对外经济贸易大学

University of International Business and Economics

毕业论文

复杂自动驾驶场景下的目标检测算法——以 OpenMPD 为例

学号 201883009

姓名 刘宇阳

学院 信息学院

专业 数据科学与大数据技术

导师 王树西

时间 2022 年 3 月 20 日

对外经济贸易大学

University of International Business and Economics

Graduation Thesis

Object Detection Algorithm in Complex Autonomous Driving Scenarios

Student ID No. 201883009

Department/School School of Information Technology and Management

Major Field Data science and big data technology

Date March 20, 2022

对外经济贸易大学
本科生毕业论文（设计）开题申请表

学　　号	201883009	姓　　名	刘宇阳
选题方式	自选☑ 指定☐ 其他☐	课题来源	学院☑ 校内☐ 校外☐
题　　目	复杂自动驾驶场景下的目标检测与语义分割—以 OpenMPD 为例		
选题意义	在过去的几十年中，自主驾驶技术在定位、感知、规划、控制等方面取得了巨大的进步。一般情况下，大多数问题都可以用现有的算法解决。然而，一些复杂场景下的定位、感知等问题仍然没有解决。同时，现在主流的自动驾驶数据集均采自国外道路场景，这些已有的著名数据集都记录的是国外的路况而与我国的路况不尽相同，这或多或少制约着中国本土的自动驾驶技术发展。因此，提出一个采于国内的包含光影变化、道路施工、U 形掉头、密集目标、目标遮挡等复杂场景的数据集并使用相关算法进行尝试、调优，对中国本土场景的自动驾驶技术发展，具有重要意义。本人水平和精力有限，因此将主要研究感知领域的目标检测和语义分割算法。由于该数据集将会开源，其他方面的技术可以留待他人研究。		
研究内容	我和我的团队伙伴在前期已经完成了数据采集、清洗和送标流程的工作，初步获取了 OpenMPD（Open Multimodal Perception Dataset）数据集。接下来，我将针对该数据集研究 2D 目标检测和语义分割算法，给出现有方法在该数据集上的基准表现，并在此基础上进行优化调整，针对该复杂场景下的数据集，使用适合的训练和优化方法，以求达到更佳的效果。		
研究基础	我已在清华大学猛狮无人驾驶实验室实习三个月，团队教会了我很多计算机视觉相关的知识，也给了我很多的支持。本文将用到的数据集在我们团队的努力下已经初步采集标注完毕，经过清洗、检验等后续工作即可使用。算力方面，清华大学无偿提供海量 GPU 资源。同时，我上个月入职蔚来汽车感知算法实习生职位，工作内容主要涉及 2D、3D 的目标检测。部门是由中科大博士任少卿带领的自动驾驶部门。任少卿是感知方向的大牛，也是目前全球论文被引用量最多的著名学者。因此我的部门和直属上级也能够给我足够的支持和指导，我有信心利用我周围的人力物力资源和自研数据集完成这篇毕业论文，同时跟我的兴趣、工作方向相互促进，共同进步。		

研究计划	2021 年 11 月到 12 月完成数据集筛选、清洗、检查以及算法实现前期构思和准备工作 2021 年 12 月到 2022 年 1 月完成目标检测和语义分割的模型在 OpenMPD 上的构建 2022 年 1 月到 2 月完成模型调优和论文撰写	
指导教师 审阅意见	同意通过	
	时间: 2021 年 11 月 11 日	指导教师签字: 王树西

毕业论文（设计）诚信声明书

本人声明：我将提交的毕业论文（设计）《 复杂自动驾驶场景下的目标检测算法 》是我在指导教师指导下独立研究、写作的成果，论文中所引用他人的无论以何种方式发布的文字、研究成果，均在论文中加以说明；有关教师、同学和其他人员对本文的写作、修订提出过并为我在论文中加以采纳的意见、建议，均已在我的致谢辞中加以说明并深致谢意。

论文作者_____刘宇阳_____（签字） 时间：2022 年 6 月 4 日

指导教师已阅_____王树西_____（签字） 时间：2022 年 6 月 4 日

毕业论文（设计）版权使用授权书

本毕业论文《 复杂自动驾驶场景下的目标检测和语义分割 》是本人在校期间所完成学业的组成部分，是在对外经济贸易大学教师的指导下完成的，因此，本人特授权对外经济贸易大学可将本毕业论文的全部或部分内容编入有关书籍、数据库保存，可采用复制、印刷、网页制作等方式将论文文本和经过编辑、批注等处理的论文文本提供给读者查阅、参考，可向有关学术部门和国家有关教育主管部门呈送复印件和电子文档。本毕业论文无论做何种处理，必须尊重本人的著作权，署明本人姓名。

论文作者：_刘宇阳_（签字） 时间：2022 年 6 月 4 日

指导教师已阅_____王树西____（签字） 时间：2022 年 6 月 4 日

对外经济贸易大学
毕业论文（设计）指导记录表

<table>
<tr><td>学生姓名</td><td>刘宇阳</td><td>学生学号</td><td>201883009</td></tr>
<tr><td>导师姓名</td><td>王树西</td><td>指导方向</td><td>算法、机器学习</td></tr>
<tr><td>题　　目</td><td colspan="3">复杂自动驾驶场景下的目标检测算法——以 OpenMPD 为例</td></tr>
<tr><td>修改情况</td><td colspan="3">修改要求</td></tr>
<tr><td rowspan="2">1</td><td colspan="3">1. 修改论文格式，按学校要求的论文格式进行编写。
2. 修改图片尺寸，换用更清晰的图片
3. 给出更为详尽的数据和文字描述</td></tr>
<tr><td colspan="2">时间：　2022 年 3 月 7 日</td><td>指导教师签字：王树西</td></tr>
<tr><td rowspan="2">2</td><td colspan="3">1. 将论文中“我”和“我们”统一为“我”
2. 注明工作中哪一部分是团队工作，哪一部分是个人工作
3. 完善实验结果和实验过程描述</td></tr>
<tr><td colspan="2">时间：　2022 年 5 月 11 日</td><td>指导教师签字：王树西</td></tr>
<tr><td>3</td><td colspan="3">1. 修改论文格式，完善目录索引和快捷查找
2. 完善实验结果，丰富实验度量指标
3. 修正部分表述不当的地方
4. 修改英文版学院名称</td></tr>
</table>

	时间： 2022 年 5 月 28 日	指导教师签字：王树西
审阅结果	通过	
	时间：2022 年 6 月 1 日	指导教师签字：王树西

说明：按实际修改情况填写，不足 3 次，可留空白；修改超过 3 次，可另加页。

目录

复杂自动驾驶场景下的目标检测算法——以 OpenMPD 为例

摘要

多模态融合技术促进了自动驾驶的发展，然而复杂环境中的感知仍然是一个具有挑战性的问题。为了解决这个问题，我所在的猛狮团队提出了一个开放多模态感知数据集（OpenMPD），这是一个针对困难场景的多模态感知数据集。与现有数据集相比，OpenMPD 更关注城市地区过度曝光或黑暗、行人和车辆密集、非结构化道路和十字路口等复杂交通场景。它通过一辆配备六个红外摄像头和四个激光雷达的车辆进行数据采集，同步获取普通图片、红外图片和 3D 点云。特别地，我们应用了 128 束激光雷达来采集获得高分辨率点云，以更好地了解 3D 环境并与传感器融合。实验室将所采视频以相等的间隔采样 15K 关键帧用于标注，所用数据集可供 2D/3D 目标检测、 2D 语义分割等相关研究。对于该数据集，我进行了目标检测的相关实验，并在目标检测领域提出了一些普通图片和红外图片进行特征融合对齐的方案来提高检测精度。本文提出的新的方法在保证速度和参数量变化不大的情况下，达到了更高的检测精度。

关键词：多模态融合 深度学习 目标检测

Object Detection Algorithm in Complex Autonomous Driving Scenarios

ABSTRACT

Multi-modal sensor fusion techniques have promoted the development of autonomous driving, while perception in the complex environment remains a challenging problem. In order to tackle the problem, we introduce the Open Multimodal Perception Data set (OpenMPD), a multi-modal perception benchmark targeted at hard examples.Comparing with existing data sets, OpenMPD focuses more on those complex traffic

scenarios in urban areas with overexposure or darkness, dense pedestrians and vehicles, unstructured roads and intersections,etc.It collects data through a vehicle equipped with six infrared cameras and four lidars, and simultaneously acquires ordinary pictures, infrared pictures and 3D point clouds. In particular, we apply a 128-beam lidar to acquire high-resolution point clouds for better understanding of the 3D environment and fusion with sensors. We sample 15K key frames of the collected video at equal intervals for annotation, and the dataset can be used for 2D/3D object detection, 2D semantic segmentation related research. For this dataset, I have conducted relevant experiments on object detection, and proposed some solutions for feature fusion and alignment of ordinary pictures and infrared pictures in the field of object detection to improve detection accuracy. The new method achieves higher detection accuracy (map) while detecting fast equally.

Keywords:Multimodal fusion,Object detection,Semantic segmentation

一、绪论

（一）研究背景

在过去几十年中，自动驾驶技术已经在感知、定位、决策、规划等领域取得了巨大的进步。大多数实际场景中的问题已经能够用现有的算法解决，然而某些复杂的场景仍然难以应对。因此，我所在的团队提出了一个全新的数据集：OpenMPD 来促进相关的研究，并且我个人在该数据集上也进行了相应探索研究。

在此之前，不同的组织已经发布了不同的数据集来促进这一领域的发展。例如，KIT（Karlsruhe Institute of Technology）和 Toyota Technological Institute at Chicago 在 2012 年发布了最早的、最知名的 KITTI 数据集，这一数据集很大程度上促进了自动驾驶技术的多个领域如场景理解，3D 目标检测、3D 追踪等驾驶场景。在那之后，也有更多的全面的数据集问世，促进了自动驾驶技术的发展。谷歌于 2019 年发布了著名的 Waymo 开放数据集，这一数据集同时使用相机、雷达和激光来进行采集，可供学者们研究 2D、3D 等相关问题。

（二）研究意义

显然，优质的数据集能够促进计算机视觉算法的发展，并且算法的性能表现与数据的质量息息相关。同时，目前的多模态融合技术正在蓬勃发展，研究者发现融合热图像、激光点云等额外信息比光使用光学照相机采集图片来做相关视觉任务，效果有明显提升。因此，多模态数据集比普通的图片数据集更有利于促进计算机视觉技术的研究。故而，提出一个多模态、数据充裕的数据集，价值斐然。此外，本文注意到经典的 KITTI、Waymo 等数据集中普遍存在小部分标注错误，这对算法表现一定是有影响的；再者，已有的经典数据集大部分来自海外，这些图片信息均采集的是国外道路，与我国本土的一些道路情况有所不同。因此，开发一个采自本土的开放多模态数据集，对中国的自动驾驶技术有很大帮助。

具体地，对于目标检测，这是自动驾驶中的基础技术之一。人类看一眼图片即可知道图片里的物体是什么、在哪里和如何与其他物体重叠遮挡的。让计算机算法完成这一任务并不是一件容易的事，但却是很有价值的。人类先进的视觉系统让我们能够轻易地感知周围的事物，并只需要付出小部分意识精力便能完成这一任务，例如开车中这一过程。快速准确的目标检测算法将使计算机实时监测周围状况，不局限于自动驾驶，这对无人机、GPS 等领域都有利。本文基于经典目标检测算法 yolov5s，提出一个融合热图像和普通图片的目标检测算法，该算法的表现优于原来的目标检测算法，并同时保证了较高推理速度和较低开销。

总的来说，本文的贡献可以总结为以下三个方面：

1.提出并使用了一个多模态采集、标注的自动驾驶数据集。

2.该数据集包含本土采集获取的弱光、强光、黑暗、重叠、密集等多种复杂

场景，有利于促进学者进行自动驾驶技术相关研究。

3.提出了一个全新的多模态融合算法，精度优于现有的主流目标检测算法。

二、文献综述

（一）国内外研究状况

得益于深度学习的不断发展，算力资源、硬件设备的不断进步，计算机视觉技术取得了飞速的发展。在目标检测领域，传统的机器学习方法很难建模人眼识别物体位置和类别这一行为。随着深度学习的发展，深度神经网络在目标检测、语义分割等任务上都发掘出了不俗的潜力。

2013 年 11 月，Ross Girshick 提出了目标检测的神经网络模型 RCNN。RCNN（Region with CNN feature）是卷积神经网络应用于目标检测问题的一个里程碑的飞跃。CNN 具有良好的特征提取和分类性能，采用 RegionProposal 方法实现目标检测问题。算法可以分为三步（1）候选区域选择。（2）CNN 特征提取。（3）分类与边界回归。区域建议 Region Proposal 是一种传统的区域提取方法，基于启发式的区域提取方法，查看现有的小区域，合并两个最有可能的区域，重复此步骤，直到图像合并为一个区域，最后输出候选区域。然后将根据建议提取的目标图像标准化，作为 CNN 的标准输入可以看作窗口通过滑动获得潜在的目标图像，在 RCNN 中一般 Candidate 选项为 1k~2k 个即可，即可理解为将图片划分成 1k~2k 个网格，之后再对网格进行特征提取或卷积操作，这根据 RCNN 类算法下的分支来决定。然后基于就建议提取的目标图像将其标准化为 CNN 的标准输入。标准卷积神经网络根据输入执行诸如卷积或池化的操作以获得固定维度输出。也就是说，在特征提取之后，特征映射被卷积和汇集以获得输出。最后一步是分类与回归，分类与边界回归：实际上有两个子步骤，一个是对前一步的输出向量进行分类（分类器需要根据特征进行训练）；第二种是通过边界回归框回归（缩写为 bbox）获得精确的区域信息。其目的是准确定位和合并完成分类的预期目标，并避免多重检测。在分类器的选择中有支持向量机 SVM，Softmax 等等；边界回归有 bbox 回归，多任务损失函数边框回归等 。

RCNN 采用的区域提议方式虽然在暴力穷举搜索方法上有所改良，但依然面临着计算量太大，运行效率低的问题。所选取的每一个候选区域都需要通过特征提取网络，重复计算太多。为了解决这一问题，Fast RCNN 和 Faster RCNN 被相继提出。

新的方法摒弃了重复卷积的思路，而是将输入的图片只经过一次卷积，得到一个特征图，然后在该特征图上设置锚框，然后直接将分类器和回归函数应用于提取后的特征图上。这样的方法完全解决了重复计算的弊端，使得目标检测算法

的效率大大提高。得益于 Faster RCNN 的高效和精确度，目前，Faster RCNN 依然是工业界广受欢迎的目标检测算法。

2015 年，YOLO 之父 Joseph Redmon 首次提出单阶段目标检测算法 YOLOv1。在此之前，目标检测算法一般被视为是两阶段模型：特征提取和分类回归。而 YOLO v1 将特征提取和分类回归用一个网络实现，使得目标检测效率有了质的飞跃。在目标检测或者说自动驾驶场景中，检测速度是很重要的指标，因为很小的延迟都可能导致较为严重的事故损失发生。在 YOLO 刚刚提出时，其检测速度明显快于两阶段模型，但由于缺乏候选区域选择这一过程，模型的精度并没有达到当时的 sota。后来的两年，Joseph Redmon 持续改进原有的 YOLO 模型，推出了 YOLOv2 和 YOLOv3。新的方法不管是速度还是精度都已经高于已有的两阶段算法。

遗憾的是，在 YOLOv3 之后，Joseph Redmon 宣布因为一些个人原因将永久退出 CV 界。因此此后再没有他的视觉算法继续问世。

Alexey Bochkovskiy 从 Redmon 手中接过了 YOLO 继续往前发展的大旗。2020 年，YOLOv4 问世，运用了最新的图像增强技术、新颖的激活函数和更大的特征提取网络等方法来提高模型表现。YOLOv4 总体上来看是一个集大成者，集成了当下多个最新的研究成果，但本身并没有非常突出的创新工作。同年，Ultralytics 公司开源了 YOLOv5 模型，在 YOLOv4 的性能上有了进一步的提升。

2021 年，旷视科技发布了 Yolox 模型。这一方法对 baseline、数据增强、参数调节、网络结构、计算效率等各个方面都进行了优化，并且同时发布了三种模型：yolos,yolom,yolol 分别对应小型、中型和大型网络。小型网络属于轻量级网络，适合部署在较小的电器上，而大型模型开销较高但能达到更好的效果。

2022 年，旷视又推出了 Yolof 目标检测算法，这篇文章的主要贡献是指出了特征金字塔结构之所以有效，并不是因为之前学者们认为的多尺度特征融合后的结果，而是将小目标、中等目标和大目标分而治之达到的。

（二）本文研究前景分析

目前主流的 2D 检测算法模型主要利用传统的 RGB 图像进行检测。相关学者经过多年的努力，已经取得了丰硕的研究成果。但存在的不足是，单纯依据 RGB 图片进行目标检测，具有不小的局限性。例如，密集目标、物体之间的遮挡、视角的变化都会影响物体的检测质量。最近几年，结合 3D 点云信息来进行 3D 检测的多模态研究也取得了不小的进展。但在多光谱融合领域，相关的研究较为匮乏，还暂无丰硕的成果和惊艳的模型亮相。因此本文的亮点在于研究方向比较新颖独特，一定程度上弥补了当前环境下这方面的不足。具体地，本文同时使用普通 RGB 图片和对应的红外图片来进行多模态融合，选用 yolov5s 作为基模型，将 RGB 和红外热图像同时输入网络，并对网络结构作了多次修改和实验，最终得出的模型比原有模型的性能表现有了一定提高，优于现有的模型。本文认为，目标之间相互遮挡一直是目标检测领域的难题。遮挡程度越高，目标检测算法的性能受影

响就越大。只使用传统的 RGB 图像很难解决遮挡等问题，但融合了热图像信息，将很大程度上解决这一困难。这是因为，当行人被其他障碍物大程度遮挡时，从 RGB 图像上看将很难发现这一行人。但是如果同时察看热图像和 RGB 图像，则很容易从热图像中发现该行人的位置。行驶中的车辆被障碍物遮挡等情况也是相同的情景。因此，本文认为，合理地融合 RGB 图像信息和热图像信息，将有利于目标检测精度的提高。

三、多模态融合算法

（一）yolov5s 模型结构

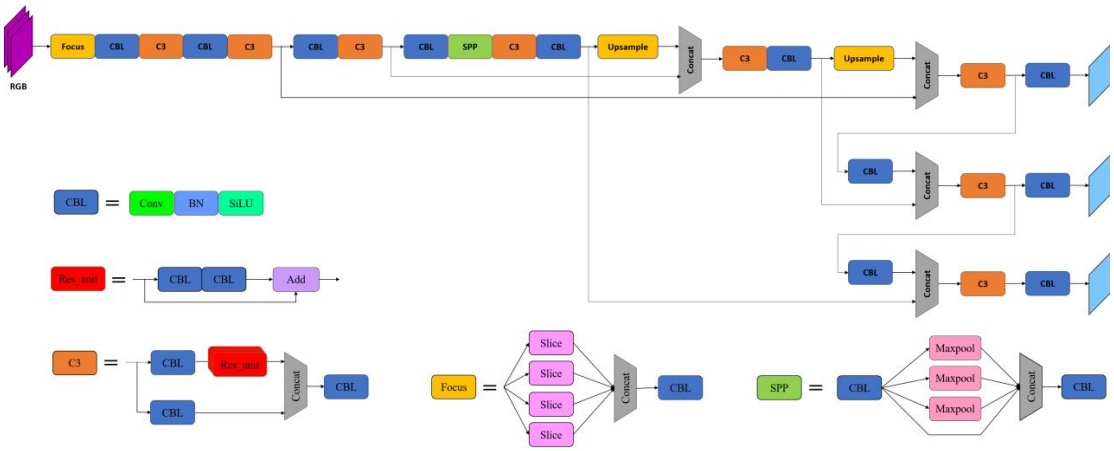

图 3.1 YOLOv5s 模型结构

由于本文工作建立在 yolov5s 之上，因此简要介绍一下原 yolov5s 模型结构。原始图像输入后，首先进行 focus 操作，focus 核心结构如下图：

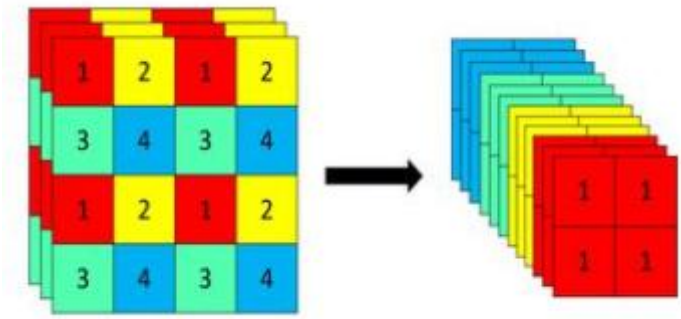

图 3.2 Focus 结构

其原始 608*608*3 的图像输入 Focus 结构，采用切片操作，先变成

304*304*12 的特征图，再经过一次 32 个 1*1*12 的卷积核的卷积操作，最终变成 304*304*32 的特征图。

之后依次经过几组 CBL+C3 结构。CBL 为 Conv+BatchNormalization+SiLU 模块的简称，分别表示卷积层、批标准化和激活函数 silu。C3 结构是一个双路径的特征提取结构，一个路径由一个 CBL 模块组成，另一路径由 2 个 CBL 模块和 1 个 res_unit 模块组成。Res_unit 模块是一个残差结构，在图 3.1 中已给出。将两条路径的计算结果 concat 到一起，得到整个 C3 模块的输出。

在经过四个 CBL 和 3 个 C3 的交替连接之后，特征图通过了一个特征金字塔池化结构 SPP（Spatial Pyramid Pooling）。特征金字塔池化结构如下：

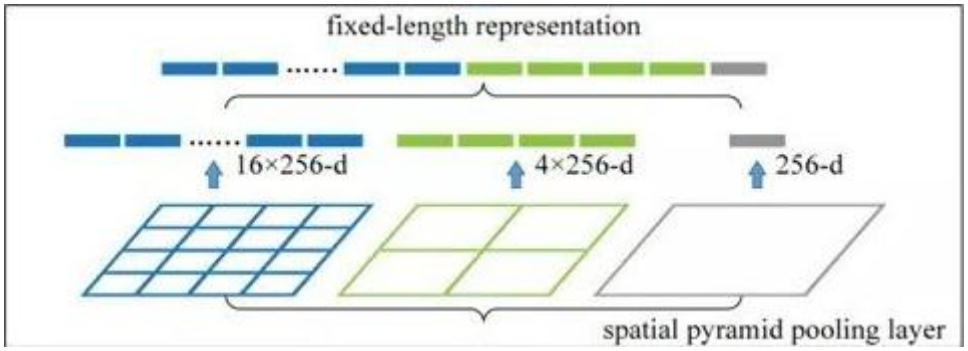

图 3.2 SPP 结构

首先对输入的特征图进行一次 1*1 卷积，将特征图的通道数进行转化得到新的特征图 x。之后对 x 分别使用最大池化，池化层参数分别为 5、9、13，分别得到三个新的特征图 y1,y2,y3，然后将原来的 x 和新的 y1,y2,y3 进行 concat 得到输出 y，最后再通过一个 1*1 的卷积层改变通道数，得到最终的输出。SPP 结构是在 yolov3 时期引入 SPP 结构的，作者 Redmon 应用了当时他人的研究成果到 yolo 之中，取得了良好的效果。SPP 结构增强了不同尺度信息的融合，不同大小的池化层能够提取大中小三种尺寸的物体信息，增强了模型对不同尺度物体的识别能力。

此后的上采样层采用最近邻采样法，将原有的特征图尺寸扩大为原来的两倍，然后依次通过几个 C3 和 CBL 层。每通过一层 C3 和 CBL 结构，特征图的尺寸减小一倍，而通道数增加一倍。Yolov3 到 Yolov5 都采用了多尺度结合的预测方法，在模型结构中，采用了 FPN 的金字塔预测结构，利用不同尺寸、深度的图像信息分别对大中小尺寸的物体进行预测。这样的模型结构充分结合了不同尺度的信息，并在预测的时候也对不同尺寸的物体进行了分治。最近，旷视最新发表的顶会论文 Yolof(You Only Look One-level Feature)详细探讨了 FPN 的作用以及为什么有效。此外，在模型中还多次融合了不同层的输出结果。通过多次的 concat 操作将浅层的信息与深层信息进行融合，也提高了模型的性能表现。总的来看，发展到 yolov5 版本，模型在向着越来越全面的信息提取方向发展。不管是深层浅层信息的融合，还是不同尺度物体信息的融合，都得到了充分的考虑。

以上是特征提取的步骤，下面再来介绍一下锚框的设置和作用。anchor 是在之前的两阶段模型中大放异彩的先进方法，因此 YOLO 也借鉴了过来。模型为经过 backbone 提取的特征图上的每个特征点设置三个不同形状的锚框。同时，在图 3.1 中可以看到，模型实际上是在深度不同、尺寸不同的三个特征图上进行的检测，因此，这三个特征图上的锚框设置的大小也是不同的。一般来说，浅层的特征图尺寸更大，包含更多小物体的信息；而深层的特征图尺寸更小，包含更多大物体的信息。因此，在浅层的大特征图上设置的锚框应该小一些，而在深层的小特征图上设置的锚框应该大一些。在训练阶段，将所有设置好的锚框与标注好的图片上的真实框做好对应，然后采用候选框线性回归的方式调整锚框使得锚框与真实框之间更加接近，模型此时学习对锚框的线性回归变换参数和分类器参数；在推理阶段，则直接使用学习到的线性回归参数对锚框进行线性变换，以便得到最终的预测框以及类别信息。

最后介绍一下 yolov5s 的损失函数，因为后面的模型优化方面也会涉及到其损失函数。Yolov5 的损失函数均由三部分组成：候选框损失、置信度损失和物体分类损失。候选框损失指的是候选框在置信度大于一定阈值时，与真实框之间的线性差值；置信度损失指的是模型预测的候选框置信度与该候选框是否有真实框与之对应的差值。候选框若有真实框与之对应，则标签值为 1,，否则标签值为 0，二者采用交叉熵损失函数进行计算。最后一部分是物体分类损失，表示模型预测的物体类别与物体真实类别之间的损失，这部分就是普通的分类模型的损失，也是使用交叉熵来进行计算。

（二）多模态融合模型改进

为了有效融合信息，我对网络结构和特征融合方式进行了大量的探索。最为首要的问题是应该如何融合 RGB 信息和热图像信息。一个自然的想法是，将 RGB 图像和红外图像 concat 成一张特征图然后输入到原来的 yolov5s 里面。遗憾的是，自然且如此简单的想法并没有得到很好的结果，相反，模型结果变得更差了。因此对模型的特征对齐和融合方式进行了调研和思考。首先我调研了特征对齐的方式，因为红外特征和 RGB 特征需要先进行对齐再融合，否则可能影响模型的判断能力。经过调研当下热门的相关融合方法，我最终选用了 plard 对齐方法并在 plard 方法上进行了改进。plard 特征对齐方法结构图如下：

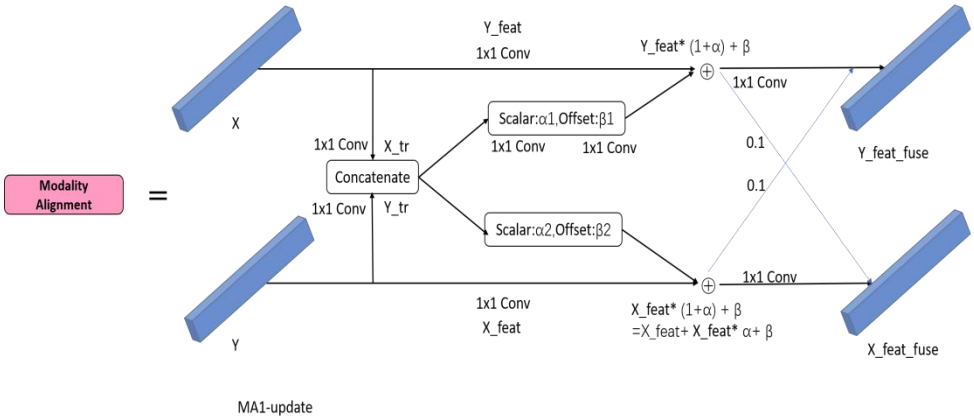

图 3.3 plard 结构

原有的 plard 方法如上图，x 和 y 分别代表 RGB 图像和热图像。为了简单起见，下面我以输入的单个分支 x 为例阐述对 x 的变换。x 和 y 首先分别进行 1*1 卷积改变通道数，然后进行 concat 得到一张特征图 z。此后，对 z 使用两个 1*1 卷积，得到两个新的特征图 alpha 和 beta。与此同时，对 x 使用另一个 1*1 卷积得到一个新的特征图。此后，根据公式 x_feat_fuse=y_feat*(1+alpha)+beta 得到 x_feat_fuse.最后，令 x=x+0.1*x_feat_fuse 得到最终的输出。对于 y 分支，也作同样的操作处理。从 plard 的特征变换来看，是将原始的输入经过一些卷积变换之后，设置一个 0.1 倍的权重，再与初始信息结合。

以上是 plard 特征对齐模块。此外，本文还考虑使用差分模块来对信息进行融合。

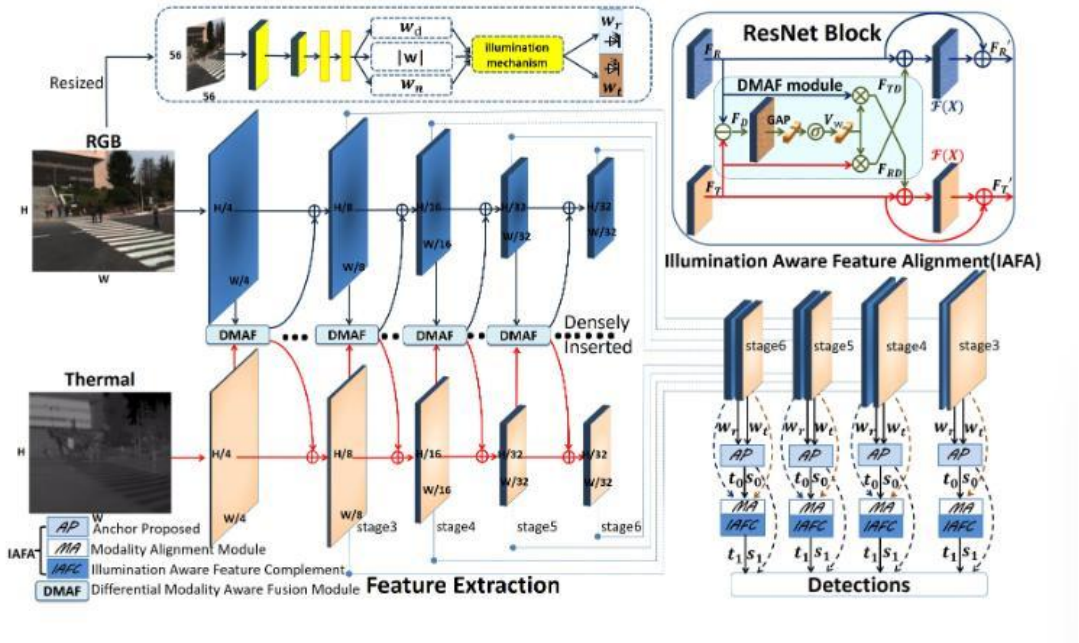

图 3.4 DMAF 结构

融合模块的结构如上图所示，还是以输入的 RGB 和热图像两个分支之一的

RGB 图像 Fr 为例。对输入的 Fr 和 Ft，首先作 Fr-Ft 得到 Fd。然后对 Fd 作全局平均池化并通过激活函数 SiLU 得到向量 Vw。用向量 Vw 对热图像 Ft 作通道注意力机制，然后与原来的 RGB 图像相加，得到中间结果 Ftd+Fr。然后中间结果通过一个函数 f(x)，该函数为 CBL 结构，也就是前文提到的卷积模块。通过 f(x)的结果与初始结果 Fr 相加，得到最终输出计算结果。对于另一热图像分支 Ft，也作相同操作。

有了以上的特征对齐和融合方案，还需要对融合位置进行探索和实验。因此，我设计并实现了如下的模型结构：

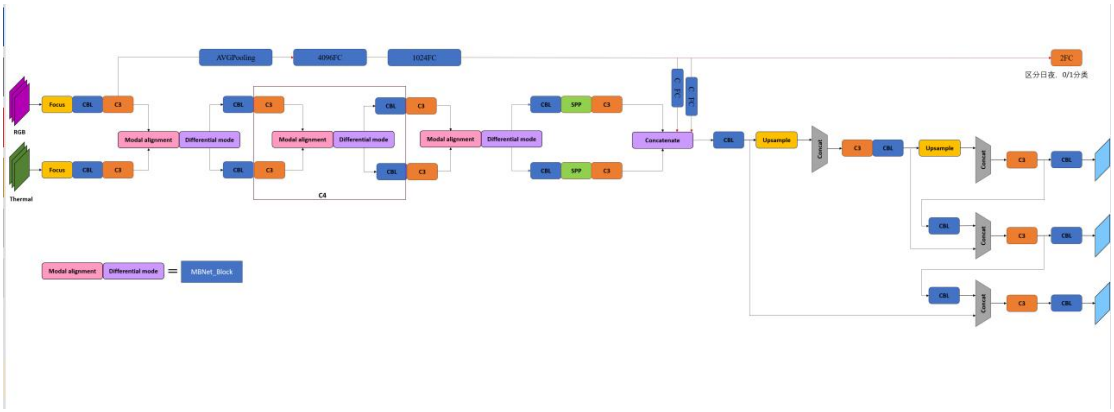

图 3.5 改进的 yolo5s 模型结构

本文初步设计了如上的网络结构，并在此基础上进行了系列实验。如上图，对于初始的网络结构，本文作了如下改动：将 RGB 单通道的结构改为 RGB 和热图像双通道结构，然后分别进行三次特征融合，融合方式均为先使用 plard 进行特征对齐，再使用 DMAF 进行特征融合。此外，因为 RGB 图像可以清晰地指示白天和黑夜，因此本文考虑了上图上半部分所示的任务驱动结构。新建一个分支，对首次经过 C3 结构提取的 RGB 图像进行展平，然后进行平均池化，再通过两个全连接层得到一个向量 m。对这一向量 m，使用一个分类全连接层连接两个输出神经元，表示模型对白天和黑夜的预测。同时对该向量 m 使用再使用两个全连接层得到两个新的输出向量 alpha 和 beta，然后使用 alpha 和 beta 对之前设计的两个路径特征图 concat 之后的融合结果进行通道注入。此后，输入 yolov5s 原有的后续结构，得到最终输出。值得一提的是，加入了任务驱动模块之后，模型的损失函数需要增加一项，即对白天和黑夜的预测，这是一个分类损失，同样使用交叉熵进行计算。值得一提的是，经过实验发现，这一损失的权重需要进行调整，系数为 0.001，因为当系数为 1 时，模型过于关注对白天和黑夜的分类而无法取得较好的目标检测结果。初步试验的结果如下表所示：

表 3.1 实验结果（1）

	MA	DMAF	TASK	P	R	MAP50
Baseline	×	×	×	0.726	0.621	0.696
Model-v1	√	√	√	0.715	0.586	0.642

为了方便表述，上表列举了 MA 和 DMAF、TASK 分支，用√表示模型中使用了该结构，用×表示模型未使用该结构。P 表示精确率，R 表示召回率，而 MAP50 则是目标检测的权威指标，一般用这个指标来衡量目标检测算法的优劣。结果是令人感到遗憾的，多次融合的模型算法各方面都有下降。因此，本文决定围绕 plard 结构作进一步探索。

我注意到，在 plard 方法中，输入的 RGB 图像和热图像是经过 concat 之后再进行的下一步操作。而在 DMAF 中则是采用的相减处理。因此产生了对这一操作的好奇，于是补充进行了如下两个对比实验：其余操作不变，用相加和相减的操作替换原有的 concat 方案，结果如下：

表 3.2 实验结果（2）

	MA	DMAF	TASK	P	R	MAP50
Baseline	×	×	×	0.726	0.621	0.696
Model-v1	√	√（concat）	√	0.715	0.586	0.642
Model-v2	√	√(add)	√	0.729	0.605	0.675
Model-v3	√	√(sub)	√	0.760	0.562	0.631

由实验结果分析可知，相加的效果更佳，但距离最终的目标——高于 baseline 还有很远。此外，为了验证 Task 在模型中的作用，我将任务分支去掉，来测试模型的表现，结果如下：

表 3.3 实验结果（3）

	MA	DMAF	TASK	P	R	MAP50
Baseline	×	×	×	0.726	0.621	0.696
Model-v1	√	√（concat）	√	0.715	0.586	0.642
Model-v2	√	√(add)	√	0.729	0.605	0.675
Model-v3	√	√(sub)	√	0.760	0.562	0.631
Model-v4	√	√（concat）	×	0.751	0.563	0.627

从实验结果来看，Task 分支具有重要作用。因为去掉任务分支之后，模型的表现大幅下降。此外，我对 plard 中生成 alpha 和 beta 的步骤产生了好奇。原来的操作是使用 1*1 卷积来生成 alpha 和 beta，而除了 1*1 之外，还可以使用可变形卷积或者长宽卷积。可变形卷积是指为卷积核学习一个线性参数，使得卷积核不再只作用于原来特征图的固定相邻区域，而是能够作用于其他位置；长宽卷积方案则是对 C*W*H 的图像分别使用 1*W 和 1*H 的卷积核进行卷积，得到相同尺寸的输出。我用这两种方式代替了 plard 方法中的 1*1 卷积，得到了下列实验结果：

表 3.4 实验结果（4）

	MA	DMAF	TASK	P	R	MAP50
Baseline	×	×	×	0.726	0.621	0.696
Model-v1	√	√（concat）	√	0.715	0.586	0.642
Model-v2	√	√(add)	√	0.729	0.605	0.675

Model-v3	√	√ (sub)	√	0.760	0.562	0.631
Model-v4	√	√（concat）	×	0.751	0.563	0.627
Model-v5	√(可变形)	√（concat）	√	0.746	0.595	0.649
Model-v6	√(WH 卷积)	√（concat）	√	0.744	0.606	0.654

从上图来看，可变形卷积和长宽卷积对模型效果都有略微的提升，但是模型效果还是不如 baseline。同时，不管是 WH 卷积还是可变形卷积，计算开销都比较大，导致模型参数激增，推理速度降低。因此，这一方案不可行。

此外我还注意到，plard 方法中的最后一步，是将特征图组合计算出的结果设置以 0.1 的权重，再与原始输入经过 1*1 卷积的结果进行相加。我猜想，在经过一系列变换之后，该计算结果已经包含了丰富的特征信息，而这个 0.1 的权重削弱了这个融合好的特征。因此尝试了下列去掉 0.1 权重的实验：

表 3.5 实验结果（5）

	MA	DMAF	TASK	P	R	MAP50
Baseline	×	×	×	0.726	0.621	0.696
Model-v1	√	√（concat）	√	0.715	0.586	0.642
Model-v2	√	√ (add)	√	0.729	0.605	0.675
Model-v3	√	√ (sub)	√	0.760	0.562	0.631
Model-v4	√	√（concat）	×	0.751	0.563	0.627
Model-v5	√(可变形)	√（concat）	√	0.746	0.595	0.649
Model-v6	√(WH 卷积)	√（concat）	√	0.744	0.606	0.654
Model-v7	√(可变形 v2)	√（concat）	√	0.753	0.581	0.647
Model-v8	√(WH卷积v2)	√（concat）	√	0.763	0.582	0.648

改进 plard 方法之后，模型的表现并没有太大的提升，但是模型的参数量更小了，因此我认为这是一个有效的改进。当然，模型的表现离高于 baseline 的期待还有不小的距离。因此我重新分析了网络结构。我发现相对于 baseline，我一上来就进行了大刀阔斧的改动，把模型改为了三次融合，这样的改进思路太过激进，因此转而思考一次融合的方式。于是设计了如下的网络结构：

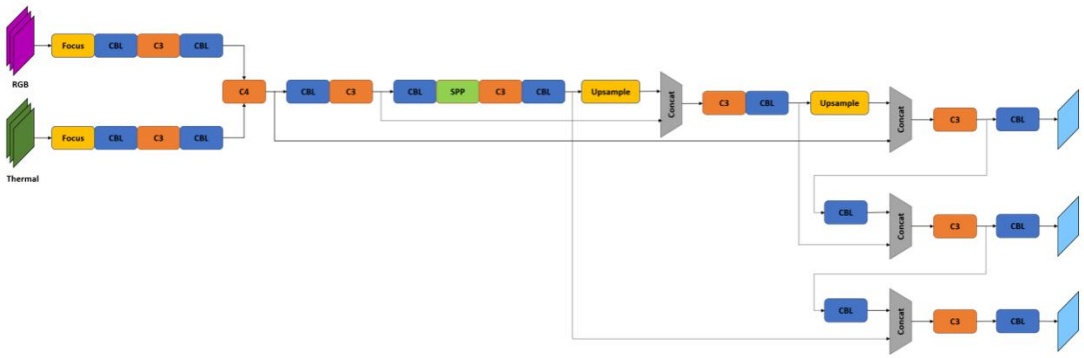

图 3.5 改进的一次融合的 yolo5s 模型结构

其中 C4 结构为结合了 MA、DMAF 的 C3 结构的改进，将两个分支通过 MA 和

DMAF 之后，再通过两层 CBL 模块和一层 Res_unit 模块，然后将两个分支进行 concat，再通过 1*1 卷积降维得到最终输出。为了方便区分，我将新的只融合一次的模型统称为 OM(One_mix)。同时我对任务驱动的有无作了对比实验。另外，两个分支通过 C4 融合的位置也作了相关探究，分别在第一个 C3 和第二个 C3 之后接 C4 融合结构，测试模型效果。另外我还将 C4 的两个分支特征融合结构变为简单的 concat+1*1 卷积，得到如下结果：

表 3.6 实验结果（6）

	MA	DMAF	TASK	P	R	MAP50	GPU 占用	ap
Baseline	×	×	×	0.726	0.621	0.696	5.65G	0.8382
Model-v1	√	√（concat）	√	0.715	0.586	0.642	6.72G	0.8174
Model-v2	√	√(add)	√	0.729	0.605	0.675	7.61G	0.8275
Model-v3	√	√(sub)	√	0.760	0.562	0.631	8.35G	0.8259
Model-v4	√	√（concat）	×	0.751	0.563	0.627	7.27G	0.8325
Model-v5	√(可变形)	√（concat）	√	0.746	0.595	0.649	5.87G	0.8341
Model-v6	√(WH 卷积)	√（concat）	√	0.744	0.606	0.654	4.66G	0.8816
Model-v7	√(可变形 v2)	√（concat）	√	0.753	0.581	0.647	6.14G	0.8182
Model-v8	√(WH卷积v2)	√（concat）	√	0.763	0.582	0.648	8.89G	0.8322
OM-v1	√	√（concat）	×	0.727	0.632	0.684	5.66G	0.8524
OM-v2	√	√（concat）	√	0.722	0.604	0.680	5.86G	0.8545
OM-v3	√	√（concat）	×	0.7	0.606	0.679	6.00G	0.8464
OM_v4	√	√（concat）	×	0.695	0.667	**0.708**	4.98G	0.8682

OM-v1 是不带任务驱动的 C4 融合，OM_v2 是带任务驱动的 C4 融合，OM-v3 是将 C4 结构的融合分支改为简单 concat 融合，而 OM _v4 融合实验为将融合位置改为在第二个 C3 之后进行融合。由实验结果可知：

在单词融合的模型中，任务分支的效果并不显著。

单词融合的模型效果远好于多次融合的模型效果。

将特征融合的位置由第一个 C3 之后改为第二个 C3 之后时，效果得到了较大的提升，且超过了 baseline 模型，达到了比原有所有模型都更佳的效果，比原有方法提高了一个点多一些。因此，这是一个迄今为止最好的方法。

四、总结

（一）结果分析与研究结论

本文首先介绍了一个新颖的多模态数据集，具有数据量丰富、数据种类多样、支持面广、包含多种复杂场景、采自本土等特点的优质数据集 OpenMPD。在该数据集基础上，本文进行了一系列研究与实验。本文首先简要介绍了 YOLOv5s 的模型结构，并使用原有的模型和 OpenMPD 数据集得出 baseline。有了这个 baseline 的结果为基准，对原有模型进行了大量的改进和实验。本文的算法致力于多模态

融合，结合红外光谱图像和 RGB 图像的信息来作目标检测，以期得到更高的检测精度。

具体地，本文首先尝试了直接将红外光谱图像和 RGB 图像进行 concat 并输入原有网络，发现这样的效果并不好。于是本文调研了当前的特征对齐和信息融合的先进方法，最终选择了 plard 方法进行特征对齐，使用 DMAF 差分融合方法进行特征融合。除此之外，本文对网络结构进行了精心的设计。最初本文尝试了多次融合的网络结构，但并没有取得预期的效果。此后，本文尝试了单次融合的模型，发现效果有了显著提升，但仍然不及现有算法的表现。

随后，本文尝试了对信息融合位置的探索，最终发现在第二个 C3 结构之后进行信息融合时，效果有了显著提高，同时模型表现已经优于了现有的方法。在这之后，本文还探究了任务驱动方案的有效性，结果发现在多次融合中表现优异的任务驱动分支在单次融合中效果不明显。

深度学习算法属于黑盒算法，可解释性较差，因此模型的效果在搭建时并不能预见到模型表现的好坏。比如，本文初步设计的三次特征对齐和融合的算法模型，从直觉和理论上来看应该优于单次融合的算法，因为不管从参数量、模型复杂度还是计算量来看都是多次融合更高，但效果却并不符合预期。在模型表现不好的时候，也往往会陷入迷茫，不知道模型表现差的原因在哪里，只能尝试性地去分析，然后做对比实验，逐一排查原因。本文首先对融合次数进行了探究，发现只融合一次的效果确比多次融合效果好；此外，本文还对模型融合的位置进行了探究，发现将融合位置调整到第二个 C3 之后，效果比在第一个 C3 之后融合好很多。

在整个研究过程中，模型的算法在前期总是无法达到期望，对很多想法抱有较高的预期，但算法效果却不尽人意。好在经过不停的探索和对比实验，最终找到了正确的方向和融合策略，使得模型的效果优于已有的所有方法，可谓是柳暗花明，苦尽甘来。

（二）未来研究方向

本文经过系列实验的探究，历经挫折，初步得出了一个优于现有检测算法的多模态融合算法。但客观来说，模型效果的提升幅度还不够大，所以学术价值较为有限。此外，根据目前的研究进展，我还有很多的方向需要去探索，也有很多的研究工作需要在未来跟进。

1.例如，在一次融合的模型中，对于 plard 方法中的输入信息的组合方式是 concat 更好，还是相加或相减更好，还有待继续实验。

2.此外，在新的模型中，我还需要考虑将任务分支加入不同的位置作通道注意力机制，因为之前的实验表面模型融合的位置对模型效果由显著影响，所以可以合理猜测任务分支其实是能够提高模型表现的，只是之前模型的任务分支连接

的位置可能并没有放置在合理的位置。因此，未来需要进一步作对比实验，寻找最优的任务分支插入位置。

3.根据之前的实验结果，模型多次融合的效果劣于单次融合，但理论上如果信息融合有利于特征提取，那么多次融合的效果应该更好。因此，本文有理由怀疑是否是模型融合方案并不是最优的，因此未来还可以进一步探寻更佳的特征融合方案。

4.在当前最优的模型中，我在特征融合的时候，是对 DMAF 模块的输出作 concat+1*1 卷积来将两个分支的信息融为一体。这样的融合方式其实是非常简单的，因此应该进一步考虑使用更为复杂的信息结合方式，比如将两个分支的信息经过多次卷积拼接之后再接 1*1 卷积，来提高模型的性能表现。

5.在YOLO模型的研究和升级过程中，引入了一个性能强大的Mish激活函数，这个激活函数跟传统的激活函数相比，计算代价并不高，但却取得了更好的效果。因此，后续工作可以用 Mish 函数代替部分网络中原有的 SiLU 函数，观察能否达到更好的效果。

以上是目前比较清晰的未来研究方向。所谓未来，并不在遥远的未来，而是在不久的将来。在未来几个月中，我将继续对该模型进行以上探索，以求获得更佳的模型效果，创造更高的学术价值。

参考文献

[1] 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.

[2] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

[3] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[4] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[5] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[6] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[7] 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.

[8] Chen Q, Wang Y, Yang T, et al. You only look one-level feature[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 13039-13048.

[9] Ge Z, Liu S, Wang F, et al. Yolox: Exceeding yolo series in 2021[J]. arXiv preprint arXiv:2107.08430, 2021.

[10] Sun P, Kretzschmar H, Dotiwalla X, et al. Scalability in perception for autonomous driving: Waymo open dataset[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 2446-2454.

[11] Geiger A, Lenz P, Stiller C, et al. Vision meets robotics: The kitti dataset[J]. The International Journal of Robotics Research, 2013, 32(11): 1231-1237.

附录 外文翻译两篇

“你只看一次”:一阶段的，实时的物体检测算法

Joseph Redmon

我们将该模型用一个单阶段的卷积神经网络实现，并在 PASCAL VOC 检测数据集上对其进行评估。网络的初始卷积层从图像中提取特征，全连通层预测输出概率和坐标。我们的网络架构受到用于图像分类的 GoogLeNet 模型的启发。我们的网络有 24 个卷积层,后面是 2 个全连接层。与 googleet 使用的初始模块不同，我们简单地使用了 1 × 1 降采样层和 3 × 3 卷积层。我们还训练了一个 YOLO 的快速版本，旨在推动快速目标检测的边界。Fast YOLO 使用的神经网络具有更少的卷积层(9 层而不是 24 层)。除了网络的规模，YOLO 和 Fast YOLO 的所有训练和测试参数都是一样的。

我们在 ImageNet 1000 类竞赛数据集上对卷积层进行预训练。对于预训练，我们使用 backbone 网络中的前 20 个卷积层，然后是平均池化层和全连接层。我们对这个网络进行了大约一周的训练，并在 ImageNet 2012 验证集上获得了 88%的前 5 名准确率，可以与 Caffe 的 Model Zoo 中的 GoogLeNet 模型不分伯仲。我们使用 darknet 框架进行所有的训练和推理。然后我们转换模型来执行检测。Ren 等人的研究表明，在预训练的网络中同时添加卷积层和连接层可以提高性能。受他们的启发之后，我们添加了四个卷积层和两个随机初始化权值的完全连接层。检测通常需要细粒度的视觉信息，因此我们将网络的输入分辨率从 224 × 224 提高到 448 × 448。我们的最后一层预测了类概率和边界框坐标。我们通过图像的宽度和高度对边界框的宽度和高度进行归一化，使它们落在 0 和 1 之间。我们将边界框 x 和 y 坐标参数化为特定网格单元位置的偏移量，因此它们也被限制在 0 和 1 之间。我们在最后一层使用线性激活函数，而其他所有层使用 leaky ReLU 激活函数。

我们使用 PASCAL VOC 2007 年和 2012 年的训练和验证数据集对该网络进行了大约 135 次的训练。在 2012 年数据的使用中，我们还将 VOC 2007 的测试数据用于训练。在整个训练过程中，我们使用的批处理大小为 64，动量为 0.9，衰减为 0.0005。我们的学习速度计划如下:对于第一个迭代，我们慢慢地将学习速度从 0.001 提高到 0.01。如果我们以较高的学习率开始，我们的模型通常会因为不稳定的梯度而发散。我们继续用 0.01 来训练 75 个 epoch，然后是 0.001 来训

练 30 个 epoch，最后是 0.0001 来训练 30 个 epoch。为了避免过拟合，我们使用了缺失和广泛的数据扩充。第一个连接层之后概率为 0.5 的丢失层阻止了层之间的协同适应。对于数据增强，我们引入了高达原始图像大小 20%的随机缩放和平移。我们还随机调整曝光和饱和度的图像高达 1.5 倍的 HSV 颜色空间。

就像在训练中一样，预测测试图像的标签只需要一次网络评估。在 PASCAL VOC 上，训练好的 YOLOv1 网络预测了每幅图像的 98 个边框以及每个边框的类概率。YOLO 在测试时非常快，因为它只需要一个网络评估，而不像基于分类器的方法。网格设计加强了边界框预测中的空间多样性。通常，一个物体落在哪个网格单元中是很清楚的，而网络只能为每个物体预测一个盒子。然而，一些大型对象或靠近多个单元边界的对象可以被多个单元很好地定位。NMS（非极大性抑制）可用于修复这些多重检测。在 R-CNN 或 DPM 等两阶段模型中，NMS 对性能的作用并不明显，但在 YOLOv1 中，它使得 MAP 增加了 23%。

YOLO 对目标边框的预测施加了很强的空间约束，因为每个网格单元只能预测两个盒子，并且只能有一个类。这种空间限制限制了我们的模型可以预测的附近物体的数量。我们的模型很难处理成群出现的小对象，比如鸟群。因为我们的模型学会了从数据中预测边界框，所以它很难推广到新的或不寻常的纵横比对象。我们的模型也使用相对粗糙的特征来预测边界框，因为我们的架构从输入图像到输出标签有多个下采样层。

目标检测是计算机视觉的核心问题。检测管道通常首先从输入图像中提取一组鲁棒特征。然后，使用分类器来识别特征空间中的对象。这些分类器要么在整个图像上以滑动窗口的方式运行，要么在图像中的某些区域子集上运行。我们将 YOLO 检测系统与几种顶级检测算法进行了比较，突出了关键的异同。

DPM（可变形分离模型）。DMP 使用滑动窗口方法来检测对象。DPM 采用不连接的管道提取静态特征，对区域进行分类，预测高分区域的边界框等。我们的系统用一个卷积神经网络取代了所有这些完全不同的部分。该网络同时进行特征提取、边界框预测、非最大抑制和上下文推理。与静态特性不同，网络内联地训练这些特性，并为检测任务优化它们。我们的统一架构带来了比 DPM 更快、更准确的模型。

R-CNN。R-CNN 及其变体使用区域提议方案，而不是滑动窗口来寻找图像中的对象。选择性搜索生成潜在的边界框，卷积网络提取特征，SVM 对边界框进行评分，线性模型调整边界框，非最大抑制消除重复检测。这个复杂管道的每个阶段都必须进行精确的独立调优，最终的系统非常慢，在测试时间时每张图像需要超过 40 秒的时间。

YOLO 和 R-CNN 有一些相似之处。每个网格单元提出潜在的边界框，并使用卷积特征对这些框进行评分。然而，我们的系统对网格单元提出了空间限制，这有助于减少对同一对象的多次检测。我们的系统也提出了更少的边界框，只有

98 个，相比之下，从选择性搜索大约要产生 2000 个区域提议。最后，我们的系统将这些单独的组件组合成一个单阶段的、联合优化的模型。

“YOLOv3”:一些增量改进

Joseph Redmon

University of International Business and Economics 1
University of International Business and Economics 1
Graduation Thesis 1
Student ID No. 201883009 1
Department/School School of information 1
Major Field Data science and big data technology 1

有时候你会把它搁置一年，你知道吗?我今年没有做很多研究。花了很多时间在推特上。稍微研究了一下 GANs。我去年设法对 YOLO 做了一些改进。但是，老实说，没有什么超级有趣的，只是一些小的改变，让它变得更好。我还帮了别人一点研究学术的忙。事实上，这就是我们今天来这里的原因。我们需要发布一些对 YOLO 的随机更新，但我们没有别的方式。所以准备好接收一份技术报告吧!科技报道最棒的一点就是它们不需要过多介绍，你们都知道我们为什么在这里。因此，本介绍的结尾将为本文的其余部分指明方向。首先我们会告诉你 YOLOv3 的内容和表现。然后告诉你我们是怎么做的。我们还会告诉你一些我们尝试过但没有成功的事情。最后，我们将思考这一切意味着什么。

所以 YOLOv3 的做法是:我们主要是从其他人那里获得好的想法。我们还训练了一个新的分类器网络，它比其他分类器更好。我们会从头开始讲解整个系统，这样你就能理解它了。

YOLOv3 使用逻辑回归预测每个边界框的有无物体的得分。如果之前的边界框与标签框的重叠大于之前的任何其他边界框，则该值应为 1。对于模型预测的有无对象的概率，我们用 0.5 作为阈值。我们的系统只为每个标签框分配一个预测边界框。如果一个边界框之前没有分配给任何标签框，那么它不会导致坐标或类预测的损失，而只计算有无对象的预测损失。

对于类别分类方法，我们使用多标签分类来预测每个边界框可能包含的类。

我们没有使用 softmax，因为我们发现它对良好的性能是没有提升作用的，因此我们简单地使用独立的逻辑分类器。在训练过程中，我们使用二元交叉熵损失进行类预测。当我们移动到更复杂的领域，如开放图像数据集时，这种方法很有帮助。在这个数据集中有许多重叠的标签(如女人和人)。使用 softmax 会假定每个盒子只有一个类，但通常情况并非如此。多标签方法可以更好地为数据建模。

我们使用一个新的网络来进行特征提取。我们的新网络是 YOLOv2 中使用的网络、Darknet-19 和新发明的残余网络的组合方法。我们的网络使用了连续的 3 × 3 和 1 × 1 卷积层，但现在加入了一些 short_cut，而且明显更大。它有 53 个卷积层，所以我们叫它——Darknet-53 ！

训练时，我们仍然训练完整的图像，没有负面挖掘或任何类似的操作。我们使用多尺度训练，大量的数据扩充，批量标准化——这些都是的标准训练方法。我们使用 Darknet 神经网络框架进行训练和测试。

YOLOv3 非常好!在 COCO 数据集上，它的 MAP 指标与 SSD 模型相当，但比 SSD 快 3 倍。

过去，YOLO 一直在为难以检测出小物体而犯难。然而，现在我们看到了这一趋势的逆转。通过新的多尺度预测，我们看到 YOLOv3 对小物体也具有相对较高的性能。然而，在中等和较大的对象上，它的性能相对较差。需要进行更多的调查才能弄清真相和原因。当我们在 MAP50 上绘制精度与速度的关系时，我们可以看到 YOLOv3 比其他探测系统有显著的优势。也就是说，它更快更好。

YOLOv3 是一个很好的探测器。它快速，准确。虽然在 MAP50—MAP95 之间，它没有那么好。但它在 MAP50 的检测标准上性能很好。为什么我们要转换度量?原来的 COCO 论文只有这句话:“一旦评估体系完成后，将会举行一个全面的评估指标的讨论会议”。Russakovsky 等人报告称，人类很难区分 0.3 和 0.5 的 IOU!训练人类用视觉检查一个 IOU 为 0.3 的候选框，并将其与一个 IOU 为 0.5 的候选框区分开来，这是出奇的困难。但也许一个更好的问题是:“我们现在有了这些目标检测器，我们要用它们做什么?”做这项研究的很多人都在谷歌和 Facebook。我想，至少我们知道这项技术掌握得很好，而且肯定不会被用来获取你的个人信息，然后卖给....等等，你是说这就是它的用途?嗯，其他大量资助视觉研究的人是军队，他们应该从来没有做过任何可怕的事情，比如用新技术杀死很多人.....我对大多数使用计算机视觉技术的人抱有很大的希望，他们只是在用它做一些快乐的、有意义的事情，比如在国家公园里数斑马的数量，或者在他们的猫在房子里闲逛的时候跟踪它们。但是，计算机视觉已经被用于可疑的用途，作为研究人员，我们至少有责任考虑我们的工作可能造成的危害，并想办法减轻它。我们欠这个世界太多了。

致谢

行文至此，我的大学生活也即将结束。回顾大学四年的时光，虽然疫情肆虐，却依旧过得丰富多彩，值得铭记。

首先我要感谢我的父母，他们给了我一个无比幸福的童年，教育我为人处世，抚养我茁壮成长。在我大学四年的时光里，他们虽然远在家乡，却依然尽全力给我所有支持，十分感谢父母的辛勤付出和无条件支持。其次，我要感谢大学四年还算努力的自己。那些实习时凌晨被困地铁站的日子，那些凌晨还在计程车上的日子，那些半夜才能回到寝室的日子，那些周末加班一整天的日子，在毕业阶段终于都有了回报。感谢那些年奋力拼搏的自己。

其次，我要感谢培养过我的公司和实验室。感谢诺信创联、滴滴出行、蔚来公司的 leader 和 team 对我算法以及工程能力的培养，感谢清华猛狮实验室对我算法和研究能力的培养。我一直是个并不出众的好学生，大多数能力和算法功底都来自于公司的鞭策和实战学习。另外我还要感谢李兵教授对我编程的启蒙。一切的知识都建立在 Python 这门语言的体系之上，而李兵教授指引我提前学习 python，打下了良好的基础，为我以后的学习工作生活都提供了莫大的帮助。

最后，我要感谢在大学生活中陪伴过我的同学、朋友们。感谢四个搞怪幽默的室友，大学四年有你们陪伴真的多了很多很多的欢乐。感谢可爱的学弟学妹们，很高兴能认识你们并一起走过大学这几年。有你们的陪伴我的大学生活更加丰富多彩，幸福完满，也经历了很多未曾设想的美好旅程。衷心祝愿你们心想事成，实现理想。

姓名：刘宇阳

2022 年 3 月 12 日