的参考。

3 深度神经网络基本结构与优化和泛化算法研究

深度学习是机器学习中一种以人工神经网络为架构对数据进行表征学习的算法，在此基础上，深度学习进一步假定这一相互作用的过程可分为多个层次，代表对观测值的多层抽象，不同的层数和层的规模可用于不同程度的抽象[75,76] 。进入 21 世纪以来，深度学习取得了长足进步，解决了以往传统技术难以解决的科学问题，并出现了爆炸式增长，其拥有对庞大数据的处理能力，并给出不错的特征表达，通过多层神经网络和适当的非线性激活函数以及反向传播算法 (BackPropagation, BP) 不断修正预测结果，以逼近样本真值。这种端到端一体的机制，与不同任务需求的相应处理模块结合，可以很好地完成诸如预测、检测识别、分割等目标。本章将介绍其中相关的基本理论，包括人工神经网络的基本结构，反向传播算法，优化、泛化算法，以及应用在视觉探测低慢小无人机任务中的卷积神经网络的基本组成模块，并在最后介绍三种典型的网络模型，以其作为本文的特征提取网络 (或启发设计)，作为后续探测算法实现的基础与保证。

3.1 人工神经网络基本原理

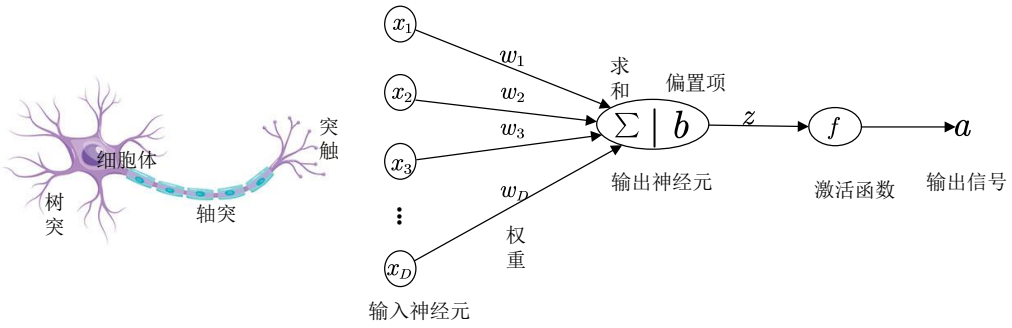
人工神经网络 (Artificial Neural Network，ANN) 是指一系列受生物学和神经科学启发的数学模型。这些模型主要是受到人类大脑工作机理的启发，对其中的人脑神经元网络进行建模抽象，通过人工构建神经元，并设计特定的拓扑结构和激励/抑制机制来连接和沟通多个神经元，以此模拟具备生物特性的神经网络[77]。神经网络一般被看成一个高度非线性的数学模型，网络的深度，非线性激活函数 (Activation Function) 的选取，以及内部大量神经元之间的连接方式，让神经网络模型可以处理较为复杂的任务。其中，各个神经元之间的权重就是需要在训练过程中去学习的参数，一般可以利用梯度下降和反向传播算法进行。

3.1.1 人工神经元模型

人工神经元 (Artificial Neuron) 是组成神经网络的基本单元，主要作用是接受来自其他神经元的信号，并根据相应的状态给出输出，传递给下一个神经元。如图 3.1 所示，人工神经元在功能设计上与生物神经元类似，其整个流程是对生物神经元的模拟抽象。假设一个神经元接收了 D 个神经元的输入 $x_1, x_2, \cdots, x_D$，其输入信号可以表示为 D 维向量 $\boldsymbol{x} = [x_1; x_2; \cdots; x_D]$，这些信号经过加权后累加在一起，得到了最终属于该神经元接收到的净输入信号 $z(z \in \mathbb{R})$。

$$z = \sum_{d=1}^{D} w_d x_d + b = \boldsymbol{w}^{\top} \boldsymbol{x} + b \tag{3.1}$$

上式中，$\boldsymbol{w} = [w_1; w_2; \cdots; w_D] \in \mathbb{R}^D$ 是与输入信号对应的权重向量，$b \in \mathbb{R}$ 是偏置项，一般是一个待学习常值参数，可以控制神经元的激活状态，增加神经网络的自由度。



(a) 生物神经元示意图 (b) 人工神经元示意图

图 3.1 生物神经元和人工神经元示意图

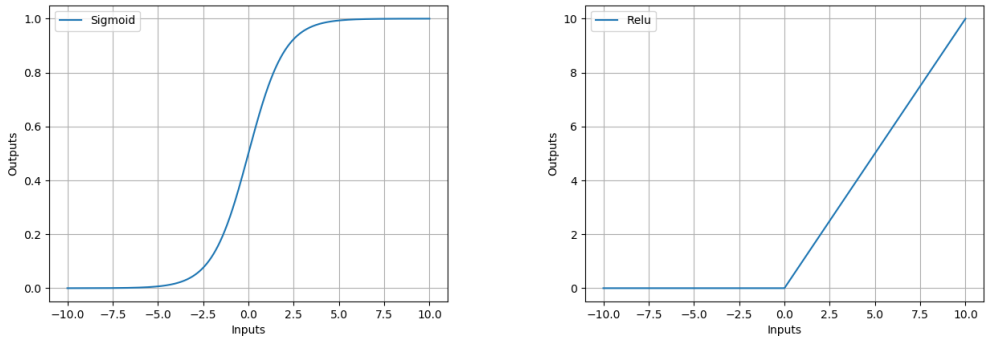
信号 z 再经过一个非线性激活函数 f 之后，得到该神经元的活性值 (Activation) a，该活性值代表接收的其他神经元输入信息后产生的信号量，用以传递给下一个神经元。

$$a = f(z) \tag{3.2}$$

3.1.2 非线性激活函数

如前所述，人工神经元的主要作用就是对输入信号做了一次仿射变换和非线性变换，其中仿射变换是为了综合局部信息，而非线性变换是为了增加网络的表征能力和学习 (逼近) 能力，没有非线性变换这一步，加深网络则失去了意义。非线性激活函数就是其中一种较为简单直接的非线性变换手段。

在本文的探测任务中，主要使用两种激活函数作为网络非线性化的手段，其中一个是 Sigmoid[78]，另一个为 ReLU(Rectified Linear Unit, 修正线性单元)[79]。前者主要用来产生预测的概率输出，一般用在网络的输出层，而后者则是用在网络的隐藏层中，它们的函数图像如图 3.2 所示，下面分别对其进行介绍。



(a) Sigmoid激活函数 (b) ReLU激活函数

图 3.2 Sigmoid, ReLU 函数图像示意图

Sigmoid 函数是一个在生物学中常见的 S 型函数，也称为 Logistic 函数，曾在经典机器学习中用以逻辑回归进行分类。其函数图像如图 3.2(a) 所示，函数定义如式 (3.3)

所示。

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{3.3}$$

Sigmoid 激活函数可以把实数域的输入压缩到 (0,1) 之间，实现“挤压”效果。该函数的因变量特性与生物神经元类似，当输入越小时，函数值越接近与 0，当输入越大时，函数值越接近与 1，有种饱和抑制的效果，当输入在 0 附近时，近似于线性函数，函数值变化明显。此外，Sigmoid 函数是连续可导的，便于神经网络的计算和训练，其输出也可以直接看做是概率分布。

但是 Sigmoid 激活函数在深度神经网络的训练中存在两个问题。一个是当输入的值过大或过小时，函数的梯度接近于 0，这样在反向传播的时候会出现梯度消失的情况，导致网络无法更新参数。另一个它的梯度值不是中心对称的，由式 (3.4) 可知，函数的梯度值是大于 0 的，因此网络在训练时会朝着正方向或负方向一直更新，导致网络收敛速度变慢，产生振荡。所以 Sigmoid 目前基本不用于网络的隐藏层中，而是作为产生概率分布的手段。

$$\frac{\partial\sigma(x)}{\partial x} = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}} - \frac{1}{(1+e^{-x})^2} = \sigma(x)(1-\sigma(x)) \tag{3.4}$$

ReLU 是神经网络中最常用的激活函数之一，其计算形式简单，导数值为常值，稳定且易于收敛。该函数图像如图 3.2(b) 所示，具体定义是一个斜坡函数，定义如式 (3.5) 所示。

$$\mathrm{ReLU}(x) = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases} = \max(0, x) \tag{3.5}$$

采用 ReLU 激活函数的神经元往往只需要进行比较操作，计算效率高，其单侧抑制 (左饱和) 也符合生物学合理性，右侧导数为 1 在一定程度上也缓解了网络训练过程中梯度消失的问题，而且 ReLU 可以导致网络具有稀疏性，即大约只有 50% 左右的神经元处理激活状态，这一点也符合人脑生物结构。因此本文的模型将在隐藏层中采用该激活函数辅助模型训练和表征。

此外，在检测模型对无人机、风筝和飞鸟这三种物体进行分类判别时，还需要利用 Softmax 激活函数[80] 来输出属于每类物体的置信度。Softmax 通常是用在单标签多分类问题上，基本上只出现在神经网络的最后一层进行概率判别，而不会用于隐藏层中实现非线性化。对于一组输入，Softmax 会通过指数求和再求权重比例的方式得出每个输入对应的预测概率值 (0-1 之间)，并保证整体的预测和为 1。其具体计算公式如式 (3.6) 所示。

$$S_i = \frac{e^{V_i}}{\sum_j e^{V_j}} \tag{3.6}$$

上式中，$\mathbf{V}$ 代表一组输入，S_i 代表其中第 i 个输入值 V_i 经过 Softmax 之后得出的概率预测值。Softmax 的定义和计算方式都较为直观，设计初衷是希望样本特征对概率的影响是乘性的。Softmax 与 Sigmoid 虽然都可以输出概率值，但是前者可以很好地处

理多分类问题，并且可以实现类间竞争，而后者则常用于二分类，因此本文将根据不同的分类需求来恰当地使用这两种激活函数。

3.1.3 前馈神经网络

前馈神经网络 (Feedforward Neural Network, FNN) 是一种有着比较直接的拓扑结构的人工神经网络，也被称为多层感知机 (Multi-Layer Perceptron, MLP)，其内部按接收信息的先后分成不同的组，每一组就是一个神经网络层 (Layer)，信息从输入到输出是单向流动的，不存在反向连接，可以用一个有向无环图来进行表示。本文的探测任务需要不断地对图像特征进行抽象表达和复用，因此所使用的皆为前馈神经网络。图 3.3 给出了一个全连接多层前馈神经网络的示意图，其中第 0 层是输入层，代表输入样本的信息，一般不计入网络的深度，最后一层是输出层，代表任务需要的预测值，其余的中间层是隐藏层，主要用来进行复杂的特征提取、表征和变换。表 3.1 给出了多层前馈神经网络模型的相关数学记号。

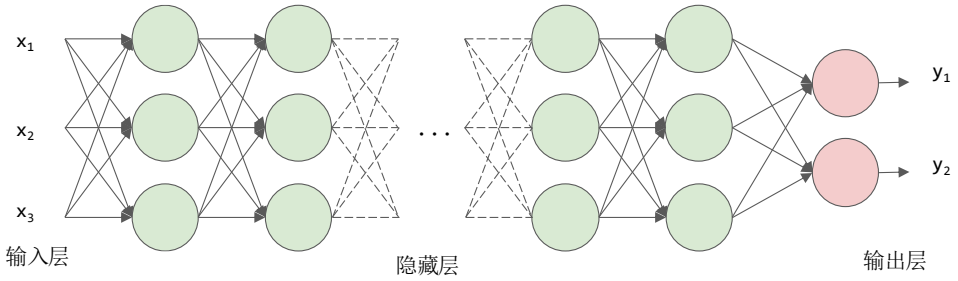


图 3.3 多层前馈神经网络示意图

表 3.1 多层前馈神经网络相关数学记号

数学记号	具体含义
L	神经网络的层数 (隐藏层加输出层)
M_l	第 l 层神经元的个数
$f_l(\cdot)$	第 l 层神经元的激活函数
$\boldsymbol{W}^{(l)} \in \mathbb{R}^{M_l \times M_{l-1}}$	第 $l-1$ 层到第 l 层的权重矩阵
$\boldsymbol{b}^{(l)} \in \mathbb{R}^{M_l}$	第 $l-1$ 层到第 l 层的偏置
$\boldsymbol{z}^{(l)} \in \mathbb{R}^{M_l}$	第 l 层神经元的净输入
$\boldsymbol{a}^{(l)} \in \mathbb{R}^{M_l}$	第 l 层神经元的输出 (活性值)

假设网络的输入是一个多维样本数组 $\boldsymbol{x}$，令 $\boldsymbol{a}^{(0)} = \boldsymbol{x}$，则网络的信息传递可以表示为：

$$\boldsymbol{a}^{(l)} = f_l\left(\boldsymbol{W}^{(l)}\boldsymbol{a}^{(l-1)} + \boldsymbol{b}^{(l)}\right) = f_l(\boldsymbol{z}^{(l)}) \tag{3.7}$$

网络根据第 $l-1$ 层的活性值 $\boldsymbol{a}^{(l-1)}$ 作为输入，计算出第 l 层的净输入活性值 $\boldsymbol{z}^{(l)}$，再经过激活函数得到第 l 的输出活性值 $\boldsymbol{a}^{(l)}$，以此类推。前馈神经网络通过这样迭代式

的逐层信息传递，得到最后的网络输出 $\boldsymbol{a}^{(L)}$。我们可以将整个网络看做成是一个复杂的复合函数 $\boldsymbol{\phi}(\boldsymbol{x};\boldsymbol{W},\boldsymbol{b})$，$\boldsymbol{W},\boldsymbol{b}$ 代表网络所有的权重矩阵和偏置项，该函数的输入是 $\boldsymbol{x}$，输出是 $\boldsymbol{a}^{(L)}$.

除了结构简单之外，多层前馈神经网络的另一个优点是具有很强的拟合能力，大部分的非线性连续函数都可以用它来进行近似逼近 (通用近似定理[81])。然而大多数情况下，复杂任务的真实映射函数是无法获知的，在深度学习理论中，为了尽量给出一个最优的近似函数，采用如图 3.4 所示的，求误差函数梯度和反向传播算法来进行网络的参数 (权重，偏置等) 更新 (梯度下降)，直至最终的误差达到设定精度。

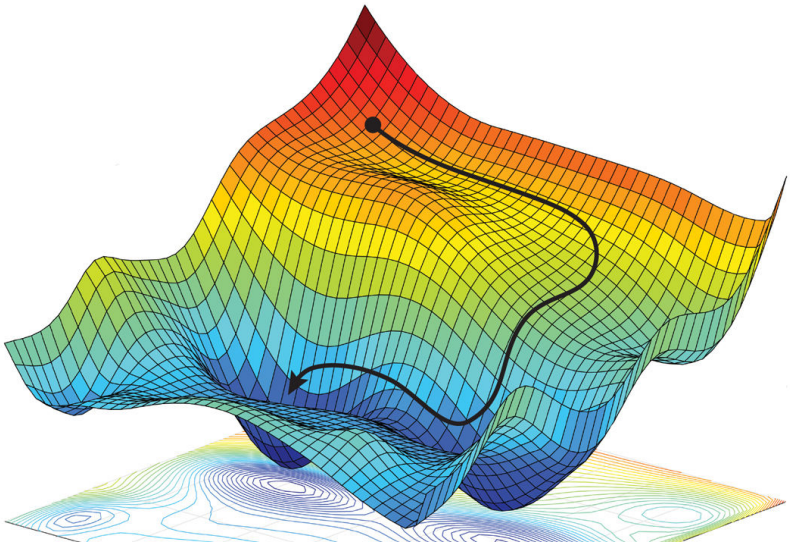


图 3.4 利用梯度下降和反向传播算法寻找网络最优的参数

3.1.4 反向传播算法

不失一般性，假定采用随机梯度下降来对前馈神经网络进行参数更新，现给定一个样本集 $(\boldsymbol{x},\boldsymbol{y})$，其中 $\boldsymbol{y}$ 是输入数据 $\boldsymbol{x}$ 对应的真值 (Ground Truth)，经过神经网络后得到的输出是 $\hat{\boldsymbol{y}}$，为了定量衡量真实值和预测值之间的误差，利用损失函数 $\mathcal{L}(\boldsymbol{y},\hat{\boldsymbol{y}})$ 来完成。现在，为了进行参数的学习，就需要计算出该损失函数对每个参数的导数，然后往梯度的负方向更新使误差变小。

对于神经网络第 l 层的参数 $\boldsymbol{W}^{(l)}$, $\boldsymbol{b}^{(l)}$，根据链式法则计算偏导数，有：

$$\begin{aligned}\frac{\partial\mathcal{L}(\boldsymbol{y},\hat{\boldsymbol{y}})}{\partial w_{ij}^{(l)}}&=\frac{\partial\boldsymbol{z}^{(l)}}{\partial w_{ij}^{(l)}}\frac{\partial\mathcal{L}(\boldsymbol{y},\hat{\boldsymbol{y}})}{\partial\boldsymbol{z}^{(l)}}\\\frac{\partial\mathcal{L}(\boldsymbol{y},\hat{\boldsymbol{y}})}{\partial\boldsymbol{b}^{(l)}}&=\frac{\partial\boldsymbol{z}^{(l)}}{\partial\boldsymbol{b}^{(l)}}\frac{\partial\mathcal{L}(\boldsymbol{y},\hat{\boldsymbol{y}})}{\partial\boldsymbol{z}^{(l)}}\end{aligned}\tag{3.8}$$

在式 (3.8) 中，$w_{ij}^{(l)}$ 是权重矩阵 $\boldsymbol{W}^{(l)}$ 中第 i 行第 j 列的元素。对于偏导数 $\frac{\partial\mathcal{L}(\boldsymbol{y},\hat{\boldsymbol{y}})}{\partial\boldsymbol{z}^{(l)}}$，其反映了第 l 层神经元对最后损失误差的影响，也间接体现出不同层的神经元对网络表达能力的贡献程度，称该偏导数为第 l 层神经元的误差项，用 $\delta^{(l)}$ 来表示，即：

$$\delta^{(l)}\triangleq\frac{\partial\mathcal{L}(\boldsymbol{y},\hat{\boldsymbol{y}})}{\partial\boldsymbol{z}^{(l)}}\in\mathbb{R}^{M_l}\tag{3.9}$$

对式 (3.9) 使用链式法则，计算得到：

$$\begin{aligned}\delta^{(l)} &\triangleq \frac{\partial \mathcal{L}(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\partial \boldsymbol{z}^{(l)}} \\ &= \frac{\partial \boldsymbol{a}^{(l)}}{\partial \boldsymbol{z}^{(l)}} \cdot \frac{\partial \boldsymbol{z}^{(l+1)}}{\partial \boldsymbol{a}^{(l)}} \cdot \frac{\partial \mathcal{L}(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\partial \boldsymbol{z}^{(l+1)}} \\ &= \operatorname{diag}\left(f_l'\left(z^{(l)}\right)\right) \cdot \left(\boldsymbol{W}^{(l+1)}\right)^{\top} \cdot \delta^{(l+1)} \\ &= f_l'\left(\boldsymbol{z}^{(l)}\right) \odot \left(\left(\boldsymbol{W}^{(l+1)}\right)^{\top} \delta^{(l+1)}\right) \quad \in \mathbb{R}^{M_l}\end{aligned} \tag{3.10}$$

上式中，$\odot$ 代表向量的点积运算，表示每个元素相乘。从上式可以看出，神经网络第 l 的误差项可以由第 $l+1$ 层的误差项计算得到，这样只要计算出了最后一层输出层的误差项，就可以逐层往前递推，得到中间每一层的误差项，这就是反向传播的具体含义。

最后进一步对矩阵求导并结合式 (3.10)，求出损失函数关于网络第 l 层权重矩阵和偏置项的结果为：

$$\begin{aligned}\frac{\partial \mathcal{L}(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\partial \boldsymbol{W}^{(l)}} &= \delta^{(l)}\left(\boldsymbol{a}^{(l-1)}\right)^{\top} \quad \in \mathbb{R}^{M_l \times M_{l-1}} \\ \frac{\partial \mathcal{L}(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\partial \boldsymbol{b}^{(l)}} &= \delta^{(l)} \quad \in \mathbb{R}^{M_l}\end{aligned} \tag{3.11}$$

下面给出利用梯度下降和反向传播算法训练前馈神经网络的具体流程。

算法 3.1 利用随机梯度下降和反向传播算法训练前馈神经网络流程

输入： 给定训练集 $\mathcal{D} = \left\{\left(\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)}\right)\right\}_{n=1}^{N}$，验证集 $\mathcal{V}$，前馈神经网络层数为 L，每层神经元数量为 $M_l, 1 \leq l \leq L$，梯度下降的步长 (学习率) 为 α.

随机初始化网络参数 $\boldsymbol{W}, \boldsymbol{b}$;
repeat
  对训练集 $\mathcal{D}$ 中的样本打乱，随机重排序;
  for $n = 1 \cdots N$ **do**
    从训练集 $\mathcal{D}$ 中选取样本 $\left(\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)}\right)$;
    前向计算网络每层的净输入 $\boldsymbol{z}^{(l)}$ 和输出的激活值 $\boldsymbol{a}^{(l)}$，直到最后的输出层;
    利用反向传播算法计算每层的误差项 $\delta^{(l)}$;
    计算每层参数的偏导数，$\forall l$，$\frac{\partial \mathcal{L}\left(\boldsymbol{y}^{(n)}, \hat{\boldsymbol{y}}^{(n)}\right)}{\partial \boldsymbol{W}^{(l)}} = \delta^{(l)}\left(\boldsymbol{a}^{(l-1)}\right)^{\top}$, $\frac{\partial \mathcal{L}\left(\boldsymbol{y}^{(n)}, \hat{\boldsymbol{y}}^{(n)}\right)}{\partial \boldsymbol{b}^{(l)}} = \delta^{(l)}$;
    更新每层参数，$\boldsymbol{W}^{(l)} \leftarrow \boldsymbol{W}^{(l)} - \alpha\delta^{(l)}\left(\boldsymbol{a}^{(l-1)}\right)^{\top}$，$\boldsymbol{b}^{(l)} \leftarrow \boldsymbol{b}^{(l)} - \alpha\delta^{(l)}$;
  end for
until 神经网络在验证集 $\mathcal{V}$ 上的错误率满足设定的精度或者不再下降;

输出： 前馈神经网络参数 $\boldsymbol{W}, \boldsymbol{b}$.

3.1.5 神经网络的优化与泛化

神经网络的优化问题往往是一个非凸优化问题，优化函数较为复杂，而目前大多使用梯度下降来实现参数更新，这种一阶优化方法，只能保证收敛到局部最小值点 (比如

图 3.4 展示的例子)，且收敛速度和效果还与参数的初始化值有关，此外还有可能出现梯度消失无法继续优化的情况。

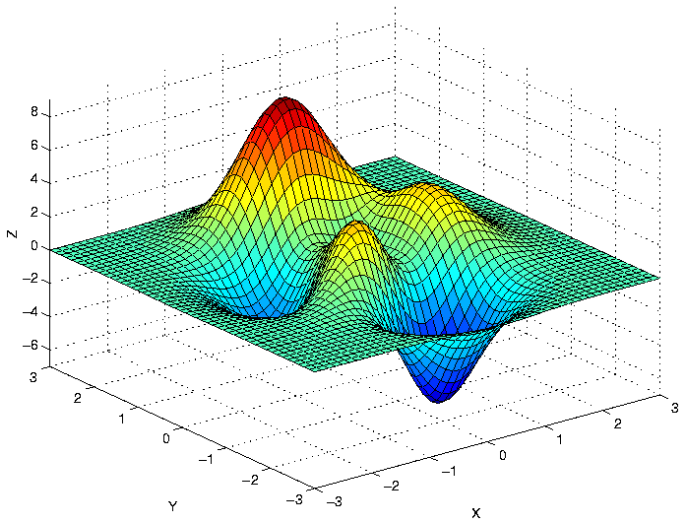


图 3.5 神经网络的优化函数可能存在多个极小值点以及鞍点，会使梯度下降陷入局部最小值

由于本文收集的低慢小无人机数据集体量较少，同时包含复杂的、数量不均衡的样本，存在不小的训练难度，为了让最终的检测模型能够更好地学习到数据中的特征和规律，尽量逃脱局部最小值点，将采用以下几种方式来辅助训练。

1). 小批量梯度下降

在具体的实现中，梯度下降可以分为随机梯度下降 (Sophisticated Gradient Descent)，批量梯度下降 (Batch Gradient Descent) 和小批量梯度下降 (Mini-Batch Gradient Descent)。这三种方法的区别主要在于一次梯度计算采用的样本数量不同，随机梯度下降是随机抽取训练样本中的一个，批量梯度下降是采取整个训练集样本，而小批量梯度下降是随机抽取固定数量的训练样本。一般用来训练神经网络的数据集都比较大，每次都计算整体样本的梯度虽然结果较为准确，更新平稳，但是代价较大，而且样本也存在冗余，仅采取一个样本计算梯度会出现噪声，带来参数更新的振荡，但是有时候这种振荡可以跳出局部最小值点。采取小批量样本的方法则是综合了前面两种方法的优点，在一定程度上即保证了训练的稳定性，又可以冲出局部最小值点，因此本文将采取这种方法进行训练。

网络每次小批量更新为一次迭代，直至所有的训练样本全部采样完毕，便完成了一次回合 (Epoch)。此外值得一提的是，在神经网络训练中，批量大小 K 一般不影响梯度的期望而且影响梯度的方差，K 越大 (一定范围内[82])，梯度方差越小，噪声也越小，训练就会越稳定，而且选取的 K 值一般是 2 的多次幂，以满足底层硬件的数据读取和存储。

2). 学习率调整

学习率 (步长)α 是梯度下降中的超参数，如果取得过大，网络就不会收敛，取得过小，网络就会收敛很慢。为了尽量让参数学习得更加充分，本文将使用学习率衰减

(Learning Rate Decay) 和预热[83](Warmup) 等调节方法。

学习率衰减是为了让网络在前期拥有大的学习率，加快收敛，后期收敛到最优值附近拥有小学习率，精细调整，避免错过和来回振荡。本文采取的衰减方法是 (假设初始学习率为 α_0，第 t 次回合后为 α_t，总回合次数为 T) 阶梯衰减 (Step Decay)，即在 $[t_1, t_2, \cdots, t_m]$ 回合时将学习率相应衰减 $[\beta_1, \beta_2, \cdots, \beta_m]$ 倍。预热是指在刚开始训练的几次迭代中，先采用较小的学习率，然后逐渐线性恢复到原始的学习率，此举是为了防止当初始学习率较大时，训练有不稳定的情况发生。预热后也可以继续选择一种衰减方法逐渐降低学习率，以保证训练的平稳。

3). 梯度自适应估计

梯度的自适应估计是为了调整优化方向，通过采用带动量 (Momentum) 的方式来综合前面一段时间的梯度来决定当前的梯度，给出参数的实际更新方向，保证模型在训练时冲出鞍点或局部最小值点，实现又快又好地收敛。本文将主要采取两种梯度自适应估计方式，即带动量的随机梯度下降优化器和结合了学习率自适应调整和动量法的 Adam[84] 优化器。

动量法是将网络的梯度看作加速度，并根据之前的动量累积 (加权移动平均) 来代替当前的真正梯度，即：

$$\Delta\boldsymbol{\theta}_t = \rho\Delta\boldsymbol{\theta}_{t-1} - \alpha\boldsymbol{g}_t = -\alpha\sum_{\tau=1}^{t}\rho^{t-\tau}\boldsymbol{g}_\tau \tag{3.12}$$

上式中，ρ 表示动量因子，一般取 0.9。在训练初期，参数梯度方向较为一致，动量法可以起到加速作用，训练后期，梯度方向不同，动量法会带来振荡效果，具有一定的减速作用，这样可以帮助模型收敛。

Adam 是一个集大成优化器，不仅可以自适应调整权重的学习率，还可以调整权重的更新方向，其具体计算公式如下：

$$\begin{aligned}
\boldsymbol{M}_t &= \beta_1\boldsymbol{M}_{t-1} + (1-\beta_1)\,\boldsymbol{g}_t \\
\boldsymbol{G}_t &= \beta_2\boldsymbol{G}_{t-1} + (1-\beta_2)\,\boldsymbol{g}_t \odot \boldsymbol{g}_t \\
\hat{\boldsymbol{M}}_t &= \frac{\boldsymbol{M}_t}{1-\beta_1^t} \\
\hat{\boldsymbol{G}}_t &= \frac{\boldsymbol{G}_t}{1-\beta_2^t} \\
\Delta\boldsymbol{\theta}_t &= -\frac{\alpha}{\sqrt{\hat{\boldsymbol{G}}_t + \epsilon}}\hat{\boldsymbol{M}}_t
\end{aligned} \tag{3.13}$$

式 (3.13) 中的 $\boldsymbol{M}_t, \boldsymbol{G}_t$ 可以看作是梯度的一阶矩和二阶矩，β_1, β_2 分别是对应的滑动平均衰减系数，通常取 0.9 和 0.99，$\hat{\boldsymbol{M}}_t, \hat{\boldsymbol{G}}_t$ 是迭代初期对 $\boldsymbol{M}_t, \boldsymbol{G}_t$ 的修正，因为前期 $\boldsymbol{M}_t, \boldsymbol{G}_t$ 的估计偏差很大，α 是初始学习率，默认设置为 0.001。

4). 网络参数初始化

除了选取适当的梯度下降优化算法之外，神经网络的参数初始化也是十分重要的环节。参数初始值选取不好，可能导致网络优化效率变低，甚至无法收敛。考虑到本文

制作的数据集较小，将以预训练初始化 (Pre-trained Initialization) 方法为主，随机初始化 (Random Initialization) 方法为辅，其中随机初始化只采用基于方差缩放 (Variance Scaling) 的 Xavier[85] 初始化和 He[86] 初始化方法。

预训练初始化是指利用在现有大规模数据集 (ImageNet, COCO 等) 上训练好的模型参数来作为自己网络的初始值，这也是迁移学习理论的一种应用，但是这个方法必须在两者模型结构完全一样的前提下才可以使用。随机初始化是为了让网络的神经元之间具有区分性，训练时避免往同一方向优化。利用基于方差缩放的随机初始化方法通过尽量保持网络输入和输出的方差一致，来缓解训练时梯度消失或者爆炸的问题，这种一致性是根据神经元的连接数量来调整。其中对于不同的激活函数，调整的方式也不尽相同，针对 Sigmoid 采用 Xavier 初始化方法，针对 ReLU 采用 He 初始化方法。表 3.2 给出了两种初始化方法的具体设置情况。

表 3.2 Xavier 初始化和 He 初始化的方差设置情况

参数初始化方法	激活函数	高斯分布 $\mathcal{N}(0,\sigma^2)$
Xavier 初始化	Sigmoid	$\sigma^2 = 16 \times \frac{2}{M_{l-1}+M_l}$
He 初始化	ReLU	$\sigma^2 = \frac{2}{M_{l-1}}$

5). 抑制网络过拟合

训练神经网络的主要目的是为了让模型应用在训练集之外的其他同分布数据上，也就是尽量提高网络的泛化性能，实现样本真实分布上的期望风险最小化。然而本文的训练集很有限，其经验风险和期望风险的最小化结果并不完全一致，很容易出现过拟合(同时学习了数据噪声，如图 3.6 所示)，导致模型的泛化能力变差。为此，本文将使用提前停止 (Early Stop)，丢弃法[87](Dropout)，加入正则项 (Regularization)，以及数据增强 (Data Augmentation) 手段增强网络的表现性能。

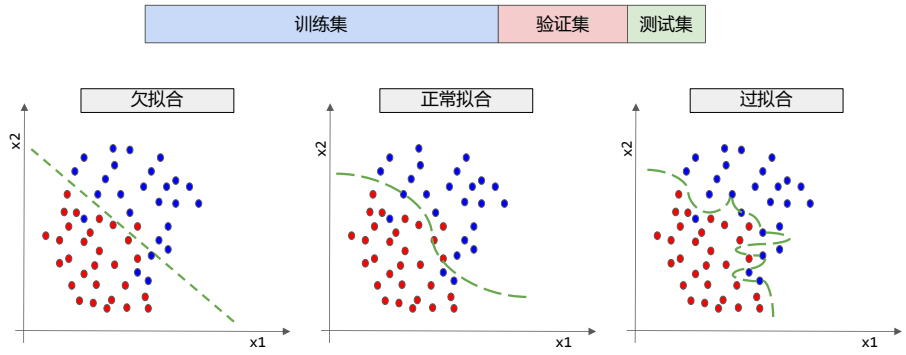
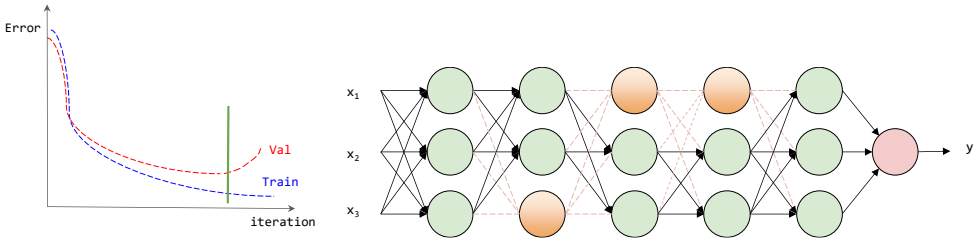


图 3.6 神经网络在训练集上误差较低，但是在验证集和测试集上可能出现欠拟合或者过拟合

提前停止是一种简单有效的人工干预网络优化方法，如图 3.7(a) 所示，一旦出现训练集误差降低，验证集误差上升的情况，表明此刻模型出现了过拟合，泛化能力开始衰减，则立刻手工停止训练。丢弃法是受生物神经元启发，如图 3.7(b) 所示，随机以一定

概率 p 让部分神经元失活来防止过拟合，这种方式可以强制网络中的神经元学习关键的特征，并包含一定的冗余性，同时也加强了网络结构的多样性和鲁棒性。此外使用丢弃法时需要注意保持训练和测试时输出的期望一致，一般在利用丢弃法训练网络时会对输出除以失活概率 p (Inverted Dropout) 来实现。



(a) 提前停止，绿线代表停止位置 (b) 丢弃法，橙色代表随机失活的神经元

图 3.7 提前停止和丢弃法示意图

正则项中最常用的是 ℓ_1 和 ℓ_2 正则项，主要目的是通过约束权重的 ℓ_1 和 ℓ_2 范数来防止部分权重过大主导了网络的输出，抑制过拟合。通过加入上述正则项，神经网络的优化问题可以变为：

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}\left(\boldsymbol{y}^{(n)}, f\left(\boldsymbol{x}^{(n)}; \boldsymbol{\theta}\right)\right) + \lambda \ell_p(\boldsymbol{\theta}) \tag{3.14}$$

上式中，N 代表训练样本 $\boldsymbol{x}$ 的总数量，$\mathcal{L}(\cdot)$ 表示损失函数，$f(\cdot)$ 为神经网络函数，$\boldsymbol{\theta}$ 表示网络待学习的参数，λ 为正则化系数，p 取值为 1 或 2。本文只使用 ℓ_2 正则项。

数据增强是通过一些手段变换训练数据，以扩大数据量和增强数据的学习难度，进而提高模型的鲁棒性和泛化性能。本文的数据增强主要用于图像上，以引入噪声，扩大数据多样性，其采取的方法有归一化，随机翻转，缩放和图像混合。

3.2 卷积神经网络基本模块

在本文的探测任务中，待学习的对象是图像数据，如果只用全连接前馈神经网络来处理，会出现两个问题，一个是图像的每个像素点都代表一个神经元，造成隐藏层的权重参数过多，难以训练，且容易过拟合；另一个是难以提取到图像的局部不变性特征 (比如尺度缩放，平移等不影响图像语义信息)，模型的表征能力不强。卷积神经网络 (Convolutional Neural Network, CNN) 是一种特殊的多层前馈神经网络，具有权重共享，局部连接等特性，可以很好地解决上述两个问题，因此将作为本文处理图像数据的主体模型。

卷积神经网络是根据生物学中的感受野机制 (Recep- tive Field) 提出的。在视觉神经系统中，不是所有视觉皮层的神经元都会接收视网膜光感受器输出的兴奋信号，那些特定区域被激活的神经元就是指神经元的感受野。人工抽象出的卷积神经网络，一般包含卷积层 (Convolutional Layer)，池化层 (Pooling Layer) 和全连接层 (Fully Connected

Layer)，这些模块使得卷积神经网络具有一定程度的平移、缩放等不变性，后来研究者根据实验进一步增补了加速网络训练优化和收敛的归一化层。图 3.8 展示了一个卷积神经网络的典型结构，下面将对其中的模块依次进行介绍。

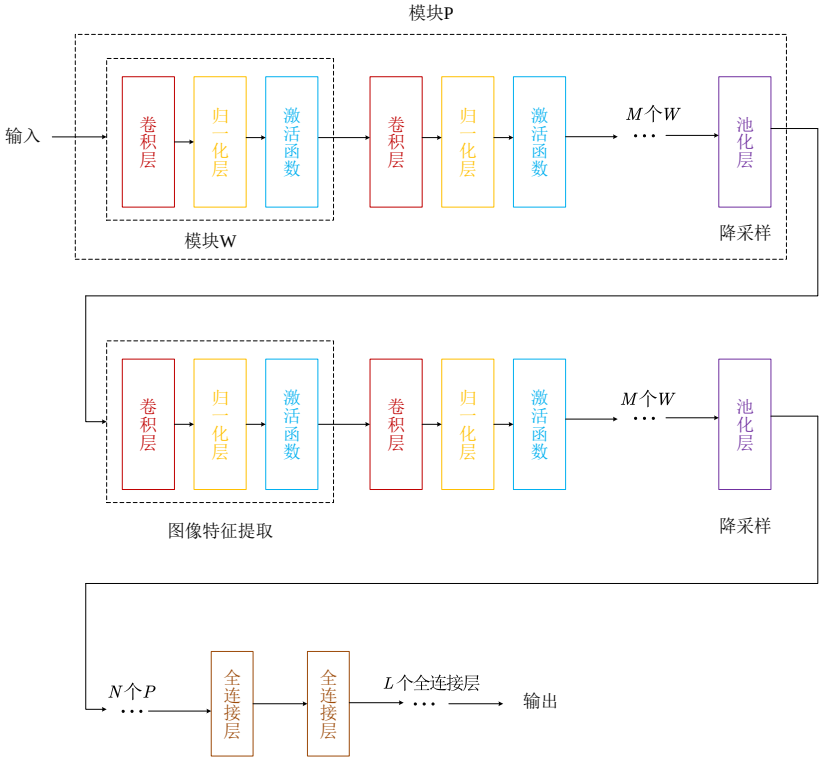


图 3.8 卷积神经网络典型结构示意图

3.2.1 二维卷积

卷积 (Convolution) 是分析数学中的一种重要运算，一维卷积主要用在信号处理中，二维卷积主要用在图像处理上。给定一个图像数据 $\boldsymbol{X} \in \mathbb{R}^{M\times N}$，一个二维卷积核 (有时也叫滤波器)$\boldsymbol{W} \in \mathbb{R}^{U\times V}$ (一般 $U < M, V < N$)，则图像的 $[[U-1:2U-1],[V-1:2V-1]]$ 区域的卷积结果为：

$$y_{ij} = \sum_{u=1}^{U}\sum_{v=1}^{V} w_{uv}x_{i-u+1,j-v+1} \tag{3.15}$$

在卷积神经网络中，二维卷积通常是方形尺寸，此外还附有步长 (Stride) 和零填充 (Zero Padding) 这两个参数。步长是指一个卷积核在图像上计算卷积时滑动到下一个区域的间隔，常用的步长有 1 和 2，当使用 2 时一般是为了在提取特征后实现下采样。零填充是在图像两端的宽高方向进行补零操作，此参数一般配合步长来设置，以控制输出的特征图尺寸。假设输入的特征图尺寸为 $W_i \times H_i$，二维卷积核大小为 $K \times K$，步长为 S，零填充为 P，则输出的特征图尺寸为：

$$\begin{aligned} W_o &= \lfloor (W_i + 2P - K)/S \rfloor + 1 \\ H_o &= \lfloor (H_i + 2P - K)/S \rfloor + 1 \end{aligned} \tag{3.16}$$

式 (3.16) 中，W_o, H_o 代表输出的特征图尺寸，$\lfloor x \rfloor$ 代表对 x 向下取整。从二维卷积的定义可以看出，第 l 层的神经元只与 $l-1$ 层中某个局部窗口的神经元相连，这个窗口的大小就是卷积核的大小，而第 l 层的神经元都共用一个卷积核参数，也就是利用一个卷积核来提取图像中的某一种局部特征，其带来的优点是局部连接和权重共享，既符合图像数据的特点，又减少了计算量。

在卷积神经网络中，二维卷积可以有效地提取出特征，经过不同卷积核的特征映射后，得到不同的特征图 (Feature Map)。通常一个卷积层包含多个卷积核，相当于包含了不同的特征提取器，卷积层中需要学习的参数即为这些卷积核的权重矩阵以及对应的偏置。为了充分利用图像的局部信息，卷积神经网络中神经元是以三维矩阵的形式组织的，大小为 $W \times H \times C$(宽度 × 高度 × 通道/宽度)，如果是彩色图像，输入的通道数就为 3(RGB)，如果是灰度图，输入的通道数就为 1。图 3.9 给出了一个带激活函数的卷积层特征映射示意图。

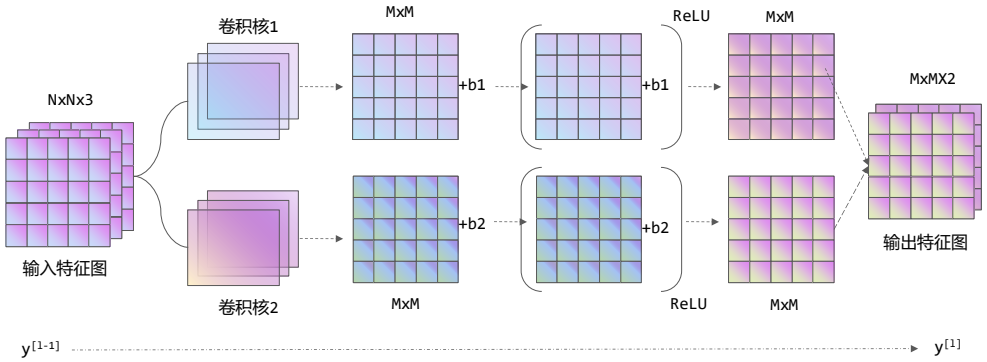


图 3.9 利用卷积核对图像特征进行提取映射

3.2.2 逐层归一化

逐层归一化 (Layer-wise Normalization) 是利用数据归一化手段对神经网络隐藏层的输入进行归一化来使得网络的训练变得稳定和高效。逐层归一化主要在两个方面起作用，一个是保证网络每层的输入稳定，提供一定尺度不变性，便于参数学习；另一个是使输入处于激活函数的不饱和区域，很好地缓解了梯度消失问题，而且还让网络优化函数变得更加平滑，以致可以更快更稳定地收敛[88,89]。本文中使用的归一化方法是批量归一化[90](Batch Normalization, BN)。

批量归一化是在输入数据的批量 (Batch Size) 维度上进行标准正态分布归一化。假设网络第 l 层的净输入为 $\boldsymbol{z}^{(l)}$，样本的批量数为 K，则该归一化方式的定义如式 (3.17) 所示。其中，$\boldsymbol{\mu}_{\mathcal{B}}$ 和 $\boldsymbol{\sigma}^2_{\mathcal{B}}$ 是第 l 层神经元净输入 $\boldsymbol{z}^{(1,l)}, \boldsymbol{z}^{(2,l)}, \cdots, \boldsymbol{z}^{(K,l)}$ 的均值和方差，$\hat{\boldsymbol{z}}^{(l)}$ 是经过批量归一化之后的结果，ϵ 是为了维持数值稳定而设置的一个很小的常数。$\boldsymbol{\gamma}$ 和 $\boldsymbol{\beta}$ 是附加的缩放和平移向量，也是该归一化层需要学习的参数，目的是为了让网络自己决定是否需要进行标准正态分布归一化的操作，如果不需要也可以进行逆变换还原回初始值。值得一提的是，在训练过程中，$\boldsymbol{\mu}_{\mathcal{B}}$ 和 $\boldsymbol{\sigma}^2_{\mathcal{B}}$ 会在每次小批量中单独计算，而预测

时，$\boldsymbol{\mu}_{\mathcal{B}}$ 和 $\boldsymbol{\sigma}_{\mathcal{B}}^2$ 会使用训练时对应的移动平均值来代替。

$$\begin{aligned}\boldsymbol{\mu}_{\mathcal{B}} &= \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{z}^{(k,l)} \\ \boldsymbol{\sigma}_{\mathcal{B}}^2 &= \frac{1}{K}\sum_{k=1}^{K}\left(\boldsymbol{z}^{(k,l)}-\boldsymbol{\mu}_{\mathcal{B}}\right)\odot\left(\boldsymbol{z}^{(k,l)}-\boldsymbol{\mu}_{\mathcal{B}}\right) \\ \hat{\boldsymbol{z}}^{(l)} &= \frac{\boldsymbol{z}^{(l)}-\boldsymbol{\mu}_{\mathcal{B}}}{\sqrt{\boldsymbol{\sigma}_{\mathcal{B}}^2+\epsilon}}\odot\boldsymbol{\gamma}+\boldsymbol{\beta}\end{aligned} \tag{3.17}$$

3.2.3 池化与全连接

池化在卷积神经网络中主要承担着特征选择的作用，从而降低特征图大小，减少网络参数量，避免过拟合，一般也会叫池化层为下采样层 (Subsampling Layer)。假设池化层接收的输入特征图组为 $\mathcal{X}\in\mathbb{R}^{M\times N\times D}$，将其中某个特征图 $\boldsymbol{X}^d\in\mathbb{R}^{M\times N}$ $(1\le d\le D)$ 划分为多个矩形区域 $R_{m,n}^d$ $(1\le m\le M', 1\le n\le N')$，这些区域可以重叠，也可以不重叠 (通过步长来控制)。池化的作用就是对这些区域进行特征选择，并采出一个值作为概括和替代。本文主要使用最大池化 (Max Pooling) 和平均池化 (Average Pooling) 两种，最大池化是选取区域 $R_{m,n}^d$ 中的最大值作为结果，平均池化是计算 $R_{m,n}^d$ 中特征值的平均值作为结果。图 3.10 展示了最大池化的计算过程。

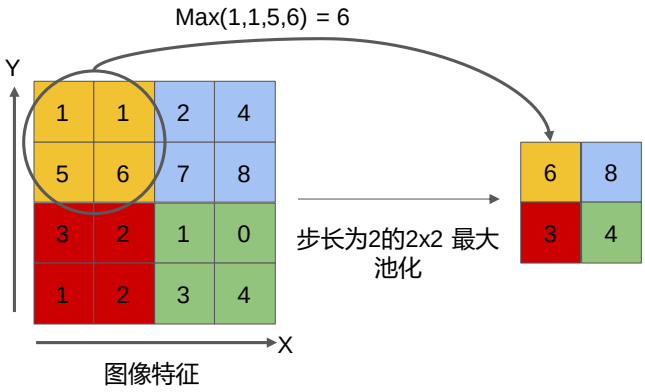


图 3.10 最大池化计算示意图

全连接层往往作为卷积神经网络的最后几层存在，目的是为了综合前面卷积层提取的局部特征，通过几个隐藏层来让网络学习全局特征，并利用相应的激活函数和损失函数，完成分类或回归任务。

3.3 典型的特征提取网络

卷积神经网络是目前应用最广泛的前馈神经网络之一，其训练方法和优化器也可以直接沿用3.1节的理论。本节将介绍几种成熟的卷积神经网络并将其作为探测模型的特征提取骨架，便于后续进行低慢小无人机的识别定位实验。

3.3.1 VGG

VGG[91] 是英国牛津大学视觉几何研究组 (Visual Geometry Group) 在 2014 提出的网络，它利用几个连续的 3×3 的卷积核来代替较大的卷积核 (比如 5×5，7×7 等)，可以在不改变总感受野大小的前提下，增加网络的非线性层，使其具有更强的表征能力，并且减少了参数量，方便网络优化。

目前比较常用的 VGG 主要是 VGG-16 和 VGG-19，两者的模型结构组成如表 3.3 所示，都是由 3×3 的卷积核和 2×2 的最大池化组成，其中每次最大池化都会将特征图尺寸缩小两倍，后面紧跟的卷积层会相应地加倍通道数，以保持网络的复杂度。网络的最后是两个全连接隐藏层和一个最后的输出层，目的是为了学习图像的全局特征，然后进行物体分类。

表 3.3 VGG-16 与 VGG-19 网络结构 (conv3-c 代表输出通道数为 c 的 3×3 大小的卷积核，fc-m 表示神经元数量为 m 的全连接层)

模型 (模块) 名称	VGG-16	VGG-19
模块 1	(conv3-64) × 2 + maxpool	(conv3-64) × 2 + maxpool
模块 2	(conv3-128) × 2 + maxpool	(conv3-128) × 2 + maxpool
模块 3	(conv3-256) × 3 + maxpool	(conv3-256) × 4 + maxpool
模块 4	(conv3-512) × 3 + maxpool	(conv3-512) × 4 + maxpool
模块 5	(conv3-512) × 3 + maxpool	(conv3-512) × 4 + maxpool
模块 6	(fc-4096) × 2	(fc-4096) × 2
输出	fc-1000 + softmax	fc-1000 + softmax

3.3.2 GoogLeNet

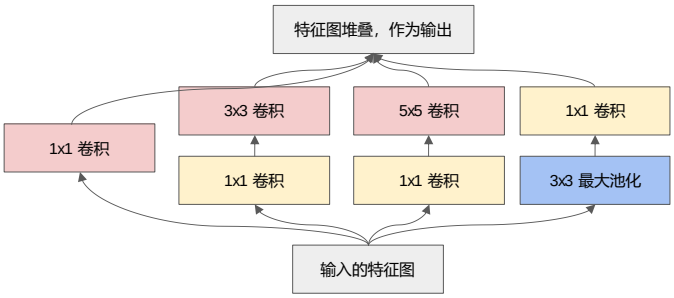


图 3.11 Inception 模块示意图

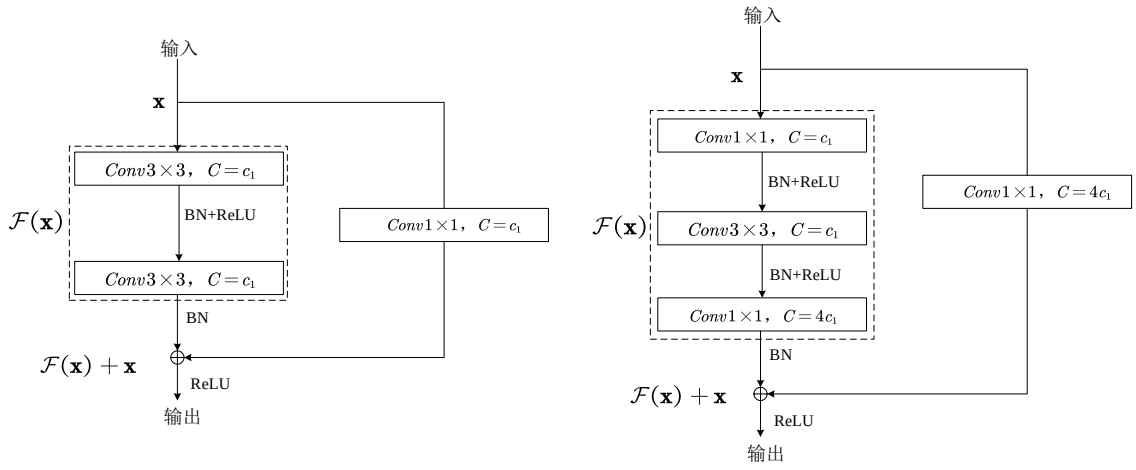
GoogLeNet[92] 是 Google 在 2014 年提出的深层卷积神经网络，它的主要关注点在卷积核大小的设置上，对同一输入使用大小不同的卷积核，可以获得不同具有不同感受野的特征图，增强了模型的特征表达。在 GoogLeNet 中，一个卷积层包含了几个大小不同的卷积核，并对这些不同卷积核操作得到的特征图堆叠起来作为最终的输出，称其

为 Inception 模块。如图 3.11 所示，该模块采用了 1×1，3×3，5×5 的卷积和 3×3 的最大池化这四组特征提取方式来平行操作，为了减少参数量，在 3×3，5×5 的卷积操作之前和 3×3 的最大池化之后使用 1×1 的卷积减少特征深度，剔除冗余信息。GoogLeNet 的另一个改进是使用了全局平均池化[93](Global Average Pooling, GAP)，即一个特征图采样一个代表特征值，单独利用特征图分类，以减少参数量。

3.3.3 ResNet

ResNet[94](Residual Network) 是微软亚洲研究院于 2015 年提出的网络模型，其通过给卷积层输出增加对应输入跨连 (Shortcut Connection) 相加的方式提高了网络信息流动的有效性，也让深层，甚至是超深层网络训练变得容易。ResNet 认为，如果网络加深的只是恒等映射层，那么最后的模型至少不会比没加深之前差，但是直接让网络来学习恒等映射也增加了其学习的负担，因此利用跨连，连接网络的输入和输出，假设该层的输入为 x，映射函数为 $\mathcal{F}(\mathrm{x})$，则输出为 $\mathcal{F}(\mathrm{x})+\mathrm{x}$，这样只需要该层学习恒等映射和输出之间的残差即可，不会进一步带来网络优化的负担。因此，也把 ResNet 的输入跨连称为残差连接 (Residual Connection)。

图 3.12 展示了 ResNet 中的典型残差单元，ResNet 就是将很多个这样的残差块串联起来的深层网络，其中 (b) 是在 (a) 上做了多层小卷积核改进，目的是为了减少参数量，防止 (超) 深层网络出现过拟合。图中的跨连输入其实不是严格意义上的净输入，而是加了一个卷积层，以应对特征图相加时维度不一致的情况。残差模块是 ResNet 最主要的贡献点，一是让梯度传导更加高效，由于残差的数值范围小，导致模型待学习参数对反向传播的误差有更敏感的响应能力，同时保持了梯度的相关性；二是残差模块可以根据训练数据集动态自适应分配实际网络深度，从几何角度看就是制造了一个更为平直的流形并在其上进行优化，导致系统的收敛性更好。表 3.4 给出了本文所使用的的 ResNet-101 网络结构配置。



(a) BasicBlock，ResNet-18和ResNet-34所用的残差模块　　(b) Bottleneck，ResNet-50, ResNet-101和ResNet-152所用的残差模块

图 3.12 ResNet 残差块示意图 (C 代表通道数)，不同深度的网络会使用不同的残差块来构建

表 3.4 ResNet-101 网络结构 (假设原图尺寸为 224×224, c 代表通道数，s 代表步长)

层组名	特征图输出大小	ResNet-101
Conv1	112×112	conv 7×7, c=64, s=2
Conv2_x	56×56	[maxpooling 3×3, s=2] + $\begin{pmatrix} 1 \times 1, c = 64 \\ 3 \times 3, c = 64 \\ 1 \times 1, c = 256 \end{pmatrix} \times 3$
Conv3_x	28×28	$\begin{pmatrix} 1 \times 1, c = 128, s = 2 \\ 3 \times 3, c = 128 \\ 1 \times 1, c = 512 \end{pmatrix} \times 4$
Conv4_x	14×14	$\begin{pmatrix} 1 \times 1, c = 256, s = 2 \\ 3 \times 3, c = 256 \\ 1 \times 1, c = 1024 \end{pmatrix} \times 23$
Conv5_x	7×7	$\begin{pmatrix} 1 \times 1, c = 512, s = 2 \\ 3 \times 3, c = 512 \\ 1 \times 1, c = 2048 \end{pmatrix} \times 3$
FC	1	fc-1000

3.4 本章小结

本章主要对深度学习的相关技术理论做了研究。首先介绍了人工神经网络的基础内容，阐述了人工神经元模型，非线性激活函数以及前馈神经网络的结构，并根据本文数据集的特点引入了对神经网络的训练与优化、泛化方法。然后研究了前馈神经网络中的卷积神经网络，并具体分析了其结构和相关模块。最后给出了几个典型的卷积神经网络模型，为后续的无人机检测提供特征提取骨架 (Backbone) 参考。

4 基于规则区域识别的视觉探测无人机算法设计与实现

如第一章所述，传统的目标检测算法具有很大的局限性，且性能不高，无法满足低慢小无人机的检测要求。随着深度学习的发展，目标检测算法的精度和通用性都得到了大大提高，本文也主要采用基于深度学习的方法来实现无人机及其他干扰物体的检测。现阶段该领域的目标检测算法主要有基于锚框 (Anchor-Based) 和无锚框 (Anhor-Free) 两种，前者又可分为二阶检测 (Two-Stage) 和一阶检测 (One-Stage)，且发展都较为成熟，性能也比较稳定；后者的检测思路较为直观简洁，是当前的研究热点，但是性能还有待提高。考虑到低慢小无人机的出现场景复杂多变，对检测准确度的要求也较高，因此本文选择基于锚框的检测方法进行研究和实现。本章将首先介绍二阶、一阶目标检测的代表性模型 Faster R-CNN 和 SSD，并在无人机数据集上进行基准 (Baseline) 实验，接着分别在可见光图片和红外热成像图片上改进相关算法，以实现对低慢小无人机的检测，最后对测试结果进行展示和分析，总结本章。

4.1 基于锚框的目标检测方法

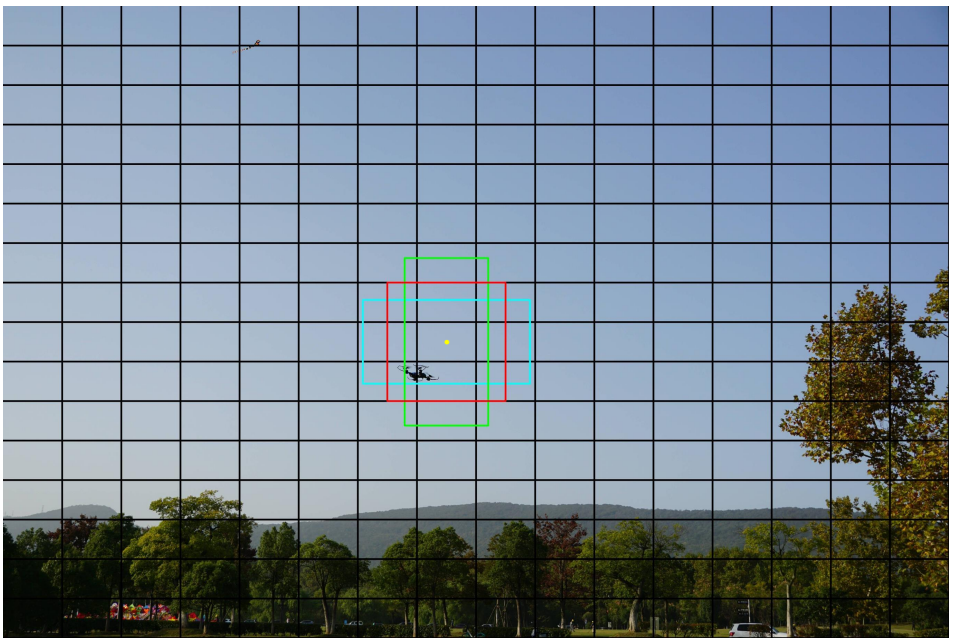


图 4.1 锚框生成示意图 (黄色点为锚框中心，红、蓝、绿框代表生成的不同宽高比的锚框)

在基于锚框的目标检测方法中，锚框机制是用来在特征图上以每个像素点为中心，生成一系列具有不同大小和宽高比的预选框。其受传统目标检测方法中滑动金字塔的启发而来，可以作为端到端检测模型的一个模块存在，具有适应目标多尺度变化，覆盖多物体 (尤其是针对重叠物体) 的特点，同时也不会引入过多的计算量。此外，锚框机制也保证了检测模型对目标的平移不变性，也就是说生成的建议区域框会跟着目标在

图像中的变化而相应变化，这使得任意位置的物体都会被检测出来。图 4.1 给出了锚框的生成示意图，假设以某个特征图来生成锚框，并人为设置锚框基础尺寸 w，一系列缩放比系数 $[s_1, \cdots, s_i, \cdots]$ 和宽高比 $[r_1, \cdots, s_i, \cdots]$，其中前两者决定了锚框的大小 (面积)，后者决定了锚框的形状，然后以每个像素为中心点，生成宽为 $w \times s_i \times \sqrt{r_i}$，高为 $w \times s/\sqrt{r_i}$ 的锚框，以此作为物体预选框。训练时，模型会计算这些生成的锚框与真值的交并比 (IoU)，并根据设定的阈值划分训练的正负样本，预测时则直接根据模型的分类与回归结果来对锚框进行筛选和变换，得到检测结果。

4.1.1 二阶目标检测模型

二阶目标检测模型的主要流程是先通过锚框机制产生大量的建议区域，然后先对这些区域进行筛选，前景/背景的二分类，并进行第一次区域框的回归，得到一系列建议区域，接着再对这些建议区域进行筛选，精细的多类别分类和第二次框回归，得到最终的检测结果。由于该算法包含了两次框回归，以及建议区域的抠取机制，因此被称为二阶算法。下面介绍二阶检测算法中的代表模型 Faster R-CNN，其结构如图 4.2 所示。

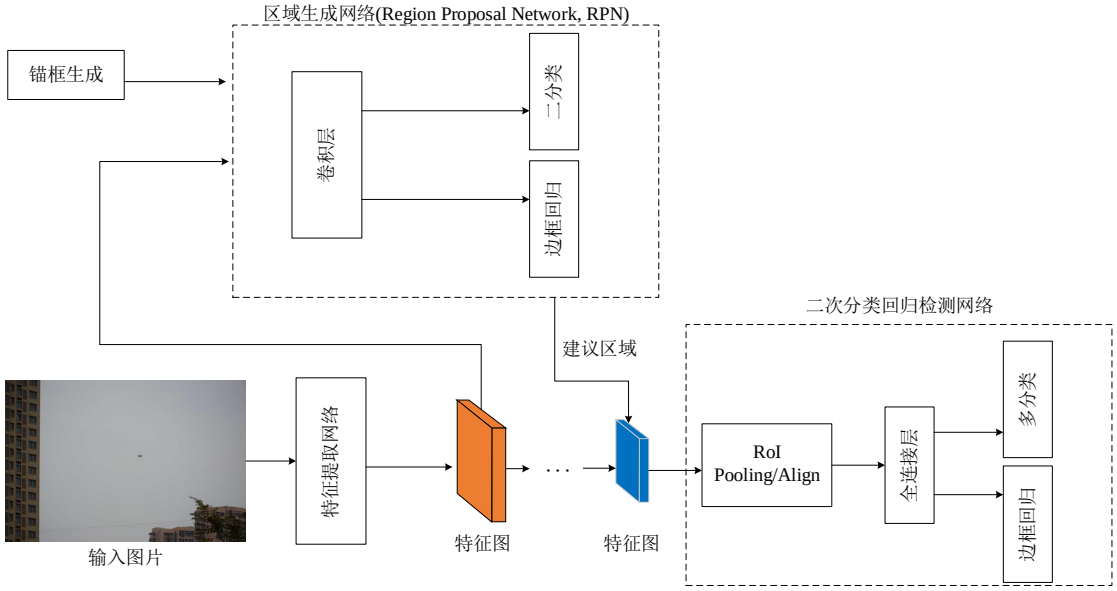


图 4.2 Faster R-CNN 检测模型结构示意图

根据图 4.2，Faster R-CNN 主要由特征提取网络，区域生成网络 (Region Proposal Network, RPN) 和二次分类回归检测网络组成。特征提取网络 (Backbone) 可以采用第二章3.3节中的 VGG-16、ResNet-101 等，以产生高层的特征图供物体分类和定位。假设 Faster R-CNN 以 ResNet-101 来对输入图像进行特征提取和降采样，当特征图缩小为原图的 16 倍后，将其作为区域生成网络的输入。模型会以该特征图的每个像素为中心，以该特征图下采样倍数即 16 为基础尺寸，$[8, 16, 32]$ 为缩放比系数，$[0.5, 1, 2]$ 为宽高比来生成锚框，如果输入原图的宽高大小为 800×600，则会生成 $800/16 \times 600/16 \times 9 = 17100$ 个锚框，基本覆盖了原图的所有区域。为了减少参数量，区域提取网络被设计成全卷积神经网络，其具体结构如图 4.3 所示。它首先对传入的特征图使用一个 3×3 的卷积降

低通道数，并进一步实现特征抽象，然后分别利用两个 1×1 的卷积实现锚框的二分类和框粗回归。在训练时，根据锚框与真值之间的交并比值会将它们划分为正样本，负样本和忽略样本，其中正样本锚框的交并比值在 [0.7,1] 之间，负样本的值在 [0,0.3] 之间，忽略样本的值在 (0.3,0.7) 之间，属于正样本或负样本的锚框参与区域生成网络分类器的训练，正样本锚框参与回归器训练，作为损失函数的计算对象。预测时，区域生成网络会给每个锚框一个属于前景类别的置信度，取前 2000 个得分最高的锚框进行非极大抑制 (Non-Maximum Suppressio, NMS) 去重叠度高的样本，然后再取其中前 1000 个锚框进入回归器进行第一次回归。

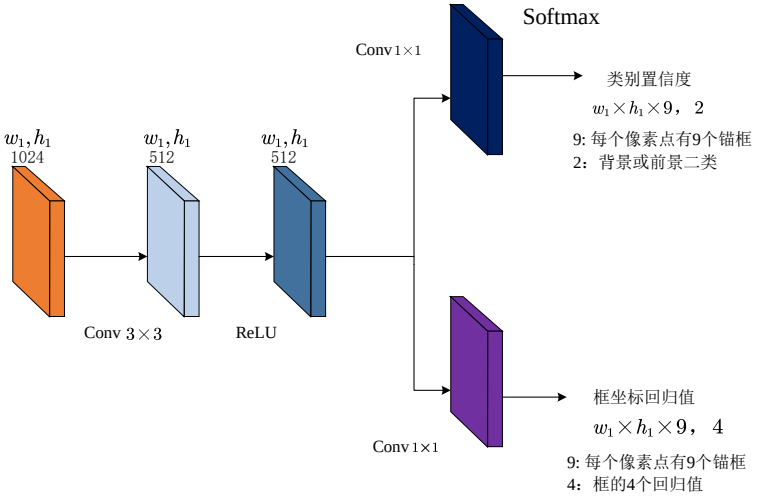


图 4.3 区域生成网络 (RPN) 结构示意图 (假设传入特征图的尺寸为 $w_1 \times h_1$，数字为特征图通道数)

锚框经过区域生成网络之后得到建议区域，然后通过二次分类回归检测网络实现区域精细的多类别分类和框的进一步回归修正，得到最后的检测结果。二次分类回归检测网络的核心是从特征图中抠取建议区域的机制，原论文是使用 RoI(Region of Interest) Pooling，后来经过 Mask R-CNN[50] 的改进后，现在一般用 RoI Align 代替，这两种方法的工作原理如图 4.4 所示。对于建议区域，先通过计算感受野 (Receptive Field) 大小得出其在待抠取特征图上的特征区域，RoI Pooling 会将该区域进行像素取整对齐，然后划分成固定的小区域，一般是 7×7，并对小区域的坐标进行第二次取整，最后对每个小区域进行最大池化，以最大的特征值作为替代。而 RoI Align 为了提高特征区域的抠取精度，放弃了两次取整的操作，直接平均划分，并且利用双线性插值来得到待最大池化区域的特征值，在后面有关的实验中，本文也将采取该方法。RoI Pooling / Align 的作用是为了让不同的建议区域都能得到同样大小的输出，以此对应后面全连接层的维度，而且也顺带缓解了两次回归上特征不对齐的问题。最后，不同于区域生成网络，二次分类回归检测网络使用两个全连接层来进行多分类和坐标回归，全连接层在这里的作用是整合局部特征，利用全局特征来得到更为精细的结果。

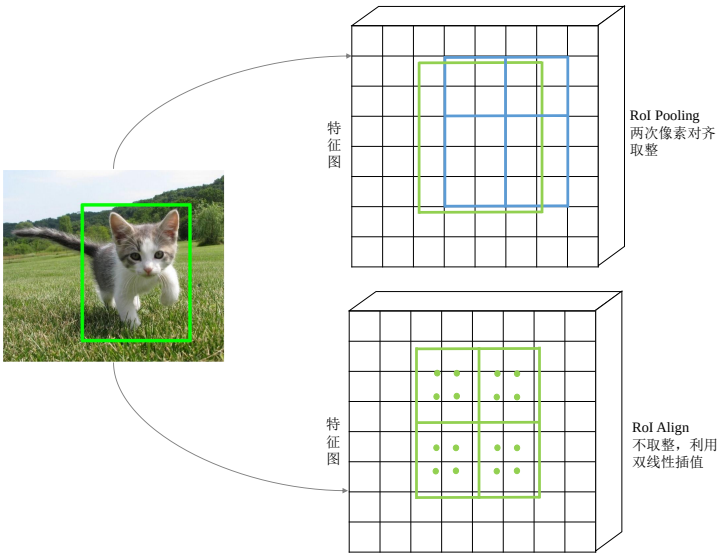


图 4.4 RoI Pooling 与 RoI Align 工作原理示意图

Faster R-CNN 的损失函数主要由两次修正中的分类损失和回归损失组成，以区域生成网络为例，其损失函数如下 (二次分类回归检测网络与其相同)：

$$L\left(\{p_i\},\{t_i\}\right)=\frac{1}{N_{cls}}\sum_i L_{cls}\left(p_i,p_i^*\right)+\lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}\left(t_i,t_i^*\right) \tag{4.1}$$

上式中，$\frac{1}{N_{cls}},\frac{1}{N_{reg}},\lambda$ 分别是小批量中正负样本总和 (一般是 256)，训练过程中正样本的数量 (一般大约是 2400)，分类与回归损失之间的权重系数，其中前两个是作为归一化系数，最后一个是为了平衡两个任务的学习强度，但是影响不大，一般取 10 即可。i 代表样本索引，p_i 代表网络预测的锚框类别置信度，p_i^* 代表该锚框所分配类别的置信度，在区域生成网络中，正样本即为 1，负样本即为 0，t_i,t_i^* 则分别代表网络预测的框回归值和真实值。

值得一提的是，Faster R-CNN 不是直接学习预测框的坐标点，而是锚框到真实框的变换值 (平移和缩放)，是一种经过编码的回归值，这样更利于网络训练，不会使损失值来回振荡。设真实框的中心点和宽高分别为 x,y,w,h，锚框或建议区域的中心点和宽高分别为 x_a,y_a,w_a,h_a，则四个真实回归值为：

$$\begin{aligned}
t_x^* &= (x-x_a)/w_a \\
t_y^* &= (y-y_a)/h_a \\
t_w^* &= \log(w/w_a) \\
t_h^* &= \log(h/h_a)
\end{aligned} \tag{4.2}$$

当得到网络预测的四个回归值 t_x,t_y,t_w,t_h 后，可以通过对式 (4.2) 进行反推得到修正后的矩形框。

$$L_{CE}(y_i,p_i)=-\sum_{i=0}^{C-1} y_i\log\left(p_i\right) \tag{4.3}$$

式 (4.1) 中的 $L_{cls}(\cdot)$ 代表分类损失函数，一般为交叉熵 (Cross-Entropy) 损失函数，计算公式如式 (4.3) 所示。其中，C 为类别数，p_i 代表样本属于第 i 类的置信度，y_i 为样本标签值，当该样本属于第 i 类时为 1，否则为 0。$L_{reg}(\cdot)$ 为回归损失函数，一般为 Smooth L1 损失函数，计算公式如下：

$$L_{\text{Smooth L1}}(x) = \begin{cases} 0.5x^2, & |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases} \tag{4.4}$$

上式中，x 为回归值。当 x 接近于 0 时，说明预测值与真值越来越接近，此时导数也会慢慢降低，减缓更新差值，否则以固定的更新速率来让网络往真值收敛，保证训练的效率。

4.1.2 一阶目标检测模型

一阶目标检测模型中的代表是 SSD(Single Shot MultiBox Detector)，其同样使用了锚框机制，但是只在特征图上进行了一次回归修正，并没有涉及到预选区域的抠取机制。相较于 Faster R-CNN，它的检测精度存在一定差距，但是推理速度较快，而且引入多尺度检测的概念，即用高层特征图检测大目标，低层特征图检测小目标，加强了模型对尺度变化的鲁棒性，启发了后续检测模型的设计。

图 4.5 给出了 SSD 的结构图，假设特征提取网络为 VGG-16，原图经过一系列卷积和池化操作，得到了具有不同尺寸和表征能力的特征图。其中浅层特征图尺寸大，图像细节信息丰富，但是语义信息较少；深层特征图尺寸小，图像细节信息丢失严重，但是语义信息丰富，因此这些特征图组成了单向的特征金字塔。SSD 类似于 Faster R-CNN 的区域生成网络 (RPN)，在这些不同层级的特征图上分配不同的锚框，利用 3×3 的卷积来进行类别和回归预测。

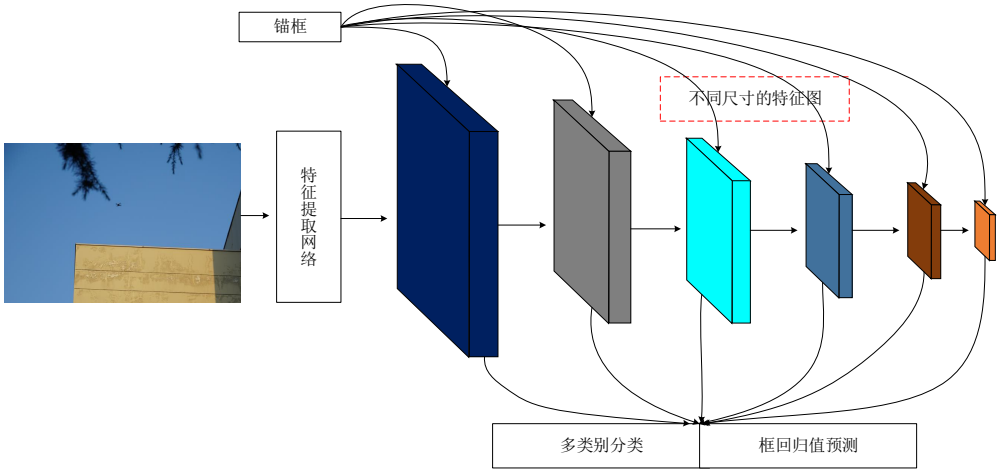


图 4.5 SSD 模型结构示意图

小物体在图像中占据的像素比较少，经过不断的特征提取和池化操作后，位置、形状等细节信息便会逐渐丢失，因此在高层的特征图上，小目标的检测效果会比较差。在

SSD 中，为了试图解决这个问题，专门拿浅层的特征图来检测小物体，拿具有大感受野的高层特征图检测大物体，实现分而治之。这种单向特征金字塔机制受传统检测方法中的图像金字塔启发而来，同时只利用一次图像特征前向计算，去除了重复的计算量，以加快处理速度。类似地，每个特征图上也分配了不同大小的锚框，浅层特征图使用小尺寸，高层特征图使用大尺寸，随着特征图大小降低，锚框的尺寸线性增加，以更好地给出参考区域，其具体分配方法如下：

$$s_k = s_{\min} + \frac{s_{\max} - s_{\min}}{m-1}(k-1), k \in [1, m] \tag{4.5}$$

式 (4.5) 中 (以特征提取网络 VGG-16 为例)，s_k 代表锚框相对于原图大小的比例，s_{min} 和 s_{max} 表示比例的最小值和最大值，分别为 0.2 与 0.9，m 代表用来检测的特征图的个数，这里为 5，第一个特征图的比例单独设置，一般为 $s_{min}/2$，即 0.1。对于宽高比，一般是 $[1, 2, 3, \frac{1}{2}, \frac{1}{3}]$，此外每个特征图还会分配一个宽高比为 1，但是比例为 $s_k^{'} = \sqrt{s_k \times s_{k+1}}$ 的特殊锚框，也就说每个特征图的每个像素都会分配 6 种不同形状的锚框，且锚框的大小随特征图缩小而变大。但是在实际使用中，第一个特征图和最后两个特征图都只使用其中的四个锚框，放弃宽高比为 $3, \frac{1}{3}$ 的锚框。

训练时，SSD 中锚框正负样本认定的交并比阈值与 Faster R-CNN 不同，SSD 直接使用 0.5 来将锚框划分为正样本和负样本，不再考虑忽略类。每个特征图的预测结果都包含 $k \times (C+4)$ 维通道，其中 k 表示锚框数，C 表示类别数 (背景也为一类)，4 表示 4 个回归值 (经过与 Faster R-CNN 中相同的编码)，也就是说输出的是特征图像素中每个锚框的类别置信度和回归值，最后再对这些所有的预测进行非极大抑制得到检测结果。

SSD 的损失函数也包括分类和回归两部分，大体上与 Faster R-CNN 中的区域生成网络差别不大，具体为：

$$L\left(\{p_i\}, \{t_i\}\right) = \frac{1}{N}\left(L_{cls}\left(p_i, p_i^*\right) + \alpha L_{reg}\left(t_i, t_i^*\right)\right) \tag{4.6}$$

上式中，i 代表小批量训练样本中的锚框索引，N 为正样本的数量，p_i, p_i^* 代表模型预测的类别置信度和真值，t_i, t_i^* 代表模型预测的回归值和真实偏移值。分类损失函数 $L_{cls}(\cdot)$ 使用交叉熵损失函数，回归损失函数 $L_{reg}(\cdot)$ 采用 Smooth L1 损失函数，皆与 Faster R-CNN 相同，两个损失函数的权重配比系数 α 采用 1，即等比例贡献损失。此外 SSD 中还使用了困难负例样本挖掘 (Hard Negative Mining) 技术，通过选取置信度误差较大的 top-k 个负样本来控制正负样本比例在 1：3 左右，防止训练过程被大量的负样本主导。

4.2 面向可见光图像的低慢小无人机检测算法

4.2.1 可见光无人机数据集基准实验

为了直观了解二阶、一阶检测模型在低慢小无人机数据集上的性能，使用 Faster R-CNN 和 SSD 在 RGB 数据集上进行基准实验，具体实验环境的配置如表 4.1 所示。此

外，本文所有的模型都将在 PyTorch[95] 下进行编写和训练，PyTorch 是由美国 Facebook 公司开发的基于 Python 语言的深度学习开源库，底层由 C++ 实现，可以提供多维张量计算，自动微分与反向传播机制，以及 GPU 加速等，非常适合进行基于深度学习相关算法的开发和实现。

表 4.1 实验环境相关配置

配置参数	型号/版本
操作系统	Ubuntu 18.04
编程语言	Python 3.6
CPU	Intel(R) Xeon(R) W-2145 CPU @ 3.7GHz
GPU	GeForce RTX 2080Ti
内存	32G
深度学习框架	PyTorch 1.3

基准实验在第二章采集的可见光数据集上进行，其中训练集 12366 张，测试集 3293 张，由于数据较少，训练时直接采用测试集每隔 1 个回合 (Epoch) 进行精度验证，以监视是否有过拟合发生，最终实验结果如表 4.1 和图 4.6，图 4.7 所示。

表 4.2 基准模型在可见光数据集上的实验结果 (AP%)

模型	AP_{drone}	AP_{kite}	AP_{bird}	mAP	FPS
FCOS	27.9	14.5	0.3	14.2	18.9
SSD(VGG-16)	60.3	22.8	1.4	28.2	**37.8**
Faster R-CNN(VGG-16)	65.2	31.0	7.8	34.7	23.1
Faster RCNN(ResNet-101)	**66.5**	**33.4**	**11.2**	**37.0**	22.6

各个模型的具体参数配置如下：

(1) 选用无锚框检测模型 FCOS，测试其在无人机可见光数据集上的性能，并与基于锚框的检测算法比较，来验证锚框机制对本文无人机数据集的有效性。FCOS 模型的特征提取网络使用原论文中的 ResNet-50，并以在 ImageNet 上预训练过的该网络参数进行初始化，其他层的参数使用表3.2中的 He 初始化方法。优化器选用带动量的随机梯度下降法，初始学习率为 0.001，动量值为 0.9，ℓ_2 正则化系数选用 0.0001，小批量数为 4，训练回合数为 20，其他参数和设置与原论文相同。

(2) 选用基于锚框的一阶检测模型 SSD，其特征提取网络使用原论文中的 VGG-16，并且也用在 ImageNet 上预训练过的进行参数初始化，其他卷积层使用 He 初始化方法。锚框宽高比设置上，为了适应本文的数据集，仅使用 $[0.5, 1, 2, 3]$。优化器选用带动量的随机梯度下降法，初始学习率为 0.000125，动量值为 0.9，ℓ_2 正则化系数为 0.0005，小批量数为 8，训练回合数为 20，其他参数和设置与原论文相同。

(3) 选用基于锚框的二阶检测模型 Faster R-CNN，其特征提取网络使用 VGG-16，参数初始化上与 (2) 相同。根据数据集中物体的尺寸分布，锚框的基础尺寸使用 $[4, 8, 16]$，而锚框的宽高比和缩放系数使用原论文的设置，即 $[0.5, 1, 2]$ 和 16。二次分类回归检测网络中使用 RoI Align，优化器使用带动量的随机梯度下降，初始学习率为 0.01，动量值为 0.9，学习率阶梯衰减系数为 0.1，每隔 5 个回合进行衰减，ℓ_2 正则化系数为 0.0005，小批量数为 2，回合数为 20，其他参数和设置与原论文相同。

(4) 将 (3) 中的特征提取网络换为在 ImageNet 上预训练过的 ResNet-101，初始学习率设为 0.001，ℓ_2 正则化系数设为 0.0001，其余与 (3) 相同。

图 4.6 SSD(VGG-16) 与 Faster R-CNN(ResNet-101) 检测效果图 (图中分别以 S 和 F 代替)

根据表 4.2 给出的 4 种模型配置在可见光低慢小无人机数据集上的检测精度结果可知，FCOS 的检测精度最低，由于它没有利用锚框提供先验，而是通过关键点来预测检测框，对于像素分布较少的小目标来说，很容易被漏检。SSD 相比 FCOS 提升很大，说明锚框机制对小目标无人机的检测很有帮助，但是像鸟类这样的极小目标，或者说是样本数量很少的目标，检测性能提升不是很明显，容易被大量的负样本掩盖，难以学习到。Faster R-CNN 相比 SSD，检测精度进一步提升，尤其在风筝和鸟类上，说明二次回归和区域抠取机制加强了回归框的质量，在一定程度上缓解了样本不均衡的问题。此外，将 Faster R-CNN 的特征提取网络换成更深更加高效的 ResNet-101 后，模型的特

征提取能力和表征能力进一步加强，也对模型的检测性能带来了一定的提升。在检测速度上，一阶检测算法由于结构简单，推理速度最快，二阶检测算法 Faster R-CNN 的两次修正操作使处理速度较 SSD 慢了一倍，但是在处理一定距离的低慢小无人机监测上，依然具有一定的实时性。

图 4.6 给出了 SSD(VGG-16) 和 Faster R-CNN(ResNet-101) 的部分测试结果图，从中可以看出，Faster R-CNN 的漏检率明显比 SSD 低，尤其是面对复杂背景的情况下，而且 Faster R-CNN 的检测框也更贴合实际。由于本文重点收集了低慢小无人机可见光图像，其尺度变化大，背景也十分复杂，因此决定以 Faster R-CNN 为基础，并做出一些改进来实现 RGB 数据集上的检测。

本小节将以 ResNet-101 为特征提取网络的 Faster R-CNN 为基础，在 RGB 低慢小无人机数据集上设计合适的检测算法。原始的 Faster R-CNN 在该数据集上存在一些误检和漏检情况，比如将风筝，水中的鸭子，楼宇等误检成无人机，当目标很小或者较大时，算法也无法很好适应，出现漏检或者检测偏差，具体如图 4.7 所示。出现这些问题的原因主要有两个，一个是无人机的尺度变化范围大 (最小像素为 10，最大像素为 193626)，而且无人机的型号，角度也各不相同；另一个则是无人机出现的背景较为复杂，光照、遮挡、以及相似物体的干扰等都会带来影响。下面针对这些情况，对算法做出相应改进。

图 4.7 Faster R-CNN 面对无人机多尺度变化和复杂背景时出现漏检和误检

4.2.2 针对多尺度问题的改进

1). 特征金字塔结构

在传统目标检测算法中，通过将原图缩放成一系列不同大小的图片来构造图像金字塔，以解决目标的多尺度问题，但是这种机制用在深度学习里会带来额外的开销，每次缩放都要重新计算特征。本章开头所说的锚框机制也是一种应对多尺度的手段，但是像 Faster R-CNN 这样只用了一个层级的特征图，会导致某些物体，尤其是小物体的信息在下采样时丢失。SSD 虽然也给出了一种金字塔检测的概念，不过由于是单向结构，浅层的特征图即使保留了小目标的信息却因为特征表达能力不够，也会导致检测效果不理想。为此，Lin[96] 等提出了一种特殊的特征金字塔结构 (Feature Pyramid Network, FPN)，其主要包括三个部分：(1) 自下而上的路径，通过卷积对图像进行特征提取，并不断下采样特征图；(2) 自上而下的路径，将 (1) 的特征图进行上采样，不断加倍其尺度；(3) 侧向相加的路径，将 (1) 和 (2) 每个对应尺度的特征图相加。自下而上是为了获取不同层级的特征图，浅层的细节丰富，深层的特征抽象程度高，语义信息强；自下而上是为了让每个尺度的特征图都具有较强的特征表达能力；侧向连接是为了补充上采样过程中的信息，防止特征和细节过于稀疏。图 4.8 展示了特征金字塔的具体结构，双向结合的设计可以使得多特征图检测的效果更加可靠和稳定。值得一提的是，上采样通过最近邻插值来实现，侧向相加的左值需要经过一个 1×1 的卷积层减少通道数以对齐右值的维度，最后相加的结果还要经过一个 3×3 的卷积来处理混叠效应，提取出有用的信息。

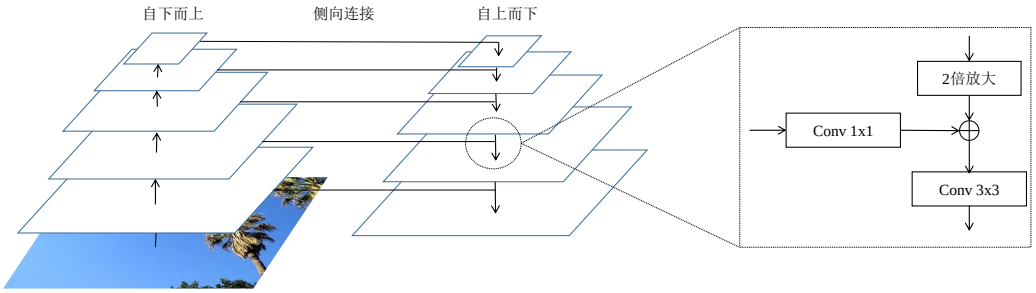


图 4.8 特征金字塔结构示意图

为了实现多尺度物体检测的分而治之，在小特征图里找大目标，在大特征图里找小目标，图 4.9 给出了使用该金字塔结构的 Faster R-CNN 模型示意图。特征提取网络采用 ResNet-101，根据表3.4，以 Conv2_x, Conv3_x, Conv4_x, Conv5_x 层组输出的特征图为自下而上特征图，以 P2, P3, P4, P5 为自上而下的特征图，它们对应原图分别缩小了 4，8，16，32 倍，其中对 P5 使用最大池化进行了 2 倍下采样，得到了 P6。区域生成网络在 P2, P3, P4, P5, P6 这每个特征图里各生成一个固定大小锚框，为了照顾到多尺度物体，本小节在 RGB 数据集中设 4 为锚框基础尺寸，$[4, 8, 16, 32, 64]$ 为锚框的缩放比系数，这五个尺度分别对应 P2, P3, P4, P5, P6，在二次分类回归检测网络中，仅在 P2, P3, P4, P5 中进行 RoI Align，以抠取建议区域的特征，特征图选择对应公式如下：

$$k = \left\lfloor k_0 + \log_2(\sqrt{wh}/L) \right\rfloor \tag{4.7}$$

上式中，L 是模型的图像矩阵输入大小，w,h 分别对应建议区域的宽高，$k_0=5$ 作为基准值，k 即为区域对应进行后续特征抠取的特征图索引。

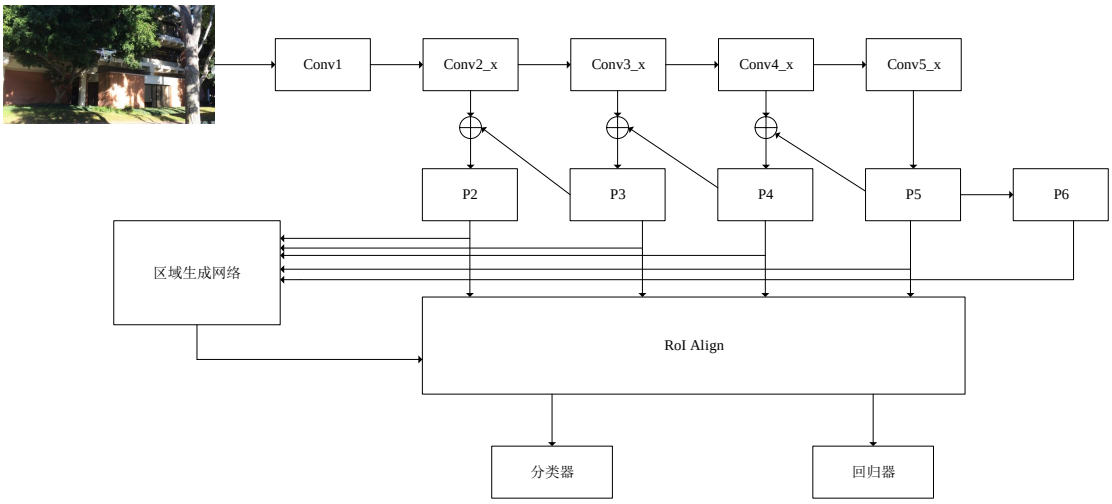


图 4.9 使用特征金字塔结构的 Faster R-CNN 模型示意图

2). 可变形卷积

在应对物体形状变化方面，传统的目标检测算法使用一些鲁棒性比较好的手工特征，比如 SIFT[97] 和 SURF[98]，但毕竟受制于人工设计，场景有限且处理速度较慢。在深度学习中，模型利用卷积核来自主提取图像特征，同时在训练时也会选择一些数据增强技巧辅助网络来容忍几何图形变化，但是这些手段“治标不治本”，只能在一定程度上缓解。此外，常用的卷积核都是矩形形状，在特征计算时只对固定的位置进行采样，因此从本质上来说，也无法很好适应几何形变。

目前一个很好的解决思路是 Dai 等人提出的可变形卷积[99]。如图 4.10 所示，其基本思想非常简洁朴素，就是给常规的卷积核引入偏移量，来打破规则区域采样的限制。可变形卷积中的偏移量是指卷积核中的每个权重对应采样的特征像素的偏移量，一般是沿宽高上的 2 维偏移，并且可以通过反向传播进行自适应学习，不需要人为提前设定。可变形卷积的计算流程如下：

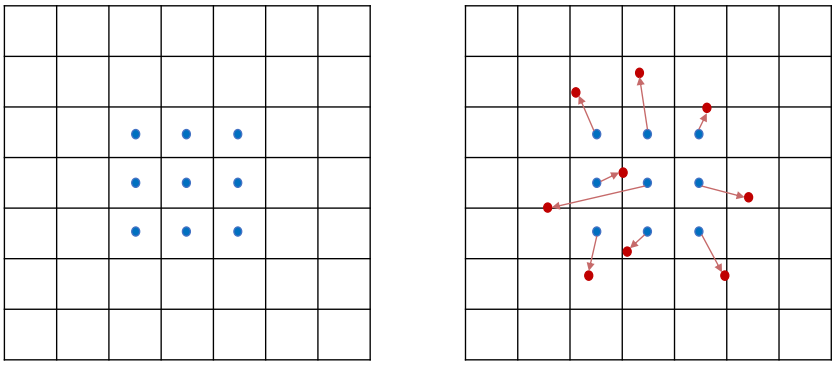
(1) 对于一个输入的特征图 (假设其维度为 (N,C_1,h_1,w_1)，N 为小批量数，C_1 为通道数，h_1,w_1 分别为该特征图的高和宽，原始的常规卷积核大小为 3×3)，首先让该特征图经过一个卷积，得到一个维度为 $(N,18,h_1,w_1)$ 的偏移图，其中 $18=2\times3\times3$，表示常规 3×3 卷积核每个权重对应原始采样像素的偏移量，前九个通道表示 x(宽方向) 偏移量，后九个表示 y(高方向) 偏移量；

(2) 由于 (1) 中得出的偏移量是浮点型小数，因此得到的偏移后的像素点没有具体数值，需要利用临近的四个像素点进行双线性插值，以此得到准确的采样像素值，同时也便于反向传播；

(3) 再利用常规的 3×3 卷积核对偏移后的像素点进行特征计算，合并后最终得到

经过了可变形卷积的特征图，该输出特征图又可以作为新的输入进入下一轮常规卷积或者可变形卷积。

图 4.11 给出了 3×3 可变形卷积的处理过程，由于多了偏移量计算这一步，拥有可变形卷积的模型会比常规的推理速度稍慢一些。本小节的无人机检测模型将在 ResNet-101 的 Conv3 _x, Conv4 _x, Conv5 _x 中以可变形卷积代替常规卷积，此时特征图尺寸已经较小，而且特征层级较高，可以实现在不拖累检测速度的前提下，试图适应数据集中物体的几何形变。



(a) 普通标准卷积（蓝色点为卷积采样点） (b) 可变形卷积（蓝色点偏移到了红色点）

图 4.10 3×3 可变形卷积示意图

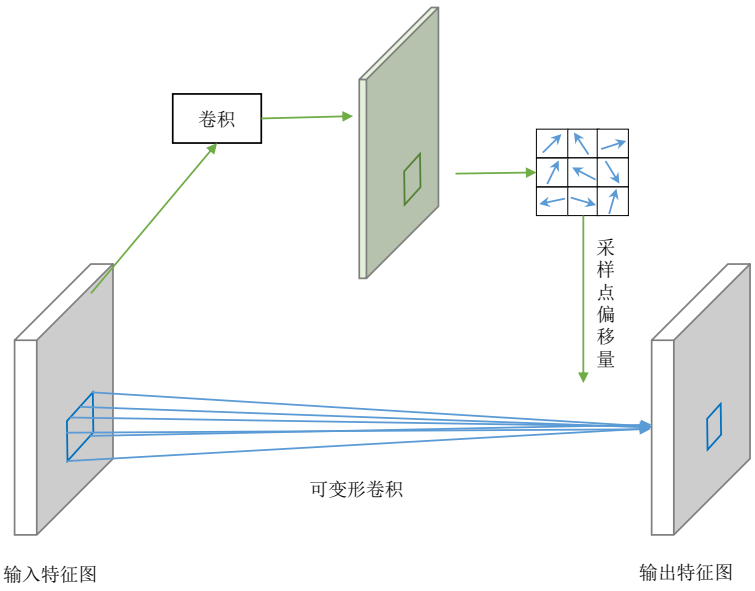


图 4.11 3×3 可变形卷积计算过程示意图

4.2.3 针对复杂背景问题的改进

1). 自注意力机制

根据通用近似定理，虽然深层神经网络可以在理论上以指定精度逼近大部分函数，但是实际上由于难优化和设计原因，很多模型都无法进行长距离信息捕捉，存在“记忆遗忘”的问题。在卷积神经网络中，模型提取的是局部特征，这些特征之间可能也存在

冗余或者某种联系，在某些任务上，有些特征是有用的，而有些则是无用的。自注意力 (Self-Attention) 机制[100–103] 就是根据人类在学习时聚焦注意力，关注主要问题和矛盾的现象而提出的一种辅助模型学习的方法，同时也可以缓解模型的复杂度，增加特征提取方式的多样性，补充更加丰富和鲁棒的语义信息。

自注意力机制在卷积神经网络中的应用主要有三个部分：(1) 进行特征图空间上的全局文本建模，将所有位置的特征聚合在一起，整合出该特征图的上下文总特征；(2) 进行特征图通道间的相关性建模，不同幅特征图之间可能存在相关性以及任务上的依赖性，通过通道间注意力建模找出相关系数；(3) 进行原特征与经过注意力建模之后的结果进行融合，实现特征聚焦和校准。

图 4.12 给出了应用在本小节检测模型中的自注意力结构示意图。输入的特征图 $\mathbf{x}$ 经过一个 1×1 卷积 ($\mathbf{W_1}$) 降低通道数至 1，调整矩阵后通过 Softmax 激活函数并与原特征图相乘，完成特征图的上下文特征聚合。接着再让其经过一个 1×1 的卷积 ($\mathbf{W_2}$)，对结果进行 LayerNorm(层归一化，与批量归一化的作用类似，只不过这里是在特征维度上进行归一化) 后使用 ReLU 激活函数，其目的是为了降低特征图通道数，减少计算量，同时进一步提取特征。然后通过一个 1×1 的卷积 ($\mathbf{W_3}$)，恢复通道数，得到校准了特征图依赖关系的全局文本特征。最后将此特征与原特征相加，实现了注意力聚焦后的特征图 ($\mathbf{z}$)，其可以表示为：

$$\mathbf{z}=\mathbf{x}+\mathbf{W_2}\,\mathrm{ReLU}\left(\mathrm{LN}\left(\mathbf{W_1}\sum_{j=1}^{H\times W}\frac{e^{W_1\mathbf{x}_j}}{\sum_{m=1}^{H\times W}e^{W_1\mathbf{x}_m}}\mathbf{x}_j\right)\right) \tag{4.8}$$

考虑到计算量和特征抽象程度，该自注意力模块与可变形卷积一样，只在 ResNet-101 的 Conv3 _x, Conv4 _x, Conv5 _x 中加入。

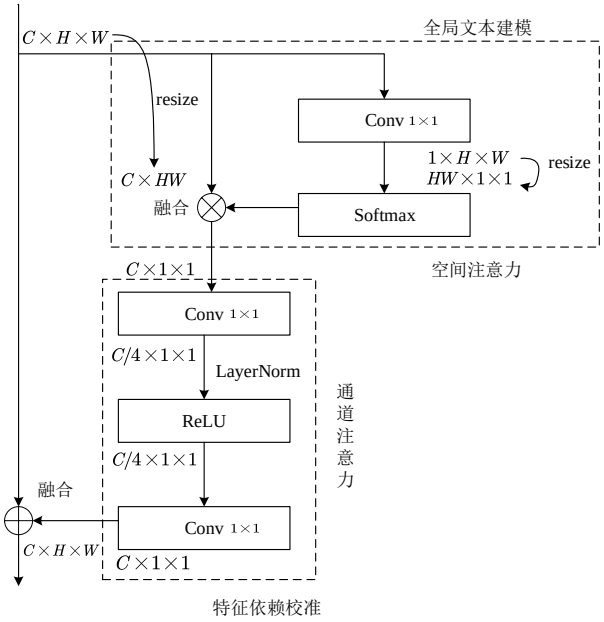


图 4.12 自注意力结构示意图 (C 代表特征图通道数，H,W 代表特征图高宽)

2). 回归损失函数与困难样本挖掘

在模型的训练方面，由于物体多尺度分布以及小目标甚至极小目标的存在，其学习难度会加大，而且收集的数据集体量较小，有些复杂的样本数量可能处于不均衡的状态，因此模型参数的更新方向会主要来自于那些常见的容易样本。为了让模型的训练过程更加全面，特别地，尽量多关注那些复杂的，困难的样本，提高检测精度，将对模型配备新的回归损失函数，并加入在线困难样本挖掘技术。

Faster R-CNN 使用 Smooth L1 作为框的回归损失函数，通过学习建议框与与真实框之间的平移缩放变换值来进行预测，然而在目标检测中，检测框好坏评价往往使用交并比 (IoU) 指标，这样就导致模型优化上的不一致，此外，L_n 范数对物体的尺寸也比较敏感，在物体尺寸差异大的数据集上可能出现某些大小的物体检测质量较差的情形。为此，该小节的检测模型将采用对物体尺度不敏感，优化方向和评价指标一致的 GIoU[104] 损失函数，其表达式如下：

$$L_{\mathrm{GIoU}}(P,G)=1-\left(\mathrm{IoU}-\frac{|C\backslash(P\cup G)|}{|C|}\right)=1-\mathrm{GIoU} \tag{4.9}$$

式 (4.9) 中，P,G 分别代表预测框和真实框，C 表示把 P,G 包含在其中的一个最小矩形框，$\frac{|C\backslash(P\cup G)|}{|C|}$ 表示 C 中没有覆盖到 P,G 的区域面积与 C 总面积的比值，而 GIoU 的定义则是 P,G 的交并比减去这个比值。

GIoU 损失函数的设计初衷是考虑到 P,G 在一开始的交集可能很小，极端的情况为 0 时无法优化 ($1-\mathrm{IoU}$ 始终为 1，无法更新梯度)。GIoU 作为一种改进后的距离，大小在-1 到 1 之间，当 P,G 的交并比为 0 时，GIoU 为负值，衡量的是这两个框之间的距离远近，此时损失值大于 1，因此在优化时便往不断拉近这种距离的方向去更新参数，直到 P,G 完全重合，GIoU 为 1，此时损失值降为 0，完美地完成了框的优化任务。

改进后的 Fatser R-CNN 中，将 GIoU 损失函数代替区域生成网络和二次分类回归检测网络里的 Smooth L1 损失函数，作为新的框回归任务指导准则。其中式 (4.9) 中 C 的确立规则如下：

$$\begin{aligned} x_1^c&=\min\left(x_1^p,x_1^g\right),x_2^c=\max\left(x_2^p,x_2^g\right)\\ y_1^c&=\min\left(y_1^p,y_1^g\right),y_2^c=\max\left(y_2^p,y_2^g\right)\end{aligned} \tag{4.10}$$

式 (4.10) 中，$(x_1^p,y_1^p),(x_1^g,y_1^g),(x_1^c,y_1^c)$ 和 $(x_2^p,y_2^p),(x_2^g,y_2^g),(x_2^c,y_2^c)$ 分别代表 P,G,C 框的左上和左下点坐标，即有 $x_1^i<x_2^i,y_1^i<y_2^i,i=p,g,c.$

在 Faster R-CNN 中，区域生成网络提供给二次分类回归检测网络的建议区域并不是全部被选取参与训练，而是通过随机采样，使得一个小批量中正负样本的比例为 1：3，其中正样本与真值的交并比需满足在 [0.5,1] 之间，负样本在 [0.1,0.5) 之间，交并比小于 0.1 的负样本被认为包含很少有用的特征，属于易分的负样本被直接忽略。这种采样的方式类似于 SSD 中的困难负例样本挖掘，在一定程度上缓解了锚框机制带来的正负样本不均衡问题。但是这样的采样方式存在两个问题，一是 0.1 阈值的选取并不具有合理性，不同的数据集可能带来不同的效果；二是正样本中也存在困难，不容易学习区

分的样本，不单单只有负样本才有，随机抽取正样本的方式存在局限性，会导致模型漏检，召回率低。另外在本文的可见光图像数据集中，无人机的尺度不一，而且存在一些相似干扰物体，正样本的学习难度会进一步加大。

为了让模型更充分地学习那些复杂的，容易错分漏检的样例，本文在 Faster R-CNN 的二次分类回归检测网络中加入在线困难样本挖掘技术[105](Online Hard Example Mining, OHEM)，让模型自己选取合适的建议区域去训练。其工作流程如下：

(1) 让二次分类回归检测网络对每个建议区域进行前向计算，得出每个区域的损失值，并按从大小排序，不进行反向传播；

(2) 对这些区域进行非极大抑制，因为有的建议区域之间可能存在很大的重叠，如果其中一个的损失很大，被认定为困难样本，那么其他的也会被同样认定，造成样本的冗余；

(3) 按照损失值的大小选取训练样本，并按照正负样本 1：3 的比例组成小批量，让二次分类回归检测网络进行训练，实现参数更新。

在线困难样本挖掘技术通过损失值来选取困难样本的方式简单有效，虽然加长了训练时间，但是并不对模型的推理速度造成影响，因此作为一项提升手段取代二次分类回归检测网络中的建议区域随机采样方法。

4.2.4 实验结果与分析

表 4.3 改进模型在可见光数据集上的实验结果 (AP%)

模型	AP_{drone}	AP_{kite}	AP_{bird}	mAP	FPS
Faster R-CNN	66.5	33.4	11.2	37.0	**22.6**
Faster R-CNN+F	71.2	45.7	13.3	43.4	20.2
Faster R-CNN+F+D	72.5	47.1	14.1	44.6	17.4
Faster R-CNN+F+D+A	72.3	48.5	14.9	45.5	15.8
Faster R-CNN+F+D+A+G	72.5	50.0	15.6	46.7	15.8
Faster R-CNN+F+D+A+G+O	**72.6**	**54.2**	**17.5**	**48.1**	15.8

为了探究上述改进对物体检测的影响，在可见光低慢小无人机数据集上做了一系列消融实验进行验证。实验环境配置如表4.1所示，特征提取网络皆采用在 ImageNet 上预训练过的 ResNet-101，以实现模型参数的初始化，其他层则使用 He 初始化方法，优化器选用带动量的随机梯度下降，初始学习率皆为 0.001，动量值为 0.9，ℓ_2 正则化系数为 0.0001，总共训练 20 个回合，并在第 16 和第 20 个回合将当前学习率分别衰减 0.1 倍，小批量数皆设为 2。表 4.3 给出了消融实验的检测结果，其中 F 代表特征金字塔，D 代表可变形卷积，A 代表自注意力机制，G 代表 GIoU 回归损失函数，O 代表在线困难样本挖掘技术。第一行是基准模型 Faster R-CNN 在可见光数据集上的检测结果，最后一行是对其改进后的检测结果，可以看出，模型对无人机，风筝和鸟这三类物体的检测 AP 分别增长了 6.1%，20.8% 和 6.3%，最终的 mAP 也增加了约 7 个点，实现了

不小的提升。图 4.13 和图 4.14 给出了该改进模型对可见光数据集中的无人机，风筝及飞鸟的部分检测样例，其较好地完成了多尺度的检测任务，同时也具有较强的复杂背景抗干扰能力，即使当这三类物体尺寸都很小时，模型也可以进行一定程度的区分，这些说明本小节改进手段的有效性和可行性。此外，由于加入了特征金字塔，可变形卷积和自注意力机制，带来了额外的计算量，所以拖慢了检测速度，改进后的模型检测速度为 15.8FPS，相比原版慢了 6.8FPS，但是依然可以满足特定场景下的检测需求。

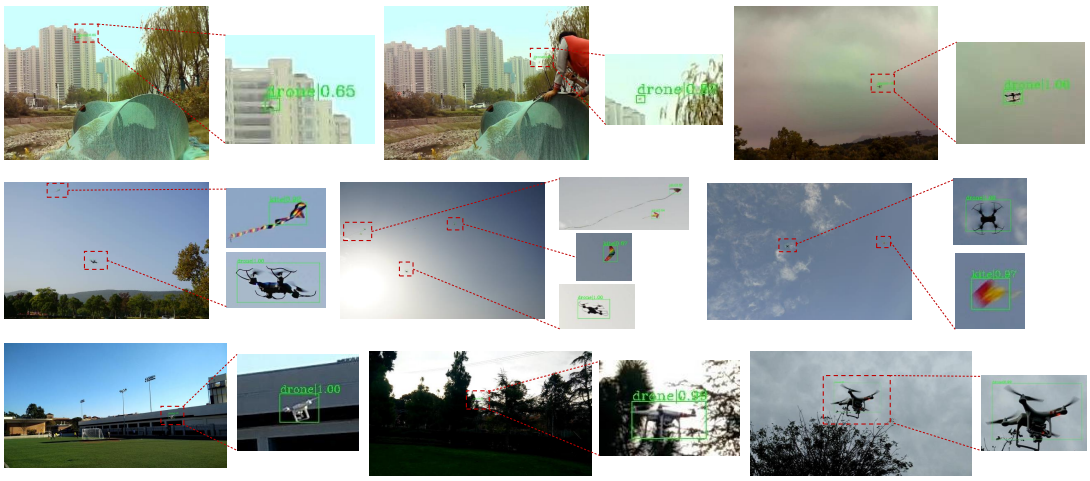


图 4.13 改进后的模型对可见光数据集中低慢小无人机检测结果样例

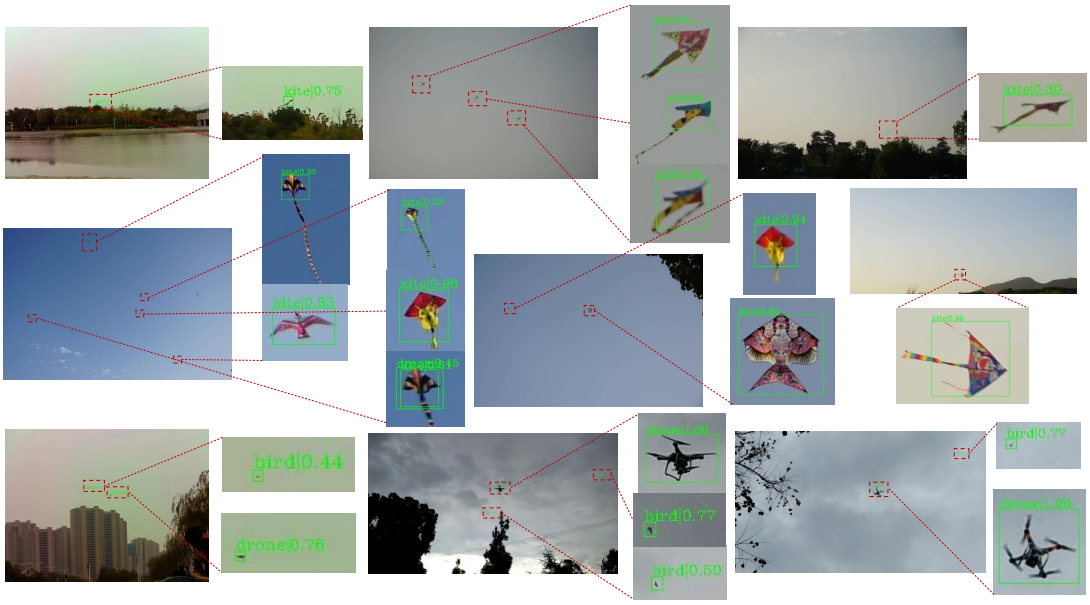


图 4.14 改进后的模型对可见光数据集中风筝和鸟类检测结果样例

对比实验的各部分消融分析如下：

(1) 特征金字塔 (F)

为了应对物体的尺度变化，首先在 Faster R-CNN 中加入特征金字塔，一方面增加小目标的位置信息，一方面实现各个大小物体检测的分而治之。如前所述，加入该模块

后，将锚框的基础尺寸设为 4，缩放比系数设为 P2、P3、P4、P5、P6 对应的原图下采样倍数，即 $[4, 8, 16, 32, 64]$，宽高比采用初始设置，即 $[0.5, 1, 2]$，则生成的锚框的面积大小范围在 $16^2 \sim 256^2$ 之间，且每种大小都具有三种形态，基本可以覆盖到可见光数据集中的所有物体。实验结果显示无人机、风筝和鸟这三类物体的检测 AP 都得到了不同程度的提升，尤其是无人机和风筝最为明显，分别增长了 4.7%，12.3%，而鸟因为数据集中样本数量较少，且分布不均，所以仅增加了 2.1%，这证明了特征金字塔对小目标和多尺度物体检测的必要性。

(2) 可变形卷积 (D)

为了进一步解决物体的多尺度和形变等问题，在 (1) 的基础上加入了可变形卷积。实验结果显示三类物体的检测 AP 继续得到了提升，分别上涨了 1.3%、1.4% 和 0.8%，说明可变形卷积可以在一定程度上缓解目标的形变问题，验证了其可行性。此外，由于可变形卷积的偏移量也是自适应学习的参数，如果进一步扩大可见光数据集的体量和质量，对其进行充分地训练，最终的效果会更好。

(3) 自注意力机制 (A)

在 (2) 的基础上引入自注意力机制以验证其有效性。风筝和鸟作为复杂背景中的干扰对象，会影响模型对无人机的检测和判断，根据表 4.3，加入该模块后，模型对风筝和鸟的检测 AP 分别提升了 1.4%，0.8%，说明其在辅助模型辨别正确物体的能力上发挥了一定的作用。无人机的检测 AP 几乎保持了不变，这可能与本小节采用的训练图片不够充足有关，同可变形卷积一样，该自注意力机制内部也具有待学习的参数，这些参数与模型处理的数据和任务有关，因此若进一步扩充可见光数据集，自注意力机制将会发挥更明显的作用。

(4) 回归损失函数 (G) 与困难样本挖掘 (O)

为了进一步加强模型的训练优化，提高复杂背景下的抗干扰和辨别能力，在 (3) 的基础上，先替换了回归损失函数，然后引入了在线困难样本挖掘技术。将 Faster R-CNN 中的 Smooth L1 损失函数换为 GIoU 损失函数后，风筝和鸟的检测 AP 分别增长了 1.5%，0.7%，验证了回归-评价指标一致优化的作用。在此基础上继续加入在线困难样本挖掘技术，风筝和鸟的 AP 又分别增长了 4.2% 和 1.9%，说明该技术让模型强制关注训练样本中的复杂难识别的样本，加大对它们的学习力度，以增加检测性能，具有可行性。值得一提的是，无人机的检测 AP 没有得到继续提升可能是因为锚框没有覆盖到所有尺寸的无人机目标，这部分物体也就没有被当成预选区域进行分类和回归，可以通过再添加几个锚框基础尺寸系数进行缓解，但是这样又会带来不小的计算开销和分布上的冗余，得不偿失。此外，训练数据体量不够大也是影响因素。

4.3 面向红外热成像的低慢小无人机检测算法

为了实现全天候检测低慢小无人机目标，本文采取白昼使用可见光探测设备，夜晚使用红外热成像探测设备的策略。在第二章中，总共收集了 5546 张红外图片，其中包

含无人机 4438 架，风筝 1707 个，鸟类 62 只。其中，无人机的最小像素为 15，最大像素为 3504，平均像素为 249；风筝的最小像素为 36，最大像素为 2278，平均像素为 366；鸟的最小像素为 28，最大像素为 760，平均像素为 236。从像素大小范围看，这三类物体的尺度变化不大，都属于小目标物体 (平均像素小于 32×32[68])。此外红外热成像图片不同于可见光图片，没有丰富的色彩和纹理特征，背景干扰相对较少，因此本小节将继续采用稳定性较好的 Faster R-CNN 作为基础模型，保证检测精度，并根据小目标和红外图像的特点对其进行改进。

4.3.1 红外无人机数据集基准实验

为了直观了解基础模型 Faster R-CNN 在红外数据集上的检测效果，先进行基准实验，实验的环境配置如表4.1所示，训练集 4021 张，测试集 1525 张，并直接采用测试集每隔 1 个回合进行精度验证，防止过拟合。基准实验结果如表 4.4 所示。

表 4.4 基准模型在红外数据集上的实验结果 (AP%)

模型	AP_{drone}	AP_{kite}	AP_{bird}	mAP	FPS
Faster R-CNN(ResNet-101)	11.2	0.6	0.0	3.9	22.6

基准实验中，Faster R-CNN 采用在 ImageNet 上预训练过的 ResNet-101 作为特征提取网络，锚框的基础尺寸为 16，缩放比系数设为 $[4, 8, 16]$，锚框的宽高比系数设为 $[0.5, 1, 2]$，二次分类回归网络使用 RoI Align 作为区域抠取方法。优化器使用带动量的随机梯度下降法，初始学习率为 0.001，动量值为 0.9，ℓ_2 正则化系数为 0.0001，小批量数为 4，总训练回合数为 20，且在第 16 和第 20 个回合分别将当前学习率衰减 0.1 倍。

基准实验结果显示 Faster R-CNN 已经无法很好地处理红外数据集中的小目标检测，而鸟这类的干扰物体，由于样本很少，且可能存在人为标注错误，导致基本无法检测出来。如上一节所述，Faster R-CNN 仅使用一个深层特征图来进行锚框生成和建议区域特征抠取，此时原始图像经过下采样操作，分辨率逐渐降低，物体的细节信息也不断丢失。而小目标本身像素就少，因此信息损失更加严重，在该深层特征图中已经基本没有代表特征了，因此必须引入第4.2.2节中的特征金字塔结构，对高层语义特征补充小目标的细节信息。此外，为了进一步保证图像的分辨率信息在特征提取时得到保留，以更好地处理小目标检测，将 ResNet-101 替换为更加具有针对性的 HRNet(High-Resolution Network)[106,107]，使得网络在任何阶段都可以保持高分辨率的表征能力。

4.3.2 基于小目标的特征提取网络

一般的卷积神经网络模型为了在特征提取过程中保持高分辨率信息，基本是采用跳级连接这种串联的方式来融合高低层特征图信息，典型的代表是分类网络 ResNet 和 DenseNet[108]，它们在不断下采样的过程中缩小图像尺寸，同时对之前的浅层特征图利用互联相加或者叠在一起的操作来给深层特征图提供图像的一些细节信息。但是这些操作实际上还是有损的，尤其下采样的操作是不可逆的，网络最后只输出一个尺度的特征

图，即使是进行了融合操作，融合的左值也可能会因为尺度原因在维度配准时丢失掉一些关键信息。对于本小节的红外数据集来说，无人机，风筝，鸟类这三类物体都属于小目标物体，为了让特征提取网络拥有尽可能多的高分辨率信息，以保证小目标的有效检测，从而引入 HRNet 网络模型。

HRNet 的结构如图 4.15 所示，其通过并联的方式，一边保持特征图的高分辨率，一边不断降低特征图的分辨率，以此形成了不同的四个阶段，同时每个阶段的末尾都会进行不同分辨率之间的特征融合，使得每一个尺度的特征图都能反复充分吸收不同并行路线中其他尺度特征图的表征信息，让网络在不断加深的过程中依然可以拥有丰富的特征表达和细节信息，从而加强后续对小目标的检测效果。

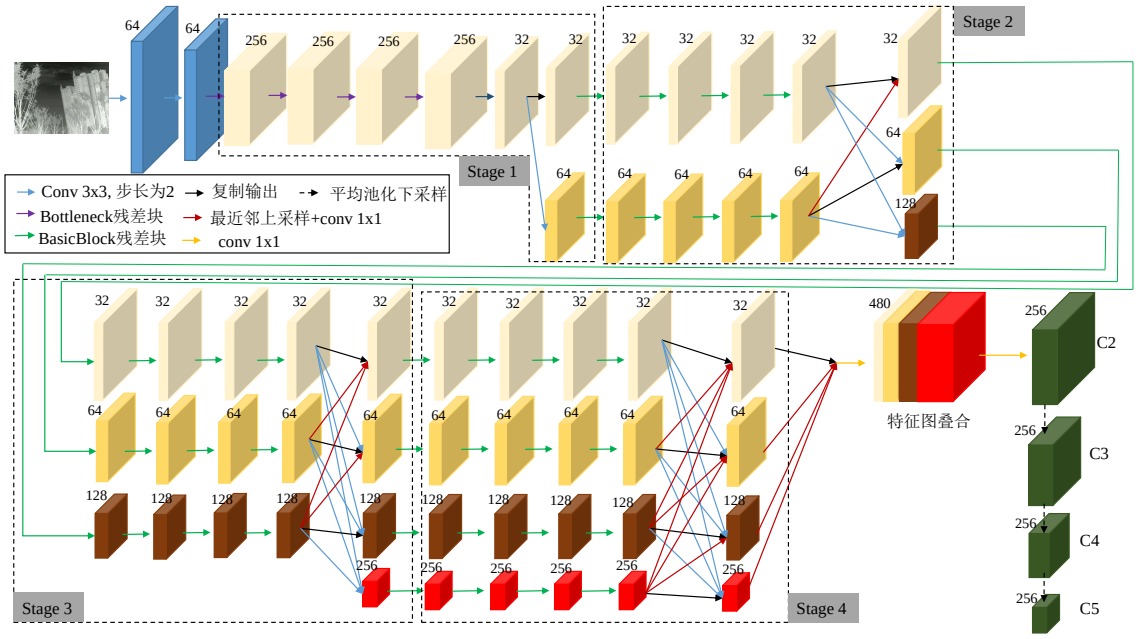


图 4.15 用于红外小目标检测的特征提取网络 HRNet 结构示意图 (数字为特征图的通道数)

根据图 4.15，将 HRNet 应用在 Faster R-CNN 中，并结合特征金字塔结构，其具体工作流程如下：

(1) 对输入原图使用两个步长为 2 的 3×3 卷积进行特征提取，并分别下采样 2 倍，然后利用图3.12中的 Bottleneck 残差结构继续提取特征 (便于网络优化训练，后同)，并保持输出通道数为 256 且分辨率相同，共进行 4 次。对最后一次该操作下输出的特征图使用一个 3×3 卷积，将通道数降低到 32。最后采用一个步长为 2 的 3×3 卷积进行下采样，开辟分辨率为原图 1/8 这条并行特征提取路线，结束 stage 1，开启 stage 2；

(2) 对分辨率为原图 1/4 和 1/8 这两个特征图分别进行 4 次基于图3.12中的 BasicBlock 残差块的特征提取操作，保持分辨率和通道数不变，然后对两条路线输出特征图进行融合：高分辨率特征图使用步长为 2 的 3×3 卷积进行尺寸降低，并扩大通道数，与低分辨率特征图相加作为最后的低分辨率特征图输出；低分辨率特征图使用最近邻插值上采样，并随后采用一个 1×1 的卷积降低通道数，与高分辨率特征图相加作为

最后的高分辨特征图输出，实现了不同路线间的多尺度特征融合。最后对原图 1/8 分辨率的特征图利用步长为 2 的 3×3 卷积实现下采样，开辟分辨率为原图 1/16 这条并行特征提取路线，结束 stage 2，开始 stage 3;

(3) 对分辨率为原图 1/4、1/8 和 1/16 这三条路线分别进行 4 次基于 BasicBlock 残差块的特征提取操作，并进行不同尺度之间的特征图融合，做法与 (2) 相同。最后对分辨率为原图 1/16 的特征图使用一个步长为 2 的 3×3 卷积进行下采样，开辟分辨率为原图 1/32 这条并行特征提取路线，结束 stage 3，开始 stage 4；

(4) 对分辨率为原图 1/4、1/8、1/16 和 1/32 这四条路线分别进行 4 次基于 BasicBlock 残差块的特征提取操作，并进行四个尺度之间的特征融合，做法与 (2) 相同。然后将后三个融合输出的不同分辨率的特征图进行上采样操作，并一起和融合后的分辨率为原图 1/4 的特征图叠在一起。接着使用一个 1×1 的卷积对其进一步提取特征，去除混叠效应，得到了最后输出的高分辨率特征图；

(5) 对 (4) 得到的特征图使用一个 1×1 卷积，将通道维数降到 256，记为 C2，并接连使用 3 个平均池化使特征图尺寸分别下采样到原图的 1/8、1/16 和 1/32，记为 C3、C4 和 C5。这四个特征图与图4.9中的 Conv2_x、Conv3_x、Conv4_x 和 Conv5_x 的作用类似，组成特征金字塔的自下而上路径，其后的自上而下路径与侧向融合做法按照图4.8进行。最后按照图4.9的结构形成基于 HRNet 特征提取网络和特征金字塔结构的 Faster R-CNN 检测模型。

4.3.3 实验结果与分析

表 4.5 改进模型在红外数据集上的实验结果 (AP%)

模型	AP_{drone}	AP_{kite}	AP_{bird}	mAP	FPS
Faster R-CNN(ResNet-101)	11.2	0.6	0.0	3.9	**22.6**
Faster R-CNN(ResNet-101)+FPN(scale=4)	57.4	67.8	0.2	41.8	20.2
Faster R-CNN(ResNet-101)+FPN(scale=2)	68.9	68.0	0.1	45.7	20.2
Faster R-CNN(ResNet-101)+FPN(scale=1)	68.3	68.4	0.1	45.6	20.2
Faster R-CNN(HRNet)+FPN(scale=2)	71.1	71.6	2.3	48.3	17.5
Faster R-CNN(HRNet)+FPN(scale=2)+OHEM	**72.7**	**74.0**	**9.1**	**51.9**	17.5

由于数据集中每幅图像基本只存在一个小目标，为了抑制正负锚框样本不均衡，增强检测模型对复杂样本的辨别能力，除了加入 HRNet 和特征金字塔之外，也引入第4.2.3节中的在线困难样本挖掘技术。为了验证改进后的模型对红外小目标的检测效果，进行了一系列对比实验。表 4.5 给出了基础模型 Faster R-CNN 有无特征金字塔结构 (FPN) 以及具有不同大小的锚框基础尺寸 (scale)，使用不同的特征提取网络和有无在线困难样本挖掘技术 (OHEM) 下的对比实验结果 (训练参数配置同基准实验)。从中可以看出，本小节提出的改进模型不论是单类别还是总体，检测精度都是最高，甚至是样本很少的鸟类，也比基准模型提高了 9 个点，说明了改进手段对小目标识别的有效

性，同时只比未改进前的模型慢了 5FPS，依然可以满足一定的实时性需求。图 4.16 和图 4.17 给出了改进后的模型在红外数据集上的部分检测结果。

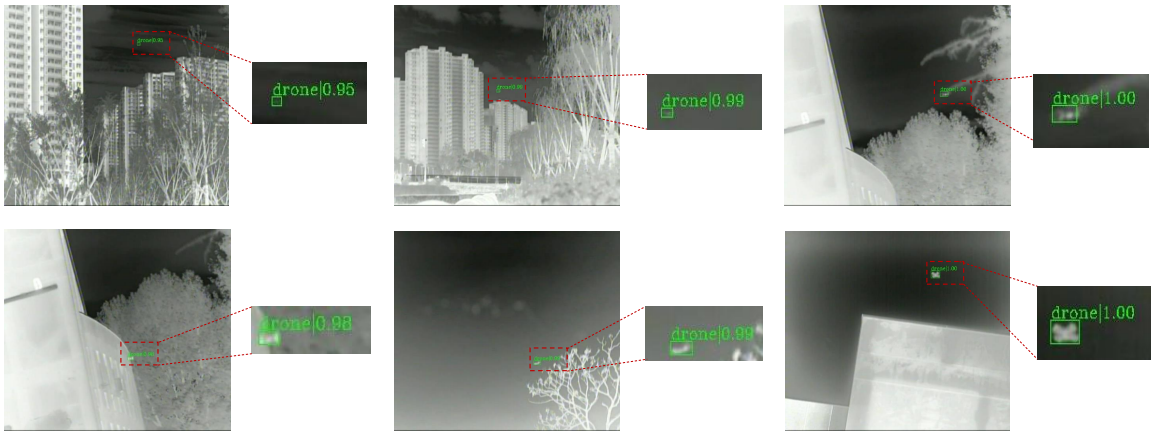


图 4.16 改进后的模型对红外数据集中低慢小无人机检测结果样例

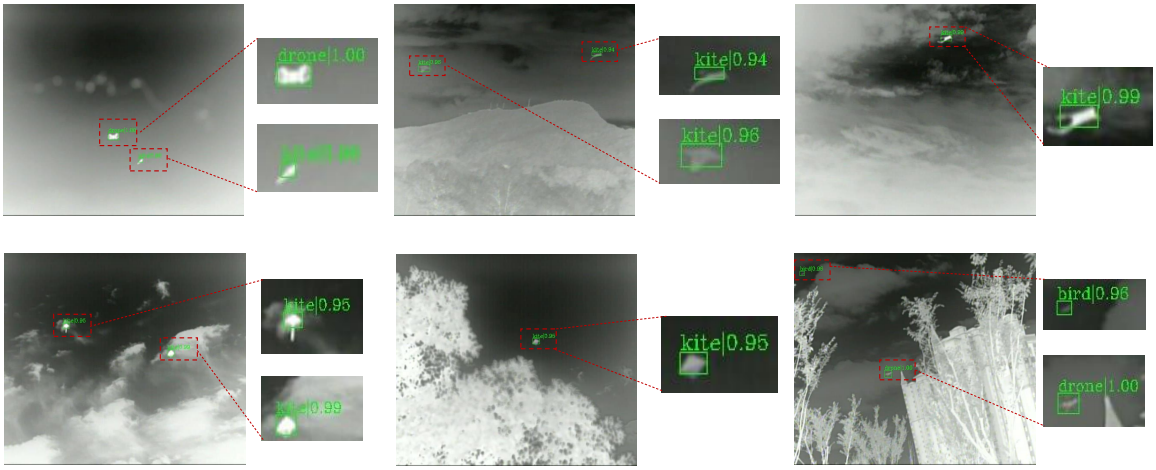


图 4.17 改进后的模型对红外数据集中风筝和鸟类检测结果样例

对比实验的各部分消融分析如下：

(1) 特征金字塔 (FPN) 和锚框基础尺寸 (scale)

在 Faster R-CNN 中加入特征金字塔结构，以增加小目标的细节信息，帮助模型检测。其直接按照可见光数据集的检测设置，将锚框的基础尺寸赋为 4，锚框的缩放比系数赋为 $[4, 8, 16, 32, 64]$，锚框宽高比赋为 $[0.5, 1, 2]$。实验结果显示检测精度得到了巨大提升，大部分无人机和风筝都可以被识别出，证明了特征金字塔对小目标检测任务的必要性。此外根据物体像素大小可知，该红外数据集的多尺度问题并不如可见光那样突出，主要问题是小尺度，因此可以将锚框的基础尺寸调小，提高预选区域的质量，方便模型学习。把锚框的基础尺寸调为 2 后，模型的 mAP 增长了 4 个点，其中无人机检测 AP 增长最为明显，证明了选取合适的锚框基础尺寸参数也可以有效辅助小目标检测。进一步地，继续调小该基础尺寸没有明显的效果提升，说明预选框的质量已经足够。

(2) 基于小目标的特征提取网络 (HRNet)

由于 ResNet-101 是一种通用的物体特征提取网络，为了保留更多的小目标信息，将其换为 HRNet，并加入特征金字塔，锚框的基础尺寸也选用 (1) 中确定的最优值 2。实验结果显示无人机，风筝和鸟的检测 AP 值都较同等参数配置下的 ResNet-101 提升了 2∼3 个点，说明其在保留图像分辨率方面效果明显，也可以促进小目标检测性能的提升。

(3) 在线困难样本挖掘技术 (OHEM)

为了加强对复杂样本的学习力度，在 (3) 的基础上加入可见光检测模型中引入的在线困难样本挖掘技术，以辅助模型更好地优化。最终的检测 mAP 达到了 51.9%，三类物体的识别效果都有不同程度的提升，其中尤其以鸟的提升最为明显，因为鸟的样本数量少，很容易被模型漏检或错检，属于典型的困难样本，而在线困难样本挖掘技术强制模型去关注这类样本，使得其识别率得到提高。

4.4 本章小结

本章通过对可见光和红外低慢小无人机数据集的检测难点进行分析，在 Faster R-CNN 上分别做出了合适的改进，设计并实现了相关的检测算法，实验证明了改进后模型的有效性，可以满足无人机的全天候监测任务。本章的内容主要分为三个部分，第一部分先分析研究了基于锚框的二阶检测算法 Faster R-CNN 和一阶检测算法 SSD 的结构与工作流程，并在可见光数据集上做了基准比较实验，确立了本章所选取的基础模型为稳定性和精度都较好的 Faster R-CNN。第二部分根据可见光数据集中待检目标存在多尺度和复杂背景的问题，针对性地引入了特征金字塔，可变形卷积，自注意力机制，GIoU 回归损失函数和在线困难样本挖掘技术，消融实验验证了改进手段的有效性，实现了可见光下的低慢小无人机检测任务。第三部分针对红外数据集中待检物体属于小尺度甚至极小尺度目标的难点，利用可以保持图像分辨率的 HRNet 作为特征提取网络，并同样加入特征金字塔和在线困难样本挖掘技术以应对小目标问题，最终的实验表明改进模型可以很好地完成红外图像下的低慢小无人机检测任务。

5 基于像素点识别的视觉探测无人机算法设计与实现

第四章利用矩形框的形式对无人机在图像中进行了定位，完成了目标检测任务，但是框中的区域除了包含无人机还包含背景，为了进一步得到图像中无人机更加精细的位置区域，本章将试图实现无人机的语义分割，获取像素级定位结果。但是由于语义分割模型的训练需要对应的物体像素标签，这项标注工作要比框标注更加费时费力，成本也更加昂贵，因此本章提出了一种基于框标注来生成像素标签的机制，并根据数据集的特点设计合适的语义分割算法，完成低慢小无人机的弱监督语义分割系统，实现无人机更加精准的图像定位，以满足不同的检测需求。

5.1 低慢小无人机弱监督语义分割系统

5.1.1 基于标注框的像素标签生成

本章所研究的语义分割算法仅在可见光图像上进行，并且只针对低慢小无人机。由于模型和训练任务与第四章的目标检测不同，重新对数据集进行了整理和划分，其中训练集 7672 张图片，验证集 354 张图片，测试集 1411 张图片。就目前通用的监督学习而言，模型需要对应任务的真值来计算损失值，然后以此指导参数更新，但是图像中物体的像素级标注费时费力，比一般的框标注难度大，一旦扩大数据集，标注的工作量和成本就会急剧增加。为了承接上一章的检测工作，同时为了降低标注带来的消耗，本章提出一种利用无人机的标注框来生成其像素标签的方法，利用这种生成的“伪标签”来监督模型学习。值得一提的是，这种弱监督标签生成的方法仅在训练集中使用，验证集和测试集的像素标签则使用第二章2.2.2节中的 Labelme 软件进行标注，以评价分割算法的精度，验证弱分割机制的有效性。

无人机像素标签生成的具体流程如下：

(1) 输入原始图像和无人机目标的框坐标，即左上点和右下点坐标；

(2) 利用 GarbCut[109] 算法，根据 (1) 中信息，抠取出无人机区域；

(3) 如果 (2) 中输出的结果为空，或者输出的区域与标注框的交并比小于 0.15，则直接用整个矩形区域作为无人机的像素标签，否则输出 (2) 的结果；

(4) 创建一个与原图同尺寸的二维标签矩阵，将 (3) 输出的区域的像素值设为 1，代表无人机标签，其他区域的像素值设为 0，代表背景标签，最后保存成真值图像供模型训练时读取。

步骤 (2) 中所使用的 Grabcut 算法是一种较为成熟的经典计算机视觉抠图算法，其作为一种基于图论的图像分割方法，通过建立吉布斯能量函数和高斯混合模型，不断迭代求解函数的最小值，得到指定区域内的前景和背景像素集合。图 5.1 给出了训练集中部分图片的无人机像素标签生成结果，从中可以看出，GrabCut 算法的抠取效果比较可靠，能够贴合无人机的形状，说明这种“伪像素标签”生成机制具有合理性，但是

GrabCut 算法也有失效的时候，比如物体被背景严重遮挡，或区分度很小时，此时便按照步骤 (3) 使用整个标注框的像素代替，作为一种折中处理的手段。

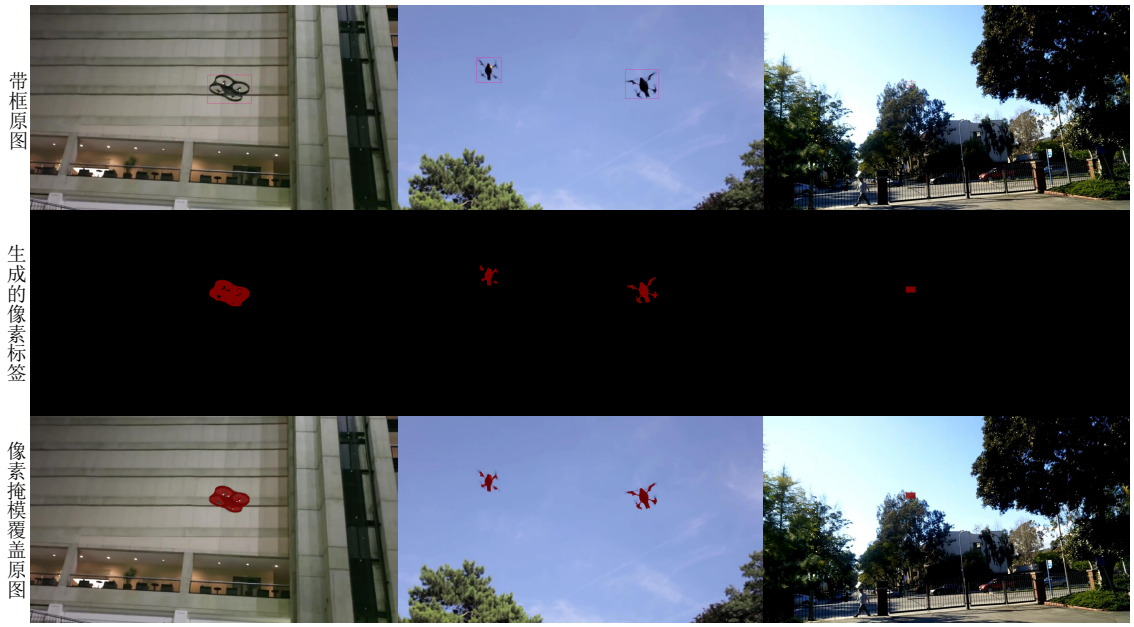


图 5.1 利用框标注和 GrabCut 算法生成无人机像素标签

5.1.2 语义分割模型

在计算机视觉中，图像分类是最基础的认知任务，神经网络也是率先在该领域取得了巨大的突破。语义分割是在图像分类上进一步提升认知难度的任务，即实现图像上每个像素点的分类，目前语义分割算法已经广泛应用于无人驾驶，遥感地图分析等领域。2014 年，Long 率先提出了 FCN[74], 一种全卷积神经网络分割模型，不久后，Ronneberger 也提出了类似的分割模型 U-Net[110], 两者基本奠定了该领域的模型设计思路，引领了后续基于深度学习的语义分割算法研究。

图 5.2 展示了 FCN 的结构示意图，以 VGG-16 为下采样的特征提取网络，作为图像特征编码器 (Encoder)，然后从 VGG-16 的最后一个特征图开始 (去除原始的全局平均池化和全连接层)，不断进行上采样，每次加倍特征图尺寸，作为图像特征解码器 (Decoder)，直到输出尺寸与输入的原图相同，最后再利用一个 1×1 的卷积得到通道数为类别数的预测特征图，逐像素点利用 Softmax 进行类别预测，完成语义分割的任务。FCN 考虑到上采样阶段会出现图像细节丢失，因此使用特征图像素值相加的方式，对编码器和解码器的每个层级进行跨连，以提高分割掩模的边缘等细节效果。U-Net 的结构与 FCN 很类似，也是采用这种下采样-上采样的方式，该模型的架构如图 5.3 所示，但是在编码器和解码器的跨连上，U-Net 使用的是叠操作，把对应层级的特征合并在一起，直接增加特征图的通道数，以此补充图像的细节信息。

值得一提的是，FCN 和 U-Net 中所使用的上采样方式并非是双线性插值，最初是考虑到特征解码的时候也需要一些特征学习和表达，因此采用了带可学习参数的转置卷积。转置卷积与一般的卷积都是一种特征的映射方式，两者在形式上存在转置关系，普

通的卷积设置步长为 2 可以实现特征图尺寸的二倍缩小，同样地，转置卷积设置步长为 2 可以实现特征图尺寸的二倍扩大，图 5.4 给出了两者的差别。

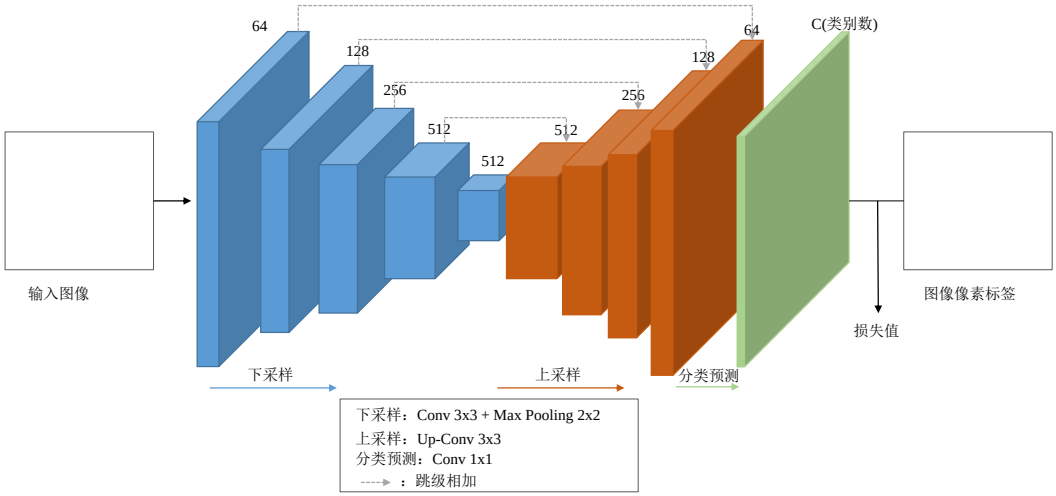


图 5.2 FCN 模型结构示意图 (数字为特征图的通道数)

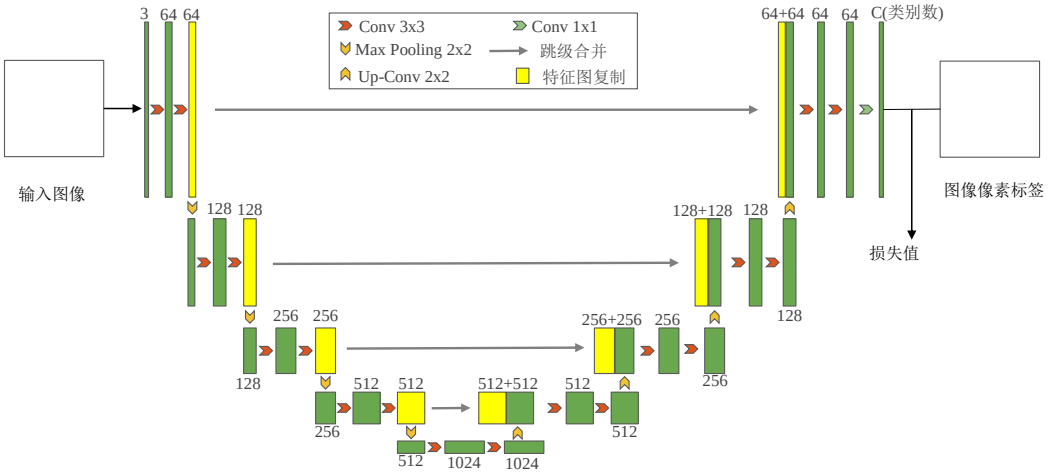
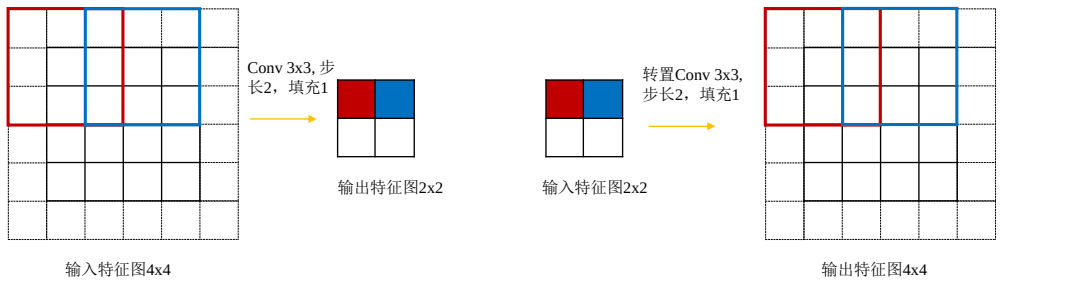


图 5.3 U-Net 模型结构示意图 (数字为特征图的通道数)



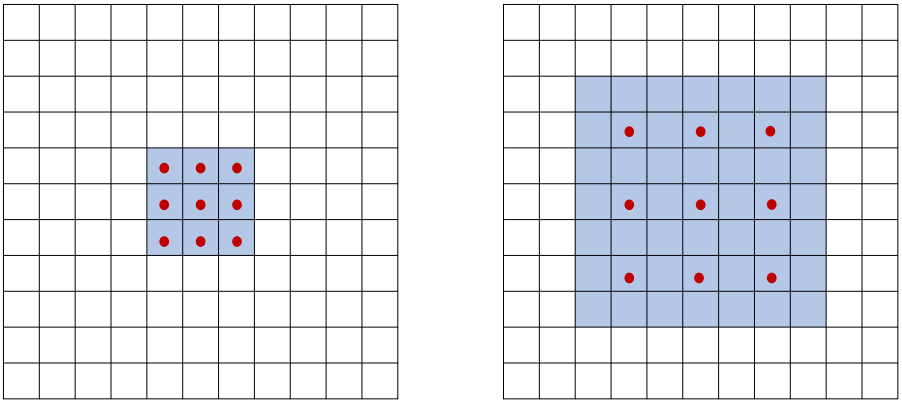
(a) 普通3x3卷积，输入与卷积核的权重进行点积

(b) 3x3转置卷积，相当于输入值给卷积核提供权重进行点积

图 5.4 普通卷积与转置卷积的差别

深层卷积神经网络中层级越高的特征图特征越抽象，其具有很好的平移不变性，处理图像分类任务非常有效，但是缺点是细节信息太少，尤其是经过池化这样的下采样操作，图像分辨率和信息逐渐降低、丢失，难以恢复，因此像 FCN 和 U-Net 这样简单的下采样-上采样结构不能很好地给出物体的边缘信息，大多情况下只能预测出粗糙的涵盖物体大部分区域的分割结果，毛刺较多。此外本章提供的像素标签并不能完全代表真实值，无人机的数据集也存在一定的尺度变化，具有不小的分割难度。为了提高无人机弱分割的性能，本章根据 DeepLab[111–114] 系列语义模型中的膨胀卷积 (Atrous Convolution) 和膨胀空间金字塔池化机制 (Atrous Spatial Pyramid Pooling, ASPP) 来设计编码器-解码器下的语义分割算法。

膨胀卷积的出现是为了取代语义分割中池化或步长大于 1 的卷积等下采样方法带来的不可逆信息损失。下采样最主要的目的是为了增大模型对图片的感受野，以不断抽象出语义信息，为了达到同样的效果同时尽量不损失图像细节信息，膨胀卷积在普通卷积的基础上设置了一个扩张率 (Dilation Rate)，以此拉开采样点的距离。图 5.5 给出了 3×3 普通卷积和膨胀卷积 (扩张率为 2) 的区别，其中红点代表卷积核采样点，淡蓝色区域代表卷积核在该特征图上的感受野，可以看出，由于扩张率的关系，3×3 膨胀卷积相当于 5×5 的普通卷积，采样区域也从 3×3 变成了 7×7，达到了在不增加权重参数的情况下，增大了模型的感受野。如果多个膨胀卷积进行组合，感受野可以进一步扩大，以代替原先的下采样层，保留图像细节信息。实际上，普通卷积也可以看成是扩张率为 1 的特殊膨胀卷积。



(a) 普通3x3卷积　　(b) 3x3膨胀卷积，扩张率为2

图 5.5 普通卷积与膨胀卷积的差别

假设原先的卷积核尺寸为 $k\times k$，扩张率为 d，则膨胀卷积对应的卷积核大小 $k^a\times k^a$ 为：

$$k^a = d\times(k-1)+1 \tag{5.1}$$

经过膨胀卷积之后的特征图输出大小依然可以按照式 (3.16) 进行计算，只需将卷积核大小换为 $k^a\times k^a$ 即可。

在处理物体多尺度问题上，一般模型会采用图像金字塔，特征金字塔等方式来解决，FCN 和 U-Net 的上采样-下采样加跨级连接的方式本质上也是特征金字塔架构的一种体现，但是由于下采样造成不可恢复的图像信息丢失导致其在语义分割任务上的效果不能尽如人意。为了使膨胀卷积在不降低图像分辨率的前提下也能很好地处理多尺度物体，本章引入了膨胀空间金字塔池化机制，其受目标检测模型 SPP-Net[46] 和 GoogLeNet 中包含多尺寸卷积核的 Inception 模块启发，在一个特征图上使用几种具有不同扩张率的膨胀卷积，实现不同的感受野捕获不同尺寸的物体，具体结构如图 5.6 所示。

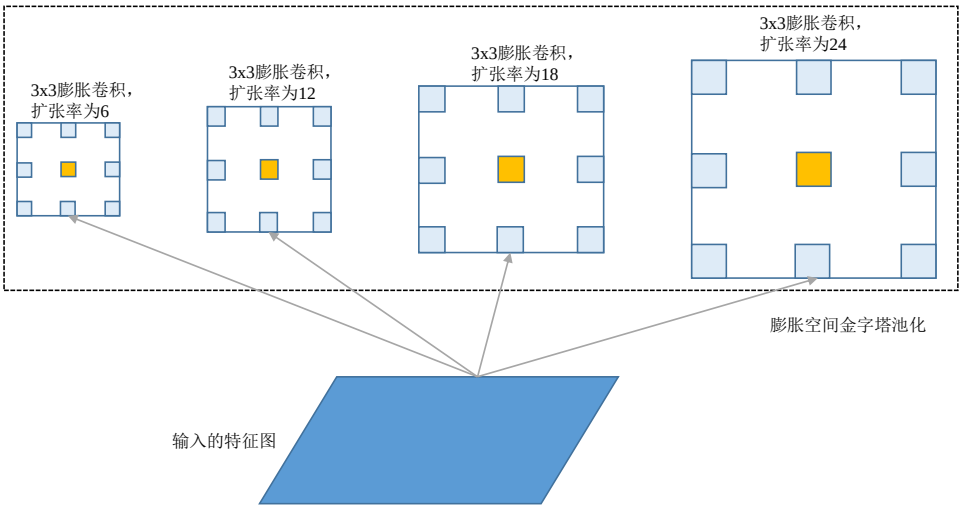


图 5.6 膨胀空间金字塔池化结构示意图

将膨胀卷积和膨胀空间金字塔池化机制结合到编码器-解码器架构中，组成本章的低慢小无人机语义分割模型，其具体结构如图 5.7 所示。该模型编码器部分的特征提取骨架采用 ResNet-101，Conv_1、Conv_2x、Conv_3x、Conv_4x 采用原先的普通卷积，并不断下采样图像，直至缩放到原图的 1/16，最后的 Conv_5x 使用扩张率为 2 的膨胀卷积增大感受野，防止图像信息进一步丢失。这里没有全部使用膨胀卷积代替普通卷积的原因是膨胀卷积是一种稀疏的采样方式，并不会计算每个像素点，单纯的叠加膨胀卷积会使得局部信息丢失，远距离之间的特征没有相关性，产生网格效应[115]，所以作为一种折衷处理，先让普通卷积充分学习局部特征，然后在较深的特征图上使用膨胀卷积避免分辨率再次降低。Conv_5x 之后的特征图则经过膨胀空间金字塔池化结构，加强对多尺度物体的处理能力，这里除了设置了 3 个不同扩张率的 3×3 膨胀卷积，还额外补充了一个 1×1 卷积和一个全局平均池化 (双线性插值后恢复原特征图尺度)，以增加特征的多样性。接着将这 5 种特征叠在一起，作为膨胀金字塔池化的输出，然后再经过一个 1×1 卷积处理降低通道数，减少计算量。至此，该语义分割模型完成了图像特征编码的作用。模型的解码器与 U-Net 类似，也是上采样和特征图的叠操作以恢复信息，第一次上采样的方式是双线性插值，尽量保留信息，并且一次性扩大 4 倍，用来与上采样后的特征图合并的 ResNet-101 中的低层特征图先经过一个 1×1 的卷积进行通道降维，然后再使用两个 3×3 的卷积对合并的特征图进特征提取和解码，去除混叠效应，

降低通道数。最后利用两个 3×3 转置卷积上采样 4 倍，恢复到原图尺寸，再通过一个 1×1 卷积进行像素级别的类别预测，完成图像特征解码的工作。

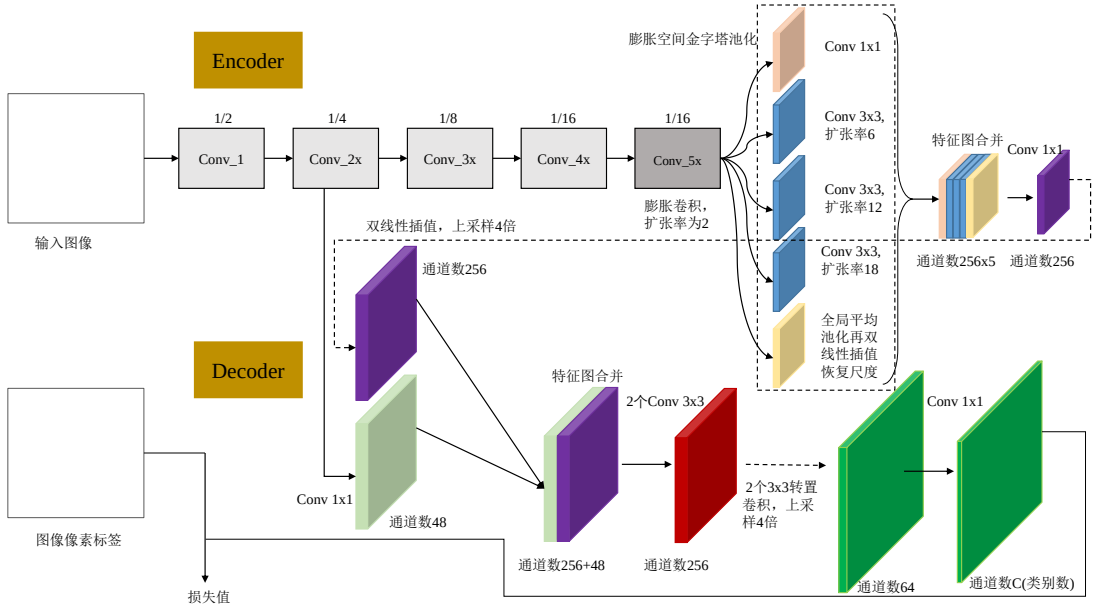


图 5.7 低慢小无人机语义分割网络结构图

5.1.3 损失函数

语义分割本质上也是个分类任务，一般常将交叉熵损失函数作为通用的监督准则，但是在本文的低慢小无人机可见光数据集中，一张图片大多只包含一个无人机物体，且总像素只占很少一部分。因此为了缓解模型训练过程中的正负样本不均衡，防止参数更新方向被背景主导，本章的语义分割模型将采用焦点损失函数[116](Focal Loss)。由于只分割无人机 (加上背景共两类)，其具体计算公式为：

$$L_{\text{Focal}}(y_i, p_i) = \sum_i -\alpha(1-p_i)^{\gamma} y_i \log(p_i) - (1-\alpha) p_i^{\gamma} (1-y_i) \log(1-p_i) \tag{5.2}$$

式 (5.2) 中，p_i 为模型最后输出的特征图 (通道数 C 为 1) 经 Sigmoid 激活函数给出的第 i 个像素的属于无人机置信度，$1-p_i$ 则为属于背景的置信度，y_i 为该像素对应的真实标签，即 1 或 0，γ 和 α 为函数的超参，用来控制样本对损失的贡献程度。焦点损失函数是对交叉熵损失函数的一种改进版本，最初是针对基于锚框的一阶检测模型而提出的，目的是为了缓解其在训练时产生大量的背景锚框而造成了正负样本的极不均衡。现假设利用焦点损失函数在含有大量负类的数据下训练网络，根据式 (5.2)，当网络错分正样本为负样本时，y_i 为 1，损失值由第一项计算得到，此时概率值 p_i 很低，$-\log(p_i)$ 较大，$(1-p_i)$ 也较大，那么系数 $(1-p_i)^{\gamma}$ 对 $-\log(p_i)$ 的缩小程度就会很小。反之，当负样本被分类正确时，y_i 为 0，损失值由第二项计算得到，p_i 较小，$-\log(1-p_i)$ 也较小，且系数 p_i^{γ} 更进一步大幅缩小 $-\log(1-p_i)$ 的值，这样一来网络的损失值就主要由被错分的正样本引导，而且 γ 越大，两者之间的损失缩放程度就会相差越大，因此在

一定程度上起到了困难样本挖掘的作用。α 作为正负样本两类损失值之间的权重配比系数，根据原论文的实验，一般与 γ 的调节方向相反。这两个超参数的取值需要手工设定，要根据不同的数据集情况进行调整，其具体数值将在5.2节的实验中确立。

与目标检测类似，语义分割的评价指标也是交并比，通过计算预测区域与真实区域之间的覆盖面积大小来衡量算法的精度，而采用分类损失函数来进行训练也会存在优化不一致的情况，为此引入 Dice 损失函数[117] 来负责交并比的优化，其计算公式如下：

$$L_{\text{Dice}}(P,G) = 1 - 2\frac{|P \cap G| + \sigma_{smooth}}{|P| + |G| + \sigma_{smooth}} = 1 - \frac{2TP + \sigma_{smooth}}{2TP + FN + FP + \sigma_{smooth}} \tag{5.3}$$

式 (5.3) 中，P, G 分别为模型预测区域和真实区域，σ_{smooth} 是额外设置的平滑项，一般取 1，用以增加数值稳定性。Dice 损失函数衡量的是预测区域与真实区域之间的轮廓相似度，也就是直接根据评价标准优化，但是正是由于针对性较强，其在训练过程中梯度变化剧烈，而且有时损失曲线可信度不高，在其他评价指标上效果不好。所以在本章的语义分割模型中，将其与焦点损失函数结合，充分吸取这两者的优点，既抑制了正负样本不均衡，又同时提高分割的效果。最终的损失函数表达式如下：

$$L = L_{\text{Dice}} + \lambda L_{\text{Focal}} \tag{5.4}$$

上式中，λ 为权重系数，因为焦点损失函数计算出的误差值一般会比 Dice 损失函数小 (具体数量级由参数 α 和 γ 决定)，为了防止在迭代过程中参数更新方向被 Dice 损失函数主导，所以利用该权重系数用来分配两个损失的贡献，其大小影响将在5.2节的实验中分析。

5.1.4 图像混合

本章使用的数据集体量较小，为了增强模型的泛化能力和鲁棒性，将采用 Mixup[118] 图像混合技术进行数据增强，引入对抗样本，防止网络过拟合。Mixup 的具体流程如下：

(1) 一个小批量数据的索引 $\mathbf{I}$ 为 $[1, 2, \cdots, i, \cdots, n]$，将其随机打乱生成新的索引 $\mathbf{I}'$，假设为 $[4, 3, \cdots, j, \cdots, k]$, $i, j, k \leq n$；

(2) 按顺序依次从 $\mathbf{I}, \mathbf{I}'$ 中取出索引，根据索引找出对应的图像，然后从 $\text{Beta}(\alpha, \alpha)$ 分布中采样出混合系数 $\lambda(\alpha > 0, \lambda \in [0, 1])$，对图像进行混合相加生成新的图像作为训练数据。假设其中两个混合的数据分别为 x_i, x_j，则新的训练样本 $\tilde{x}$ 为：

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j \tag{5.5}$$

(3) 按照 (2) 的混合索引和混合系数，对混合前的图像标签做同样的软化破坏处理，即：

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j \tag{5.6}$$

Mixup 这种数据增强手段给模型提供了一种更加平滑的而不确定性估计，可以控制网络的复杂度，从而提高模型的性能。图 5.8 给出了两幅无人机图像经过 Mixup 混合后的结果。

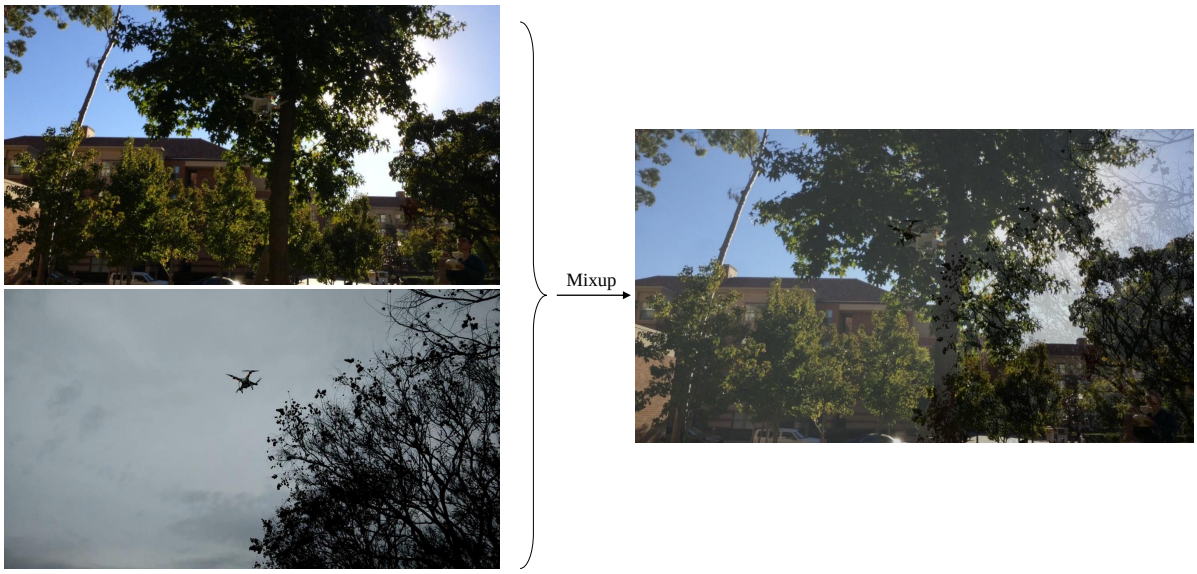


图 5.8 Mixup 图像混合效果示意图 ($\alpha = 0.5, \lambda = 0.292$)

5.1.5 分割掩模后处理

为了解决语义分割模型上采样，尤其是转置卷积而导致预测结果出现的毛刺现象，以及一些孤立的假阳性区域，使用全连接件随机场[119](Conditional Random Fields, CRF) 来对分割掩模进行平滑。它接收模型的预测和原图，通过不断迭代最小化能量函数输出优化后的分割结果。其能量函数为：

$$E(\boldsymbol{x}) = \sum_{i} \theta_i (x_i) + \sum_{ij} \theta_{ij} (x_i, x_j) \tag{5.7}$$

上式中，x_i, x_j 表示 i, j 位置处的像素标签，$\theta_i (x_i)$ 为一元势能函数，$\theta_{ij} (x_i, x_j)$ 为二元势能函数，两者的表达式如下：

$$\theta_i (x_i) = -\log P (x_i) \tag{5.8}$$

$$\begin{aligned} \theta_{ij} (x_i, x_j) &= \mu (x_i, x_j) \sum_{m=1}^{K} w_m \cdot k^m (f_i, f_j) \\ \mu (x_i, x_j) &= 1 \text{ if } x_i \neq x_j \text{ else } 0 \end{aligned} \tag{5.9}$$

上式中，$P (x_i)$ 表示语义分割模型输出的位置 i 处像素的类别概率值，概率值越高，一元势能值越小，保证分类准确率。二元势能函数考虑任意两个标签不相同的像素点，刻画像素之间的相关性，$k^m (f_i, f_j)$ 表示依赖于图像某种特征 (f) 的高斯核函数，在这里主要选取像素点的位置和色彩强度两种特征，w_m 为对应的权重。高斯核函数的具体计算公式如下：

$$k^m (f_i, f_j) = w_1 \exp\left(-\frac{\|p_i - p_j\|^2}{2\sigma_\alpha^2} - \frac{\|I_i - I_j\|^2}{2\sigma_\beta^2} \right) + w_2 \exp\left(-\frac{\|p_i - p_j\|^2}{2\sigma_\gamma^2} \right) \tag{5.10}$$

式 (5.10) 中，p, I 分别代表像素的位置和颜色强度，$\sigma_\alpha, \sigma_\gamma, \sigma_\beta$ 是核函数的超参数。公式的第一部分又被称为外观核函数 (Apperance Kernel)，用以重新估计像素所属类别，第二部分又被称为平滑核函数 (Smooothness Kernel)，负责移除小的虚警孤立区域。

图 5.9 给出了全连接条件随机场对 U-Net 模型预测的优化结果，从中可以看出经过分割掩模后处理，区域的轮廓更加鲜明且符合实际。本章实验涉及到的语义分割模型都将默认采用全连接条件随机场作为标准后处理手段，参数上，$w_1 = 10, w_2 = 3, \sigma_\alpha = 80, \sigma_\beta = 13, \sigma_\gamma = 3$, 迭代次数为 5。

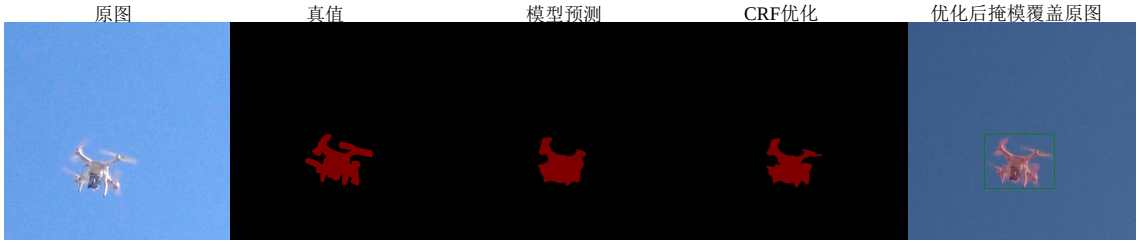


图 5.9 全连接条件随机场优化结果示意图 (为便于比较对图像经过截取处理)

5.2 实验结果与分析

实验环境配置如表4.1所示，首先进行 FCN, U-Net 和低慢小无人机语义分割模型 (以 S-Net 表示，下同) 的对比试验，优化器均采用 Adam, 初始学习率设置为 0.001，小批量数为 2，模型的权重参数使用 He 初始化方法，并采用交叉熵损失函数训练 10 个回合，最终的预测结果都经过全连接条件随机场进行优化。三个模型的测试精度如表 5.1 所示。

从表 5.1 可以看出，S-Net 的分割效果最好，其精度比 FCN 高出 9 个点，比 U-Net 高出将近 7 个点。性能提升的主要原因：一是采用 ResNet-101 了这种更深的特征提取网络，使得模型编码器部分的特征更加鲁棒和全面；二是膨胀卷积和膨胀空间金字塔池化机制在不进一步降低图像分辨率的前提上，扩大了感受野，且可以容忍物体一定的尺度变化；三是特征解码器部分同时采用双线性插值，特征融合和转置卷积，以保留图像重要特征，补充图像细节并实现特征到标签的解码和预测。图 5.10 给出了对比试验的部分测试样例，在背景复杂或者尺度变化的情况下，S-Net 依然可以很好地分割出来，但是当无人机目标很小的时候，三个模型的分割效果都不是很好，主要原因是物体占据的像素很少，而交叉熵损失函数作为通用的分类任务监督标准，无法很好地适应本章的无人机数据集。

表 5.1 低慢小无人机语义分割模型对比实验

模型	FCN	U-Net	S-Net
mIoU(%)	53.6	55.9	**62.5**

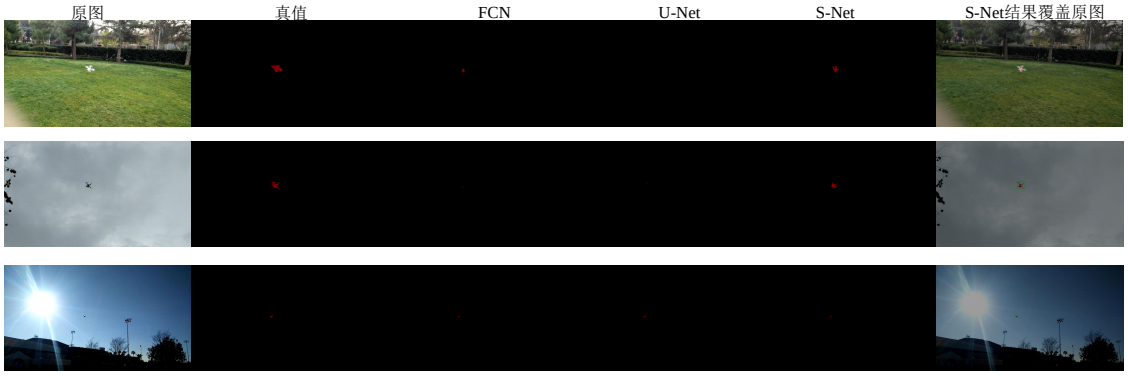


图 5.10 对比实验结果样例

下面，将5.1.3节的新损失函数和5.1.4节的 Mixup 加入到 S-Net 中，以进一步提升无人机的分割效果。

(1) 焦点损失函数

测试数据集对焦点损失函数中的 γ 和 α 参数敏感度，寻找最优的组合。以 Adam 为优化器，初始学习率设为 0.001，小批量数为 2，设置拥有不同 (γ, α) 数值组合的焦点损失函数分别对 S-Net 训练 10 个回合，测试的结果如表 5.2 所示。

表 5.2 焦点损失函数最优参数寻找

(γ, α)	(2,0.25)	(2,0.75)	(2,0.95)	(1,0.75)	(1,0.95)	(0,0.95)
mIoU(%)	55.4	62.6	63.3	62.9	64.1	**64.4**

从上表反映的趋势来看，γ 越大，模型的精度越低，增大 α，模型的精度越高。参数 γ 通过幂的形式控制对模型错分样本的惩罚力度，但是可能由于概率进行幂运算之后损失值的数量级大幅降低，导致反向传播时模型参数更新出现偏差，而 α 可以看作一个权重系数，对正负样本的损失进行配比，加大 α 的值使得模型主要关注正样本的分类，忽略掉那些易分的负样本。根据表格结果，且继续提高 α 后精度没有实质性的提升，因此本章将选取 (0,0.95) 作为 (γ, α) 的最终取值。

(2) 焦点损失函数与 Dice 损失函数组合

焦点损失函数的损失值一般比 Dice 损失函数小得多，为了防止模型的梯度方向被后者主导，进行一组实验来探究模型精度对焦点损失函数的权重系数 λ 的敏感性。模型的训练配置与 (1) 相同，(γ, α) 取最优值 (0,0.95)，测试的结果如表 5.3 所示。

表 5.3 焦点损失函数权重系数对模型精度的影响

λ	1	10	100	1000
mIoU(%)	62.3	63.5	**64.9**	64.7

当 λ 较小时，焦点损失函数的贡献不足，导致正负样本的学习力度不够，而当 λ 过大时，整体的损失函数又退化成焦点损失函数，无法凸显 Dice 损失函数对分割掩模轮廓的优化效果，因此本章将采取中间值 100 作为最终权重系数 λ 的取值。

(3) Mixup 图像混合

Mixup 中 Beta(α, α) 分布里的参数 α 直接按照原论文的建议设为 0.4，将其作为一种增强模型泛化能力的手段加入到模型中，并实施消融实验 (Ablation Study) 证明其有效性。实验结果如表 5.4 和图 5.11 所示。

表 5.4 消融实验 (F 代表焦点损失函数，D 代表 Dice 损失函数，M 代表 Mixup)

模型	S-Net	S-Net+F	S-Net+F+D	S-Net+F+D+M
mIoU(%)	62.5	64.4	64.9	**65.8**

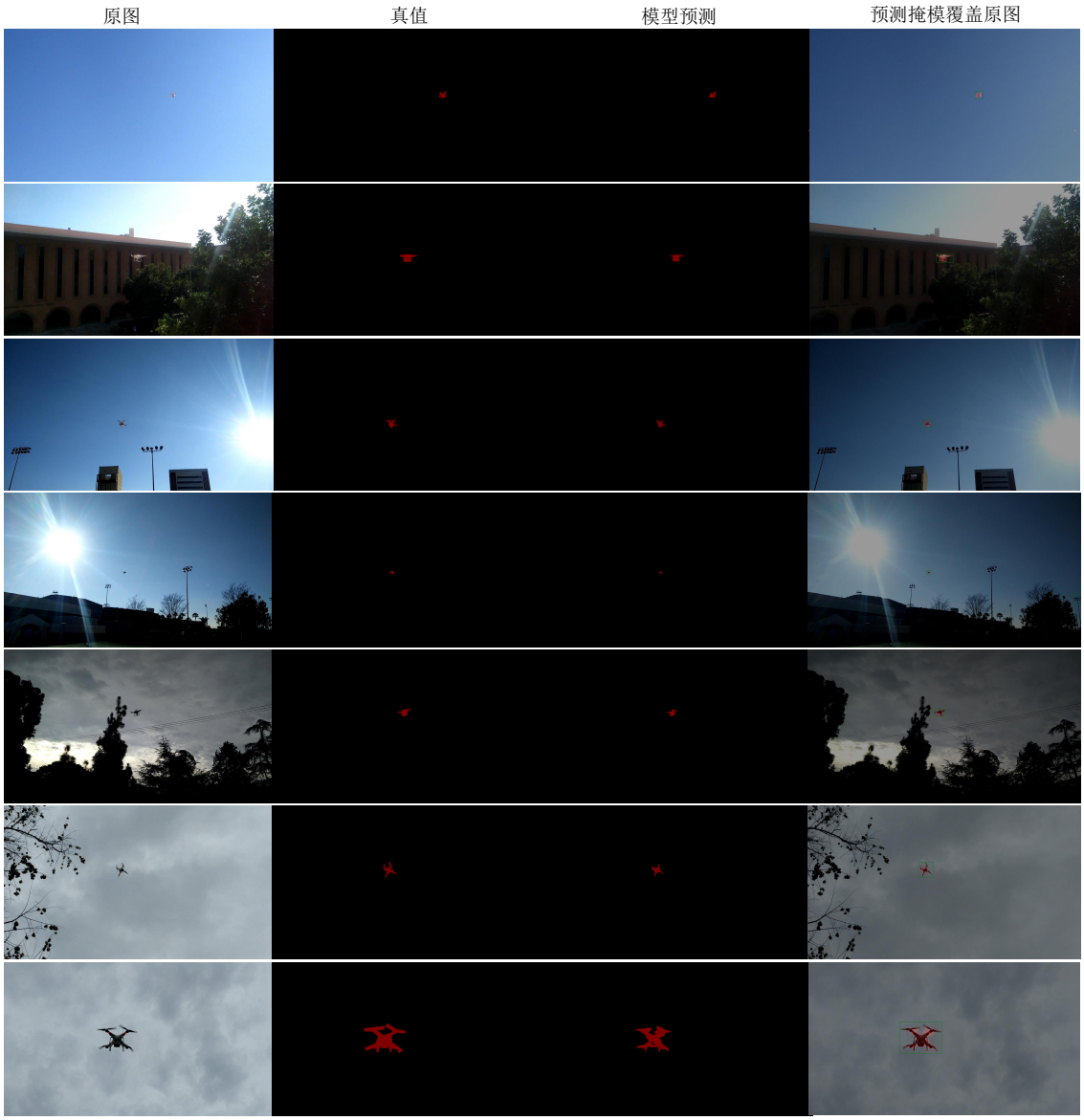


图 5.11 低慢小无人机弱监督语义分割结果示例

实验所用的损失函数均采用前文确立的最优值，在其基础上加入 Mixup 图像混合技术后，模型的精度提升了近 1%，说明了增加对抗样本训练可以控制网络的复杂度，并

与其他手段一起提升模型的性能。此外，由于显卡显存的限制，训练时的小批量数最多只能达到 2，后续如果可以加大，那么生成的对抗样本将会具有更加丰富的多样性，从而进一步促进分割的精度增长。

图 5.11 展示了表 5.4 中 S-Net+F+D+M 组合下的无人机分割结果样例。改进后的模型对处于复杂背下，比如强光或弱光，以及大小存在多尺度变化的无人机都具有良好的分割效果，预测的像素掩模也贴合无人机的形状，(极) 小目标也可以实现大部分中心区域的识别，这些主要是膨胀空间金字塔池化机制，焦点损失函数，Dice 损失函数，Mixup 以及更深的特征提取网络组合下起的作用，这也说明了本章所做工作的有效性。此外，值得一提的是，本章的语义分割算法是在弱监督的背景下进行的，即以通过标注框获取的“伪像素标签”这种不精确的区域，而不是真值来指导模型训练，实验证明，本章提出的弱监督分割系统依然具有十分可观的效果，既减少了图像像素标注的成本，又能够达到不错的分割精度，进一步加深了对低慢小无人机的检测识别研究，以满足不同的任务需求。

5.3 本章小结

本章在第四章目标检测的基础上，进一步加深低慢小无人机探测识别技术的研究深度，提出了一种弱监督语义分割系统实现低慢小无人机的像素级定位。首先考虑到像素标注相比框标注带来成本的增加，利用 GrabCut 算法抠取出无人机的像素区域生成“伪像素标签”，作为模型的监督信息。接着分析研究了典型语义分割模型的结构特点和缺陷，通过引入膨胀卷积和膨胀空间金字塔池化机制确立了本章的语义分割模型。然后针对低慢小无人机数据集中存在的小目标、多尺度、复杂背景等问题加入了焦点损失函数，Dice 损失函数，Mixup 图像混合技术和全连接条件随机场模块。最后通过实验证明了语义分割模型和改进手段的有效性，表明本章提出的低慢小无人机弱监督语义分割系统的可行性。

6 总结与展望

6.1 论文工作总结

随着低慢小无人机使用数量的急剧上升，与之对应的监管和反制措施的需求也日益增加。图像探测作为其中的一种手段，具有直观、高效、成本低廉、容易部署维护的特点。本文开展了基于深度学习的视觉探测低慢小无人机技术研究，并且针对现实场景中无人机可能为 (极) 小目标，出现多尺度范围变化，以及背景复杂易被干扰的情况，分别设计并实现了面向可见光和红外热成像的目标检测算法，以满足全天候监测。此外，为了试图加深低慢小无人机的探测效果，提出了一种基于标注框弱监督语义分割系统，实现了像素级定位，并在可见光数据集上进行了实验验证。

本文的主要工作内容如下：

(1) 低慢小无人机数据集制作。概述了权威公开数据集 PASCAL VOC 和 COCO，并根据 PASCAL VOC 的标准分别收集了不同场景，不同无人机形态和不同距离下的 156569 张可见光图像，5546 张红外热图像，随后划分了训练集，验证集和测试集，并进行了相应的标注，分析了算法的评价指标，保证了后续的算法实验和分析可以顺利进行。

(2) 深度神经网络基本结构分析及其优化、泛化算法研究。引入了本文所使用的前馈神经网络，对其中的神经元，非线性激活函数，用于参数更新的反向传播算法做了建模，比较和分析。然后根据低慢小无人机的数据集特点，深入研究了合适的诸如梯度下降、学习率调整等模型优化方法和诸如正则化、数据增强等的提高泛化能力的技术手段。紧接着剖析了具体应用于无人机图像数据的卷积神经网络的基本模块，最后阐述了用于本文探测任务中负责图像特征提取的几种卷积神经网络模型。

(3) 设计并实现了面向低慢小无人机的目标检测算法。首先研究了基于锚框的二阶检测算法 Faster R-CNN 和一阶检测算法 SSD 的工作原理，并分析了两者的优缺点，在进行了基准实验后，确立了以稳定性和准确率都较好的 Faster R-CNN 为基础模型的结论。然后针对可见光数据集中低慢小无人机和风筝，飞鸟这两类相似物体存在多尺度以及复杂背景干扰的问题下，分别使用特征金字塔、可变形卷积和自注意力机制、GIoU 回归损失函数、在线困难样本挖掘技术对 Faster R-CNN 进行了改进，并通过实验证明了改进后的模型可以很好地完成可见光数据集下的检测任务。紧接着在可见光检测模型研究的成果上，以 Faster R-CNN 为基础，针对红外数据集待检物体属于小目标，甚至极小目标的情况，除了引入了特征金字塔和在线困难样本挖掘技术来增加小目标的细节信息，辅助模型充分学习外，还使用可以保持特征图高分辨率的 HRNet 作为新的特征提取网络，以最大程度地保持小目标的位置信息，提升检测效果，最后的实验表明了改进手段的有效性和必要性。这两种模态数据下的检测模型基本满足了全天候监测低慢小无人机的需求。

(4) 设计并实现了面向低慢小无人机的弱监督语义分割算法。为了加深对低慢小无人机的定位效果，丰富视觉探测手段，提出了一种弱监督语义分割系统。首先利用目标检测中的标注框来生成“伪像素标签”，作为模型的监督信号，并大大降低了进行人工像素标注的成本，然后在分析了典型语义分割算法 FCN, U-Net 的基础上，给出了一种适用于本文的语义分割模型。接着针对低慢小无人机多尺度以及出现场景复杂的特点，引入了焦点损失函数，Dice 损失函数和图像混合技术，以加强算法的性能，最后通过实验，验证了该弱监督语义分割算法的可行性，拓宽了本文的研究内容和工作力度。

6.2 论文工作展望

由于时间，精力和个人能力的原因，本文的研究工作还存在一些问题，后续还可以从以下几个方面进行提升，拓展和加深：

(1) 进行多模态融合。本文为了实现全天候监测，采取白昼使用可见光，夜晚采用红外热成像的策略，单独分析这两种数据集的特点，然后针对性地设计不同的检测结构。这两个模态的数据并没有直接混合放在一起，并构建一个统一的模型来进行检测，这是因为神经网络目前还无法直接处理多模态的原始数据，很难既同时学习到可见光图像、红外图像的特征提取，又能给出两者的融合方式，况且也很难设计出指导网络学习某种数据融合方式的损失函数。但是可见光和红外这两种模态的数据又各有好处，可以相互结合，尤其是低光照，能见度低的白昼环境，红外可以弥补可见光的一些不足。后续可采用的一种折中方法是使用两个特征提取网络分别提取这两类模态的图像特征，然后分配权重进行特征图叠加，以实现融合，最后送入一个检测器进行无人机检测，从而增强模型的鲁棒性。此外，值得注意的是，为了让两类数据的特征可以融合，需要对可见光和红外摄像头进行标定，确保视野一致，保证两者的图像内容完全一样。

(2) 增大数据集体量。继续收集不同场景，不同光照，不同距离，不同程度遮挡等情况下的不同型号的低慢小无人机图像，以加深整体数据的覆盖面，让模型可以训练地更加充分，这也是提升模型性能最直接有效的方式之一。

(3) 加快模型的推理速度。可以考虑结合二阶和一阶检测算法的结构，设计出一种既包含级联回归，又可以舍弃费时的区域抠取操作的模型，同时拥有较快的处理速度和检测精度。也可以从模型压缩的角度入手，对模型进行剪枝，进而实现模型量化，模型部署。

(4) 引入视频检测。低慢小无人机的出现是一个连续的过程，存在一定的飞行轨迹和规律，而针对单帧图像的检测舍弃了时间这个维度。今后可以考虑使用其他的深度神经网络模型，比如循环神经网络，来一次性处理连续的视频图像，从而实现前后检测结果之间的自洽，在一定程度上消除漏检。

致　谢

三年的研究生生涯即将结束，正值硕士论文完成之际，谨向那些一直帮助我，鼓励我，支持我的人表示最诚挚的谢意！

首先我要感谢我的导师苏岩教授，苏老师治学严谨，知识渊博，眼界开阔，为人正直大方，与苏老师相处的这几年使我受益匪浅。苏老师为我提供了一个优渥，宽松和自由的科研环境，并时常关心询问我的学习方向和进度，当我迷茫和焦躁时，耐心地给予我帮助和指导，让我没有后顾之忧。同时苏老师身上的这种热情，乐观积极的人生态度也不断影响着我，激励我不断前行。

感谢 MEMS 惯性技术研究中心的周同副教授，周老师在我的选题，开题和论文写作的过程中提供了极大的帮助和许多深刻，富有建设性的意见，并且无私地向我提供设备支持我的实验研究。感谢朱欣华老师在百忙之中审阅我的论文，并对我论文整体结构和写作逻辑给出了很多有用的意见。

感谢教研室的姚速锐、郭明环、林晨、赵志鑫、曹豫、丁垒、徐川等师兄，感谢他们在我平时的学习上、生活上对我的帮助和指导，感谢郭俊幸、张奔、王欢、许荆宇等师弟对我科研工作的支持。感谢 MEMS 组同级的刘雨晨、王俊杰、刘雨东、孙伟容，孙敏杰、朱红赛等人，他们与我彼此之间相互扶持，一同创造了难忘的回忆，是我人生中宝贵的财富。

感谢我的室友章天平、张宏乐和李昱龙，好朋友占志远、凌丽阳、李远志和曹也，感谢他们在我繁重的学习生活之余的陪伴和关心。

最后，我要感谢我的父母，感谢他们对我的培育，包容，支持和无私的关爱，没有他们就没有我现在的一切，愿我的父母能够一直健康幸福下去。

参考文献

[1] 吴浩, 徐婧, 李刚. 民用无人机探测与反制技术现状及发展析 [J]. 飞航导弹, 2020:1–7.

[2] 刘子豪. 基于深度学习的无人机检测算法研究 [D]. 上海: 东华大学, 2020.

[3] 屈旭涛, 庄东晔, 谢海斌. “低慢小”无人机探测方法综述 [J]. 指挥控制与仿真, 2020:1–8.

[4] 朱英. 我国无人机实名登记数量逾 33 万架 [EB/OL]. http://www.gov.cn/xinwen/2019-05/31/content_5396457.htm.

[5] 韩晓飞, 蒙文, 李云霞等. 激光防御低慢小目标的关键技术分析 [J]. 激光与红外, 2013, 43(8):867–871.

[6] 蒋镕圻, 白若楷, 彭月平. 低慢小无人机目标探测技术综述 [J]. 飞航导弹, 2020(09):100–105.

[7] 马雯, 叱干小玄. 反无人机技术发展研究 [J]. 航空兵器, 2020:1–6.

[8] Pieraccini M, Miccinesi L, Rojhani N. RCS measurements and ISAR images of small UAVs[J]. IEEE Aerospace and Electronic Systems Magazine, 2017, 32(9):28–32.

[9] 向凡夫, 郝冬青, 吴鹏. 针对低慢小目标的雷达信号处理算法 [J]. 指挥控制与仿真, 2019(4):11.

[10] Xu D, Zhang H. Study of Low-altitude Slow and Small Target Detection on Radar[C]//2017 5th International Conference on Machinery, Materials and Computing Technology. Beijing: Clausius Scientific Press Journal, 2017.

[11] 李琴, 黄卡玛. 低空小型无人机雷达探测距离仿真分析 [J]. 无线电工程, 2018, 48(4):303–307.

[12] Patel J S, Fioranelli F, Anderson D. Review of radar classification and RCS characterisation techniques for small UAVs ordrones[J]. IET Radar, Sonar & Navigation, 2018, 12(9):911–919.

[13] 吕文. 基于信号识别和 TDOA 定位的无人机监测方法研究 [J]. 中国无线电, 2016(6):71–73.

[14] Shoufan A, Al-Angari H M, Sheikh M F A, et al. Drone pilot identification by classifying radio-control signals[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(10):2439–2447.

[15] Bisio I, Garibotto C, Lavagetto F, et al. Unauthorized amateur UAV detection based on WiFi statistical fingerprint analysis[J]. IEEE Communications Magazine, 2018, 56(4):106–111.

[16] Harvey B, O’ Young S. Acoustic detection of a fixed-wing UAV[J]. Drones, 2018, 2(1):4.

[17] 王威, 安腾飞, 欧建平等. 无人机被动音频探测和识别技术研究 [J]. 声学技术, 2018(01):89–93.

[18] 常先宇. 基于声阵列的无人机实时检测和定位系统及方法 [D]. 杭州: 浙江大学, 2019.

[19] Anwar M Z, Kaleem Z, Jamalipour A. Machine learning inspired sound-based amateur drone detection for public safety applications[J]. IEEE Transactions on Vehicular Technology, 2019, 68(3):2526–2534.

[20] 陈超帅, 王世勇等. 大疆无人机目标红外辐射特性测量及温度反演 [J]. 光电工程, 2017, 44(4):427–434.

[21] Andraši P, Radišić T, Muštra M, et al. Night-time detection of uavs using thermal infrared camera[J]. Transportation Research Procedia, 2017, 28:183–190.

[22] 刘连伟, 杨淼淼, 邹前进等. 无人机红外辐射建模与图像仿真 [J]. 红外与激光工程, 2017, 46(6):207–213.

[23] 高学志. 无人机红外图像和可见光图像配准融合算法研究 [D]. 哈尔滨: 哈尔滨理工大学, 2019.

[24] 卢鑫鑫. 基于深度学习的无人机检测算法研究 [D]. 武汉: 华中科技大学, 2019.

[25] Uijlings J R, Van De Sande K E, Gevers T, et al. Selective search for object recognition[J]. International Journal of Computer Vision, 2013, 104(2):154–171.

[26] 白文鹏. 基于深度学习的小尺度目标检测研究与实现 [D]. 西安: 西安理工大学, 2020.

[27] Lienhart R, Maydt J. An extended set of haar-like features for rapid object detection[C]//Proceedings. International Conference on Image Processing. Washington: IEEE, 2002:I–I.

[28] Wang X, Han T X, Yan S. An HOG-LBP human detector with partial occlusion handling[C]//2009 IEEE 12th International Conference on Computer Vision. Washington: IEEE, 2009:32–39.

[29] Zhu Q, Yeh M C, Cheng K T, et al. Fast human detection using a cascade of histograms of oriented gradients[C]//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington: IEEE, 2006:1491–1498.

[30] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington: IEEE, 2005:886–893.

[31] Ojala T, Pietikäinen M, Mäenpää T. Gray scale and rotation invariant texture classification with local binary patterns[C]//European Conference on Computer Vision. Berlin: Springer, 2000:404–420.

[32] Ahonen T, Hadid A, Pietikäinen M. Face recognition with local binary pat-

terns[C]//European Conference on Computer Vision. Berlin: Springer, 2004:469–481.

[33] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2):91–110.

[34] Liu C, Yuen J, Torralba A. Sift flow: Dense correspondence across scenes and its applications[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 33(5):978–994.

[35] Cortes C, Vapnik V. Support-vector networks[J]. Machine learning, 1995, 20(3):273–297.

[36] Lin C F, Wang S D. Fuzzy support vector machines[J]. IEEE Transactions on Neural Networks, 2002, 13(2):464–471.

[37] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[C]//Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington: IEEE, 2001:I–I.

[38] Safavian S R, Landgrebe D. A survey of decision tree classifier methodology[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1991, 21(3):660–674.

[39] Divvala S K, Efros A A, Hebert M. How important are “deformable parts” in the deformable parts model?[C]//European Conference on Computer Vision. Berlin: Springer, 2012:31–40.

[40] Felzenszwalb P F, Girshick R B, McAllester D, et al. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 32(9):1627–1645.

[41] Girshick R, Iandola F, Darrell T, et al. Deformable part models are convolutional neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE, 2015:437–446.

[42] Ouyang W, Wang X. Joint deep learning for pedestrian detection[C]//Proceedings of the IEEE International Conference on Computer Vision. Washington: IEEE, 2013:2056–2063.

[43] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems. New York: Curran Associates, 2012:1097–1105.

[44] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE, 2009:248–255.

[45] 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. Washington: IEEE, 2014:580–587.

[46] 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, 2015:1904–1916.

[47] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE International Conference on Computer Vision. Washington: IEEE, 2015:1440–1448.

[48] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in Neural Information Processing Systems. New York: Curran Associates, 2015:91–99.

[49] Dai J, Li Y, He K, et al. R-fcn: Object detection via region-based fully convolutional networks[C]//Advances in Neural Information Processing Systems. New York: Curran Associates, 2016:379–387.

[50] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE International Conference on Computer Vision. Washington: IEEE, 2017:2961–2969.

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

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

[53] Law H, Deng J. Cornernet: Detecting objects as paired keypoints[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018:734–750.

[54] Duan K, Bai S, Xie L, et al. Centernet: Keypoint triplets for object detection[C]//Proceedings of the IEEE International Conference on Computer Vision. Washington: IEEE, 2019:6569–6578.

[55] Tian Z, Shen C, Chen H, et al. Fcos: Fully convolutional one-stage object detection[C]//Proceedings of the IEEE International Conference on Computer Vision. Washington: IEEE, 2019:9627–9636.

[56] 郑婷婷, 杨雪, 戴阳. 基于关键点的 Anchor Free 目标检测模型综述 [J]. 计算机系统应用, 2020, 29(8):1–8.

[57] 张鹏飞. 低空空域无人机入侵检测研究 [D]. 西安: 长安大学, 2019.

[58] 岳子涵. 无人机管制系统中的视觉识别技术研究 [D]. 武汉: 华中科技大学, 2019.

[59] 于鹏. 基于无人机图像的识别算法研究 [D]. 哈尔滨: 哈尔滨工程大学, 2018.

[60] 虞晓霞, 刘智, 耿振野, et al. 一种基于深度学习的禁飞区无人机目标识别方法 [J]. 长春理工大学学报 (自然科学版), 41.

[61] 何志祥, 胡俊伟. 基于深度学习的无人机目标识别算法研究 [J]. 滨州学院学报, 2019(2):17–23.

[62] 张辉, 张文武. 基于 CNN 的无人机目标检测算法比较与分析 [C]//第八届中国指挥控制大会. 北京: 电子工业出版社, 2020:288–292.
[63] 甘雨涛. 卷积神经网络在低空空域无人机检测中的研究 [D]. 成都: 电子科技大学, 2019.
[64] 蒋兆军, 成孝刚, 彭雅琴等. 基于深度学习的无人机识别算法研究 [J]. 电子技术应用, 2017, 43(7):84–87.
[65] 周光兵. 反无人机系统中目标识别和跟踪算法研究 [D]. 青岛: 中国石油大学, 2018.
[66] The 1st Anti-UAV Workshop & Challenge[EB/OL]. https://anti-uav.github.io/.
[67] Everingham M, Eslami S A, Van Gool L, et al. The pascal visual object classes challenge: A retrospective[J]. International Journal of Computer Vision, 2015, 111(1):98–136.
[68] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//European Conference on Computer Vision. Berlin: Springer, 2014:740–755.
[69] PASCAL Visual Object Classes Challenge 2007 (VOC2007) Annotation Guidelines[EB/OL]. http://host.robots.ox.ac.uk/pascal/VOC/voc2007/guidelines.html.
[70] 周志华. 机器学习 [M]. 北京: 清华大学出版社, 2016: 30–31.
[71] Wang Y, Chen Y, Choi J, et al. Towards visible and thermal drone monitoring with convolutional neural networks[J]. APSIPA Transactions on Signal and Information Processing, 2019, 8.
[72] LabelImg is a graphical image annotation tool and label object bounding boxes in images[EB/OL]. https://github.com/tzutalin/labelImg.
[73] Image Polygonal Annotation with Python (polygon, rectangle, circle, line, point and image-level flag annotation)[EB/OL]. https://github.com/wkentaro/labelme.
[74] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE, 2015:3431–3440.
[75] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8):1798–1828.
[76] Schmidhuber J. Deep learning in neural networks: An overview[J]. Neural Networks, 2015, 61:85–117.
[77] 邱锡鹏. 神经网络与深度学习 [M]. 北京: 机械工业出版社, 2020: 79–90.
[78] Saul L K, Jaakkola T, Jordan M I. Mean field theory for sigmoid belief networks[J].

Journal of Artificial Intelligence Research, 1996, 4:61–76.

[79] Tóth L. Phone recognition with deep sparse rectifier neural networks[C]//2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Washington: IEEE, 2013:6985–6989.

[80] Liu W, Wen Y, Yu Z, et al. Large-margin softmax loss for convolutional neural networks.[C]//ICML. New York: The International Machine Learning Society, 2016:7.

[81] Hornik K, Stinchcombe M, White H, et al. Multilayer feedforward networks are universal approximators.[J]. Neural Networks, 1989, 2(5):359–366.

[82] Keskar N S, Mudigere D, Nocedal J, et al. On large-batch training for deep learning: Generalization gap and sharp minima[J]. arXiv preprint arXiv:1609.04836, 2016.

[83] Goyal P, Dollár P, Girshick R, et al. Accurate, large minibatch sgd: Training imagenet in 1 hour[J]. arXiv preprint arXiv:1706.02677, 2017.

[84] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[85] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks[C]//Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. New York: The Society for AI and Statistics, 2010:249–256.

[86] He K, Zhang X, Ren S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification[C]//Proceedings of the IEEE International Conference on Computer Vision. Washington: IEEE, 2015:1026–1034.

[87] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1):1929–1958.

[88] Bjorck N, Gomes C P, Selman B, et al. Understanding batch normalization[C]//Advances in Neural Information Processing Systems. New York: Curran Associates, 2018:7694–7705.

[89] Santurkar S, Tsipras D, Ilyas A, et al. How does batch normalization help optimization?[C]//Advances in Neural Information Processing Systems. New York: Curran Associates, 2018:2483–2493.

[90] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[J]. arXiv preprint arXiv:1502.03167, 2015.

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

[92] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of

the IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE, 2015:1–9.

[93] Lin M, Chen Q, Yan S. Network in network[J]. arXiv preprint arXiv:1312.4400, 2013.

[94] 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. Washington: IEEE, 2016:770–778.

[95] Paszke A, Gross S, Massa F, et al. Pytorch: An imperative style, high-performance deep learning library[C]//Advances in Neural Information Processing Systems. New York: Curran Associates, 2019:8026–8037.

[96] 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. Washington: IEEE, 2017:2117–2125.

[97] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2):91–110.

[98] Bay H, Tuytelaars T, Van Gool L. Surf: Speeded up robust features[C]//European Conference on Computer Vision. Berlin: Springer, 2006:404–417.

[99] Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision. Washington: IEEE, 2017:764–773.

[100] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. New York: Curran Associates, 2017:5998–6008.

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

[102] Wang X, Girshick R, Gupta A, et al. Non-local neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE, 2018:7794–7803.

[103] Cao Y, Xu J, Lin S, et al. GCNet: Non-local networks meet squeeze-excitation networks and beyond. arXiv 2019[J]. arXiv preprint arXiv:1904.11492.

[104] Rezatofighi H, Tsoi N, Gwak J, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE, 2019:658–666.

[105] Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining[C]//Proceedings of the IEEE Conference on Computer

Vision and Pattern Recognition. Washington: IEEE, 2016:761–769.

[106] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE, 2019:5693–5703.

[107] Sun K, Zhao Y, Jiang B, et al. High-resolution representations for labeling pixels and regions[J]. arXiv preprint arXiv:1904.04514, 2019.

[108] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE, 2017:4700–4708.

[109] Rother C, Kolmogorov V, Blake A. " GrabCut" interactive foreground extraction using iterated graph cuts[J]. ACM transactions on Graphics, 2004, 23(3):309–314.

[110] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical Image Computing and Computer-assisted Intervention. Berlin: Springer, 2015:234–241.

[111] Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected crfs[J]. arXiv preprint arXiv:1412.7062, 2014.

[112] Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(4):834–848.

[113] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[J]. arXiv preprint arXiv:1706.05587, 2017.

[114] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European conference on computer vision. Berlin: Springer, 2018:801–818.

[115] Wang P, Chen P, Yuan Y, et al. Understanding convolution for semantic segmentation[C]//IEEE Winter Conference on Applications of Computer Vision. Washington: IEEE, 2018:1451–1460.

[116] 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. Washington: IEEE, 2017:2980–2988.

[117] Milletari F, Navab N, Ahmadi S A. V-net: Fully convolutional neural networks for volumetric medical image segmentation[C]//4th International Conference on 3D Vision. Washington: IEEE, 2016:565–571.

[118] Zhang H, Cisse M, Dauphin Y N, et al. mixup: Beyond empirical risk minimization[J]. arXiv preprint arXiv:1710.09412, 2017.

[119] Krähenbühl P, Koltun V. Efficient inference in fully connected crfs with gaussian edge potentials[C]//Advances in Neural Information Processing Systems. New York: Curran Associates, 2011:109–117.

附　录

攻读硕士学位期间发表的论文和出版著作情况：

[1] **Zhentao Yu**, Tong Zhou, Yan Su, High-Precision Visual Localization and Dense Mapping Based on Visual SLAM for Indoor Environment [C]//IEEE 5th International Conference on Computer and Communications (ICCC), 2019, Chengdu, China.

[2] 周同，**余振滔**，国家发明专利《基于弱监督学习的痰涂片结核杆菌语义分割方法及系统》，申请号：202010804731.7.

[3] 周同，**余振滔**，国家发明专利《一种基于层次回归的轻量级图像探测农业驱鸟方法及系统》，申请号：202010804741.0.

[4] 周同，郭俊幸，**余振滔**，国家发明专利《一种基于图像探测的无人机识别方法》，申请号：202010366536.0.