第十届“泰迪杯”
数据挖掘挑战赛

优秀作品

作品名称：农田害虫图像识别

荣获奖项：特等奖并获泰迪杯

作品单位：南京理工大学

作品成员：郑睿智 盛猛猛 蔡鑫浩

指导老师：范金华

基于深度学习的农田害虫定位与识别研究

摘　要

农作物病虫害一直是影响粮食产量的一个主要原因，往往会由于发现不及时，以及错误判断病虫害种类，不能很好处理。基于虫情测报灯的虫情信息采集，提供了大量的农作物害虫图像。但是由于害虫图像背景复杂，有些目标过小，人工识别存在误判漏判，并且效率十分低下。而基于视觉手工特征的方法鲁棒性差，不适用于复杂的害虫待识别环境。近些年来，基于深度神经网络的人工智能技术得到快速发展，得益于算力的提升和大规模数据集的出现，深度学习已经在目标检测、人脸识别、语音识别等领域飞速发展 。虫情测报灯收集到了大量害虫图片，为基于深度学习的目标检测算法进行害虫识别提供了可能性。

通过分析害虫图片，我们发现数据集中存在严重的害虫种类长尾分布和待检测目标过小的问题。为了使得切分后的数据集能够更加适合模型的训练，我们首先先对数据集进行了扩充处理，主要工作是提取一部分数量少和面积小的目标，通过 copy and paste 和常规在线数据增强方法进行离线的数据增强。我们将原图像中 576 个目标扩充至 4401 个目标，然后进行了 8:2 训练集和测试集的划分，成功地提高了模型的鲁棒性。

为了能准确地识别出害虫的位置和类别，我们结合了当下主流的两阶段目标检测框架 Cascade Mask RCNN 进行实现，并使用 Swin-Transformer 模型作为我们的特征提取网络。相比于一阶段目标检测在速度上的性能，农作物害虫识别更偏向于对准确性的要求，所以我们使用两阶段目标检测算法作为基本的检测框架。Cascade Mask RCNN 在 Faster RCNN 的基础上解决了 misalignment 的问题，并添加了 multi-stage 的结构，并为每一个 stage 设立了不同的 IOU 阈值。Swin-Transformer 使用了基于窗口的自注意力机制，将 Transformer 模型很好地应用到了视觉任务中来。在此之后，我们结合 K-means 聚类实现锚框长宽比的确定，然后把 Smooth L1 损失函数和 Soft NMS 分别加入到网络的训练中。最后，我们用 SWA 和多模型融合进一步提升网络性能。

关键词：病虫害识别；目标检测；数据增强；多模型融合；随机权重平均

目 录

第一章 绪论

1.1 研究背景

中国是世界范围内的农业大国，耕地面积广大，占世界耕地面积的 7%，每年粮食产量可达 1.3 万亿斤以上，但是我国仍需要每年从外国进口大量大豆、玉米等农作物粮食，所以粮食产量现在对我们来说仍然十分重要。然而，农作物病虫害却深深地损害着农业的发展，是我国主要农业灾害之一，农作物在遭受病虫害之后，其整体生理机能会大大下降，从而导致植株瘦小，无法达到最优生产状态，进而会产量不高、经济效益低。农作物病虫害具有种类多、影响大、并时常暴发成灾的特点，对我国的国民经济和农业生产造成重大损失。农作物常见的有以下种类的病虫害：稻飞虱、小麦锈病、棉蚜、稻纹枯病、稻瘟病、蝗虫、麦类赤霉病等，并且已成为严重影响我国农业生产的重大病虫害。

目前针对农作物害虫的处理，还是基于人工的判断，这往往可能病虫害已经在范围内发作，并不能及时的进行病虫害的防治。除此之外，一些病虫害的识别往往需要很高的专家知识，对于经验较少的农民并不能准确发现和判断病虫害种类，由此可能造成农药的滥用和误用，以及农作物中会产生农药残留，影响农田的生态环境。由于农作物害虫的多样性和复杂性，简单的人工检测方式已经很难满足现代化农业科技发展的需求。

为此，随着目前农业科学技术的发展，各种崭新技术已经渐渐融入到日常的农业生产之中，能够最大程度减少人工检测这类繁琐操作，并且效率高，可以很及时的进行病虫害的警告和分析。尤其是人工智能深度学习领域，目标检测相关算法更能很好的结合到病虫害的检测和识别之中，是大规模农业生产虫害预防的有效手段。

1.2 问题分析

基于虫情测报灯的虫情信息采集，为深度学习提供了所必须的大数据训练基础，虫情测报灯可以在无人监管的情况下，实现自动的诱集、杀虫、虫体分散、拍照等作业，可以通过照片的形式采集出当前可能出现的病虫害的种类和病虫害的严重程度，如图 1-1 所示：

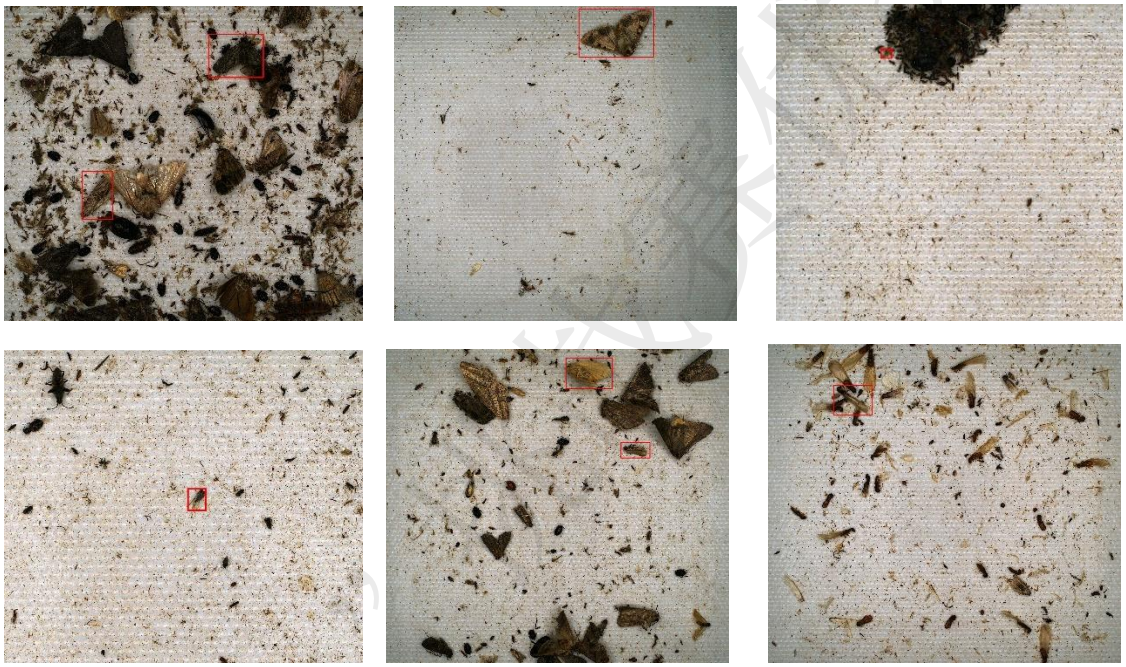


图 1-1 害虫目标定位图像

从图 1-1 中我们可以看到虫情测报灯的虫情信息采集到的照片，在白色背景下能够直接的看到害虫所在位置，获取农作物的受虫害影响程度以及害虫的种类和数量等情况，但是从图中也可以看出以下几个问题：

（1）每张图片害虫的种类并不单一，并且图片中的噪声点很多，会对目标害虫造成混淆。

（2）基于虫情测报灯采集的数据，对大多数的害虫种类来说，目标占据面积比较大，容易分别，但是对于一些目标小的害虫，与背景产生混淆，人眼很难准确判别，所以需要考虑小目标的单独处理；

（3）对于大量图片而言，并不是每张图片都会收集到害虫信息，在其中进行筛选也是一项繁琐的操作，更适合使用目标检测算法进行预测和重复操作；

（4）虫情分析相对来说更偏向于精确度的考虑，所以两阶段目标检测算法更适合虫情分析的应用场景。

1.3 相关研究

随着计算机视觉技术的飞速发展，目前与机器学习相关的人工智能相关技术得到了普遍应用，目标检测、人脸识别、语义分割、自然语音处理以及视频处理等领域相关算法也逐渐成熟，在各个现实工业领域得到快速应用，并且能够很大程度的提升以往人工的效率，更大程度的方便工业和技术发展。尤其是大数据时代和运算能力的提升，更是进一步的促进了人工智能的飞速发展。

当然，与农业技术结合也是人工智能应用的领域之一。针对农田害虫图像的识别，已经有一些方法进行实现，他们往往是基于传统的图像处理技术实现。谢成军[4]提出一种基于图像稀疏编码与空间金字塔模型相结合的害虫图像表示与识别方法。该方法能够利用大量非标注的自然图像块，用来构造过完备学习字典，并且运用该学习字典，从而实现对害虫图像的多空间稀疏表示。并且结合多核学习，设计了一种害虫图像识别算法。张洪涛[3]等人提出针对害虫目标的二值化图像提取出面积、周长、复杂度等 7 个形态学特征 并进行归一化处理；建立了9种害虫的模板库及隶属度函数并基于最小最大的原则进行模糊决策分析；对稻纵卷叶螟、棉铃虫等田间危害严重的9种害虫进行识别分类的识别率达86%以上。

深度学习神经网络相关算法，也很好的结合到农作物病虫害的相关应用。李玲[1]等人提出基于深度学习中数字图像识别的理论，构建了深层卷积神经网络，并使用网络模型对苹果树叶片进行分类试验，基于深度学习 MobileNet，修改输出的全连接层尺寸，搭建了 MobileNet 苹果树叶分类模型，实现了 Alternaria__Boltch（斑点落叶病）、Brown__Spot（褐斑病）、Grey__Spot（灰斑病）、Mosaic（花叶病）、Rust（锈病）5 种苹果树病害的识别。范世达等人[2]提出为解决柑橘种植过程中黄龙病检测不及时、检测成本较高的问题，初步探寻基于深度学习的柑橘黄龙病远程诊断方法。通过架设在田间的设备采集柑橘植株图像信息，利用深度学习相关算法构建柑橘黄龙病病害识别模型，在柑橘生长过程中实现黄龙病在线实时监测与病害远程诊断。

1.4 解决方案

在上述背景研究、问题分析和相关研究的基础之上，我们针对本次农作物害虫识别任务，采用基于两阶段的目标检测算法 Casecade RCNN 和结合 Swin-Transformer 的 Mask RCNN 算法作为基本框架，进行模型融合，并使用离线数据增强和在线数据增强组合，以及 SWA（随机权重平均）进一步进行性能的提升。

针对 1.2 提出的四个问题，相应采用以下解决方案：

1) 目前主流的目标检测算法都是基于两阶段的方法（Fast RCNN、Faster RCNN、Mask RCNN 等）以及一阶段的方法（YOLO 系列、SSD 等），相比于对识别速度的要求，农作物害虫检测更偏向于算法识别精确度的要求，故选用主流的两阶段目标检测方法作为模型：Cascade Mask RCNN，并使用目前在目标检测领域效果最好的 Swin-Transformer 模型作为我们的特征提取层。
2) 针对某些类别目标太小以及个别类别样本数目，我们采用 copy and paste 方法离线数据增强进行扩充数据集，并且采用随机反转、旋转、高斯噪声等在

线数据增强进一步在训练之前处理数据集。以及对训练集和测试集数据使用多尺度进行训练和预测，增加对小目标的识别效果。

3) 为了能更好的提升模型的识别效果，进一步使用每个模型不同的 epoch 的权值进行 SWA，并把使用两个模型 SWA 后的网络权重的预测结果进行 NMS，获取模型融合后的检测结果。

结合以上问题的提出和问题分析，具体解决方案如图 1-2 算法流程图所示：

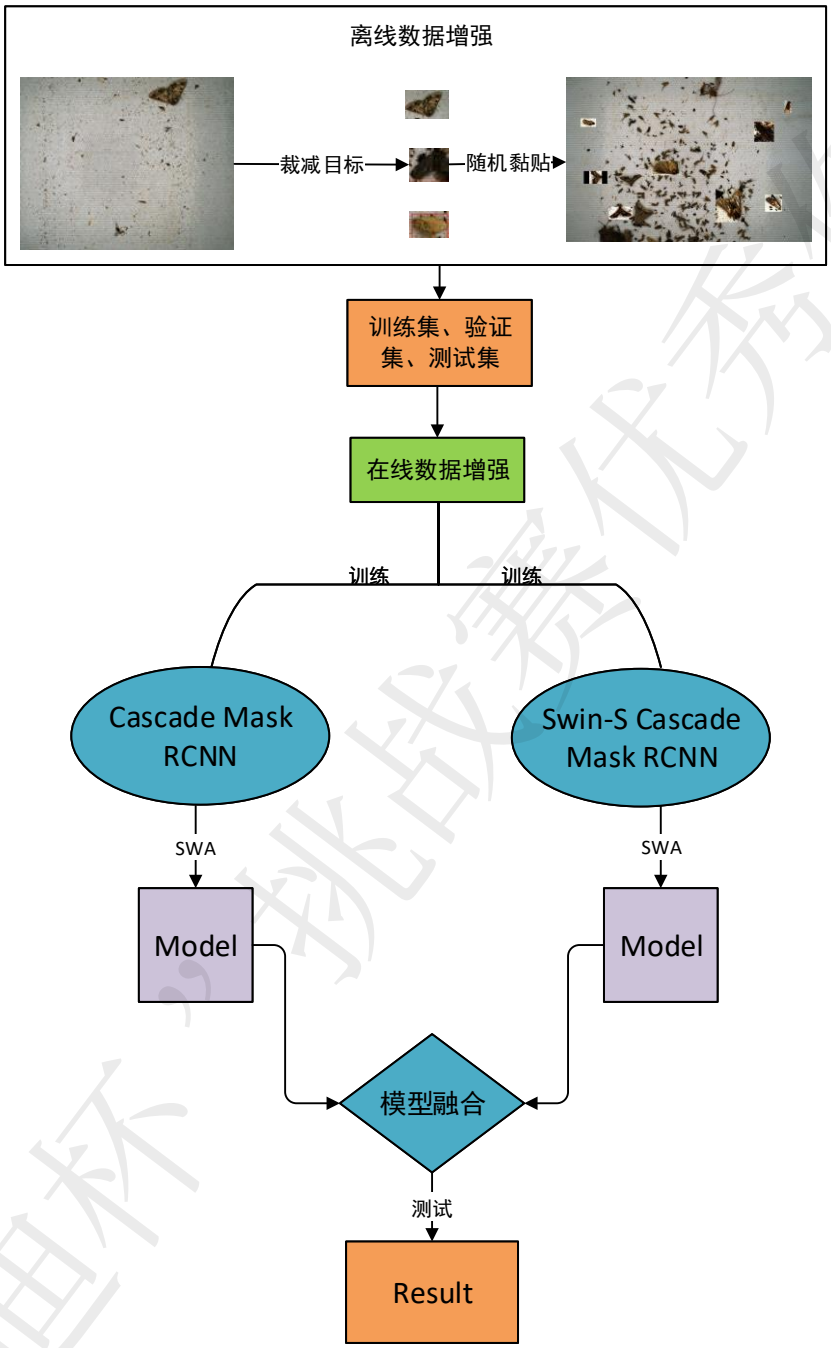


图 1-2 农作物害虫识别实现流程

1.5 章节安排

首先，根据前面相关内容的分析，我们针对农作物的害虫识别研究思路主要从下面几节相应介绍：

1) 第二章主要是介绍我们针对农作物害虫图像数据集的处理过程；
2) 第三章主要介绍使用到的目标检测模型 Cascade Mask RCNN 和特征提取网络 Swin-Transformer，以及我们针对农作物害虫识别所做出的改进和创新部分，并且叙述了如何训练起来两个模型，和如何使用 SWA、多模型融合，进一步提升性能。
3) 第四章主要介绍我们针对我们模型应用害虫检测中的试验部分，以及研究各种创新点有效性，所进行的消融研究。
4) 第五章主要是总结我们针对农作物害虫识别所做出的工作，与进一步展望。

第二章 数据分析处理

原数据集中含有 3015 张图片，根据图片虫子位置详情表提供的信息得知，其中有 576 张含有目标的图片，1637 张不含害虫的背景图片，还有 802 张用于测试最终结果的图像。通过分析数据集内的目标分布，大致发现数据集中存在的问题是：长尾数据分布问题、多尺度变换问题等，并针对这些问题进行数据集扩充，离线数据增强等操作，具体内容如下。

2.1 数据分析

2.1.1 长尾数据分布问题

我们统计了图像中含有目标的数量和分布情况，总计含有 1019 个目标。同时，我们发现所提供的样本数据中，存在典型的长尾数据分布情况。如图 2-1、图 2-2 所示，八点灰灯蛾、褐飞虱属和白背飞虱分别占总数据量的 24%、15%和 12%，属于数据集中的头部类，而歧角螟、草地螟、甘蓝夜蛾、线委夜蛾、紫条尺蛾、瓜绢野螟、稻螟蛉、地老虎、水螟蛾、豆野螟和干纹冬夜蛾所含有的总数不足总样本数的 5%，所有类别的样本数均只有几位数，属于数据集中的尾部类。

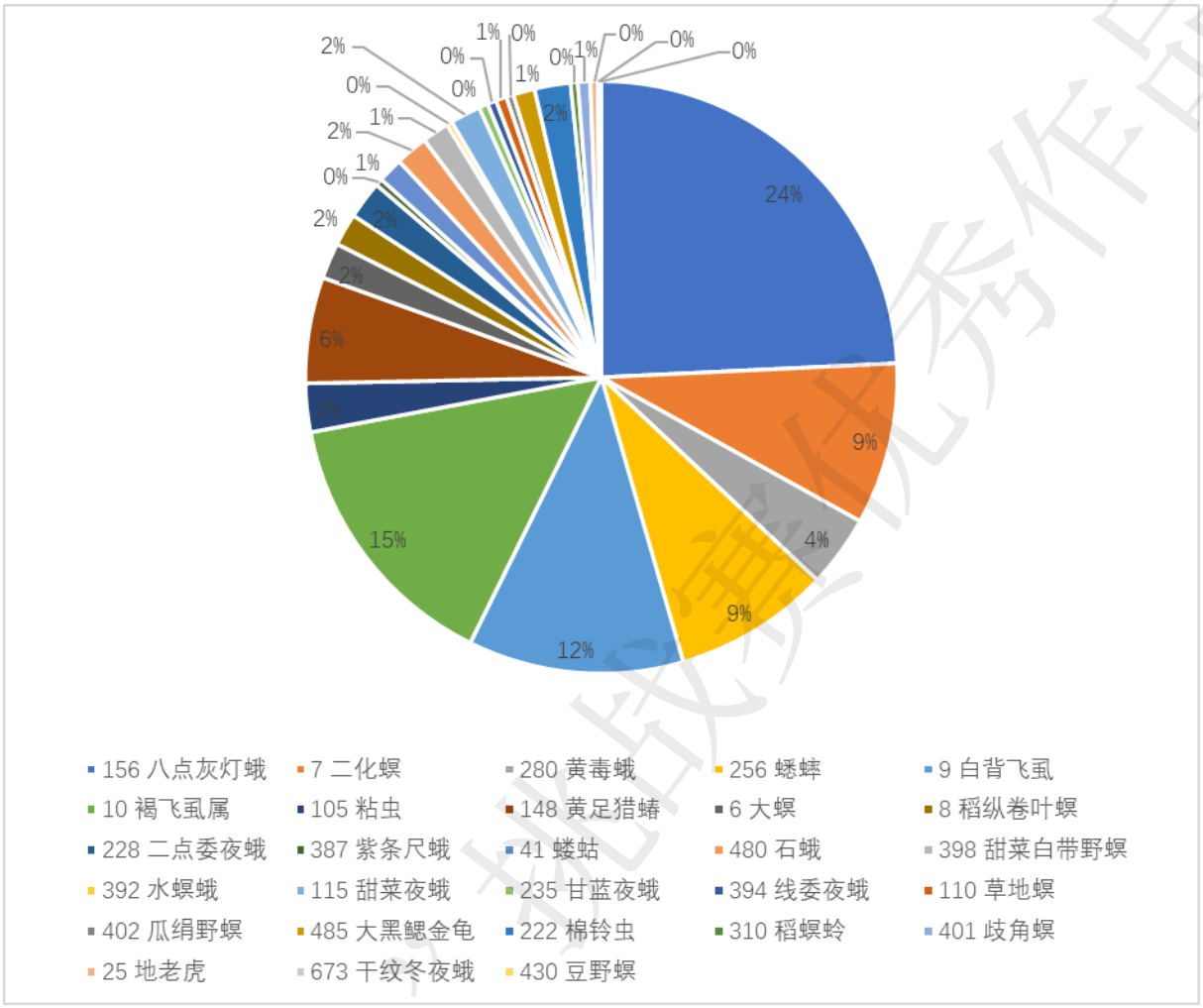


图 2-1 类别分布扇形图

长尾数据分布情况将会带来以下几种问题。首先是物体类别的不均衡问题，对于数据量充分的头部类别，模型显然能够地更好地学习该类目标的特征，而对于样本数目不足的尾部类显然达不到相同的效果。最终导致实际预测会向多数类侧重。然后，数据集中的尾部类目标数目普遍只有个位数，甚至有两类只有 1 个目标。这导致数据集划分训练集，验证集和测试集的过程中出现了问题。一般而言，针对长尾数据的方法有重采样和重加权。其中，重采样方法是增加尾部类目标在训练中出现的次数。对

于这类极端的长尾分布情况，单纯地进行重采样有可能导致模型仅学习到映射的特征，产生对该类过拟合的新问题。因此，我们采用对于目标的离线数据增强方法以应对这类问题。重加权的方法是从损失函数的角度出发，主要内容是增加尾部类损失计算的权重，使得头部类样本类带来的收益递减。

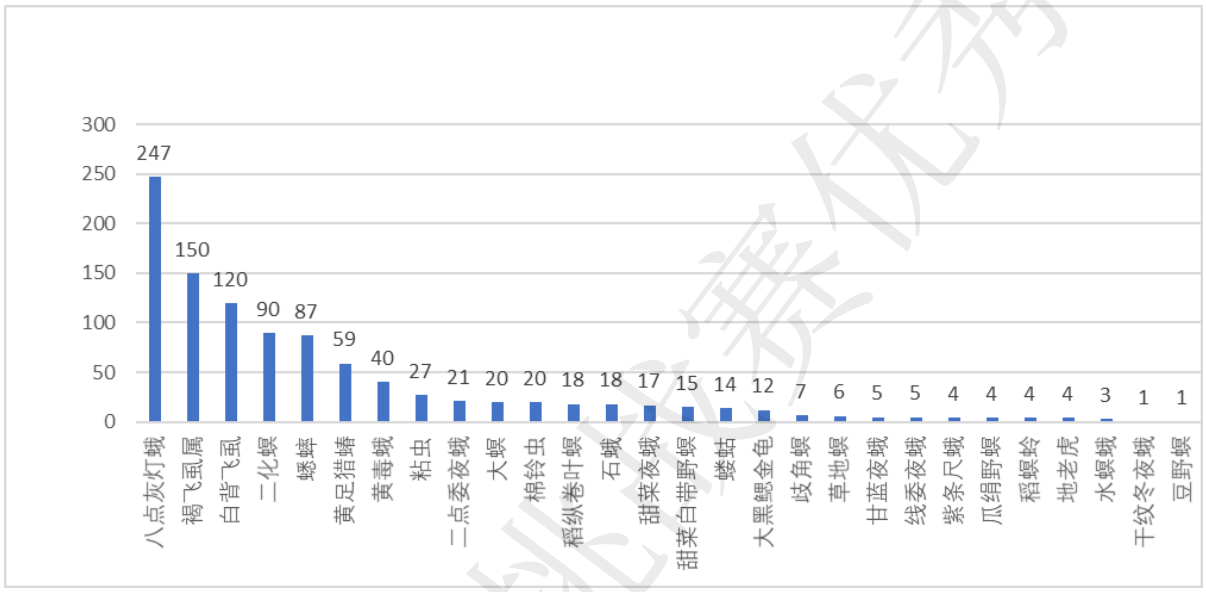


图 2-2 类别分布条形图

2.1.2 多尺度变化问题

我们关注了各类昆虫标注的面积信息(面积的计算方式为标注的长宽乘积)，发现了存在不同类别间面积差异较大。如图 2-3 所示，八点灰灯蛾、干纹冬夜蛾和蝼蛄这三类害虫与白背飞虱和褐飞虱属在面积平均值的比例上差距达到 40 倍以上。

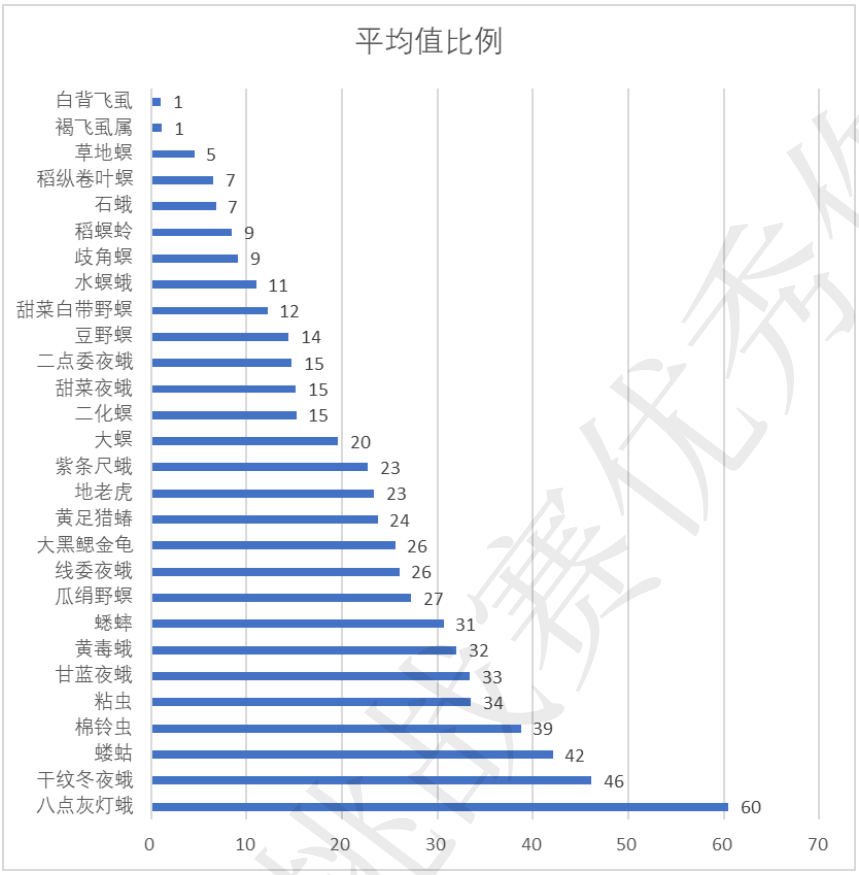


图 2-3 类间样本平均面积比

同时，我们也发现了同一类别间存在类似的问题。如 2-4 所示，八点灰灯蛾、二化螟和黄毒蛾本身就存在类内面积的差距就达到 10 倍以上，而且对于部分类还由于样本数目少的原因，我们无法得知这些类是否存在类内尺度不均匀的问题。

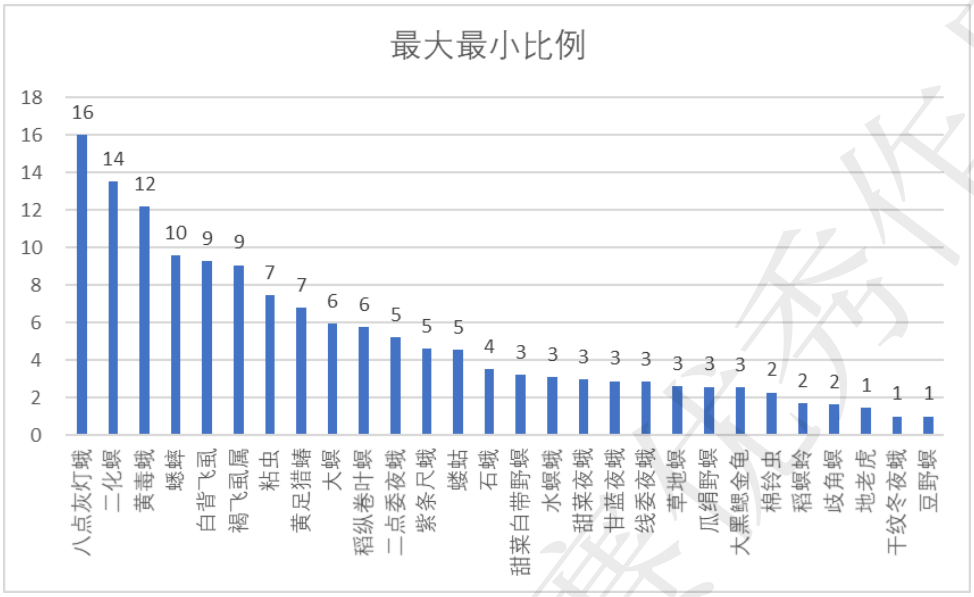


图 2-4 类内样本最大、最小面积比

以上提及的两个问题，类间和类内的多尺度问题的存在会给检测目标的召回率方面带来比较大的挑战，从而进一步影响到检测的准确率。多尺度问题也意味着现有样本中存在很多小目标。小目标自身像素点少在整图中占比低，还存在堆叠和遮挡等情况，这使得模型漏检可能性明显增加。针对数据集中存在的多尺度问题，通常采用特征金字塔的方法，将不同尺度的特征图进行融合，尽可能保留高分辨率特征图的特征；还可以通过设定不同尺度的锚框以适应不同尺寸的目标。

2.2 数据预处理

在数据分析一章中，我们通过分析得知了数据集中存在极端的长尾分布问题，需要通过离线数据增强平衡各类别目标数量后，才可以进行数据集训练集和测试集的划分。我们在预处理的工作分为两部分，首先是离线数据增强主要内容有常规数据增

强、copy and paste 和目标与背景融合操作。

2.2.1 常规数据增强

通过对数据集中的图像分析后，如图 所示，我们发现部分照片之间存在光照不均和有反光干扰的情况，这说明实际的拍摄环境光各异，模型需要额外学习不同光照条件下的昆虫特征。

图 2-5 不同光照下的图像效果

同时，各类昆虫的头部朝向各异，姿态也不相同。为了提升目标检测模型的鲁棒性，更加适应真实世界的数据，我们选择对数据进行随机翻转、随机对比度调整、随机亮度调整等数据增强操作，以达到增强数据多样性的效果。以某一类目标为例，裁减效果如 2-6 所示。

图 2-6 常规数据增强

2.2.2 Copy and Paste

在处理长尾数据分布问题上一般采用重采样方法解决，目的是为了提供样本数目少的尾部类在模型训练中更多的占比。但是，该类方法通常由于尾部类样本的单一性的原生问题而效果不佳。Copy and Paste 数据增强方法的思想是由 Golnaz Ghiasi 等人提出的一种简单而有效的数据增强方法。如 2-7 所示，作者通过实验论证了该方法能够建立在已有的数据增强方法之上带来提升，并且针对小目标的识别有一定的帮助。

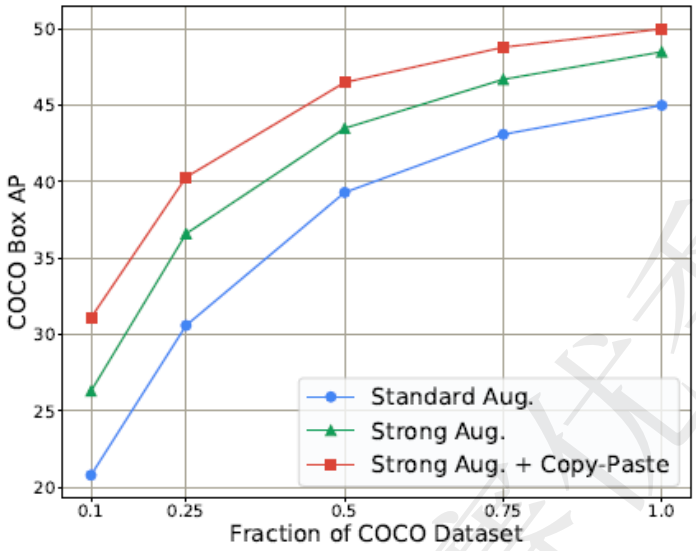


图 2-7 copy and paste 在 COCO 数据集中的效果

Copy and Paste 数据增强所基于的简单操作是，首先把需要平衡的类别的目标裁剪下来，然后随机挑选一些背景图片，把这些裁剪下来的目标进行随机粘贴在背景图片上，以此达到解决尾部类别样本数目少的问题，效果如图 2-8 所示：

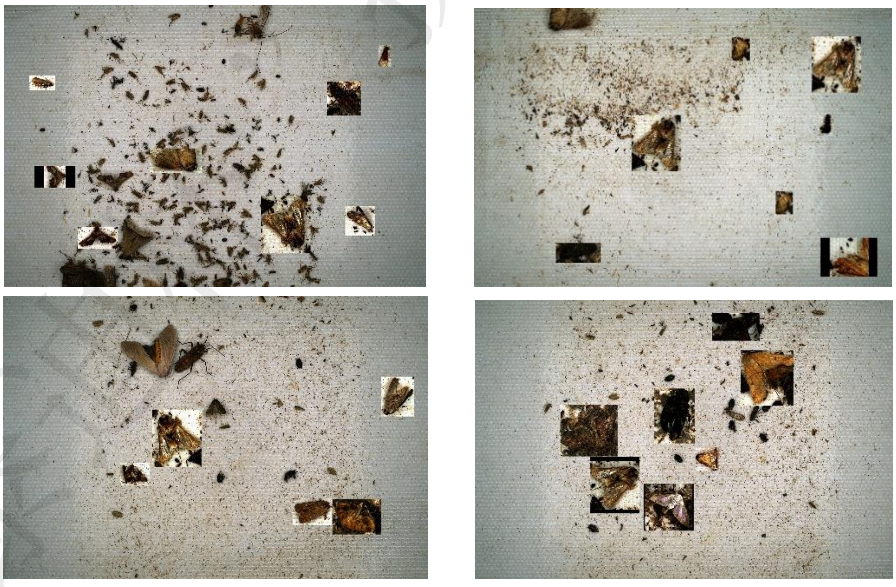


图 2-8 Copy and Paste 数据增强效果

在基本常规的数据增强，例如翻转、旋转、亮度等方法的基础之上，通过 copy and paste 数据增强操作有效地扩充了各类样本的数目，具体如表 2-1 所示。我们不仅提升了尾部类的占比，还丰富了模型在尾部类所学习到的特征，有效地平衡了各类数据在特征空间的占比，提升了网络模型的鲁棒性。

表 2-1 原数据集和扩充数据集详解

编号	名称	原数量	扩充后数量
6	大螟	20	174
7	二化螟	90	175
8	稻纵卷叶螟	18	174
9	白背飞虱	120	175
10	褐飞虱属	150	175
25	地老虎	4	174
41	蝼蛄	14	172
105	粘虫	27	173
110	草地螟	6	170
115	甜菜夜蛾	17	172
148	黄足猎蝽	59	169
156	八点灰灯蛾	247	247
222	棉铃虫	20	168
228	二点委夜蛾	21	170
235	甘蓝夜蛾	5	162
256	蟋蟀	87	167
280	黄毒蛾	40	164
310	稻螟蛉	4	156
387	紫条尺蛾	4	154
392	水螟蛾	3	147

394	线委夜蛾	5	148
398	甜菜白带野螟	15	136
401	歧角螟	7	132
402	瓜绢野螟	4	121
430	豆野螟	1	103
480	石蛾	18	121
485	大黑鳃金龟	12	110
673	干纹冬夜蛾	1	92
	总数	1019	4401

2.2.3 目标与背景融合

数据增强的大部分工作都是为了解决目标检测任务中的分类部分，解决定位任务的关键在于训练网络模型区分前景与背景的能力。如果单一地采用 copy and paste 方法的确可以有效地提升检测分类的准确率，但是有可能因为前景和背景的差异过大，导致模型学习到的定位能力偏弱。我们在采用对部分类目标 copy and paste 的过程中，扩大目标图像背景在标注中的占比，降低了原有背景的比例。这一方法迫使网络去学习目标背景的特征，从而更好地区分前景与背景。

针对数据集中存在的多尺度问题，通常采用特征金字塔的方法，将不同尺度的特征图进行融合，尽可能保留高分辨率特征图的特征；还可以通过设定不同尺度的锚框以适应不同尺寸的目标。

第三章 基于两阶段目标检测算法多模型融合的农作物害虫研究

3.1 算法确定

在完成上述任务分析和数据的预处理之后，我们首先可以确定任务的解决方案是基于深度学习的目标检测方法进行，不同于传统图像处理和传统机器学习算法与简单神经网络的实现，近些年来基于深度学习的目标检测算法针对各种复杂应用场景表现性能更好。

对于目标检测，基于深度学习的主流模型大致分为两类：（1）两阶段目标检测算法：Fast RCNN、Faster RCNN、Mask RCNN、Cascade Mask RCNN 等，这些方法首先产生候选区域（region proposals），然后对候选区域进行分类；（2）一阶段检测算法：不需要 region proposals 阶段，直接产生目标的类别概率和位置坐标值，如 Yolo 系列和 SSD。

然而，两阶段目标检测算法在准确度上有优势，而一阶段算法在速度上有优势，结合之前的问题分析和背景研究，农作物病虫害分析更适合选择两阶段目标检测算法，下表展示了一些两阶段算法在 COCO 数据集上的测试性能表现：

表 3-1 二阶段目标检测算法在 coco 数据集上的性能表现

模型	box mAP
Cascade R-CNN（Resnet-50）	41.0
Cascade Mask R-CNN（Resnet-50）	41.9
Cascade R-CNN（Swin-S）	48.5
Cascade Mask R-CNN（Swin-S）	51.9

结合这些主流目标检测算法，为了更好的解决小样本目标检测，我们首先采用多感知器和通过递增的 IOU 阈值分阶段训练方法的 Cascade Mask R-CNN，以及包含滑窗操作和具有层级设计的 Swin-S Cascade Mask R-CNN 作为基本目标检测模型，然后对每个模型多阶段权值进行 SWA，为了获得更好的性能，并进一步使用多模型融合操作。

3.2 算法基础

3.2.1 Faster RCNN

在介绍 Cascade Mask RCNN 之前，我们先简单回顾一下传统两阶段目标检测算法中的 Faster RCNN[7]，本次工作中使用了以 Faster RCNN 为基础的网络模型，因此有必要对 Faster R-CNN 的整体架构和性能进行详细的分析与讨论。

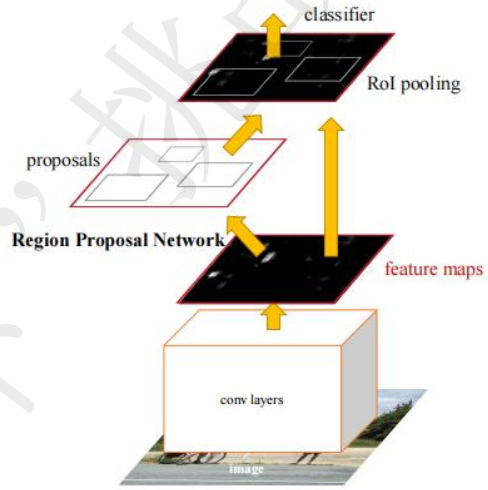


图 3-1 Faster R-CNN 的网络框架

Faster R-CNN 由 Ross B. Girshick 与 2016 年提出。在结构上，Faster RCNN 将特征抽取，proposal 提取，边框回归、分类等步骤都整合在了一个网络中，使得综合性能有

较大提高，在检测速度方面尤为明显。模型的综合性能有较大提高，尤其是在检测速度方面。

Faster R-CNN 可以分为四个部分（如下所示）：

1) 主干特征提取网络。作为一种基于 CNN 网络目标检测方法，Faster RCNN 首先使用一组基础的卷积层提取图像的特征图。该特征图随后被共享用于后续 RPN 层和全连接层。

2) 候选区域网络。RPN 网络用于生成候选区域。该层通过激活函数判断锚框属于正样本或者负样本，再利用边框回归修正锚框获得精确的候选框。

3) Roi Pooling。该层收集输入的特征图和候选框，综合这些信息后提取候选框区域特征图，送入后续全连接层判定目标类别。

4) 分类。利用候选区域特征图计算候选框内目标的类别，同时再次通过边框回归获得检测框最终的精确位置。

3.2.2 FPN

多尺度检测在目标检测中变得越来越重要，对小目标的检测尤其如此。现在主流的目标检测方法很多都用到了多尺度的方法。特征金字塔 Feature Pyramid Network (FPN)[11]是精心设计的多尺度检测方法，如图 3-2 所示。为了更好地识别到目标较小的害虫，我们采用了 FPN 进行多尺度的训练和测试。下面对 FPN 进行介绍。

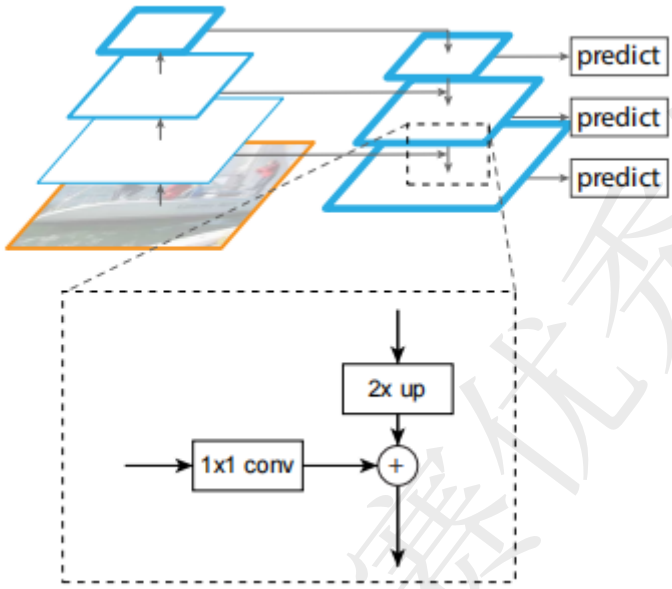


图 3-2 FPN 结构

FPN 结构中包括自下而上，自上而下和横向连接三个部分，如图 3-2 所示。这种结构可以将各个层级的特征进行融合，使其同时具有强语义信息和强空间信息。这样做的好处是：对每一种尺度的图像进行特征提取，能够产生多尺度的特征表示，并且所有等级的特征图都具有较强的语义信息，甚至包括一些高分辨率的特征图。

自下而上（Bottom-up pathway）的过程是下采样（down-sampling）的过程。其作为前馈 backone 的一部分，每一步都进行 step = 2 的下采样。越高层的特征图中的语义信息和空间信息更强，每个检测锚点的感受野也就更大。

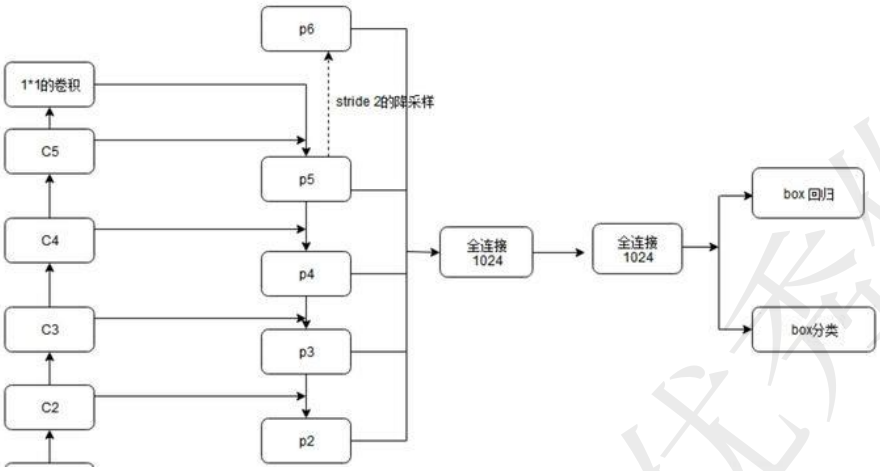


图 3-3 Faster RCNN 中的 FPN 结构

图 3-3 是 Faster RCNN 中的 FPN 结构，左列 Resnet 用每级最后一个 Residual Block 的输出。FPN 用 2~5 级参与预测(第 1 级的高级语义信息不足)，{C2,C3,C4,C5}表示 conv2，conv3，conv4 和 conv5 的输出层(最后一个残差 block 层)作为 FPN 的特征，分别对应于输入图片的下采样倍数为{4，8，16，32}。

自顶向下的过程 Top-down pathway 通过上采样（up-sampling），放大到上一个 stage 的特征图一样的大小。和自底向上的过程形成对应。上采样中使用了最近邻插值法。使用最近邻值插值法，可以在上采样的过程中最大程度地保留特征图的语义信息(有利于分类)，从而与 bottom-up 过程中相应的具有丰富的空间信息(高分辨率，有利于定位)的特征图进行融合，从而得到既有良好的空间信息又有较强烈的语义信息的特征图。

3.2.3 Mask RCNN

首先我们来看基于 Faster RCNN FPN 结构存在的一些问题。在 Faster RCNN 中，存在两次整数化的过程：

1) region proposal 提议的检测框 (x,y,w,h) 信息通常是小数，但是为了方便操作会把它整数化。
2) 将整数化后的边界区域平均分割成 k × k 个单元，对每一个单元的边界进行整数化。

事实上，经过上述两次整数化，此时的候选框已经和最开始回归出来的位置有一定的偏差，这个偏差会影响检测或者分割的准确度。在论文里，作者把它总结为“不匹配问题”（misalignment）。Mask R-CNN[8]提出了 RoIAlign 的方法来取代 ROI pooling，RoIAlign 可以保留大致的空间位置。

为了解决这个问题，ROI Align 方法取消整数化操作，保留了小数，使用以上介绍的双线性插值的方法获得坐标为浮点数的像素点上的图像数值。但在实际操作中，ROI Align 并不是简单地补充出候选区域边界上的坐标点，然后进行池化，而是重新进行设计。

下面通过一个例子来讲解 ROI Align 操作。如图 3-4 所示，虚线部分表示 feature map，实线表示 ROI，这里将 ROI 切分成 2x2 的单元格。如果采样点数是 4，那我们首先将每个单元格子均分成四个小方格（如红色线所示），每个小方格中心就是采样点。这些采样点的坐标通常是浮点数，所以需要对采样点像素进行双线性插值（如四个箭头所示），就可以得到该像素点的值了。然后对每个单元格内的四个采样点进行 maxpooling，就可以得到最终的 ROI Align 的结果。

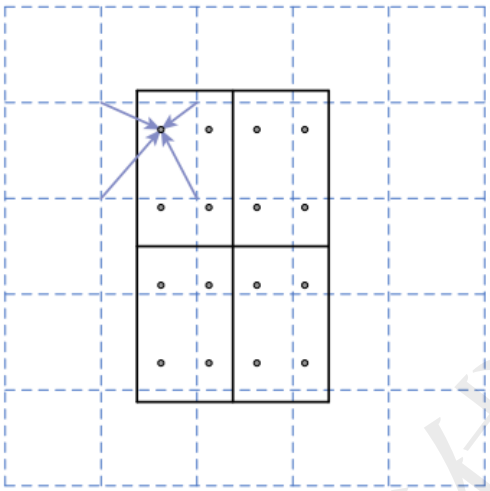


图 3-4 ROI Align

使用了 ROI Align 操作解决了 misalignment 问题后，基于多尺度 FPN 的检测模型性能可以得到很大提升。

3.2.4 Cascade Mask RCNN

在目标检测中，涉及到 IOU 阈值的选取，用来定义正样本和负样本。IOU 阈值的选取也会带来相应的性能的改变，会有以下几个问题：

1) 如果使用较低的 IOU 阈值，那么会学习到大量的背景框，产生大量的噪声预测；
2) 如果采用较高的阈值，随着 IOU 阈值的增加，正样本的数量会呈指数级的减小，因此产生过拟合。并且推理过程中出现于 IOU 的误匹配，也就是在训练优化感知器的过程中的最优 IOU 与输入 proposal 的 IOU 不相同，出现误匹配，这样很大程度上降低了检测精度，检测器的表现往往会变得很差；

3) 错误匹配问题：检测器通常在候选框自身的 IOU 值与检测器训练的 IOU 阈值较为接近的时候才会有更好的结果，如果二者差异较大那么很难产生良好的检测效果；

基于如上问题，Cascade Mask RCNN 提出使用多阶段级联的感知器，并且这些感知器通过递增的 IOU 阈值分级段训练。一个感知器输出一个良好的数据分布来作为输入，训练下一个高质量感知器，这样对于假阳性的问题会好很多，在推理阶段使用同样的网络结构合理的提高了 IOU 的阈值，而不会出现之前所说的错误匹配问题。

Cascade Mask RCNN 基于级联的方法如下图 3-5 所示：

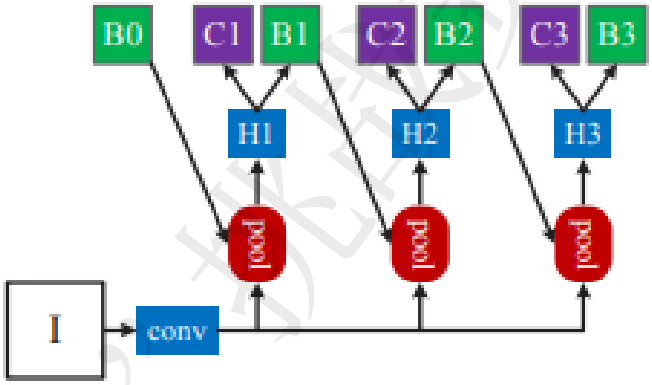


图 3-5 Cascade Mask RCNN 框架

图 3-5 中的 conv 代表主干特征提取网络的卷积层，pool 代表特征提取，H 代表网络的 head 部分，B 代表检测框，C 代表分类结果，B0 代表 proposals。从图中我们能进一步了解到 Cascade Mask RCNN 的机制，针对 RPN 提出的 proposals 大部分质量不高，导致没办法直接使用高阈值的检测器，Cascade Mask RCNN 使用级联回归作为一种重采样的机制，逐阶段提高 proposal 的 IoU 值，从而使得前一个阶段重新采样过的 proposals 能够适应下一个有更高阈值的阶段。

基于级联操作的 Cascade Mask RCNN 可以解决上述 IoU 阈值选取所面临的问题：

1) 每一个阶段的检测器都不会过拟合，都有足够满足阈值条件的样本。

2) 更深层的检测器也就可以优化更大阈值的候选框。

3) 每个阶段的网络头不相同，意味着可以适应多级的分布。

在推理阶段，虽然最开始 RPN 提出的候选框质量依然不高，但在每经过一个阶段后质量都会提高，从而和有更高IoU阈值的检测器之间不会有很严重的错误匹配问题。

3.2.5 Swin Transformer

我们将特征提取网络选择为 Swin-Transformer[6]。作为 2021 ICCV 最佳论文，Swin-Transfoermer 在各大 CV 任务上都取得了很好的性能，其性能优于 DeiT、ViT 和 EfficientNet 等主干网络，已经替代经典的 CNN 架构，成为了计算机视觉领域通用的主干特征提取网络。

Transformer 首先在 NLP 领域取得了巨大的成功，但将其运用在图像领域面领着以下挑战：（1）视觉实体的方差较大，例如同一个物体，拍摄角度不同，转化为二进制后的图片就会具有很大的差异。同时在不同场景下视觉 Transformer 性能未必很好；（2）图像分辨率高，像素点多，Transformer 基于全局自注意力的计算导致计算量较大，如果采用 ViT 模型，自注意力的计算量会与像素的平方成正比。

为了解决以上两个问题，Swin-Transformer 基于了 ViT 模型的思想，创新性的引入了滑动窗口机制，让模型能够学习到跨窗口的信息，同时通过下采样层，使得模型能够处理超分辨率的图片，节省计算量以及能够关注全局和局部的信息，如图 3-6 所示。其中滑窗操作包括不重叠的 local window，和重叠的 cross-window。将注意力计算限制在一个窗口（window size 固定）中，一方面能引入 CNN 卷积操作的局部性，另一方面

能大幅度节省计算量，它只和窗口数量成线性关系。通过下采样的层级设计，能够逐渐增大感受野，从而使得注意力机制也能够注意到全局的特征。

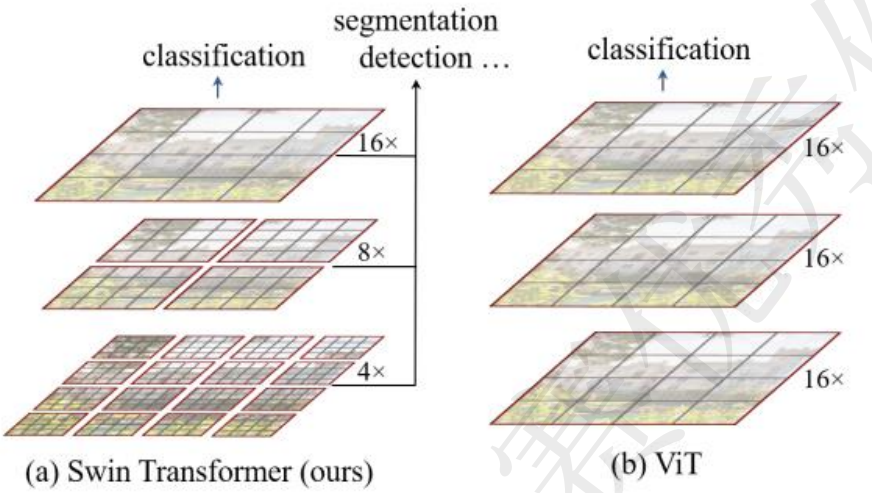


图 3-6 Swin Transformer 与 ViT 对比

Swin-Transformer 的模型结构，如图 所示。

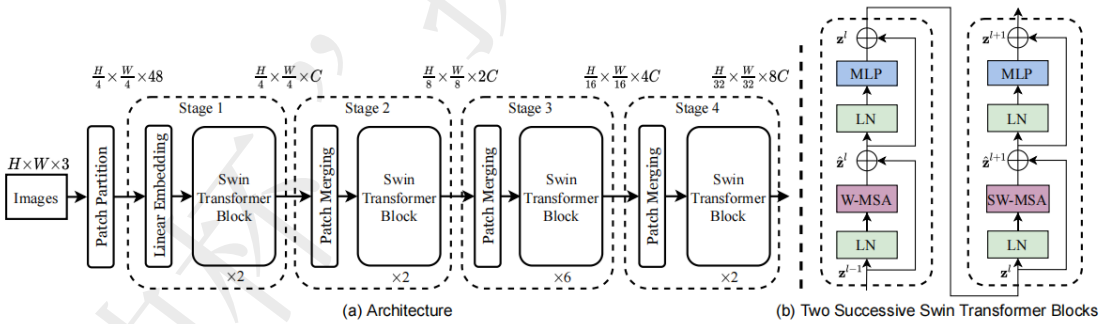


图 3-7 Swin Transformer 模型结构

整个模型采取层次化的设计，一共包含 4 个 Stage，除第一个 stage 外，每个 stage 都会先通过 Patch Merging 层缩小输入特征图的分辨率，进行下采样操作，像 CNN 一样逐层扩大感受野，以便获取到全局的信息。

1) 网络输入大小为 H×W×3 图像，首先进入 Patch Partition 层，通过一个卷积核大小为 4 的卷积层，将输出大小变为 H/4×W/4×48，并输入到 Linear Embedding 层中。其中 48 为通道数；

2) 在 Linear Embedding 层中，将通道数调整为 C；

3) 接下来每一个 Stage 都由 Patch Merging 和多个 Swin Transformer Block 组成。Patch Merging 模块在每一个 Stage 开始阶段进行下采样，降低图像分辨率。

4) Swin Transformer Block 模块如图 3-7 右边所示，由 LayerNorm，Window Attention，Shifted Window Attention 和 MLP 模块组成。下面我们将详细介绍 Window Attention 以及 Shifted Window Attention

传统的 Transformer 都是基于全局来计算注意力的，因此计算复杂度十分高。而 Swin Transformer 则将局部的注意力计算限制在每个窗口内，进而减少了计算量。

首先来对比这两种方法的复杂度。假设每个窗口包含 M×M 个 patch，一张图像可以划分为 h×w 个 patch。则基于全局的自注意力计算复杂度和而基于窗口的自注意力计算复杂度分别为

$$\Omega(\mathrm{MSA}) = 4hwC^2 + 2(hw)^2C \tag{3.1}$$

$$\Omega(\mathrm{W-MSA}) = 4hwC^2 + 2M^2hwC \tag{3.2}$$

其中 C 为该层的通道数， M 为 patch 的大小，一般为定值 7。前者是关于 hw 的二次函数，而后者是 hw 的线性函数。这也导致了当图像的分辨率很大时，基于全局的自注意力不再适用。

Window Attention 机制，只使用基于窗口的自注意力缺乏跨窗口的联系，这限制了模型的空间建模能力，所以 Swin-Transformer Block 在 Window Attention 的基础上又加入了 Shifted Window Attention 环节，如图 3-8 所示：

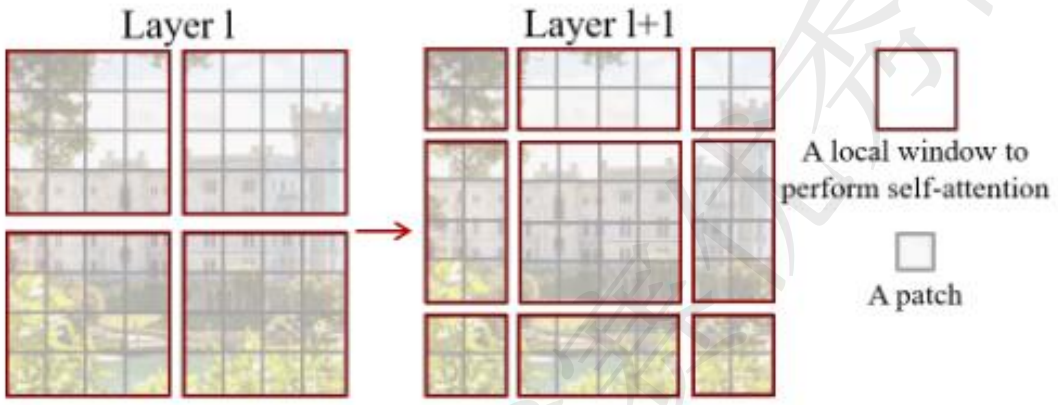


图 3-8 Shifted Window Attention 方法示意图

在 l 层中（图左）采用规则的窗口划分方案，并在每个窗口内计算自注意力。在 l+1 层中，窗口移动，由原来的 4 个窗口变为 9 个窗口。在新窗口中自注意力跨越了 l 中窗口的边界，实现了跨窗口的联系。使用移位窗口分区方法，连续的 Swin Transformer 块计算为:

$$\begin{aligned}\hat{\mathrm{z}}^{l} &= \mathrm{W}-\mathrm{MSA}\left(\mathrm{LN}(\mathrm{z}^{l-1})\right)+\mathrm{z}^{l-1},\\ \mathrm{z}^{l} &= \mathrm{MLP}(\mathrm{LN}(\hat{\mathrm{z}}^{l}))+\hat{\mathrm{z}}^{l},\\ \hat{\mathrm{z}}^{l+1} &= \mathrm{SW\text{-}MSA}\left(\mathrm{LN}(\mathrm{z}^{l})\right)+\mathrm{z}^{l},\\ \mathrm{z}^{l+1} &= \mathrm{MLP}(\mathrm{LN}(\hat{\mathrm{z}}^{l+1}))+\hat{\mathrm{z}}^{l+1},\end{aligned} \tag{3.3}$$

$\hat{\mathrm{z}}^{l}$ 表示(S)W-MSA 模块的输出，z^{l} 表示 MLP 模块的输出。W-MSA 和 SW-MSA 分别表示使用规则和移位窗口分区配置的基于窗口的多头自注意。移窗分割方法引入

了前一层相邻非重叠窗口之间的联系，在图像分类、目标检测、语义分割等方面都是有效的。

根据 Linear Embedding 层输出的特征维度和第三个阶段中 Swin-Transformer 模块重复的次数，模型的大小不同。可分为 Swin-T、Swin-S、Swin-B、Swin-L，具体如表 1-2 所示：

表 1-2 Swin-Transformer 的 不同结构

模型	特征维度 C	layer numbers
Swin-T	96	{2, 2, 6, 2}
Swin-S	96	{2, 2, 18, 2}
Swin-B	128	{2, 2, 18, 2}
Swin-L	192	{2, 2, 18, 2}

其中 C 表示 Linear Embedding 层输出的特征维度。由于硬件条件的限制，本次我们选择的模型是 Swin-S。

3.3 模型训练

如第二章数据处理操作，由于样本数目的不均衡，个别样本数目仅有一类，为了更好的训练并测试性能，首先基于样本数目特别少的类进行裁剪，并且采用 copy and paste 的方法扩充数据集，除此之外，类别之间的不平衡导致的长尾效应也会影响训练，所以进一步我们使用样本平衡操作。将不同模型在训练集上试验，并对比验证集上的性能，以此进行消融试验。

模型的试验是基于 mmdetection 框架下实现，mmdetection 能够很好的继承主流目标检测框架，通过数据集的配置和模型配置文件的更改，快速的将目标检测算法适应到各种应用场景和数据集之上，进行参数的调整和实现。

3.3.1 Cascade Mask RCNN 模型训练

基于 mmdetection 可以快速的实现 Cascade Mask RCNN 的复现，但是不同应用场景下的参数信息都需要微调，为了更好的性能，我们进行了调整 ancho 比例，设置损失函数种类，不同级联结构的 IoU 阈值，多尺度，等一系列操作。

以下是基于 Cascade Mask RCNN 所进行的修改：

1) 由于图片中的标注框拥有不同比例，各个图片中存在不同尺度和长宽比（scales and ratios）的物体，在训练过程中学习适应不同物体的形状比较困难，通过设置不同的尺度的先验框，就有更高的概率出现对于目标物体有良好匹配度的先验框。结合 YOLOv2 使用 k-means 聚类获取宽高比，我们根据标注文件设置锚框数量和不同的锚框的比例。通过加载标注信息，返回所有目标的长宽信息，对训练样本的标注框进行 k-means 聚类操作，输出计算得到的锚框，期望的长宽比例数。

2) 为了更好的提高目标检测的准确率，采用 Smooth L1 损失函数。较之于 L1 损失函数，Smooth L1 损失函数对离群点、异常值不敏感，梯度变化相对更小，训练时不容易跑飞。Smooth L1 如下公式所示

$$smooth_{L1}(x) = \begin{cases} 0.5x^2 & if\ |x| < 1 \\ |x| - 0.5 & otherwise \end{cases} \quad (3.4)$$

smooth L1 在 x 较小时，对 x 的梯度也会变小，而在 x 很大时，对 x 的梯度的绝对值达到上限 1，也不会太大以至于破坏网络参数。

3) 因为害虫检测多涉及小目标，针对三个级联操作的初试阈值选取，分别设置层 0.3，0.4，0.5，以便能够更好的检测到小目标，防止过多的漏检情况的产生。

3.3.2 Swin-S Cascade Mask RCNN 模型训练

同样，Swin-S Cascade Mask RCNN 模型的训练也是基于 mmdetection 实现，除了基本框架外，和 Cascade Mask RCNN 实现的 K-Means 聚类设置宽高比，Smooth L1 损失函数，以及为级联部分选取更小的阈值等相似操作。

除此之外，在 Swin-S Cascade Mask RCNN 的训练过程中，还结合了 Soft NMS（软化非极大抑制），NMS（Non-Maximum Suppression，非极大抑制）是检测模型的标准后处理操作，用于去除重合度（IoU）较高的预测框，只保留预测分数最高的预测框作为检测输出。在传统的 NMS 中，最高预测分数预测框重合度超出一定阈值的预测框会被直接舍弃，这样不利于相邻物体的检测。Soft NMS 的改进方法是根据 IoU 将预测框的预测分数进行惩罚，最后再按分数过滤。配合 Deformable Convnets，Soft NMS 在 MS COCO 上取得了当时最佳的表现。

3.4 模型融合

3.4.1 SWA（随机权重平均）

SWA（随机权重平均）与 SGD（随机梯度下降）有异曲同工之妙，深度神经网络的典型训练过程是使用 SGD 进行优化一个损失函数，同时使用一个衰减的学习率（Step，Cosin 等），直到收敛为止。但是每次学习率循环结束之后，都会产生局部极小值，并且趋向于在损失面的边缘区域累计，在边缘区域上的损失值可能都很小，无法陷入更优的局部极值。如图 3-9 所示：

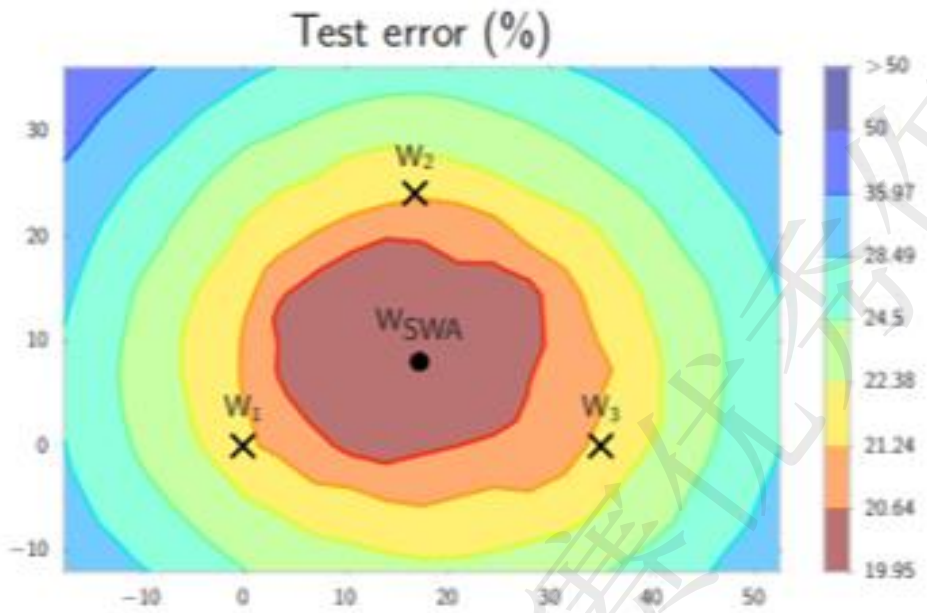


图 3-9 SWA 和 SGD 对比

SWA 取 SGD 轨迹的多个点（W_1、W_1、W_1）的简单平均值，按照一个周期或者不变的学习率，将对比于 SGD 会有更好的泛化效果，可以找到更广的最优值域。

SWA 加入了周期性滑动平均操作，来限制权重的变化，解决 了传统 SGD 在反向传播过程中的权重振荡问题。 SGD 是依靠当前 batch 的数据来更新参数，每一个 epoch 都会调整一次参数，随机挑选的梯度方向极有可能不是最佳梯度方向，甚至与最佳梯度方向有一个很大的夹角，这样大刀阔斧调整的参数，极其容易振荡。而 SWA 限制更新频率，对周期内参数取滑动均值，这样就解决 了 SGD 的问题。

SWA 结合到 Cascade Mask RCNN 和 Swin-S Cascade Mask RCNN 之中，将最后学习率不变时的相应多个 epoch 模型的权重进行平均，用平均后的模型进行检测，SWA 操作实现过程如下图：

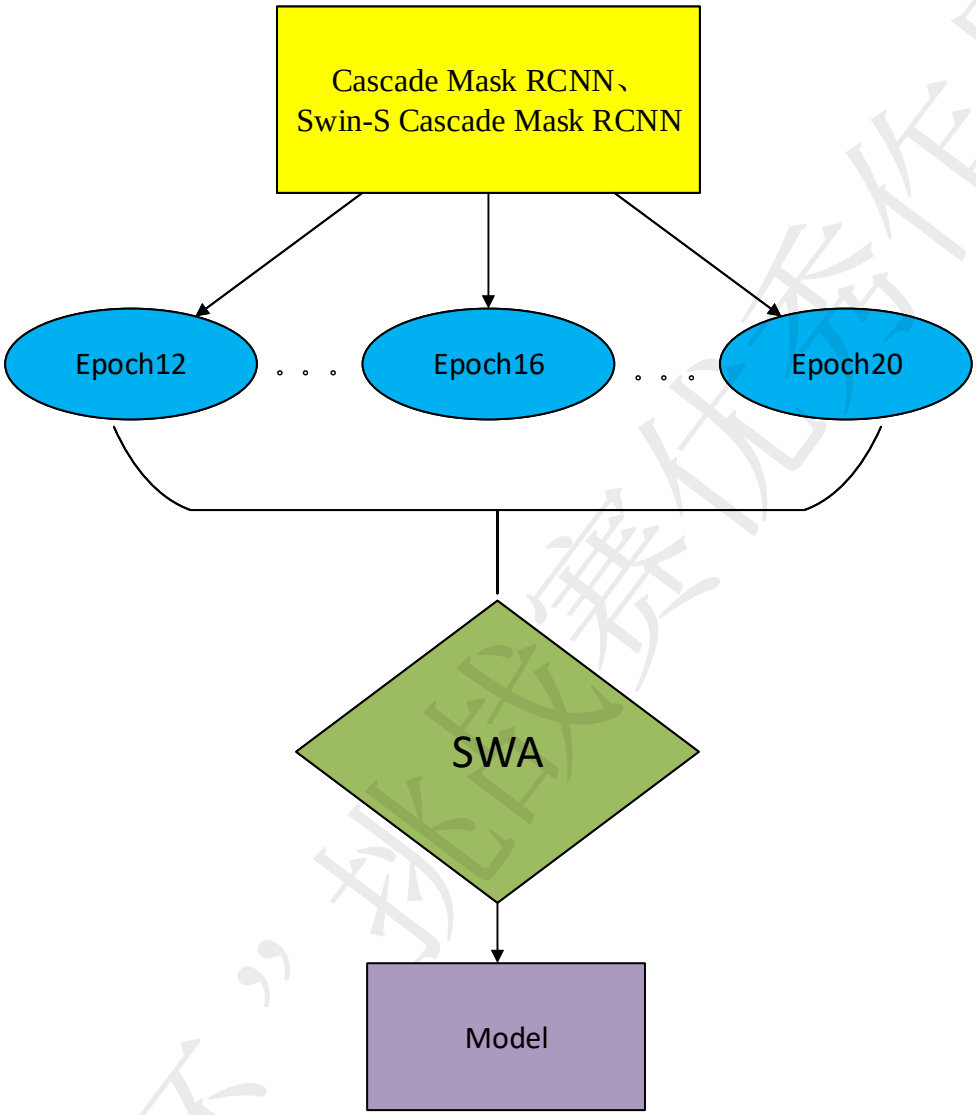


图 3-10 SWA 实现流程

3.4.2 多模型融合

有 2.3 的模型训练过程之后，得到了 Cascade Mask RCNN 和 Swin-S Cascade Mask RCNN 在害虫图像训练集上的训练结果，不同的模型基于不同的算法以及不同的初始化和更新策略等，所以不同模型必然会产生不同的学习能力，针对不同的害虫图像表现不一样的准确度，对两个模型进行融合是必要的。

使用两个目标检测算法经过 SWA 后的权重进行测试，对获得的测试结果进一步使用 NMS，获取两个模型检测结果的冰机，防止漏检，如此操作对一些样本量少的目标，尤其小目标的识别，产生进一步的性能提升。

模型融合的示意图 3-11 如下所示：

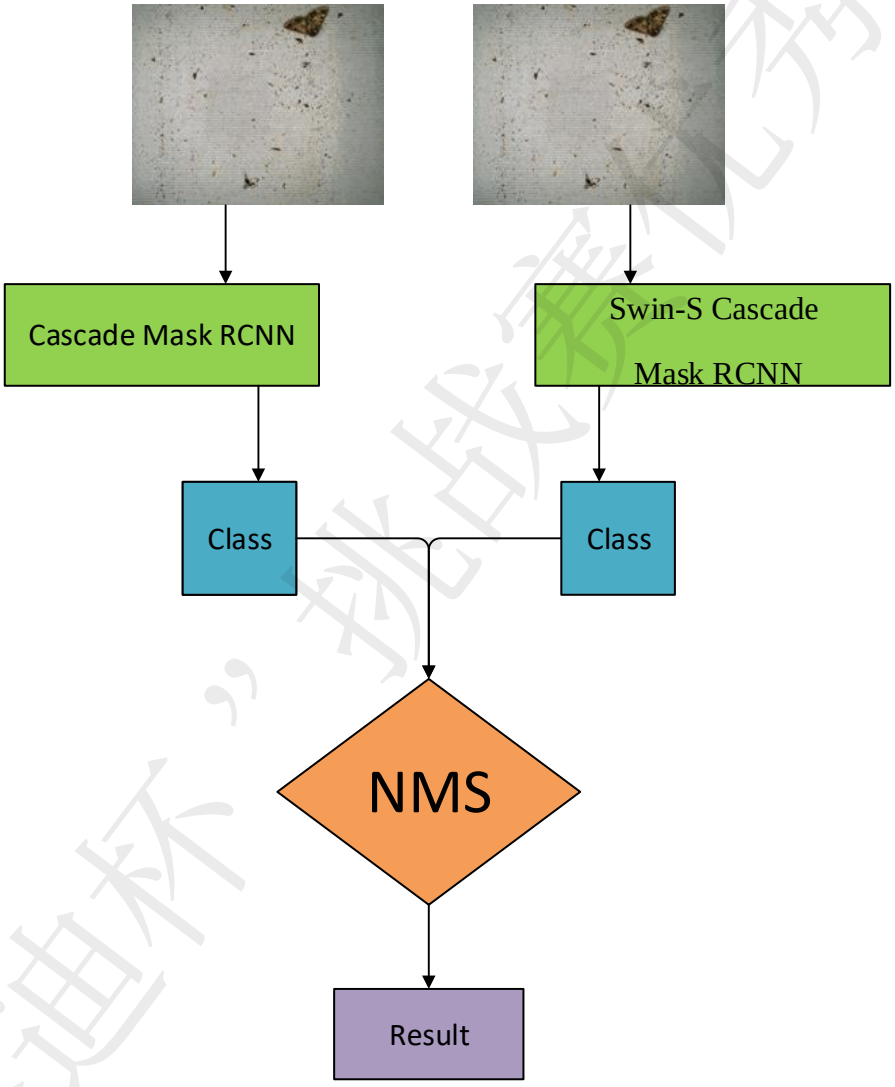


图 3-11 多模型融合流程

第四章 实验结果分析

4.1 实验配置与数据处理

4.1.1 实验配置

本篇论文的实验都是基于 Ubuntu 系统下进行，使用 GPU 和 CPU 作为基础硬件，具体配置运行环境如表 2-1 所示：

表 2-1 实验配置运行环境

名称	环境配置
GPU	NVIDIA 2080Ti*2
CPU	Intel(R) Xeon(R) Silver 4210 CPU @ 2.20GHz
操作系统	Ubuntu 18.04
编程语言	Python 3.7
CUDA	Cuda 11.3
深度学习框架	Pytorch 1.11

4.1.2 评价指标

目标检测任务可以划分为分类和定位两大任务，其中定位任务采用召回率 Recall 进行评估；分类任务采用准确率 Precision 进行评估：

$$Recall = \frac{TP}{TP+FN} \tag{4.1}$$

$$Precision = \frac{TP}{TP+FP} \tag{4.2}$$

其中，TP 表示检测器输出结果中正确的个数，FP 表示检测器输出的结果中错误的个数。同时，为了更好地平衡分类和定位任务的重要性，我们引入了以准确率和召回

率为变量的 PR 曲线。最终，我们的评估指标转换为单个类别表示为 AP，整体准确率表示为 mAP：

$$AP_j = \int_0^1 p(r)dr \tag{4.3}$$

$$mAP = \frac{1}{m}\sum_m AP_j \tag{4.4}$$

其中，mAP 的计算是采用设定阈值 0.5:0.95 的 mAP 值取平均值得到的。

在通过第二章的离线数据增强方法以后，我们将原图像中的 1019 个目标拓展为 4401 个目标。由于原图像的样本数目较少，为了提升模型的预测能力，同时避免过拟合的问题出现，我们将原有的验证集和测试集合并，以提升模型的鲁棒性。最终，我们以 7:1 的比例划分了训练集和测试集，分别含有 3751 和 650 个目标。

4.1.3 数据集划分

在通过第二章的离线数据增强方法以后，我们将原图像中的 1019 个目标拓展为 4401 个目标。由于原图像的样本数目较少，为了提升模型的预测能力，同时避免过拟合的问题出现，我们将原有的验证集和测试集合并，以提升模型的鲁棒性。最终，我们以 7:1 的比例划分了训练集和测试集，分别含有 3751 和 650 个目标。

4.2 实验结果

我们分别对 Cascade Mask RCNN 和 swin transformer 这两类模型，进行了训练和测试，并且通过 mmdetection 的检测程序得到了丰富和详细的算法指标。下面我们将依次具体展示模型的各项指标。

4.2.1 Cascade Mask RCNN

在 CascadeRCNN 模型中，我们使用了 Resnet-50 作为模型的主干特征提取网络，采用了 0.25, 0.5, 1.0, 2.0,和 4.0 总计 5 种不同尺度的锚框，并采用了两种不同尺度的图像输入方式。我们使用了 AdamW 作为神经网络的 optimizer，初始学习率 0.001，并采用线性衰减的方式。我们使用交叉熵损失函数和 SmoothL1 损失函数分别作为分类和定位任务的损失函数。对于模型输出结果的后处理方法，我们则采用 Soft NMS。模型总计训练 100 个周期，随机数种子为 42，训练每个周期采用 3 倍的训练集输入方法，以提升训练的稳定性，相当于训练 300 个周期。

下面展示的是模型的 mAP 和召回率的变化值。如图 所示，模型的 mAP 和召回率在 15 个周期之后达到稳定，20-60 个周期属于模型微调阶段，第 60 个周期以后模型各项指标趋于稳定。

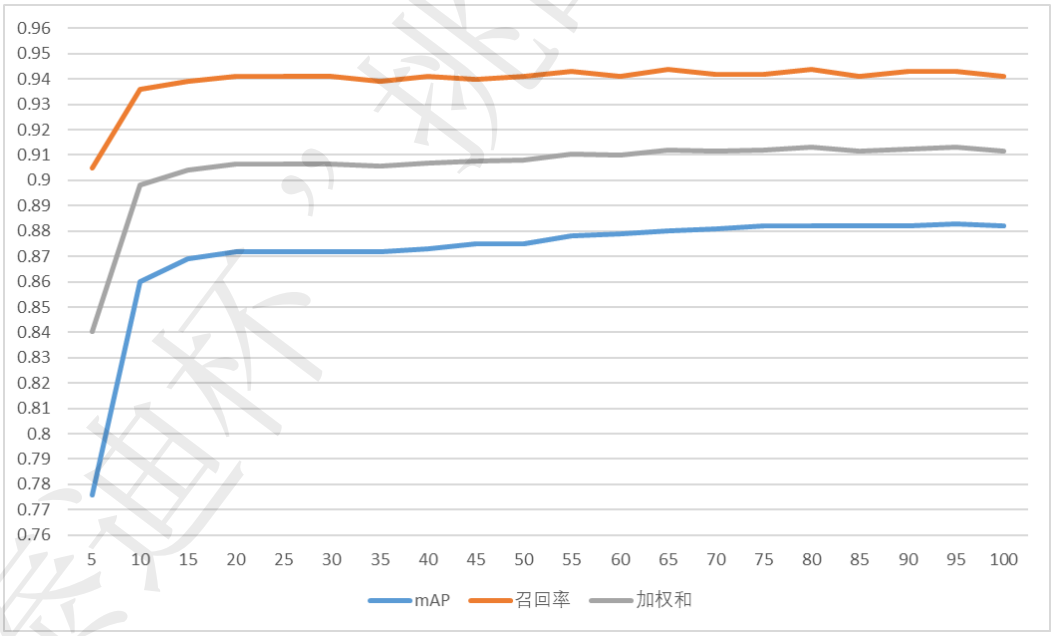


图 4-1 Cascade Mask RCNN 训练结果

最后，我们选取了第 65、80 和 95 个周期的权重值，作为随机权重 SWA 的输入。在表-中，我们将展示了效果最佳的第 80 个周期的各类别害虫准确率的情况。

表 4-2 各害虫类别识别性能

编号	AP	编号	AP	编号	AP
6	**0.995**	7	**0.491**	8	**0.906**
9	**0.424**	10	**0.287**	25	**0.989**
41	**0.995**	105	**0.916**	110	**0.927**
115	**0.929**	148	**0.884**	156	**0.754**
222	**0.979**	228	**0.936**	235	**0.990**
256	**0.734**	280	**0.954**	310	**0.984**
387	**0.959**	392	**0.983**	394	**0.984**
398	**0.941**	401	**0.970**	402	**0.965**
430	**0.993**	480	**0.880**	485	**0.881**
673	**0.865**				

4.2.2 Swin-S Cascade Mask RCNN

在 Swin-S Cascade Mask RCNN 模型中，为了兼顾检测的准确率和速率，我们使用了基于 Swin-Transformer 的预训练主干的 small 版本。在这一版本中，我们同样采用了多尺度锚框的设计方式。在损失函数设计，optimizer 和模型后处理方法上，我们均体现了一致性。模型总计训练 30 个周期，随机数种子为 42，训练每个周期采用 3 倍的训练集输入方法，以提升训练的稳定性，相当于训练 90 个周期。如图 所示，Swin-S Cascade Mask RCNN 模型收敛速度明显快于 Cascade Mask RCNN，模型的 mAP 和召回率在 10 个周期之后达到稳定，10-15 个周期属于模型微调阶段，第 15 个周期以后模型各项指标趋于稳定。

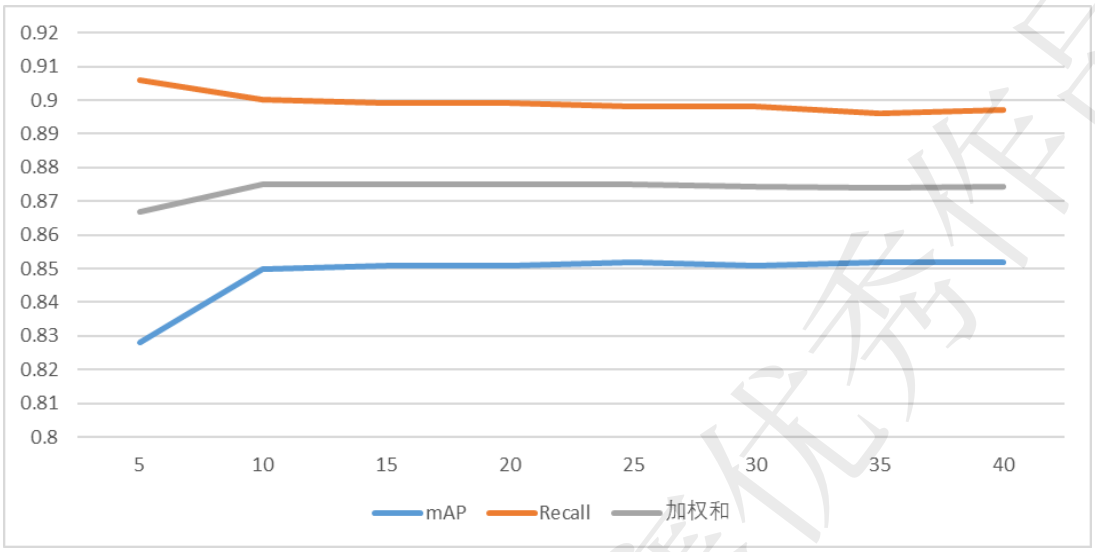


图 4-2 Swin-S CascadeRCNN 训练过程

结合 SWA 之后 Swin-S Cascade Mask RCNN 模型的识别结果如下表所示：

表 4-3 各类害虫识别结果

编号	AP	编号	AP	编号	AP
6	**0.978**	7	**0.463**	8	**0.877**
9	**0.260**	10	**0.220**	25	**0.960**
41	**0.986**	105	**0.900**	110	**0.887**
115	**0.911**	148	**0.843**	156	**0.785**
222	**0.963**	228	**0.870**	235	**0.944**
256	**0.687**	280	**0.977**	310	**0.955**
387	**0.960**	392	**0.978**	394	**0.950**
398	**0.946**	401	**0.992**	402	**1.000**
430	**0.936**	480	**0.765**	485	**0.864**
673	**0.996**				

4.2.3 消融试验

从上述实验结果能够看到，我们的方法能够以很高的精度进行农田害虫的定位和识别。为了能够更好的对比论文中使用到的各种创新点的性能表现，我们进一步进行了消融试验，分别对比了在相同的 backbone——Resnet50 以及相同的多尺度 FPN 策略下，Faster RCNN、Casecade RCNN 和 Swin-S Cascade Mask RCNN，以及三种模型结合 SWA 和 NMS 的性能表现，消融试验结果如下表所示：

表 4-4 消融试验结果

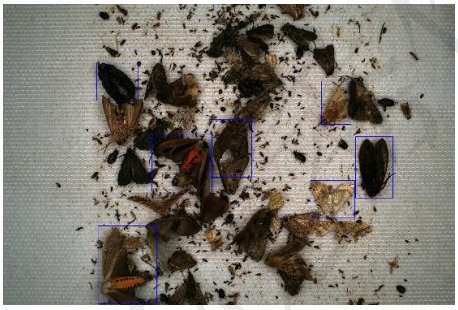
Faster RCNN(Resnet50)	Mask RCNN(Resnet50)	Casecade Mask RCNN(Resnet50)	Casecade RCNN(Swin-S)	SWA	NMS	**mAP**
√						**0.753**
	√					**0.788**
		√				**0.805**
			√			**0.821**
		√		√		**0.823**
			√	√		**0.844**
		√	√	√		**0.856**
		√	√	√	√	**0.876**

从表中我们可以看到，在基本的模型设置下，Faster RCNN、Casecade RCNN 和 Swin-S Cascade Mask RCNN 中，Swin-S Cascade Mask RCNN 的 mAP 最高，能达到 0.831，并且进一步对比 Casecade RCNN 和 Swin-S Cascade Mask RCNN 结合 SWA 和 NMS 操作之后，性能都进一步的提高，最终能够在验证集上的 mAP 能达到 0.876 的性能。从消融试验中，我们可以看到，采用的基本的目标检测框架和所进行的 SWA 和 NMS 多模型融合，都能对结果的提升产生正面的影响。

4.2.4 识别效果展示

为了直观的看到在测试集上的识别效果，我们在消融试验之外，在测试图片上直接显示出识别出的目标位置，如图 4-3 所示：

原图：01589.jpg



识别结果

原图：00028.jpg

识别结果

原图：02773.jpg

识别结果

图 4-3 识别效果显示 1

如图 4-3 所示，我们把官方验证集上的部分检测结果也直观的展示出来，具体的分类情况和定位数值在 result2.csv 和 result3.csv 中均有详细的体现。

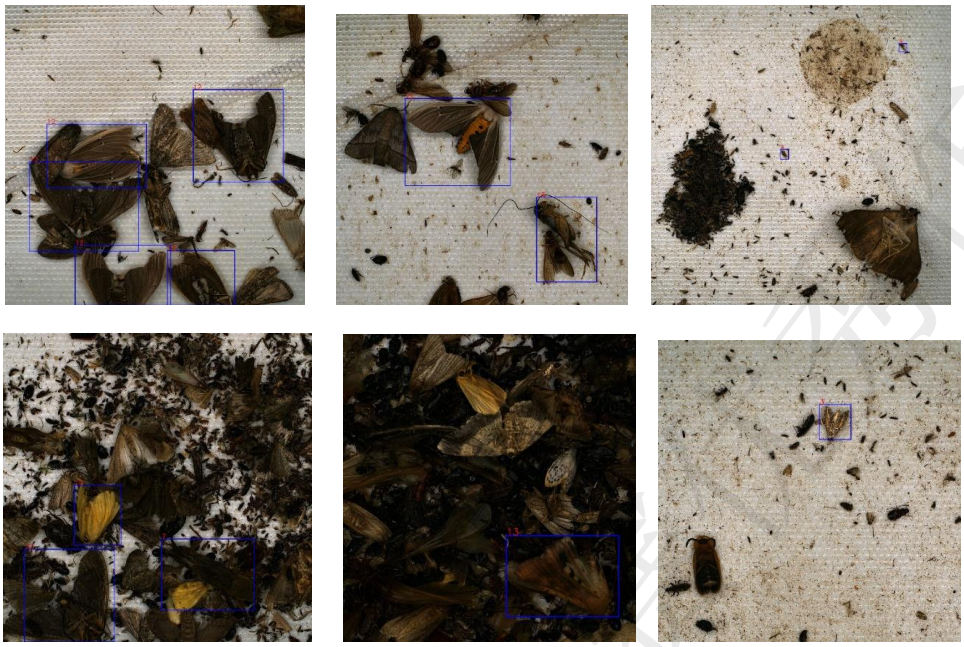


图 4-4 识别结果展示 2

从上面两个识别结果图 4-3、图 4-4 可以看到，无论是简单背景下还是复杂背景下，我们的模型都能够很好的检测到害虫所在位置，并且针对一些小目标害虫，也能够达到不错的识别效果。

第五章 总结与展望

5.1 总结

农作物病虫害识别的主要特点是害虫图片数据集的长尾效应和小目标，以及少样本的处理问题。针对数据集中的问题，我们采用了 copy and paste 的方法，拓展了数据集，并且进一步的使用旋转、噪声、翻转、亮度对比度调节等常规的在线数据增强，以此来平衡数据集不同类别样本数目的分布，以及增加数据集的复杂性，并且采用训练和测试时的多尺度训练，把图片等比例放大或者缩小，来增加对小目标的识别效果。

针对数据处理后的害虫图片，我们使用两阶段目标检测算法：Cascade Mask RCNN 和 Swin-S Cascade Mask RCNN，进行害虫的定位与识别。Cascade R-CNN 采用多感知器和递增的 IOU 阈值方法的分阶段训练，Swin-S Cascade R-CNN 使用包含滑窗操作、具有层级设计。除了基本模型之外，为了更好的适应害虫识别，我们首先使用基于 k-means 聚类的锚框长宽比生成，能够更好的符合数据集中目标检测框的不同宽高比。并且为了让模型更好的识别小目标，我们使用 Soft NMS，并且使用 Smooth L1 损失函数替换掉原来的 L1 损失函数。

最后，为了进一步提升网络的性能，我们分别对两个网络不同 epoch 的权重进行 SWA（随机权重平均），然后使用 SWA 之后的 Resnet50 Cascade Mask RCNN 训练模型和 Swin-S Cascade Mask RCNN 训练模型进行多模型融合，对两个网络分别的检测结果进行 NMS 处理，得到最终的检测结果。

5.2 优缺点分析

优点：

- 通过对农作物害虫图片的数据集处理和进行数据增强，能够有效的应对长尾效应和多尺度目标问题；
- 选择 Resnet50 Cascade Mask RCNN 和 Swin-S Cascade Mask RCNN 作为主要模型，更能贴合农作物害虫识别的目标检测任务；
- 采用 K-means 聚类完成锚框长宽比的确定，使用 Smooth L1 损失函数和 Soft NMS，更能符合害虫识别的应用场景；
- 基于随机权重平均和多模型融合的方法，让网络识别效果更好，缓解误检和漏检的可能性。

缺点：

- 由于采用 Swin-S Cascade Mask RCNN 和 Resnet50 Cascade Mask RCNN 两个模型进行训练，以及采用训练时和测试时的多尺度，导致模型的训练和检测，需要耗费大量的显存资源以及时间成本；
- 基于 copy and paste 的数据扩充方法，还是有一定程度存在与背景不符的可能性，存在少数图片的重叠情况，应该进一步考虑。

5.3 展望

我们提出的基于农作物害虫图片的目标检测识别方法，完全是基于所提供的农作物害虫识别图片进行实现，由于缺乏专家知识和相关专业技能，并不能很好的处理数据集中可能存在的错误标记情况，也不能很好的考虑到一些害虫图像中可能存在的其他特点，能够应用到数据处理和相关算法之中。

为了进一步减少人工标注数据集的繁复和昂贵的代缴，基于半监督学习的方法，将是我们未来一个很好的尝试工作，利用主流的半监督方法，通过有标签的部分数据集，来为无标签数据集产生伪标签，并且加入到训练集中训练网络。这将大大的减少人工标注的费用，省时高效。但是，半监督方法对没有标注的数据集，生成的伪标签的准确率，比较依赖有标注数据集的训练。因此，我们所完成任务为未来半监督学习铺平了道路。目前，我们所提出的方法，仍然有进一步上升的空间，期待未来基于更好的数据处理方法和更好的模型，能够实现更好的害虫识别的效果，为农作物病虫害的处理提供更好的效果。

参 考 文 献

[1] 李玲,李艳乐,郭海丽,苏明敏.基于卷积神经网络的苹果树病虫害识别[J].南方农机,2022,53(08):16-19.

[2] 范世达,马伟荣,姜文博,张辉,王金振,李琦,何鹏博,彭磊,黄兆波.基于深度学习的柑橘黄龙病远程诊断技术初探[J].中国果树,2022(04):76-79+86+133.DOI:10.16626/j.cnki.issn1000-8047.2022.04.015.

[3] 张红涛,胡玉霞,赵明茜,邱道尹,张孝远,张恒源.田间害虫图像识别中的特征提取与分类器设计研究[J].河南农业科学,2008(09):73-75.

[4] 谢成军,李瑞,董伟,宋良图,张洁,陈红波,陈天娇.基于稀疏编码金字塔模型的农田害虫图像识别[J].农业工程学报,2016,32(17):144-151.

[5] Zhaowei Cai, & Nuno Vasconcelos (2017). Cascade R-CNN: Delving into High Quality Object Detection arXiv: Computer Vision and Pattern Recognition.

[6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, & Baining Guo (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows.. arXiv: Computer Vision and Pattern Recognition.

[7] Shaoqing Ren, Kaiming He, Ross Girshick, & Jian Sun (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks IEEE Transactions on Pattern Analysis and Machine Intelligence.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, & Ross Girshick (2017). Mask R-CNN IEEE Transactions on Pattern Analysis and Machine Intelligence.

[9] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, & David Lopez-Paz (2017). mixup: Beyond Empirical Risk Minimization arXiv: Learning.

[10] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay K. Vasudevan, & Quoc V. Le (2019). AutoAugment: Learning Augmentation Strategies From Data computer vision and pattern recognition.

[11] Lin, Tsung-Yi, et al. "Feature pyramid networks for object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.