**华侨大学**

**本科毕业论文**



|  |  |
| --- | --- |
| **题目：** | **基于YOLO v5的** |
|  | **无人机目标检测算法研究** |

**学 院** 计算机科学与技术学院

**专 业** 计算机科学与技术

**届 别** 2018届

**学 号** XXX

**姓 名** CS20220147

**指导老师**  XXX

**华侨大学教务处印制**

2022 **年**5 **月**

摘 要

目标检测是从图片中搜寻特定目标，并提取出位置、类别等信息的技术。传统目标识别技术存在区域选择策略没有针对性、算法时间复杂度高、手工设计的特征鲁棒性较差等缺点，使得检测效果不理想。‎[25]随着深度学习技术的进步，深度学习在目标检测领域上的应用的算法也层出不穷，出现了许多优秀的深度学习目标检测算法，并让目标检测在各种工业领域，如城市交通、视频追踪、人脸检测，得到了丰富应用。

本文选择YOLO v5算法作为基础算法，应用在天津大学机器学习与数据挖掘实验室公布的VisDrone无人机数据集上‎[5]。本文引入了增大特征图分辨率、多尺度融合训练的优化方法，提升了YOLO v5算法在VisDrone数据集上的分类和定位性能。

本文研究内容如下：

第一、对空中无人机数据集VisDrone的分析。通过可视化方法分析数据集的标签数量、标签分布来确定YOLO v5算法的性能调优方法。

第二、YOLO v5算法的原理和实现的研究。通过阅读相关论文和YOLO v5的源码，理解了YOLO v5的原理和算法实现，完成了YOLO v5在VisDrone数据集上的模型训练。

第三、分析YOLO v5算法在VisDrone数据集上的实验结果，结合相关理论分析，引入了增大特征图分辨率和多尺度训练这两种目标检测提升分类和定位方法，在VisDrone数据集上训练。结合相关性能评估方法，分析实验结果并得出结论。

关键词：目标检测；VisDrone无人机数据集；YOLO v5；增大特征图分辨率；多尺度训练

ABSTRACT

Object detection is a technology that searches for a specific object from a picture and extracts information such as location and category. Traditional target recognition technology has shortcomings such as untargeted region selection strategy, high algorithm time complexity, and poor robustness of hand-designed features, which make the detection effect unsatisfactory. With the development of deep learning technology, algorithms for the application of deep learning in the field of target detection are emerging in an endless stream. Many excellent deep learning target detection algorithms have emerged, and target detection has been used in various industrial fields, such as urban Face detection has been widely used.

This paper selects the YOLO v5 algorithm as the basic algorithm, and applies it to the VisDrone UAV dataset published by the Machine Learning and Data Mining Laboratory of Tianjin University. This paper introduces an optimization method of increasing the resolution of feature maps and multi-scale fusion training, which improves the classification and localization performance of the YOLO v5 algorithm on the VisDrone dataset.

The research contents of this thesis are as follows:

First, the analysis of the aerial drone dataset VisDrone. The performance tuning method of YOLO v5 algorithm is determined by analyzing the number of labels and label distribution of the dataset by visual method.

Second, research on the principle and implementation of the YOLO v5 algorithm. By reading the relevant papers and the source code of YOLO v5, I understood the principle and algorithm implementation of YOLO v5, and completed the model training of YOLO v5 on the VisDrone dataset.

Third, analyze the experimental results of the YOLO v5 algorithm on the VisDrone data set, combined with relevant theoretical analysis, introduce two methods of target detection to improve classification and positioning, increasing the resolution of the feature map and multi-scale training, and train on the VisDrone data set. . Combined with relevant performance evaluation methods, the experimental results are analyzed and conclusions are drawn.

**Keywords:** Object detection; VisDrone UAV dataset; YOLO v5; Increase feature map resolution; Multi-scale training

目录

[摘 要 I](#_Toc102318994)

[ABSTRACT III](#_Toc102318995)

[目录 V](#_Toc102318996)

[第一章 绪 论 1](#_Toc102318997)

[1.1 研究背景及意义 1](#_Toc102318998)

[1.2 国内外研究现状 2](#_Toc102318999)

[1.3 主要研究内容 3](#_Toc102319000)

[1.4 本文篇章结构 3](#_Toc102319001)

[第二章 目标检测算法相关研究 5](#_Toc102319002)

[2.1 目标检测概述 5](#_Toc102319003)

[2.2 YOLO算法概述 6](#_Toc102319004)

[2.2.1 YOLO v1概述 6](#_Toc102319005)

[2.2.2 YOLO v3概述 7](#_Toc102319006)

[2.3 YOLO v5算法介绍 8](#_Toc102319007)

[2.3.1 输入端 8](#_Toc102319008)

[2.3.2 骨干网络 9](#_Toc102319009)

[2.3.3 颈部(Neck)网络 10](#_Toc102319010)

[第三章 基于YOLO v5的无人机目标检测方法 13](#_Toc102319011)

[3.1 无人机数据集分析 13](#_Toc102319012)

[3.1.1 训练集分辨率统计 14](#_Toc102319013)

[3.1.2 标签数量、真实框尺寸分布统计 15](#_Toc102319014)

[3.1.3 数据集图片观察 15](#_Toc102319015)

[3.2 数据集预处理 16](#_Toc102319016)

[3.2.1 数据集转化为YOLO v5训练格式 16](#_Toc102319017)

[3.2.2 数据增强 16](#_Toc102319018)

[3.3 YOLO v5在VisDrone数据集上的训练方法 17](#_Toc102319019)

[3.3.1 增大特征图分辨率 18](#_Toc102319020)

[3.3.2 多尺度训练 19](#_Toc102319021)

[3.3.3 早停法(Early Stopping) 19](#_Toc102319022)

[第四章 实验内容及结果分析 21](#_Toc102319023)

[4.1 实验设计及细节 21](#_Toc102319024)

[4.1.1 增大特征图分辨率实验 21](#_Toc102319025)

[4.1.2 多尺度训练 21](#_Toc102319026)

[4.1.3 实验平台 21](#_Toc102319027)

[4.1.4 模型的超参数 22](#_Toc102319028)

[4.1.5 本文遇到并解决的问题 22](#_Toc102319029)

[4.2 评价指标 23](#_Toc102319030)

[4.2.1 分析概述 23](#_Toc102319031)

[4.2.2 评价指标介绍 23](#_Toc102319032)

[4.2.3 实验结果展示 26](#_Toc102319033)

[4.2.4 实验结果分析 29](#_Toc102319034)

[第五章 实验结论及未来展望 31](#_Toc102319035)

[5.1 实验结论 31](#_Toc102319036)

[5.2 论文改进方向 31](#_Toc102319037)

[5.2.1 实验增大训练图像分辨率的上限 31](#_Toc102319038)

[5.2.2 对多尺度训练的优点的定量分析 31](#_Toc102319039)

[5.3 展望 32](#_Toc102319040)

[参考文献 33](#_Toc102319041)

[致 谢 35](#_Toc102319042)

# 绪 论

## 研究背景及意义

目标检测是计算机视觉的重点核心问题之一，也是研究理解图像高层语义信息的基础。目标检测的主要任务是根据输入的视频或者图像，找到所有感兴趣的目标，然后给出它们的位置和类别信息。‎[26]其中，输出的位置信息通常以预测边界框（Bounding Box）的形式体现，并显示该边界框判定的目标类别和置信度分数。传统的目标检测算法通常的框架是区域选择、特征提取和分类器分类３个模块组成，但是因为存在区域选择策略没有针对性、算法时间复杂度高、手工设计的特征鲁棒性较差等缺点，使得检测效果不理想。‎[25]2012年，AlexNet‎‎[4]获得了2012年度ImageNet‎[5]大规模视觉识别挑战赛冠军，并且其检测效果更令传统算法难以望其项背，使大众的注意力重新聚焦到深度学习神经网络上。因此深度学习引入目标检测领域之事水到渠成，致使目标检测技术取得突飞猛进的进步。现如今深度学习目标检测技术主要可分为两个方向，分别是以R-CNN(Region-based Convolutional Neu-ral Network) ‎‎[6]‎[7]‎[8]系列为代表的基于候选框 (Region Proposal)的两阶段深度学习目标检测算法，和以YOLO( You Only Look Once)‎[9]‎[10]‎[11]‎[12]‎[14]系列‎为代表的基于回归方法的单阶段深度学习目标检测算法。‎[25]

目前基于深度学习的目标检测在视频目标实时检测、人脸识别、行人检测、车辆检测、交通标识检测和医学影像识别等领域有广泛的应用。‎[25]其中基于无人机视觉的目标检测具有巨大的商业潜力和广泛的应用领域。

与常见的目标检测应用相比，无人机检测因为其在空中的特点，所以拥有更广阔的视野，并且可以检测的区域更大。而且同样是空中拍摄，相比于直升机拍摄，无人机的使用成本更低、适应环境的能力更强。因此被广泛使用于拥挤交通路口直播、电力巡检、灾害地区搜救等领域。并且无人机检测场景中的目标检测技术已广泛应用于许多实际应用中，如农业植保、警务安防和灾害救援等。

而无人机数据集相比于COCO‎[1]等常见目标检测数据集，具有视角高、视野广、目标小且密集等特点‎[2]。另外，由于无人机数据集中目标遮挡现象也十分常见‎[2]，这些问题都影响了无人机自动化应用。因此发展无人机目标检测技术势在必行。

## 国内外研究现状

在研究背景及意义中，目标检测技术难以应用于无人机上的一个原因就是空中拍摄视角导致存在大量小物体目标。因此，许多研究通过针对小目标检测精度的目标来改造神经网络模型。Sommer、Schuchert等人在《Fast Deep Vehicle Detection in Aerial Images》‎[3]中在基于Fast RCNN和Faster RCNN基础上通过修改区域生成网络(Region Proposal Network, RPN)锚框和用作特征图的最后一个卷积层的输出分辨率来补偿航空图像的较低空间分辨率和因此产生的较小的物体目标。而YOLO v2则根据数据集中目标的尺寸分布，用K-means聚类锚框尺寸取代手工设计锚框的方法，来提升小目标的检测精度，提高模型的召回率。

但是这些算法主要是针对数据集的固定尺寸进行优化，对于多尺度的场景，则仍然不能解决问题。因此多种研究针对多尺度检测的问题，在Wei Liu、Dragomir Anguelov等人提出的《SSD: Single Shot MultiBox Detector》‎[16]中通过提取特征提取网络中的不同分辨率的特征图的预测结果，来预测不同大小的目标。但是浅层网络提取的特征语义信息缺乏，因此对小物体目标的检测精度仍然不理想。而Lin T Y、Dollar P、Girshick R等人在《Feature Pyramid Networks for Object Detection》‎[18]中将不同分辨率的特征层进行堆叠，得到融合了多个尺度信息的特征图，从而实现对多个尺度的物体目标的检测。

## 主要研究内容

第一、对空中无人机数据集VisDrone的分析。通过可视化方法分析数据集的标签数量、标签分布来确定YOLO v5算法的性能调优方法。

第二、YOLO v5算法的原理和实现的研究。通过阅读相关论文和YOLO v5的源码，理解了YOLO v5的原理和算法实现，完成了YOLO v5在VisDrone数据集上的模型训练。

第三、分析YOLO v5算法在VisDrone数据集上的实验结果，结合相关理论分析，引入了增大特征图分辨率和多尺度训练这两种目标检测提升分类和定位方法，在VisDrone数据集上训练。结合相关性能评估方法，分析实验结果并得出结论。

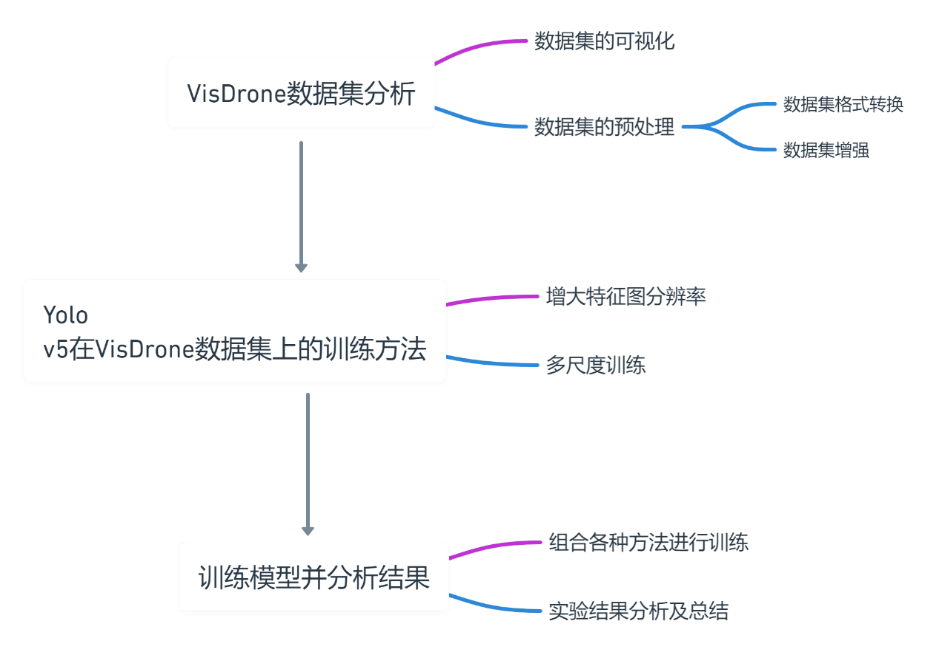


图1.1 本文研究路线

## 本文篇章结构

第一章介绍目标检测背景和意义，国内外无人机目标检测的研究现状，并简略概述本文主要的研究内容。此章结尾给出本文结构。

第二章 对目标检测算法进行概述，并介绍YOLO系列算法和详述本文使用的YOLO v5算法。

第三章 概述本文使用的VisDrone数据集，然后对其进行分析和对预处理方法介绍。并阐述了本文在YOLO v5算法基础上采用的增大特征图分辨率和多尺度训练两种模型优化方法。

第四章 阐述了本文的实验内容和细节，并对实验结果进行分析。

第五章 对实验进行总结并给出本文个人认为的改进方向和对无人机目标检测未来的展望。

# 目标检测算法相关研究

## 目标检测概述

在传统目标检测算法中，对于不同尺度的目标，常见的方法是将原图进行特征提取，构建出不同分辨率的图像金字塔，然后在每层图像金字塔中使用滑动窗口检测算法即将检测器按一定的步长在图像金字塔中滑动，以此来检测目标。用此方法可以在图像金字塔的底部来检测出小目标。也可以只用一次原始图像，用不同分辨率的分类器检测目标，以求在比较小的窗口分类器中检测到小目标。

深度学习是一种能够通过监督、半监督或无监督的训练方法自动学习训练集中隐藏的数据特征结构的多层神经网络算法。‎[25]相比于人工设计手工特征，深度学习比手工设计的特征提取算法更擅长从输入图像中提取出更高阶、更抽象化的层级特征‎[15]。因此深度学习抽取的特征比传统方法提取的具有更强的表征能力和鲁棒性。‎[27]因此，深度学习被越来越多的研究者应用于计算机视觉领域的研究和应用中。特别是在目标检测领域，大量优秀的基于深度学习的目标检测算法层出不穷。目前，已有的目标检测算法存在两种分类标准:1)是否存在显式的区域建议;2)是否显式定义先验锚框‎。‎[27] 其中，根据分类标准1)可将目标检测算法分为两阶段(Two-stage)和单阶段(One-stage)算法。其中，两阶段目标检测算法或者也可以说是基于区域建议的目标检测算法的原理是先从输入图像中提取出感兴趣区域，然后再对此区域内的局部图片进行类别分类。此类方法的代表性算法有R-CNN、Fast R-CNN等。相比于两阶段目标检测算法，单阶段目标检测主要是基于逻辑回归的算法。此类算法直接跳过了生成感兴趣区域的阶段，直接对输入图像进行回归，一次性得到类别和位置信息。而在单阶段目标检测算法中的代表性算法有YOLO、SSD‎[16]等。



图2.1 目标检测算法框架

根据是否存在显式的区域建议分类标准，可绘制如上图1.1的目标检测算法的框架流程图。

## YOLO算法概述

YOLO（You Only Look Once: Unified, Real-Time Object Detection）‎[9]，是Joseph Redmon和Ali Farhadi等人于2015年提出的基于单阶段的目标检测算法，开创了单阶段目标检测的赛道。其中影响深远的主要是两个算法，分别是YOLO v1‎[9]和YOLO v3‎[11]，本小节将主要介绍此两种算法。

### YOLO v1概述

与两阶段目标检测算法不同，YOLO直接对输入图像进行回归，一次性得到类别和位置信息。因为所有的检测流程是端对端的，即输入图像直接一次性检测得到类别和位置信息，所以可以直接针对神经网络直接进行优化。

YOLO采用了自顶而下，分而治之的思想，直接从输入图像中一次性预测到边界框的坐标，还有边界框中的是否包含物体的置信度和物体的类别。因此拥有检测速度快、泛化能力强的优点，而且相比于两阶段检测能够更好的避免误检为背景的错误。

输入的图像在YOLO算法中被划分为个区域方格。在每个区域方格中都有B个边界框，每个边界框中都包含五个参数，其分别为。其中表示了边界框的中心与区域方格单元边界的相对距离，而是边界框相对于整张输入图片的比值。在YOLO中置信度则被定义为：‎[9]

其中代表模型预测网格内存在物体的概率，而表示预测边界框和真实框之间的相关性，详见下文评价指标介绍。并且YOLO为每个网格预测了C种条件类别概率，表示该网格存在物体的条件下属于某个类别的概率。因此这些预测值被编码为的张量。

### YOLO v3概述

YOLO v3（YOLOv3: An Incremental Improvement）‎[11]的发布相当于苹果公司发布iphone4手机，将基于深度学习的目标检测算法从能用提升到了好用的地步，至今仍被大量工业界使用。YOLO v3提出了如下几种重大改进：

#### 骨干特征提取网络使用Darknet53

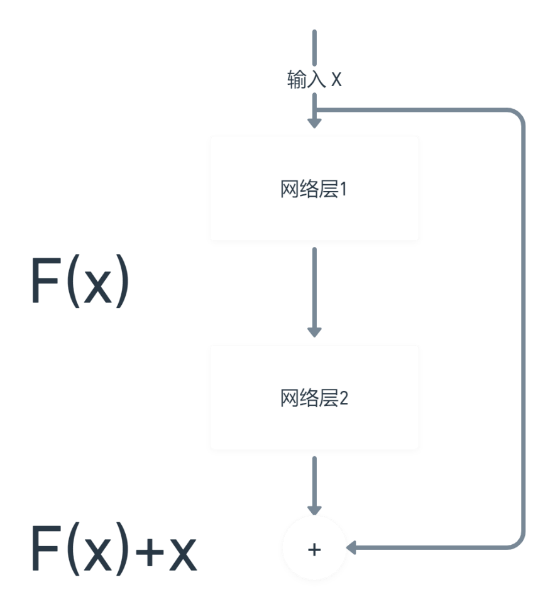


图2.2 残差结构

Darknet53‎[11]在Darknet-19‎[10]的基础上增加了大量的残差结构Residual，并且使用步长为2，卷积核大小为3×3卷积层Conv2D代替池化层Maxpooling2D。其中，残差结构是Kaiming He 及其同事在论文《Deep Residual Learning for Image Recognition》‎[17]中提出的一种结构。即将一些深层网络短路，然后拼接，解决网络层数增加带来的退化问题。结构如上图2.2。

Darknet53加入的残差结构增加网络的深度，因此能够提取更高级别的图像语义特征，同时残差的结构避免了深度神经网络带来的梯度消失或爆炸问题。同时残差结构单元里边输入首先会经过一个1×1的卷积层将输入通道降低一半，然后再进行3×3的卷积，这在相当程度上帮助网络减少了计算量，使得网络的运行速度更快，效率更高。

另外，Darknet53使用了通过步长为2的卷积取代池化层的最大池化来实现网络的下采样。因为相比于人为设定好的先验知识（选取覆盖范围内最大的值，认为最大值包含的信息是最多的），而步长为2的卷积层参数是通过学习得到的，增加了不确定性。这种不确定性增加了网络的学习能力。

#### 多尺度融合

由于深层的特征图包含的语义信息丰富，但是目标位置信息粗略，而浅层特征图包含的目标位置信息明确，但是噪音较多。‎[28]因此采用FPN‎[18]的思想，将深层特征图上采样后和特征图进行融合，然后再在多尺度的融合特征图上分别做检测，实现对大、中、小目标的检测。

## YOLO v5算法介绍

YOLO v5是本文采用的算法，因此另起一小节进行详述。YOLO v5相比于上文的YOLO v1和YOLO v3，YOLO v5的结构优化主要可分为三个部分，分别是输入端、骨干网络和颈部(Neck)网络。‎[29]本小节将对这三个部分进行阐述。

### 输入端

该阶段主要是对输入图像的一些预处理。YOLO v5使用了Mosaic‎[12]数据增强、集成自适应锚框计算‎[10]的技巧。

#### Mosaic数据增强

Mosaic数据增强就是通过对四张图片像素级数据增强（翻转、缩放、光学畸变）后，按照四个位置排列同时每张图片随机擦除，最后四张图片拼接在一起送入神经网络中进行训练。

#### 集成自适应锚框计算

YOLO系列在YOLO v2引入了先验框的概念，同时使用了K-means聚类获得自适应先验框。在YOLO v5中，作者首次将自适应锚框计算集成到训练代码中，而不是之前的单独计算先验框尺寸。

### 骨干网络

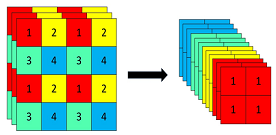


图2.3 Focus模块图像处理步骤

该阶段YOLO v5采用了CSPDarknet53骨干特征提取网络，相较于YOLO v3的Darknet53，作者结合了CSPNet‎[19]的思想，将数据输入分为两个部分，一个部分经过网络层计算，另一部分直接短路，然后两个部分进行组合。这样能够减少计算量和增强梯度表现。YOLO v5通过CSPDarknet53主要减少了计算的瓶颈和内存的消耗，并在训练图片进入骨干特征提取网络前增加了一层Focus模块。如上图2.3，Focus模块能够降低图片的宽高，扩充其维度。Focus流程是抽取输入图片宽高方向每隔一个像素抽样一次，然后将相隔像素进行组合，最后图片的宽高缩小两倍，通道数扩充了4倍。Focus层的主要目的是减少层数，减少参数，减少FLOPS，减少CUDA内存，增加前向和后向传播速度，同时最小化对mAP的影响。

### 颈部(Neck)网络

颈部(Neck)网络通常位于骨干网络和检测头输出网络之间，因此被喻为颈部(Neck)网络。它的作用是提升骨干网络提取出的特征的多样性和鲁棒性。YOLO v5主要在颈部(Neck)网络应用了SPP‎[20]模块和FPN+PAN‎[18]‎[21]结构。

#### SPP模块

SPP（Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition）即空间金字塔池化‎[20]，能够将任意尺寸的输入，生成固定尺寸的输出。

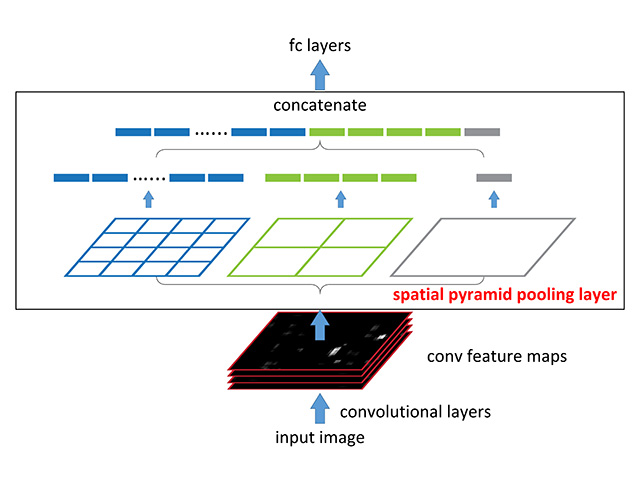


图2.4 空间金字塔池化网络结构

如上图2.4，SPP网络的具体操作是将输入的特征图划分为不同大小的网格，然后用最大池化提取出块中最大值，即每一个网格提取出一个特征。最后将所有块提取的特征组合就得到了固定大小的特征向量。

#### FPN+PAN结构

FPN自上而下的融合多层特征图，将浅层特征通过下采样和深层特征做融合得到进行预测的特征图。FPN自顶向下传达强语义特征。而PAN又自底而上的传递强定位信息，从不同的骨干层得到的每个特征图进行组合。

# 基于YOLO v5的无人机目标检测方法

## 无人机数据集分析

现如今，配备摄像头的无人机已广泛和深度应用于包括农业、航空摄影、快速交付和监视等多个领域中。对于这些领域里无人机拍摄的大量图片的处理成为一个新的难题。并且空中无人机数据集存在以下几种问题：

#### 数据集少

应用了深度学习的目标检测算法模型需要通过大量的数据得到局部最优解，但是无人机多为俯视视角，但是大多数的目标检测训练集为平视视角，不适合作为无人机目标检测算法的训练数据集，并且也用这些平视视角的预训练模型也会收到视角的影响。

#### 无人机在拍摄视角上存在一定的局限性

无人机通常以俯视的视角进行监视拍摄，特征提取角度过于单一。而且无人机离地面距离较远，待检测目标相对来说呈现出小目标的特点，小目标往往像素比较少，特征单一，在检测的时候容易出现漏检或者误检的情况。这就给无人机检测增加了一定的难度。

以上的这些因素都给无人机检测带来巨大的挑战，传统的目标检测算法很难提取到足够好的特征来进行检测，而基于深度学习的无人机检测使得检测的准确率和适用范围获得了大幅度的改善。因此，引入计算机视觉技术颇为重要。为此，中国天津大学机器学习与数据挖掘实验室的 AISKYEYE 团队制作了VisDrone无人机数据集。‎[5]

该数据集由400片视频组成，包含265,228帧和10,209张静态图像，由各种无人机摄像头拍摄，涵盖了许多的方面，包括位置（取自中国相隔数千公里的 14 座不同城市）、环境（城市和乡村）、物体（行人、车辆、自行车等）和密度（稀疏和拥挤的场景）。数据集是使用各种无人机平台（即不同型号的无人机）、在不同场景以及各种天气和照明条件下收集的。这些视频分割的图片包含手动注释的260万个边界框，例如行人、汽车、自行车和三轮车。为了更好地利用数据，VisDrone数据集还提供了一些重要的属性，包括场景可见性、对象类别和遮挡。因此本文使用VisDrone数据集作为实验数据集。

而在设计实验之前，本文根据阅读相关文献和代码学习得到的经验，对数据集的分析能够帮助我们理解数据的分布和增强对模型的优化方向的认识。

### 训练集分辨率统计

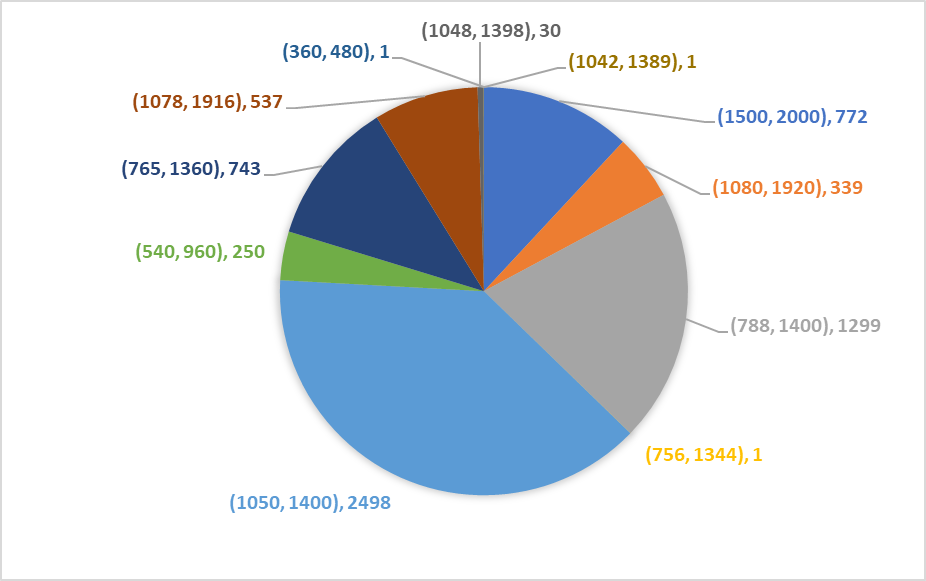


图3.1 VisDrone训练集图像分辨率统计图

本文统计了训练集图像分辨率分布，并绘制了上图3.1。其中，第一项表示分辨率，第二项表示数量。通过对训练集分辨率的统计，可以为后文的增大数据集分辨率的具体最优参数值的设定提供帮助。

### 标签数量、真实框尺寸分布统计

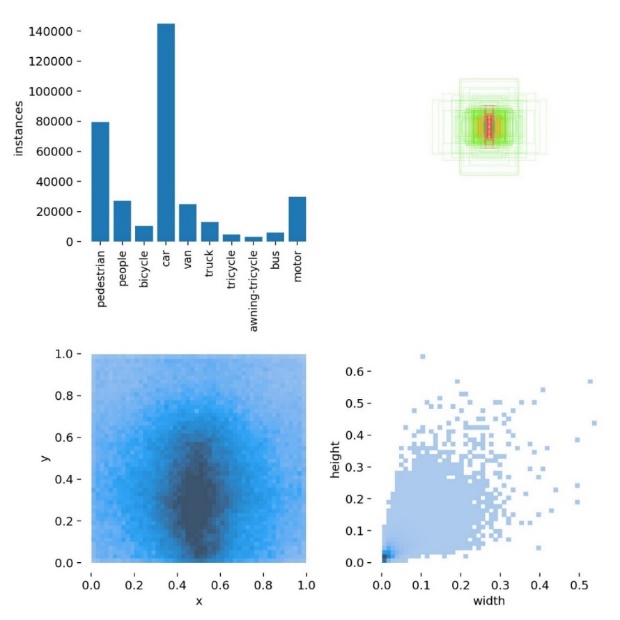


图3.2 标签数量、真实框尺寸分布图

本文统计了标签数量、真实框尺寸的分布，并绘制了上图3.2。

### 数据集图片观察



图3.3 数据集图片

浏览数据集的图片可以帮助本文优化模型在数据集上的性能。如上图3.3。

## 数据集预处理

### 数据集转化为YOLO v5训练格式

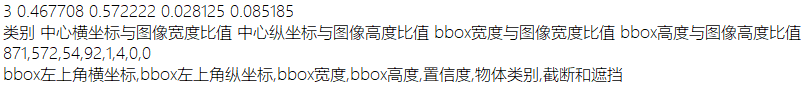
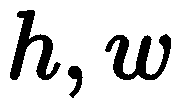
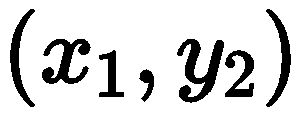
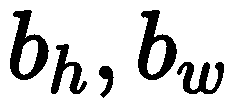
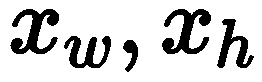


图3.4 YOLO数据集格式和VisDrone数据集格式

从VisDrone官网下载数据集后，官方已经划分好了训练集和验证集还有测试集。其中训练集含有6471张图片，验证集含有548张图片，测试集含有1610张照片。‎[30]根据图3.4可知，VisDrone标注的数据格式和YOLO格式不一致。因此本文通过脚本将VisDrone格式转化为YOLO格式。

假设图像的高和宽分别为，bbox的左上角坐标为，bbox高度和宽度分别为，则可求得bbox中心坐标与图像宽高比值为：

假设bbox的宽高与图像宽高比值为，则可求对应关系为：

### 数据增强

通过对数据集的镜像、随机调整亮度和对比度等像素级操作，扩充数据集大小。然后采用了上文介绍的

Mosaic数据增强方法。通过对四张图片像素级数据增强（翻转、缩放、光学畸变）后，按照四个位置排列同时每张图片随机擦除，最后四张图片拼接在一起送入神经网络中进行训练。

## YOLO v5在VisDrone数据集上的训练方法

从图3.1可知，训练集中存在大量高于分辨率的图像，因为YOLO v5默认训练图片分辨率为，而YOLO v5在训练时会将训练图片缩放到分辨率，因此会影响模型对小物体的检测性能。

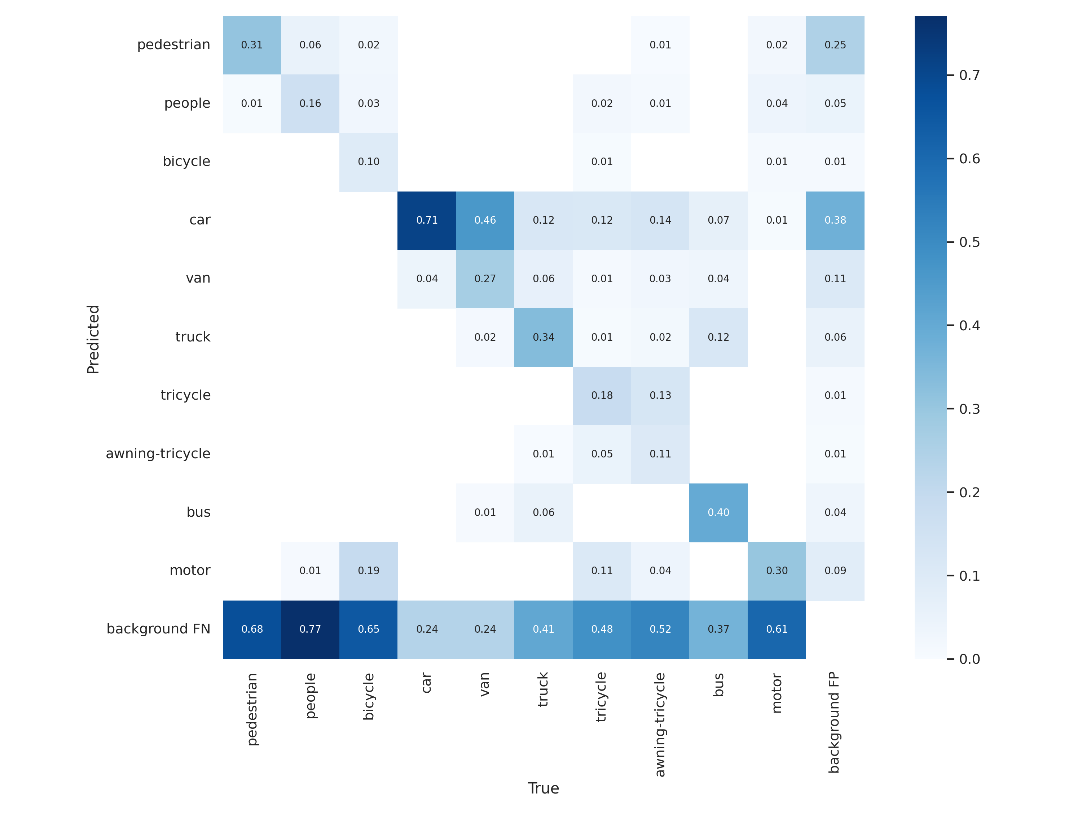


图3.5 YOLO v5在VisDrone数据集测试结果混淆矩阵

而从图3.2和图3.3可知，VisDrone训练集中多为小物体，且数据集中的真实框多数为只占图像面积10%以下的尺寸。本文还绘制了YOLO v5在VisDrone数据集上基准测试的模型检测结果的混淆矩阵‎[22]‎[23]，如上图3.5

可以从中发现YOLO v5在VisDrone数据集上对小物体的分类精度不佳。因此，本文引入了增大特征图分辨率和多尺度训练两种优化方法。

### 增大特征图分辨率

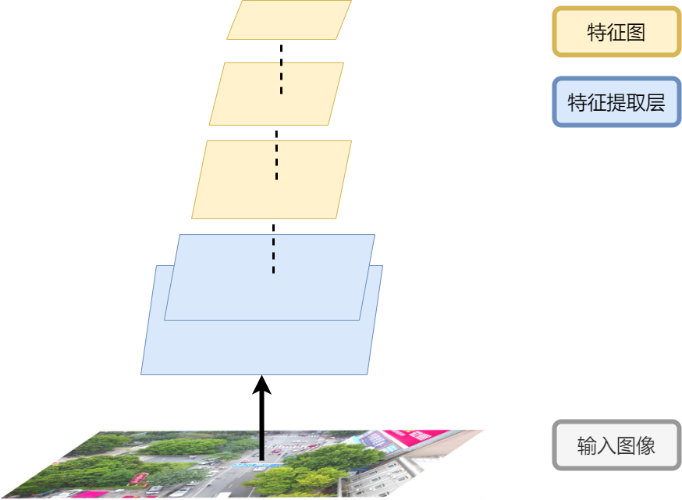


图3.6 CSPDarknet特征图概要

深度学习目标检测本质上是通过大批量数据的训练，让神经网络学习到图像的特征并加以识别。如下图所示，YOLO v5通过CSPDarknet53特征提取网络分别获得了下采样8倍、16倍和32倍的特征图。如上图3.6。

因为残差结构使深层神经网络得以广泛应用，能够从图像中获得更强的语义信息，但是由于小物体目标尺寸往往只有几个像素，在下采样过程中会被缩小至消失。并且由于VisDrone数据集大部分是高分辨率图像，而YOLO v5默认的训练分辨率为. 因此训练图片先会被缩小到大小，然后进行训练。此过程中会导致图片中的小物体变得更小，从而影响YOLO v5模型对VisDrone数据集的检测效果。

因此，本文很直观的想到提高训练图像的分辨率，可是模型是如何处理增大的输入呢？

答案是在颈部(Neck)网络中介绍的SPP模块,SPP模块使得模型对输入的数据尺度没有任何限制，因为SPP结构能够将任意的输入尺寸输出为固定尺寸。所以本文实验只需要调整YOLO v5对图像预处理时缩放的尺寸大小即可。

### 多尺度训练

多尺度训练即对输入图像进行缩放，从而达到扩大数据集大小，使模型降低对目标大小的敏感性，增强了模型检测性能的鲁棒性。同时也可以在模型测试阶段引入多尺度，然后通过NMS后处理融合多个尺度的检测，在牺牲一部分时间的基础上获得更高的精度。

### 早停法(Early Stopping)

早停法，实际上也是一种正则化的策略，可以理解为在网络训练不断逼近最优解的过程种（实际上这个最优解是过拟合的），在梯度等高线的外围就停止了训练，所以其原理上和L2正则是一样的，区别在于得到解的过程‎[24]。

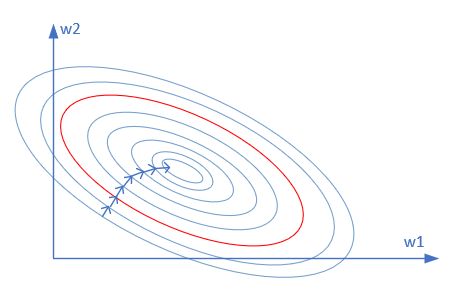


图3.7 梯度下降等高线图

如上图3.7所示，假设红色是本文需要求的局部最优解，我们模型训练的过程就是不断逼近红线的过程。但是网络训练在逼近最优解过程中，往往会过拟合，具体表现为训练集的损失在不断下降，而验证机的损失却在上升。因此，引入早停法就是防止过拟合的一种手段。目的就是为了使模型在到达红线附近时就停止训练。

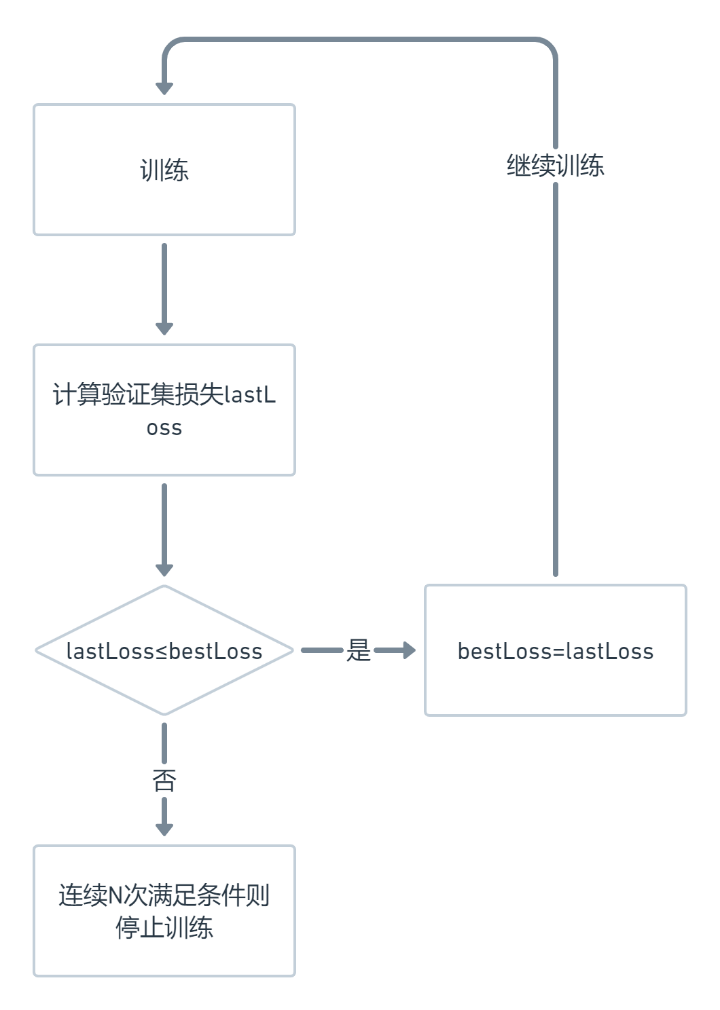


图3.8 早停法流程图

具体算法如上图3.8，记录训练过程中最小的验证集损失，当连续N次epoch（通常N取10次或者更多次）都大于最小的验证集损失时，则认为模型已经达到最小的损失了，便可以停止训练。

# 实验内容及结果分析

## 实验设计及细节

### 增大特征图分辨率实验

根据‎3.1.1节对VisDrone训练集图片分辨率的统计，在6471张训练图片中，仅有一张图片长边小于YOLO v5默认训练图片分辨率640，且只有3.9%(251张)图片短边小于640。这意味着，如果按照YOLO v5默认训练图片分辨率640进行训练，所有的训练集都会被缩放，导致小目标物体变得更小，不利于模型的学习。因此，本文设计了和两个分辨率进行训练。

### 多尺度训练

本文设计的缩放因子为50%，即假设训练分辨率为，则模型会在[0.5S,1.5S]步长为32的闭区间里随机选择一个数值，然后缩放为此分辨率。

因此最终本文一共设计了五组实验，分别是YOLO v5在VisDrone数据集上的基准实验。YOLO v5在增大训练图片分辨率为960和1280的两组实验，YOLO v5在分辨率为640和1280下引入多尺度训练的两组实验。然后所有训练得到的模型在VisDrone2021-DET test set-dev测试集进行测试。

### 实验平台

因为个人笔记本的显卡为GTX1060，无法承担实验需求。本文通过购买谷歌Colab云服务的会员，利用Colab随机提供的NVIDIA T4或者P100显卡进行训练。

### 模型的超参数

本文在 Pytorch 1.10.0 和cuda 11.6环境上训练。本文所有的模型都使用 NVIDIA P100或者T4 GPU 进行训练和测试。在训练阶段，本文使用预训练自ImageNet数据集上的 YOLOv5m 的预训练模型。本文使用了早停法用来判断模型是否收敛，前 2 个 epoch 用于热身。本文使用 adam 优化器进行训练，并使用 3e-4 作为 cosine lr的初始学习率。本文分别使用32、16、8作为模型训练图片分辨率为640、960、1280时的Batch size大小。

### 本文遇到并解决的问题

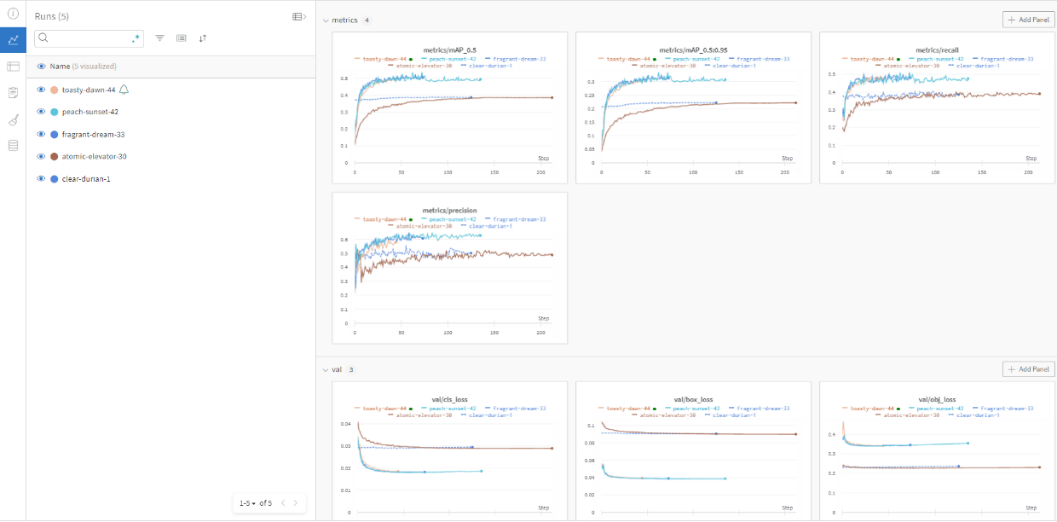


图4.1 wandb可视化训练网页界面

本文引入了wandb库来可视化训练过程，展示训练的中间结果。如上图4.1。

因为本文使用了谷歌Colab平台作为训练平台，而Colab平台存在着长期无交互或者连续运行时长达到12小时，就会终止服务的问题。本文一次的实验训练时长通常为60个小时以上，因此本文的解决方式是首先使用网页脚本定时刷新网页，造成正在与Colab交互的假象，避免判定为长期无交互行为。第二，原理上模型的训练是对模型进行数个epoch（世代）的正向和反向传播。因此个人设置为每隔5个epoch（世代）就保存到硬盘，如果中断，则从硬盘中加载模型继续训练。

通过观察训练耗时，发现有一部分时间用于加载训练数据上面。因此，本文引入了缓存的计算机思想，通过将训练数据先加入内存中作为Cache来减少数据加载的I/O时间。

## 评价指标

### 分析概述

目前业界对于深度学习图像目标检测相关算法模型的评价，通常是继续使用机器学习领域的评价指标，即选择使用查准度(Precision)、召回率(Recall)、P-R曲线等指标进行评价。但是对于目标检测领域的问题，如若单单选择将两个评价标准拼凑使用则会出现一定弊端。因此适用于目标检测算法模型的多种评价指标被讨论制定，形成评价标准体系，用于进行模型性能讨论。本节内容围绕实验结果的分析及相关原因展开，介绍多种评价指标、给出模型产生相关结果分析。

### 评价指标介绍

#### 交并比IoU‎[21]

IoU是一个简单的测量标准，只要是在输出中得出一个预测范围(bounding boxes)的任务都可以用IoU来进行测量。为了可以使IoU用于测量任意大小形状的物体检测，需要ground-truth bounding boxes（即在训练集图像中标出要检测物体的范围）和算法得出的预测范围。即IoU这个标准用于测量真实和预测之间的相关度，相关度越高，该值越高。

故本文假设真实框为A，预测框为B，面积用S表示，则IoU的计算公式为：

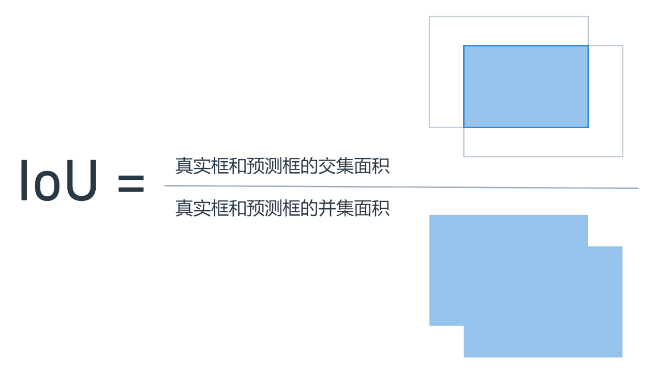


图4.2 IoU计算公式

#### 查准度、召回率‎[22]‎[23]

查准度和召回率的计算基于真正例(true positive)、假正例(false positive)、真反例(true negative)和假反例(false negative)‎[23]这四个值。这四个值为如下解释：

目标检测中规定某个类别的IoU大于某个阈值，则视为正例P（Positive），否则为反例（Negative）。而类别预测正确，则为T（True），预测错误，则为F（False）。而实验结果的正反例和预测结果的正误两两组合可作为目标检测结果查准度、召回率和mAP等指标的先决条件。可作如下表4.1，方便理解。

表4.1 真正例 (true positive)、假正例 (false positive)、真反例 (true negative) 、假反例 (false negative)‎[23]

|  |  |  |
| --- | --- | --- |
| 真实情况 | 预测结果 | |
| 正例 | 反例 |
| 正例 | TP（真正例） | FN（假反例） |
| 反例 | FP（假反例） | TN（真反例） |

查准度为所有预测为正样本中，真正是正样本的比例，查准度是针对预测结果的评价指标，它代表的是有多少是真正的正样本存在被预测为正的样本中。被预测为正的样本有两种可能，一种就是把正类预测为正类(TP)，另一种就是把负类预测为正类(FP)，也就是即：

召回率为所有真正是正样本中，找到了的正样本的比例，而召回率是针对数据集全体而言的，它代表的是样本中的正例被预测正确的所占的比例。因此也有两种可能，一种是将数据集中的正类预测成正类(TP)，另一种就是把数据集中的正类预测为负类(FN)，即：

#### P-R曲线和AP（Average Precision）‎[22]‎[23]

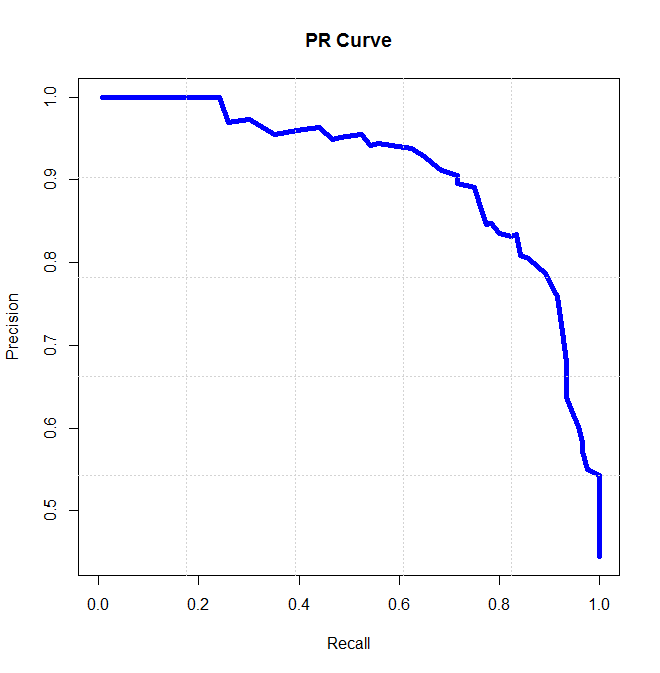


图4.3 P-R曲线

通常情况下，查准率和召回率是一对矛盾的值。因此为了能更好的评估模型的性能，本文引入了P-R曲线和AP（Average Precision）。P-R曲线即以查准率为纵轴、召回率为横轴作图得到的曲线。

本文作如上图4.3以方便理解。而AP值指P-R曲线的面积，这里本文用P(r)表示P-R曲线，则AP的计算公式为：

但是，在实际应用中，获取的值都是离散值，因此需要对P-R曲线进行平滑处理，处理公式如下：

然后将平滑后的P-R曲线进行计算。

#### 均准率（mean Average Precision, mAP）‎[22]‎[23]

即所有类别的AP值的平均值。

#### 混淆矩阵‎[22]‎[23]

混淆矩阵也被称为误差矩阵，是表示精度评价的一种标准格式，用n行n列的矩阵形式来表示。混淆矩阵中的横轴是模型预测的类别数量统计，纵轴是数据真实标签的数量统计。其中对角线，表示模型预测和数据标签一致准确率。对角线上数字越偏向1，在可视化结果中颜色越深，说明模型在该类的预测准确率越高。如上文图3.5所示。

### 实验结果展示

通过图3.2，本文认为提高YOLO v5在小物体的检测能力是优化YOLO v5算法在VisDrone数据集上的性能的关键。因此，本文引入了增大特征图分辨率和多尺度训练的方法，并在YOLO v5在VisDrone数据集上的训练方法中阐述了其可能性。下面是本文在VisDrone数据集上，在YOLO v5的基础上引入这两种方法，对实验结果的展示和分析。

表4.2 模型mAP与推理速度

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 方法 | 分辨率 | mAP(%) | | 速度(ms/(32,3,S,S)) | | | |
|  | mAP@.5 | mAP@.5:.95 | 预处理 | 推理 | NMS | 总和 |
| YOLO v5 | 640 | 32.6 | 18.3 | 0.2 | 13.5 | 5.5 | 19.2 |
| YOLO v5+多尺度训练 | 640 | 31.9  (↓0.7) | 17.8(↓0.5) | 0.2 | 13.4 | 3.9 | 17.5  (↓1.7) |
| YOLO v5 | 960 | 39.9  (↑7.3) | 23.2(↑4.9) | 0.3 | 24.1 | 6.7 | 31.1  (↑11.9) |
| YOLO v5 | 1280 | 43.5  (↑10.9) | 25.9(↑7.6) | 0.8 | 38.7 | 6.9 | 46.4  (↑27.2) |
| YOLO v5+多尺度训练 | 1280 | 43.9  (↑11.3) | 26.0(↑7.7) | 0.5 | 38.9 | 8.6 | 48.0  (↑28.8) |

表4.3 模型P-R和mAP

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 方法 | 分辨率 | mAP(%) | | P-R(%) | |
|  | mAP@.5 | mAP@.5:.95 | Precision | Recall |
| YOLO v5 | 640 | 32.6 | 18.3 | 43.1 | 35.9 |
| YOLO v5+多尺度训练 | 640 | 31.9  (↓0.7) | 17.8(↓0.5) | 45.1 | 34.2 |
| YOLO v5 | 960 | 39.9  (↑7.3) | 23.2(↑4.9) | 49.7 | 42.2 |
| YOLO v5 | 1280 | 43.5  (↑10.9) | 25.9(↑7.6) | 53.1 | 44.9 |
| YOLO v5+多尺度训练 | 1280 | 43.9  (↑11.3) | 26.0(↑7.7) | 55.0 | 44.1 |

本篇文章一共设置了五组实验，训练结果在VisDrone2021-DET test set-dev测试集上得到如下性能评估结果，如上表4.2和表4.3，其中红色文字表示相对于基准模型性能的变化。

### 实验结果分析

#### 增大特征图分辨率

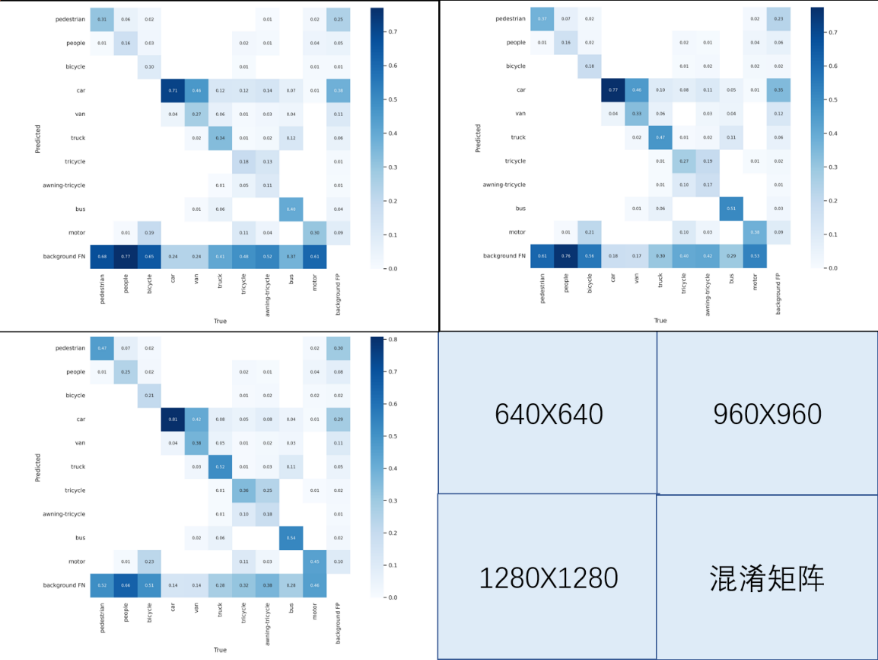


图4.4 多种分辨率训练模型的混淆矩阵

根据图4.4，可发现通过增大特征图分辨率能够大幅增加YOLO v5在VisDrone数据集上的mAP性能，但是同样会增加计算时间。

本文还根据评估结果绘制了模型训练分辨率分别为640、960和1280时的混淆矩阵，如上图4.4，通过混淆矩阵发现，增大模型训练分辨率可以提高模型对小物体的分类精度。

#### 多尺度训练

表4.4 模型训练epoch数量

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 方法 | 分辨率 | mAP(%) | | 总训练epoch数 |
|  | mAP@.5 | mAP@.5:.95 |
| YOLO v5 | 640 | 32.6 | 18.3 | 200 |
| YOLO+多尺度训练 | 640 | 31.9(↓0.7) | 17.8(↓0.5) | 125(↓75) |
| YOLO v5 | 1280 | 43.5(↑10.9) | 25.9(↑7.6) | 135 |
| YOLO+多尺度训练 | 1280 | 43.9(↑11.3) | 26.0(↑7.7) | 72(↓63) |

本文根据实验结果，绘制了表4.4。其中蓝色表示加入多尺度训练后总的训练epoch数相比于不加多尺度训练方法的模型训练epoch数减少的量。从表4.4中可以看到，多尺度训练并不能稳定提高模型的性能，加入多尺度训练造成了mAP性能1%的波动，但是可以缩短模型训练时间。因此可实验表明，相比于不使用多尺度训练的模型，使用了多尺度训练的模型能够减少的世代（epoch）即减少的训练时间，并且能够达到相近分类和定位精度。

# 实验结论及未来展望

## 实验结论

本篇文章通过实验得到两种优化方法与YOLO v5结合在VisDrone数据集上的性能，证明了提高训练图片分辨率能够提升YOLO v5在Visdrone数据集上的检测精度。在分辨率基础上，1.5倍训练分辨率()能够提升22.4%的mAP@.5()，会增加11.9ms的耗时。因此，如果对于时间不敏感的应用可以引入增大训练图片分辨率的方法来提升分类和定位精度。

通过对多尺度训练的实验，本文得出结论，可以通过多尺度训练的方式，用更少的训练时间达到相近的性能。

## 论文改进方向

### 实验增大训练图像分辨率的上限

本篇文章只验证了增大训练图像分辨率对模型分类和定位精度的提升，没有验证增大训练图像分辨率的上限。随着训练图像分辨率的增加，训练所要求的显存也会增加，因此会造成Batch size的减少，而当Batch size过小时，模型将无法训练。

### 对多尺度训练的优点的定量分析

由于个人数学水平有限，本篇文章没能详细阐述多尺度训练的定量分析，证明多尺度训练对模型训练的加速作用。

## 展望

相信随着硬件技术的进步，现阶段通过时间换取精度的方法能够真正落地应用于工业界。无人机技术也越来越成熟，广泛应用于警务安防、灾害救援和农业植保等领域。深度学习目标检测能够对无人机领域进行赋能，针对无人机拍摄的图片的需要大量人工进行处理的痛点，实现无人机拍摄、自主处理分析、自主决策的全链路闭环。本文相信，无人机领域的目标检测正处于蓝海阶段，有着宽广的应用空间和巨大的能量。

参考文献

1. Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//European Conference on Computer Vision. Springer, Cham, 2014: 740-755.
2. 贾俊立. 基于深度学习的无人机影像目标检测算法研究[D].大连:大连理工大学, 2021.
3. Sommer L W, Schuchert T, Beyerer J. Fast deep vehicle detection in aerial images[C]//2017 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2017: 311-319.
4. Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 25.
5. Deng J. A large-scale hierarchical image database[J]. Proc. of IEEE Computer Vision and Pattern Recognition, 2009, 2009.
6. Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 580-587.
7. Girshick R. Fast r-cnn[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1440-1448.
8. Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in Neural Information Processing Systems, 2015, 28.
9. Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 779-788.
10. Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE Conference onComputer Vision and Pattern Recognition. 2017: 7263-7271.
11. Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.
12. Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.
13. Zhu P, Wen L, Bian X, et al. Vision meets drones: A challenge[J]. arXiv preprint arXiv:1804.07437, 2018.
14. Kuznetsova A, Maleva T, Soloviev V. Detecting apples in orchards using YOLOv3 and YOLOv5 in general and close-up images[C]//International Symposium on Neural Networks. Springer, Cham, 2020: 233-243.
15. Nauata N, Hu H, Zhou G T, et al. Structured label inference for visual understanding[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(5): 1257-1271.
16. Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European Conference on Computer Vision. Springer, Cham, 2016: 21-37.
17. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.
18. Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2117-2125.
19. Wang C Y, Liao H Y M, Wu Y H, et al. CSPNet: A new backbone that can enhance learning capability of CNN[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020: 390-391.
20. He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.
21. Li H, Xiong P, An J, et al. Pyramid attention network for semantic segmentation[J]. arXiv preprint arXiv:1805.10180, 2018.
22. Zhang P, Su W. Statistical inference on recall, precision and average precision under random selection[C]//2012 9th International Conference on Fuzzy Systems and Knowledge Discovery. IEEE, 2012: 1348-1352.
23. 周志华.机器学习[M].南京:清华大学出版社,2016.
24. LeCun Y, Bengio Y, Hinton G. Deep learning[J]. nature, 2015, 521(7553): 436-444.
25. 刘洪江,王懋,刘丽华, 等.基于深度学习的小目标检测综述[J].计算机工程与科学,2021,43(08):1429-1442.
26. 罗会兰,陈鸿坤.基于深度学习的目标检测研究综述[J].电子学报,2020,48(06):1230-1239.
27. 李柯泉,陈燕,刘佳晨, 等.基于深度学习的目标检测算法综述[J].计算机工程,2022,38(05):1-17.
28. 汪亚妮, 汪西莉. 基于注意力和特征融合的遥感图像目标检测模型[J]. 激光与光电子学进展, 2021, 58(2): 0228003.
29. 涂沛驰,傅钰雯,熊宇璇,杨健晟.基于YOLOv5网络模型的火焰检测[J].智能计算机与应用,2022,12(03):158-161.
30. 吕晓君,向伟,刘云鹏.基于强化底层特征的无人机航拍图像小目标检测算法[J].计算机应用研究,2021,38(05):1567-1571.

致 谢

我的这篇毕业论文的完成，首先应当归功于指导老师XXX老师。他无论是在论文开题、资料整理还是在论文的撰写等各个方面都给予了大量的指导和帮助，令我不但完成了论文，也学到了许多书本上学不到的知识，受益匪浅，特致以深深的感谢。

在毕业论文完成的过程中，更是走的很困难。完成了这篇论文之时，最深的感触就是时间紧张，专业知识很是缺乏，在此感谢帮助过我的人，没有他们的帮助，我是不可能顺利完成论文的。特别是感谢指导老师在完成整个毕业论文的过程中，给予我悉心的指导和帮助。老师的严谨的治学态度、勤奋的工作作风、平易近人的处世风范，给我留下了深刻的印象，将会在我今后的学习和生活中时刻影响我。值此论文完稿之际，特此向老师致以衷心的感谢！

本科中，我参与了Robomaster比赛，从其中我受益匪浅。感谢Robomaster团队指导老师XXX老师和各位一起参赛的队友们。因为XXX老师的指导和帮助，我步入了计算机视觉的大门。因为各位队友的帮助，让我有了计算机视觉的工程化经验。

同时，也感谢各位评审老师在百忙之中抽出宝贵时间来审阅我的论文，谢谢各位老师的指导。

在我即将结束毕业论文的写作，即将离开华侨大学的时候，不经意间我发现自己喜欢上了深度学习目标检测的研究方向。我从机械学院转到计算机学院，三年的学习生活对于我，真得是一场艰苦的转型、一次艰难的历练，我从中收获的是一份辛苦的光荣、一段得以受益终生的经历。

感谢所有授我以业的老师。没有老师们的辛苦教学，我是无法学会一门门计算机课程，实现自我价值的探索。同时，我也要感谢原机械学院的辅导员XXX老师，感谢XXX老师四年对我的照顾。

感谢所有与我共享阳光与快乐的同学。实验室里、图书馆里、餐桌前、运动场上、宿舍楼内，无数个朝夕相处的快乐都将是别后梦里的珍藏。愿我们以感恩的心在未来的日子里共同成长。

感谢给予我平静与温暖的家人、爱人和朋友。感谢父母，感谢姨妈姨伯，一直激励着我奋进，感谢老弟，使我时刻有着为人兄长的自我约束。最后，请把四年的经历与收获献给我一生的爱人，谢谢！

谨以此致谢，路还长，唯勤勉前行。

XXX

2022 年 4 月 23 日