计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(2):190−198 [doi: 10.15888/j.cnki.csa.008923]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于改进 YOLOv3 的集卡车头防砸检测①

张柏阳[1], 赵　霞[1], 包启睿[2]

[1](同济大学 电子与信息工程学院, 上海 201804)
[2](利物浦大学 计算机科学学院, 利物浦 L693BX)
通信作者: 赵　霞, E-mail: zhaoxia@tongji.edu.cn

摘　要: 在自动化港口集装箱起重机作业流程中, 集卡车头防砸检测是不可或缺的一个环节. 针对在此环节采用人工确认方法效率低和基于激光扫描方法耗费高、系统复杂的问题, 本文提出一种基于作业场景视频图像和深度学习的算法对集卡车头进行目标检测. 建立集卡车头样本数据集, 采用 DCTH-YOLOv3 检测模型, 通过模型迁移学习方法进行样本训练. DCTH-YOLOv3 模型是本文提出的一种改进 YOLOv3 算法模型, 该算法改进了 YOLOv3 的 FPN 结构提出一种新的特征金字塔结构—AF_FPN, 在高、低阶特征融合时通过引入具有注意力机制的 AFF 模块聚焦有效特征、抑制干扰噪声, 提高了检测精度. 另外, 使用 CIoU loss 度量损失替代 L2 损失, 提供更加准确的边界框变化信息, 模型检测精度得到进一步提升. 实验结果表明: DCTH-YOLOv3 算法在 GTX1080TI 上检测速率可达 46 fps, 相比 YOLOv3 算法仅降低了 3 fps; 检测精度 $AP_{0.5}$ 为 0.9974、$AP_{0.9}$ 为 0.4897, 其中 $AP_{0.9}$ 相比 YOLOv3 算法提升了 16.4%. 本研究算法相比 YOLOv3 算法, 精度更高, 更能满足自动化作业对集卡防砸检测高精度、快识别的要求.

关键词: 计算机视觉; 集卡车头检测; 深度学习; DCTH-YOLOv3; 智能制造; 目标检测

引用格式: 张柏阳,赵霞,包启睿.基于改进 YOLOv3 的集卡车头防砸检测.计算机系统应用,2023,32(2):190–198. http://www.c-s-a.org.cn/1003-3254/8923.html

Anti-crash Detection of Container Truck Head Based on Improved YOLOv3

ZHANG Bo-Yang[1], ZHAO Xia[1], BAO Qi-Rui[2]

[1](College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China)
[2](Department of Computer Science, University of Liverpool, Liverpool L693BX, the United Kingdom)

Abstract: In the process of automated crane operations for port containers, the detection of container truck heads is an indispensable link. To solve the problem of low efficiency by manual confirmation and high costs and complex systems by the laser scanning method, this study proposes an algorithm based on video images of operation scenes and deep learning for target detection of container truck heads. Specifically, upon the construction of a sample data set of container truck heads, the DCTH-YOLOv3 detection model is used, and sample training is performed through the method of model migration learning. The DCTH-YOLOv3 model is an improved YOLOv3 model proposed in this study. The algorithm improves the FPN structure of YOLOv3 and proposes a new feature pyramid structure—AF_FPN. During the fusion of higher- and lower-order features, the AFF module with the attention mechanism is introduced to focus on effective features and suppress interference noise, which increases the accuracy of detection. In addition, the metric CIoU loss is used to replace L2 loss to provide more accurate boundary box change information and further improve the model detection accuracy. The experimental results indicate that the detection rate of DCTH-YOLOv3 can reach 46 fps on GTX1080TI, which is only 3 fps lower than that of YOLOv3. The detection accuracy can reach $AP_{0.5}$ 0.9974 and $AP_{0.9}$ 0.4897, in which $AP_{0.9}$ is 16.4% higher than that of YOLOv3. Compared with the YOLOv3 algorithm, the proposed algorithm has higher accuracy and can better meet the requirements of automatic operations for high accuracy and fast identification in

① 收稿时间: 2022-06-14; 修改时间: 2022-07-12; 采用时间: 2022-07-29; csa 在线出版时间: 2022-09-26
CNKI 网络首发时间: 2022-11-15

the anti-collision detection of container trucks.

Key words: computer vision; detection of container truck head; deep learning; DCTH-YOLOv3; intelligent manufacturing; object detection

随着港口自动化技术的不断成熟, 自动化集装箱码头正逐渐取代传统码头. 这也是未来港口的重点发展方向. 自动化集装箱门式起重机 (以下简称“自动化轨道吊”) 是自动化码头的主力集装箱装卸设备之一, 其工作效率影响整个码头的运营. 目前, 当自动化轨道吊卸箱到集卡车上时, 需要人工进行安全确认以免砸损集卡车头. 这样一来, 完整的自动化流程则会被迫中断, 进而严重影响作业效率. 因此, 一套高效稳定的集卡防砸检测系统是提高自动化轨道吊作业效率的关键.

基于二维激光扫描检测技术[1] 是目前常用的一种集卡防砸检测方法. 该技术通过激光传感器可获取到集卡的轮廓特征并以此计算出集卡车头的位置以及车头与传感器的距离, 从而确定落箱的安全区域. 该套系统在理想环境下, 正确识别率能达到 95% 以上, 但也存在缺点: 首先, 该套系统的位置计算模型建立在激光射出的扫描平面为绝对水平的理论基础上, 但在实际作业中存在诸多扰乱水平的因素, 如设备磨损、安装支架松动、轨道沉降等; 其次, 该套系统中的激光传感器价格昂贵, 多车道检测时更是需要安装多个设备才能满足要求; 最后, 激光传感器在大雨、大雪、大雾、灰尘环境下, 容易出现误检. 针对上述方案稳定性不足的问题, 胡荣东等人[2] 提出一种基于三维激光的检测方案, 通过三维激光雷达采集吊具下方集装箱的三维点云和吊具的运动姿态参数, 再根据姿态参数对三维点云进行转换得到全面点云并在点云中确定集装箱下落区域的范围, 最后通过判断该范围内是否存在障碍物进行防砸检测. 但该方法对硬件设备要求较高. 为了平衡研发成本与经济效益, 张俊阳等人[3] 提出一种基于深度学习的视觉检测方案, 通过摄像头截取包含吊具和集卡车头的图片, 再将图片送入 Mask R-CNN 神经网络算法[4] 进行集卡车头定位. 但该方法并不适合此场景: 一方面, Mask R-CNN 网络是一种基于像素级的监督学习检测算法, 在网络的训练过程中需要大量的语义分割标签样本, 然而这种标签样本在特殊场景任务中获取是非常困难的; 另一方面, Mask-RCNN 在单块 TitanX 上的检测速度只能达到 8 fps 左右, 无法满足该场景下实时检测任务的要求.

针对以上问题, 本文提出一种新的神经网络算法 DCTH-YOLOv3 (detection of container truck head based on YOLOv3) 从而实现集卡车头的识别与定位. DCTH-YOLOv3 通过改进 YOLOv3 网络[5], 在 FPN 结构[6] 中引入注意力融合模块 AFF (attention fusion feature) 将高层特征与低层特征进行融合时, 聚焦和强化有效特征、抑制干扰特征, 获得一种新的特征金字塔结构 AF_FPN (attention fusion feature pyramid network), 该结构虽然在一定程度上降低了网络的检测速度但有效地提升了网络的检测精度; 另外, 针对目标位置回归损失函数, 采用 CIoU loss[7] 度量损失替换 L2 损失, 提供更加准确的边界框的数据信息, 从而进一步提高了模型的检测精度. 在测试数据集上, 本文所提算法 $AP_{0.5}$ 为 0.99, $AP_{0.9}$ 为 0.49 相比于 YOLOv3 算法提升了近 16.4% 左右.

1 目标检测算法

目标检测技术通过从图像中定位出所有感兴趣的目标 (物体) 并识别出它们所属种类, 该技术是计算机视觉技术的一个重要分支. 早期的目标检测技术主要依靠统计学的方法提取特征, 例如 HOG[8]、SIFT[9] 等方法对检测图像进行特征提取, 再使用 SVM[10]、AdaBoost[11]、DPM[12] 等分类检测器对上述提取的特征进行目标检测. 然而不是所有的物体都能设计出有效的特征提取方法, 因此这种方法在工程应用中存在很大的局限性. 深度学习则在训练过程中利用数据本身自动学习如何提取有效特征, 这种方法有效地打破了手动设计特征提取方法的局限, 从而极大地促进了目标检测技术的发展. 目前这种基于深度学习的目标检测方法已成为目标检测技术的主流研究方向. 近年来, 基于卷积神经网络[13] 的目标检测方法更是取得了突破性进展, 已被广泛应用于自动驾驶、视频监控、机器人视觉、工业自动化等领域.

基于深度学习的主流目标检测算法大致分为两类: 一类称为二阶段算法, 这种算法基于区域建议算法, 先从图片中找出包含检测目标的区域作为预选框, 再将预选框内的特征信息送入下一阶段的卷积神经网络进

行目标的分类和位置回归, 其代表算法有 R-CNN[14]、Fast-RCNN[15]、Faster-RCNN[16] 等; 另一类称为一阶段算法, 该算法将整幅图送入神经网络, 直接检测出图中目标的类别及相关位置信息, 其代表算法有 SSD[17]、YOLO 系列[18] 等. 前者算法的检测精度较高, 但检测的速度更慢; 后者的检测精度较差, 但检测速度能够达到实时检测的要求. 鉴于集卡车头检测任务实时高精度的要求, 本文设计的网络选取 YOLOv3 为基础. YOLOv3 是一种一阶段的目标检测算法, 集检测精度高和速度快于一身, 已被广泛应用于工业检测任务[19–21].

2 集卡车头检测算法设计

集卡车头目标检测算法的流程如图 1 所示, 主要包括准备数据、模型搭建、训练优化、实验分析. 准备数据包括图像采集、样本标注、数据增强; 模型搭建和训练优化部分主要为模型创建、网络训练、性能评估; 在实验分析部分通过测试数据集对训练好的模型进行验证分析.

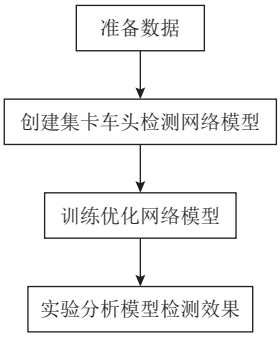


图 1 集卡车头防砸检测算法流程

2.1 数据

本文为一个露天环境下的检测, 因此图片中背景相对复杂; 另外检测目标的规格也并不统一, 集卡车头的形状、颜色、尺寸、结构、特征存在很大差异. 如图 2 部分数据图像所示, 第 1 行包括白天、黑夜、雨天等多种天气场景, 第 2 行中的集卡车头只有后半部分, 存在目标被遮挡的情况, 第 3 行中的场景受不同程度的阴影和光照的干扰.

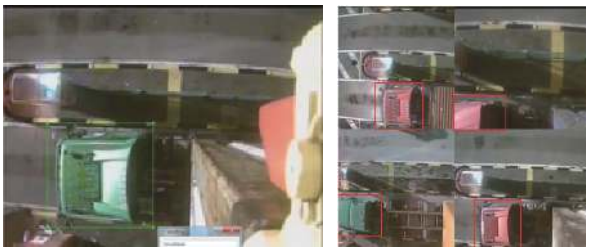
图 2 部分数据图像

该任务为一个特殊工业场景下的检测, 没有可以直接使用公开数据集. 为确保数据的多样性, 本文从不同作业场地的视频源中采集了各种场景的作业图像, 共计 1 124 张. 使用 LabelImg 工具对样本进行人工标注 (如图 3(a) 所示), 并将标注后的样本集按照 80%、10%、10% 的比例随机划分为训练数据集、验证数据集和测试数据集. 其中验证数据集用于帮助网络在训练过程中选择合适的参数, 测试数据集用于比较不同模型的检测效果. 虽然训练数据集包含 900 张图片, 但对于实际检测任务来说远远不够, 为防止过拟合 , 本文采用数据增强的方法对训练数据集进行扩充, 扩充后训练集图片达到 18 万张左右, 增加训练样本多样性: 包括随机水平镜像翻转、随机色域扭曲、随机尺寸缩放、随机长宽扭曲和 Mosaic 方法[22]. 其中 Mosaic 数据增强的方法如图 3(b) 所示, 将从训练集中随机选取 4 张图片拼接组合为一张新的图片, 这种方法丰富了图片的背景信息. 另外, 图片被输入网络前, 将分辨率统一调整为 416×416 大小.

2.2 网络结构

YOLOv3 主要由 Darknet53 主干特征提取网络和预测网络两部分组成, 结构如图 4 所示 (输入图片尺寸为 416×416). 整个网络主要由一系列的 1×1 和 3×3 的 DBL 卷积模块交替构成. 在主干特征提取网络中使用 5 个 RES 残差模块 (个数分别为 1、2、8、8、4). 在预测网络部分, 利用 FPN 网络 (feature pyramid network) 结构, 将主干特征提取部分得到的 8 倍、16 倍、32 倍下采样特征图进行一系列的卷积、上采样、和拼接操作, 得到大小为 13×13、26×26 和 32×32 的预测框, 实现多尺度预测.

YOLOv3 在进行特征融合时, 只是将两种特征在通道方向上进行拼接融合. 如果能依据检测任务特点, 在融合时对高层语义特征信息和低层空间特征信息各有侧重, 融合效果会更好. 基于这个思想, 本文提出一种改进的 YOLOv3 算法模型 DCTH-YOLOv3. 该模型使用具有注意力机制的 AFF 模块来融合高、低阶特征, 聚焦有效特征、抑制干扰噪声, 得到一种新的特征金字塔结构 AF_FPN (attention fusion feature pyramid network).



(a) 数据标注 (b) Mosaic 效果

图 3 数据集标注与 Mosaic 数据增强

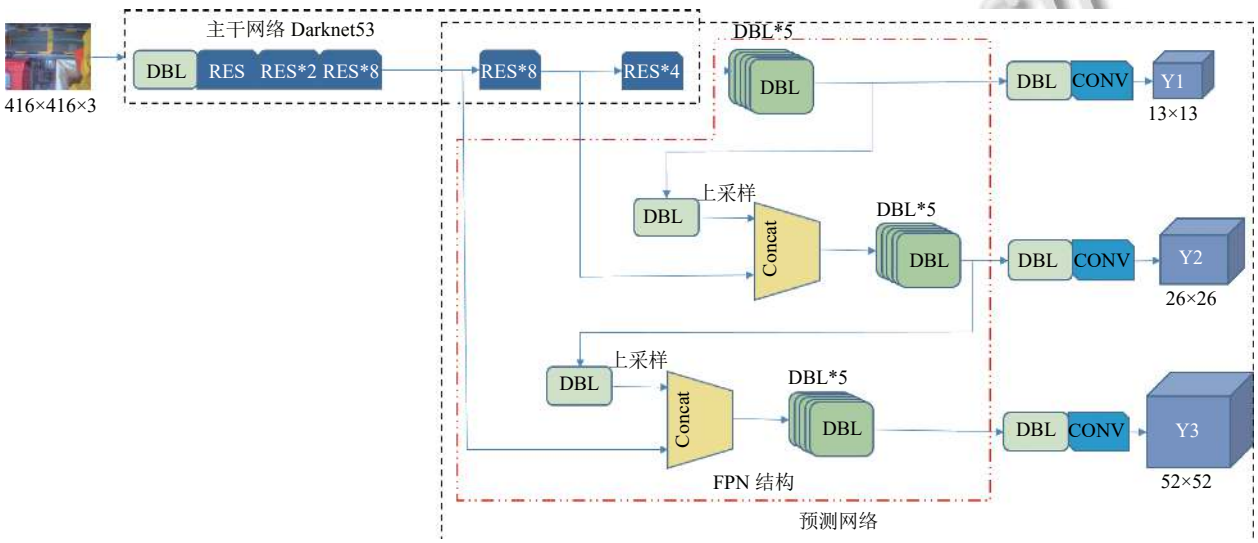


图 4 YOLOv3 网络结构

与 YOLOv3 网络一样, DCTH-YOLOv3 算法模型 (如图 5 所示) 主要由主干特征提取网络 Darknet53 和预测网络两部分组成, 预测网络部分包括 Y1、Y2、Y3 三个层级分支, 实现多尺度预测. 为了获得更加合理的融合特征, 本文提出的 AF_FPN 结构在预测网络部分的 Y2、Y3 分支中: 首先, 将主干特征提取部分得到的

8 倍、16 倍下采样的特征分别经过一个 1×1 的 DBL 模块进行升维, 得到的特征大小分别为 52×52×512、26×26×1024; 其次, 将 Y1、Y2 分支中获得的高阶特征 (DBL*5 模块的卷积结果, *5 在此表示紧接 5 个 DBL 模块) 先经过一个 1×1 的 DBL 卷积再进行上采样, 上采样后的特征大小分别为 26×26×1024、52×52×512; 最后, 将尺寸和通道数相同的高阶特征和低阶特征一起送入 AFF 模块处理, 获得新的融合特征.

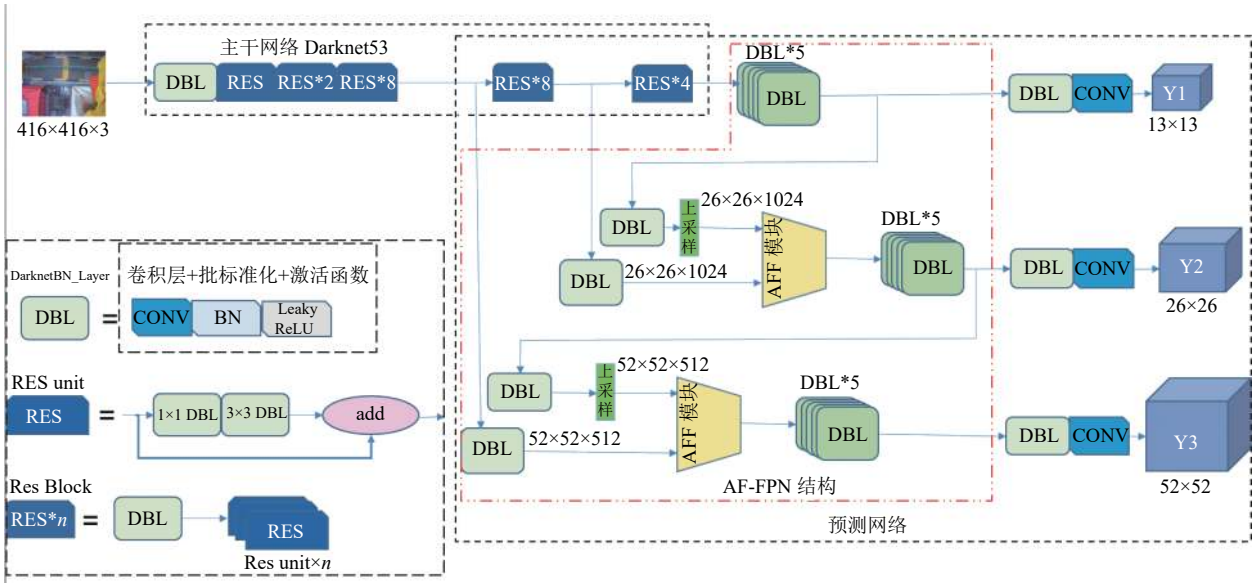


图 5 DCTH-YOLOv3 网络结构

特征融合是来自不同层或分支的特征的组合, 是现代网络体系结构中重要的一部分. 它通常通过简单线性操作来实现, 例如求和 (summation) 或串联 (concatenation). AFF (attentional feature fusion) 是一种具有学习能力的注意力特征融合模块[23], 利用 MS-CAM (multil-scale channel attention module) 以一种动态的学习方式将两种特征进行融合. AFF 的结构框图如图 6(a) 所示, 先将特征 X 和特征 Y 线性相加融合获得一个新的特征, 再将这个新特征送入 MS-CAM 模块提取注意力权重, 最后根据下列公式获得注意力融合特征 Z:

$$Z = M(X \oplus Y) \odot X + (1 - M(X \oplus Y)) \odot Y \tag{1}$$

其中, $\oplus$表示初始特征线性融合, 将两个特征相对应的元素依次求和; $M(\cdot)$ 表示经多尺度通道注意力模块 (MS-CAM) 处理得到的融合权重, 为 0–1 之间的数值; $(1–M(\cdot))$ 为图 6(a) 中虚线操作, 使得网络在 X 和 Y 之间选择对任务更为重要的特征进行聚焦; $\odot$表示特征加权融合, 将初始特征与融合权重逐元素相乘.

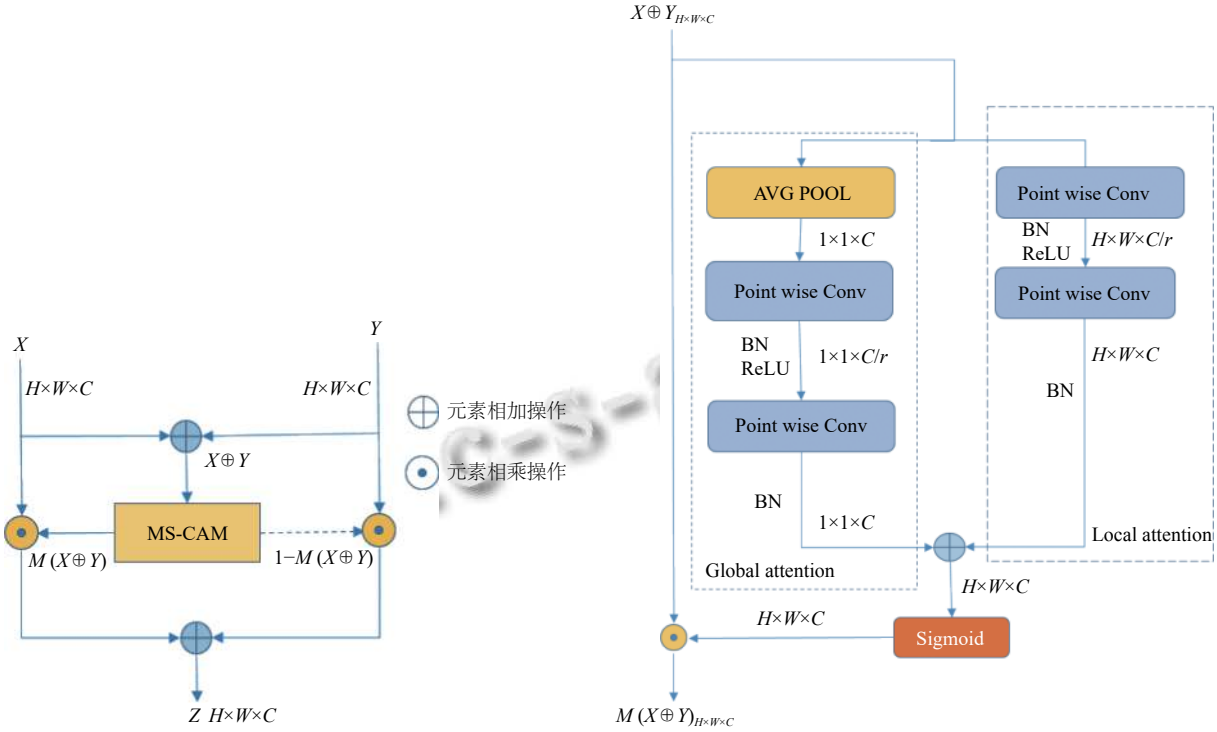
多尺度通道注意力模块 MS-CAM (multil-scale channel attention module), 通过两个不同的分支来提取通道注意力. 其工作原理如图 6(b) 所示, 由局部特征注意力通道和全局特征注意力通道组成, 并在空间上使用 Attention 进行多尺度特征融合. 在局部特征注意力通道中: 先使用 1×1 点卷积将输入特征$X \oplus Y$通道数减小 r 倍, 之后紧接一个 BN 层和 Leaky ReLU 层; 再用 1×1 点卷积将通道数恢复成与原输入特征一致, 后面再接一个 BN 层; 1×1 的逐点卷积能够有效地关注通道的尺度问题. 全局特征注意力通道与局部特征注意力通道唯一不同的是对输入特征先进行一次 GAP (global average pooling) 操作. 最后, 将计算之后的权重值对输入特征做注意力操作, 得到新的特征$M(X \oplus Y)$, 公式如式 (2):

$$M(X \oplus Y) = (X \oplus Y) \odot \sigma(L((X \oplus Y)) \oplus G((X \oplus Y)) \tag{2}$$

其中, $L(\cdot)$ 表示局部通道获得的权重信息, $G(\cdot)$ 表示全局通道获得的权重, $\oplus$表示将两种权重信息相加, $\odot$表示两个特征图对应元素相乘, $\sigma(\cdot)$表示 Sigmoid 激活函数. 目前在卷积神经网络中常用的注意力机制的注意力权值都是通过全局通道注意力机制生成的, 例如 SENet[24] 和 ResNeSt[25]. 这种机制忽略了局部特征信息, 对小目标和位置回归的检测效果没有 MS-CAM 好.

图 5 中 DBL (DarknetBN layer) 模块是 YOLO 系列网络的标准卷积模块, 自 YOLOv2[26] 开始, YOLO 将归一化、加速收敛和避免过拟合的方法变为 BN (batch

normalization), 并将 BN 层和 Leaky ReLU 层接在每一卷积层之后. RES 模块为一种残差网络结构, 子段由卷积核大小分别为 1×1 和 3×3 的 DBL 模块组成, 使用 shortcut 的连接方式使每个小段利用残差结构进行训练, 这种残差结构能够有效地解决深层网络在训练时出现梯度消失的问题[27], 从而提高网络的学习能力.



(a) AFF 模块 (b) MS-CAM 模块

图 6 注意力特征融合模块

2.3 损失函数

在 YOLOv3 中, 使用 L2 损失来计算预测框和真实框之间的位置和大小差异. 由于定义边框信息的这 4 个值并不是相互独立的, 仅使用 4 个距离值 (框的中心坐标信息和框的长宽信息) 来衡量预测框的好坏并不精确. 为了提升网络的定位准确性, DCTH-YOLOv3 在训练时将目标的回归损失函数设计为更加合理的 CIoU loss. CIoU loss 在计算回归损失时将边框视为一个整体, 有效地解决了 L2 损失法无法精确反应预测框与真实框的重合程度的问题; 其次, 成功解决了 IoU loss 方法在预测框与实际框无重合时梯度为零使网络无法学习的难题; 最后, 这种方法还将目标的形状、偏移方向以及相对位置等信息考虑了进来. 通常情况下, 训练图片中的负样本 (背景) 数量都会远远超过正样本 (目标) 数量. 为了有效解决正负样本不均衡的问题, 本文在计算置信度损失函数部分引入 focal loss 损失函数[28]. 这种方法在计算置信度损失时, 通过降低易分负样本的 loss 权重聚焦困难样本, 有效地解决了训练过程中困难样本因数量少而导致其 loss 被大量简单负样本 loss 淹没的问题. 置信度和类别损失使用二值交叉熵 (binary cross entropy) 的方法来计算, 如式 (3) 所示:

$$BCE = -\sum_{i=1}^{n} \hat{y}_i \log y_i + (1-\hat{y}_i)\log(1-y_i) \tag{3}$$

其中, $\hat{y}_i$为真实值, y_i为预测值, 当真实值和预测值越接近时 BCE 值越小. DCTH-YOLOv3 的损失函数一共分为 4 个部分:

1) CIoU loss 损失函数:

$$loss_{\mathrm{CIoU}} = c\left(1 - IoU + \frac{\rho^2(b,b^{gt})}{d^2} + \alpha v\right) \tag{4}$$

$$v = \frac{4}{\pi^2}\left(\tan^{-1}\frac{w^{gt}}{h^{gt}} - \tan^{-1}\frac{w}{h}\right)^2 \tag{5}$$

$$\alpha = \frac{\mu}{(1-IoU)+\mu} \tag{6}$$

其中, c 为真实目标置信度, CIoU 原理如图 7 所示, IoU 为预测框 (predict box) 与真实框 (ground truth box) 的

交并比, ρ为预测框和真实框中心点的距离, d 为包含预测框和真实框的最小 box 的对角线长, w^{gt} 、h^{gt} 、w、h 分别为真实框的宽高和预测框的宽高.

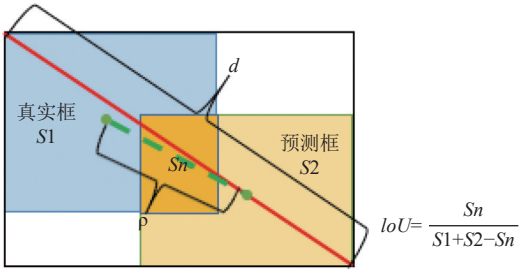


图 7 CIOU 原理示意图

2) 置信度 (confidence) 损失函数:

$$loss_c = c \times BCE_c + (1-c) \times BCE_c \times ignore \tag{7}$$

其中, BCE_c 为 c 的二值交叉熵, $ignore$ 为 IoU 低于一定阈值的但确实存在的物体.

3) $focal\_loss$ 损失函数:

$$focal\_loss = \alpha(|c-\hat{c}|)^{\gamma} \tag{8}$$

其中, $\hat{c}$为预测目标置信度, α、γ为超参数, 本文将α设置为 1, γ设置为 2.

4) 类别损失函数:

$$loss_{cls} = c \times BCE_{cls} \tag{9}$$

其中, cls 为类概率, BCE_{cls} 为 cls 的二值交叉熵.

因此, 模型总体损失函数为:

$$loss = loss_{\mathrm{CIoU}} + focal\_loss \times loss_c + loss_{cls} \tag{10}$$

2.4 模型训练

本文基于 TensorFlow 2.0 框架完成 DCTH-YOLOv3 网络的搭建, 训练网络的计算机硬件配置如下: CPU 型号为 ARM r7-2700X, GPU 型号为 NVIDIA GeForce GTX 1080Ti, 运行内存为 16 GB; 操作系统为 Ubuntu 20.04, Python 环境为 3.6.8.

首先在 COCO 数据集[29] 上进行了预训练处理, 并将预训练模型参数迁移至 DCTH-YOLOv3 网络. 使用冻结学习的方法更新模型参数, 相关超参数设置如表 1 所示, 整个训练轮次为 200 个 epochs: 前 100 个 epochs 为冻结阶段, 此阶段不更新模型的骨干特征提取网络 (Darknet53 网络) 参数, 初始学习率为 0.001, 批处理为 8; 后 100 个 epochs 为解冻阶段, 此阶段对 Darknet53 网络参数也进行更新, 初始学习率为 0.000 1, 批处理设为 4. 整个训练过程使用自适应动量估计梯度下降算法 (Adam)[30] 来更新权重, 学习动量 beta1 设为 0.9, beta2 设为 0.999; 权重衰减正则项系数设为 0.000 5; 使用衰减指数为 0.94 的指数衰减法更新学习率.

表 1 训练数据

参数类型	参数名称	参数值
基本参数	Beta1	0.9
	Beta2	0.99
	Weight Decay	0.000 5
冻结阶段参数	Batch	8
	初始学习率	0.001
	迭代轮次	100
解冻阶段参数	Batch	4
	初始学习率	0.000 1
	迭代轮次	100

2.5 实验

本文使用目标检测任务中常用的评价指标 AP (average precision) 对模型的检测精度进行评价, 主要包括 $AP_{0.9}$ 和 $AP_{0.5}$. 另外模型的检测速度和参数总量也是重要的评价准则.

首先对 CIoU loss 损失函数和 AF_FPN 结构分别进行了消融实验, 以 YOLOv3 网络为比较基准, 然后分别将 CIoU loss 损失函数和 AF_FPN 结构添加到 YOLOv3 的网络中, 保证训练和测试方法一致, 对比在同一份测试数据集上的 $AP_{0.9}$ 和检测速度以及模型参数总量, 得到的消融实验数据如表 2 所示.

表 2 消融实验数据

方法	CIoU loss	AF_FPN	$AP_{0.5}$	$AP_{0.9}$	检测速度 (fps)	参数总量 (MB)
YOLOv3	—	—	0.999 8	0.325 7	49	234.89
YOLOv3-CIoU	√	—	0.999 4	0.406 8	49	234.89
YOLOv3-AF_FPN	—	√	0.999 6	0.412	46	236.50
DCTH-YOLOv3	√	√	0.997 4	0.489 7	46	236.50

由表 2 可以看出, 原始 YOLOv3 网络在测试数据集上的 $AP_{0.9}$ 为 32.57%, 检测速度和参数总量分别为 49 fps 和 234.89 MB, 融合 CIoU loss 损失函数方法和 AF_FPN 结构的 $AP_{0.9}$ 为 48.97%, 检测速度和参数总量分别为 46 fps 和 236.5 MB; DCTH-YOLOv3 网络增加了 1.61 MB 参数量以及降低了 3 fps 检测速度, 但网络的检测精度提升 16.4%. 将 CIoU 方法和 AF-FPN 分别引入 YOLOv3 网络之后, $AP_{0.9}$ 均有不同程度的提

升, 其中 CIoU loss 损失函数只是修改了网络的训练方法, 因此并不会给网络带来额外的负担, 只影响网络的训练过程, 使网络学习到更好的模型参数, 模型的参数总量和检测速度与 YOLOv3 网络一样分别为 234.89 MB 和 49 fps. 由此消融实验可以说明, 在 YOLOv3 网络中引入 CIoU loss 损失函数和 AF-FPN 结构, 能够有效地提高网络的检测精度, 对预测框回归精度的改善更为明显. 另外, 在消融实验中 DCTH-YOLOv3 的 $AP_{0.5}$ 高达 99.74%, 尽管相对于 YOLOv3 下降了 0.24%, 但这精度足以满足本文的检测任务, 因此并未对此展开进一步的分析.

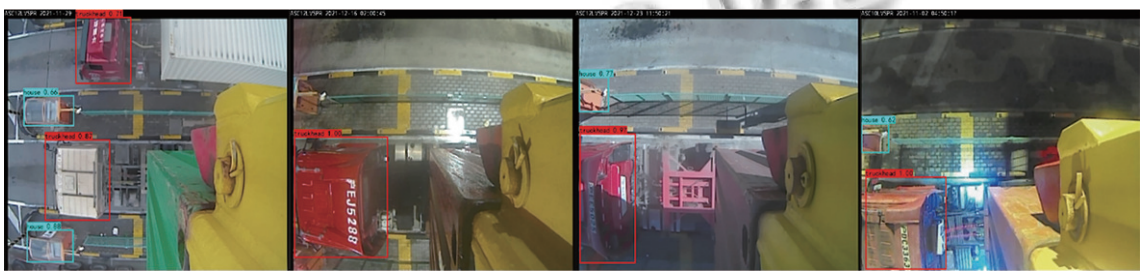
为进一步证明 DCTH-YOLOv3 网络的性能, 本文将网络与 YOLOv4 和 YOLOv5 进行实验对比, 对比结果如表 3 模型性能对比所示. 为确保公平, 在训练时, 所有模型采用一样的训练方法, 超参数设置与表 1 一致. 由表 3 可以看出, DCTH-YOLOv3 网络的参数总量和检测速度较 YOLOv5 分别增加了 58.38 MB 和降低了 10 fps, 但在测试数据集上 DCTH-YOLOv3 的 $AP_{0.9}$ 比 YOLOv5 高 2.71%, 比 YOLOv4 高 15.64%.

表 3 模型性能对比

方法	$AP_{0.5}$	$AP_{0.9}$	检测速度 (fps)	参数总量 (MB)
YOLOv3	0.9998	0.3257	49	234.89
YOLOv4	0.997	0.3333	41	244.16
YOLOv5	0.9999	0.4826	56	178.12
DCTH-YOLOv3	0.9974	0.4897	46	236.50

另外, 在对检测效果可视化比较时, 本文分别将 YOLOv3 网络和 DCTH-YOLOv3 网络对多个实际作业场景进行检测. 检测效果如图 8 所示, 其中检测场景依次包括晴日白天、雨天黑夜、白天遮挡、夜晚光影. YOLOv3 网络的检测效果如图 8(a) 所示, 虽然车头也都被正确检测到, 但是预测框的信息相对比较粗糙. DCTH-YOLOv3 网络的检测效果如图 8(b) 所示, 在以上各种场景, 对集卡车头都能做出更加精准的检测, 尤其是预测框的回归. 比如在图 8 的第 1 个场景中, 明显地可以看出图 8(b) 中的白色集卡车头预测框比图 8(a) 中的更加接近实际车头的尺寸. 另外在第 3 个和第 4 个场景中, 只有 DCTH-YOLOv3 网络正确检测出了集卡车头附近的小岗亭.

(a) YOLOv3 检测效果



(b) DCTH-YOLOv3 检测效果

图 8 检测效果对比图

从上述实验和可视化比较结果可见, 针对集卡车头本文提出的 DCTH-YOLOv3 算法模型具备检测精度高、检测速度快、鲁棒性强等优点. 该算法能够满足集卡车头防砸检测系统对集卡车头高精定位、实时检测的要求, 为最终的防砸决策提供可靠的数据.

3 结论

针对集卡车头防砸检测研究不足及存在的问题, 本文提出了一种改进的 YOLOv3 算法应用于集卡车头防砸检测. 首先改进 YOLOv3 的 FPN 结构, 在特征融合时引入注意力机制, 聚焦有效特征, 抑制干扰噪声, 提高了模型检测精度, 其次改进预测框的损失函数, 提

高预测框的检测精度, 使得模型检测精度得到进一步提升, 最终 $AP_{0.9}$ 提高了 16.4%. 但是目前该网络在一些罕见的环境下, 对集卡车头的检测效果不佳, 后续将尝试进一步扩充训练数据, 以及优化网络结构, 提高网络的鲁棒性和检测精度, 以增加模型的实际应用价值.

参考文献

1 陈培. 2D 激光传感器在自动化集装箱码头设备上的应用. 港口科技, 2021, (8): 44–48. [doi: 10.3969/j.issn.1673-6826.2021.08.009]

2 胡荣东, 文驰, 彭清, 等. 基于三维激光的集卡防吊起检测方法、装置和计算机设备: 中国, CN113376651A. 2021-09-10.

3 张俊阳, 吴翔, 单磊, 等. 集卡车头防砸保护系统及方法、计算机存储介质、龙门吊: 中国, CN112580517A. 2021-03-30.

4 He KM, Gkioxari G, Dollár P. Mask R-CNN. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 386–397. [doi: 10.1109/TPAMI.2018.2844175]

5 Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv: 1804.02767, 2018.

6 Lin TY, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 2117–2125.

7 Wang XE, Song J. ICIoU: Improved loss based on complete intersection over union for bounding box regression. IEEE Access, 2021, 9: 105686–105695. [doi: 10.1109/ACCESS.2021.3100414]

8 Dalal N, Triggs B. Histograms of oriented gradients for human detection. Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego: IEEE, 2005. 886–893.

9 Lowe DG. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2004, 60(2): 91–110. [doi: 10.1023/B:VISI.0000029664.99615.94]

10 Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge: Cambridge University Press, 2000.

11 Sun XW, Zhou HB. Experiments with two new boosting algorithms. Intelligent Information Management, 2010, 2(6): 386–390. [doi: 10.4236/iim.2010.26047]

12 Felzenszwalb PF, Girshick RB, McAllester D, *et al*. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627–1645. [doi: 10.1109/TPAMI.2009.167]

13 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe: ACM, 2012. 1097–1105.

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

15 Li JN, Liang XD, Shen SM, *et al*. Scale-aware fast R-CNN for pedestrian detection. IEEE Transactions on Multimedia, 2018, 20(4): 985–996.

16 Ren SQ, He KM, Girshick R. 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]

17 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.

18 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 779–788.

19 王晓宇, 张长伦, 何强, 等. 基于 YOLOv3 的建筑工地目标检测研究. 计算机科学与应用, 2021, 11(11): 2788–2794.

20 朱晨阳, 冯虎田, 欧屹. 基于 YOLO3 的人脸自动跟踪摄像机器人系统研究. 电视技术, 2018, 42(9): 57–62, 84. [doi: 10.16280/j.videoe.2018.09.012]

21 刘洋, 姜涛, 段学鹏. 基于 YOLOv3 的复杂天气条件下人车识别方法的研究. 长春理工大学学报 (自然科学版), 2020, 43(6): 57–65.

22 Bochkovskiy A, Wang CY, Liao HYM. YOLOv4: Optimal speed and accuracy of object detection. arXiv:2004.10934, 2020.

23 Dai YM, Gieseke F, Oehmcke S, *et al*. Attentional feature fusion. Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2021. 3560–3569.

24 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141.

25 Zhang H, Wu CR, Zhang ZY, *et al*. ResNeSt: Split-attention networks. Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New Orleans: IEEE, 2022. 2735–2745.

26 Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 7263–7271.

27 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.

28 Lin TY, Goyal P, Girshick R, *et al*. Focal loss for dense object detection. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 2980–2988.

29 Lin TY, Maire M, Belongie S, *et al*. Microsoft COCO: Common objects in context. Proceedings of the 13th European Conference on Computer Vision. Zurich: Springer, 2014. 740–755.

30 Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv:1412.6980, 2014.

(校对责编: 牛欣悦)