涉密论文 □　公开论文 □

浙江大学

本 科 生 毕 业 论 文（设计）

题目　复杂场景下抽烟行为检测技术研究

姓名与学号　曹闵丞，3180100191

指导教师　王总辉

年级与专业　18 级计算机科学与技术

所在学院　计算机科学与技术学院

提交日期　2023/05/21

浙江大学本科生毕业论文（设计）承诺书

1. 本人郑重地承诺所呈交的毕业论文（设计），是在指导教师的指导下严格按照学校和学院有关规定完成的。

2. 本人在毕业论文（设计）中除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得 **浙江大学** 或其他教育机构的学位或证书而使用过的材料。

3. 与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

4. 本人承诺在毕业论文（设计）工作过程中没有伪造数据等行为。

5. 若在本毕业论文（设计）中有侵犯任何方面知识产权的行为，由本人承担相应的法律责任。

6. 本人完全了解 **浙江大学** 有权保留并向有关部门或机构送交本论文（设计）的复印件和磁盘，允许本论文（设计）被查阅和借阅。本人授权 **浙江大学** 可以将本论文（设计）的全部或部分内容编入有关数据库进行检索和传播，可以采用影印、缩印或扫描等复制手段保存、汇编本论文（设计）。

作者签名：　　导师签名：

签字日期：　　年　月　日　　签字日期：　　年　月　日

致 谢

随着毕业论文相关工作的完成，我在浙江大学的本科生涯也即将结束。回顾毕设的完成过程就如同这四年的缩影，让我深刻地回想起自己在本科生涯的自我提升，以及一路上是如何受到他人的支持与帮助。

首先特别感谢王总辉老师在本次毕业论文的每一个环节中，给与我的指导和帮助。在选题到开题的阶段，协助我确定研究方向；在实验阶段，王老师腾出宝贵的时间，关心我的进度以及是否在论文的完成过程中遇到了什么困难，并且就问题上细心的给予我指导；在报告的撰写部分，王老师就论文的篇章结构、内容、格式和文字叙述上，给与了我宝贵的改进建议。使得我论文能够内容完整且合乎规范，因而得以顺利完成毕业论文。

感谢实验室里的所有学姐学长们，在实验的过程中，他们给予了我许多宝贵建议以及指导。帮助我解决在实验理论以及相关工具使用上遇到的问题，帮助我深入理解相关理论以作为模型改进的重要理论依据。其中特别感谢杨勇学长，从毕业论文的开始阶段就给与了我莫大的帮助。帮助我快速确定研究方向与思路以及了解目标检测的背景知识，使得我能够顺利完成数据集标注、厘清检测任务的难点以及目标检测模型的改进工作。并为我的开题报告和毕业论文的书写提供了改进建议，让我最后完成的论文是完整且专业的。

同时，我要向陪伴我的家人和室友们表达衷心的感谢。虽然由于研究领域的不同，无法就毕业论文的学术部分提供太多的帮助。但感谢你们这些年的支持与关怀，特别是毕业论文的完成过程中，陪伴我一同度过这段期间面临的困难与压力，在心灵方面给与了我最大的关心与支持。

最后，我要感谢学校开设毕业论文的课题。经过毕业论文的洗礼，我在学术

能力方面获得了极大的自我提升，比如解决问题的学术思维、文献阅读能力、代码和框架的操作能力、以及与开源社区间的交流方面都得到了显著的提升，特别是与开源社区间的交流，通过学习如何向原作者提问、如何利用社区中的提问解决自己遇到的问题、如何通过社区里面开发者的分享，学习到大量的改进思路以及对计算机视觉领域的前沿研究的心得。这些提升不论是继续深造还是进入职场，都是极其受用的。再次感谢学校通过这个课题，让我能够得到一个这么好的机会提升自己。

摘　要（中文）

在公共场所的吸烟行为会产生未熄灭烟蒂和二手烟，存在影响公共安全与健康的隐患，因此需要严格管控。通过覆盖范围广、能远距离及时通报的基于监控摄像头的自动化抽烟行为检测技术可以取代传统覆盖范围小的人工检测方法。但是由于监控场景下目标较小及场景复杂导致的外在干扰使得复杂场景下的抽烟行为检测工作面临巨大的挑战。本文对现有抽烟检测相关工作以及前沿的单阶段目标检测相关工作进行了调研，并基于调研结果对数据集和 YOLOv8s 模型进行改进。在数据集设计部分，本文香烟结合部分人脸进行标注，降低模型将背景中相似物体预测为香烟的可能性，并通过将香烟复制黏贴在脸部等合理位置，增加模型可学习的样本数。在目标检测模型改进部分，本文通过改进通过改进特征融合网络以及损失函数中的预测框相似度度量，改善当前目标检测模型对于中小型目标的检测能力，并通过自注意力机制提升模型在遮挡、角度变化以及图像干扰下的鲁棒性。实验结果表明，上述改进方法有效改善了复杂场景下抽烟行为的检测效果，并且模型表现优于 YOLOv7、YOLOv8s、DAMO-YOLO、YOLOv8m 这些常见的前沿目标检测模型。

关键词：抽烟行为检测；单阶段目标检测；自注意力机制；特征融合网络；边界框回归损失

Abstract （英文）

Smoking in public places can create hazards such as unextinguished cigarette butts and secondhand smoke, which can affect public safety and health. Therefore, strict control is necessary. Automated smoking behavior detection technology based on surveillance cameras that cover a wide range and can report in real-time from a distance can replace traditional manual detection methods with limited coverage. However, the small size of the target and the complexity of the scene in surveillance scenarios pose significant challenges to smoking behavior detection in complex scenes. This article surveys existing smoking detection-related work and cutting-edge single-stage object detection-related work, and improves the dataset and YOLOv8s model based on the survey results. In the dataset design part, this article annotates cigarettes with some faces to reduce the possibility of the model predicting similar objects in the background as cigarettes. Additionally, by copying and pasting cigarettes in reasonable positions such as on the face, the number of samples that the model can learn is increased. In the object detection model improvement part, this article improves the feature fusion network and the similarity measurement of predicted boxes in the loss function to improve the detection ability of the current object detection model for small and medium-sized targets. Furthermore, the model's robustness is enhanced under occlusion, angle changes, and image interference through self-attention mechanisms. Experimental results show that the above improvement methods effectively improve the detection effect of smoking behavior in complex scenes. Moreover, the model performs better than common cutting-edge object detection models such as YOLOv7, YOLOv8s, DAMO-YOLO, and YOLOv8m.

Key words: smoke behavior detection, Single-stage object detection, Self-attention mechanism, feature fusion network, bounding box regression loss.

目 录

第一部分 毕业论文

第一部分

毕业论文（设计）

1 绪论

1.1 背景

吸烟是导致火灾的主要原因之一，对人们的生命和财产安全构成严重威胁。特别是在公共场所，吸烟存在引发重大火灾事故的潜在风险。此外，公共场所吸烟还会让周围的人吸入二手烟，对他们的肺部健康和整体健康造成危害。因此，在办公区域、工厂、工地等场所，为了上述原因，严禁吸烟。

我国政府已经采取了多项措施来控制吸烟，比如实施公共场所全面禁烟等以减轻和消除抽烟的危害和隐患。这些措施有一定的成效，但仍存在很大的局限性。以往只能通过人工巡逻对这类违规行为进行控管，但是人力巡逻需要耗费大量的人力物力，并且容易被违规者察觉。通过基于摄像头自动化、实时的抽烟行为自动检测算法可以节省大量人力，同时监控更大的范围，并且由于是实时自动化捕捉，可以实时通报。

随着深度学习的快速发展，目前目标检测模型在很多领域都取得了显著成就，例如瑕疵检测、医学影像诊断、无人驾驶等。可以通过目标检测解决传统只能通过人力巡逻公共场合抽烟的问题，提高检测效率。然而与普通的目标检测任务不同，吸烟检测需要对烟头和人脸的存在以及人的姿势进行综合分析，检测出吸烟行为。因此吸烟检测的实现面临以下问题：（1）由于烟头是小型目标，容易被忽略或者将其他类似的物体检测为烟头，在复杂的环境下进行抽烟行为检测容易产生误报以及漏报。（2）监控用的硬件通常是监控摄像头，存在解析度以及取样帧率不足的硬件限制，且监控摄像头一般为俯瞰视角，并且监控的室内场所为复杂场景，待检测的目标可能会出现角度变化大、尺度变化大以及遮挡等问

题。这从数据层面对抽烟行为的检测带来了难度。（3）如果单纯基于姿势检测以检测吸烟，很容易将捂嘴巴等行为误检测为吸烟。

1.2 国内外研究现况

随着深度学习的快速发展，特别是计算机视觉相关技术在很多领域都取得了显著成就，因此研究人员也尝试将相关技术应用于抽烟行为检测领域。例如 Chien 等人[1]利用卷积网络进行特征提取，然后结合时序信息对行为进行分类。Yang 等人[2]提出 Fast TLAM（fast two-level attention model）检测网络，通过对 EdgeBox 生成出的候选区域进行过滤，将包含前景物体的候选区域传递给下一阶段的卷积神经网络（CNN），过滤及分类出仅包含待检测目标的图像块和分类结果，最终使用 K-means 算法聚类，对分类结果归类以完成检测过程。Tang 等人[3]通过 K-means 算法和增加小型目标检测层的方法改进 YOLOv5s 算法对小型目标的检测能力，以达到有效检测画面中香烟的目的。

然而这些方法采用的数据基本为网路爬取的抽烟行为图像，因此并没有充分的考量到实际应用场景中可能会发生的问题，例如监控摄像头的低分辨率导致可利用特征稀少，以及实际场景下物体可能会受到遮挡、角度变化等干扰。因而导致在监控场景下这些抽烟行为检测模型的效果不佳。

1.3 本文研究的意义和目的

现有的工作虽提出了多种目标检测模型，并且在医学图像、自动驾驶等领域都取得了显著成就。但现有的工作在复杂场景下的抽烟检测任务上仍具有一定的局限性。烟头包含的特征信息较少，因此难以提取特征，对定位精度的要求高，并且难以消除复杂场景下周围背景的干扰，导致烟头难以检测。即便结合人脸及

姿势等辅助信息，仍然会因为烟头的特征信息较少而导致模型将普通人脸或易混淆动作（例如捂嘴）判定为抽烟行为。另外，现有的工作对低质量的数据具有较差的泛化性，在低分辨率、受到环境干扰、遮挡以及目标在不同视角和方向的情况下，图像内容会受到干扰，导致可利用特征大幅减少，导致一般的目标检测模型表现不佳等问题[4]。由于上述的原因，一般的目标检测算法在抽烟检测场景下的表现往往不尽如人意。

综上所述，本研究的目的是解决当前目标检测模型在上述场景中存在的问题。通过对现有目标检测模型在抽烟行为检测方面的精度和召回率进行调研和实验分析，本文旨在深入研究这些模型的设计和使用技术与其性能之间的关系。同时，本文将分析模型需要改进的方面以及改进方法，以提升抽烟行为检测的准确性和效果。

1.4 本文主要工作

本文主要工作包含设计和制作自定义的复杂场景下抽烟行为检测数据集，以及对现有的目标检测模型进行改进以更好的解决抽烟行为的检测难点。

数据集方面，通过改进标注策略、数据增强以及负样本设计，更好的避免传统只标注烟的方法会导致训练出的模型具有严重的误检和漏检现象，并且解决了原数据中正样本数量和多样性不足的问题。

模型改进部分，通过改进边界框回归损失函数中交并比的度量方法，降低交并比计算对于尺度的敏感性，以提高边界框回归损失函数对较小型的目标的收敛速度和精度。通过改进模型的特征融合模块，额外融入浅层特征图，从而提供更丰富的图像与定位信息，增加模型对于小型目标或细微特征的检测能力，并通过

引入重参数化卷积提高训练效果。最后在特征融合网络末端加入 Transformer 编码器，可以更好的捕获像素间的长距离依赖关系，通过更聚焦于目标整体而非局部特征，改善遮挡、受干扰图像以及易混淆动作（例如捂嘴）的检测问题。

通过以上改进方法解决抽烟行为检测中对于小型目标（例如烟头）、细微特征、易混淆动作的误检以及目标尺度变化大的检测问题。同时在低分辨率或有干扰的环境下，能够保证模型检测的鲁棒性。

1.5 本文结构与章节安排

本章通过对研究的背景进行概述，提出抽烟行为检测的研究意义与难点，并且总结出研究目的，最后对本文主要工作进行大致介绍。本文后续内容的组织结构如下：第二章将就本文研究中使用的技术背景知识以及重要模型与算法进行介绍，包括现有的抽烟检测工作、目标检测模型的理论基础与结构、YOLOv8 等前沿单阶段目标检测模型的技术特点、数据增强、改进跨尺度融合能力的颈部网络 GFPN 以及 DETR。第三章将对本文工作进行详细介绍，具体包括总体方案设计、数据集设计、颈部网络设计、改进边界框回归损失函数以及 DETR 检测头部的方法设计。第四章将介绍本文研究的实验结果以及结果分析。最后第五章对本文工作进行总结，并分析研究工作的局限、改进空间以及将来进一步研究的展望。

2 相关技术介绍

2.1 目标检测器

目标检测器通常由三个部分组成，第一个部分是基于卷积神经网路的骨干网络，用于图像特征提取。第二个部分是颈部模块，用于将骨干网络在不同阶段提取的不同尺度的特征图进行特征融合。最后一个部分是检测头，用于预测物体的类别和边界框。

常见的骨干网络包括 ResNet[5]、MobileNet[6]、CSPDarknet[7]、Swin Transformer[8]等，这些网络具有强大的特征提取能力，可以将提取的特征信息应用于图像分类、目标检测、语义分割等下游任务。

颈部的设计是为了更好地利用骨干网络提取的特征。对骨干网络在不同阶段提取的特征图进行多尺度特征融合，这样可以更好的融合深层特征图的高级语义信息以及浅层特征图的空间定位信息，提升目标的定位精度，以及对多尺度目标的检测能力。常用的颈部网络有：FPN[9]、PANet[10]、BiFPN[11]，这些方法通过反复使用各种上下采样、张量拼接来设计聚合策略，并且由多个从深层到浅层聚合，然后再从浅层到深层的聚合路径所组成。

检测头部的设计是为了对目标进行分类与定位。头部通常分为单阶段目标检测器和双阶段目标检测器。双阶段检测器先在第一个阶段生成一定数量的目标候选框（region proposal），然后第二个阶段再对各个候选框进行分类，常见的双阶段检测器有 R-CNN 系列[12,13]。而单阶段检测器直接使用预定义的不同比例和长宽比的区块来定位目标，同时对目标进行定位与分类，因此具有明显的速度优势，但精度一般不及双阶段检测器。常见的单阶段检测器有 YOLO 系列[16, 17, 20, 25]、

FCOS [14]和 SSD[15] 。

2.2 YOLO 系列相关前沿工作调研

YOLOv5 因为其易于训练、具有高度可拓展性，并且通过将模型缩放为不同大小，方便不同应用场景的部署,因此受到了广泛的应用。同时对于后续的 YOLO 相关工作具有重要的启发意义，后续的相关工作基本都参考了 YOLOv5 的框架式结构，或是直接基于 YOLOv5 的框架进行修改，并通过整合当下计算机视觉领域的前沿技巧，以达到改进 YOLO 系列模型的目标检测性能的目的。

Tph-YOLOv5[16] 基于 YOLOv5，对其网络结构进行了修改，以提升在无人机空拍场景下的目标检测能力。为了检测微小物体，作者将浅层的 P2 特征图输入到颈部网络中，生成一个额外用于检测微小物体的检测头。由于该特征图具有较为丰富的空间定位信息以及较小的感受野，可以聚焦于小型目标并且提供更精确的空间定位，使得额外的检测头对微小型目标更敏感，进而目标提升小型目标的检测表现。同时作者将 YOLOv5 原始版本中的某些瓶颈模块替换为 Transformer 编码器，利用 Transformer 编码器来捕获全局像素间的长距离依赖关系，提升在密集或遮挡等场景下的检测能力。

YOLOv6[17]采用基于 Anchor-Free 的设计，在特征图上使用网格替代锚框，解决锚框的尺寸是人工预先定义的因而导致与目标之间的尺度不匹配问题，以及过多的冗余锚框导致的正负样本不平衡以及增加非最大值抑制等后处理的开销。同时在骨干网络以及颈部参考 RepVGG[18]使用了重参数化卷积，使用多分支网络进行训练，推理时通过将训练后的参数重参数化为适用于单路网络的等价参数作为单路网络的参数进行推理。这样可以有效结合多分支网络的训练效果，以及

单路网络快速、节省内存的优点。并且使用 Task Aligned Assigner[19]标签匹配策略。通过计算每个实例的锚框（或是网格）的对齐度，使用 Top-K 算法选择每个实例中具有最大值的 k 个锚点或网格作为正样本，并将其他锚点或网格作为负样本，然后通过损失函数进行训练。使用解耦头部将单阶段目标检测模型的检测头中，被耦合的分类与回归的任务进行解耦，解决分类与回归的冲突问题，加快模型收敛速度并且提高检测精度。并且采用了知识蒸馏，利用规模较大、性能较好的模型提供监督讯息训练较为轻量的模型，以增加轻量化模型的性能。最后，YOLOv6 与 YOLOv5 类似，它为不同的工业应用场景，提供了多种不同缩放比例的模型。

YOLOv7[20]则通过 E-ELAN[21]模块通过对最短的最长梯度路径进行控制以及并行更多的网络分支，让更深层的网络可以获取更丰富的梯度信息以有效的学习和收敛。并且通过去除重参数化卷积中的恒等连接（Identity connection），避免恒等连接破坏 ResNet[5]中的残差结构和 DenseNet[22]中的张量拼接，导致模型精度降低。此外还通过深度监督（Deep Supervision）技巧，通过不参与推理过程的辅助头部网络以及通过辅助损失（auxiliary loss）协助训练浅层网络权重，能够提升训练效果并且不会对推理时间造成影响。

DAMO-YOLO[23]则采用 GiraffeDet[24]重量级颈部，轻量主干网络的设计范式。通过神经架构搜索（NAS）技术来搜索最有效的网络架构，以此最大化轻量级主干网络的性能。同时简化 GiraffeDet 的 GFPN 颈部网络结构以减少特征图的冗余复用，更好的符合实时检测的需求。此外 DAMO-YOLO 也引入了 ELAN、重参数化卷积以及知识蒸馏等前沿技巧以改善训练效果和性能表现。

YOLOv8 继承了 YOLOv5 的框架结构并且整合当下计算机视觉领域的前沿

技巧。C2f 模块在原有 C3 模块的基础上，参考 YOLOv7 的 ELAN 模块的思想，通过并行更多的梯度分支，使得梯度信息更为丰富，从而获得更好的精度和训练表现。边界框回归损失函数的部分导入了 DFL[25]（Distribution Focal Loss），通过将边界框的定位问题转为概率分布问题，使用交叉熵函数让标注附近的概率密度尽可能的提高，使得网络可以快速的聚焦目标位置周围区域的分布。检测头部分，YOLOv8 参考 YOLOv6 使用基于 Anchor-Free 和解耦头部的头部设计，并且也使用了 Task Aligned Assigner 标签匹配策略。YOLOv8 通过整合当下计算机视觉领域的前沿技巧，以及通过对网路结构的微调，在实时目标检测器领域取得了 SOTA（state-of-the-art）级别的表现。

综上所述，本文基于以下原因选择 YOLOv8 作为改进目标：（1）YOLOv8 的框架继承了 YOLOv5 易于训练、拓展以及部署等特性；（2）Anchor-Free 设计可以避免锚框的带来的尺度不匹配问题。对于 YOLOv8 本文采取以下方法进行改进：（1）加入 YOLOv7 的重参数化卷积，在基本不增加推理时间的前提下，提升训练时的表现与精度；（2）引入 DAMO-YOLO 的重量级颈部设计，通过改进颈部网络以获得更强的特征融合能力，提升检测精度以及对跨尺度目标的检测能力；（3）参考 TPH-YOLOv5 融合 P2 层特征图来提升对小型目标的定位能力和保留更多图像中的细微特征；（4）引入 Transformer 编码器改善场景中密集对象、目标角度变化、遮挡以及图像受到噪声、模糊等干扰下的检测问题。

2.3 数据增强

数据增强是通过特定策略，对数据集的内容进行变换，来增加数据集的规模以及内容多样性的方法。常用的数据增强方法包含进行几何变换（翻转、裁剪、旋转、缩放等）、色彩空间变换（改变亮度、饱和度、对比度、对数变换、伽马变换等）、高斯模糊以及随机噪声等方法。

YOLOv8 除了使用传统的数据增强方法外，还使用Mosaic以及 MixUp作为数据增强方法。Mosaic 通过将四个图像拼接在一起，增加了训练样本的数据量以及背景多样性，并且通过拼接可以在一张图片上同时计算四张图片的数据，因此可以使用较小的批处理大小实现与较大的批处理大小同等的训练效果，节省内存开销。MixUp 则是通过随机选择两个样本从训练图像中进行随机加权求和，样本的标签也对应加权求和。这样有助于提升模型在噪声样本以及对抗样本下的表现提升，提升模型的泛化能力和鲁棒性。

此外针对小型目标占整体的图像的比例较小以及样本数量较少的问题。Kisantal 等人[26]提出复制增强的方法，通过将小型目标在图像中进行多次合理的复制黏贴，扩充数据集中小型目标样本的数量，提升小型目标检测表现的效果。

本文除了使用 YOLOv8 原有的图像增强方法外，还通过对数据集进行高斯模糊、随机噪声、对数变换、伽马变换。改善模型在复杂场景下画面常见的模糊、图像噪声、暗部细节不足、图像整体亮度过亮或过暗等情况下的检测能力。对于抽烟行为样本较少的问题，本文通过复制增强的方式。在考量了香烟与抽烟行为之间的空间以及尺度相关性后，结合尺寸自适应缩放以及姿势检测，以合理的尺寸复制黏贴在合理的位置上，达到有效扩充数据集中小型目标样本数量的目的。

2.4 Generalized-FPN

目前的目标检测模型主要基于重视骨干的设计范式，并且骨干网路主要是针对图像识别任务而设计，因此强调网络的特征提取能力。但是目标检测任务并不需要那么完整的特征信息，而是需要改进特征融合的能力，以提升语义信息与空间定位信息间的融合以及在目标尺度变化大的场景下的检测能力。

对此，Jiang 等人[24]通过实验证实，颈部在目标检测任务中比骨干网络更重要，并且设计了 GFPN（Generalized-FPN）颈部模块以解决既有的特征融合方法如 FPN、PANet、BiFPN 跨尺度连接不足的问题。

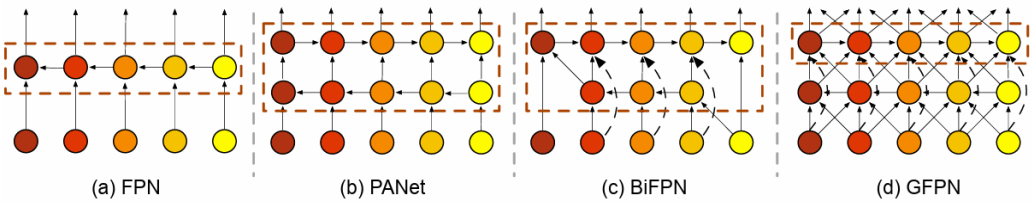

图 2.1 不同的特征金字塔网络设计，GFPN 相较于其他特征金字塔网络，含有更多跳层以及跨尺度链接，可以更充分有效地进行多尺度信息融合[24]

GFPN 包括更多的跳层和跨尺度连接，可以更充分有效地进行多尺度信息融合。在跨尺度连接方面，设计了 Queen-fusion 跨尺度融合模块将同层与邻近较深的一层与较浅一层特征图使用 Queen-fusion 进行特征融合，通过融合更多特征图层，可以更好的克服多尺度的变化。同时作者等人设计了 $\log_2 n$ 跳层连接方法，第 l 层最多只需要接收 $\log_2 l+1$ 的来自先前层的特征图。同时，在反向传播的过程中，$\log_2 n-\text{link}$ 将最短距离从 l 提升到了 $\log_2 l+1$ 。因此 $\log_2 n-\text{link}$ 能够避免 GFPN 在通过堆积网络深度以取得更好的学习效果时，出现梯度消失或梯度爆炸的现象，并确保特征图的复用是有效而非冗余的。

2.5 DETR

传统的目标检测器主要依赖于卷积神经网络（CNN），通过手动设计的 Anchor 机制和非极大值抑制（NMS）来完成目标定位与分类。然而，DETR（DEtection TRansformer）[27]的提出为目标检测领域带来了革命性的变化。DETR 采用基于 Transformer 的编码器-解码器架构，彻底摆脱了对 Anchor 和 NMS 等手动组件的依赖，同时通过引入匈牙利匹配损失，直接预测一对一的目标集，显著简化了目标检测的流程。这种创新设计使得 DETR 具备许多吸引人的特性，但也暴露了一些问题：一方面，DETR 需要长达 500 个训练周期才能达到令人满意的性能；另一方面，其 query 设计较为模糊，未能充分发挥其潜力。

针对这些问题，研究者们提出了一系列基于 DETR 的改进模型，显著提升了其性能和效率：Deformable DETR[28]：通过重新设计注意力模块，仅聚焦于参考点周围的采样点，从而提高了交叉注意力的计算效率。DN-DETR[29]：通过分析发现 DETR 训练收敛缓慢的原因在于早期阶段二分图匹配的不稳定性。为此，引入去噪组技术（denoising group），大幅加速了模型收敛速度。

在上述改进模型的基础上，DINO[30]进一步推动了类 DETR 模型的发展。通过结合 Object365 数据集进行检测预训练，并采用 Swin Transformer 作为主干网络，DINO 在 COCO val2017 数据集上实现了 63.3 AP 的最新（SOTA）性能。DINO 不仅在类 DETR 模型中展现出最快的训练收敛速度和最高精度，还证明类 DETR 检测器能够达到甚至超越经典检测器的表现。

3 复杂场景下抽烟行为检测工作

3.1 总体方案与流程

本文工作的目标是通过构建专门的数据集以及改进 YOLOv8 的网络结构，以提高目标检测模型对复杂场景下抽烟行为的检测能力。

数据集方面，通过数据增强以及负样本设计的方法，改善数据可用特征少，正样本不足的问题，并且提升模型在信号干扰以及低分辨率的场景下的鲁棒性。

网络结构方面，针对现有模型较为不足的小型目标、跨尺度检测以及抗遮挡等干扰的能力。本文通过改进颈部架构，使得模型可以更有效的融合深层与浅层的特征图，以提升跨尺度目标检测能力以及目标的定位精度；通过改进边界框回归损失，改进既有的交并比度量方法不利于较小尺度目标的问题，改善较小尺度目标的边界框的回归过程，进而提升对小型目标的检测能力；还在颈部末端的低分辨率特征图上使用 PSA 注意力替代基於 Transformer 编码器降低训练和推理成本，同时保持和原有 DETR 架构接近的性能表现。並且使用 DETR 檢測頭部替代原有的 YOLO 检测头部 (仅作为辅助分支提供监督信息)，通过自注意力机制对长距离特征依赖关系的建模能力获取丰富的上下文信息，同时能够动态调整感受野，聚焦于目标整体而非局部特征，从而可以在遮挡、图像噪声以及监控视角导致的目标角度变化等会导致可利用特征减少的场景下具有较好的鲁棒性[31]，同时也不需要 NMS 后处理过程，理论上拥有更快的运行速度。

本章先对于总体设计方案进行介绍，然后在本章剩余部分详细介绍不同方案的设计方法，以及不同模块的具体工作方法。

3.2 复杂场景下抽烟行为检测的数据集设计

由于监控场景下的抽烟检测的既有问题，例如抽烟的人比例较少，导致正样本的数量少于负样本、香烟尺寸过小、画面受到噪声干扰、以及不同视角导致的

香烟图像间特征差异较大，可利用特征较少等问题，导致复杂场景下的抽烟行为难以检测。

如果采用现有的检测方案[1-3]只标注香烟的方法，容易因为上述原因发生严重的误检与漏检的情况。误检可以通过与人脸同时检测，在后处理阶段通过判定烟头样本与人脸的交并比来判断是否抽烟，不过漏检的问题无法通过后处理解决。

其他行为检测常用的检测方法，例如姿势检测和时序检测同样无法在抽烟行为检测上取得良好的效果。姿势检测方面，由于监控是以俯瞰视角进行录像、室内的人多半为坐姿，并且身体受到室内摆设以及其他人的遮挡，因此不易采集到完整的人体关键点，并且抽烟姿势呈现多样性，导致姿势与抽烟行为的相关性不高，因此姿势检测并不适用。时序检测方面[1]，由于监控摄像头的更新率不高，以及监控场景需要大规模数据处理等原因。导致更新率不高，进而导致动作的连续性低下，因此时序检测也同样不适用。因此本文直接采取基于抽烟外观的目标检测方法进行抽烟行为检测，而不与其他行为检测方法进行级联。

在标注上，通过结合部分人脸，以及手部等周边图像，作为上下文信息以及约束。由于人脸相较于香烟更是较大的目标并且具有较明显的特征，可以有效提升查全表现。同时由于标注范围有人脸特征，可以作为约束条件避免将背景图像误检为香烟，提升抽烟行为检测的精度。

数据增强方面，本文使用复制增强的方法进行数据增强，通过将一定数量的经过自适应缩放后的香烟图像合成在人嘴部关键点，以增加正样本数量。同时采用常见的高斯模糊、随机噪声、对数变换、伽马变换。以提升在复杂场景下画面常见的模糊、图像噪声、暗部细节不足、图像整体亮度过亮或过暗等现象下的检测能力，同时增加数据的复杂度避免过拟合的情况发生。

负样本设计方面，为了避免本文的标注方法会导致将普通人脸误检为抽烟的情况，需要使用普通人脸作为负样本。但由于原数据的非抽烟者远多于抽烟者（比例约为八比一）因此同时标注人脸会导致正负样本不平衡的问题，导致检测的效果下降。为了使负样本的数量可控，本文参考 COCO [32]数据集中背景图像的方法，使用无标注图像作为背景帮助模型能够有效区分图像的前景与背景。通过在数据集中截取非抽烟者的人脸作为无标注的背景图像。由于模型训练的过程也会去学习无标注的背景图像，因而达到负样本的目的，同时数量可以控制，避免正负样本数量不平衡的问题。

3.3 目标检测模型设计方案

3.3.1 颈部结构设计

（1）GFPN 颈部结构存在的问题

通过使用 GFPN（Generalized-FPN）将同层以及临近的多个尺度的特征图都进行了融合，可以更充分的交换深层语义信息和浅层空间信息[24]。通过 $\log_2 n$ 跳层连接提供了更有效的信息传输，可以将网络拓展的更深，而不用担心梯度消失或梯度爆炸的问题。相较于 PANet 和 BiFPN，GFPN 具有更好的精度表现。然而由于 GFPN 使用 3x3 卷积进行特征融合不论精度表现还是性能开销都无法取得良好的表现。同时 Queen-Fusion 模块包含冗余的上采样操作以及特征图融合，大幅增加了推理时间。此外，GFPN 直接使用全部的骨干网络提取的特征图，进而导致特征融合网路较为复杂。上述原因导致 GFPN 网络具有远高于 PANet 的运算量、延迟以及参数量。直接使用 GFPN 替换 YOLO 系列上常见的 PANet 网络，不利于监控场景的大规模数据处理以及具有实时性的应用场景。

（2）特征融合模块设计

因此本文针对 GFPN 的既有问题进行改进。在特征融合模块中，通过参考 CSPNet 架构，通过将输入分作两个通道维度只有一半的分支，并且只有一个分支会作为主分支通过瓶颈模块[7]，另外一个分支相当于 ResNet 中的跳层直连（Shortcut）。通过这样的设计，相比等价的单路网络，可以减少总计算量并且还可以基于 ResNet 的思想利用重复的梯度信息提升网络的学习效果[5]。而瓶颈模块部分，则通过参考 ELAN 网络结构，通过并行更多的梯度分支，使得梯度信息更为丰富，从而获得更好的精度和训练表现[21]。

然后，参考 RepVGG 引入重参数化思想，使用重参数化卷积替换掉原有的卷积。在训练时采取多分支结构，然后推理时将训练得到的参数重参数化给单路网络结构进行推理。这样可以综合多分支网络能够获取更丰富的梯度信息以提升训练表现和单路网络内存占用少，推理速度快，单路结构较方便剪枝等操作的灵活性的优势[18]。重参数化卷积的结构如图 3.1（a）所示。

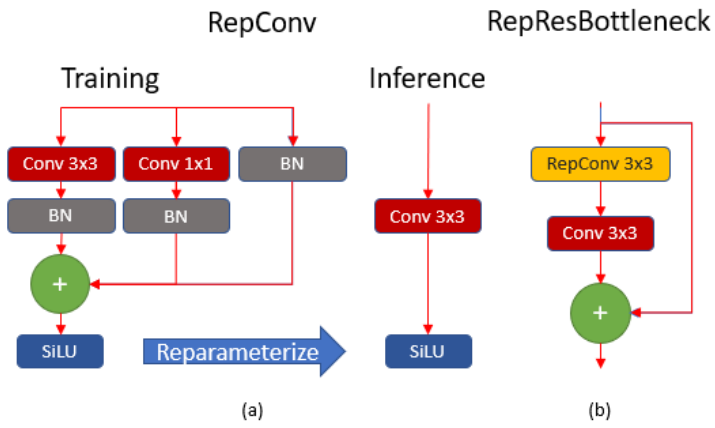

图 3.1 重参数化卷积和重参数化瓶颈模块的结构。a）为重参数化卷积，使用多分支网络训练，重参数化为单路网络进行推理。b）为重参数化瓶颈模块，只有将第一个卷积进行替换，避免残差结构被破坏

而重参数化卷积主要通过三个步骤实现：（1）通过网络优化中常见的 Fuse 操作，将卷积与批量归一化层（Batch Normalization, BN）融合为一个带偏置的卷积，而恒等分支中的 BN 层则设置 1×1 且卷积核为 1 的卷积进行融合；（2）将另外两个分支的 1×1 卷积通过填充 0 的方法转换为等价的 3×3 卷积（3）将三分支的 3×3 卷积中的权重与偏置进行相加运算，合并为只有一个 3×3 卷积且不带有 BN 层的结构。

本文将重参数化卷积用于替换特征融合模块内部的瓶颈模块中的卷积块，同时本文基于 YOLOv7 的研究结果，只取代瓶颈模块中的第一个卷积。由于 YOLOv7 的研究指出具有残差结构或张量拼接操作（Concatenation）的层，重参数化卷积中的恒等连接（Identity Connection）会破坏 ResNet 中的残差结构和 DenseNet 中的张量拼接操作，然而这些网路架构为不同特征图提供了更多梯度的多样性[20]。以残差网络结构为例，残差网路结构中拟合函数的值为输入 x 加上期望输入 $f(x)$，即 $f(x)+x$ 。通过将原先的网路转为残差部分，并将与输出与输入的恒等映射相加，可以改善残差部分在前向传播的过程中，由于经过多层网络结构而出现的信息损失与扭曲，从而有效避免网络退化，进而提升网络训练表现。本文通过只取代瓶颈模块中残差部分的第一个卷积，避免其残差结构被破坏，同时可以利用重参数化卷积提升训练效果。而融入重参数化卷积的 C2f 模块本文命名为 C2fRep。重参数化瓶颈模块的图如图 3.1（b）所示。C2fRep 的结构如图 3.2 所示。

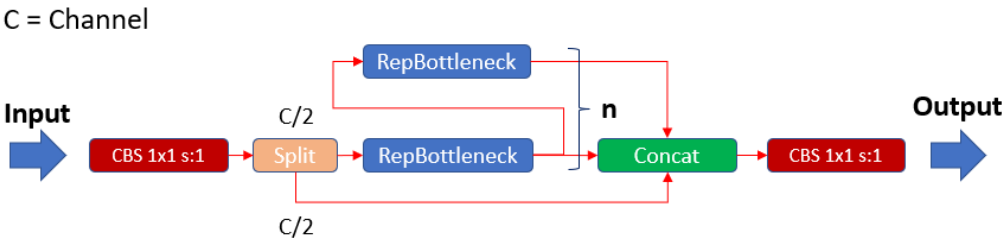

图 3.2 加入重参数化瓶颈模块的 C2fRep 特征融合模块的结构

（3）颈部网络设计

DAMO-YOLO 的论文中通过实验发现多余的上采样操作是导致延迟的主要原因。同时融入过多的特征图层也是导致 GFPN 网络的特征融合过程过于复杂的原因[23]。因此本文参考 YOLOv8 的 PANet 架构，颈部只有直接使用较深层的 P3、P4、P5 特征图去通过特征融合网络分别映射到对应三种尺寸特征图的检测头上。同时由于在 GFPN 中，较深层的特征图以及特征融合模块都会经过上采样操作后，与当前的特征图层进行融合。本文对较深一层的特征图经过上采样操作后直接与当前特征图层进行融合的路径进行删除，只保留经过特征融合后的那一张特征图，减少冗余的上采样操作以及多融合一张特征图导致的计算开销。

此外，通过将 P2 的浅层特征图与小尺度特征检测头所对应的特征图进行融合，利用浅层特征图保留的图像细节信息以及定位信息较多的特点，可以避免小型目标的图像信息在多次卷积下采样的过程中丢失，导致目标难以检测。同时，更精确的定位信息可以帮助模型更精确的定位小型目标，避免边界框的偏移导致小型目标难以检测。此外，浅层特征图具有较小的感受野，可以聚焦于局部细节，避免较大的感受野会受到周围上下文信息影响，过度聚焦于无关的特征上影响检测效果。通过将 GFPN 的跨尺度连接与 $\log_2 n$ 跳层连接设计与 YOLOv8 原有的

PANet 进行结合，提升了模型的目标定位精度，以及跨尺度融合能力，同时相较于原有的 GFPN 大幅降低了计算开销。

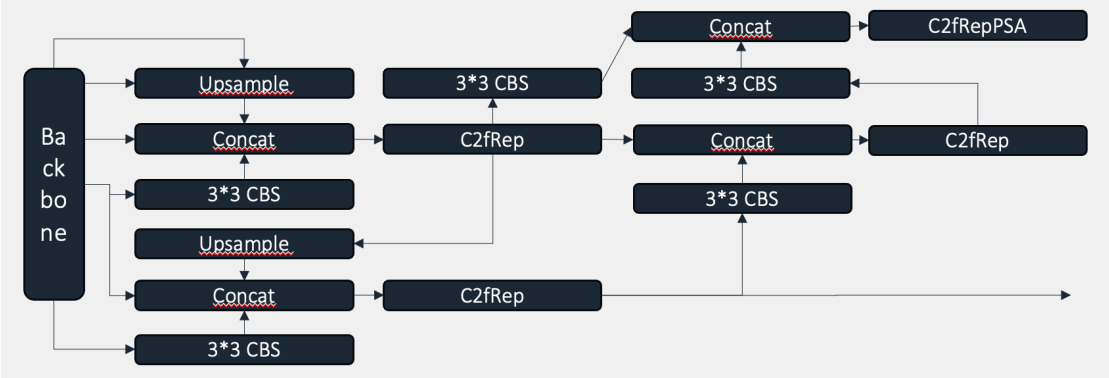

图 3.3 改进后的模型颈部网络架构。

（4）高效编码器

与使用 Transformer 作为编码器的 DETR 相反，本文利用了 YOLO's Neck 的纯卷积架构，该架构在初始阶段进行了预训练，以编码多尺度特征。然后，这些编码的特征被馈送到特征投影模块中，以将它们与隐藏的维度对齐。由于颈部的强大的多尺度特征提取能力，在一开始就通过有效的预训练获得，编码器可以为解码器提供高质量的键，值和建议的边界框。与 DETR 的随机初始化多尺度层和 Transformer 编码器相比，本文的纯卷积结构实现了显着的速度。特征投影公式可概述如下：

$$\begin{aligned} S_1 &= Proj(P_3, P_4, P_5) \\ S_2 &= Concat(S_1) \\ Q &= K = V = S_2 \end{aligned} \tag{3-1}$$

（5）DETR 检测头部

本文采用 YOLOv8 作为模型中的一对多分支，YOLOv8 的三个多尺度层为一对一分支提供了多达 8400 个查询，这些查询可用于生成候选边界框，并作

为解码器的键和值。与 DETR 不同，YOLO 受益于一对多训练方法，使这些查询在第一阶段训练中能够得到更充分的监督。因此，一个强大的颈部模块得以训练，用于向解码器提供多尺度信息，从而使模型实现更优性能。

本文的解码器采用了与 DETR 类似的架构，通过 Transformer 中的自注意力机制捕捉不同查询之间的关系，从而建立分数差异以抑制冗余边界框。在解码器的每一层中，查询会被逐步优化，最终生成一对一的目标集。这种设计大大简化了目标检测过程，并消除了对非极大值抑制（NMS）的依赖。此外，由于 Transformer 解码器提供的全局感知能力，与 DETR 类似，本文则表现出了更强的分类能力，有助于提升对抽烟这类特征不明显，以混淆行为的检测能力。

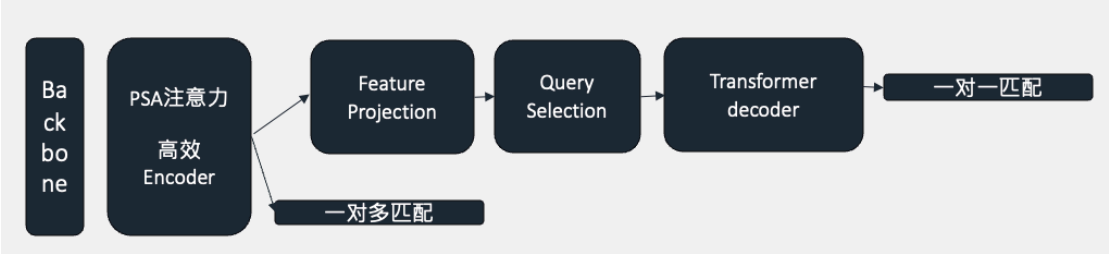

图 3.4 改进后的模型检测头网络架构。

3.3.2 边界框回归损失函数设计方案

YOLOv8 采用的边界框回归损失函数为 CIoU[35]。CIoU 除了考虑预测框与真实框之间的重叠率外，还综合考量了中心点距离以及长宽比一致性，通过考量中心点距离，将两个目标框之间的距离最小化，提高收敛速度；通过考量长宽比一致性解决了多个候选预测框中心点重合的时候难以选出最优解的问题。提升了边界框回归过程的稳定度以及收敛的精度，其计算式如下列公式所示。

$$CIoU = IoU - \frac{\rho^2(b,b^{gt})}{C^2} \tag{3-2}$$

$$\alpha = \frac{v}{(1-IoU)+v} \tag{3-3}$$

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

中心点距离部分，ρ 代表预测框 b 与真实框 b^{gt} 中心点之间的欧氏距离，c 则代表两框交集区域的对角线长度。长宽比一致性部分，v 是作为两框之间长宽比一致性的参数，α 则是作为 v 的权重系数。

但是计算重叠率所使用的交并比对于小型目标的边界框的位置偏移敏感。对此，Wang 等人[36]提出了将归一化 Wasserstein 距离用于边界框回归的损失函数上的方法。基于中心点定位的度量标准更适用于小型目标检测，因此作者使用二维高斯分布对边界框进行建模，中心点的权重最高，权重在从中心往边缘的过程中逐渐降低。将度量标准则从边界框之间的重叠程度转为两个高斯分布之间的分布距离，计算两个分布间的归一化 Wasserstein 距离作为真实分布与预测分布间的相似性度量标准。通过归一化 Wasserstein 距离，可以更好度量小型目标预测框与真实框之间的相似度，从而提升对小型目标的检测能力，并且该损失函数是连续而非离散的，可以更好的指导网络的训练。但是 Li 等人[27]发现对于一般尺度的目标，基于交并比的度量方法表现仍优于基于中心点定位的分布距离度量方法。因此本文通过实验，发现将 0.8 的比例保持原有的交并比度量，剩余的 0.2 则使用归一化 Wasserstein 距离度量，可以在抽烟行为检测任务上取得较好的表现。归一化 Wasserstein 距离的算法如下列算式所示。

$$NWD(\mathcal{N}_a, \mathcal{N}_b) = \exp\left(-\frac{\sqrt{W_2^2(\mathcal{N}_a, \mathcal{N}_b)}}{C}\right) \tag{3-5}$$

$$W_2^2(N_a, N_b) = \left(\left[cx_a, cy_a, \frac{w_a}{2}, \frac{h_a}{2}\right]^{\mathrm{T}}, \left[cx_b, cy_b, \frac{w_b}{2}, \frac{h_b}{2}\right]^{\mathrm{T}}\right) \tag{3-5}$$

(cx, cy, w, h) 分别代表边界框的中心点横坐标、纵坐标、宽度以及长度，c 是

常数默认值为 2.5，$\mathcal{N}_a$，$\mathcal{N}_b$ 是通过两个边界框的 (cx, cy, w, h) 来建模的高斯分布。$W_2^2(\mathcal{N}_a, \mathcal{N}_b)$ 为两个高斯分布间的分布距离度量。

本文最终采用的新的边界框相似性度量如下列公式所示。

$$IoU_{NWD} = 0.8\,(1 - IoU) + 0.2(1 - NWD(\mathcal{N}_a, \mathcal{N}_b)) \tag{3-6}$$

4 实验结果与分析

4.1 数据集制作与预处理

（1）数据集数据分布

通过将原始数据中过多的不包含抽烟行为的图像进行删除，确保包含抽烟行为的图像数量与未包含抽烟行为的图像数量相同。抽烟行为数据集总计有 8680 张图片，由真实场景的监控图像与网络爬取的抽烟行为图像组成。其中真实场景图像共计 7388 张，网路爬取的抽烟行为图像则共计 1292 张。将数据随机分为训练集共计 6996 张图片以及训练集共计 1684 张图片，并且为了确保测试结果是符合真实场景的，本文只将网络爬取的图像加入到训练集中，确保测试集包含的都是来自真实场景的图像。然后本文对数据分别采用只标注香烟的标注方法以及将香烟融合周围特征如部分脸部与手部的标注方法。数据分布如表 4.1 所示。

表 4.1 数据集的数据分布情况

数据集类型	数据总数	包含抽烟行为的图片总数	抽烟标注总数	占总体数据的百分比
训练集	6996	3498	3871	80.6%
测试集	1684	842	858	19.4%

（2）数据集标注策略

通过实验得知，采用只标注香烟的方法，无论精确率还是召回率的表现都远不及将香烟融合周围特征如部分脸部与手部的标注方法。虽然误检可以通过与人脸同时检测，在后处理阶段通过判定烟头样本与人脸的交并比来判断是否抽烟，不过召回率低的问题没有办法解决，因此可以判定标注烟的标注方法不适用于复杂场景下的抽烟行为检测。使用 YOLOv8s 对标注策略改进前后的数据集进行测试的结果如表 4.2 所示。

表 4.2 修改标注策略前后，在 YOLOv8s 上检测表现的对比

标注方法	精确度	召回率	mAP_{50}	mAP_{95}
只标注烟	47.3	49.8	46.3	13.5
烟 + 周围特征	70.5	67.2	71.0	41.7

图 4.1 只标注烟的方法在真实场景下除了容易将背景中不相关的物体误检为香烟外，还容易漏检真实场景下特征不明显的香烟样本

通过实验结果，本文最终决定使用香烟融合周围特征的标注策略。

在标注类别上，本文只标注抽烟行为一个类。通过预训练的人脸检测模型，本文发现以训练集为例。总计有 24276 个人脸样本（平均每张 3.47 张人脸）远

多于抽烟行为样本的 3817 个。如果使用人脸作为抽烟行为的负样本，会导致正负样本不平衡的问题。抽烟行为样本受到过多人脸样本的约束，而导致检测不到抽烟人脸。

（3）数据集标注工具与标注格式

使用标记软件 LabelImg 以 YOLO 格式进行数据集标注。标注文件后缀名为 txt，文件名称与对应图片名称相同。文件内容共计五列，分别为类别（以数字代表，从类别 0 开始）以及以相对于图像大小的比值进行记录的标注框中心点的横坐标、中心点的纵坐标、标注框的宽度、标注框的高度。

（4）数据增强

同时通过对数据集进行数据增强的方法，进一步提升模型检测的能力以及泛化能力。数据增强包含三个部分。首先是采用常见的高斯模糊、随机噪声、对数变换、伽马变换。以改善在复杂场景下画面常见的模糊、图像噪声、暗部细节不足、图像整体亮度过亮或过暗等情况下的检测能力。其次是对于抽烟行为样本不足的情况，采用复制增强的方法进行数据增强。考虑到香烟与人物具有位置相关性，同时在监控场景下脸部相对容易找到关键点。因此先通过 BlazePose[37]进行简单的单人姿势识别找到嘴部或手部关键点。并基于两肩之间的距离约为香烟长度的三倍，并且考量到监控摄像头的视角问题，本文将香烟的图像长度等比例缩放为两肩关键点的 0.5 倍。香烟的尺寸自适应公式如下列公式所示：

$$L_{ciga} = \sqrt{(P_{x\ of\ left\ shoulder} - P_{x\ of\ right\ shoulder})^2 + (P_{y\ of\ left\ shoulder} - P_{y\ of\ right\ shoulder})^2} \qquad (4.1)$$

L_{ciga} 代表香烟长度，P 代表关键点坐标的横坐标值或纵坐标值。

将香烟自适应尺寸缩放后，本文可以将香烟合成在人物的嘴部关键点。通过综合考量香烟与人物的位置相关性以及对香烟的尺寸进行自适应缩放，复制增强可以更接近真实场景，进而取得更好的效果。本文选择训练集中的 229 张图片进行复制增强。复制增强的效果如图 4.2 所示。

图 4.2 复制增强的效果图

（5）负样本设计

在负样本设计的部分。直接使用不包含抽烟行为的未标注图像作为背景图片来作为约束，会因为图像中无关的背景信息过多而导致约束能力不足。通过标注普通人脸作为负样本来进行约束，则会因为普通人脸样本过多而导致正负样本不平衡，导致检测不到抽烟人脸。于是本文选择折衷的方法，截取非抽烟者的人脸作为无标注的背景图片。由于模型训练的过程也会去学习无标注的背景图片，因而达到负样本的目的，同时数量可以控制，避免正负样本不平衡的问题。通过将1000 张不包含抽烟行为的图片，替换为人脸图像，在基本不影响召回率的情况下，精确率获得大幅的提高。分别使用增强前和增强后的训练集对 YOLOv8s 进行训练，然后对测试集进行测试的结果如表 4.3 所示。

表 4.3 训练集经过增强前后，使用 YOLOv8s，在原有测试集上的表现

数据是否增强	精确度	召回率	mAP_{50}	mAP_{95}
增强前	0.705	0.672	0.71	0.417
增强后	0.751	0.666	0.735	0.428

（5）数据集样本分布分析

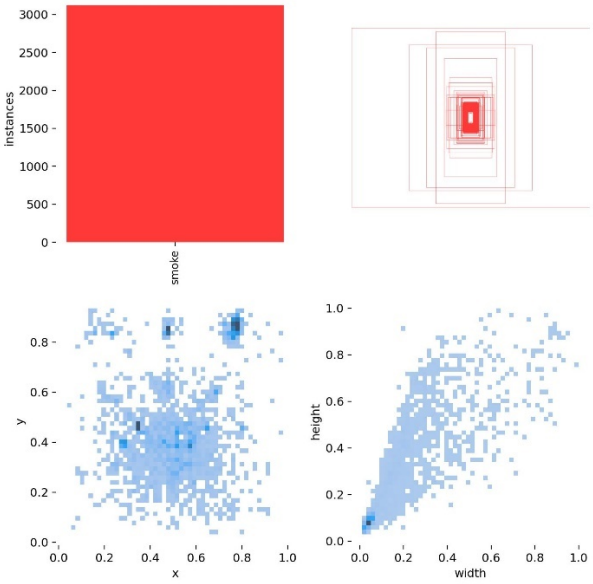

图 4.3 数据集样本可视化结果，左上角为标签类别分布；右上角为标注框的形状分布；左下角为标注样本的归一化空间分布；右下角为标注样本与图像间相对长宽比例的分布

样本分布的可视化结果可以看出，由于本文标注的主要为包含烟的局部人脸，从标注框形状分布可以看出形状基本相符，同时可以从较为密集的分布看出，大部分标注样本的尺寸较小。从样本空间分布图可以看出样本基本分布在图像中心的大区域，以及在图像中多处呈现点状分布，这是因为本文的数据集是由真实监控场景以及网路爬取的抽烟图像所组成，高密度的点状分布是因为室内监控监控场景中人物的分布受到室内布局的限制，因此抽烟样本会高度集中于室内有位置的地方。而中心处的大范围分布则是因为网路爬取图像大部分为类似肖像的形式，导致抽烟样本会呈现在中心处的大范围分布。从标注样本与图像间相对长宽比例的分布可以看出，虽然因为增加了网路爬取图片的关系，使得样本相对长宽占比的分布有所增广，但是多数样本仍然为小型目标，根据训练时会将图像缩放到 640×640 的尺寸来计算，标注样本的平均大小为 42×75，接近 COCO 数据集对于小型目标的定义，即尺寸小于 32×32 的物体。

4.2 模型评价指标

训练后的模型需要性能指标来衡量模型在任务中的表现，并且性能指标可以客观和全面地将多个模型在特定的基准下进行比较，提供分析模型综合表现以及改进的依据。对目标检测模型的性能表现，通常使用以下指标：

交并比（Intersection over Union，IoU）。衡量预测框和真实框之间相似度的指标，定义为预测框和真实框之间的交集面积和并集面积的比值，交并比的最大值为1，代表两个框之间完全重合。

除了用于作为预测框和真实框之间的相似度度量之外，交并比还被用来作为目标检测时的阈值。如果样本的预测框与真实框之间的交并比大于阈值，则判断为真阳性（TP），代表预测正确；如果样本的预测框与真实框之间的交并比低于阈值甚至没有交集，则判断为假阳性（FP），代表预测错误。

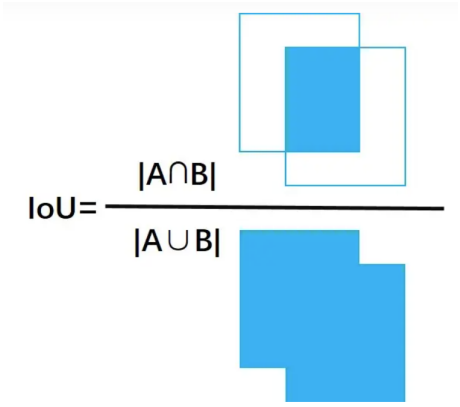

图 4. 4 交并比的定义

混淆矩阵（Confusion Matrix）。通过综合考虑假阳性（误检）以及假阴性（漏检）的情况，可以避免准确率（Accuracy）的计算只考虑样本总数，而导致数据集在被正样本或负样本主导的情况下，精度指标虚高或是虚低的情况。以混肴矩阵为基础还可以延伸出精确率（Precision）、召回率（Recall）、F1 分数等性能指标，以及提供不同分类之间的混淆情况。方便本文对模型的精度、查全率、精准度和

召回率之间的平衡程度以及检测模式有更全面的理解。混淆矩阵的结构如表 4.4 所示。

表 4.4 混淆矩阵的结构

真阳性（True Positive, TP） 图像：阳性 预测值：阳性 正确预测	**假阳性（False Positive, FP）** 图像：阴性 预测值：阳性 错误预测（误检）
假阴性（False Negative, FN） 图像：阳性 预测值：阴性 错误预测（漏检）	**真阴性（True Negative, TN）** 图像：阴性 预测值：阴性 正确预测

准确率（Accuracy）。其计算公式如下列公式所示：

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{4.2}$$

精确率 (Precision)。其计算公式如下列公式所示：

$$P = \frac{TP}{TP+FP} \tag{4.3}$$

召回率 (Recall)。其计算公式如下列公式所示：

$$R = \frac{TP}{TP+FN} \tag{4.4}$$

F1分数，其计算公式如下列公式所示：

$$F1 = \frac{2PR}{P+R} \tag{4.5}$$

平均精确度(Average Precision, AP)。为了更全面的对模型的精准度和召回率之间的平衡程度进行评估。Dalal 等人[38]提出 AP 指标。通过不同置信度阈值下

的精确度和召回率组合，以精确率为纵坐标召回率为横坐标，绘制一条 P-R（Precision-Recall）曲线。对不同召回率 r 相对应的精确度 $p(r)$ 求积分，求得曲线与坐标轴之间的面积作为 AP 的值。并且对每个分类的 AP 求平均，就是均值平均精确度 (mean Average Precision，mAP)，AP 的计算公式如下列公式所示。

$$AP = \int_0^1 P(r)dr \tag{4.6}$$

mAP 还可以按照特定的交并比区间来计算，比如 mAP_{50} 是指在样本的交并比大于0.5的条件下的 mAP，mAP_{50-95} 则代表 mAP_{50}、mAP_{55}、mAP_{60}、到 mAP_{95}（从0.5开始以0.05为步长增加）的平均值。

4.3 目标检测模型训练与结果分析

4.3.1 实验设置

实验环境方面，本文使用 PyTorch 实现，Python 版本为 3.8.13，Pytorch 版本为 2.0.0 并且在 A800 上进行训练。

训练超参数设置方面，最大迭代次数设置为 150 个 epoch，并且设置前 3 个 epoch 用于热身。批处理大小为 32。使用 SGD 优化器进行训练，初始学习率设置为0.01，Momentum 设置为 0.937，最后一个epoch的学习率衰减为初始学习率的 0.01。

预训练权重部分，本文使用来自官方的 YOLOv8s 的预训练模型作为预训练权重进行迁移学习，因为修改后的 YOLOv8 和 YOLOv8s 的骨干网络结构相同。可以从 YOLOv8s 转移部分权重到修改后的 YOLOv8 上。由于预训练模型使用了大规模的数据集，例如 ImageNet[39]上进行训练，对于通用的特征具有良好的

学习能力，可以迁移到不同的数据上进行训练，并且提供模型训练一个较好的初始点，而非随机初始化。因此通过在部分层加载预训练权重可以加速收敛过程并且提升训练效果，特别是数据集样本数较少或者困难样本较多的情况。

4.3.2 模型训练过程

本次实验基于上节提到的实验环境与超参数配置对原先的YOLOv8s模型与改进后的YOLOv8模型进行对比。评价指标为置信度阈值为0.25下的精确率、召回率以及 mAP_{50} 和 mAP_{95}。图 4.5 、图 4.6、图 4.7、图 4.8 分别为改进前后的模型之间训练 Box Loss、验证 Box Loss、 mAP_{50} 以及 mAP_{95} 随迭代次数的变化图。纵坐标为评价指标的数值，横坐标为当前迭代次数。绿色折线代表改进后模型的结果，蓝色折线代表原先YOLOv8s模型的结果。如图所示，随着迭代次数的增加， mAP_{50} 和 mAP_{95} 的数值逐步升高，同时 Loss 数值也在不断降低，并且在训练后期趋于平缓，代表模型可以收敛并且已经完成了对数据的拟合。 从图中的曲线变化可以看出，改进后的模型在训练过程中具有更高的 mAP_{50} 和 mAP_{95}以及较低的训练和验证 Box Loss，Box Loss 表示的是模型的边界框坐标定位的误差，越低表示模型的边界框定位的越精确。改进后的模型 Loss略低于原先的模型，代表改进后的模型具有更精确的定位能力，这与本文改进模型的方向相符。而随着模型迭代次数的增加，验证 Loss 整体呈现先下降后上升，最终收敛的趋势，原因可能为过拟合或是训练集与测试集之间数据分布不一致，导致训练后期模型拟合大部分样本后，loss 的计算被测试集中与训练集内容差异较大的样本所主导，导致 loss 的值重新回升。不过这些少数的困难样本对于验证过程中整体的精度计算影响有限，甚至会出现 Loss 上升，精度表现反而提升的情况[40]。基于上述原因，为了选取泛化能力最强的模型权重，YOLOv8 通过式 4.7 作为评价指标，选取 mAP 表现最好的那一轮迭代的模型权重最为最佳模型权重。

$$fitness = 0.9 \times mAP_{95} + 0.1 \times mAP_{50} \tag{4.7}$$

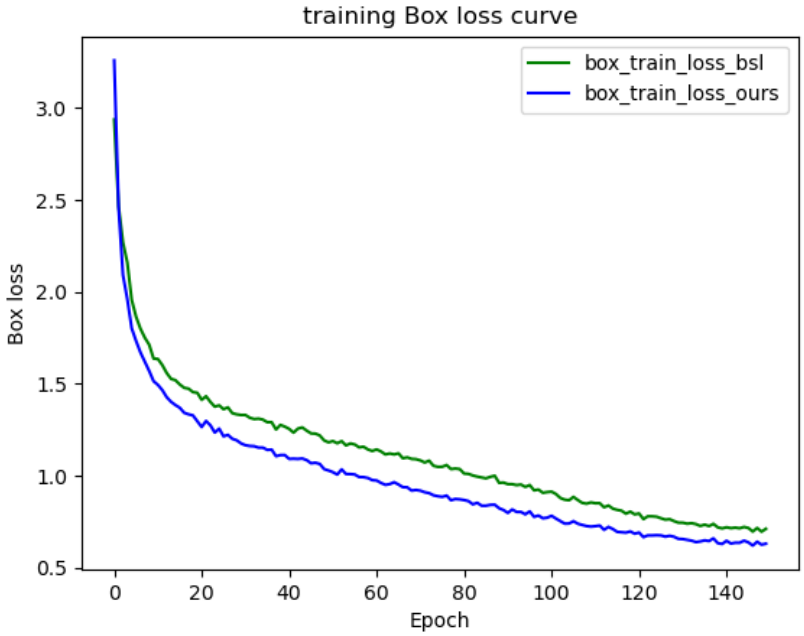

图 4.5 改进前后的模型在训练集上的边界框 loss 变化曲线对比图（bsl 代表改进前的模型，ours 代表本文改进后的模型）

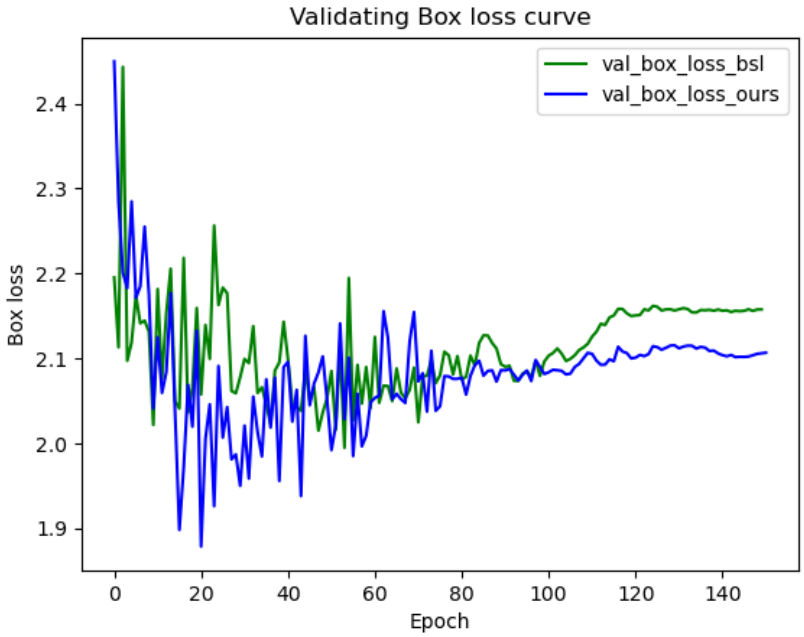

图 4.6 改进前后的模型在测试集上的边界框 loss 变化曲线对比图

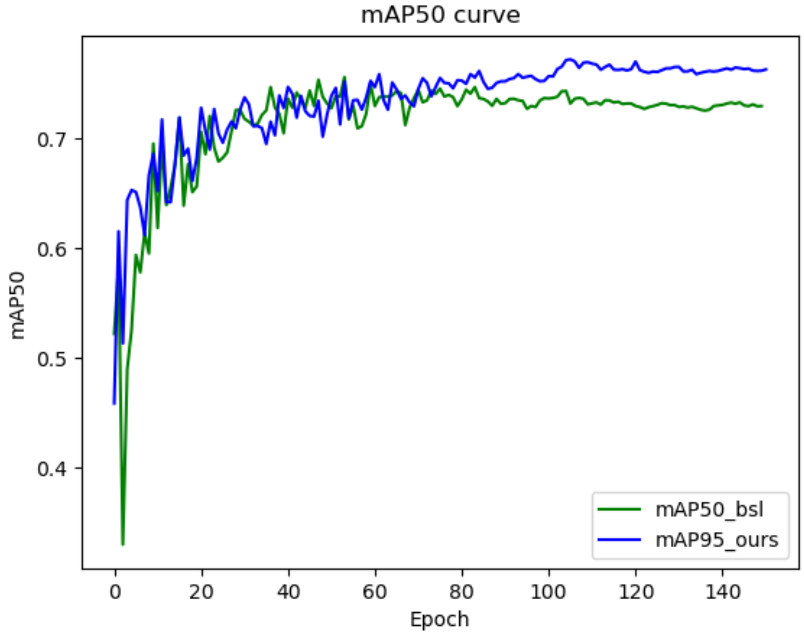

图 4.7 改进前后的模型在测试集上的 mAP50 变化曲线对比图

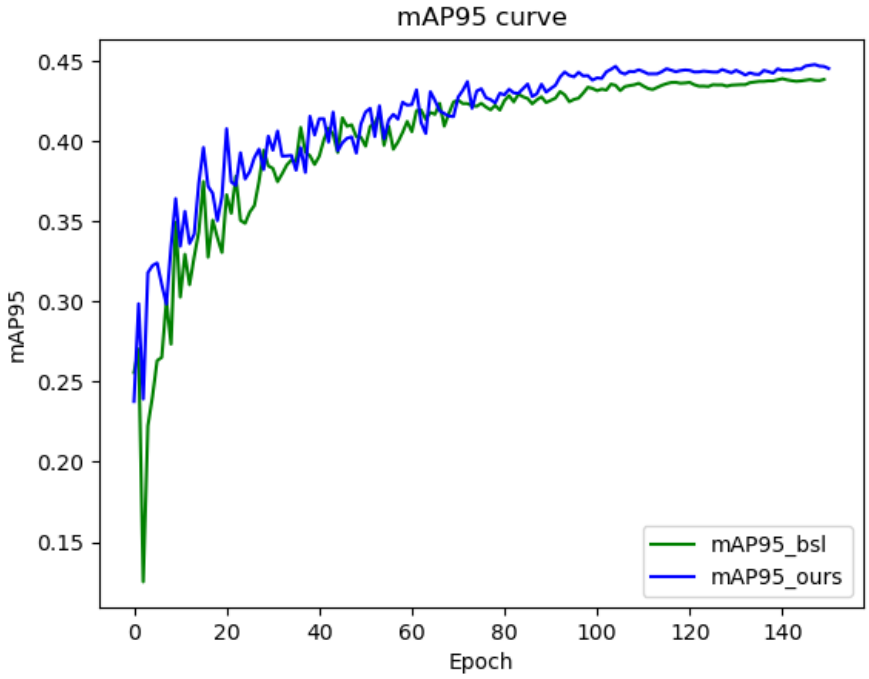

图 4.8 改进前后的模型在测试集上的 mAP95 变化曲线对比图

4.3.3 模型训练结果与分析

在抽烟行为的检测任务中，通过精确度-召回率曲线以及模型打印结果。可以看出改进后的模型在测试集中精确率为 76.6%，召回率为 69.5%，mAP_{50} 为 0.781，mAP_{95} 为 0.457。由于 YOLOv8 是采用可以获得最高 f1 分数的精确度与召回率组合，来计算打印的精确率和召回率，以代表在这个组合下精确率与召回率可以取得最佳平衡，因此打印的数值无法代表实际应用场景下的精确率和召回率表现。所以本文使用混淆矩阵来了解特定阈值条件下的模型检测的性能表现。置信度阈值为 0.25，交并比阈值在 0.45 的条件下的混淆矩阵如图 4.11 所示。在混淆矩阵的阈值条件下，可以计算出精确率为 63.6%，召回率为 82%。

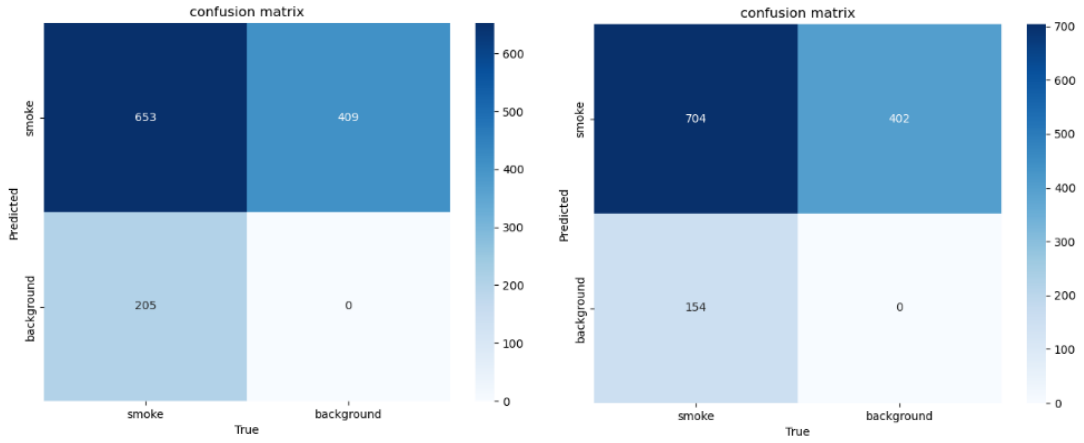

图 4.9 抽烟行为检测测试集的混淆矩阵。左子图为改进前的模型的混淆矩阵，右子图为改进后模型的混淆矩阵

而修改前的模型在测试集中精确率为75.1%，召回率为66.6%，mAP_{50} 为0.725，mAP_{95} 为0.428。修改前后的模型的混淆矩阵如图 4.9 所示。

在混淆矩阵的阈值条件下，可以计算出精确率为61.5%，召回率为76.1%。由此可见，修改后的模型不论是在精度还是查全表现上皆优于原先版本。

图 4.12-4.15 为本文针对不同检测上的难点，挑选出对应的困难样本，用以对比模型改进前后的效果。

图 4.12 模型改进前后对目标角度变化大场景下的检测结果。左图为改进后的结果，右图为改进前的结果

图 4.12 为模型改进前后在目标角度变化较大的场景下的检测效果对比图。在监控场景下，由于视野范围较广，人脸以及香烟的角度变化范围较大，角度变化会导致目标的可用特征减少以及增加与一般样本的特征差异，进而导致目标难以被检测。画面中红色框代表被预测为抽烟的目标，数字则代表置信度，也就是模型估计目标预测正确的概率。改进后的模型相较原有模型，预测目标的置信度上升了3%，并且可以多预测到下方的人的抽烟行为，比较改进前后的检测效果可以看出改进后的模型，利用自注意力机制能够动态调整感受野，对于角度变化大而导致的可用特征减少现象具有更强的鲁棒性[31]。

图 4.13 模型改进前后对目标的易混淆动作的检测结果。左图为改进后的结果，右图为改进前

的结果

图 4.13 为出现了易混淆动作而容易导致误检的图片。由于香烟的特征信息较少，因此容易受到周围上下文信息的影响，特别是捂嘴这些与抽烟动作相似的易混淆动作。原有的模型将右下角的人误检为正在抽烟，而改进后的模型则不会将易混淆动作误检为抽烟行为。比较改进前后的检测效果可以看出改进后的模型通过提升对细节的处理能力，以及更多的全局上下文信息，使得模型能够更好的捕捉具有区别性的特征和特征之间的依赖关系。进而可以更好的区分抽烟行为与易混淆动作之间的差别。

图 4.14 模型改进前后对小型目标的检测结果。左图为改进后的结果，右图为改进前的结果

图 4.14 为模型改进前后对于小型目标的检测效果对比图。小型目标包含的特征信息较少，同时受到边界框偏移以及图像干扰的影响更大，导致小型目标难以检测。相较于改进前的模型，对于与镜头较近的目标置信度下降了5%，并且都将目标的易混淆动作误判为抽烟行为，但是改进后的模型可以多检测到离镜头最远的小型目标。比较改进前后的检测效果可以看出模型具有更强的跨尺度目标检测能力。

图 4.15 模型改进前后对模糊目标的检测结果。左图为改进后的结果，右图为改进前的结果

图 4.15 为模型改进前后对于模糊目标的检测效果对比图。小型目标由于包含的特征信息较少，在图像干扰下受到干扰的有效特征占比较大，因此对于图像干扰敏感。由于该场景图像较为模糊，同时具有强光干扰以及图像对比度的问题，导致可用特征较不明显，因此导致了改进前的 YOLOv8s 模型无法检测到抽烟行为。但是改进后的模型仍然可以检测到场景中的抽烟行为。比较改进前后的检测效果可以看出改进后的模型对于受干扰图像的检测能力有了一定程度的提升，即使在香烟和人脸的局部特征受到严重干扰，仍然可以通过周围的上下文信息判断抽烟行为。

综上所述，可以得知模型改进结果基本符合预期。通过第三章提及的改进技巧，改进后的模型在目标定位精度、跨尺度目标检测、图像细节的提取、易混淆动作的判别以及在遮挡、角度变换和图像干扰场景下检测的鲁棒性，相比改进前的模型取得了明显的提升。从精度来看，精确率、mAP_{50} 和 mAP_{95} 分别达到了 0.766、0.781、0.457，同时在使用与预测阶段相同置信度阈值的情况下，在测试集上的召回率高达79.2%。各项性能指标相较于改进前都有显著的成长，证明了改进方法的可行性。

4.4 与其他模型间的比较

为了证明第三章中改进技巧的有效性，本文分别选取了工业应用中常见的YOLOv5 以及当前单阶段目标检测模型中性能较为突出的模型，例如 DAMO-YOLO、YOLOv8以及DAMO YOLO，并使用较小的缩放规模以符合抽烟行为检测场景下的实时性要求以及避免模型的参数过多导致在小型数据集上容易过拟合。同时本文还加入 YOLOv7 以及 YOLOv8m 这两个较大型的网络，以代表本文的模型在抽烟检测场景下,能够以更少的参数量以及运算量，获得与之同等甚至是更好的检测能力。通过与其他基线模型进行比较，本文的模型具有明显的性能优势，模型之间性能表现的对比如表 4.5 所示。

表 4.5 不同模型之间的性能对比

目标检测模型	参数量（M）	FLOPs（G）	P（%）	R（%）	mAP_{50}（%）	mAP_{95}（%）
YOLOv5s	7.2	16.5	72.3	69.4	72.8	39.4
YOLOv7	36.9	104.7	73.4	70.6	73.1	39
YOLOv8s	11.2	28.6	75.1	66.6	73.5	42.8
YOLOv8m	25.9	78.9	75.5	70.1	75.7	45
DAMO-YOLO	16.3	37.8	78.6	64	74.3	44.6
Ours	13.6	29.7	76.6	69.5	78.1	45.7

相较于原有的YOLOv8s模型，本文的模型分别在mAP_{50}以及mAP_{95}的表现上，高于改进前的模型 3.6 以及 1.9 个百分点，证明了本文的改进方法通过提升模型的跨尺度目标检测能力、图像细节提取、目标定位能力以及获取全局上下文信息的能力，在检测效果上优于改进前的 YOLOv8。

与 YOLOv7 相比，本文认为 YOLOv7 的参数量过大，导致在本文小型的数据集上难以取得良好的训练效果，同时 YOLOv7 属于比较学术化的版本，较为强调与计算机视觉前沿研究结果间的整合，而非工业场景的需要的容易训练和部署。

与 YOLOv8m 相比，本文对于网路拓展的策略并非增加骨干网络的特征堆叠以增加特征提取能力，而是参考 GFPN 以及 DAMO-YOLO 重颈部(Heavy Neck)的范式，着重于增加颈部网络的跨尺度连接以及跳层链接以提升特征融合能力。从结果来可以看出，虽然mAP_{95}略低于 YOLOv8m 0.3个百分点，但是本文的模型在mAP_{50}的表现上优于YOLOv8m 1.7个百分点，并且运算量以及参数量皆远低于YOLOv8m，代表着重于特征融合的设计范式在抽烟行为检测这种尺度变化大的检测任务中，相较于传统着重于特征提取的设计范式，可以取得更好的效果和效率。

相较于 DAMO-YOLO 本文的模型在精度以及查全能力间取得了更好的平衡，本文虽然参考了 DAMO-YOLO 的重颈部范式，但仍然采用 YOLOv8s 原有的骨干网络。从结果可以看出，可能是因为骨干网络过度轻量化导致特征提取的不足，DAMO-YOLO-S 的查全表现普通，因此即使精确率上本文低于 DAMO-YOLO-S 2 个百分点，但由于本文保留了 YOLOv8s 特征提取能力较强的骨干网络，在特征提取与特征融合之间取得了更好的平衡，因此取得了更有优异的表现，通过上述实验结果可以得出本文的模型相较于现有的前沿模型，在抽烟行为检测场景下具有更好的性能表现。

4.5 消融实验

通过消融实验可以分析模块或参数的变化，与模型性能表现之间的关联。在抽烟行为检测数据集上的消融实验结果如表 4.6 所示。

表 4.6 在抽烟行为检测数据集上的消融实验结果

目标检测模型	参数量（M）	FLOPs（G）	P（%）	R（%）	mAP_{50}（%）	mAP_{95}（%）
YOLOv8s	11.2	28.6	75.1	66.6	72.5	42.8
YOLOv8s + NWD0.5	11.2	28.6	67.7	72	72.7	42.7
YOLOv8s + NWD0.2	11.2	28.6	74.1	68.6	73.9	43.1
YOLOv8s+our neck+NWD0.2	13.3	25.3	75.3	69.6	75.7	44
DETR + Full Encoder+NWD0.2	17.6	44.3	76.9	69.6	78.3	46.2
DETR+our neck+NWD0.2	13.6	29.7	76.6	69.5	78.1	45.7

首先是添加归一化 Wasserstein 距离到原有的 YOLOv8s 模型的边界框回归损失函数的交并比计算中。本文参考了同样使用了归一化 Wasserstein 距离的 YOLO-FaceV2[41]实验中效果较好的两个参数设置，分别为将交并比的权重设置为 0.5，归一化 Wasserstein 距离的权重设置为0.5，以及将交并比的权重设置为 0.8，归一化 Wasserstein 距离的权重设置为 0.2，通过消融实验的结果可以看出，虽然在边界框相似度度量上采用较高的归一化 Wasserstein 距离占比可以有更高的查全表现，召回率为 0.72，但是精度表现普通，精确率只有 67.7%，导致 mAP_{50} 和 mAP_{95} 相比于改进前有所下降。因此本文采用将交并比的权重设置为 0.8，归一化 Wasserstein 距离的权重设置为 0.2 的设置，在保有一定程度的跨尺度检测能力提升之外，也不会过多的影响对普通尺度目标的检测精度。

最后是添加使用DETR檢測頭，使用 Transformer 替换掉原有的檢測頭部可以增加模型的检测能力，通过自注意力机制的长距离特征依赖关系的建模能力获取丰富的上下文信息，使得模型能够倾向关注物体整体，而非局部特征，特别是

对于监控场景下如遮挡、角度变换和图像干扰等检测难点有了明显的改善。因此显著的提升了检测表现，mAP 相较于 YOLOv8s 提升 5.6%。并且可以发现高效编码器，与基于Transformer的编码器模块相比，性能只有微幅降低，但在参数量和运算量上有一定程度的下降。

5 总结与展望

5.1 总结

现有的研究工作提出了多种目标检测模型，并且通过与计算机视觉领域的前沿研究进行集成，使得目标检测模型可以适用于更多的应用场景以及具有更好的检测表现。有鉴于当前抽烟检测的研究相对不足，并且设计上明显与现实应用场景脱节。本文主要通过设计与制作适用于复杂场景下抽烟行为检测的数据集，以及使用 DETR 至架构改进 YOLOv8s 目标检测模型这两个方面，来改进现有相关工作的不足。

数据集方面，对于复杂场景下的抽烟行为的难点以及相关研究的不足进行分析，设计与制作适用于复杂场景下抽烟行为检测的数据集，通过改进标注策略、数据增强以及负样本设计。本文成功在不改变训练集数据规模，即训练图片总数，以及测试集内容的前提下。成功提升了在改进前的 YOLOv8s 模型上的检测表现，证明数据集设计方法的有效性。

模型改进方面，根据对现有工作的调研，本文选取 YOLOv8s 为改进对象，并且引入重量级颈部设计范式增加更多的跨尺度连接和跳层连接，提升模型的特征融合能力和训练效果。通过将重参数化卷积引入原有的特征融合模块，能够在基本不增加推理时间的情况下，利用多分支网络的优点提升训练效果。此外，还

在颈部末端的低分辨率特征图上使用 PSA注意力替代 Transformer 编码器降低训练和推理成本，同时保持和原有 DETR 架构接近的性能表现。並且使用DETR檢測頭部替代原有的 YOLO 检测头部（仅作为辅助分支提供监督信息），並且还可以通过自注意力机制的长距离特征依赖关系的建模能力获取丰富的上下文信息，使得模型能够倾向关注物体整体，而非局部特征，进而有效提升遮挡、目标角度变换和图像干扰等场景下的检测能力。

本文通过将改进后的模型与现有前沿模型进行比较。实验结果表明，在复杂场景下的抽烟行为检测任务下，改进后的模型无论是检测性能还是性能开销都取得一定的优势，并通过消融实验进一步说明改进模块是如何影响性能表现。证明了改进模型的特征融合能力、损失函数中的边界框相似度度量，以及引入自注意力机制能够有效提升目标检测模型在复杂场景下的抽烟行为检测任务下的表现。

5.2 展望

本文工作通过改进模型的方法，成功提升模型在抽烟行为检测的表现，并且相比于现有的前沿模型具有一定程度的性能提升。但是根据章节 4.3.3 的实验结果，在控制错误检测方面，以及模型的查全能力方面仍然存在改进空间。因此本文认为可以在以下三个方面进行更深入的研究。

数据集方面，由于网络爬取图像与监控图像间的特征差异过大，导致训练集与测试集之间的数据分布不一致，进而导致章节 4.3.2 中出现的测试集 loss 曲线出现异常以及过拟合的情况。需要更注意数据的来源以及内容，同时深入研究主流公开数据集的设计方法，了解其是如何设计正负样本以及样本的尺度、空间

分布，确保模型学习的样本多样性足够，并且能够确保训练集和测试集的分布能够尽可能一致避免测试集 loss 出现异常情况。

数据后处理方面可以通过通过加权框融合[42]（Weighted Box Fusion, WBF）将不同训练设置导致不同检测倾向的模型的检测结果，通过加权融合预测框的方法有效结合不同模型检测的侧重点，通过模型集成的方法进而提升模型的检测能力与泛化性。此外还可以通过在预测阶段也导入数据增强[43]，通过在测试时将数据进行切片或是创建使用不同增强方法的副本，使得样本能够以更容易被检测的形式被模型预测，进而提升预测表现。

模型训练部分，可以在模型训练的时候采用知识蒸馏[44]（Knowledge Distillation）的方法将学习能力强的复杂模型作为教师模型去监督简单的学生模型的训练。以实验章节 4.2.2 中 YOLOv8m 的实验结果为例，较大规模的模型一般能够对数据进行更充分的学习，因此能够取得更好的表现。通过知识蒸馏将数据的真实标签与教师模型较为精确的预测结果作为监督信息去训练学生模型，可以使较为简单或是训练资源较为不足的模型，能够取得接近复杂模型的性能。

参考文献

[1] Chien T C, Lin C C, Fan C P. Deep learning based driver smoking behavior detection for driving safety[J]. Journal of Image and Graphics, 2020, 8(1): 15-20.
[2] Yang Z, Yao D. Fast TLAM: High-precision fine grain smoking behavior detection network[C]//2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP). IEEE, 2020: 183-188.
[3] Tang J, Liu S, Zheng B, et al. Smoking behavior detection based on improved YOLOv5s algorithm[C]//2021 9th International Symposium on Next Generation Electronics (ISNE). IEEE, 2021: 1-4.
[4] Cheng G , Yuan X , Yao X , et al. Towards Large-Scale Small Object Detection: Survey and Benchmarks[J]. 2022.
[5] 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.
[6] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.
[7] 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.
[8] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.
[9] 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.
[10] Liu S, Qi L, Qin H, et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8759-8768.
[11] Tan M, Pang R, Le Q V. Efficientdet: Scalable and efficient object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10781-10790.
[12] Ren S , He K , Girshick R , et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149.
[13] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.
[14] Tian Z, Shen C, Chen H, et al. Fcos: Fully convolutional one-stage object detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 9627-9636.
[15] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.

[16] Zhu X, Lyu S, Wang X, et al. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 2778-2788.
[17] Li C, Li L, Jiang H, et al. YOLOv6: A single-stage object detection framework for industrial applications[J]. arXiv preprint arXiv:2209.02976, 2022.
[18] Ding X, Zhang X, Ma N, et al. Repvgg: Making vgg-style convnets great again[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 13733-13742.
[19] Feng C, Zhong Y, Gao Y, et al. Tood: Task-aligned one-stage object detection[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE Computer Society, 2021: 3490-3499.
[20] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[J]. arXiv preprint arXiv:2207.02696, 2022.
[21] Wang C Y, Liao H Y M, Yeh I H. Designing Network Design Strategies Through Gradient Path Analysis[J]. arXiv preprint arXiv:2211.04800, 2022.
[22] 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. 2017: 4700-4708.
[23] Xu X, Jiang Y, Chen W, et al. DAMO-YOLO: A Report on Real-Time Object Detection Design[J]. arXiv preprint arXiv:2211.15444, 2022.
[24] Jiang Y, Tan Z, Wang J, et al. GiraffeDet: a heavy-neck paradigm for object detection[J]. arXiv preprint arXiv:2202.04256, 2022.
[25] Li X, Wang W, Hu X, et al. Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 11632-11641.
[26] Kisantal M, Wojna Z, Murawski J, et al. Augmentation for small object detection[J]. arXiv preprint arXiv:1902.07296, 2019.
[27] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to end object detection with transformers. In Computer Vision
[28] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In International Conference
[29] Feng Li, Hao Zhang, Shi guang Liu, Jian Guo, Lionel Ming shuan Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 13609–13617, 2022. on Learning Representations.
[30] Hao Zhang, Feng Li, Siyi Liu, Lei Zhang, Hang Su, Jun-Juan Zhu, Lionel Ming shuan Ni, and Heung yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. ArXiv, abs/2203.03605, 2022.

[31] Naseer M M, Ranasinghe K, Khan S H, et al. Intriguing properties of vision transformers[J]. Advances in Neural Information Processing Systems, 2021, 34: 23296-23308.
[32] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014: 740-755.
[33] Zheng Z, Wang P, Liu W, et al. Distance-IoU loss: Faster and better learning for bounding box regression[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(07): 12993-13000.
[34] Wu H, Xiao B, Codella N, et al. Cvt: Introducing convolutions to vision transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 22-31.
[35] Zheng Z, Wang P, Liu W, et al. Distance-IoU loss: Faster and better learning for bounding box regression[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(07): 12993-13000.
[36] Wang J, Xu C, Yang W, et al. A normalized Gaussian Wasserstein distance for tiny object detection[J]. arXiv preprint arXiv:2110.13389, 2021.
[37] Bazarevsky V, Grishchenko I, Raveendran K, et al. Blazepose: On-device real-time body pose tracking[J]. arXiv preprint arXiv:2006.10204, 2020.
[38] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). Ieee, 2005, 1: 886-893.
[39] 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. Ieee, 2009: 248-255.
[40] Ishida T, Yamane I, Sakai T, et al. Do we need zero training loss after achieving zero training error?[J]. arXiv preprint arXiv:2002.08709, 2020.
[41] Yu Z, Huang H, Chen W, et al. YOLO-FaceV2: A Scale and Occlusion Aware Face Detector[J]. arXiv preprint arXiv:2208.02019, 2022.
[42] Solovyev R, Wang W, Gabruseva T. Weighted boxes fusion: Ensembling boxes from different object detection models[J]. Image and Vision Computing, 2021, 107: 104117.
[43] Akyon F C, Altinuc S O, Temizel A. Slicing aided hyper inference and fine-tuning for small object detection[C]//2022 IEEE International Conference on Image Processing (ICIP). IEEE, 2022: 966-970.
[44] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv:1503.02531, 2015.

附 录

表附 1

本次论文的实验环境

操作系统	Ubuntu 18.04
处理器	Intel(R) Xeon(R) Gold 6230N
显卡	NVIDIA A800

作者简历

姓名：曹闵丞　性别：男　民族：汉　出生年月：2000-03-26　籍贯：台湾省台北市

学术经历：

2018.09-2023.06　浙江大学攻读学士学位

获奖情况：无

参加项目：无

发表的学术论文：无

本科生毕业论文（设计）任务书

一、题目：复杂场景下抽烟行为检测技术研究

二、指导教师对毕业论文（设计）的进度安排及任务要求：

时间	进度安排
3.26 之前	1.对目标检测算法进行调研，了解目前的模型的网络架构、性能表现以及应用场景与局限。 2.对复杂场景下抽烟行为的检测问题进行分析，了解其实现上的难点及需求。同时对数据集进行整理， 3.编写开题报告。
3.26-4.10	1.基于调研结果与实验，选择适用于复杂场景下抽烟行为的检测问题的目标检测算法，以其作为基线以及优化目标。 2.学习 Transformer 以及数据增强方法以及深入了解目标检测模型中颈部网络以及损失函数的实现与细节。
4.11-5.7	1.通过先前的深入了解的相关改进技巧，完成模型的设计。 2.对所建立的模型进行实验，对实验结果进行分析，并基于对实验的分析，进一步完善模型的设计以提升结果性能。
5.8-5.21	进行毕设论文的写作、准备答辩。

1.阅读前沿论文，了解吸烟行为检测和目标检测模型的发展现状。

2.设计并制作复杂场景下的吸烟行为检测数据集。

3.设计并实现复杂场景下的吸烟行为检测视频检测模型。

4.在自己制作的数据集上进行一定规模的实验验证。

起讫日期 **22** 年 月 日 至 **23** 年 月 日

指导教师（签名）_____职称_____

三、系或研究所审核意见：

负责人（签名）_____

年 月 日

毕 业 论 文（设计） 考 核

一、指导教师对毕业论文（设计）的评语：

指导教师(签名）

年 月 日

二、答辩小组对毕业论文（设计）的答辩评语及总评成绩：

成绩比例	文献综述/中期报告占（10%）	开题报告占（15%）	外文翻译占（5%）	毕业论文（设计）质量及答辩占（70%）	总评成绩
分值					

答辩小组负责人（签名）

年　　月　　日