计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(3):159−168 [doi: 10.15888/j.cnki.csa.008359]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

改进 YOLOv5 的油田作业现场安全着装小目标检测①

田　枫, 贾昊鹏, 刘　芳

(东北石油大学 计算机与信息技术学院, 大庆 163318)
通信作者: 刘　芳, E-mail: 942857522@qq.com

摘　要: 针对油田作业现场监控视频中的工人安全着装小目标检测效果较差的问题, 提出了改进 YOLOv5 的油田场景规范化着装检测方法 Cascade-YOLOv5 (C-YOLOv5). 首先搭建 YOLO-people 与 YOLO-dress 级联的小目标检测网络, 定位行人目标, 然后裁剪出行人区域并进行尺度变换, 最后对行人进行安全着装检测; 为了充分融合浅层与深层特征信息, 在各级网络中使用 4 个不同尺度的卷积特征层来预测待检测目标. 最后在原始图像中用不同颜色的框标出行人以及行人的着装部件类别, 从而判定行人是否着装规范. 实验证明, 相比原始 YOLOv5 算法, C-YOLOv5 方法不仅满足实时性的要求, 而且检测的 mAP 提升了 2.3%. 同时, 融合了深浅层信息的改进方法有效地增强了特征的表征能力, 提高了小目标的检测精度.

关键词: 改进 YOLOv5; 着装检测; 多尺度融合; 小目标检测; 级联网络

引用格式: 田枫,贾昊鹏,刘芳.改进 YOLOv5 的油田作业现场安全着装小目标检测.计算机系统应用,2022,31(3):159–168. http://www.c-s-a.org.cn/1003-3254/8359.html

Small Target Detection in Oilfield Operation Field Based on Improved YOLOv5

TIAN Feng, JIA Hao-Peng, LIU Fang

(School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)

Abstract: Given the poor performance on the small target detection of clothing safety in video surveillance for oilfield operation, this study proposes a standardized clothing detection method based on Cascade-YOLOv5 (C-YOLOv5), an improvement from YOLOv5. Firstly, a small target detection network cascading with YOLO-people and YOLO-dress is built to locate the pedestrian target. Then the pedestrian area is cut out and transformed in scale to detect the clothing safety of pedestrians. To fully integrate the shallow and deep feature information, this study adopts four convolutional feature layers with different scales to predict the undetected targets. Finally, in the original image, different color frames are used to mark the types of pedestrians and their clothing parts, determining whether the pedestrians are dressed properly. Experimental results show that compared with the original YOLOv5 algorithm, the C-YOLOv5 method not only meets the real-time requirement but also improves the detection mAP by 2.3 percentage points. At the same time, the improved method of fusing deep and shallow information effectively enhances the representation ability of features and promotes the detection accuracy of small targets.

Key words: improved YOLOv5; clothing detection; multi-scale fusion; small target detection; Cascade network

① 基金项目: 黑龙江省自然科学基金 (LH2021F004); 黑龙江省高等学校教改工程 (SJGZ20200037); 东北石油大学研究生教育创新工程 (JYCX_11_2020); 黑龙江省省属本科高校基本科研业务费 (KYCXTD201903);东北石油大学引导性创新基金 (2020YDL-11);黑龙江省优秀青年科学基金 (YQ2020D001); 黑龙江省教育科学规划重点课题 (GJB1421113)
收稿时间: 2021-05-08; 修改时间: 2021-06-08; 采用时间: 2021-06-21; csa 在线出版时间: 2022-01-24

© 中国科学院软件研究所 http://www.c-s-a.org.cn

油田作业现场中工作人员的着装不规范问题经常出现, 例如油田工人工作时未佩戴安全帽、未穿着劳保服等, 从而导致一些不安全的事件时而发生, 因此, 规范化安全着装检测在推进企业智能化安全管理、保障人员生命财产安全等方面起着至关重要的作用[1]. 对于规范化安全着装检测任务, 由于油田作业现场的监控摄像头覆盖范围广、拍摄距离远, 画面中人员目标的尺寸较小, 导致着装部件目标所占的像素数较少、携带的特征信息很弱, 提高了检测难度. 因此, 提高小目标检测精度是规范化安全着装检测任务的关键. 本文针对油田作业现场规范化安全着装小目标检测问题, 基于 YOLOv5 设计了一种 YOLO-people 和 YOLO-dress 级联的小目标检测网络, 先定位行人目标进行尺度变换后再对行人进行安全着装检测. 同时结合多尺度特征融合的思想, 将 YOLOv5 中原来的 3 尺度检测扩展为 4 尺度检测, 充分融合深浅层特征信息, 进一步提高小目标检测精度, 降低漏检率.

1 相关工作

当前目标检测算法主要分为两阶段方法和单阶段方法. 经典的两阶段方法[2–4] 首先利用区域候选网络来提取候选目标信息, 再利用检测网络对候选目标集合进行位置和类别的判定, 具有很高的检测精度, 但由于两阶段方法推理速度慢、处理效率低, 很难达到实时性检测的要求. 单阶段检测算法如 SSD[5] 和 YOLO[6,7] 系列, 这些算法不需要对特征区域进行提取, 只需在特征图上进行边界框的回归, 仅仅一步即可预测最终结果, 算法推理速度能够达到实时. 这些方法在通用数据集上已经取得了较好的检测效果, 但对于某些特定场景比如油田作业现场, 由于监控探头拍摄范围广、目标距离远, 导致人员身上的着装部件在监控画面下所占像素较少, 携带的特征信息十分有限, 缺少明显的纹理、形状等特征, 使得通用的目标检测算法在规范化安全着装检测任务上表现不佳, 在进行小目标检测具有一定的局限性. 因此, 主流的目标检测方法在处理规范化着装检测任务尤其是针对小目标的检测, 仍有较大的提升空间.

目前基于深度学习的目标检测算法普遍针对具有一定比例或尺寸的中大型目标, 难以适用复杂背景下的小目标, 小目标检测的难点总结成以下 3 个方面.

(1) 深度神经网络特征提取和小目标尺寸之间的问题. 对于如油田作业现场等特定场景下的监控探头, 由于拍摄范围广、目标距离远, 导致待检测目标尺寸较小, 而且多数目标检测网络由于池化层的存在, 在特征提取时能够提取到的特征信息逐渐减少, 导致深层特征对小目标的表达能力较弱, 检测效果欠佳.

(2) 小目标检测公共数据集较少且目标尺寸分布不均匀. 现有的目标检测算法大多基于 Pascal VOC 或 COCO 数据集进行小目标检测研究, 数据集中小目标的分布十分不均匀且占比不多, 导致在训练过程中难以充分学习小目标物体的特征, 对小目标的泛化能力弱、检测精度较低.

(3) 相关检测算法缺乏通用性. 目前对小目标检测的研究大多应用于特定场景下, 如疫情期间对人脸口罩的识别、无人机视角检测地理空间小目标, 道路交通方面识别交通标志等, 由于检测任务的多样性和复杂性使算法在实际应用中面临很大挑战, 难以迁移.

目前针对小目标检测, 基于单阶段的检测算法[8–10] 大多通过多尺度特征融合的方式充分利用包含较多细节信息的高分辨率底层特征, 卷积神经网络能够自动学习不同层次结构的图像信息特征, 底层特征保留了很多局部的图像细节信息如轮廓、边缘和纹理等, 有助于更好地进行目标定位. 深层特征图感知细节信息的能力较差, 但包含了更多深层次的语义信息, 有助于更好地进行目标分类. 多尺度特征融合采用自顶向下的横向连接的方式融合底层特征与深层特征, 融合后的特征具有更强的描述性, 有利于小目标的检测. Lin 等人[11] 提出特征金字塔网络 (FPN), 结构如图 1 所示. 深层特征图经过上采样后与较低层特征图做元素级别的相加, 上采样时采用双线性插值法增强了网络提取多尺度特征的能力. FPN 结构作为基础模块, 在各类目标检测算法中被普遍使用并取得了极大成功; Woo 等人[12] 改进了 FPN 上采样方式, 在深层特征图进行上采样时用反卷积层代替双线性插值法并获得了更好的性能; Li 等人[13] 将注意力机制引入 FPN 结构, 使经过全局池化后的深层特征对底层特征图的通道进行加权, 来引导不用层级信息之间的融合; 郑秋梅等人[14] 改进 YOLOv3 中的 FPN 结构, 将其在 3 个尺度检测增加至 4 个尺度, 充分融合提取的底层特征信息和深层语义信息, 同时改进原有的损失函数, 最终提升对车辆小目标的识别效果.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

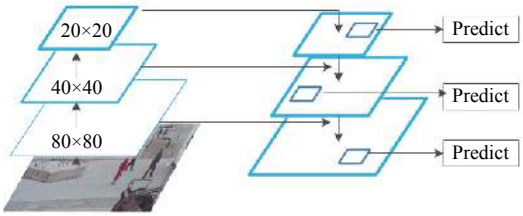

图 1 FPN 结构

针对规范化安全着装小目标检测问题, 分析小目标检测的特点, 本文提出了 Cascade-YOLOv5 网络 (C-YOLOv5), 首先搭建 YOLO-people 与 YOLO-dress 级联的小目标检测网络, 定位行人目标, 然后裁剪出行人区域并进行尺度变换, 最后对行人进行安全着装检测; 同时采用多尺度特征融合方法, 更改 YOLOv5 中的 PANet 结构, 采用 4 个不同尺度的特征层对目标进行预测, 使底层与深层特征信息充分融合, 有效增强小目标的检测效果; 并采用旋转、裁剪、缩放等数据扩充方式处理数据集, 最终提升小目标检测准确率, 减少漏检、误检.

2 YOLOv5 原理

由于油田作业现场规范化着装检测任务对小目标的检测精度要求较高, 本文经过在速度和精度方面的衡量, 最终选择 YOLOv5 作为本文方法的基础模型, 其图像推理速度最快达到 0.007 s, 即每秒可处理 140 帧, 满足油田作业现场视频监控图像实时检测需求, 同时有非常轻量级的模型大小, 适合监控视频下目标检测任务的模型部署. YOLOv5 的网络结构主要由输入端、Backbone、Neck 和 Prediction 四部分组成, YOLOv5 网络结构如图 2 所示.

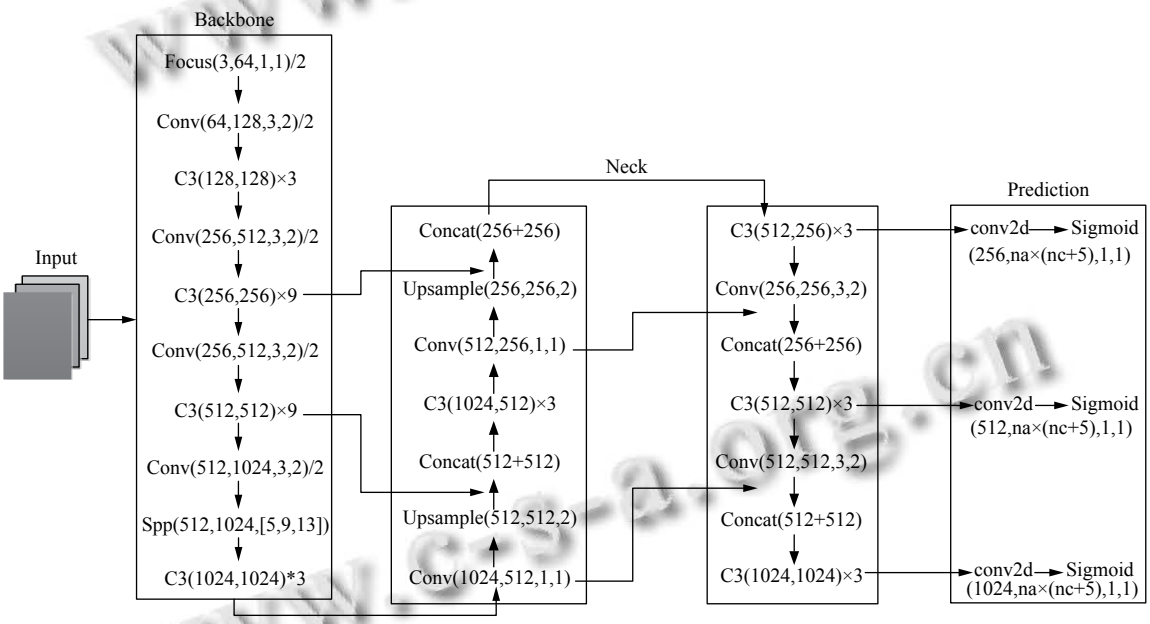

图 2 YOLOv5 网络结构

在输入端包括 Mosaic 数据增强、图片自适应缩放和自适应锚框 3 个部分, 其中 Mosaic 增强方法能够有效提高小目标检测效果, 适用于规范化安全着装检测任务.

在 Backbone 特征提取网络部分, 包括 CSP 结构和新增的 Focus 结构. YOLOv5 中设计了两种 CSP 结构, CSP1_X 结构应用于 Backbone 主干网络中, 另一种 CSP2_X 结构则应用于 Neck 中, 二者结构具体如图 2 中所示. Focus 结构中比较关键的是切片操作, 原始 640×640×3 的图像, 先复制 4 份, 然后通过切片操作将这 4 个图片切成了 4 个 320×320×3 的切片, 然后使用 Concat 从深度上连接这 4 个切片, 输出为 320×320×12, 之后再通过卷积核数为 32 的卷积层, 生成 320×320×32 的输出, 最后经过 Batch_norm 和 Leaky_ReLU 将结果输入到下一个卷积层. 提高了特征图上每个点的感受野, 减少原始信息的丢失且减少了计算量加快了检测速度.

在 Neck 结构中采用了 PANet 结构, 由 FPN+PAN

© 中国科学院软件研究所 http://www.c-s-a.org.cn

结构组成. FPN 利用自上而下的方式通过对特征图进行上采样融合. PAN 结构采用自底向上的方式通过对特征图进行下采样融合, PANet 结构如图 3 所示.

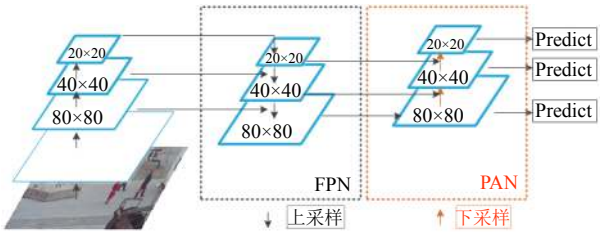

图 3 PANet 结构

在输出端 YOLOv5 采用了 GIoU_Loss 来作为损失函数, 并采用加权 NMS 来解决遮挡严重导致检测不准确的问题. GIoU_Loss 如式 (1) 所示:

$$L_{GIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} \tag{1}$$

其中, b与b^{gt}表示预测框与真实框的中心点, $\rho(\cdot)$表示欧式距离, c表示预测框与真实框的最小外接矩阵的对角线距离. 同时, YOLOv5 通过 FPN+PAN 结构, 最终形成 3 个不同尺度的特征层来预测待检测目标, 以 640×640 分辨率的图像为输入, 输出 3 个尺度规格分别为 20×20, 40×40, 80×80 的特征图, 这种采用多尺度特征融合的检测方法在检测图像中不同尺寸的目标时具有较好的鲁棒性, 但仅利用 3 个尺度的特征层进行预测, 并不能充分的利用底层特征信息, 从而会导致小目标位置信息丢失, 不利于油田作业现场着装部件小目标检测.

针对 YOLOv5 的不足之处以及油田作业现场安全着装小目标检测任务的特点, 本文提出了改进的 C-YOLOv5 方法, 最后通过对比实验验证本文方法的有效性.

3 Cascade-YOLOv5 级联网络算法原理

3.1 Cascade-YOLOv5 整体网络搭建

对实际油田作业场景进行分析, 安全着装检测任务有如下几个特点: 一是由于实际油田作业场景中监控摄像头摆放位置较高, 导致拍摄画面中作业人员目标较小, 且监控摄像头长期暴露在空气中, 监控画面会有轻微的模糊现象, 都会给规范化着装检测带来一定的难度; 二是由于作业人员身上的着装部件相对于作业人员目标尺寸更小, 例如安全帽、作业上衣、作业裤子、作业鞋子、手套等部件, 这也会给检测的准确性带来影响; 三是由于外在环境的影响, 如光照变化、天气变化、场景遮挡以及监控视角的变化等这些因素也会使小目标检测的难度变得非常大.

基于以上这些考虑, 由于通用目标检测算法对中大型目标有较好的检测效果. 因此对比于直接在监控画面中检测较小尺寸的着装部件, 先进行作业人员较大目标检测, 再进行着装检测能够有效地缓解部分小目标难以被检测到的问题. 于是本文以 YOLOv5 为基础网络来搭建安全着装小目标检测的级联网络 Cascade-YOLOv5 (C-YOLOv5). 该网络由行人检测网络 YOLO-people 和着装检测网络 YOLO-dress 两部分组成. C-YOLOv5 具体检测流程如图 4 所示.

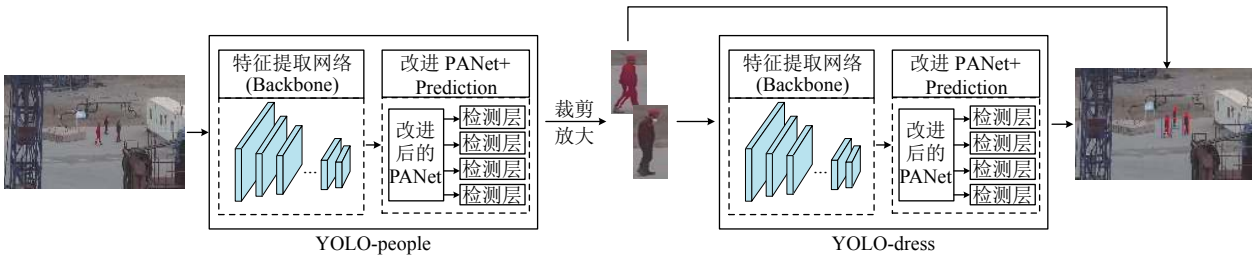

图 4 C-YOLOv5 检测流程

以摄像头源流为输入, 首先利用 YOLO-people 网络将输入图片的像素放缩至 640×640, 将整幅图片划分为 $S \times S$ 个网格, 分别在 20×20、40×40、80×80 和 160×160 四种不同大小特征图上进行预测, 在每个特征图上采用跨邻域网格匹配策略, 从目标中心点所在的当前网格的上下左右 4 个网格中找到离目标中心点最近的两个网格, 再加上当前网格共 3 个网格进行匹配, 产生更多的 anchor 边界框, 每个网格中都会预测 3 个不同尺寸的边界框, 每个边界框会产生一个 $5+C$ 维的向量, 其中 5 表示 5 维向量, 包含边框坐标 (4 个数值), 类别置信度 (1 个数值). C 表示预测类别总数.类别置信度计算方法如式 (2) 所示:

© 中国科学院软件研究所 http://www.c-s-a.org.cn

$$Confidence = Pr(object) \times IOU_{\text{pred}}^{\text{truth}} \tag{2}$$

其中, $Pr(object)$是指某个网格中是否存在目标对对象, 如果存在则为 1, 不存在则为 0; $IOU_{\text{pred}}^{\text{truth}}$是指目标预测框和真实框的交集和并集的比例. 当网络检测完成时, 特征图将被重新映射至原图中, 同时绘制出预测框, 在原图中预测框(*DetectionResult*)和真实标注框 (*Ground-Truth*) 之间的面积重合度将以像素为坐标值来计算, 这一过程如式 (3) 所示:

$$IOU_{\text{pred}}^{\text{truth}} = \frac{DetectionResult \cap GroundTruth}{DetectionResult \cup GroundTruth} \tag{3}$$

如果网格中包含目标, 网格还需要预测该目标属于第 i 类的概率$Pr(class_i|Object)$, 即目标分类条件概率. 通过计算目标类别条件概率$Pr(class_i|Object)$与预测框置信度(*Confidence*)的乘积可得目标位置及类别置信度, 能够综合评价预测边界框的性能. 具体计算如式 (4) 所示:

$$Pr(class_i|Object) \times Confidence = Pr(class_i|Object) \times IOU_{\text{pred}}^{\text{truth}} \tag{4}$$

同时, 损失函数使用 CIoU_Loss[15] 作为 BoundingBox 回归的损失, 如式 (5) 所示:

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha\upsilon \tag{5}$$

b与b^{gt}表示预测框与真实框的中心点, υ用来描述预测框和真实框长宽比的比例一致性, 如式 (6) 所示:

$$\upsilon = \frac{4}{\pi^2}\left(\arctan\frac{\omega^{gt}}{h^{gt}} - \arctan\frac{\omega}{h}\right)^2 \tag{6}$$

α是权重函数, 用来平衡比例, 如式 (7) 所示:

$$\alpha = \frac{\upsilon}{(1 - IoU) + \upsilon} \tag{7}$$

CIoU_Loss 函数增加了相交尺度的衡量方式, 有利于解决边界框有时不重合的问题, 同时考虑了边界框中心点距离的信息与边界框宽高比的尺度信息, 使网络会有更快更好的收敛效果.

最后, 选择置信度最高的预测边界框作为结果的检测框, 同时去除冗余框, 即使用了非极大值抑制 (NMS)[16] 的方法. 然后, 将 YOLO-people 检测出的行人目标区域根据其左上角和右下角坐标, 从原始图像中裁剪并进行放大处理, 同时对每一个放大后的行人目标区域做锐化、对比度增强等处理, 增强行人目标的边界信息表示, 有利于规范化着装检测. 然后将每个行人区域送入 YOLO-dress 网络进行规范化着装检测, 检测流程和 YOLO-people 几乎相同, 最后, 在原始的输入图像中, 用不同颜色的框标出行人以及行人的着装部件类别.

3.2 Cascade-YOLOv5 子网络搭建

对于背景复杂的图像, 在进行目标检测时小目标在图像中只占用很少量的像素点, 检测网络从这些像素中能提取出的语义信息非常有限. 根据图像梯度上升法[17], 利用不同层提取出的特征对原始图像进行重构, 得出富含细节信息的低层特征比富含语义信息的高层特征可以更好地协助目标检测的结论.

本文以 YOLOv5L 版本的网络结构为基础, 搭建 YOLO-people 与 YOLO-dress 级联的小目标检测网络, 每级网络从多尺度融合方式进行改进. 原始 YOLOv5 通过 3 个不同尺度的特征图进行目标物体的检测, 输出 3 个尺度规格分别为 20×20、40×40、80×80 的特征图, 但是仅利用 3 个尺度的特征, 对浅层信息的利用并不充分, 会导致部分小目标信息丢失. 因此针对 C-YOLOv5 两级子网络, 分别将原来的 3 尺度检测扩展为 4 尺度, 构成 4 个检测分支结构, 如图 5 所示, 输入尺寸为 640×640, 每个分支分别从 Backbone主干网络中提取特征, 将从主干网络中提取的深层特征图在 FPN 中进行上采样并与网络前期相应尺寸的底层特征图由深入浅融合成有效信息, 实现了 Concat 连接操作. 但是在 FPN 网络中会存在一个问题, 浅层特征图在向深层传递信息时, 难以与高层特征图进行融合, 因此, 在 FPN 特征金字塔网络的基础上实现 PAN 网络结构, 增加了自底向上 (bottom-up) 路径, 将 FPN 网络中融合后的特征图继续进行自底向上的下采样融合, 实现特征图的反向融合, 这样会得到更丰富的特征信息. 利用 PAN 结构的 bottom-up 路径将 FPN 网络中融合后的特征图再次自底向上融合并进行下采样操作, 实现特征图的反向融合, 最后在 4 个尺度的融合特征图上分别做独立的检测. 改进后的多尺度融合可以从浅特征层中学习较强的位置特征, 经过 PAN 结构的 bottom-up 路径再次融合使得深层特征可以进行更确切的细粒度检测. 通过融合更多尺度的浅层特征信息, 增强路径聚合网络 (PANet) 的特征表达能力, 提升小目标的检测精度, 降低漏检率.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

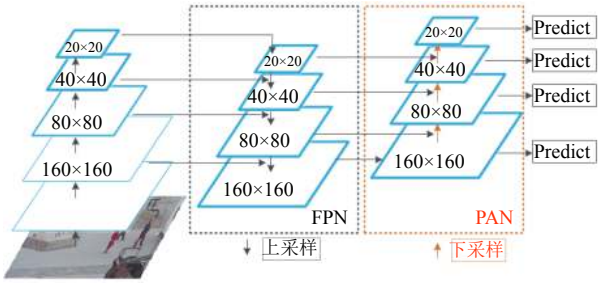

图 5 改进 PANet 的多尺度特征融合方式

YOLO-people 和 YOLO-dress 的网络整体结构基本相同, 如图 6 所示. 差别在于检测的目标和网络的输入分辨率有所不同, YOLO-people 用于检测行人目标, 网络输入分辨率为 640×640, YOLO-dress 用于检测着装部件, 网络输入分辨率为 320×320, 图像信息量的减少加快了检测速度. 最后, 由于 anchor box 的数量和大小对检测的精度和速度有直接影响, 因此采用 K-means 聚类方式生成更适合着装部件检测的锚点框, 通过 K-means 聚类选取合适的 IOU 分数, 可以在模型召回率和精确度之间取得平衡. 根据 IOU 和 anchor box 的关系, 针对 C-YOLOv5 两级子网络分别生成 12 个 anchor box.

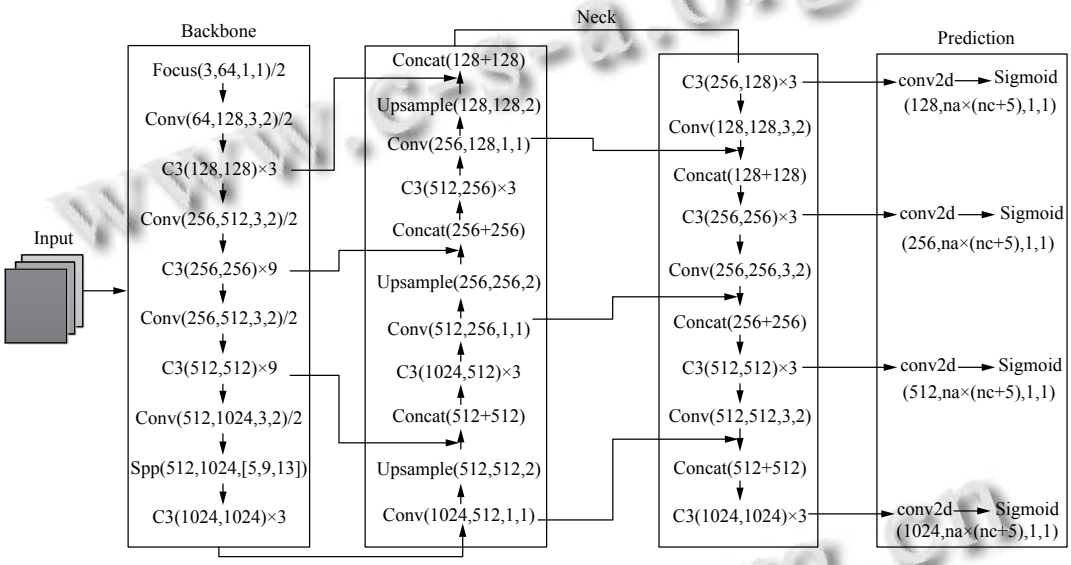

图 6 YOLO-people (YOLO-dress) 整体网络结构

4 实验与结果分析

4.1 数据集构建

4.1.1 数据集预处理

由于油田作业现场要求佩戴安全帽、身穿红色劳保服, 故本文所采用的数据集全部源于某油田作业现场多个监控视角下所拍摄的监控画面, 对每一个摄像头设置每间隔半个小时抓取一张图片, 最后再删除部分没有工人的图片. 为了提高数据集的泛化性, 对 4 500 张样本采取图像反转、对比度变化等数据增强方式, 这样可以有效地避免了训练模型过拟合问题, 提高模型泛化能力, 避免了因数据样本量少导致的训练不充分的问题, 最终获得 9 000 张数据样本图, 用于行人检测网络 YOLO-people 的训练. 着装检测网络 YOLO-dress 的训练图像是根据每幅图像中行人的标记框坐标, 从原始图像中切割出行人目标区域再进行尺度变换得到, 最终每幅图像中都只有一个行人, 共 27 000 张样本图. 行人检测网络和着装检测网络部分训练样本图如图 7、图 8 所示.

4.1.2 图片标注

本文依据规范化安全着装检测目标, 将设计行人、安全帽、非安全帽、安全上衣、非安全上衣、安全裤子、非安全裤子共 7 类检测标签, 参照 PASCAL VOC[18] 目标检测数据集标准格式, 首先对每张训练样本中出现的行人目标进行标记, 采用 Labelimg 图像标注工具标注外围框, 将外围框的左上角坐标 (x_{min}, y_{min})、右上角坐标 (x_{max}, y_{max}) 以及目标类别 person 等相关信息进行记录, 得到行人检测数据集后, 根据标记框坐标将每张训练样本图中的行人目标进行裁剪并放大, 得

© 中国科学院软件研究所 http://www.c-s-a.org.cn

到每幅图像只有一个行人, 对得到的每幅图像依据上述方式进行其余 6 个类别的标注, 最终得到着装检测数据集. 将两类数据集分别按 8:2 划分为训练集和测试集, 行人和 6 类着装类别目标在两类数据集中所占图像数量分布如图 9 所示.

图 7 工人检测网络训练样本图

图 8 着装检测网络训练样本图

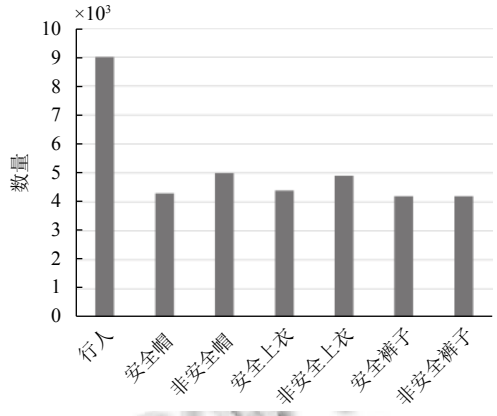

图 9 数据分布量

4.2 网络训练

本文训练和测试实验均在以下工作条件下完成: Ubuntu 20.04 系统; CPU 处理器 Intel Xeon Gold 6244; 运行内存 128 GB; 显卡为 NVIDIA Tesla V100. 采用基于 PyTorch 框架的 YOLOv5 进行多尺度融合方法改进, 并利用改进后的 YOLOv5 网络搭建 YOLO-people 和 YOLO-dress 网络进行级联, 从而实现安全着装检测, 训练时需要对两个子网络单独进行训练. YOLO-people 采用行人数据集训练, YOLO-dress 采用着装数据集训练, 通过 K-means 聚类算法, 在行人数据集上, YOLO-people 选用 4 个尺度的 12 个 anchor box 分别为 (11, 25)(14, 44)(23, 22), (30, 52)(32, 80)(34, 53), (45, 99)(53, 66)(86, 104), (106, 171)(151, 244)(170, 321). 在着装数据集上, YOLO-dress 选用 4 个尺度的 12 个 anchor box 分别为 (5, 6)(7, 9)(12, 10), (10, 13)(16, 30)(33, 23), (47, 52)(75, 79)(96, 102), (116, 90)(148, 187)(372, 324). 以训练 YOLO-people 为例, 输入分辨率设置为 640×640, 初始学习速率设置为 0.000 5, 速度衰减因子为 0.000 5, batchsize 设置为 64, NMS 设置为 0.5, 每迭代 100 次保存一次模型, 最后修改网络训练相关参数开始模型训练.

图 10 为两级子网络训练的损失值收敛曲线, 可以看出损失值随着迭代步数降低, 迭代步数为 290 左右时, 损失值趋于稳定, YOLO-people 网络训练 loss 值为 0.024, YOLO-dress 网络训练 loss 值为 0.005.

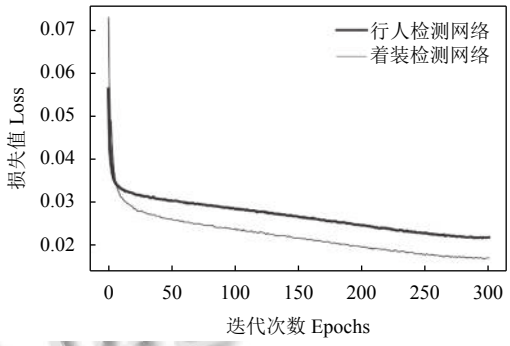

图 10 C-YOLOv5 子网络训练损失值曲线

4.3 评价指标

本文对于规范化安全着装检测任务采用查准率 (*Precision*)、查全率 (*Recall*)、平均精度 (*AP*)、平均精度均值 (*mAP*) 及每秒传输帧数 (FPS) 作为模型的评价指标.

查准率 (*Precision*) 衡量的是所有预测为正样本的结果中, 预测正确的比率, 在规范化安全着装检测任务中 *Precision* 代表模型正确预测某类着装部件的个数 (TP) 占预测该类着装部件总个数 ($TP+FP$) 的比值, 该指标如式 (8) 所示:

$$Precision = \frac{TP}{TP+FP} \tag{8}$$

查全率 (*Recall*) 衡量的是所有正样本中被预测正

© 中国科学院软件研究所 http://www.c-s-a.org.cn

确的比率. 在规范化安全着装检测任务中 *Recall* 代表模型正确预测某类着装部件的个数(TP)占该类着装部件总数($TP+FN$)的比值, 该指标如式 (9) 所示:

$$Recall = \frac{TP}{TP+FN} \tag{9}$$

平均精度 (*AP*) 用来计算单类别模型平均准确度, 对于目标检测任务, 每一类都可以计算出 *Precision* 和 *Recall* 并得到一条 P-R 曲线, 曲线下的面积就是 *AP* 的值. 平均精度均值 (*mAP*) 就是对所有类别的 *AP* 值求平均值, N 代表类别总数. 该指标如式 (10) 所示:

$$mAP = \frac{\sum AP}{N} \tag{10}$$

4.4 模型评估

实验 1. 利用 YOLOv5 算法模型和 YOLO-people、YOLO-dress 算法模型进行对比实验, 其中 YOLO-people 和 YOLO-dress 算法表示利用了 4 个尺度的多尺度特征融合检测方式进行实验, YOLOv5 没有进行任何改动. 在自制行人检测数据集和着装检测数据集上进行实验, 采用 *mAP* 和 FPS 作为评价指标. 测试结果如表 1, YOLO-people 模型在保证实时检测的情况下, 相比 YOLOv5 模型的 *mAP* 提高了约 0.7%, YOLO-dress 模型相比 YOLOv5 的 *mAP* 提升了约 0.6%.

表 1 C-YOLOv5 子网络对比实验

算法模型		传输速率 (f/s)	*mAP* (%)
检测工人	YOLOv5	90.45	92.12
	YOLO-people	82.60	92.82
检测着装	YOLOv5	81.93	91.75
	YOLO-dress	70.46	92.35

图 11(a) 和图 11(b) 为 YOLOv5 和 YOLO-people 在同一测试图像上的检测效果对比. 从图中可以看出, YOLO-people 网络可以检测出更多的小目标行人. 图 12(a) 和图 12(b) 为 YOLOv5 和 YOLO-dress 在同一测试图像上的检测效果对比. 从对比图像中看出, 原始 YOLOv5 对于画面中两个人的安全帽佩戴情况没有检测出来, 而 YOLO-dress 模型全部检测准确, 相比 YOLOv5 的 3 尺度检测效果更好. 以上消融实验说明改进的 4 尺度模块对小目标检测的准确率有一定的提升.

实验 2. 为了验证级联网络检测方式对小目标检测的有效性, 本文的级联网络测试实验采用两个原始 YOLOv5 网络进行实验, 在检测出行人目标的基础上检测着装部件, 在本文的工人检测数据集和着装检测数据集上进行实验, 采用 *mAP* 和 FPS 作为评价指标, 测试结果如表 2 所示. 实验证明, 通过级联网络检测的方式, 在保证实时的情况下, *mAP* 相比于 YOLOv5 算法提升 1.05%, 由此证明, 级联网络检测方式对于小目标的检测效果提升确实有效.

(a) YOLOv5

(b) YOLO-people

图 11 YOLO-people 检测效果测试

(a) YOLOv5

(b) YOLO-dress

图 12 YOLO-dress 检测效果测试

表 2 级联网络对比实验

算法模型	YOLOv5	YOLOv5级联
传输速率 (f/s)	81.93	42.48
mAP (%)	91.75	92.80

© 中国科学院软件研究所 http://www.c-s-a.org.cn

图 13(a) 和图 13(b) 为 YOLOv5 和 YOLOv5 级联算法对同一图像的测试结果. 从图中可以看出, 原始 YOLOv5 算法没有检测出画面中一个人的安全帽和工作上衣的穿着情况, 而级联检测方式对画面中的行人都进行了正确的规范性着装检测, 无漏检情况.

(a) YOLOv5

(b) YOLOv5-级联

图 13 级联网络对比测试

实验 3. 通过整合多尺度融合改进和级联网络检测方式, 最终形成本文的 C-YOLOv5. 在整体测试中, 本次实验对比了 YOLOv5 算法在自制着装数据集上的检测效果, 根据 FPS 和 *mAP* 评价指标评估本文方法的有效性, 如表 3 所示.

表 3 整体改进对比实验

算法模型		YOLOv5	C-YOLOv5
传输速率 (fps)		81.93	36.24
AP (%)	helmet	91.90	94.83
	no helmet	91.45	94.27
	safe wear	92.10	93.89
	no safe wear	92.04	93.45
	safe trousers	92.63	94.38
	no safe trousers	90.38	93.36
mAP (%)		91.75	94.03

通过对比实验结果表明, 原始 YOLOv5 在自制着装数据集上的传输速率达到了 81.93 f/s, 检测精度 *mAP* 达到了 91.75%. 而本文提出的 C-YOLOv5 方法在自制数据集上的传输速率达到 36.24 f/s, 检测精度 *mAP* 达到了 94.03%, 相比于 YOLOv5 提高了约 2.3%, 这是因为 C-YOLOv5 在改进了多尺度融合方式的同时采用了级联检测, 不仅增强了小目标的特征信息, 而且通过级联的方式极大减少了背景干扰等影响因素, 降低了小目标检测的难度, 进一步提升了油田作业现场安全着装小目标的检测效果.

图 14 展示了本文 C-YOLOv5 算法在自制数据集上的检测效果与 YOLOv5 算法的对比. 从对比效果图中可以看出, 对于规范化着装检测任务, C-YOLOv5 算法在监控视角拍摄距离较远、人员目标尺寸较小的情况下, 能够很好地检测出油田工人的规范着装情况, 在精确度和查全率方面具有较好的效果, 在小目标检测中表现较好.

(a) YOLOv5

(b) C-YOLOv5

图 14 整体改进对比测试

5 结语

针对油田作业现场的监控视频中人员安全着装小目标检测问题, 本文提出了一种改进的 Cascade-YOLOv5 (C-YOLOv5) 方法, 该算法首先针对原始 YOLOv5 算法进行改进多尺度特征融合方式, 从原来的 3 尺度检测扩展为在 4 个尺度的融合特征图上分别做独立的检测, 改进后的模型充分利用了浅层特征信息并融合高层语义信息, 增强了路径聚合网络的特征表达能力, 提高了小目标的检测精度. 然后在此基础上, 用两个多尺度特征融合改进后的 YOLOv5 算法进行级

© 中国科学院软件研究所 http://www.c-s-a.org.cn

联, 先检测工人目标, 然后将目标进行尺度变换处理后进行着装检测, 进一步提高远距离视频监控中的安全着装检测效果. 本文方法在自制数据集上与 YOLOv5 算法的对比实验, 证明该方法既能保证实时检测速度, 也提高了小目标检测精度. 本文方法最终实现了油田现场对于工作人员的智能化、自动化的规范化安全着装检测, 对油田作业现场安防智能监控发展有着重要的价值, 且本文方法适用于不同场景下的小目标检测任务.

参考文献

1 刘欣宜, 张宝峰, 符烨, 等. 基于深度学习的污染场地作业人员着装规范性检测. 中国安全生产科学技术, 2020, 16(7): 169–175.

2 Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of 2014 IEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 580–587.

3 Girshick R. Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 1440–1448.

4 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]

5 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot multibox detector. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 21–37.

6 Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. arXiv: 1804.02767, 2018.

7 Redmon J, Farhadi A. YOLOv3: An incremental improvement. IEEE Conference on Computer Vision and Pattern Recognition. 89–95.

8 Cui LS, Ma R, Lv P, *et al*. MDSSD: Multi-scale deconvolutional single shot detector for small objects. Science China Information Sciences, 2020, 63(2): 120113. [doi: 10.1007/s11432-019-2723-1]

9 Fu CY, Liu W, Ranga A, *et al*. DSSD: Deconvolutional single shot detector. IEEE, 2017. 2881–2890.

10 Li Z, Zhou F. FSSD: Feature fusion single shot multibox detector. arXiv:1712.00960, 2017.

11 Lin TY, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 936–944.

12 Woo S, Hwang S, Kweon IS. StairNet: Top-down semantic aggregation for accurate one shot detection. Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision. Lake Tahoe: IEEE, 2018. 1093–1102.

13 Li HC, Xiong PF, An J, *et al*. Pyramid attention network for semantic segmentation. Proceedings of British Machine Vision Conference 2018. Newcastle: BMVA Press, 2018. 285.

14 郑秋梅, 王璐璐, 王风华. 基于改进卷积神经网络的交通场景小目标检测. 计算机工程, 2020, 46(6): 26–33.

15 Zheng ZH, Wang P, Ren DW, *et al*. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. IEEE, 2021.

16 Neubeck A, van Gool L. Efficient non-maximum suppression. Proceedings of the 18th International Conference on Pattern Recognition. Hong Kong: IEEE, 2006. 850–855.

17 Mahendran A, Vedaldi A. Understanding deep image representations by inverting them. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 5188–5196.

18 Everingham M, van Gool L, Williams CKI, *et al*. The Pascal visual object classes (VOC) challenge. International Journal of Computer Vision, 2010, 88(2): 303–338. [doi: 10.1007/s11263-009-0275-4]

© 中国科学院软件研究所 http://www.c-s-a.org.cn