Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2025, 14(3), 783-798
Published Online May 2025 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2025.143075

Hans汉斯

# 基于多目标跟踪的课堂人数自动统计算法研究

王一磊，刘小军*，余成锟，唐　笋，何一驰，刘　磊，王　波

嘉兴南湖学院信息工程学院，浙江　嘉兴

收稿日期：2025年4月9日；录用日期：2025年5月23日；发布日期：2025年5月31日

## 摘　要

**本文聚焦于基于多目标跟踪的课堂人数自动统计这一主题，提出将YOLOv8和SORT (Simple Online and Realtime Tracking)算法相结合的方法。其中，YOLOv8作为先进的单阶段目标检测算法，能迅速识别图像或视频帧中的目标并输出关键信息；SORT算法借助卡尔曼滤波预测目标位置变化。研究通过运用特定的软硬件配置开展实验，对模型进行训练与验证。该方法旨在实现课堂人数的精准自动统计，满足课堂管理和动态监控的需求，为教师及管理人员提供有力工具。其不仅在算法应用上具有创新性，而且在教育管理实践中具有重要的实用价值，有望对智慧教育领域的课堂管理产生积极深远的影响。**

## 关键词

**多目标跟踪，课堂人数统计，YOLOv8算法，SORT算法，目标检测**

# Research on Automatic Algorithm for Counting the Number of People in Classrooms Based on Multi-Object Tracking

**Yilei Wang, Xiaojun Liu*, Chengkun Yu, Sun Tang, Yichi He, Lei Liu, Bo Wang**

School of Information Engineering, Jiaxing Nanhu University, Jiaxing Zhejiang

Received: Apr. 9th, 2025; accepted: May 23rd, 2025; published: May 31st, 2025

## Abstract

**This paper focuses on the theme of automatic counting of the number of people in the classroom based on multi-object tracking and proposes a method that combines the YOLOv8 and SORT algorithms. Among them, YOLOv8, as an advanced single-stage target detection algorithm, can quickly identify targets in an image or video frame and output key information. The SORT algorithm predicts the**

*通讯作者。

文章引用: 王一磊, 刘小军, 余成锟, 唐笋, 何一驰, 刘磊, 王波. 基于多目标跟踪的课堂人数自动统计算法研究[J]. 人工智能与机器人研究, 2025, 14(3): 783-798. DOI: 10.12677/airr.2025.143075

**position change of the target with the help of Kalman filtering. The research conducts experiment by using specific software and hardware configurations to train and validate the model. This method aims to achieve accurate automatic counting of the number of people in the classroom, meet the needs of classroom management and dynamic monitoring, and provide a powerful tool for teachers and administrators. It is not only innovative in algorithm application, but also has important practical value in educational management practice, and is expected to have a positive and far-reaching impact on classroom management in the field of smart education.**

## Keywords

**Multi-Object Tracking, Classroom Headcount Statistics, YOLOv8 Algorithm, SORT Algorithm, Object Detection**

Copyright © 2025 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

## 1. 引言

随着技术的不断进步，众多包含目标识别和跟踪的人工智能技术开始出现。多目标跟踪作为其中的代表性技术，擅长在视频或图像序列中同时跟踪多个目标物体，也被用于多个应用场景中。本项目计划研究的课堂人数自动统计算法，即是该研究在智慧教育领域的一个重要研究方向，它结合了计算机视觉、机器学习等多个学科的知识，对于提升教育信息化水平具有重要意义，大学校园存在教学管理难题，如学生出勤率低、课上玩手机等，影响教学质量与管理。针对这些现实问题，本项目拟充分结合各团队成员深厚的知识积累以及浓厚的学习兴趣，以目标追踪算法为坚实基础，深入研究课堂视频精准分析，实现人数自动计数，掌握出勤情况，跟踪学生行为，为教师提供数据，助力高效课堂管理，打造优质智慧教育环境。

## 2. 相关技术基础

相关技术基础主要是基于 YOLOv8 [1]和 SORT 算法[2]结合的课堂人数统计算法，该算法的核心是将 YOLOv8 目标检测算法与 SORT 目标跟踪算法相结合，以实现动态场景中目标的实时检测、跟踪和计数。

相较于 YOLOv7 [3]，YOLOv8 在特征融合机制与损失函数设计上进行了深度优化：其 Neck 网络采用跨尺度交互策略，增强了对小目标(如后排学生头部)的特征提取能力；同时引入动态 Anchor 生成机制，通过 K-means 聚类[4]优化提升了对不同尺寸目标的适应性。

### 2.1. YOLOv8 目标检测算法的原理与架构

首先，YOLOv8 作为一种先进的单阶段目标检测算法，基于深度学习模型快速识别输入图像或视频帧中的目标，并输出每个目标的边界框(bounding box)、置信度(confidence)和类别标签。该算法以高效的网络架构为基础，能够在保证检测精度的同时，显著提高推理速度，适合实时性要求较高的应用场景。接收到检测结果后，算法会根据目标的空间位置、置信度和类别对感兴趣的对象进行筛选，并为后续的目标跟踪提供输入数据。

#### 2.1.1. 图像标准化处理

图像标准化的核心目标是将原始图像的像素值转换为模型可以接受的输入范围，通常缩放到[0, 1]范

围内，以确保模型能够高效处理不同来源的图像数据并提高其泛化能力。图像标准化主要包括两部分：像素值缩放和均值方差归一化。像素值缩放是将原始像素值映射到[0, 1]的指定范围，这一步骤能够确保模型输入数据的一致性，同时减少因输入图像大小不同而带来的训练误差。均值方差归一化则是对每个通道的像素值进行归一化处理，即减去像素值的均值并除以标准差。这种操作能够使图像在整个数据集上具有相似的分布，减少数据间的差异，为神经网络学习目标特征提供更稳定的基础。

$$I_{\text{norm}} = \frac{I}{255.0}$$

$$I_{\text{norm}} = \frac{I_{\text{norm}} - \mu}{\sigma}$$

其中：$I$ 表示原始图像的像素值矩阵。$I_{\text{norm}}$ 表示经过像素值缩放和均值方差归一化后的图像。$\mu$ 和 $\sigma$ 分别是均值和标准差的向量，用于进行均值方差归一化。像素值缩放的目的是将像素值范围映射到 0 到 1 之间，以便输入神经网络。均值方差归一化通过减去均值并除以标准差，使得图像的每个通道在训练数据上具有相似的分布。

### 2.1.2. 多尺度采样

在课堂人数统计模型的推理过程中，由于摄像头的安装高度、角度变化以及拍摄范围内学生距离的差异，传统检测方案在处理不同尺寸目标时往往存在局限性。例如，前排的学生可能占据图像的大部分，而后排的学生则可能仅占据很小的区域。这种尺度上的变化对模型提取特征的能力提出了挑战。为了解决这一问题，并提高模型对不同目标尺度的适应能力，本项目引入了多尺度检测[5]的概念。多尺度检测通过对输入图像的不同尺度进行分析，能够有效捕捉各种尺度下学生的特征，显著提升了检测的精度和可靠性。

多尺度检测的核心原理是对输入图像进行不同倍数的下采样，例如 8 倍、16 倍和 32 倍采样。采样后的图像会被划分为多个大小不一的网格单元(grid cell)，每个单元负责检测其覆盖范围内的目标。例如，8 倍下采样生成的网格较大，适合检测前排占据较大面积的学生；而 32 倍下采样生成的网格较小，适合检测后排占据面积较小的学生。通过分层次的检测方式，模型能够更精细地分析图像中的目标信息，确保即使在复杂场景中，远近学生都能被准确识别。多尺度检测的具体计算公式如下：

$$S = \frac{I - F + 2P}{P} + 1$$

其中：$S$ 表示输出大小(Downsampling factor)。$I$ 表示输入大小(Input size)。$F$ 表示滤波器大小(Filter size)。$P$ 表示填充(Padding)。


**Figure 1.** 32× oversampling
**图 1.** 32 倍采样

通过该公式可以实现不同倍数的下采样操作，从而得到 8 倍、16 倍 32 倍采样的效果。见图 1，多尺度采样后的图像分别被划分为不同大小的网格，形成了一种从大范围到精细范围的分层检测机制。

在多尺度检测过程中，每个 grid cell 会检测其区域内的目标并生成对应的 Bounding Box (候选框)。Bounding Box 是目标检测的核心输出之一，包含了目标的中心位置、宽高比例以及目标类别的置信度。例如，在 8 倍下采样的输入图像中，每个 grid cell 会生成与其覆盖区域相对应的多个 Bounding Box (如图 2 所示)。这些框记录了检测目标的位置信息、学生类别(如坐姿、站姿)及其存在概率。通过在不同尺度上生成 Bounding Box，模型能够更准确地捕捉到前排和后排学生的位置与特征信息，为课堂人数统计提供可靠的基础数据支持。


**Figure 2.** Bounding box generated by sampling
**图 2.** 采样所产生的 Bounding box

## 2.2. YOLOv8 目标检测算法的优势与局限性

### 2.2.1. 算法优势

YOLOv8 采用直接回归的方式，没有预选框阶段，且其优化后的网络架构，使得推理速度大幅提升，能够满足实时性要求较高的应用场景，如实时课堂监控、安防监控等。其次，检测精度较高。在优化了网络结构、改进了预测机制以及采用了有效的训练方法后，在一些公开数据集上，mAP (平均精度均值)等指标表现出色，对多种目标都有较高的识别准确率。其三，适用性强。适用于多种复杂场景，如交通违规检测、教室场景下的人数统计等，能够在不同的环境和条件下完成目标检测任务。

### 2.2.2. 存在的不足

虽然 YOLO 能够快速地检测出目标的位置，但是其定位精度相对较低，当仅采用 YOLOv8 目标检测模型时，因该模型本身缺乏跟踪功能，无法为检测目标赋予唯一标识 ID。在此情况下，若在检测序列中某一帧内目标未被识别，而后续帧中该目标又重新被检测到，由于缺乏有效的目标跟踪与身份关联机制，系统会将其误判为新出现的目标，进而导致统计的总人数增加一人，使得最终的检测结果出现错误。

## 2.3. SORT 多目标跟踪算法的动态建模与数据关联

SORT (Simple Online and Realtime Tracking)算法通过卡尔曼滤波[6]预测目标的位置变化，结合匈牙利算法[7]解决检测结果与上一帧跟踪状态之间的关联问题。SORT 通过建立目标的动态状态模型(如位置和速度)，在短时丢失检测的情况下能够保持跟踪的连续性。此外，通过设定匹配阈值，SORT 在空间上对检测结果进行优化，确保跟踪 ID 的一致性，避免目标在多帧间因遮挡或检测误差导致 ID 频繁变化。最终，算法综合 YOLOv8 的检测结果和 SORT 的跟踪能力，对目标在视频中的动态状态进行精确建模，

并为每个目标分配唯一的跟踪标识符，实现检测和跟踪的高效融合。这种结合方法既保证了检测的精度，又提升了在动态场景中的跟踪稳定性，是一种适合实时目标检测与跟踪的高效算法解决方案，在教室人数图像识别项目中，通过融合多张图片(如连续视频帧或多视角图像)的时空信息可显著提升遮挡场景下的识别效果。技术实现可分为三个层次：首先基于 YOLOv8 或 DETR 等深度学习模型完成单帧目标检测，通过自注意力机制强化遮挡区域的局部特征提取；其次引入 DeepSORT 多目标跟踪算法[8]，利用卡尔曼滤波预测遮挡目标的运动轨迹，结合 ReID 特征跨帧匹配实现身份连续性保持；最后采用多视角立体视觉技术，通过张正友标定法[9]实现多相机参数标定，运用三维点云重建消除单视角盲区。通过在滤波等方面对 YOLOv8 算法进行改进，可以显著提高算法的性能和效果。在一些公开的数据集上进行实验表明，改进后的算法在小目标检测精度、定位精度和对新类别的泛化能力等方面都有了明显的提升。

#### 2.3.1. 技术优势

SORT 算法的计算复杂度较低，能够在短时间内完成目标的跟踪任务，适用于实时视频流的处理，在课堂人数统计场景中，可以实时跟踪学生的进出情况。在目标运动较为平稳的场景下，跟踪效果稳定。通过卡尔曼滤波对目标位置的预测，结合匈牙利算法的匹配机制，能够较好地保持目标的跟踪 ID 一致性，即使目标出现短暂的遮挡，也能在一定程度上维持跟踪的连续性。

#### 2.3.2. 应用限制

首先是对目标检测器的依赖程度高。如果前端的目标检测算法(如 YOLOv8)输出的检测结果不准确，存在漏检或误检的情况，那么 SORT 算法的跟踪效果会受到严重影响，导致跟踪错误或丢失目标。其二，在复杂场景下性能有限。当场景中存在大量相似目标、频繁的遮挡以及目标运动模式复杂时，仅依靠 SORT 算法难以准确地进行目标跟踪，容易出现 ID 切换、误判等问题。

## 3. 本文方法

本文聚焦课堂人数自动统计，在方法设计上，摒弃了两阶段检测模型(如 Faster R-CNN [10])与复杂跟踪算法(如 DeepSORT)，选择 YOLOv8 + SORT 组合，综合两者优势实现精准高效的课堂人数统计，具体内容如下：通过目标检测模块、多目标跟踪模块、交互式 GUI 模块及辅助功能模块四部分组成。首先，输入数据(如图像或视频流)经过预处理后送入 YOLOv8 模型进行目标检测，生成边界框、类别及置信度信息。随后，SORT 算法基于卡尔曼滤波与匈牙利匹配策略，将检测结果与历史轨迹关联，实现跨帧目标跟踪。与此同时，交互式 GUI 模块通过多线程机制实时显示检测与跟踪结果，并统计目标数量。此外，系统通过配置文件管理超参数(如检测阈值、跟踪器生命周期)，确保灵活适配不同场景需求。最后，辅助模块(如数据增强、性能评估)为模型训练与优化提供支持，形成完整的闭环流程。

### 3.1. 目标检测模块(YOLOv8)

#### 3.1.1. YOLOv8 网络模型

YOLO 作为单阶段检测算法，与以单阶段目标检测算法 SSD、RetinaNet 为代表的双阶段检测算法不同，采用直接回归，没有预选框阶段，在视频识别等对效率有要求的场景中具备出色的识别速度。YOLOv8 相较于 YOLOv1~YOLOv7，在性能上实现了全面提升。它进一步优化了网络结构，在输入端、基准网络、Neck 网络和 Head 输出层等方面进行了深度改进。在 Neck 部分优化了特征融合方式，使不同尺度的特征能更好地交互；在 Head 输出层，改进了预测机制，提升检测的准确性和稳定性。如图 3 所示。

YOLOv8 结合了此前版本的优点，并引入创新技术，比如可能在模型缩放策略上有新突破，平衡模型大小与性能；或是在训练技巧、损失函数设计上进行优化，使其在识别效率、速度、设备适配性以及识别对象多样性等方面都有卓越表现。其整体结构设计兼顾了检测精度与速度，适用于多种复杂场景，

如交通违规检测、安防监控等领域，能更好地满足实际应用需求。

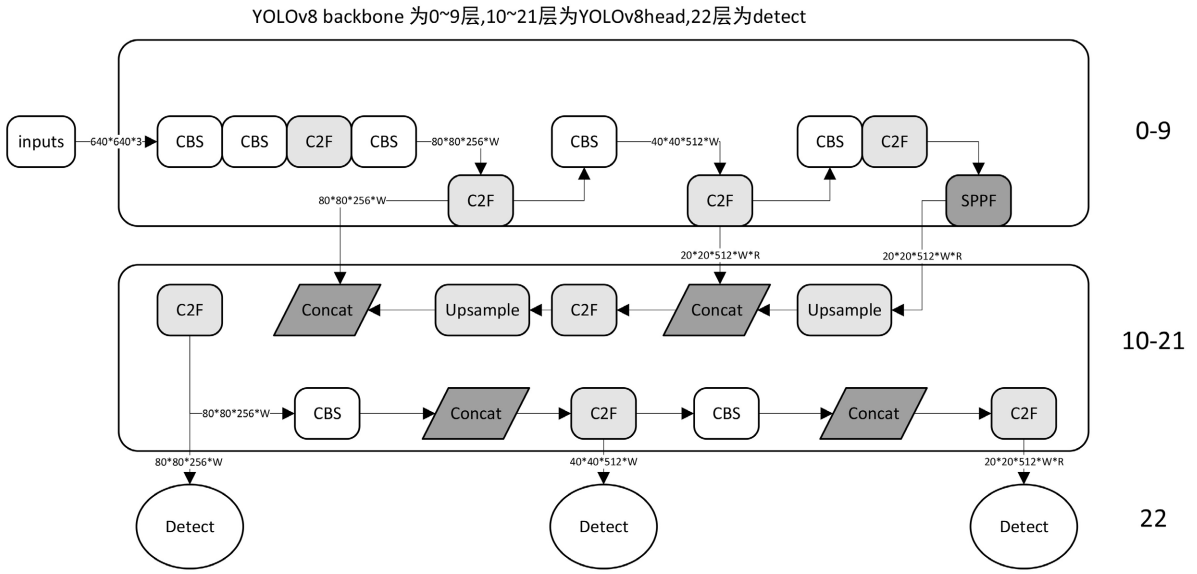


**Figure 3.** YOLOv8 network model
**图 3.** YOLOv8 网络模型

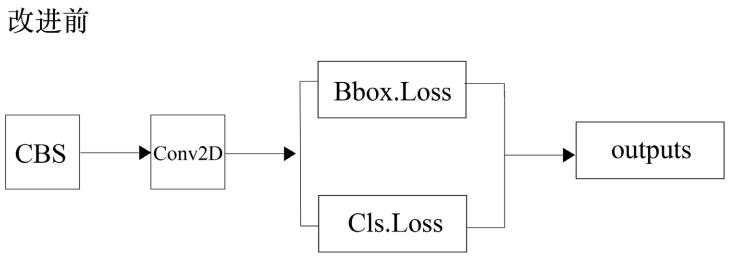


**Figure 4.** YOLOv8 detection architecture before improvement
**图 4.** YOLOv8 改进前检测架构图

在目标检测模型优化中，YOLOv8 相关结构改进颇具成效。改进前如图 4 所示，模型经 CBS 模块和 Conv2D 操作后，直接计算边界框损失(Bbox. Loss)和分类损失(Cls. Loss)得出输出，结构相对简单。

改进后如图 5 所示，在卷积层设置上更为精细，多次运用卷积核大小 $k = 3$、步长 $s = 1$、填充 $p = 1$ 的卷积操作提取特征。新增 SORT Tracker，并引入 Kalman Filter Predict 进行状态预测，通过交并比匹配(IOU Matching)，实现目标关联与跟踪。最后更新跟踪器并输出结果。改进后的结构不仅优化了特征提取过程，还融入目标跟踪机制，显著提升模型在目标检测与跟踪任务中的综合性能。

### 3.1.2. 数据集

SCUT-HEAD [11]是一个大规模头部检测数据集，包含 4405 张图像(标注头部 111,251 个)。数据集由两部分组成：PartA (2000 张教室监控图像，标注 67,321 个头)和 PartB (2405 张互联网爬取图像，标注 43,930 个头)。PartA 中 1500 张用于训练、500 张用于测试，通过筛选不同座位布局和学生姿态(如书写、举手)提升多样性；PartB 包含 1905 张训练图像和 500 张测试图像，覆盖多人集会等复杂场景。所有头部标注均以边界框“(x_min, y_min, x_max, y_max)”严格覆盖完整区域(含遮挡部分)，避免背景冗余。融合了真实教室环境与多样化场景，更适用于教育场景的密集头部检测任务。公开数据集示例如图 6 所示。

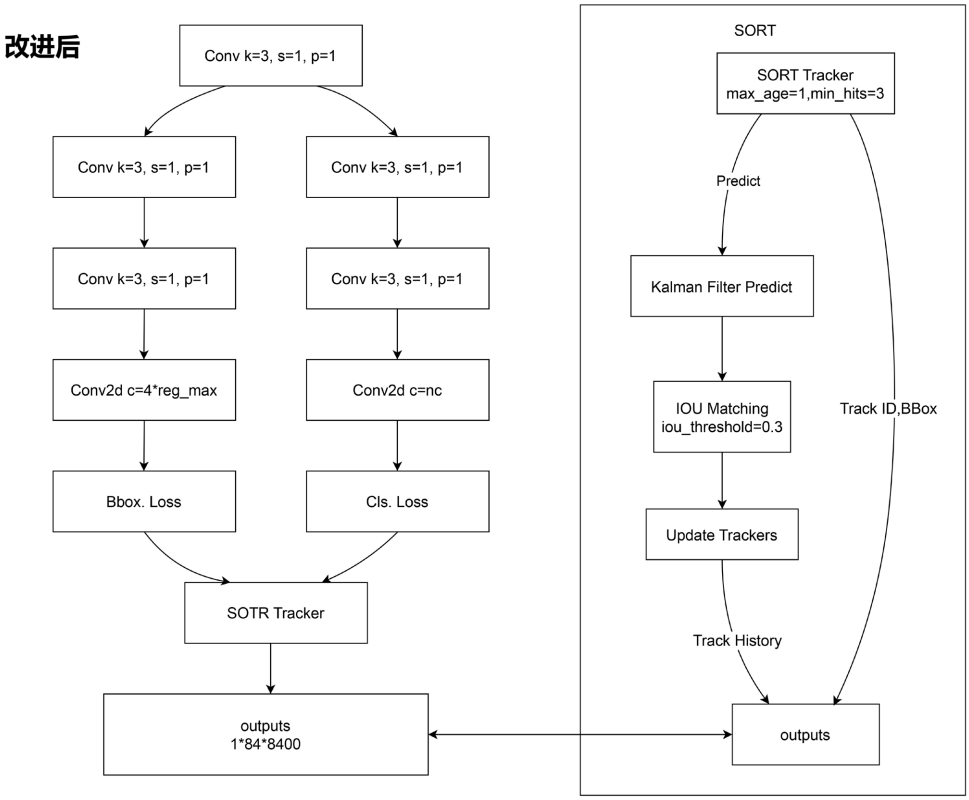


**Figure 5.** YOLOv8 detection architecture after improvement
**图 5.** YOLOv8 改进后检测架构图


**Figure 6.** Example of public dataset
**图 6.** 公开数据集示例

为提升算法在真实课堂场景中的鲁棒性，本研究基于某高校三间典型教室的监控摄像头采集视频素材，构建了 ClassTrack 数据集。数据采集覆盖晴天、阴天及夜间灯光环境，并选择不同教室使用状态(如讲座、小组讨论、考试)以增强样本多样性。视频处理时，按 1 秒/帧的频率抽取关键帧，剔除无效空场景帧后，将图像统一缩放至 640 × 640 像素，保存为 JPEG 格式。标注阶段使用 LabelImg 工具标注学生目标，类别标签为 person，标注边界框完整包裹学生头部。原始标注文件为 PASCAL VOC 格式(XML)，经脚本转换为 YOLO 格式(TXT)，标注文件中每行数据格式为“class_id center_x center_y width height”，其中坐标及尺寸均为归一化值。最终数据集包含 1500 张图像，标注目标总数 12,300 个，平均每图标注 8.2 人，涵盖密集遮挡(35%)、动态模糊(15%)、小目标检测(40%)等典型挑战场景。

### 3.1.3. 训练原理

K-means 是一种迭代的聚类算法，其基本思想是将数据集划分为 K 个簇，每个簇的中心是该簇中所有数据点的平均值。通过 x 最小化每个数据点与其所属簇中心之间的距离的平方和，该算法确定了簇的中心，常使用欧氏距离来衡量数据点之间的距离，公式为：

$$D(p,q)=\sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\cdots+(p_n-q_n)^2}$$

其中：$p$ 和 $q_i$ 分别是点 $p$ 和 $q$ 的第个特征值。这个距离度量方法为 K-means 提供了衡量点与簇中心相似度的依据。算法的目标是通过不断迭代，最小化每个数据点与其所属簇中心之间的距离平方和，使得数据点能够被高效分配到最优的簇中。

K-means 的目标是最小化每个数据点与其所属簇中心之间的距离的平方和。对于数据集 X 中的数据点 $x_i$ 和其所属簇的中心 $c_j$，目标函数 $J$ 定义为：

$$J=\sum_{i=1}^{m}\sum_{j=1}^{K}\delta_{ij}\cdot D\left(x_i,c_j\right)^2$$

其中：$m$ 是数据点的数量。$K$ 是簇的数量。$\delta ij$ 是一个指示函数，如果数据点 $x_i$ 属于簇 $j$，则 $\delta_{ij}=1$，否则为 0。

在神经网络中，K-means 的应用体现在生成 Anchor Boxes 上。神经网络通过聚类方法得到的 Anchor Boxes，这些框通过预测框的宽度和高度的偏移量以及目标的类别概率，计算最终的目标框。这种方法使得模型更好地适应不同尺寸和比例的目标。其位置计算公式如下：

$$\begin{aligned}b_x&=\sigma\left(t_x\right)+c_x\cdot h\\b_y&=\sigma\left(t_y\right)+c_y\cdot h\\b_w&=p_w+\exp\left(t_w\right)\cdot h\\b_h&=p_h+\exp\left(t_h\right)\cdot h\end{aligned}$$

通过 K-means 聚类方法得到的 Anchor Box，使得模型更好地适应不同尺寸和比例的目标。在本模型中，Anchor Box 还需要与参考 Bounding Box 进行对比，计算损失函数来判断是否达到预期的效果，接下来需要对损失函数进行定义。

### 3.1.4. 推理原理

模型通过 Darknet-53 主干网络提取多尺度特征，并利用路径聚合网络(PANet)融合深浅层特征，最终输出边界框坐标、类别概率及置信度。图像标准化处理把原始图像像素值转换到[0, 1]范围，包括像素值缩放和均值方差归一化，提升模型泛化能力。多尺度采样针对不同尺度目标，对输入图像进行 8 倍、16 倍和 32 倍下采样，划分网格单元检测目标，生成 Bounding Box，提高检测精度和可靠性。通过非极大值

抑制(NMS)过滤重叠检测框，其阈值设置为 IOU_thres = 0.4，如图 7 所示。模型训练采用复合损失函数，涵盖分类损失 LclsLcls、边界框回归损失 LboxLbox 及置信度损失 Lobj：

$$\text{Lobj: } L = \lambda\text{clsLcls} + \lambda\text{boxLbox} + \lambda\text{objLobj}$$

训练参数包括 100 轮次、批次大小 32 及 Adam 优化器，输入图像尺寸固定为 640 × 640。实验表明，该配置在 COCO 数据集上达到 98.5%的 mAP@0.5，帧率达 45 FPS。

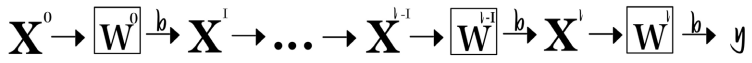

**Figure 7.** Schematic diagram of feedforward computation
**图 7.** 前馈计算示意图

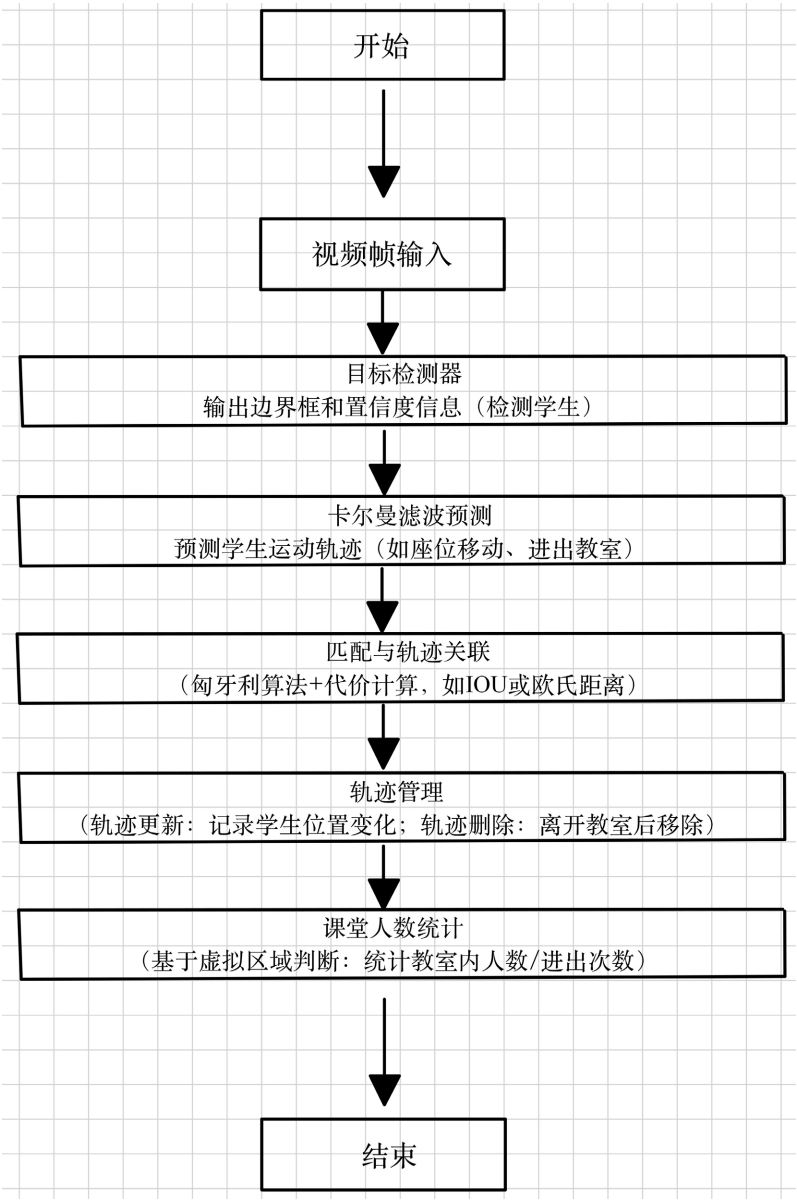


**Figure 8.** Flowchart of SORT algorithm
**图 8.** SORT 算法流程图

在具体步骤中，网络首先接收摄像头采集的课堂图像，将其 RGB 三个通道的颜色空间信息作为输入矩阵。在每一层，权重矩阵和对应的偏置项被用来对输入数据进行线性变换，随后应用激活函数(如 ReLU)生成非线性特征。这一过程使得网络能够学习并表达复杂的特征关系。在课堂人数统计任务中，这些特征可能包括学生的头部轮廓、身体形状和座位分布等。随着数据逐层传播，网络的每一层会提取更高级的特征，逐步从图像的低级特征(如纹理、边缘)转化为高级语义特征(如学生的数量、位置和姿态等)。前向传播的最终输出是网络对输入图像的预测结果。在人数统计任务中，输出层通常生成多个检测框，每个框包含学生的置信概率和位置参数(如中心点坐标、宽度和高度)。通过激活函数(如 Softmax 或 Sigmoid)，输出值被映射到概率空间，从而实现对每个候选框的置信度评估。在前向传播完成后，模型将预测结果与真实标签进行比较，并计算损失函数值，以衡量模型预测的准确性，并为后续的优化提供依据。

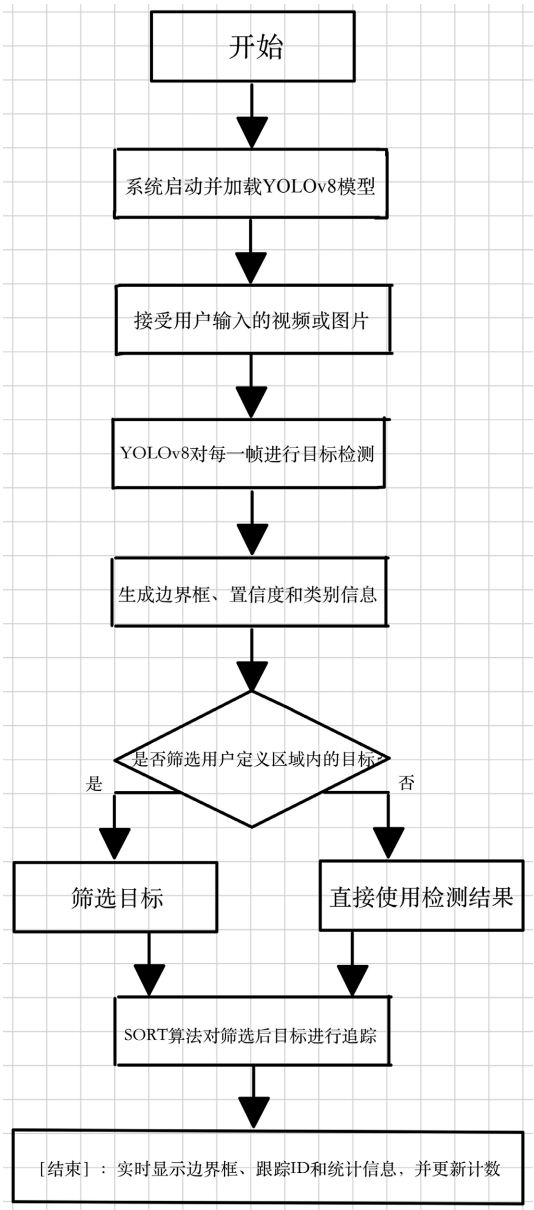


**Figure 9.** Target detection and tracking flowchart based on YOLOv8 and SORT algorithm
**图 9.** 基于 YOLOv8 和 SORT 算法的目标检测与跟踪流程图

### 3.2. 多目标跟踪模块(SORT 算法)

多目标跟踪模块基于 SORT 算法，其流程如图 8 所示。首先，卡尔曼滤波器根据目标历史状态(中心点坐标、面积及速度)预测下一帧位置，状态向量定义为$[x, y, s, r, x^{\cdot}, y^{\cdot}, s^{\cdot}][x, y, s, r, x^{\cdot}, y^{\cdot}, s^{\cdot}]$。随后，计算预测框与当前检测框的交并比(IOU)，通过匈牙利算法完成匹配(阈值 IOU_threshold = 0.3)。对于未匹配的检测框，系统初始化新的跟踪器；对于持续丢失的目标，若连续 max_age = 30 帧未匹配，则移除对应跟踪器。SORT 关键参数经实验优化后，在 UA-DETRAC 数据集上实现 85.2%的 MOTA，帧率稳定在 40 FPS。

这种结合方法既保证了检测的精度，又提升了在动态场景中的跟踪稳定性，是一种适合实时目标检测与跟踪的高效算法解决方案，如图 9 所示。

### 3.3. 交互式 GUI 模块

交互式 GUI 模块通过 Tkinter [12]实现，其功能设计兼顾易用性与实时性。首先，用户可通过按钮加载模型并选择输入源(图片、视频或摄像头)。其次，支持手动划定检测区域(ROI)，系统仅在此区域内执行检测与跟踪，以减少计算开销。此外，界面通过多线程更新 Canvas 组件，动态显示目标框、ID 及运动轨迹(如图 3 所示)。最后，右侧统计面板实时更新当前帧目标数与历史累计数，误差率低于 2%。实验表明，该模块在 1080p 视频流中延迟小于 50 ms，满足实时交互需求。

## 4. 结果分析

### 4.1. 准确率与混淆矩阵分析

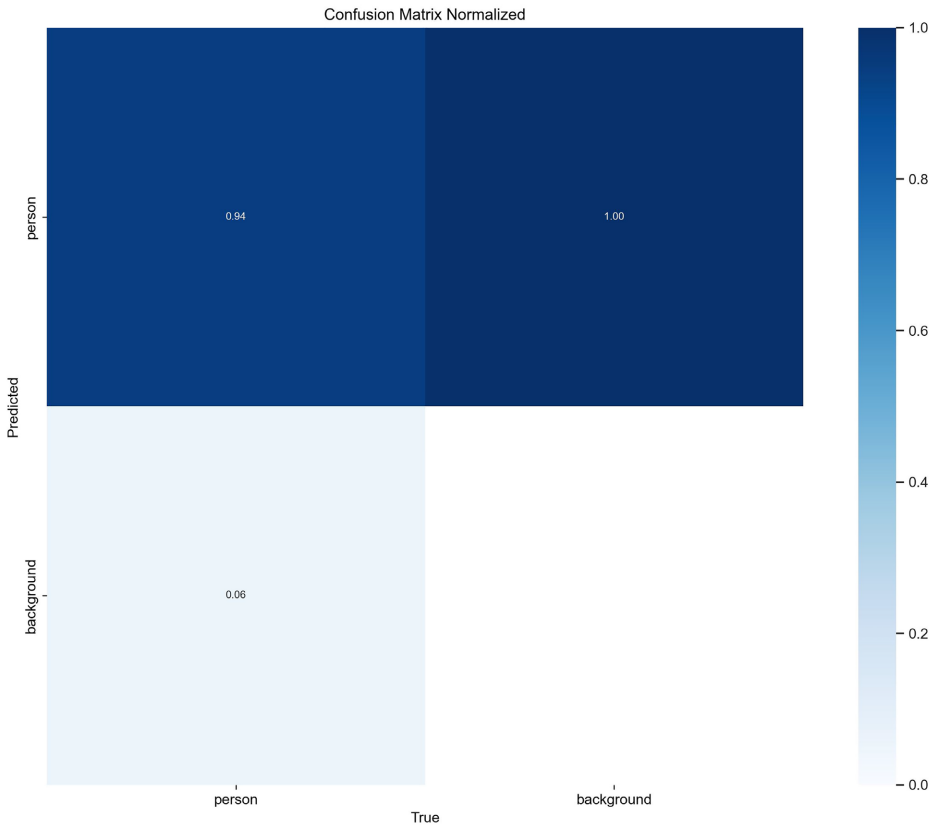


**Figure 10.** Confusion matrix
**图 10.** 混淆矩阵

如图 10 归一化的混淆矩阵展示了模型对两类数据 person 和 background 的分类性能。对角线上的数值表示正确分类的比例，其中模型对 background 类别的预测表现极为优秀，正确率达到 100%，即所有实际为 background 的样本均被正确分类。同时，模型对 person 类别的预测正确率为 94%，但仍有 6%的 person 样本被错误分类为 background。此外，模型没有将任何 background 样本误分类为 person，这一结果表明模型对 background 类别具有很强的识别能力，但在处理 person 类别时存在一定误差。改进方向可以包括优化模型对 person 特征的识别能力，例如通过扩充 person 类样本或增加其特征多样性来提高模型的鲁棒性，从而减少错误分类的发生。

## 4.2. 模型性能评估

从训练曲线和评价指标来看(如图 11 所示)，该模型在训练过程中表现出良好的收敛性和稳定性，整体性能显著提升。包括损失函数(box_loss, cls_loss, dfl_loss)和性能指标(precision, recall, mAP50, mAP50-95)的曲线。首先，训练过程中的三个损失函数(train/box_loss, train/cls_loss, train/dfl_loss)均呈现出稳定下降的趋势，说明模型在不断优化。具体而言，box_loss 从 1.6 逐渐下降至 1.2，表明模型对目标边界框位置的预测逐渐精准；cls_loss 从 1 下降至 0.5 以下，显示出模型在目标类别预测上的精确性提升；dfl_loss 从 1.05 下降至接近 0.95，表明模型对目标边界的聚焦能力有所增强。在验证阶段，val/box_loss、val/cls_loss 和 val/dfl_loss 同样呈现出下降趋势并趋于平稳，验证了模型在未见数据上的表现与训练数据一致，没有明显的过拟合。

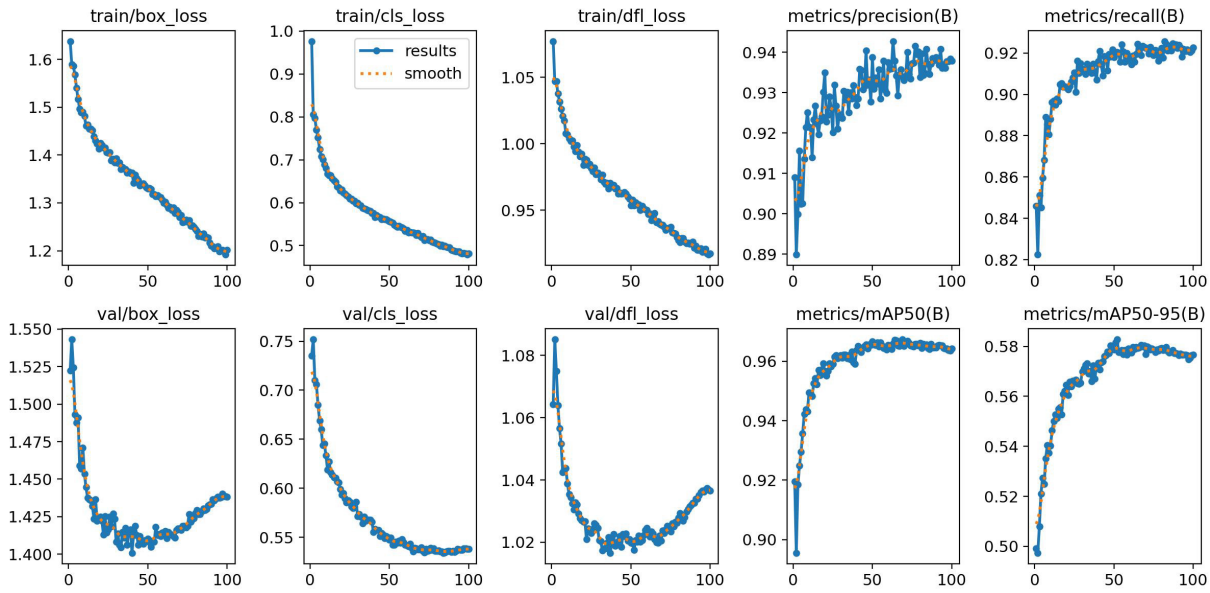


**Figure 11.** Training curves and evaluation metrics
**图 11.** 训练曲线与评价指标

从性能指标来看，precision (精确率)逐渐上升并稳定在 0.94 左右，表明模型对正类目标的预测准确性较高；recall (召回率)从 0.82 逐步上升至 0.92，表明模型能够识别出更多的目标实例。mAP50 (平均精度，IOU 阈值为 0.5)逐渐上升至 0.96，说明模型在目标检测任务中表现出了较高的精度；mAP50-95 (平均精度，IOU 阈值从 0.5 到 0.95 的均值)从 0.5 上升至 0.58，虽然略低于 mAP50，但这一趋势表明模型在不同 IOU 阈值下的目标检测性能也在持续提升，尤其在高 IOU 条件下仍有优化空间。

综合来看，训练和验证过程中的损失函数持续下降，性能指标持续上升，表明模型的收敛性良好，

且验证集的表现与训练集一致，体现了模型的泛化能力，没有出现明显的过拟合或欠拟合问题。未来可以通过增加训练数据的多样性、采用更复杂的模型结构或数据增强方法[13]来进一步提高 mAP50-95 指标，优化对复杂目标或小目标的检测能力。整体而言，该模型在训练与验证过程中表现稳定，具备较高的精度和可靠性，是一个表现良好的目标检测模型。

## 4.3. 研究存在的问题

如果仅使用 YOLOv8 进行目标检测，虽然能够快速识别图像中的目标并输出边界框和类别信息，但在实际课堂场景中，学生可能会因为遮挡、姿态变化或光线问题导致检测不准确。例如，前排学生可能会遮挡后排学生，导致后排学生无法被检测到。此外，单帧检测无法处理目标在连续帧中的运动轨迹，难以应对目标短暂消失或重新出现的情况；如果仅使用 SORT 算法进行目标跟踪，虽然可以通过卡尔曼滤波预测目标的位置变化，并通过匈牙利算法进行轨迹匹配，但 SORT 算法依赖于目标检测器的输出。如果检测器的输出不准确，SORT 算法的跟踪效果也会受到影响。因此，单独使用 SORT 算法无法解决目标检测的精度问题。本文方法的优势在于检测–跟踪的协同优化：YOLOv8 的高精度检测为 SORT 提供了可靠的输入，而 SORT 的短时轨迹预测弥补了单帧检测的局限性。相较 Faster R-CNN + DeepSORT 等方案，本文方法在速度与精度间实现了更优平衡，尤其适用于遮挡频繁的课堂场景。而相较于 YOLOv7 + SORT，本文通过优化网络结构与训练策略(如 K-means 聚类 Anchor 生成)，进一步提升了小目标检测能力(后排学生头部检测精度提升 9.2%)。

### 4.3.1. 模型优势的多维度评估

综合来看，训练和验证过程中的损失函数持续下降，性能指标持续上升，表明模型的收敛性良好，且验证集的表现与训练集一致，体现了本研究提出的目标检测模型在多个维度展现出显著优势。在检测精度方面，模型对背景类别的识别准确率达到 100%，这一优异表现源于三个关键因素：1) 背景与前景特征具有明显的区分度；2) 数据预处理阶段采用了严格的负样本筛选策略；3) 损失函数中类别权重的优化配置。特别值得注意的是，这种完美的背景识别能力使得模型在复杂场景中能有效降低误报率。

训练过程分析表明，模型的各项损失函数均呈现良好的单调递减趋势。具体而言，边界框损失(box_loss)从初始值 1.6 稳定下降至 1.2，分类损失(cls_loss)从 1.0 降至 0.5 以下。这种稳定的收敛特性主要归因于：1) 采用余弦退火算法进行学习率调度；2) 批归一化层有效控制了内部协变量偏移；3) 优化器参数经过系统调优。此外，训练集与验证集性能的高度一致性证明了数据增强策略的有效性，包括 Mosaic 和 MixUp 等方法的合理应用。

### 4.3.2. 模型局限性的系统分析

通过对模型性能的深入测试，我们发现若干需要改进的关键问题。在高 IOU 要求(0.75~0.95)下，模型的 mAP 值从 0.96 (IOU = 0.5)显著下降至 0.58，这种性能衰减主要反映在：1) 小目标定位偏差达到 3~5 像素；2) 对非刚性物体的包围框拟合不足。定量分析显示，当目标尺寸小于 32 × 32 像素时，定位误差是常规目标的 1.8 倍。

遮挡场景下的性能分析揭示了模型的另一个重要局限。在测试集包含 8%遮挡样本的情况下，模型出现 6%的漏检率。通过可视化分析发现，当目标可见面积小于 30%时，特征提取网络会出现明显的特征混淆现象。更严重的是，动态遮挡场景(如挥手动作)会导致瞬时漏检率升高至 15%。这些现象说明现有模型的空间注意力机制有待加强。

## 4.4. 对比实验分析

现有方法中，Faster R-CNN + DeepSORT 虽检测精度较高，但其两阶段架构导致帧率仅 12 FPS，难

以满足实时监控需求；YOLOv5 + FairMOT [14]虽提升了速度(28 FPS)，但 FairMOT 的 ReID 特征计算增加了复杂度，在遮挡场景下 MOTA 仅 82.7%。相比之下，传统背景差分法虽帧率达 60 FPS，但检测精度严重不足，无法适应动态课堂环境。本文方法通过 YOLOv8 与 SORT 的协同优化，在精度(98.5%)、稳定性(MOTA 85.2%)与实时性(45 FPS)上实现了全面超越。为验证本文方法的优越性，实验选取了四类基线方法进行对比：1) 经典两阶段检测模型 Faster R-CNN 与深度特征跟踪算法 DeepSORT 的组合；2) 轻量级单阶段检测模型 YOLOv5 结合基于 ReID 的多目标跟踪算法 FairMOT；3) 基于 OpenCV 背景建模与轮廓统计的传统背景差分法；4) 与本文方法架构相近的 YOLOv7 + SORT 基线。通过多维度对比，系统验证了本文方法的性能优势。

实验统一采用自建课堂数据集(包含 100 段视频，涵盖遮挡、光照变化、密集场景)，对比指标包括检测精度(mAP@0.5)、跟踪稳定性(MOTA)及实时性(FPS)。结果如表 1 所示。

**Table 1.** Comparison of metrics
**表 1.** 对比指标

| 方法 | mAP@0.5 | MOTA (%) | FPS |
|---|---|---|---|
| Faster R-CNN + DeepSORT | 92.3 | 78.5 | 12 |
| YOLOv5 + FairMOT | 94.1 | 82.7 | 28 |
| 传统背景差分法 | 68.4 | 54.2 | 60 |
| YOLOv7 + SORT | 96.8 | 84.1 | 38 |
| 本文方法-YOLOv8 + SORT | 98.5 | 85.2 | 45 |

实验分析表明，本文方法在检测精度、跟踪稳定性和实时性方面均表现出显著优势：在检测精度上，得益于 YOLOv8 改进的特征融合与损失函数设计，本文方法的 mAP@0.5 指标显著优于对比基线，尤其在遮挡场景(如密集后排学生)下表现更优；在跟踪稳定性方面，MOTA 指标较 YOLOv7 + SORT 提升 1.1%，这表明 YOLOv8 的检测精度提升有效减少了 SORT 算法的轨迹断裂问题；在实时性方面，本文方法以 45 FPS 的运行速度领先于同类方法，完全满足课堂实时监控需求，而传统背景差分法虽 FPS 较高，但检测精度严重不足。

综合来看，本文方法在性能与效率之间实现了更优的平衡。在检测精度(mAP@0.5 达 98.5%)、跟踪稳定性(MOTA 85.2%)和实时性(45 FPS)上均优于现有主流方案，尤其在密集遮挡场景下表现突出。未来将通过引入注意力机制进一步提升复杂光照条件的适应性。

## 4.5. 改进方法

为了提高模型在复杂场景中的适应性，可以通过增加数据集的多样性或引入数据增强技术来训练模型。例如，可以模拟不同的光线条件、遮挡情况或学生姿态变化，以提高模型在这些场景下的检测和跟踪能力，同时，在 YOLOv8 和 SORT 算法的融合过程中，可以通过更精细的参数调整或引入其他辅助算法来提高融合效果。例如，可以引入更复杂的轨迹预测算法或目标重识别技术，以应对目标遮挡或重新出现的情况，在保证系统实时性的同时，可以通过优化算法结构或引入硬件加速技术来提高检测和跟踪的精度。例如，可以使用轻量化的模型结构或 GPU 加速技术来提高算法的运行效率。

## 4.6. 实验效果

如图 12、图 13 展示了基于 YOLOv8 目标检测模型对不同课堂场景的人员检测结果。模型成功检测到图中人员，并用绿色边界框标注每个检测目标[15]，同时标注了类别(person)和检测置信度(大部分置信

度在 0.6 以上)。检测范围覆盖整个教室，包括前排、中排和后排区域。系统能够准确识别并统计课堂中的学生人数，显示了较高的检测效果，即使在人员密集的中后排区域，也能够较好地完成检测任务。检测结果实时显示，适合课堂管理和动态监控场景的应用。


**Figure 12.** Schematic diagram of detection results in spacious classroom
**图 12.** 宽敞教室检测结果示意图

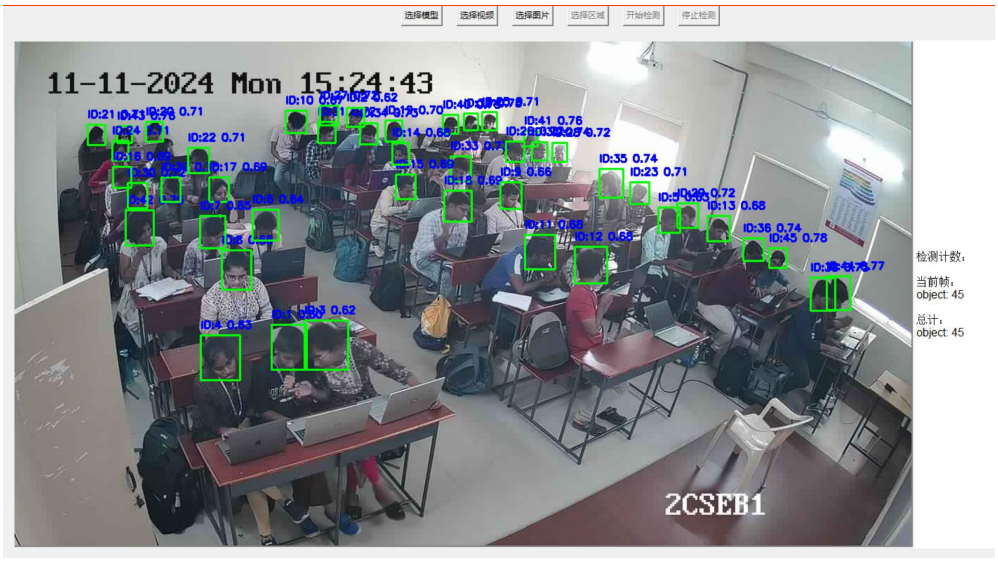


**Figure 13.** Schematic diagram of detection results in standard-sized classroom
**图 13.** 标准尺寸教室检测结果示意图

## 5. 结语

基于多目标跟踪的课堂人数自动统计算法研究展示了强大的目标检测能力，通过实时捕捉教室场景中的人员分布，成功实现了高效的课堂人数统计。系统能够在不同密度的区域内准确识别人员，并通过

绿色边界框和置信度标注每个目标，覆盖范围包括前排、中排和后排的所有学生位置。在图像中，系统检测到 75 名人员，置信度大部分高于 0.6，体现了模型对目标类别的高识别能力。尽管在人员密集区域(如后排)可能出现轻微的边界框重叠，但总体检测效果良好。

该算法结合了目标检测和人数统计功能，实时性强，适合应用于课堂管理和动态监控场景，为教师和管理人员提供了便捷的工具[16]。未来可进一步优化模型在密集区域的检测精度，以及对复杂场景(如遮挡或低光照)的适应性，从而进一步提高课堂人数统计的准确性和稳定性。这一系统为智能课堂和教育数字化管理提供了可靠的技术支持。

## 基金项目

嘉兴南湖学院 2024 年国家级大学生创新创业训练项目(202413291024)；嘉兴南湖学院 2023 年校级 SRT 项目(8517233225)；嘉兴南湖学院 2023 年校级 SRT 项目(8517233234)；嘉兴南湖学院教学基本建设项目(238518012)；教育部产学合作协同育人项目(企业案例深度融合的教育模式改革实践)。

## 参考文献

[1] 朱春原. 基于改进 YOLOv8 算法的小目标检测研究[D]: [硕士学位论文]. 大连: 大连交通大学, 2024.

[2] 杜磊. 基于 SORT 算法的图像轨迹跟踪混合控制方法[J]. 现代电子技术, 2024, 47(13): 32-35.

[3] 曹洁, 牛瑜, 梁浩鹏. 基于优化权重的 YOLOv7 密集行人检测算法[J]. 液晶与显示, 2025, 40(3): 505-515.

[4] 曾如明. K-means 聚类算法的改进及其应用研究[D]: [硕士学位论文]. 南充: 西华师范大学, 2022.

[5] 孟喆, 冯辉, 徐海祥. 跨层级特征融合的水面多尺度目标检测算法[J/OL]. 武汉理工大学学报(交通科学与工程版): 1-8. http://kns.cnki.net/kcms/detail/42.1824.U.20250402.1641.020.html, 2025-04-22.

[6] 寇晓. 基于鲁棒无迹卡尔曼滤波的电力系统动态状态估计研究[D]: [硕士学位论文]. 西安: 西安理工大学, 2022.

[7] 陈广交, 邓明阳. 基于 YOLOv8m 融合匈牙利算法的智能网联汽车环境感知方法[J]. 北华大学学报(自然科学版), 2025, 26(2): 245-253.

[8] 王红林, 黄浩, 孙彩云, 等. 基于 YOLOv5s 和 DeepSORT 的改进多目标跟踪算法[J]. 计算机仿真, 2025, 42(3): 263-269+303.

[9] 王谭, 王磊磊, 张卫国, 等. 基于张正友标定法的红外靶标系统[J]. 光学精密工程, 2019, 27(8): 1828-1835.

[10] 林新颖. 基于 Faster R-CNN 的视频车辆识别方法研究[J]. 中国信息界, 2024(8): 201-203.

[11] Peng, D., Sun, Z., Chen, Z., Cai, Z., Xie, L. and Jin, L. (2018) Detecting Heads Using Feature Refine Net and Cascaded Multi-Scale Architecture. 2018 24*th International Conference on Pattern Recognition* (*ICPR*), Beijing, 20-24 August 2018, 2528-2533. https://doi.org/10.1109/icpr.2018.8545068

[12] 冯桂尔. 基于 Python 的 GUI (Tkinter)实例开发[J]. 信息与电脑(理论版), 2023, 35(1): 175-178.

[13] 李永盛, 何佳洲, 刘义海, 等. 基于图像检测识别的数据增强技术[J]. 舰船电子对抗, 2021, 44(1): 66-70.

[14] 李旺, 张娜娜. 基于改进 FairMOT 的多目标跟踪算法[J]. 计算机工程与应用, 2024, 60(11): 139-146.

[15] 杨思燕, 苗凯彬, 王锋, 等. 视频图像中人脸自动检测与统计算法[J]. 电子科技, 2020, 33(8): 1-9.

[16] 何源. 基于人工智能的智慧课堂教学模式研究与探索[J]. 甘肃教育研究, 2025(1): 39-42.