基于深度学习的人脸检测算法研究与实现

数媒2101班：康玥瑶 指导教师：李湘眷 论文类型：毕业论文

摘要：近年来人工智能技术发展迅速，人脸检测应用广泛，但传统方法在复杂场景下、姿态变化等情况下存在不足。本研究对YOLOv10和Faster R-CNN两种深度学习算法展开研究。在IMDB-WIKI 500k+和 WIDER FACE数据集基础上构建更全面的人脸数据集，进行模型训练与测试，对比分析两种算法的性能，并基于PyTorch框架和PyQt5开发前端界面，将两种算法进行集成并实现图像、视频的人脸检测。实验结果表明，YOLOv10适用于对实时性和综合检测能力要求高的场景，Faster R-CNN更适合检测精度要求较严苛的场景。

关键词：深度学习；人脸检测；YOLOv10；Faster R-CNN

1 基于YOLOv10的人脸检测

1.1 数据集采集和预处理

本研究数据集部分来源于IMDB-WIKI 500k+[3]、WIDER FACE数据集[4]并进行选取，另一部分通过使用OIDv4工具包从Google Open Images中抓取相关图像并进行筛选。数据集包含大量不同场景、不同姿态和光照条件下的人脸图像，并且已经进行了标注，能够为模型训练提供丰富的数据资源。部分数据集图片如图1.1所示，部分已标注的人脸数据如图1.2所示。最终，共采集到了2500多张人脸图像数据，其中验证集400张，训练集2000张。

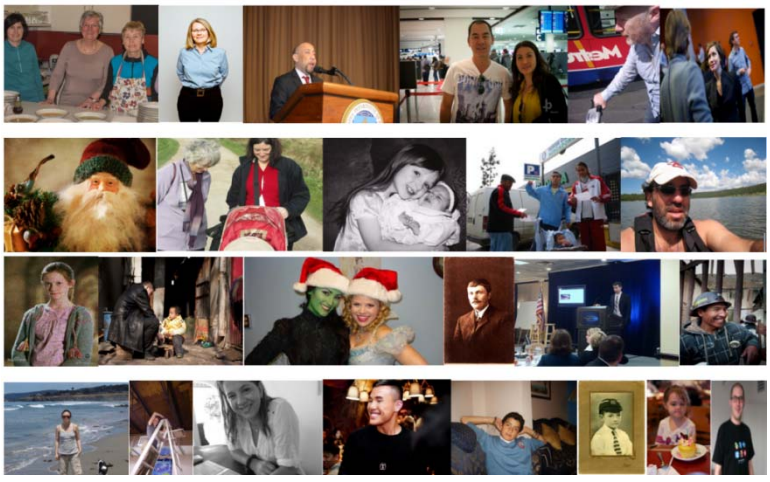


图1.1 部分数据集图片

图1.2 部分数据集图片

在训练前需要对采集到的数据需要进行预处理。本研究对采集到的2500余张图像进行了系统化的预处理流程，对其进行了数据清洗和数据增强。如图1.3为对焦不准或模糊导致的低质量图像。如图1.4采用了数据增强技术，包括随机翻转、裁剪、亮度调整、对比度调整等。

图1.3 对焦不准或模糊导致的低质量图像

(a)水平翻转处理效果图　　(b)垂直翻转处理效果图

（c）裁剪处理效果图　　（d）亮度处理效果图

图1.4 数据增强处理效果图

1.2 模型架构

YOLOv10作为单阶段目标检测算法，通过将图像划分为网格单元，直接在各单

元中预测人脸边界框与类别概率，以端到端的简洁计算流程实现高效检测。该模型兼具高检测速度与实时性优势，适用于动态场景下的人脸快速定位，能在复杂背景中快速确定人脸区域的位置与尺寸。但由于人脸在姿态、遮挡、光照变化下的多样性，仅依靠YOLOv10的检测结果可能存在特征提取不精细的问题，尤其在小尺寸人脸或复杂姿态下易出现定位偏差。检测的模型基本结构如图1.5[1]所示。

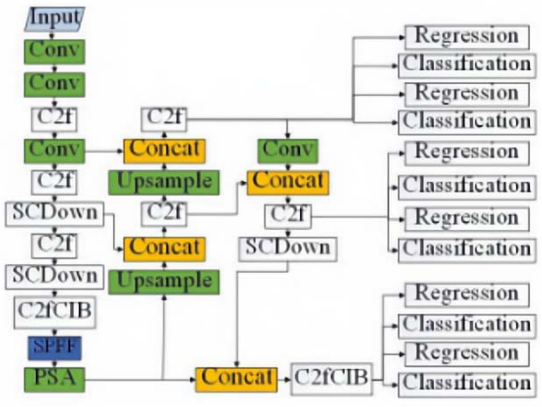


图1.5 YOLOv10基本结构图

1.3 模型训练和测试

在数据采集和预处理后，开始训练YOLOv10模型。先使用数据集预训练权重初始化模型参数，以加快收敛、减少资源消耗。接着设置训练参数，初始学习率0.001，迭代100次，批量大小4。其损失函数包含边界框回归、类别分类和置信度损失，通过最小化三者加权和更新参数。训练过程中，分类损失（Classification Loss）采用Focal Loss解决正负样本不平衡问题，如式3.1所示：：

$$L_{cls} = -\alpha_t(1-p_t)^{\gamma} log(p_t) \quad (3.1)$$

其中，p_t为预测概率，α_t为类别权重，γ为调制因子。

边界框回归损失（Bounding Box Loss）采用CIoU Loss优化定位精度，同时考虑重叠面积、中心点距离和长宽比，如式3.2所示：

$$L_{box} = 1-IoU + \frac{\rho^2(b,b^{gt})}{c^2} + \alpha v \quad (3.2)$$

其中，IoU为交并比阈值，ρ^2为预测框与真实框中心点的欧氏距离平方，c为包含两框的最小外接矩形对角线长度，$\boldsymbol{\alpha}$和v为长宽比相关参数。

分布聚焦损失（Distribution Focal Loss,DFL）用于优化回归值的概率分布，提升边界框的精细定位能力，公式为：

$$L_{dfl} = -\sum_{i=0}^{n-1} y_i log(p_i) \quad (3.3)$$

其中，y_i为目标分布的真实值，p_i为预测的概率分布。

训练时将预处理的训练数据输入模型，正向传播计算输出，根据损失函数算损失值，定期保存权重。

如图1.6的loss曲线显示，box_loss、cls_loss和dfl_loss的训练集与验证集值均随

训练轮次增加而下降，表明模型在目标定位、分类及边界框分布处理上学习有效。但需注意两集合损失差异，以防过拟合或欠拟合，保证模型泛化能力。

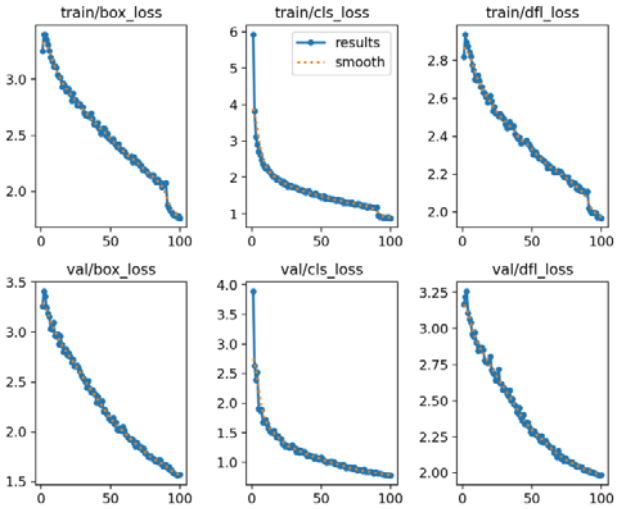
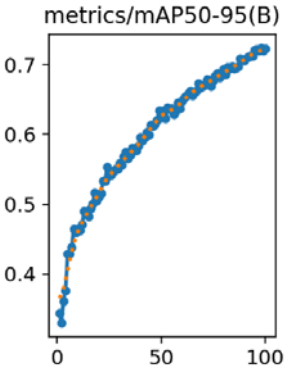
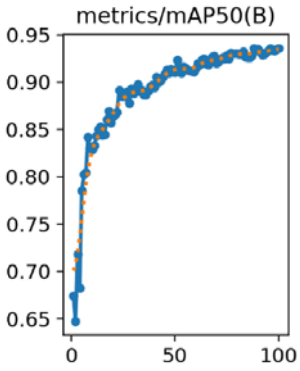


图1.6 YOLOv10模型训练loss图

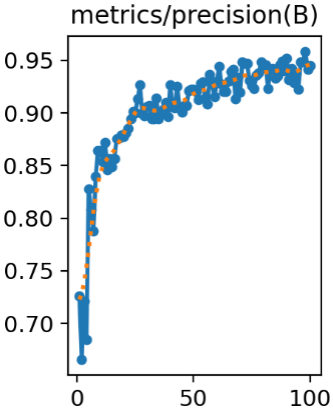
如图1.7展现YOLOv10的性能变化。mAP50-95、mAP50 随轮次增加持续上升，模型检测精度提升；precision（精度）、recall（召回率）也逐步上升并趋于稳定，说明模型在识别真阳性、平衡精确与召回方面优化。整体来看，训练过程中模型检测能力不断增强，已具备较好性能。



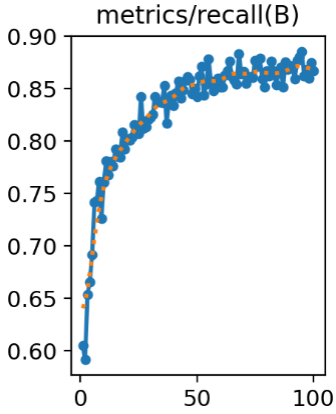
（a）YOLOv10模型训练mAP50-95曲线图



（b）YOLOv10模型训练mAP50曲线图



（c）YOLOv10模型训练precision的曲线图



（d）YOLOv10模型训练recall的曲线图

图1.7 YOLOv10的性能指标曲线图

训练完成后进行测试，准备格式与训练数据一致的测试数据，用训练好的模型推理，得到边界框位置、类别和置信度，还分析了模型在光照、姿态变化及遮挡等不同场景下的检测效果。YOLOv10的部分检测效果如图1.8所示。

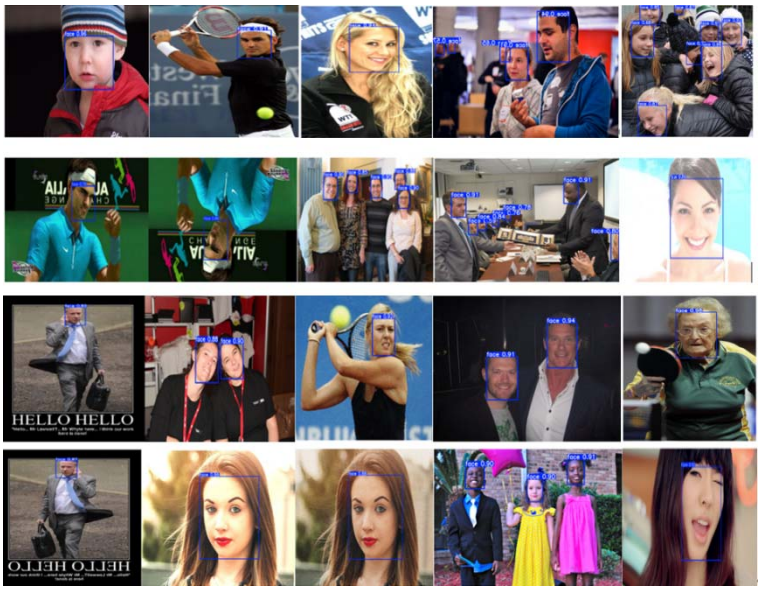


图1.8 YOLOv10人脸检测结果图

2 基于Faster-R-CNN的人脸检测

2.1 数据采集与预处理

数据采集流程与YOLOv10系统保持一致，确保数据集的同源性，便于后续对比实验。针对Faster R-CNN训练耗时较长的特点，对训练集进行二次筛选，剔除标注模糊的样本，提高训练效率。

2.2 模型架构

Faster R-CNN作为两阶段目标检测算法，先通过区域建议网络（RPN）生成潜在人脸候选区域，再对候选区域进行特征细化与分类回归。该模型通过深度特征提取与多层级语义分析，在遮挡、多角度或小目标人脸场景中表现出更高的检测精度，但计算复杂度较高，实时性较差。检测算法原理图如图2.1[2]所示。

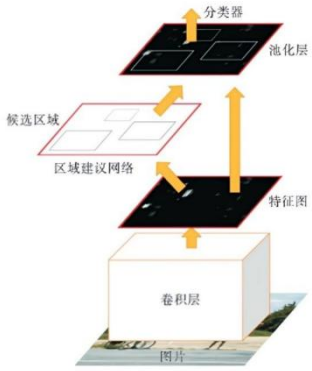


图2.1 Faster R-CNN基本结构图

2.3 模型训练与测试

选用预训练模型提取通用图像特征以加速训练收敛。骨干网络提取特征图后，RPN通过滑动窗口在特征图上生成多个锚框，并对其进行前景/背景分类和边界框回归，进一步生成高质量的目标候选区域[5]，并将候选区域映射到特征图，经ROI Pooling[6]生成固定大小特征向量，由全连接层完成目标分类和边界框精修。测试时，图像经骨干网络得特征图，RPN生成候选区域，非极大值抑制筛选后，经ROI Pooling和全连接层分类回归，输出目标类别、边界框坐标等信息。

如图2.2展示Faster R-CNN训练时，train loss（训练损失）、val loss（验证损失）及对应平滑损失（smooth train loss、smooth val loss）随训练轮次（Epoch）的变化。起始阶段损失值高，随Epoch增加，各类损失总体呈下降趋势，说明模型在训练过程中不断优化，对数据的拟合和泛化能力逐步提升。

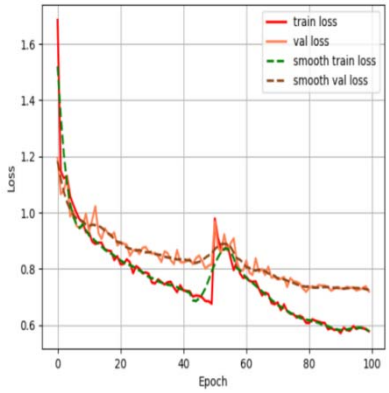


图2.2 Faster R-CNN训练Loss图

如图2.3显示Faster R-CNN训练阈值0.5时，训练集mAP初始近0，10轮达0.7后缓升至接近0.8，期间在0.7-0.8波动。表明前期学习快、精度提升大，后期优化难但仍有空间，波动或受数据分布与参数影响，整体检测精度在训练集上逐步提高。

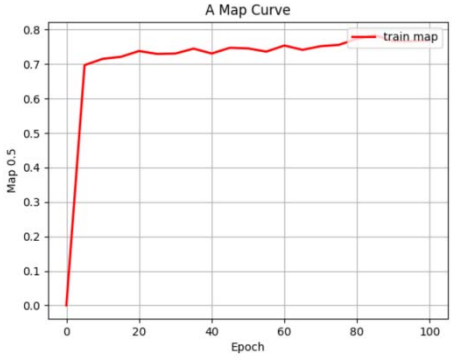


图2.3 Faster R-CNN训练mAP图

人脸检测测试结果图如图2.4所示：

图2.4 Faster-R-CNN人脸检测结果图

3 软件设计与算法对比分析

3.1 软件设计

本系统基于YOLOv10和Faster R-CNN，选用了Python作为开发。系统操作简便，用户体验友好，系统提供丰富的检测参数设置功能，赋予用户根据不同应用场景灵活调整检测策略的能力。在算法选择层面，用户可通过直观的下拉菜单，轻松在YOLOv10（适合实时场景）与Faster R-CNN（适合高精度需求场景）之间切换，满足多样化的检测需求。并含有置信度阈值与交并比（IoU）阈值的动态设置。同时检测结果展示（显示标签名称与置信度、用时、目标数目、可进行目标选择、明确类型、呈现置信度及目标位置 ）、检测结果与位置信息记录以及操作功能（能打开图片、文件夹、视频、摄像头，可进行保存和退出操作），人脸检测系统功能模块如图3.1所示。

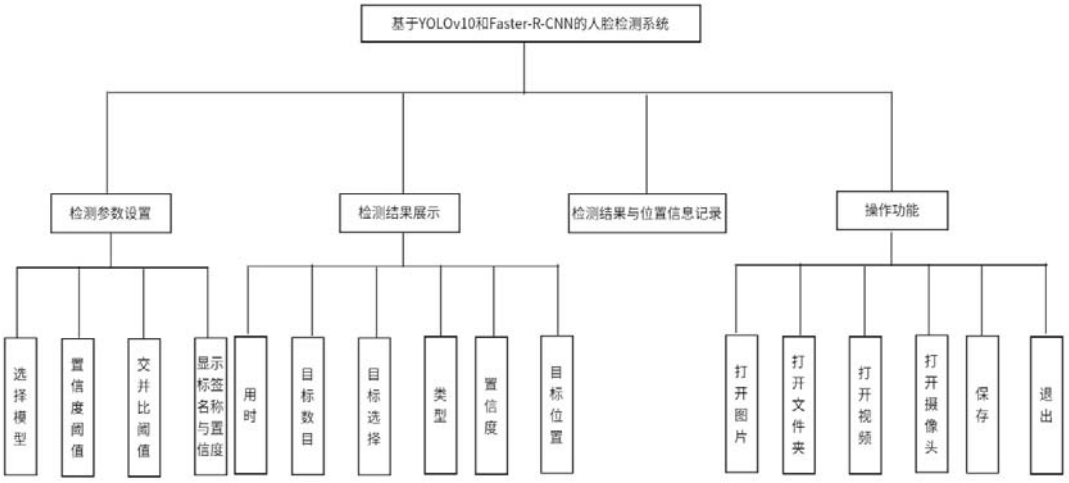


图3.1 系统架构图

该软件具有对照片、文件夹、视频，摄像头实时画面进行检测的功能，检测界面如图3.2所示。

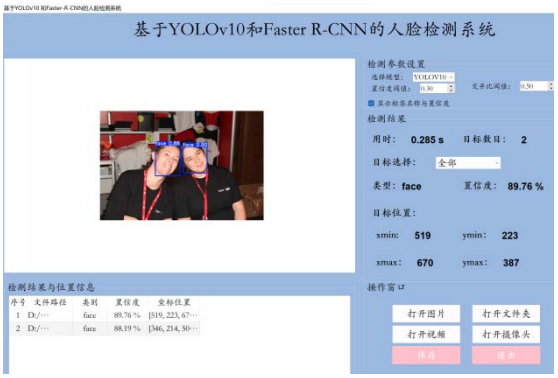


图3.2 人脸检测系统界面效果图

3.2 评估指标

评估指标采用目标检测领域通用标准：平均精度均值（mAP）衡量模型综合性能，计算不同IoU阈值下各类别平均精度的均值；精确率（Precision）反映预测为正样本中实际为正的比例，如式3.1所示：

$$Precision = \frac{TP}{TP + FP} \tag{3.1}$$

其中，*TP*（真正例）是模型正确预测为正的样本数，*FP*（假正例）是模型错误预测为正的样本数，该*Precision*值反映预测为正样本中实际为正的比例。

召回率（Recall）表示实际正样本中被正确检测出的比例，如式3.2所示：

$$Recall = \frac{TP}{TP + FN} \tag{3.2}$$

其中，*FN*（假负例）是模型错误预测为负的样本数，该*Recall*值表示实际正样本中被正确检测出的比例。

*F*1值为精确率和召回率的调和平均数，综合评估模型性能，如式3.3所示：

$$F1 = 2 \times \frac{Precision + Recall}{Precision \times Recall} \tag{3.3}$$

3.3 对比实验设计

在相同实验数据集上对YOLOv10和Faster R-CNN分别进行训练和测试展开对比实验，采用相同的评估指标对两个模型的性能进行对比分析。然后，基于前端界面从图像输入多样性展开对比实验，通过“打开图片”功能，选取含模糊、低分辨率、复杂背景的人脸图像，记录YOLOv10与Faster R-CNN的检测人脸数量、位置准确性、置信度等，分析算法对非理想数据的适应性。在实验过程中，不断调整置信度阈值（0.2、0.4、0.8），观察两种算法检测结果变化，对比低阈值下的多检测或多误检与高阈值下的少误检差异。

3.4 结果对比与分析

YOLOv10和Faster R-CNN两个模型在测试数据集下各项性能指标曲线对比图，如图3.3所示，其中包含Loss对比，Precision对比，Recall对比，F1分数对比，不同阈值下的mAP对比。综合各项指标的曲线图进行分析发现在大量数据集上训练，YOLOv10在收敛效率、检测速度、精度上整体优于Faster R-CNN，更适配对实时性与精度均有要求的场景；Faster R-CNN也有独特优势，但综合性能稍逊，不过在特定对两阶段精细处理依赖高的场景，有其应用价值。

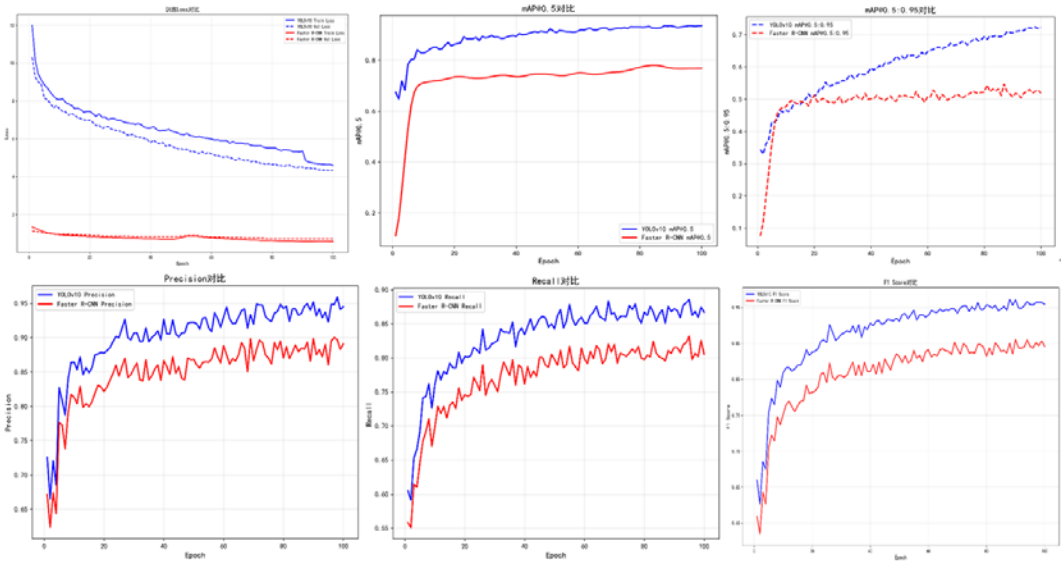


图3.3 YOLOv10和Faster R-CNN训练指标对比图

如图3.4所示为YOLOv10和Faster R-CNN两个模型前端界面，设置相同的置信度阈值和交并比阈值，选取两张噪点比较多的照片，分别通过YOLOv10和Faster R-CNN检测模型进行检测，YOLOv10结果显示两张照片的检测用时分别是0.158s和0.156s，置信度分别为89.73%和89.63%，Faster R-CNN结果显示两张照片的检测用时分别是37.594s和22.331s，置信度分别为99.83%和99.79%。由此说明YOLOv10检测速度快，能快速给出检测结果，满足对实时性要求较高的场景需求，而Faster R-CNN模型结构复杂，训练和运行需要较多计算资源支持，检测速度较慢，但相对置信度而言，虽然检测精度都比较高，但在此条件下Faster R-CNN模型的检测精度更高。

（a）Faster R-CNN的检测图　　（b）YOLOv10的检测图

（c）Faster R-CNN的检测图　　（d）YOLOv10的检测图

图3.4 YOLOv10与Faster R-CNN检测对比图

设置0.2、0.4、0.8不同置信度阈值时，YOLOv10检测人脸数从6个减至4个，如图3.5所示，因单阶段模型对阈值敏感，阈值升高会舍弃低置信度结果；Faster R-CNN始终检出6个，如图3.6所示，因两阶段模型置信度高，对阈值变化不敏感，在该场景下能稳定检测所有目标，体现二者对置信度阈值敏感性的显著差异。

（a）置信度为0.2的检测图　（b）置信度为0.4的检测图　（c）置信度为0.8的检测图

图3.5 YOLOv10不同阈值的效果图

（a）置信度为0.2的检测图　（b）置信度为0.4的检测图　（c）置信度为0.8的检测图

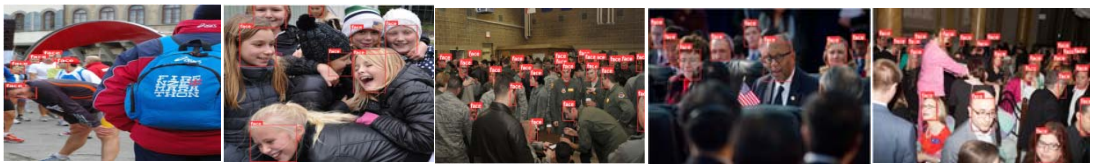
图3.6 Faster R-CNN不同阈值的效果图

如图3.7和图3.8分别选取十张复杂场景作为人脸检测目标，并通过表3.1和表3.2展示的结果来看，二者在不同场景下呈现出不同性能特点。

在复杂场景的人脸检测任务中，YOLOv10与Faster R-CNN表现出明显不同的性能表现。而如表3.1和表3.2显示，在检测用时方面，YOLOv10具有明显优势，用时集中在0.118-0.249s，因其轻量化模型架构，能快速完成检测，适合实时性需求高的场景；而Faster R-CNN用时大幅度增加，处于19.237-34.207s之间，检测速度慢，在对响应速度敏感的场景中应用受限。同时YOLOv10的检出率受人脸数目影响大，图（g）8个人脸时检出率达100%，但图（h）24个人脸的检出率降低至34.29%，高密度人脸

易干扰检测效果；最高置信度在79.79%-92.79%之间，不过结果可靠性仍有提升空间。显示Faster R-CNN检出率在中低人脸密度（如图（d）、图（e）、图（g））下表现稳定，高达100%，但图（h）人脸检出率也降至54.29%，在面对极高密度人脸仍然受限；不过其最高置信度具有明显优势，大都在95%以上，部分图（如图（a）、图（b）、图（c））高达99%以上，检测结果可信度高。整体而言，YOLOv10适合追求快速检测的场景，Faster R-CNN在精度要求高的复杂场景下更具价值，实际应用可依据场景对速度、精度的侧重灵活选择。

（a）讲座场景 （b）聚会场景 （c）聚餐场景 （d）教学场景 （e）会议场景



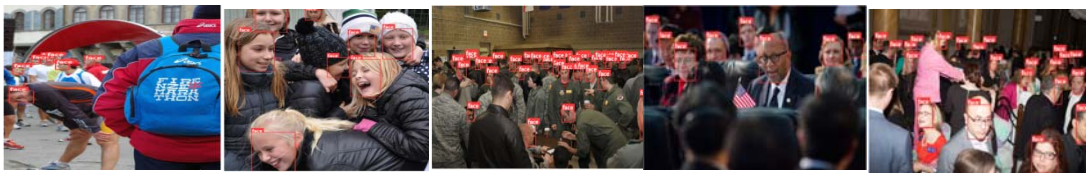
（f）街道场景 （g）玩耍场景 （h）集合场景 （i）会坛场景 （j）密集场景

图3.7 复杂场景下YOLOv10的人脸检测结果

表3.1 复杂场景下YOLOv10的检测性能

图片	用时（s）	人脸数目（个）	检出率	最高置信度
图（a）	0.123	8	50.00%	90.37%
图（b）	0.123	6	85.71%	91.61%
图（c）	0.128	4	40.00%	92.79%
图（d）	0.118	7	87.50%	92.27%
图（e）	0.130	4	80.00%	88.97%
图（f）	0.128	5	62.50%	83.50%
图（g）	0.249	8	100.00%	92.38%
图（h）	0.128	24	34.29%	79.79%
图（i）	0.127	8	80.00%	90.87%
图（j）	0.126	22	68.75%	90.07%

（a）讲座场景 （b）聚会场景 （c）聚餐场景 （d）教学场景 （e）会议场景



（f）街道场景 （g）玩耍场景 （h）集合场景 （i）会坛场景 （j）密集场景

图3.8 复杂场景下Faster R-CNN的人脸检测结果

表3.2 复杂场景下Faster R-CNN的检测性能

图片	用时（s）	人脸数目（个）	检出率	最高置信度
图（a）	21.299	12	75.00%	99.94%
图（b）	19.991	6	85.71%	99.96%
图（c）	21.237	5	50.00%	99.97%
图（d）	20.740	8	100.00%	99.92%
图（e）	20.276	5	100.00%	99.88%
图（f）	22.334	7	87.50%	95.87%
图（g）	19.964	8	100.00%	99.64%
图（h）	27.299	38	54.29%	99.03%
图（i）	34.207	8	80.00%	99.96%
图（j）	31.637	22	68.75%	99.87%

4 结论

随着深度学习在计算机视觉领域的应用日益深入，其逐渐成为人脸检测的主流方案。本文聚焦YOLOv10和Faster R-CNN两种算法，构建综合数据集开展对比研究，通过PyTorch框架开发集成界面实现功能可视化。实验表明，单阶段YOLOv10以简洁计算流程实现实时检测优势，适用于响应速度场景；两阶段Faster R-CNN借精细架构在复杂场景下实现高精度检测。模型层面，前者轻量易部署，后者虽资源消耗大但泛化能力强。后续可在复杂场景鲁棒性与模型轻量化方向深化研究。