**Software Engineering and Applications** 软件工程与应用, 2022, 11(6), 1191-1203
Published Online December 2022 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2022.116121

Hans汉斯

# 基于人物交互的学生课堂行为识别研究

**周珍玉，秦　学**

贵州大学大数据与信息工程学院，贵州　贵阳

收稿日期：2022年10月18日；录用日期：2022年11月28日；发布日期：2022年12月8日

## 摘　要

**学生课堂行为分析是评估课堂教学效果的有效方法，现有的学生行为识别研究仅针对学生自身进行识别，而对学生与周围物品的交互关注不够。基于此，本文提出了一种基于人物交互的学生课堂行为识别研究方法，通过分析教室监控视频，检测出学生和物品目标，并基于人与物的交互关系来识别其课堂行为。首先，考虑到笔和手机等小目标物品所占像素和可提取的有效特征较少，提出了一种改进YOLOv5s的目标检测方法，解决随着网络层数的叠加，小目标的特征信息逐渐消失导致漏检的问题。然后，为解决教室环境中目标数量较多，目标与目标间存在遮挡等因素导致网络难以提取特征的问题，引入Triplet注意力机制，增强网络提取特征的能力。接下来，采用VSGNet网络识别人与物的交互关系以确定行为类别。最后，在自制教室数据集和公开数据集进行了多组对比实验，实验结果表明，与原YOLOv5s网络相比，改进后的网络在自制和公开数据集上mAP分别提升了3.06%和3.2%，召回率分别提升了3.1%和4.2%，验证了改进方法的有效性。**

## 关键词

**YOLOv5s，学生行为识别，人物交互，Triplet Attention**

# Research on Students' Classroom Behavior Recognition Based on Human-Object Interaction

**Zhenyu Zhou, Xue Qin**

School of Data & Information Engineering, Guizhou University, Guiyang Guizhou

Received: Oct. 18th, 2022; accepted: Nov. 28th, 2022; published: Dec. 8th, 2022

## Abstract

**Student classroom behavior analysis is an effective way to assess the effectiveness of classroom**

文章引用: 周珍玉, 秦学. 基于人物交互的学生课堂行为识别研究[J]. 软件工程与应用, 2022, 11(6): 1191-1203.
DOI: 10.12677/sea.2022.116121

**instruction. Existing research on student behavior identification focuses only on the identification of students themselves, and not enough attention is paid to students' interactions with surrounding objects. Based on this, this paper proposes a research method for recognizing students' classroom behaviors based on Human-Object interaction, by analyzing classroom surveillance videos, detecting students and object targets, and recognizing their classroom behaviors based on person-object interaction. Firstly, considering that small target items such as pens and cell phones occupy fewer pixels and can extract fewer effective features, a target detection method is proposed to improve YOLOv5s to solve the problem of missing detection due to the gradual disappearance of feature information of small targets as the layers of the network are superimposed. Then, in order to solve the problem that it is difficult for the network to extract features due to a large number of targets in the classroom environment and the occlusion between targets and targets, the Triplet attention mechanism is introduced to enhance the ability of the network to extract features. Next, the VSGNet network is used to identify human-object interactions to determine behavior categories. Finally, multiple sets of comparison experiments were conducted on the self-made classroom dataset and the public dataset. The experimental results showed that the improved network improved the mAP by 3.06% and 3.2% on the self-made and public datasets, respectively, and improved the recall by 3.1% and 4.2%, respectively, compared with the original YOLOv5s network, which verified the effectiveness of the improved method.**

## Keywords

**YOLOv5s, Students' Classroom Behavior Recognition, Human-Object Interaction, Triplet Attention**

Copyright © 2022 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

课堂是学校教学教育的主要场所，课堂教学评价对提高教学质量具有重要的意义，而学生的课堂行为表现是进行课堂教学评价的重要依据[1]。在传统的课堂教学中，学习者评价主要来源于学习结果和人为观察，难以进一步了解和评价学习者的学习过程。近年来，智慧课堂成为了提高教学质量和教学效率的一种新方式，而学生的课堂行为识别是实施智慧课堂的重要一环。很多研究者采用深度学习的方法，对学生课堂行为进行精细分类，提供更有价值的课堂行为分析结果。Wu 等人[2]利用感兴趣区域(Region Of Interest, ROI)和人脸跟踪算法，识别学生在课堂上的站立和举手行为。Ge 等人[3]利用 VGG16 网络在 ImageNet 数据集上进行预训练，然后将训练结果迁移到学生课堂行为识别任务中，识别学生的课堂行为。魏艳涛等人[4]提出基于 ResNet50 预训练模型的迁移学习，通过迁移学习得到深度学习网络识别出学生看书、睡觉等行为。Zhou 等人[5]从学生行为图像中提取人体骨骼的关键信息，并结合 10 层深度卷积神经网络(CNN-10)来识别学生的课堂行为。

但在人员密集的教室环境中，如果学生动作较相似，仅通过深度学习的方法对学生行为进行识别，很难准确识别出学生的行为。而学生与周围物品的交互占据了大多数的学生行为活动，因此检测和识别每个人与周围物品的交互方式对学生行为识别有着重要作用。本文研究一种两阶段的学生课堂行为识别方法，包含检测和识别两个阶段。第一阶段的目标是检测出图片中的人、书、手机、笔等四种目标类别。第二阶段则通过识别人与物品之间的交互关系，并依次确定看书、记笔记、玩手机、听课四种行为类别。由于第一阶段检测到的目标及其位置信息是第二阶段输入，因此在检测阶段网络的准确性是正确识别学

生行为的基础和关键所在。

针对这一需求，本文重点研究并提出了一种改进 YOLOv5s 的目标检测方法。首先针对教室场景下，人员密集且尺度跨越较大，同时手机、笔等目标相对较小，YOLOv5s 检测小目标时存在性能较差这一问题，本文在 YOLOv5s 的 Head 部分添加了一个适合检测小目标的检测头，增强网络对小目标特征的提取能力，同时将浅层网络特征与深层网络特征相融合，强化小目标特征。其次，为了解决教室场景下由于遮挡等导致网络难以提取特征的问题，在 CBL (Conv + BN + LeakyRelu)模块中加入 Triplet (Triplet Attention)注意力机制，细化特征的同时保留上下文特征，提高网络检测精度。最后，使用 VSGNet (Visual-Spatial-Graph Network)判断第一阶段检测到的人与物品间存在的交互关系，进一步验证改进工作的有效性。

## 2. 相关研究工作

### 2.1. 目标检测

现有的目标检测网络都取得了较好的检测效果，但相比于中、大目标，小目标存在尺度较小、可提取有效特征较少等问题。针对这一问题，许多研究人员着重研究提升小目标检测性能的方法。Xiao 等人[6]使用 Faster RCNN 检测图片中的苹果这一小目标，针对原网络中 RPN (Region Proposal Network)模块的滑动窗口太小，容易丢失有效信息的问题，增加了两个不同尺寸的滑动窗口。该方法虽然在准确率和召回率上取得了较好的效果，但是检测速度欠佳。为了提升网络的检测速度，一些研究人员优化了诸如 YOLO、SSD 等单阶段目标检测网络，此类网络相比于双阶段网络，检测速度和准确率都有了明显提升。Chen 等人[7]优化了 SSD 网络，在特征提取模块中将 Mobile-Net_v1 作为 SSD 网络特征提取的主干模块，取代了原网络中的 VGG16 模块，从而提高了网络对沃柑这一小目标的检测性能。Ye 等人[8]为解决 YOLOv3 在小目标检测方面表现较差，漏检率较高，误检率较高的问题。借鉴自适应空间特征融合的思想，充分利用高级特征的语义信息和低级特征的细粒度、边缘、纹理特征，融合高级特征图和浅层特征图。同时，采用 K 均值聚类算法生成更适合的锚框，进一步提高网络对小目标的检测精度。王程等人[9]优化了 YOLOv4 网络，采用深度可分离卷积代替原网络残差结构中的传统卷积，同时对 FPN (Feature Pyramid Networks)进行了改进，增强小目标多尺度特征学习，以提高网络检测精度及实时性。Yan 等人[10]为了增加小目标特征多样性和不变性，对数据集进行了预处理，提高 YOLO 网络在处理小目标时的性能。相比于 YOLO 网络，SSD 网络中先验框的大小需要手工设置，不能直接通过学习获得，导致调试过程非常依赖经验，其检测精度也低于前者[11]。

为了进一步提高网络检测精度，抑制无关信息的干扰，许多研究人员将注意力机制整合到目标检测网络中。Yang 等人[12]提出了扩张卷积注意力模块(Dilated-CBAM)，将该模块应用于残差网络的骨干，扩大了感受野，降低了网络参数量，节省了网络训练的时间和空间，同时强化了图像中的有效信息，削弱无效信息，将全局特征和局部特征融合，使得网络能提取到更丰富的特征。Wen 等人[13]提出了一种残差信道注意网络，在残差学习路径中引入了信道注意机制，构建了残差信道注意模块，有效避免了信息的丢失，提升了网络检测精度。

基于以上分析，本文采用 YOLOv5s 网络对教室下的目标进行检测。YOLOv5s 是在 YOLOv4 [14]的基础上进行了优化，总体性能明显提高。其网络的改进主要有：Backbone 部分采用 Focus 结构进行切片操作，降低了计算量；Neck 部分采用自上而下的特征金字塔结构加强了特征信息，同时在此基础上结合自下而上特征金字塔结构，使网络获得了更加丰富的特征信息；采用加权 NMS (Non Maximum Suppression)的方式，遮挡目标检测精度有了一定提升。YOLOv5 主要有四个版本网络，分别是 YOLOv5s、YOLOv5m、

YOLOv5l、YOLOv5x，其中 YOLOv5s 特征图深度最浅，宽度最窄，后面三种在此基础上不断加深、加宽，它主要由 Backbone、Neck、Head 三个部分组成，其网络结构图如图 1 所示。

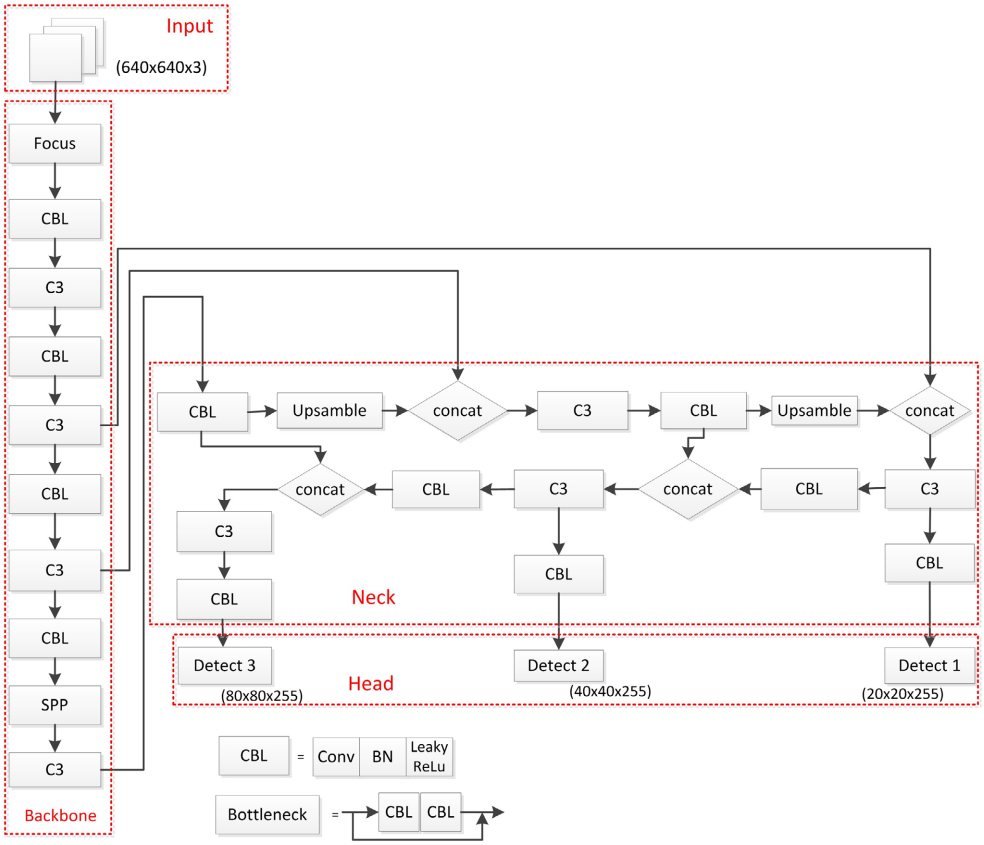



**Figure 1.** YOLOv5s network structure diagram
**图 1.** YOLOv5s 网络结构图

## 2.2. 人物交互检测

识别一个人的动作行为，不仅仅只是检测出单个目标对象，还需要识别每个人如何与周围物品交互[15] [16]。人物交互(Human Object Interaction, HOI)检测旨在检测人与物的位置信息，以及识别人与物之间的交互关系[17]。2015 年 Malik 等人[18]首次提出了“视觉语义角色标注”这一概念，即：对细粒度的动作进行推理，并检测这个动作的各种语义角色。此后许多研究人员也开始关注这一方向，并做出了相关研究。Kolesnikov 等人[19]提出了 BAR-CNN (Box Attention R-CNN)模型，利用链式规则将概率模型分解，将人与物的空间位置关系进行编码，定位输入图像中的所有目标，并检测与该目标交互的所有其他物品。Shao ZP 等人[20]使用多流特征优化网络，细化了网络中提取到的人、环境和物品的视觉特征。

本文使用 VSGNet 网络识别教室场景下的学生课堂行为。VSGNet [21]网络首先从人和物中提取视觉特征，然后利用人与物的空间位置信息对特征进行细化，最后利用图卷积对人与物之间的结构交互进行建模。VSGNet 网络主要由视觉分支、空间注意分支、图卷积分支三个分支组成，其网络结构图如图 2 所示。

视觉分支在人、物检测框上，首先利用感兴趣区域池化 RoI pooling (Region of interest pooling)提取特征，然后再经过残差块 Residual、全局平均池化 GAP (Global Average Pooling)操作，输出人和物的视觉特

征向量。同时周围的物品、背景以及其他的人都可以帮助检测相互作用，所以从整个输入图像中提取特征，输出上下文的视觉特征向量。将所有的视觉特征向量连接起来，并通过一个全连接层进行投影，最后得到人－物对的视觉特征向量。

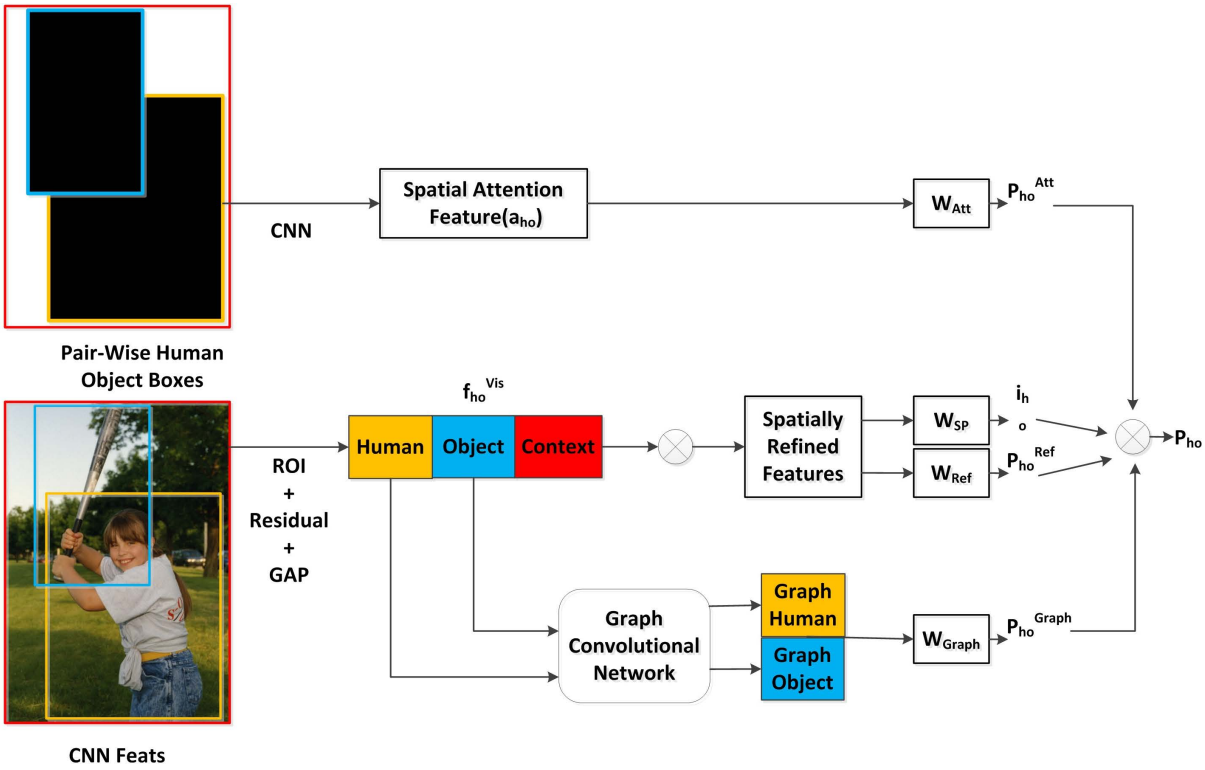



**Figure 2.** VSGNet network structure diagram
**图 2.** VSGNet 网络结构图

空间注意分支主要学习人与物品之间的空间相互交互模式。利用人和物品检测框的位置信息，生成二进制映射，然后使用卷积提取空间注意力特征，得到注意特征向量，即得到人－物对的空间关系。将视觉特征向量与注意特征向量相乘，从而细化具有空间配置的视觉特征向量，利用该特征向量预测人－物对的交互得分。

图卷积分支使用一个图卷积网络来学习人与物之间的结构联系。将人和物看作节点，他们的关系作为边，通过遍历和更新图中的节点，提取节点间结构关系的特征。最后将三个分支的概率相乘，得到最终预测的动作类别概率。

## 3. 人物交互的学生课堂行为识别网络设计与改进

### 3.1. 学生课堂行为识别网络设计

学生课堂行为识别网络结构如图 3 所示。首先，对学生上课视频进行数据预处理后，将图片从 YOLOv5s 的 Input 处输入，依次经过 Backbone 部分进行特征提取；Neck 部分将多个不同分辨率的特征进行融合，以丰富网络的特征；Head 部分以四种尺度大小分别为 160 × 160、80 × 80、40 × 40、20 × 20 的特征图检测小、较小、中、大目标。接着，输出检测结果，包括所需目标的位置信息以及检测框。然后，将检测结果输入 VSGNet 网络的三个分支中，空间注意分支利用人和物的位置信息得到空间注意力

特征向量；视觉分支从图片中提取人和物的特征，得到视觉特征向量；图卷积分支提取人和与之交互物品之间的结构关系特征。最后将三个分支概率相乘得到各动作类别概率。

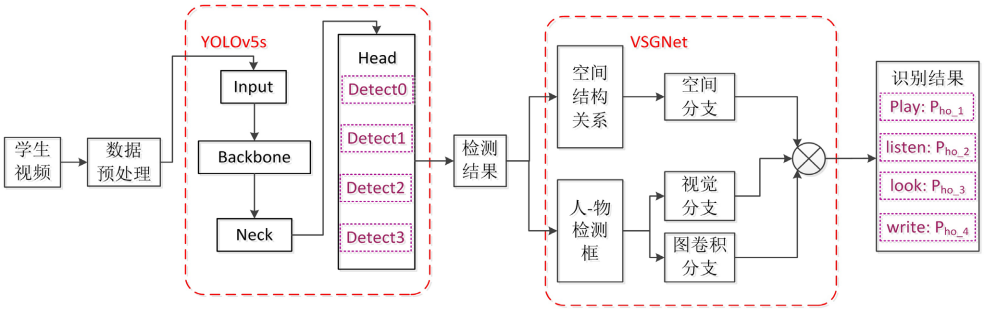



**Figure 3.** Network structure of student classroom behavior recognition
**图 3.** 学生课堂行为识别网络结构图

## 3.2. 小目标检测网络的改进

原 YOLOv5s 网络中采用 80 × 80 的检测头检测小目标，即用每格以 8 × 8 的感受野学习特征，导致低于 8 × 8 像素的特征信息不能被网络学习到。因此，本文在 YOLOv5s 的 Head 中增加了一个大小为 160 × 160 的检测头检测小目标，以提高网络对小目标的检测能力。虽然在一定程度上会增加计算开销，但在特征融合阶段，160 × 160 检测层的特征图维数相对较低，参数数量的增加只集中在预测层，使得参数数量的增加相对有限。与原始图像相比，新增的检测头能学习到 4 × 4 像素的特征。

另外，虽然深层网络包含大量的语义信息，但是由于小目标所占像素较少，其特征信息会随着网络的加深逐渐消失。所以，本文在 Backbone 的第一个 C3 模块后引入 160 × 160 小目标检测头，将浅层网络中学习到的浅层特征图与深层网络特征图进行拼接，实现特征融合，强化小目标特征。改进后网络结构如图 4 所示，增加的检测头为长虚线框所示。

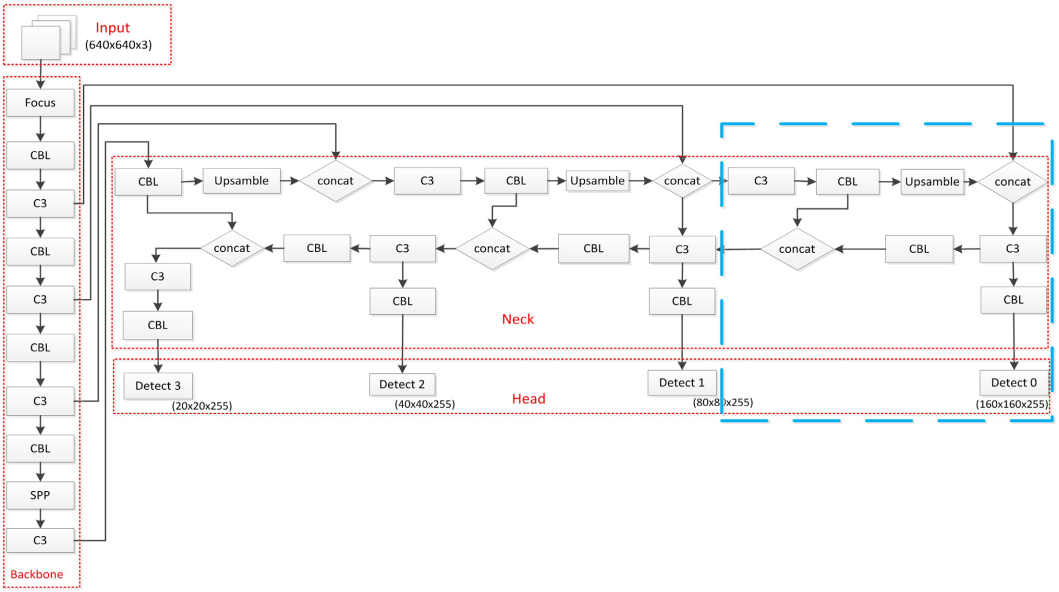



**Figure 4.** Improved YOLOv5s network structure diagram
**图 4.** 改进后 YOLOv5s 网络结构图

### 3.3. 基于注意力特征提取网络的改进

Triplet 注意力是由 Misra 等人[22]提出的一种轻量型注意力机制，不同于 CBAM [23] (Convolutional Block Attention Module)在计算通道注意力时通过降维捕获通道之间的非线性局部依赖性关系。Triplet 注意力在几乎不增加参数量的同时实现了跨通道交互且不降低维度，从而消除了通道和权重的间接对应关系。其网络结构图如下图 5 所示。

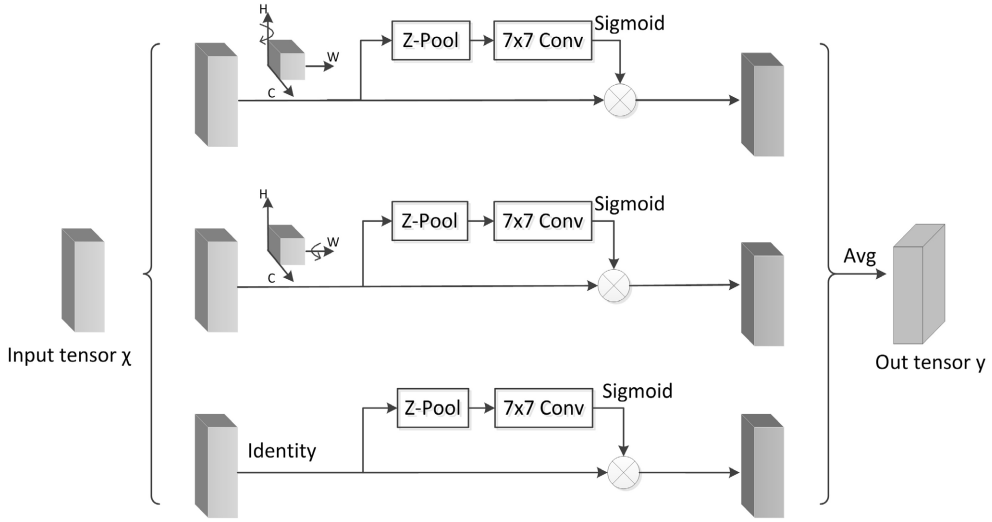



**Figure 5.** Triplet attention network structure diagram
**图 5.** Triplet 注意力网络结构图

Triplet 注意力主要由三个平行分支组成。在第一个分支中，输入张量 $\chi$ 首先沿 H 轴逆时针旋转 90˚得到旋转张量 $\hat{\chi}_1$，此时形状为(W × H × C)。然后 $\hat{\chi}_1$ 经过 Z-Pool 降维后，形状变为(2 × H × C)，再通过 7 × 7 的卷积后张量形状变为(1 × H × C)，接着张量通过 Sigmoid 激活函数生成了注意力权重。最后输出的张量沿着 H 轴顺时针旋转 90˚保持和输入的形状一致。自此，建立了 C 维度和 H 维度间的交互。

类似的，在第二个分支中，输入张量 $\chi$ 沿着 W 轴逆时针旋转 90˚得到旋转张量 $\hat{\chi}_2$，形状为(H × C × W)，再经过 Z-Pool 后，形状变为(2 × C × W)。然后通过 7 × 7 的卷积后张量形状变为(1 × C × W)，接着张量通过 Sigmoid 激活函数来生成注意力权重，最后沿着 W 轴顺时针旋转 90˚，使得输出张量与输入张量的形状一致。由此建立了 C 维度和 W 维度间的交互。

第三个分支中，为了建立 H 维度和 W 维度间的交互，输入张量 $\chi$ 通过 Z-Pool 将张量降为 2 维，即张量简化为(2 × H × W)。然后经过 7 × 7 卷积得到形状为(1 × H × W)的张量，最后通过 Sigmoid 生成了注意力权重，并将其作用于 $\chi$。其中 H、W、C 分别表示高度、宽度、通道数，Z-Pool 主要作用是将 C 维度的张量缩减至 2 维，并将该维上的平均汇集特征和最大汇集特征连接起来，丰富张量表示，同时缩小深度使计算量更小，可以表示为对输入张量进行最大池化和平均池化。

最后，输入张量经过三个分支后相加再进行平均池化，得到输出张量 y。

在对真实教室场景下的目标进行检测时，仅依靠原网络中 CBL 模块的简单叠加，并不能显著提高网络提取特征的能力。而 Triplet 注意力模块能提取到高阶语义特征，并且不会增加参数量。因此，本文将 Triplet 注意力模块加入 CBL 模块中。CBL 模块结构图如图 6 所示，由标准卷积层、归一化批处理 BN 层以及 Leaky relu 激活层构成。本文在 CBL 的 BN 层后加入 Triplet 注意力模块，在经过标准卷积和批量归

一化后，由 Triplet 注意力模块进行特征增强同时保留上下文特征，改进后的 CBL 记为 CBL_Triplet，CBL_Triplet 模块结构图如图 7 所示。


**Figure 6.** CBL structure diagram
**图 6.** CBL 结构图


**Figure 7.** CBL_Triplet structure diagram
**图 7.** CBL_Triplet 结构图

## 4. 实验分析

### 4.1. 数据集的选择

本文实验数据集来源于真实教学场景下的课堂监控视频，考虑到学生上课时在连续时间内变化幅度较小，故每间隔 50 帧抽取一张图片，同时为了丰富数据集的多样性，选取不同教室上课视频进行图片抽取，共计抽取 1500 张图片，每张图片包含 8~10 位学生。通过观察学生上课视频，选取学生典型的四种课堂行为，分别是看书、记笔记、玩手机、听课。

COCO 数据集由微软团队构建的可用于目标检测的大型数据集，包含 150 万个对象实例、33 万张图片、91 个物体类别以及 80 个目标类别。相比于 PASCAL VOC 数据集，COCO 数据集背景更复杂，小目标数量更多，检测难度更大。为了验证本文改进目标检测网络的有效性，在 COCO 数据集中随机抽取小目标较多的 person、book、bird、cell phone 四种目标类别图片，共计 16,560 张图片，按照训练集：测试集：验证集 = 8:1:1 的比例进行训练和测试。

### 4.2. 性能评价指标

本文实验性能评价指标采用准确率($P$) Precision、召回率($R$) Recall、以及 $mAP$ 进行计算。准确率 $P$、召回率 $R$、$mAP$ 计算公式如式(1)~式(3)所示。

$$P = \frac{S_{TP}}{S_{TP} + S_{FP}} * 100\% \tag{1}$$

$$R = \frac{S_{TP}}{S_{TP} + S_{FN}} * 100\% \tag{2}$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{3}$$

准确率 $P$ 表示正确预测的样本数量占总预测样本数量的百分比；召回率 $R$ 表示预测正确的样本数占标注正确类别的总样本数的百分比，即样本类别被模型正确预测的个数；$mAP$ 表示 IoU 设为 0.5 时所有类别的平均准确率。其中，$TP$ 表示正样本被预测为正样本，$FP$ 表示负样本被预测为正样本，$FN$ 表示正样本被预测为负样本。$S_{TP}$ 表示把正样本预测正确的个数，$S_{FP}$ 表示负样本被预测为正样本的个数，$S_{FN}$ 表示正样本未被正确预测的个数。

### 4.3. 检测网络实验结果分析

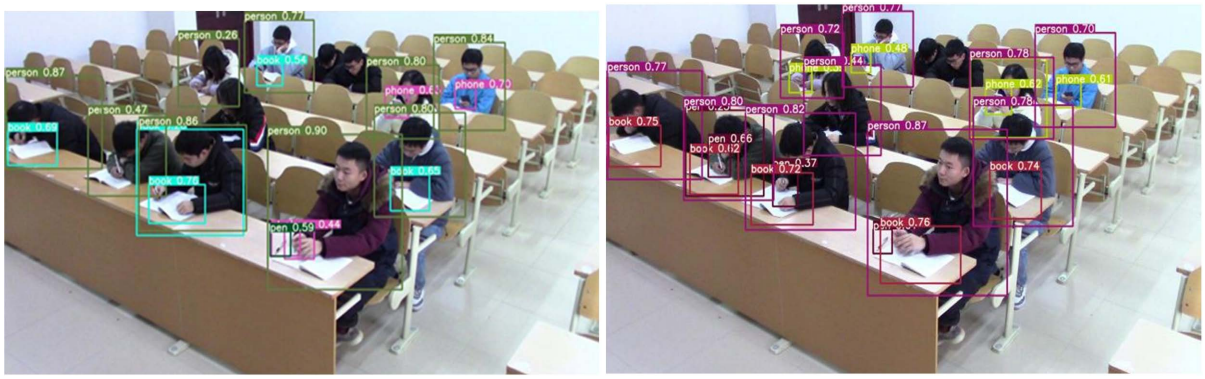
#### 4.3.1. 自制教室数据集对比实验

原网络与改进后网络实验结果如表 1 所示。实验结果表明，虽然增加一个检测头使得网络大小增加了 0.63 MB，每张图片的检测时间增加了 0.0072 s，但在牺牲参数量和增加少许检测时间的基础上精度和召回率都有了明显提升。改进后的检测网络 YOLOv5s 与原网络相比 mAP 由 82.42%提升到 85.49%，提升了 3.06%，召回率 R 由 82.60%提升到 85.70%，提升了 3.1%。实验证明改进后 YOLOv5s 网络对教室场景下的目标检测性能有所提升。

**Table 1.** Experimental results of self-made dataset
**表 1.** 自制数据集实验结果

| Networkmodel | mAP/% | P/% | R/% | 网络大小/MB | Average time/s |
|---|---|---|---|---|---|
| YOLOv5s | 82.42 | 83.30 | 82.60 | 6.76 | 0.0092 |
| Our_YOLOv5s | 85.49 | 82.94 | 85.70 | 7.39 | 0.0164 |

改进前后网络对遮挡小目标检测结果对比如图 8 所示。从图中可以看出，改进后的 YOLOv5s 网络能检测出笔这一遮挡小目标(如第一排中间两位正在记笔记的同学)，漏检问题明显得到改善，检测到的目标数量也有所提升，这对于后续人物交互网络识别学生行为至关重要。



(a) (b)

**Figure 8.** Comparison of detection results of occlusion small targets by network before and after improvement. (a) CBL_Triplet structure diagram; (b) Improved YOLOv5s detection results
**图 8.** 改进前后网络对遮挡小目标检测结果对比图。(a) YOLOv5s 检测结果；(b) 改进后 YOLOv5s 检测结果

#### 4.3.2. 公开数据集对比实验

为了验证改进 YOLOv5 网络的有效性和可靠性，在公开数据集 COCO 中选取具有数量多、有遮挡、小目标等属性的 16,560 张图片进行实验。实验结果如表 2 所示。实验结果表明，Our_YOLOv5s 与 YOLOv5s 相比，虽然网络大小增加了 0.631 MB，平均检测时间增加了 0.0097 s，但是准确率和召回率分别由 30.21%、31.10%提高到 33.41%、35.30%，分别提升了 4.5%和 4.2%。

检测网络改进前后 mAP 值对比如图 9 所示，从图 9 可以看出，虽然训练刚开始时改进网络 mAP 值没有原网络值高，但是随着迭代次数的增加，网络逐渐平稳，改进后的网络 mAP 值逐渐提升直至超越原网络。实验结果表明改进后的网络精度和召回率在公开数据集上有提升。

**Table 2.** Experimental results of public datasets
**表 2.** 公开数据集实验结果

| Networkmodel | mAP/% | P/% | R/% | 网络大小/MB | Average time/s |
|---|---|---|---|---|---|
| YOLOv5s | 30.21 | 45.00 | 31.10 | 6.744 | 0.0075 |
| Our_YOLOv5s | 33.41 | 49.50 | 35.30 | 7.375 | 0.0172 |

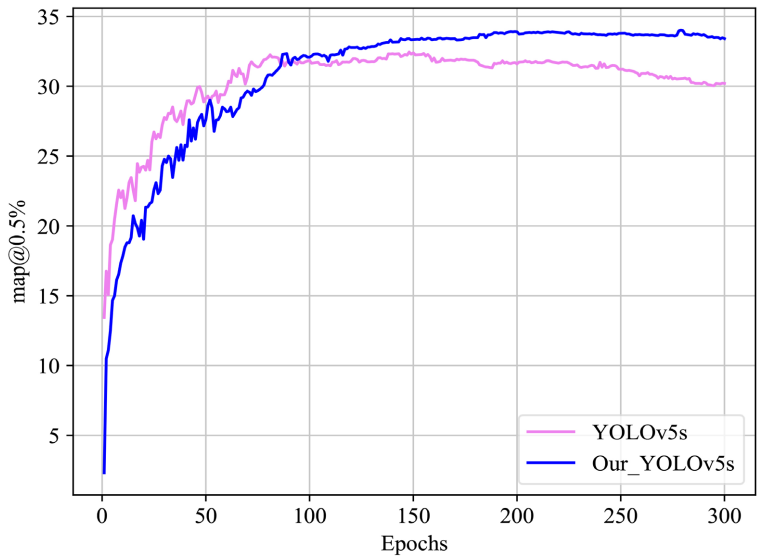



**Figure 9.** Comparison of mAP values before and after network improvement
**图 9.** 网络改进前后 mAP 值对比图

### 4.3.3. 消融实验

为了验证改进的模块对网络整体性能的影响，本文对改进模块进行了消融实验，其实验结果如表 3 所示。实验结果表明添加 Head，网络大小仅增加了 0.631 MB，网络的检测精度却提高了 2.57%。Triplet 模块的加入，网络大小仅增加 0.017 MB，网络精度提升了 1.64%，这表明 Triplet 模块在几乎不增加计算开销的同时还能提高网络检测精度。在添加 Head 的基础上引入 Triplet 模块，网络的检测性能更佳，检测精度提高了 3.2%。

**Table 3.** Ablation experiment
**表 3.** 消融实验

| Head | Triplet | mAP/% | 网络大小/MB |
|---|---|---|---|
| | | 30.21 | 6.744 |
| √ | | 32.78 | 7.375 |
| | √ | 31.85 | 6.761 |
| √ | √ | 33.41 | 7.375 |

### 4.3.4. 不同模型的对比实验

为了进一步验证改进工作的有效性和可靠性，将改进的 YOLOv5s 网络与当前主流的目标检测网络 YOLOv3 和 Faster RCNN 进行对比实验。为了保证实验的公平性，对比实验在相同的实验环境下进行。实验结果如表 4 所示。从表 4 可看出，改进后的 YOLOv5s 网络相比于其它网络总体性能更优。

**Table 4.** Comparative experiments of different network models
**表 4.** 不同网络模型对比实验

| Network model | mAP/% | Average time/s |
|---|---|---|
| YOLOv5s | 30.21 | 0.0075 |
| YOLOv3 | 30.10 | 0.0291 |
| Faster RCNN | 28.75 | 0.1982 |
| Our_ YOLOv5s | 33.41 | 0.0172 |

### 4.4. 学生行为网络实验及结果展示

将学生上课视频中抽帧得到的图片输入到本文改进的 YOLOv5s 网络中，输出图片中学生以及手机、笔等物品的检测框及位置坐标信息，检测框包含目标的置信度。接着，将检测结果输入 VSGNet 三个分支中，初始学习率和批处理大小分别设置为 0.01 和 8。然后，将来自 VSGNet 网络不同分支的所有预测结果相乘。最后，确定行为类别。

学生行为识别结果如图 10 所示。实验结果表明，本文提出的网络能识别出学生看书(look)、记笔记(write)、玩手机(play)、听课(listen)四种课堂行为，平均准确率 mAP 可达到 58.4%。但是，现有的数据集中，“记笔记”这一行为数量较少，导致网络稳定性较差，后续还需要扩充数据集。

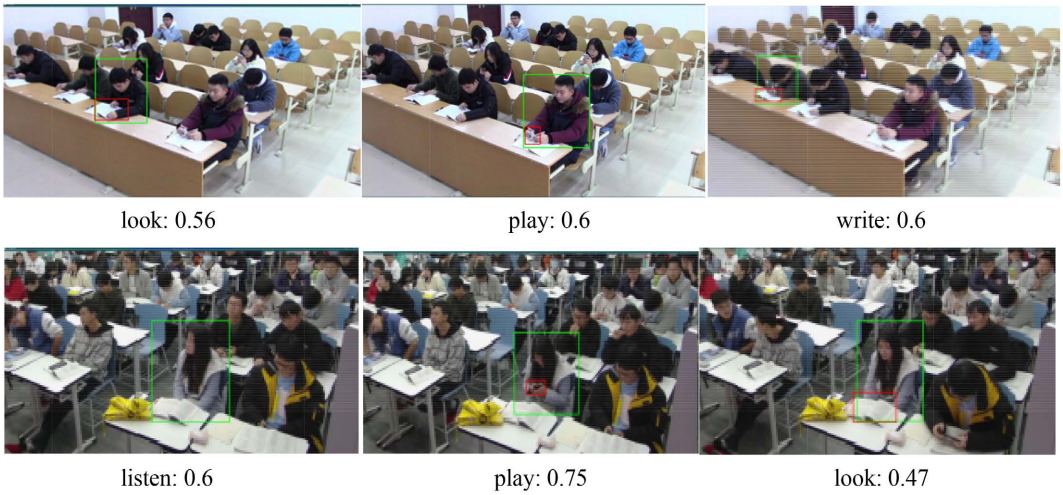



**Figure 10.** Results of student behavior recognition
**图 10.** 学生行为识别结果

## 5. 结束语

本文提出了一种基于人物交互的学生课堂行为识别研究方法，重点对学生四种典型的课堂行为即：看书、记笔记、玩手机、听课进行了识别。研究过程中，本文对实际应用场景中存在的一系列问题，优化和改进了网络。首先，针对检测阶段教室环境中目标数量多、目标与目标存在遮挡、小目标漏检等问题，本文提出了一种改进的 YOLOv5s 目标检测方法。通过增加一个大小为 160 × 160 的小目标检测头，检测 4 × 4 像素的小目标区域，同时实现浅层特征与深层特征的融合。接着在 CBL 模块中加入 Triplet 注意力机制，增强网络提取特征的能力，进一步提升网络检测精度。然后，使用 VSGNet 网络识别人和物交互关系以确定行为类别。实验结果表明，改进后的 YOLOv5s 网络，有效改善了小目标漏检问题，同时

提升了网络检测精度。

目前，本文对学生行为识别的研究，虽然取得了一定的效果，但本文是通过学生上课视频进行抽帧得到的图片，不能实时识别，实时性将会是后续的主要研究方向。

## 参考文献


[1] 高杨凡. 基于人体骨架和深度学习的学生课堂行为识别研究[D]: [硕士学位论文]. 武汉: 华中师范大学, 2021.

[2] Wu, B., Wang, C.-M., Huang, W., *et al*. (2021) Recognition of Student Classroom Behaviors Based on Moving Target Detection. *Traitement du Signal*, **38**, 215-220. https://doi.org/10.18280/ts.380123

[3] Ge, C., Ji, J.-Q. and Huang, C.-F. (2022) Student Classroom Behavior Recognition Based on OpenPose and Deep Learning. *International Conference on Intelligent Computing and Signal Processing* (*ICSP*), Xi'an, 15-17 April 2022, 576-579. https://doi.org/10.1109/ICSP54964.2022.9778501

[4] 魏艳涛, 秦道影, 胡佳敏, 等. 基于深度学习的学生课堂行为识别[J]. 现代教育技术, 2019, 29(7): 87-91.

[5] Zhou, J., Feng, R., Guang, L., *et al*. (2022) Classroom Learning Status Assessment Based on Deep Learning. *Mathematical Problems in Engineering*, **2022**, Article ID: 7049458. https://doi.org/10.1155/2022/7049458

[6] Xiao, X.-S. and Tian, X.X. (2021) Research on Reference Target Detection of Deep Learning Framework Faster-RCNN. *International Conference on Data Science and Business Analytics* (*ICDSBA*), Changsha, 24-26 September 2021, 41-44. https://doi.org/10.1109/ICDSBA53075.2021.00017

[7] Chen, L. and Wang, S.-G. (2021) Identification and Detection of Picking Targets of Orah Mandarin Orange in Natural Environment Based on SSD Model. *Eurasia Conference on IOT, Communication and Engineering* (*ECICE*), Yunlin, 29-31 October 2021, 439-442.

[8] Ye, K.-Q., Fang, Z.-B., Huang, X.-J., *et al*. (2022) Research on Small Target Detection Algorithm Based on Improved Yolov3. *International Conference on Mechanical, Control and Computer Engineering* (*ICMCCE*), Harbin, 25-27 December 2020, 1467-1470.

[9] 王程, 刘元盛, 刘圣杰. 基于改进 YOLOv4 的小目标行人检测算法[J]. 计算机工程, 2022: 1-9. https://doi.org/10.19678/j.issn.1000-3428.0063623

[10] Yan, M.Y. and Sun, J.B. (2022) A Dim-Small Target Real-Time Detection Method Based on Enhanced YOLO. *International Conference on Electrical Engineering, Big Data and Algorithms* (*EEBDA*), Changchun, 25-27 February 2022, 567-571. https://doi.org/10.1109/EEBDA53927.2022.9745012

[11] Redmon, J. and Farhadi, A. (2018) YOLOv3: An Incremental Improvement.

[12] Yang, J.Y. and Jiang, J. (2021) Dilated-CBAM: An Efficient Attention Network with Dilated Convolution. *International Conference on Unmanned Systems* (*ICUS*), Beijing, 15-17 October 2021, 11-15. https://doi.org/10.1109/ICUS52573.2021.9641248

[13] Wen, X., Pan, Z.-X., Hu, H.-Y., *et al*. (2022) An Effective Network Integrating Residual Learning and Channel Attention Mechanism for Thin Cloud Removal. *IEEE Geoscience and Remote Sensing Letters*, **19**, 1-5. https://doi.org/10.1109/LGRS.2022.3161062

[14] Bochkovskiy, A., Wang, C.-Y. and Liao, H.-Y.M. (2020) YOLOv4: Optimal Speed and Accuracy of Object Detection.

[15] Xu, B.J., Li, J.N., Wong, Y.K., *et al*. (2020) Interact as You Intend: Intention Driven Human Object Interaction Detection. *IEEE Transactions on Multimedia*, **22**, 1423-1432. https://doi.org/10.1109/TMM.2019.2943753

[16] Siadari, T.S., Han, M. and Yoon, H. (2020) Three-Stream Network with Context Convolution Module for Human-Object Interaction Detection. *ETRI Journal*, **42**, 230-238. https://doi.org/10.4218/etrij.2019-0230

[17] Gao, C. and Zou, Y.L. (2018) iCAN: Instance-Centric Attention Network for Human-Object Interaction Detection. 2018 *IEEE Conference on Computer Vision and Pattern Recognition*, *CVPR* 2018, Salt Lake City, 18-22 June 2018, 1-13.

[18] Gupta, S. and Malik, J. (2015) Visual Semantic Role Labeling.

[19] Kolesnikov, A., Kuznetsova, A., Lampert, C., *et al*. (2019) Detecting Visual Relationships Using Box Attention. 2019 *IEEE/CVF International Conference on Computer Vision Workshop* (*ICCVW*), Seoul, 27-28 October 2019, 1749-1753. https://doi.org/10.1109/ICCVW.2019.00217

[20] Shao, Z.P., Hu, Z.Y., Yang, J.Y., *et al*. (2022) Multi-Stream Feature Refinement Network for Human Object Interaction Detection. *Journal of Visual Communication and Image Representation*, **86**, Article ID: 103529. https://doi.org/10.1016/j.jvcir.2022.103529

[21] Ulutan, O., Iftekhar, A. and Manjunath, B.S. (2020) VSGNet: Spatial Attention Network for Detecting Human Object

Interactions Using Graph Convolutions. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 13614-13623. https://doi.org/10.1109/CVPR42600.2020.01363

[22] Misra, D., Nalamada, T., Arasanipalai, A.U. and Hou, Q.B. (2020) Rotate to Attend: Convolutional Triplet Attention Module. *IEEE Winter Conference on Applications of Computer Vision* (*WACV*), Snowmass Village, 1-5 March 2020, 3138-3147. https://doi.org/10.1109/WACV48630.2021.00318

[23] Woo, S., Park, J., Lee, J.-Y. and Kweon, I. (2018) Cbam: Convolutional Block Attention Module. 15*th European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1