软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2022,33(7):2633−2645 [doi: 10.13328/j.cnki.jos.006289]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于级联注意力与点监督机制的考场目标检测模型*

田卓钰[1], 马 苗[1,2,3], 杨楷芳[1]

[1](陕西师范大学 计算机科学学院, 陕西 西安 710119)

[2](现代教学技术教育部重点实验室(陕西师范大学), 陕西 西安 710062)

[3](空天地海一体化大数据应用技术国家工程实验室, 陕西 西安 710129)

通信作者: 马苗, E-mail: mmthp@snnu.edu.cn

摘 要: 智慧考场是智慧校园的重要组成部分, 准确、快速地检测考场中的学生状态, 是智慧考场应用的基本任务和关键环节. 标准化考场中的考生分布相对密集且成像尺寸差异较大, 而现有目标检测算法未充分考虑真实考场的环境特征, 很难精确、实时地检测出考生目标, 加之大部分目标检测算法需对不同目标手工设计先验锚框, 模型部署范围受限. 针对以上问题, 提出一种高效的无锚框全卷积目标检测模型. 该模型采用全卷积网络对输入图像进行逐像素预测, 在可能存在目标的区域回归其包围框. 在该模型中, 设计了基于级联注意力的特征增强模块, 通过逐级细化修正特征增强特征图的判别性, 有效地提高考生目标识别精度; 另一方面, 针对真实考场中大量交叠目标检测问题, 提出了点监督机制, 以进一步提升交叠多目标的识别效果; 最后, 在构建的标准化考场检测专用数据集上, 对所提模型进行验证. 实验结果表明, 与当前最先进的目标检测模型相比, 针对真实复杂的考场环境特征提出的基于级联注意力和点监督机制的全卷积目标检测模型的 mAP 指标为 92.9%, 检测速度为 22.1 f/s, 泛化能力突出, 综合效果最优.

关键词: 目标检测; 智慧考场; 无锚框方法; 注意力机制; 点监督机制

中图法分类号: TP391

中文引用格式: 田卓钰, 马苗, 杨楷芳. 基于级联注意力与点监督机制的考场目标检测模型. 软件学报, 2022, 33(7): 2633–2645. http://www.jos.org.cn/1000-9825/6289.htm

英文引用格式: Tian ZY, Ma M, Yang KF. Object Detection Model for Examination Classroom Based on Cascade Attention and Point Supervision Mechanism. Ruan Jian Xue Bao/Journal of Software, 2022, 33(7): 2633–2645 (in Chinese). http://www.jos.org.cn/1000-9825/6289.htm

Object Detection Model for Examination Classroom Based on Cascade Attention and Point Supervision Mechanism

TIAN Zhuo-Yu[1], MA Miao[1,2,3], YANG Kai-Fang[1]

[1](School of Computer Science, Shaanxi Normal University, Xi'an 710119, China)

[2](Key Laboratory of Modern Teaching Technology of Ministry of Education (Shaanxi Normal University), Xi'an 710062, China)

[3](National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, Xi'an 710129, China)

Abstract: Smart examination classroom is an important part of smart campus, and accurately and quickly detecting students in the examination classroom is a basic task of building a smart classroom. However, due to the dense distribution and imaging difference of the examinees in an examination classroom, most of the existing object detection methods can not precisely detect all the examinees in real-time. Moreover, most of the object detection methods rely on predefined anchor boxes, which are lack of portability. Aiming at the above problems, this study proposes an efficient one-stage object detection model based on fully convolutional network, which is

* 基金项目: 国家自然科学基金(61877038, 61801282, U2001205); 陕西师范大学研究生创新团队项目课题(TD2020044Y); 空天地海一体化大数据应用技术国家工程实验室开放课题(20200201)

收稿时间: 2020-05-08; 修改时间: 2020-08-08, 2020-12-03; 采用时间: 2020-12-03

anchor-free, with a prediction on the input image in pixel-level. In this model, a feature enhancement module is firstly designed based on cascade attention, which can effectively enhance the discriminability of the feature map by gradually refining and modifying the features. Secondly, in order to enable the network to distinguish overlapping objects in the examination classroom, a point supervision mechanism is proposed. Finally, this study verifies the above model on the special dataset of standardized examination classroom. With the cascade attention module and point supervision mechanism, the proposed model achieves 92.9% in mAP at the speed of 22.1 f/s, and is superior to most the state-of-the-art detection models. Especially, for object detection in new classroom environments, the proposed model achieves the best results.

Key words: object detection; smart examination classroom; anchor-free method; attention mechanism; point supervision mechanism

考试是考核学习者知识水平和能力的主要途径. 为了有效防范考生作弊行为、规范考场秩序, 维护考试的公平和公正, 我国目前主要采用人防和技防相结合的方式, 即现场考官巡查与电子视频监控的方式进行监考. 然而, 由于考场监控视频的数据量庞大、冗余信息过多, 传统考场监控系统效率极低, 监考人员往往会由于劳动强度大而导致视觉疲劳, 无法保证对多个考场监控画面进行高效监测, 从而难以发现考生的异常行为. 因此, 如何建设智慧考场, 运用先进的计算机视觉技术服务于现行的各类考试, 实现考生行为的智能化监控, 对于减轻监考人员的压力、维护考场秩序和保证考试公平具有重要的现实意义.

考场目标检测技术在智能化监控考生情况中发挥着重要作用. 然而, 目前国内外关于考场智能化监测技术的研究较少. 2010 年, Lu 等人[1]针对考场内的部分遮挡和背景杂波问题, 提出了基于时空形状和流相关的考生异常行为检测方法. 2017 年, 张银霞等人[2]利用背景差分法与帧差法相结合的方法来分析视频监控中的考生行为, 实现异常行为的自动识别、检测和分类. 但是这些方法均采用传统的图像处理技术检测异常情况, 存在着准确率低、计算量大、速度慢等问题, 无法满足实时精准的需求.

自 2012 年 Hinton 等人[3]提出的深度卷积神经网络(deep convolutional neural networks, DCNN) AlexNet 获得 ImageNet 挑战赛[4]冠军以来, 其优异性能引发了深度学习研究热潮, 产生了一系列基于卷积神经网络(convolutional neural networks, CNN)的通用目标检测算法. 这些算法可大致分为两阶段检测[5−7]与单阶段检测两类[8−12].

两阶段算法通常包括候选区域的生成及分类两个步骤, 代表性工作包括: 2014 年, Girshick 等人[5]提出了 RCNN (region-based convolutional neural networks)模型, 利用选择性搜索生成候选区域, 之后通过卷积神经网络提取候选区域的特征, 然后用 SVM (support vector machine)对候选区域进行分类; 次年, 为进一步降低计算复杂度, Girshick 等人[6]又提出了 Fast R-CNN 模型, 通过共享卷积特征图和采用感兴趣区域池化层(region of interest pooling, RoI-pooling)来提取各区域特征; 2017 年, Ren 等人[7]提出了 Faster R-CNN, 即利用区域候选网络(region proposal network, RPN)生成候选区域, 从而实现端到端的检测, 提高了检测速度和精度. 然而,两阶段算法仍不能满足实时类应用的速度要求.

为此, 研究人员提了 YOLO (you only look once)[8−10]和 SSD (single shot multibox detector)[11]等单阶段算法.单阶段算法利用卷积神经网络对整幅图像提取特征, 并直接预测回归目标的类别与位置. 例如, YOLO 将图像划分为多个网格, 并直接预测各个网格中目标的类别和位置[8−10]; SSD 采用 Faster R-CNN 中的先验锚框策略,并在多个尺度的特征图上共同检测目标[11], 在保证速度的同时, 达到了与 Faster RCNN[7]相当的精度; 为提高单阶段算法的检测精度, 2017 年, RetinaNet 模型中的分类损失函数 *Focal Loss* 有效地解决了正负样本不均衡问题, 显著提升了准确率[12]. 然而, 以上算法大多采用基于先验锚框(anchor)的思想, 不仅需要人工设计繁琐的锚框参数, 而且当结合多尺度架构时会变得更加复杂.

近年来, 人们提出了许多无锚框(anchor-free)的目标检测算法, 避免手工设计锚框, 适用于检测分布密集或形状差异较大的物体. 此类无锚框的目标检测算法可分为关键点检测和分割检测两类. 对于关键点检测类算法, 2018 年, Law 等人[13]提出的 CornerNet 模型通过预测包围框的左上角和右下角两个角点检测目标, 并提出了能够更好地定位包围框角点的角池化层(corner pooling). 2019 年, Zhou 等人提出的 CenterNet 模型将目标视作一个点, 通过回归得到其他目标属性, 例如尺寸、3D 位置、方向, 甚至姿态[14]. 同年, Zhou 等人提出的 ExtremeNet 模型通过预测目标的上下左右这 4 个极值点标记目标, 有效避免了强行使用矩形框包围物体带来

的问题[15]. 但是这些基于关键点检测方法的训练时间较长, 也无法很好地检测分布密集的目标. 对于分割检测类算法, Huang 等人在 2015 年提出的 DenseBox 模型使用全卷积网络(fully convolutional network, FCN)与图像金字塔策略, 直接预测不同尺度的目标包围框和置信度[16]. 2019 年, Kong 等人提出的 FoveaBox 模型通过预测类别相关的语义图来表示目标存在的概率, 然后在每个可能存在目标的位置生成与类别无关的包围框[17]. 同年, Tian 等人提出了基于 FCN 的逐像素目标检测模型 FCOS, 采用中心度(center-ness)来抑制检测到的低质量包围框[18].

综上分析, 考虑到真实考场监控视频中考生目标分布相对密集, 且因考生就坐位置与成像设备间的距离导致成像尺寸差异大, 传统基于锚框的算法难以对其进行精确检测, 故此类无锚框的检测算法更适合于考场目标检测.

因此, 本文针对真实复杂场景下考场目标检测中存在的问题, 提出了一种基于 FCOS 的无锚框考场目标检测模型——基于级联注意力与点监督机制的全卷积单级目标检测模型. 在该模型中, 我们在 FCOS 的基础上设计了一种级联注意力模块(cascade attention module, CAM)来增强语义特征, 同时提出了一种点监督机制(point supervision mechanism, PSM)来解决 FCOS 无法处理交叠目标的问题, 以有效地提升其检测性能, 满足实时的高质量考场目标检测需求.

1 全卷积单阶段目标检测模型

FCOS (fully convolutional one-stage object detection)[18]是一种基于 FCN 的逐像素目标检测模型, 它先对输入图像进行特征提取, 然后以特征图中各个像素点为中心进行目标分类与包围框回归. 不同于常用的目标检测算法, 在 FCOS[18]的检测流程中, 无须使用预先定义的锚框或者候选区域, 有效地减少了计算复杂度. 此外, FCOS 模型采用的逐像素预测的方式, 使之适用于目标密集分布时的检测问题.

如图 1 所示, FCOS 由骨干网络、特征金字塔模块和全卷积检测头部这 3 部分组成, 其中, "/s" (s=8,16,..., 128)代表各特征层对于输入图像的下采样率, 输入图像尺寸为 1000×600. 骨干网络通常采用 ResNet[19]进行特征提取, 将提取得到的一系列特征图{C_3,C_4,C_5}输入至特征金字塔模块中进行多尺度特征融合, 最后将不同尺度的特征图{P_3,P_4,P_5,P_6,P_7}输入 Head 模块, 进行目标分类与回归.

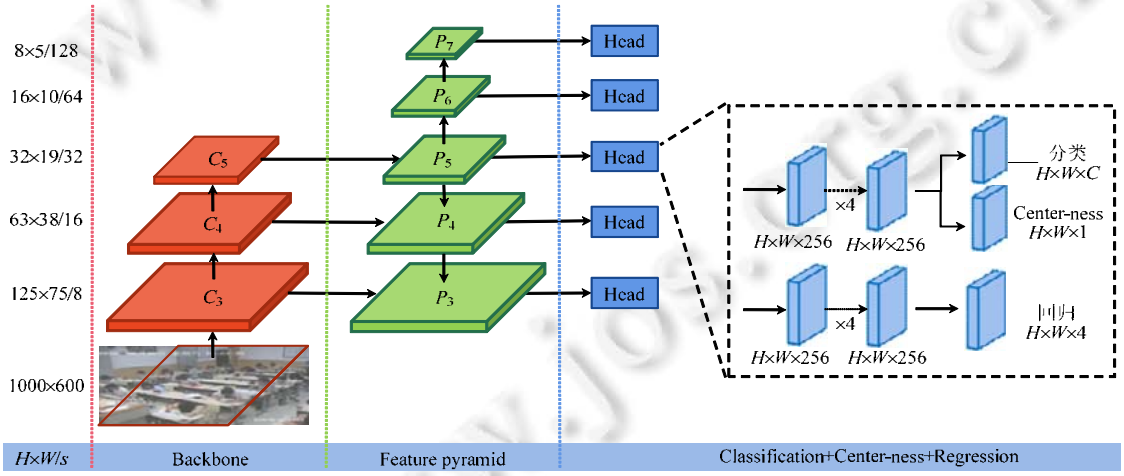


图 1 FCOS 网络结构图

Head 模块包含分类和回归两大分支, 其中, 前者包含分类得分图和 Center-ness 热力图, 后者包含距离信息图. 具体地, 将大小为 H×W×C 的多尺度特征图输入分类分支中, 经过 4 次 1×1 卷积操作后, 得到分类得分图与 Center-ness 热力图. 分类得分图包含 C 个通道, 其中, C 表示当前数据集中目标的类别数. 该图在各点预测存在各类别目标的概率, 大于置信度阈值的点被认为存在目标. 在训练阶段, 不同于传统模型需要使用先

验锚框与真实框进行 IoU (intersection over union)计算来筛选正样本，FCOS 将输入图像的所有像素都视为训练样本. 若位置(x,y)落入任何真实框(ground truth bounding box)，则认为它是一个正样本，将其类别标记为这个真实框的类别；否则记为负样本. Center-ness 热力图的通道数为 1，该图负责预测各点距所属目标中心点的距离，其中，红色、蓝色和其他颜色分别表示 1, 0 及它们之间的实数值，如图 2 所示. 在训练阶段其真实标签按公式(1)生成，距离目标中心越近，其值越高. 该模块可以有效地抑制低质量包围框:

$$center\text{-}ness=\sqrt{\frac{\min(l^*,r^*)}{\max(l^*,r^*)}\times\frac{\min(t^*,b^*)}{\max(t^*,b^*)}} \tag{1}$$

回归分支的输出是一个四通道的距离信息图，其通过预测一个 4D 向量 $\boldsymbol{v}^*=(l^*,t^*,r^*,b^*)$来回归目标位置. 其中，$l^*,t^*,r^*,b^*$表示该像素点$(x,y)$到 4 条边框的距离，如图 2 所示. 各像素点的回归目标位置表示为

$$\begin{cases} l^*=x-x_0^{(i)},\ t^*=y-y_0^{(i)}, \\ r^*=x_1^{(i)}-x,\ b^*=y_1^{(i)}-y \end{cases} \tag{2}$$

其中，$(x_0^{(i)},y_0^{(i)})$和$(x_1^{(i)},y_1^{(i)})$分别表示包围框左上角和右下角的坐标.

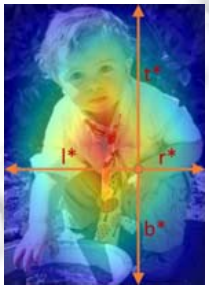


图 2 Center-ness 示意图[18]

2 基于级联注意力与点监督机制的全卷积单级目标检测模型

本文通过改进全卷积单级检测模型 FCOS[18]中的特征提取网络和引入点监督分支来提高 FCOS[18]对考场目标检测的精度和速度，整体模型结构如图 3 所示. 首先，在改进的特征提取网络中，我们增加了基于级联注意力的特征增强模块 CAM，以逐级细化增强骨干网络所提取得到的特征图$\{C_3,C_4,C_5\}$；其次，将增强后的感兴趣目标特征输入至特征金字塔模块进行多尺度特征融合，得到多尺度的特征图$\{P_3,P_4,P_5,P_6,P_7\}$；最后，为了更好地检测这些特征图中对应的交叠目标，我们在 Head 模块中引入点监督分支 PSM，实现对各目标中心点区域的学习，更准确地定位并检测密集遮挡目标，完成复杂目标群的分类与回归任务.

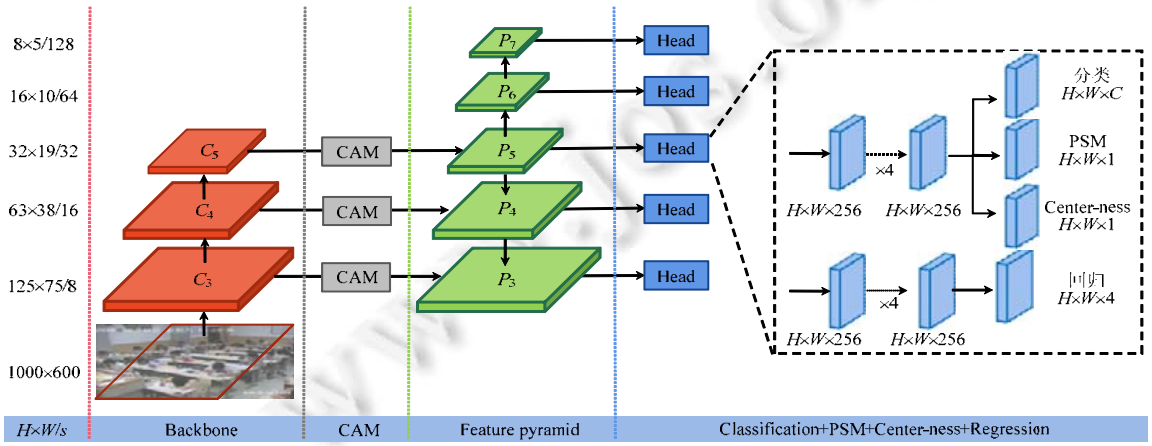


图 3 改进的网络结构

2.1 级联注意力模块

注意力在人类感知中起着重要作用. 人类的视觉处理系统(human visual system, HVS)会有选择地专注于图像的感兴趣区域, 而忽略其他信息[20]. 受此启发, 计算机视觉中注意力机制(attention module)的基本思想是: 使模型能够关注感兴趣区域的特征, 而忽略其他区域的特征. Hu 等人[21]提出的 SENet 将注意力机制引入到通道维度上, 使得模型能够自动获取各个特征通道的重要程度, 并根据该重要程度提升包含有用信息的特征通道的关注度, 同时抑制无用的特征通道, 以提升网络性能. Woo 等人[22]提出的轻量的注意力模块同时引入通道注意力机制(channel attention module)和空间注意力机制(spartial attention module), 自适应地细化特征图, 从而提升特征提取能力. Vaswani 等人[23]提出的多头注意力(multi-head attention)模块通过并联多个注意力机制的方式, 使模型在不同的特征空间中提取图像的重要特征.

为了整合不同尺度的感兴趣细节信息, 受级联思想启发, 我们利用级联结构把空间注意力机制和通道注意力机制相结合, 得到级联注意力模块 CAM. 如图 4 所示, 该模块首先以骨干网络的特征图作为输入, 通过空间注意力机制得到具有空间注意力的一级特征; 然后将该特征与原始输入特征图拼接并进行两次 1×1 卷积操作, 得到二级特征; 重复该操作, 将二级特征再次与原始输入特征图拼接并卷积,得到三级特征; 最后, 对一级特征、二级特征、三级特征进行拼接融合, 并将融合后的特征输入至通道注意力机制中, 抑制无用通道进一步增强特征.

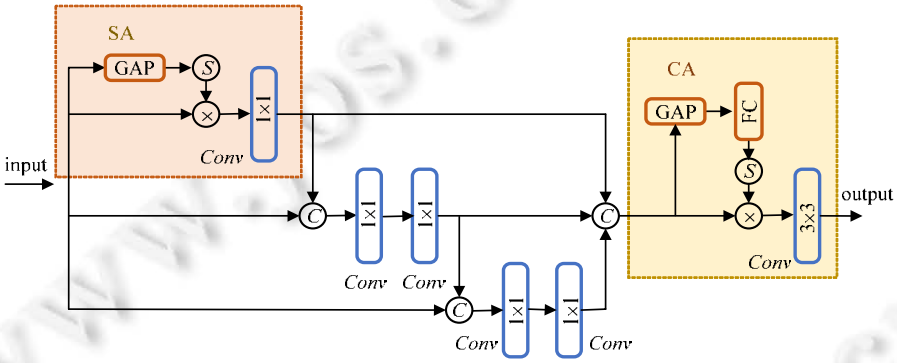


图 4 级联注意力模块

空间注意力机制通过对输入特征图进行全局平均池化与 *sigmoid* 函数激活得到空间权值图, 将该权值图与原始输入特征相乘并进行 1×1 的卷积操作后, 得到具有空间注意力的一级特征. 通道注意力机制, 同样采用全局平均池化进行特征整合, 将整合后特征输入全连接层并进行 *sigmoid* 函数激活, 得到通道特征向量, 再将该特征向量与输入特征图相乘, 进行一次 3×3 的卷积得到最终的输出, 具体定义为

$$\begin{cases} F' = SA(F) \\ F'' = Conv(F \oplus F') \\ F''' = Conv(F \oplus F'') \\ F_{output} = CA(F' \oplus F'' \oplus F''') \end{cases} \tag{3}$$

其中, F 为输入特征图, F', F'', F'''分别表示一级特征、二级特征和三级特征. SA 和 CA 分别表示空间注意力机制和通道注意力机制, 符号⊕表示特征图拼接操作, *Conv* 表示两次 1×1 的卷积操作.

受级联思想的启发, CAM 逐级修正并提取特征. 一级特征包含更多的细节语义, 而随着逐级修正, 高级特征具有更好的判别性信息. 融合 3 个级别的特征, 得到包含不同层次的特征图, 可以在保持细节语义的同时, 增强特征的判别性. CAM 采用了 ResNet[19]中 shortcut 的方法, 将原始输入特征作为先验与各级特征进行拼接, 从而避免网络在逐级修正特征时矫枉过正引入异常噪声. 将 CAM 插入骨干网络与特征金字塔网络之间, 可以使特征金字塔网络自适应地获得更显著的特征, 并且通过级联的方式反复地融合修正所提取的特征,

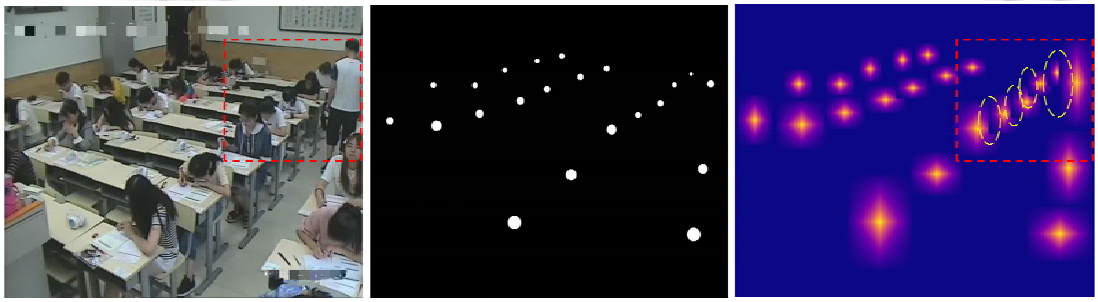
利于后续检测.

2.2 点监督机制

在 FCOS[18]中，若标注的不同目标对应的真实框重叠时，重叠部分任意位置(x,y)都将会被映射到原图中各个目标对应的真实框，则该位置被认为是模糊样本. 因此，重叠的真实框可能会在训练过程中造成难以处理的歧义. 尽管在 FCOS 中采用多级预测的方法在一定程度上缓解了该问题，但这种模糊性仍会导致模型性能下降. 另外，FCOS 中使用 Center-ness 抑制低质量包围框. 然而在教室监控场景下，目标分布密集，大多数目标会相互遮挡，仅使用多级预测和 Center-ness 的方法难以确定包围框的边界，容易造成误判. 为了更好地处理密集目标和遮挡目标，我们提出了一种点监督机制 PSM.

如图 3 所示，本文引入了一个与 Center-ness 分支并联的分支 PSM 负责预测各目标的中心点. 为避免正负样本不平衡问题，在训练阶段，我们以一定的置信度 P 认为各目标中心点附近的区域也属于该目标的中心点. 我们首先计算目标包围框中各像素距其中心点 C 的距离，计算该距离与目标框短边长度的比值，得到归一化距离，取该距离小于$(1–P)$的所有像素形成中心点区域；将中心点区域的标签置为 1，其余区域置为 0；最后，使用 Binary Cross Entropy 损失函数监督训练该分支，以增强特征图中的各目标中心点处的激活程度. 通过引入该点监督机制，能够使网络捕捉到目标的中心点及其附近的信息. 由于 FCOS 是以像素为中心来回归包围框，因此通过加强中心区域的响应，能够更好地在考生座位密集的考场监控场景下处理交叠的目标，解决考试场景下目标检测包围框回归不准确的问题.

如图 5 所示，我们根据点监督机制和 Center-ness 的生成公式(1)对图 5(a)中目标所在位置进行计算，得到图 5(b)、图 5(c)，其中，图 5(b)中的白色区域代表训练正样本且值为 1、其余区域为背景且值为 0；图 5(c)中黄色、蓝色和其他颜色分别表示 1, 0 及它们之间的值，数值越大，代表越靠近目标中心. 在图 5(a)中，真实考场中存在大量目标分布密集、遮挡严重的区域，且因该区域中相邻目标交叠比大而难以区分，造成包围框判断错误. 利用 Center-ness 虽然能够在一定程度上抑制低质量包围框，然而如图 5(c)中黄色曲线框所示的区域可以看出，在目标交叠处仍明显存在歧义，导致无法区分被遮挡目标. 在应用本文提出的 PSM 分支后，图 5(b)中显示的各目标分布清晰，较好地解决了此类问题.



(a) 考生遮挡区域较大　(b) 点监督机制的训练类标　(c) Center-ness 的训练类标

图 5　考场场景下，点监督、中心度分支类标对比

2.3 损失函数

本文提出的端到端的目标检测模型，在训练阶段，采用下式作为多任务损失函数，进行监督训练:

$$L=L_{Cls}+\lambda L_{Reg} \tag{4}$$

其中，L_{Cls} 和 L_{Reg} 分别表示分类损失和回归损失，参数λ用来表征回归损失的重要性(文中令λ=1).

(1) 分类损失：由得分图损失(focal loss)[12]、Center-ness 损失[18]和点监督损失这 3 部分组成，定义为

$$L_{Cls}=\underbrace{-\alpha(1-p_{score})^{\gamma}\log(p_{score})}_{\text{得分图损失}}\underbrace{-\log(p_{centerness})}_{\text{Centerness损失}}\underbrace{-\log(p_{point})}_{\text{点监督损失}} \tag{5}$$

其中，α为 Focal Loss 中平衡正负样本的权重因子[12]，γ为 Focal Loss 中平衡难易样本的权重因子[12]，本文后续

实验取α=0.25, γ=2.0; P_{score}, $P_{center\text{-}ness}$ 和 P_{point} 分别表示分类得分图、Center-ness 热力图以及点监督分支各个位置的预测值.

(2) 回归损失: 采用 IoU 损失[24]对回归分支进行监督, 损失函数定义为

$$L_{Reg}=-\log\frac{|Area_{predict}\cap Area_{groundtruth}|}{|Area_{predict}\cup Area_{groundtruth}|} \tag{6}$$

其中, $Area_{predict}$ 表示网络预测的包围框所在区域, $Area_{groudtruth}$ 表示真实目标框所在区域.

3 数据集及评价指标

为验证所提模型的有效性, 我们在标准化考场考生检测专用数据集和标准化考场考生扩充测试集上对其进行评估.

3.1 实验数据集

为了检验本文模型在标准化考场监控下的目标检测性能, 我们建立了标准化考场考生检测专用数据集: EPD 数据集(examinee position detection dataset)[25]. EPD 数据集的图像来自 2016–2018 年间 435 个标准化考场的监控视频, 共包括 880 幅图像, 其中, 700 幅用于模型训练, 180 幅用于模型测试.

EPD 数据集采用国际上通用的开源数据标定工具 LabelImg[26]对真实的考试场景图像进行标定, 标注结果保存为与 Pascal VOC[27]格式相同的 XML 标注文件. 为确保数据集标定的质量, 在数据标注前, 我们对标注人员进行了统一的标注规范培训. 由于标准化考场监控图像中主要目标为考生和监考教师, 因此 EPD 数据集中的目标标签设置为“person”和“background”两类. 使用 LabelImg 工具具体标定时, 一个标注框内标记且仅标记一个考生目标, 标注区域为包含桌面以上考生部分或任意位置监考教师的最小矩形框; 最后, 根据一个考试场景内所有目标的标记生成相应的标注文件. 图 6 给出了 LabelImg 工具标注的一个训练样本.

图 6 用 LabelImg 工具标注的一个训练样本示例

为了检验所提模型在实际应用中的泛化能力, 我们用 2019 年 228 个考场的监控视频形成标准化考场扩充测试集: ETD 测试集. ETD 测试集包含 710 幅图像, 全部用于目标检测模型的泛化能力测试, 见表 1.

表 1 实验数据集

	训练集/幅	测试集/幅	年份/年	考场数量/个
EPD 数据集	700	180	2016, 2017, 2018	435
ETD 测试集	–	710	2019	228

3.2 评价指标

本文拟选取平均精度(mean average precision)和每秒检测帧数(frames per second)作为考场目标检测的评

价指标, 分别表示为 mAP 和 FPS (f/s).

为了综合评价目标检测模型的准确性, 我们利用文献[27]中对 VOC 2010 使用的平均精度 mAP 作为评价指标, 计算公式为

$$mAP=\int_0^1 P(R)\mathrm{d}R \tag{7}$$

其中,

$$P=\frac{TP}{TP+FP}\times 100\%,\ R=\frac{TP}{TP+FN}\times 100\%\ .$$

TP, FP 和 FN 分别表示真阳性(true positive)、假阳性(false positive)和假阴性(false negative).

在检测速度方面, 我们采用每秒检测帧数对目标检测模型进行定量的客观评价, 计算公式为

$$FPS=\frac{FrameNum}{ElapsedTime} \tag{8}$$

为了验证级联注意力模块和点监督机制的有效性以及点监督中心点区域置信度对模型性能的影响, 实验取 5 次实验 mAP 的平均值, 并以 mAP 的平均值±mAP 标准差和每秒检测帧数 FPS 作为模型的性能定量分析指标.

4 实验结果

下面我们设计了 3 类实验, 分别验证所提模型的整体性能、级联注意力模块的有效性和点监督机制的有效性. 在这些实验中: (1) 第 4.1 节以目标检测中的经典两阶段和单阶段算法为参考, 在标准化考场数据集及其扩充数据集上进行性能对比, 以分析 SSD[11], RetinaNet[12], Faster-RCNN[7], Cascade-RCNN[28]与本文模型之间的性能差异; (2) 第 4.2 节以点监督机制中不同的中心点区域置信度为研究对象, 比较分析置信度不同的设计情况对点监督机制性能的影响; (3) 第 4.3 节以级联注意力模块为研究对象进行可视化分析, 并将其与经典注意力模块 SE 模块[21]、CBAM 模块[22]和多头注意力模块[23]进行性能对比, 最后分析验证本文模型中的级联注意力模块扩展至其他模型上的性能.

本文实验全部在 Ubuntu 16.04 系统, Pytorch 框架下实现, 硬件环境为 GPU(NVIDIA GeForce GTX 1080Ti), 内存 64 GB. 关于实验数据的预处理及参数设置主要包括: (1) 将 EPD 数据集中的图像作为输入图像, 大小调整为 1000×600; (2) 用 ImageNet 上预先训练的权重初始化骨干网络; (3) 训练阶段采用 SGD 作为模型的优化器; (4) 令模型训练轮数 *epoch*=48, *batchsize*=4, 学习率初值=0.01(第 36 轮迭代起, 其值下降 10 倍); (5) 在训练阶段, 仅使用随机水平翻转的数据增强方法, 未使用其他复杂的增强方法.

4.1 目标检测模型的整体性能

本节实验比较分析 SSD[11], RetinaNet[12], Faster-RCNN[7], Cascade-RCNN[28]与本文所提模型在 EPD 数据集上的性能差异, 结果见表 2.

表 2 目标检测模型性能比较(EPD 数据集)

网络模型	骨干网络	*mAP*(%)	检测速度(f/s)	*p* 值
Faster-RCNN[7]	ResNet50	89.4±0.321	8.2	<<0.0001
Faster-RCNN-I[26]	ResNet50	90.2±0.235	8.2	<<0.0001
Faster-RCNN-II[26]	ResNet50+FPN	91.9±0.249	14.5	<<0.0001
Cascade_RCNN[28]	ResNet50+FPN	92.5±0.141	6.3	0.0035
SSD[11]	ResNet50+FPN	89.6±0.305	20.3	<<0.0001
RetinaNet[12]	ResNet50+FPN	91.3±0.339	18.5	<<0.0001
FCOS[18](baseline)	ResNet50+FPN	91.1±0.205	22.4	<<0.0001
FCOS_with_CAM	ResNet50+FPN+CAM	91.7±0.358	22.1	0.0002
FCOS_with_PSM	ResNet50+FPN	92.4±0.297	22.4	0.0064
FCOS_with_CAM_PSM	**ResNet50+FPN+CAM**	**92.9±0.195**	**22.1**	**–**

为了保证对比实验结果的可靠性和公平性, 本文均采用 ResNet50[19]作为各模型的骨干网络, 且训练参数

设置和数据增强方式完全相同. 关于对比模型中的先验锚框参数选取, 本文采用 RetinaNet[12]中先验锚框的设置方法, 即在特征图中各点处设置 3 种长宽比例{0.5,1,2}以及 3 种大小{$(2^0,2^{1/3},2^{2/3})$×默认尺寸}的 9 个先验锚框. 根据不同的特征层, 各先验锚框默认尺寸分别设置为{$32^2,64^2,128^2,256^2,512^2$}. 为体现各个模型之间检测精度的细微差别和稳定性, 我们在表 2 最后一列给出了独立样本 T 检验的 p 值, 以对比本文模型 FCOS_with_CAM_PSM 与其他模型间的显著差异. 不难发现, 所得 T 检验结果的 p 值均小于 0.05, 说明本文模型的统计结果有显著意义且性能最优. 部分考场的目标检测结果图 7 表明, 即使在目标密集的复杂场景下, 本文模型也能够很好地检测大小不同及部分遮挡严重的目标.

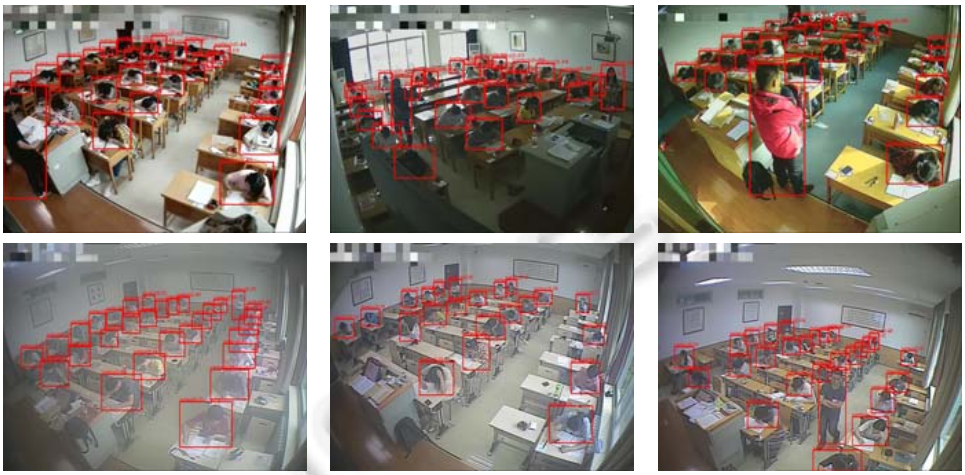


图 7 部分考场的目标检测结果

原始的 FCOS 模型在所有对比模型中取得了最快的检测速度 22.4 f/s, 但其检测精度仅为 91.1%, 明显低于两阶段算法. 下面我们对 FCOS 引入本文所提出的两个改进策略: 在 FCOS 上单独添加了 CAM 后, 其检测精度上升至 91.7%; 在 FCOS 上单独引入点监督机制 PSM 后, 其检测精度上升为 92.4%. 两个改进策略的提升显著, 且并未影响到检测速度. 同时, 引入这两个改进策略后, FCOS 的检测精度达到 92.9%, 且保持着 22.1 f/s 的检测速度, 在所有对比模型中性能最优.

与主流的单阶段算法 SSD[11]和 RetinaNet[12]相比, 本文模型检测精度更高, 且检测速度更快. 与两阶段算法相比, 本文模型在精度上分别优于 Faster-RCNN[7]模型、采用 RoI-Align[29]的 Faster-RCNN[7]模型(本文将之表示为 Faster-RCNN-I, 方便后续对比实验表达)3.91%, 2.99%, 并且速度是 Faster-RCNN[7]模型的 1.7 倍. 对于同时使用 FPN (feature pyramid networks)[30]与 RoI-Align 的 Faster-RCNN 模型(本文将之表示为 Faster-RCNN-II, 方便后续对比实验表达), 本文模型优于其检测精度 1.09%, 并且速度较该模型提高了 0.5 倍. 即便与牺牲速度而专注于精度的 Cascade-RCNN[28]相比, 本文模型精度也略高, 且检测速度是该模型的 2.5 倍. 综合考虑检测精度和每秒检测帧数两个指标, 本文模型优于其他检测模型, 更加适合真实考场目标检测任务. 此外, 我们的模型采用无锚框的方法, 无须人工设计先验锚框的大小尺寸, 且能根据不同的场景自适应调整.

为了进一步检验本文模型的优越性, 我们可视化地对比了本文模型、单阶段模型 RetinaNet[12]以及两阶段模型 Faster-RCNN-II[29]的检测结果, 如图 8 所示. 其中, 黄色虚线框表示回归不准确的区域, 红色虚线框表示误检区域.

图 8 表明, 本文采用的无锚框方法可以更好地表示不同尺寸的物体; 另一方面, 本文模型中的点监督分支与级联注意力模块使得模型可以更好地区分相邻目标, 减少背景信息干扰, 降低误检率.

为验证本文模型对全新考场环境中的目标检测效果, 以测试模型的泛化能力, 我们将于 EPD 数据集上训练的模型在 ETD 测试集上直接进行测试, 结果见表 3.

表 3 表明, 本文模型泛化能力突出,检测精度明显优于其他几种模型; 尤其与 Cascade-RCNN[28]模型相比,

检测精度接近的情况下，本文模型的检测速度是其 2.5 倍.

本文模型 RetinaNet[12] Faster-RCNN-II[29]

图 8 检测结果对比

表 3 目标检测模型性能比较(ETD 测试集)

网络模型	骨干网络	mAP (%)
Faster-RCNN[7]	ResNet50	81.5±0.402
Faster-RCNN-I[29]	ResNet50	82.1±0.297
Faster-RCNN-II[29]	ResNet50+FPN	85.0±0.288
Cascade-RCNN[28]	ResNet50+FPN	86.6±0.187
SSD[11]	ResNet50+FPN	81.4±0.327
RetinaNet[12]	ResNet50+FPN	83.8±0.607
FCOS[18](baseline)	ResNet50+FPN	82.7±0.283
FCOS_with_CAM_PSM	**ResNet50+FPN+CAM**	**86.4±0.217**

4.2 点监督中心点区域的置信度对模型性能的影响

本节实验比较不同点监督中心点区域的置信度对考场目标检测模型性能的影响. 实验设计包含中心点区域置信度的不同数值设置，记录模型在 EPD 数据集的 *mAP* 数值，结果见表 4.

表 4 点监督中心点区域的置信度对模型性能的影响

网络模型	骨干网络	置信度	mAP (%)
FCOS_PSM	ResNet50+FPN	0.7	92.1±0.245
FCOS_PSM	ResNet50+FPN	0.8	92.4±0.297
FCOS_PSM	ResNet50+FPN	0.9	91.8±0.288

表 4 表明，点监督中心点区域置信度数值从 0.9 降为 0.8 后，模型的检测精度由 91.8%上升至 92.4%；然而继续将中心点区域置信度数值从 0.8 降为 0.7 后，模型的检测精度略微下降为 92.1%. 这一现象是由于当置信度设置为 0.9 时，正样本数过少导致训练时正负样本严重不平衡，影响了网络的性能. 当置信度设置为 0.7 时，中心点区域较大. 然而考场教室场景下考生位置密集、目标相互交叠，很难区分相近的小目标影响了性能. 因此，针对本文的考场目标检测模型，我们将中心点区域置信度数值设置为 0.8.

4.3 级联注意力模块对模型性能的影响

本节实验先对级联注意力模块 CAM 的级数设置进行分析；然后对 CAM 进行可视化分析，并将其与经典的注意力模块——SE 模块[21]、CBAM 模块[22]、多头注意力模块[23]分别进行性能对比；最后，在 Faster-RCNN-II[29]和 RetinaNet[12]模型上引入 CAM，以验证其通用性.

(1) 对 CAM 中的级数设置进行性能分析，比较不同 CAM 级数对考场目标检测模型性能的影响. 实验设

计包括 CAM 级数={1,2,3,4}时检测结果的精度和速度, 结果见表 5.

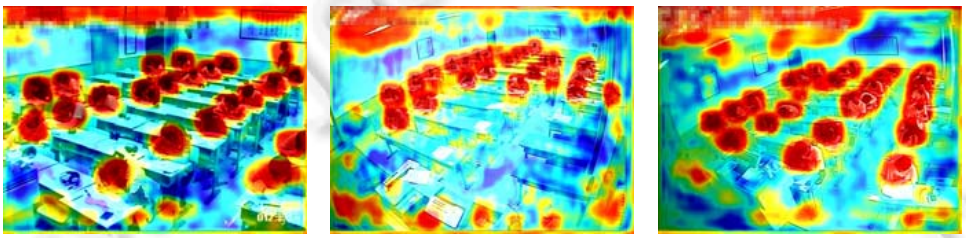
表 5 不同 CAM 级数的检测结果性能对比

网络模型	CAM 级数	骨干网络	mAP (%)	检测速度(f/s)
FCOS_with_CAM	1	ResNet50+FPN+CAM	92.2±0.327	22.6
FCOS_with_CAM	2	ResNet50+FPN+CAM	92.7±0.228	22.2
FCOS_with_CAM	3	ResNet50+FPN+CAM	92.9±0.195	22.1
FCOS_with_CAM	4	ResNet50+FPN+CAM	93.0±0.365	20.3

表 5 表明, 随着 CAM 级数的增加, 检测性能逐渐提升. 然而, CAM 级数过大时, 级数增加带来的检测精度提升幅度变小, 且导致检测速度明显降低. 因此, 综合考虑模型性能, 本文采用三级 CAM.

(2) 可视化地对比使用 CAM 的 FCOS_with_CAM_PSM 模型的特征图和不使用 CAM 的 FCOS_with_PSM 模型的特征图. 图 9 给出了分辨率较高的 P_3 层特征图的可视化对比结果.

(a) 未使用 CAM 的可视化特征图



(b) 使用 CAM 的可视化特征图

图 9 关于 CAM 有效性的可视化对比实验

图 9 充分表明, CAM 能够提升检测模型区分相邻目标和抑制噪声的能力, 使目标区域具有明显的判别性.

(3) 为验证 CAM 的优越性, 将之与 3 种经典的注意力模块[21−23]进行对比实验, 其中, 多头注意力模块中 Head 个数=3, 且各 Head 采用 Self-Attention 的形式[31], 结果见表 6.

表 6 CAM 与常见注意力模块的性能对比

网络模型	骨干网络	mAP (%)	检测速度(f/s)
FCOS_with_CAM_PSM	ResNet50+FPN+CAM	92.9±0.195	22.1
FCOS_with_SE_PSM[21]	ResNet50+FPN+SE	92.3±0.122	22.7
FCOS_with_CBAM_PSM[22]	ResNet50+FPN+CBAM	92.4±0.230	22.3
FCOS_with_Multi-Head[23,31]	ResNet50+FPN+Multi-Head	92.7±0.329	15.0

表 6 表明, 本文提出的 CAM 较 SE 模块、CBAM 模块和多头注意力模块的检测精度分别提升了 0.6%, 0.5% 和 0.2%. 这是由于 CAM 在为网络增加注意力的同时使用类残差结构, 确保更多原始信息被逐级细化, 进一步提升了检测效果. 不难发现: CAM 与多头注意力模块相比, 检测速度具有明显优势.

(4) 为验证 CAM 的通用性, 在 Faster-RCNN-II[29], RetinaNet[12]模型上引入 CAM 进行性能对比. 实验结果如表 7 所示.

表 7 表明, 在两阶段模型 Faster-RCNN-II[29]上增加 CAM 后, 检测精度提高了 0.97%, 而检测速度仅降低了 1.6 f/s; 对于引入 CAM 的单阶段算法 RetinaNet[12], 检测速度基本没有降低, 而检测精度提高了 1.3%. 我们

提出的级联注意力模块具有通用性，能够有效地细化特征图，可以便捷地插入任意检测模型中，以提高其检测精度.

表 7 CAM 对其他模型性能的影响

网络模型	骨干网络	mAP (%)	检测速度(f/s)
Faster-RCNN-II[29]	ResNet50+FPN	91.9±0.249	14.5
Faster-RCNN-II[29]	ResNet50+FPN+CAM	92.8±0.232	12.9
RetinaNet[12]	ResNet50+FPN	91.3±0.339	18.5
RetinaNet[12]	ResNet50+FPN+CAM	92.5±0.305	17.4

5 结 论

本文以标准化考场中的考生目标为研究对象，在无锚框单级检测框架 FCOS 内，提出了基于级联注意力模块与点监督机制的全卷积考场目标检测新模型. 该模型的优点主要体现在: (1) 采用无锚框的方法，避免手工设计锚框，有效地降低了计算损耗; (2) 充分考虑了考场目标的分布特征及场景复杂性，提出了级联注意力模块来逐级修正并增强特征; (3) 引入点监督机制，有效地解决了交叠多目标的检测难问题. 实验与模型分析结果表明，本文所提模型在检测考生目标时，检测准确性和检测速度性能突出，适用于复杂真实考场环境的实时目标检测任务.

在未来的工作中，我们将继续探索并优化本文模型，进一步提高更多类型目标的检测性能，并结合行为识别技术对考生动作进行异常行为分析，实现异常行为和物品检测，推动人工智能在智慧考场和智慧校园中的应用进程.

References:

[1] Yong L, Dongjian H. Video-based detection of abnormal behavior in the examination room. In: Proc. of the 2010 IEEE Int'l Forum on Information Technology and Applications. 2010. 295−298.

[2] Zhang YX, Ma XC, Yang JB, Xu XN. The examinee's abnormal behavior detection and recognition in video based on Kalman filter. Journal of Qiqihar University, 2017, 33(6): 16−19 (in Chinese with English abstract).

[3] Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. In: Proc. of the Int'l Conf. on Neural Information Processing Systems. 2012. 1097−1105.

[4] Deng J, Dong W, Socher R, Li L, Li K, Li FF. ImageNet: A large-scale hierarchical image database. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2009. 248−255.

[5] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2014. 580−587.

[6] Girshick R. Fast R-CNN. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 1440−1448.

[7] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137−1149.

[8] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 779−788.

[9] Redmon J, Farhadi A. Yolo9000: Better, faster, stronger. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 7263−7271.

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

[11] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC. SSD: Single shot multiBox detector. In: Proc. of the European Conf. on Computer Vision. 2016. 21−37.

[12] Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 2980−2988.

[13] Law H, Deng J. Cornernet: Detecting objects as paired keypoints. In: Proc. of the European Conf. on Computer Vision. 2018. 734−750.

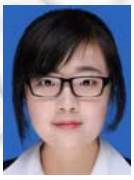
[14] Zhou X, Wang D, Krhenbühl P. Objects as points. arXiv: 1904.07850, 2019.

[15] Zhou X, Zhuo J, Krhenbühl P. Bottom-up object detection by grouping extreme and center points. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 850−859.
[16] Huang L, Yang Y, Deng Y, Yu Y. Densebox: Unifying landmark localization with end to end object detection. arXiv: 1509.04874, 2015.
[17] Kong T, Sun F, Liu H, Jiang Y, Shi J. Foveabox: Beyond anchor-based object detector. arXiv: 1904.03797, 2019.
[18] Tian Z, Shen C, Chen H, He T. FCOS: Fully convolutional one-stage object detection. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2019. 9627−9636.
[19] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 770−778.
[20] Larochelle H, Hinton GE. Learning to combine foveal glimpses with a third-order Boltzmann machine. In: Proc. of the Int'l Conf. on Neural Information Processing Systems. 2010. 1243−1251.
[21] Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-excitation networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 7132−7141.
[22] Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional block attention module. In: Proc. of the European Conf. on Computer Vision. 2018. 3−19.
[23] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. In: Proc. of the Int'l Conf. on Neural Information Processing Systems. 2017. 5998−6008.
[24] Yu J, Jiang Y, Wang Z, Cao Z, Huang T. Unitbox: An advanced object detection network. In: Proc. of the ACM Int'l Conf. on Multimedia. 2016. 516−520.
[25] Tao LL. Research on the detection method of students in examination classroom based on SSD [MS. Thesis]. Xi'an: Shaanxi Normal University, 2020 (in Chinese with English abstract).
[26] Tzutalin. LabelImg. Git code. 2015. https://github.com/tzutalin/labelImg
[27] Everingham M, Gool LV, Williams CKI, Winn J, Zisserman A. The pascal visual object classes (VOC) challenge. Int'l Journal of Computer Vision, 2010, 88(2): 303−338.
[28] Cai Z, Vasconcelos N. Cascade R-CNN: Delving into high quality object detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 6154−6162.
[29] He K, Gkioxari G, Dollar P, Ross G. Mask R-CNN. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 2961−2969.
[30] Lin TY, Dollar P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 936−944.
[31] Zhang H, Goodfellow I, Metaxas D, Odena A. Self-attention generative adversarial networks. In: Proc. of the Int'l Conf. on Machine Learning. 2019. 7354−7363.

附中文参考文献:

[2] 张银霞, 马小川, 杨季彪, 徐雪南. 基于卡尔曼滤波的考生异常行为检测与识别. 齐齐哈尔大学学报(自然科学版), 2017, 33(6): 16−19.
[25] 陶丽丽. 基于 SSD 的考场考生检测方法研究 [硕士学位论文]. 西安: 陕西师范大学, 2020.

田卓钰(1996－), 女, 硕士, 主要研究领域为目标检测, 场景分析.



杨楷芳(1987－), 女, 博士, 副教授, CCF 专业会员, 主要研究领域为视频压缩编码, 视频/图像质量评估.

马苗(1977－), 女, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为图像处理, 机器学习, 灰色理论和群体智能的应用.