工程科学学报
Chinese Journal of Engineering

DS-YOLOv5：一种实时的安全帽佩戴检测与识别模型

白培瑞 王瑞 刘庆一 韩超 杜红萱 轩辕梦玉 傅颖霞

DS-YOLOv5: A real-time detection and recognition model for helmet wearing

BAI Peirui, WANG Rui, LIU Qingyi, HAN Chao, DU Hongxuan, XUANYUAN Mengyu, FU Yingxia

引用本文:
白培瑞, 王瑞, 刘庆一, 韩超, 杜红萱, 轩辕梦玉, 傅颖霞. DS-YOLOv5：一种实时的安全帽佩戴检测与识别模型[J]. 工程科学学报, 2023, 45(12): 2108-2117. doi: 10.13374/j.issn2095-9389.2022.11.11.006
BAI Peirui, WANG Rui, LIU Qingyi, HAN Chao, DU Hongxuan, XUANYUAN Mengyu, FU Yingxia. DS-YOLOv5: A real-time detection and recognition model for helmet wearing[J]. *Chinese Journal of Engineering*, 2023, 45(12): 2108-2117. doi: 10.13374/j.issn2095-9389.2022.11.11.006

在线阅读 View online: https://doi.org/10.13374/j.issn2095-9389.2022.11.11.006

工程科学学报, 第 45 卷, 第 12 期: 2108−2117, 2023 年 12 月
Chinese Journal of Engineering, Vol. 45, No. 12: 2108−2117, December 2023
https://doi.org/10.13374/j.issn2095-9389.2022.11.11.006; http://cje.ustb.edu.cn

DS-YOLOv5: 一种实时的安全帽佩戴检测与识别模型

白培瑞[1)], 王　瑞[1)], 刘庆一[1)], 韩　超[1)], 杜红萱[1)], 轩辕梦玉[1)], 傅颖霞[2)✉]

1) 山东科技大学电子信息工程学院, 青岛 266590　2) 山东科技大学能源与矿业工程学院 , 青岛 266590
✉通信作者, E-mail: fyx22@163.com

摘　要　基于视频分析技术对生产现场人员安全帽佩戴情况进行自动检测与识别是保障安全生产的重要手段. 但是, 复杂的现场环境和多变的外界因素为安全帽检测与识别的精确性提出挑战. 本文基于 YOLOv5 模型的框架, 提出一种 DS-YOLOv5 安全帽检测与识别模型. 首先, 利用改进的 Deep SORT 多目标跟踪的优势, 提高视频检测中多目标检测和有遮挡的容错率, 减少漏检情况; 其次在主干网络中融合简化的 Transformer 模块, 加强对图像的全局信息的捕获进而加强对小目标的特征学习; 最后在网络的 Neck 部分应用双向特征金字塔网络 (BiFPN) 融合多尺度特征, 以便适应由摄影距离造成的目标尺度变化. 所提模型在 GDUT-HWD 和 MOT 多目标跟踪数据集上进行了验证实验, 结果表明 DS-YOLOv5 模型可以更好地适应遮挡和目标尺度变化, 全类平均精度(mAP)可以达到 95.5%, 优于其他常见的安全帽检测与识别方法.

关键词　目标检测; 安全帽佩戴; YOLOv5; Deep SORT; 自注意力机制
分类号　TP391.4;TU714

DS-YOLOv5: A real-time detection and recognition model for helmet wearing

BAI Peirui[1)], *WANG Rui*[1)], *LIU Qingyi*[1)], *HAN Chao*[1)], *DU Hongxuan*[1)], *XUANYUAN Mengyu*[1)], *FU Yingxia*[2)✉]

1) College of Electronic and Information Engineering, Shandong University of Science and Technology, Qingdao 266590, China
2) College of Energy and Mining Engineering, Shandong University of Science and Technology, Qingdao 266590, China
✉Corresponding author, E-mail: fyx22@163.com

ABSTRACT　Automatic detection and recognition of safety helmet wearing based on video analysis is important to ensure production safety. It is inefficient to supervise whether workers wear safety helmets by manual means. With the advancement of deep learning, using computer vision to assist in the detection of safety helmet-wearing holds significant research and application value. However, complex environments and variable factors pose challenges in achieving accurate detection and recognition of safety helmet usage. Helmet-wearing detection methods are generally classified as traditional machine learning and deep learning methods. Traditional machine learning methods employ manually selected features or statistical features, resulting in poor model stability. Deep learning–based methods are further divided into “two-stage” and “one-stage” methods. The two-stage method has high detection accuracy but cannot achieve real-time detection, while the one-stage counterpart is the reverse. Achieving accuracy as well as real-time detection is an important challenge in the development of video-based helmet detection systems. Accurate and quick detection of helmets is essential for effective real-time monitoring of production sites. To address these challenges, this paper proposes DS-YOLOv5—a real-time helmet detection and recognition model based on the YOLOv5 models. The proposed model solves three main problems: First, insufficient global information extraction problem of convolutional neural network (CNN) models. Second, the lacking robustness of the deep SORT for multiple targets and occlusion problems in video scenes. Third, the inadequate feature extraction of multiscale targets. The DS-YOLOv5 model takes advantage of the improved Deep SORT multitarget tracking algorithm to reduce the rate of missed detections in multitarget detection and occlusion and increase the error tolerance in video detection. Further, a simplified transformer

收稿日期: 2022−11−11
基金项目: 国家自然科学基金资助项目(61471225)

module is integrated into the backbone network to enhance the capture of global information from images and thus enhance feature learning for small targets. Finally, the bidirectional feature pyramid network is used to fuse multiscale features, which can better adapt to target scale changes caused by the photographic distance. The DS-YOLOv5 model was validated using the GDUT-HWD dataset by ablation and comparison experiments. In these experiments, the tracking capability of the improved Deep SORT is compared with the YOLOv5 model using the public pedestrian dataset MOT. The results of the comparison of the five one-stage methods and four helmet detection and recognition models revealed that the proposed model has better capability for dealing with occlusion and target scale. Further, the model achieved mean average orecision (mAP) of 95.5%, which is superior to that of the other helmet detection and recognition models.

KEY WORDS object detection; helmet wearing; YOLOv5; deep SORT; transformer

在不同行业的施工和生产现场, 使用基于视频分析技术对安全帽佩戴情况进行自动检测与识别, 是保障人员生命安全的重要措施. 但生产现场视频采集环境往往更复杂, 目标遮挡、光照不均匀和目标尺度差异大等原因, 都会对基于视频的安全帽佩戴情况自动检测与识别提出挑战. 安全帽佩戴检测与识别算法分为基于传统机器学习和基于深度学习的方法. 传统机器学习方法通常采用背景减法、人体检测和安全帽检测识别等步骤, 利用的是手工选定特征或统计特征. 刘晓慧和叶西宁[1] 提出利用肤色辅助确定安全帽位置, 提取 Hu 矩特征向量, 再用支持向量机对安全帽进行识别分类. Li 等[2] 针对位置固定的监控场景下的运动目标, 采用 Vibe 对运动背景进行分割处理, 使用实时的人体分类框架 C4 对人体定位, 最终通过颜色特征判别实现安全帽检测.

基于深度学习的方法进一步分为“两阶段”方法和“单阶段”方法[3]. “两阶段”方法即由算法提取特征后进行候选区域生成, 然后使用分类器进行分类回归. Yogameena 等[4] 利用 Faster R-CNN 检测带标记的摩托车目标, 再利用卷积网络模型和空间转换器识别头盔. Ferdous 等[5] 设计以 ResNet50 为主干融合特征金字塔网络 (FPN), 采用分类和回归模型对安全帽进行分类和定位. “两阶段”方法的优势是可以有效提升检测精度, 但是难以满足实时性检测的要求. “单阶段”方法使用端到端策略对图像进行目标位置的检测和分类. SSD(Single shot multibox detector)[6] 模型和 YOLO(You only look once)[7] 模型是“单阶段”算法的典型代表. SSD 模型结合回归和 anchor 思想实现多尺度预测, 但密集采样会造成模型难收敛, 对小目标检测能力较差. 吴等[8] 使用结合反向渐进注意机制的 SSD 网络对安全帽进行检测, 使用 H-SVM 对安全帽颜色分类, SSD-RPA512 模型达到了 83.89% 的准确率. 徐先峰等[9] 使用 MobileNet 为主干构建 MobileNet-SSD 模型, 将安全帽检测速度提高了 10.2 倍. YOLO 模型采用边检测边分类的策略, 具有计算效率高的优势, 其升级版本得到相关领域的广泛关注. YOLOv3[10] 提出 Darknet53 网络, 使用 FPN 架构和多尺度融合策略, 提高了对小目标检测的精度. YOLOv4[11] 的主干采用具有不同层间交叉的 CSPDarknet53, 采用 Mosaic 数据增强方法和自我对抗训练策略以提高网络的检测与识别性能. YOLOv5[12] 通过增加模型检测规模和数据优化处理, 采用 DIoU-NMS 边界盒抑制措施提高小目标的检测精度. YOLOv5 模型轻量且易移植, 让很多研究在其基础上都得到了良好的检测性能[13−14]. 针对复杂环境下检测也成为新的研究领域, 例如, 黄林泉等[15] 提出 YOLOv3 结合 Deep SORT 多目标跟踪技术实现跳帧检测, 结合可微图像处理模块提高目标检测在恶劣天气的适应性. 受到自然语言处理中 Transformer[16] 应用的启发, 使用 CNN 的架构与 Transformer 结合能对图像中全局信息和局部信息统一建模, 得到灵活高效的网络模型. Zhu 等[17] 针对无人机采集图片的尺度变化和运动模糊问题, 提出了 TPH-YOLOv5 模型, 结合 Transformer 和注意力模块 CBAM 提高预测输出效果, 提高小目标的检测性能. 前人研究证明结合注意力机制和特征融合机制可提高检测器的学习能力进而提升检测性能. 上述安全帽检测包含以下几种难题: (1)CNN 模型通常关注局部信息忽略全局信息, 而安全帽佩戴检测往往要伴随人体或人脸特征信息进行学习; (2)主干网络的特征提取往往针对单目标情况进行优化, 在视频场景下的检测伴随着多目标和遮挡问题, 算法缺乏鲁棒性; (3)对多尺度目标特征提取不充分.

本文针对以上难题, 提出一种基于 Deep self-attention YOLOv5(DS-YOLOv5) 的安全帽检测模型, 主要改进如下:

(1)使用 YOLOv5 模型为基础, 融合改进的 Deep SORT 多目标跟踪算法, 改进的 Deep SORT 使用

DIoU 提高检测框与预测框匹配准确度, 加强视频检测中前后帧的数据关联, 能改善遮挡和多目标造成的漏检错检问题;

(2) YOLOv5 主干网络融合 Transformer 模块, 将图片转成序列向量提取局部特征之间的关系, 加强对图像全局信息的捕获和注重上下文信息的联系;

(3) YOLOv5 颈部网络部分使用加权双向特征金字塔网络 (BiFPN) 融合多尺度特征, 根据特征的重要程度赋以权重, 提高小目标的检测准确度. 本文构建一种能够应对复杂环境, 提高安全帽佩戴情况检测精度和实时性的深度学习模型 DS-YOLOv5. 实验结果表明, 该模型的多个定量指标优于目前主流的安全帽佩戴检测模型, 其 mAP 可以达到 95.5%.

1　相关背景

1.1　YOLOv5 模型结构与工作原理

YOLO 模型的核心思想是把物体检测问题转化为回归问题, 用一个卷积神经网络从输入图像直接预测边界框位置和目标的类别概率. YOLOv5 结构主要分为输入端、主干(Backbone)、颈部网络(Neck)和预测头(Prediction)四个部分. 输入端采用 mosaic 数据增强, 主干采用 Focus 结构和 CSPDarknet53(Cross stage partial darknet53)结构相结合的方案, CSPDarknet53 是特征提取的核心网络, 借助残差块实现了对特征图的快速降维, 在不损失检测精度的前提下, 提升对特征提取的能力和检测速度. 颈部网络采用空间金字塔池 (Spatial pyramid pooling, SPP) [18] 和路径聚合网络 (Path aggregation network, PANet) [19] 的结构, 加强来自不同特征层的特征聚合, 提高网络检测不同尺度目标的能力. 输出端采用 CIoU_Loss 和 DIoU_NMS 操作, 输出对象坐标和分类结果. YOLOv5s 设置了两种 CSP 模块, 在主干应用 CSP1_X 结构, 在 Neck 部分应用 CSP2_X 结构, CSP1_X 比 CSP2_X 多设有残差结构, 可以增加层与层之间反向传播的梯度值, 避免因为网络过深带来的梯度消失, 加强网络提取特征的能力.

当输入端接收图像后, YOLOv5 模型首先将每幅图像划分为 7×7 的网格, 每个网格作为一个预测盒, 用于捕获落入网格单元中心的目标. 每个网格单元输出 M 个预测边界框, 每个边界框的预测输出位置信息和置信度 C(Confidence), C 主要反映预测盒与 Ground Truth 之间的交并比 (Intersection over union, IOU) 得分. 置信度得分公式为:

$$C = P_r(\text{object}) \times \text{IOU}_{\text{pred}}^{\text{truth}} \tag{1}$$

式中, P_r(object)表示目标出现在单元格的条件类概率, 规定单元格中有目标其值取 1, 否则取 0. $\text{IOU}_{\text{pred}}^{\text{truth}}$表示预测边界框与 Ground Truth 的交并比, C 反映预测类别的概率和边界框定位的准确性. 根据网络深度和宽度参数的不同, YOLOv5 提供了 YOLOv5s、YOLOv5m、YOLOv5l 和 YOLOv5x 四种模型配置文件. YOLOv5s 是网络深度和特征图宽度最小的网络, 考虑到面向工业应用, 本文采用 YOLOv5s 为基础模型进行研究.

1.2　Deep SORT 多目标跟踪算法

深度简单在线实时跟踪 (Deep simple online and realtime tracking, Deep SORT) 的核心思想是利用传统的卡尔曼滤波和匈牙利算法分别负责跟踪目标的状态估计和轨迹分配问题[20]. 通过引入视频序列中检测结果的数据关联, 整合目标的运动信息和外观信息, 对丢失目标进行重识别. 改进的 Deep SORT 算法流程如图 1 所示. Deep SORT 中使用卡尔曼滤波器与匀速运动和线性观测模型来观察目

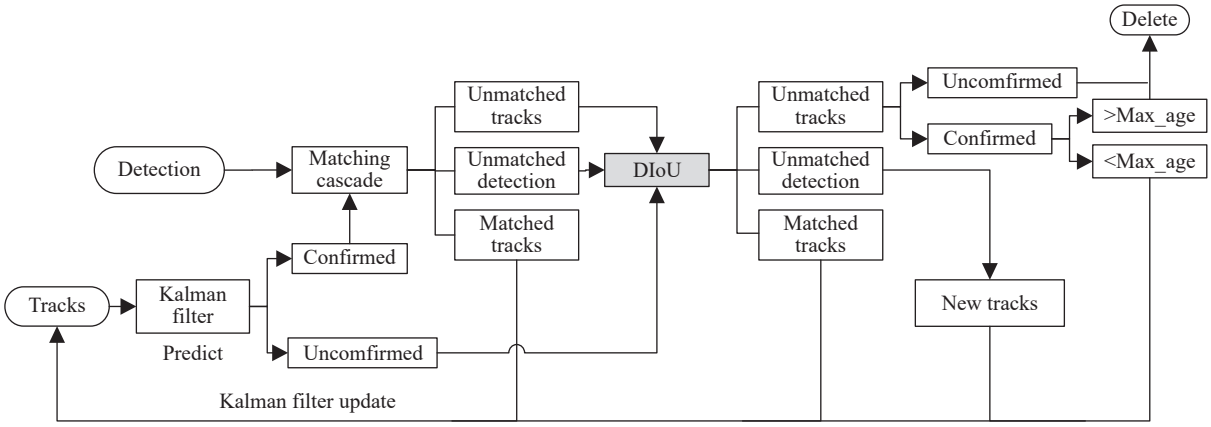

图 1　改进的 Deep SORT 算法流程图

Fig.1　Flow chart of the improved Deep SORT algorithm

标的状态. 对于输入的轨迹, 卡尔曼滤波器通过预测状态与上一时刻的对比和检测框信息的校正, 估计每个目标位置的均值和协方差矩阵, 再通过卡尔曼滤波框架最优地求解速度分量对目标状态进行更新[21]. 匈牙利算法将前一帧的跟踪框与当前帧的检测框关联, 通过特征信息和马氏距离来计算代价矩阵. 首先, 基于已存在目标运动状态的卡尔曼预测结果与检测结果之间的马氏距离进行运行信息的关联, 马氏距离如式 (2):

$$l_1(i,j)=(\boldsymbol{d}_j-\boldsymbol{y}_i)^{\mathrm{T}}\boldsymbol{S}_i^{-1}(\boldsymbol{d}_j-\boldsymbol{y}_i) \tag{2}$$

式中, $(\boldsymbol{y}_i,\boldsymbol{S}_i)$表示第$i$个跟踪到测量空间的投影, 其中$\boldsymbol{y}_i$为当前时刻的预测框, $\boldsymbol{S}_i$为协方差矩阵, $\boldsymbol{d}_j$表示第j个检测框.

当目标遮挡或者镜头视角抖动时, 借助外观信息对检测框d_j计算出对应的外观特征描述符$\boldsymbol{r}_j$, 即重识别网络提取的 128 维单位特征向量. 对于每个轨迹 k, 都保留在最后 100 个外观描述符作为集合$\{\boldsymbol{R}_i\}$, 通过计算第i个跟踪框和第j个检测框最小余弦距离来减少跟踪误差, 余弦距离如式 (3):

$$l_2(i,j)=\min\left\{1-\boldsymbol{r}_j^{\mathrm{T}}\boldsymbol{r}_k^{(i)}\mid \boldsymbol{r}_k^{(i)}\in\boldsymbol{R}_i\right\} \tag{3}$$

级联匹配将卡尔曼滤波器判断为确认的轨迹框与检测框在卷积网络进行特征提取, 根据轨迹框与检测框的l_1和l_2, 进行匹配确认. 再利用距离交并比 (Distance-IoU, DIoU)[22] 对级联匹配中的未匹配轨迹和未匹配检测以及卡尔曼预测未确定的轨迹, 进行再次匹配, 对确认状态的轨迹框设置最大匹配寿命 Max_age, 并对超过 Max_age 轨迹框删除处理, 避免因短暂性遮挡而造成的漏检.

Deep SORT 算法中 IoU 匹配方式回归速度快, 但是无法衡量预测框与轨迹框的重叠方式. 本研究使用 DIoU 关联匹配, 它在 IoU 的基础上引入一个惩罚项, 对重叠情况更加敏感, 训练过程损失收敛更快, 减少遮挡情况下 ID 切换的问题, 有更好的跟踪效果. 检测框和目标框之间的归一化距离损失函数公式为:

$$L_{\mathrm{DIOU}}=1-\mathrm{IOU}+\frac{\rho^2(b,b_{\mathrm{GT}})}{c^2} \tag{4}$$

式中, b和b_{GT}分别代表检测框和预测框的中心点, ρ代表两个中心点之间的欧氏距离. c 代表同时覆盖检测框和预测框的最小矩形的对角线距离.

1.3 Transformer

Transformer 是谷歌团队提出基于注意力结构来处理序列模型相关问题的模型, Transformer 的基本结构由编码组件、解码组件和连接(position-wise FFN)组成. 编码器中有自注意力层和前馈层, 解码器中除了有自注意力层和前馈层之外, 还有一个解码器注意力层用于关注层之间的交叉信息, 它推翻了以往自然语言处理通常需要 CNN 结构的思想. Transformer 最初应用于自然语言处理领域, 近来越来越多地被应用到计算机视觉领域. 在目标检测领域, Carion 等 [23] 提出了一种端到端的目标检测模型(Detection transformer, DETR), 将 CNN 与 transformer 架构相结合直接预测最终的检测结果, 唯一的不足是需要大量的数据进行训练才能得到良好的精度. VIT(Vision transformer) 是一种基于 Transformer 进行图像块序列预测的分类模型, 在多个图像识别基准数据集上获得优越的性能[24]. 研究表明 Transformer 在计算机视觉领域有优秀的学习能力.

受到 VIT 的启发, 我们将 YOLOv5 主干中部分原始卷积模块替换为 VIT 中的 Transformer Block, Transformer 将图像切块输入, 借助对图像间的空间信息, 提高主干的特征提取能力, 最后使用 YOLOv5 检测头进行分类. VIT 模型包含 Embedding 层、Transformer Encoder 和 MLP(Multilayer perceptron) Head. Linear 层主要负责将二维图像转化为序列向量作为 Transformer 的输入, Transformer Encoder 功能是捕获序列向量之间的相互关系, MLP Head 用于最终的分类.

(1)Embedding 层.

对输入的安全帽图像$\boldsymbol{x}\in\boldsymbol{R}^{H\times W\times C}$分为 N 个不重叠的二维图像块$\boldsymbol{x}_P\in\boldsymbol{R}^{N\times(S^2\times C)}$, 其中 $\boldsymbol{R}$ 是包含所有图像块的集合, 并使用卷积核大小为 D 的网络对图像块编码生成序列向量$\boldsymbol{x}_r=\{\boldsymbol{r}_{\mathrm{class}},\boldsymbol{r}_1,\boldsymbol{r}_2,\cdots,\boldsymbol{r}_N\}$用于 Transformer 的输入, 图像块数量与原图像之间的关系为:

$$N=HW/S^2 \tag{5}$$

其中, H, W, C 为原图像的高、宽和通道数, S 为图像块的尺寸大小, $\boldsymbol{r}_{\mathrm{class}}$是为防止在多层感知机(MLP)学习全局特征不够充分而加入的分类向量. 图像块的序列向量代表图像块的内容信息, 而自注意力机制处理序列还需要向量间的位置关系, 因此在原文中还引入了位置编码$\boldsymbol{E}_{\mathrm{poc}}$用于捕获图像块之间的位置关系, $\boldsymbol{E}$ 是实现线性映射的矩阵. 因此在 Embedding 层处理后, 特征向量 $\boldsymbol{F}$ 表示为:

$$\boldsymbol{F}=\boldsymbol{x}_r+\boldsymbol{E}_{\mathrm{poc}} \tag{6}$$

(2)Transformer Block 和 MLP Block.

Transformer Block 主要由 Transformer 编码器

组成，作用是对序列向量提取局部特征之间的关系，首先对 Embedding 层的输出进行归一化处理，经多头注意力层（Mutli-head self-attention，MSA）提取得到当前节点的全局特征向量，经过 Dropout 和残差输出相加，进一步 Norm 处理，最终经过多层感知机模块（MLP block）获得分类特征. MLP block 主要由全连接层、GELU 激活函数和 Dropout 层组成，序列向量经过 MSA 融合特征信息，最终选择一项特征作为全局特征送入 MLP 中进行分类.

2　DS-YOLOv5 模型网络结构与工作流程

2.1　YOLOv5 模型结构

2.1.1　Backbone

模块如图 2 所示 C3TR 模块结构图，将原 C3 残差模块中的 Resunit 替换为 Transformer 模块，C3TR 模块由卷积层和 Transformer block 经 Concat 操作构成，由于模块参数复杂度过高，为了降低训练成本仅在 YOLOv5 主干网络的最后一个阶段进行加强特征提取，图 2 中模块中 k、s、n 分别代表卷积神经网络卷积核、步长以及包含 Bottleneck 的数量，T 代表 C3 中设有残差模块，F 代表 C3 中没有残差模块. 同时在研究中发现 Norm 能削弱模型的特征表征能力，如图 3 所示，将 Transformer 简化成图 3 的 Transformer block，首先去除 MSA 和 MLP 模块后的归一化处理，仅使用 Dropout 加速网络收敛；其次 MLP block 去除激活函数，使用全连接层和 Dropout 进行分类处理，可加强特征全局表达的能力，同时降低一定计算成本. 简化后 Transformer block 减少了计算量，同时可以加强捕获全局特征的能力，其多头注意力机制还可以获得上下文的语义信息，对于低分辨率图像有良好的处理能力，可以应对小目标的特征提取.

2.1.2　Neck

传统的 PANet 可融合不同层级间的特征且各层级是逐级通道顺序连接，这种结构忽略了不相邻层级之间的特征信息关系，遗失细节信息. 本文使用 BiFPN[25] 对提取的特征进行跨层级融合，通过对特征赋予不同的权重进行有重点的融合，可增强对有益特征的针对性，同时抑制冗余特征的干扰. 图 4 为 PANet 和 BiFPN 实现不同层级特征传递与融合的示意图，自下到上为 3 ~ 7 层特征图（P3 ~ P7），(a) 是 PANet 结构，(b) 是 BiFPN 结构.

BiFPN 实现了深浅层特征双向融合，跨尺度残差连接增强不同网络层之间特征信息的传递，有

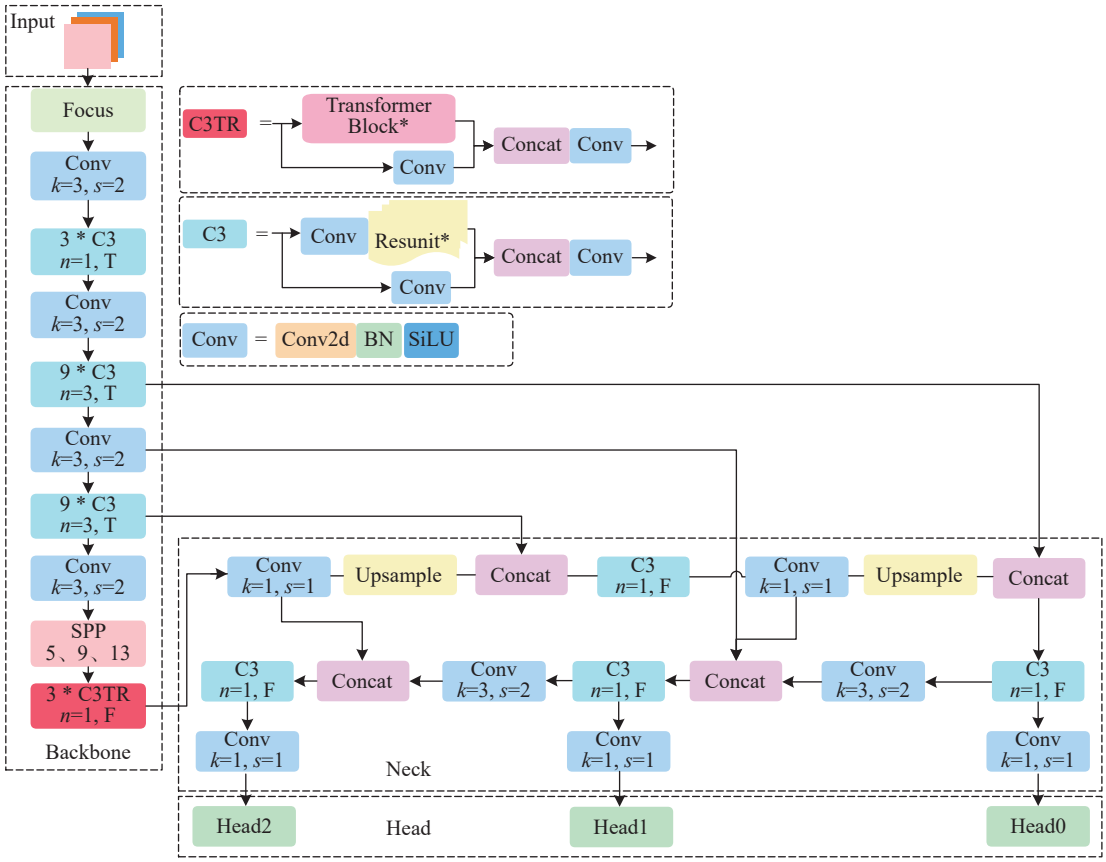

图 2　改进的 YOLOv5 的网络结构

Fig.2　Improved network architecture of YOLOv5

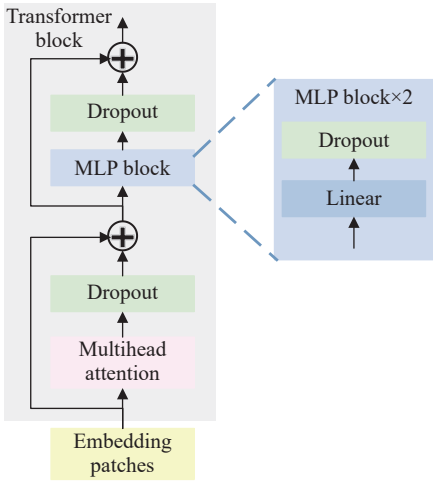

图 3 Transformer 模块

Fig.3 Transformer block

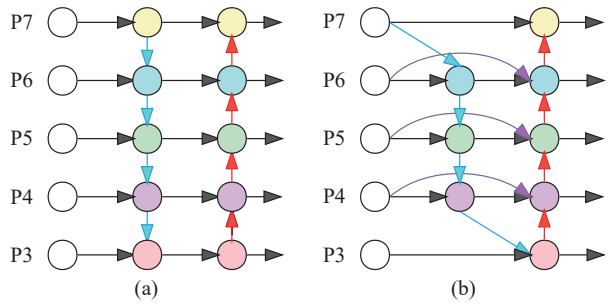

图 4 PANet 与 BiFPN 结构示意图. (a) PANet; (b) BiFPN

Fig.4 Schematic diagram of the Path Aggregation Network and bidirectional feature pyramid network structures: (a) PANet; (b) BiFPN

利于小尺寸安全帽的特征学习, 而且与 PANet 相比, 它移出了没有特征融合的节点, 进一步减少模型的计算量. 此外, BiFPN 针对特征融合增加了一个可学习的权重, 可以有效聚合不同尺度的特征并进行特征输出. BiFPN 以更轻量的结构同时提升模型对有益特征的敏感性, 带来模型性能的提升.

2.2 DS-YOLOv5 工作流程

DS-YOLOv5 模型的工作流程: 首先利用带标签的安全帽数据集, 对 YOLOv5 网络和改进的 Deep SORT 特征提取网络分别进行训练得到安全帽检测模型和安全帽特征模型, 为视频序列的检测识别做好模型权重. 最后, 将待检测视频输入 YOLOv5 进行检测, 得到安全帽的检测框信息和分类信息, 再将检测框送入改进的 Deep SORT 进行特征提取进行建模标准化, 对目标轨迹进行卡尔曼预测和更新, 经过级联匹配和 IOU 匹配得到跟踪结果, 输出安全帽位置信息和类别信息.

3 实验与分析

为了验证 DS-YOLOv5 模型的性能, 本文采用 GDUT-HWD 数据集进行测试. GDUT-HWD 数据集考虑包括场景、照明、个人姿态和遮挡等复杂环境, 包含 3174 张图像, 18893 个实例, 小目标实例数占比达到了 47.4%, 环境更复杂、分类更精细. 实验的硬件配置为 Intel(R) Xeon(R) Silver 4210R CPU @ 2.40 GHz 和 2.39 GHz NVIDIA GTX1070GPU, 模型搭建在 Pytorch 深度学习框架. 所有 YOLO 模型的训练最大周期为 300, Batch-size 为 8, 学习率为 0.0001, 在训练过程中采用早停策略避免过拟合. 评价指标采用多个定量指标, 分别为精确率 P、召回率 R、F_1 分值、平均精确度 AP 和全类平均精度 mAP 和每次迭代花费的时间 Spend Time. 各指标的定义如下:

$$P = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{7}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{8}$$

式中, 真阳性 (True positive, TP) 代表实际为正例, 预测为正; 真阴性 (True negative, TN) 实际为负例, 预测为负; 假阳性 (False positive, FP) 为实际为负例, 预测为正; 假阴性 (False negative, FN) 实际为正例, 预测为负. 一般来说精确度和召回率是互相矛盾的两个值, 定义 F_1 说明了精确度和召回率的实际平均情况, 公式为:

$$F_1 = \frac{2PR}{P+R} \tag{9}$$

对于数据集给定的类别, 先得到其精确率/召回率(P/R)曲线, 平均精确度 AP 为使用积分法计算 P/R 曲线与坐标轴所围面积, mAP 是将 AP 进行加权平均而得到, 评价指标 AP 和 mAP 计算方法如公式 (10) 和 (11) 所示:

$$\mathrm{AP} = \int_0^1 P(R)\mathrm{d}R \tag{10}$$

$$\mathrm{mAP} = \frac{1}{\mathrm{class}} \sum_{i=1}^{\mathrm{class}} \mathrm{AP}_i \tag{11}$$

式中, $P(R)$ 是关于 P 和 R 函数关系, class 表示检测数据集的类别数.

3.1 消融实验

消融实验基于 YOLOv5 模型为基础模型, 对 Deep SORT、BiFPN 结构、Transformer 模块和 Deep SORT 进行分析, 了解各模块对模型性能提升起到的作用以及结构改进的有效性.

表 1 列出了在 GDUT-HWD 数据集上的消融实验结果, 可以看到, 加入 Transformer、BiFPN 和 Deep SORT 分别使 mAP 提升 1.2%、1.4% 和 0.8%, 最终

表 1　基于 GDUT-HWD 数据集的消融实验结果

Table 1　Results of the ablation experiments based on the GDUT-HWD dataset

Model	mAP/%	*P*/%	*R*/%	F_1/%	Param/10^6	Spend time/min
YOLOv5	92.5	92.1	90.6	91.4	7.1	6.3
YOLOv5 + BiFPN	93.9	92.5	91.6	92.0	7.5	6.1
YOLOv5 + Transformer	93.7	92.6	97	94.7	8.1	9.1
YOLOv5 + Deep SORT	93.3	89.3	91.9	90.6	7.1	6.3
YOLOv5 + BiFPN + Transformer	94.4	92.1	99	95.3	8.2	8.8
YOLOv5 + BiFPN + Transformer + Deep SORT	95.5	89.6	98	93.6	8.2	8.8

模型 mAP 提升了 3%. 加入 Transformer 的模型召回率有明显的提高, 进一步证明本模块对捕获全局特征有良好能力, 减少小目标漏检情况. Deep SORT 模型将使用视频图像验证其识别跟踪可靠性. F_1 反映出模型精度和召回率的实际平均情况, 本模型加入 BiFPN 和 Transformer 后 F_1 下降 1.7%, 仍比 YOLOv5 原模型提高了 2.2%, mAP 能达到 95.5%, 此时模型检测精度最高. 根据消融实验结果, 增加模块后网络复杂度增加, Spend Time 也随之增加, 而使用 BiFPN 模块后可以比原模型节约 3% 的训练时间. 我们将表 1 中的最后一组模型称为 DS-YOLOv5.

如表 2 对改进前后 Deep SORT 跟踪性能对比. 使用行人数据集 Market-1501 对改进的 Deep SORT 重识别网络进行训练, 训练集包含了 751 个类别行人[26]. 表 2 中评价指标 MOTA(Multi-object tracking accuracy) 为衡量跟踪器的准确率, MOTP (Multi-object tracking precision) 为检测器的准确率, IDs (Identity switch) 为 Ground Truth 分配 ID 发生切换的次数 (表中向上箭头表示值越大性能越好, 向下箭头表示值越小性能越好). 实验结果可以看到 MOTA 提高了 2.7, IDs 在原有基础上降低的 16%.

表 2　Deep SORT 改进前后的实验结果对比

Table 2　Comparison results for improved Deep SORT

Model	MOTA↑	MOTP↑	IDs↓
Deep SORT	48.5	77.1	48
Improved Deep SORT	51.2	77.6	40

为了验证模型的可靠性, 使用 MOT16 训练集中视频 MOT16-09 对改进前后 Deep SORT 进行跟踪性能测试[27], 对测试的 MOT16-09 视频每 10 帧抽取一次, Deep SORT 改进前后在 MOT16-09 测试对比如图 5 所示, 可视化中展示了三个场景, 图 5 (a) 和 (d) 为场景 1、图 5 (b) 和 (e) 为场景 2、图 5 (c) 和 (f) 为场景 3. 场景 1 中, 改进的模型检测到左下角漏检目标并修正了 (a) 中与背景相似误判为蓝色安全帽的小目标; 场景 2 中, 面对右侧密集且互相遮挡的小目标, 改进后的算法几乎无漏检情况; 场景 3 中, 对白色头发的目标, 改进的算法修正了原算法误判的白色安全帽的检测. 可以看到改进后的 Deep SORT 在跟踪性能得到改善, 可以应对实时检测中目标被遮挡或多目标、小目标的

图 5　Deep SORT 与改进模型在 MOT16-09 测试对比. (a) Deep SORT 在场景 1 的检测结果; (b) Deep SORT 在场景 2 的检测结果; (c) Deep SORT 在场景 3 的检测结果; (d) 改进模型在场景 1 的检测结果; (e) 改进模型在场景 2 的检测结果; (f) 改进模型在场景 3 的检测结果

Fig.5　Comparison of Deep SORT and improved Deep SORT in MOT16-09 detection: (a) detection results of Deep SORT in scenario 1; (b) detection results of Deep SORT in scenario 2; (c) detection results of Deep SORT in scenario 3; (d) detection results of improved Deep SORT in scenario 1; (e) detection results of improved Deep SORT in scenario2; (f) detection results of improved Deep SORT in scenario 3

情况.

3.2 对比实验

本节使用五种单阶段算法和现有的安全帽识别方法进行对比实验, 表 3 列出了 5 种算法在 GDUT-HWD 数据集上的检测与识别指标. 表中的 P_{Red}、P_{White}、P_{Yellow} 和 P_{Blue} 分别代表佩戴红色、白色、黄色和蓝色安全帽的准确率, P_{None} 表示未佩戴安全帽的准确率, P 表示模型检测精度, R 为模型召回率, Weight 为模型权重大小, Time 为单张图片检测时间, fps 表示模型部署在开发板 Jetson Xavier NX 上每秒能检测的帧数. 在输入端, YOLOv3 系列的图像分辨率为 416×416, YOLOv4 和 YOLOv5 图像分辨率为 640×640. YOLOv3-tiny 速度更快, 但各项指标均低于其他 YOLO 系列模型. YOLOv4 召回率仅为 81.1%, 在实际检测中不能检测到所有安全帽, 模型权重 237.4 MB, 模型训练时间长且难以部署. 在本实验进行对比的“单阶段”算法中, YOLOv5 模型为最轻量、检测效率最好的模型, 原模型权重仅有 14.2 MB, 在测试集平均每张检测时间 2 ms, 改进后的 DS-YOLOv5 以少量的参数数量和计算量为代价, 提高了 3% 的检测精度, 本模型 mAP 值和召回率都处于领先地位.

对比实验中使用模型在主机上测试单张检测时间和模型在嵌入式开发板 Jetson Xavier NX 部署测得的 fps, 对模型效率和实用性进行直观表示. SSD 和 YOLOv4 因网络结构和模型过大难以在嵌入式设备进行检测, YOLOv3-tiny 的检测速度优于 YOLOv3 模型, YOLOv4 由于模型过大, 难以在开发板上实现实时检测. YOLOv5 模型部署后 fps 可以达到 25, 本模型在改进后 fps 为 18. 对于非节奏极其快速的场景视频外, fps 达到 15 是保持视频画面连续的最低帧率. 参数增加在一定程度影响 DS-YOLOv5 模型的检测速度, 但仍能达到视频检测的需求.

为了验证 DS-YOLOv5 有效性, 将 DS-YOLOv5 模型与研究安全帽检测工作[8−9,28−29] 中所提及的模型在 GDUT 测试集上进行对比实验, 从表 4 中可以看到, 对比其他模型 DS-YOLOv5 的 mAP 分别提升了 10.8%、1.8%、2.4% 和 9.4%. 最新的 YOLOXs 在准确率上处于领先地位, 但本模型的召回率和 mAP 值最高. 模型检测速度与模型大小决定了模型是否很好地应用于实际场景下, 在表 4 中可以看到 DS-YOLOv5 模型不论是在模型大小、模型参数或检测时间上都有明显优势, 模型参数量小代表模型空间复杂度低, 有利于在边缘化设备进行部署, 实验中 YOLO 系列和改进的 MobileNet-SSD 网络参数量有明显优势, DS-YOLOv5 模型仅比文献[9] 模型参数量多 0.1 MB. DS-YOLOv5 模型大小为各文献模型的 1%、45%、43% 和 64%, 本模型网络参数量较小且性能优越, 可满足工业场景下的实时性要求.

图 6 给出了几种复杂环境下 DS-YOLOv5 模

表 3 基于 GDUT 数据集的对比实验结果

Table 3 Results of comparison experiments based on the GDUT dataset

Model	P_{None}/%	P_{Red}/%	P_{White}/%	P_{Yellow}/%	P_{blue}/%	P/%	R/%	mAP/%	Time/ms	Weight/MB	fps
SSD512[6]	74.8	78.8	79.5	86.3	80.8	83.5	79.2	81.6	36.8	34.6	—
YOLOv3[10]	82.4	92.4	75.7	81.9	94.4	86.9	80.4	86.2	14.5	84.3	2
YOLOv3-tiny[10]	74.1	82.8	75.3	80.8	85.0	84.5	76.3	79.6	6.4	28.2	12
YOLOv4[11]	83.4	94.2	86.1	92.0	95.7	92.4	81.1	90.3	14.2	237.4	—
YOLOv5[12]	90.2	93.6	91.4	92.6	93.4	92.1	90.6	92.5	2.0	14.2	25
DS-YOLOv5	92.5	95.3	96.2	98.6	95.0	89.6	98.0	95.5	2.2	15.7	18

表 4 安全帽检测模型对比实验

Table 4 Results of the comparison of different safety helmet detection models

Detection models	P/%	R/%	mAP/%	Weights/MB	Param/10^6	Time/ms
SSD-RPA512[8]	73.3	74.4	84.7	158.9	52.6	42.4
MobileNet-SSD[9]	84.3	90.5	86.1	24.3	8.1	19.3
Improved YOLOX[28]	94.2	90.3	93.7	34.5	8.9	17.7
Improved YOLOv4-tiny[29]	93.2	88.4	92.9	36.5	13.2	14.9
DS-YOLOv5	89.6	98.0	95.5	15.7	8.2	2.2

图 6　复杂环境下 DS-YOLOv5 对安全帽佩戴情况检测与识别结果. (a) 光线不足; (b) 遮挡情况; (c) 目标尺度差异

Fig.6　Detection and recognition results of the DS-YOLOv5 model for the wearing of safety helmets in complex environments: (a) underlighting conditions; (b) occlusion conditions; (c) different target scales

型对安全帽佩戴情况检测与识别结果. 图中展示被检测安全帽的检测框、分类和置信度分数, 其中图 6(a) 光线不足场景; 图 6 (b) 在多个目标互相遮挡场景; 图 6(c) 由于拍摄距离造成安全帽尺度差异场景, 检测结果表明 DS-YOLOv5 融合多尺度特征对特征学习充分, 在图 6 (a) 光线不足的环境下仍能得到正确的检测结果, 加入改进的 Deep SORT 有利于应对遮挡情况和尺寸差异较大的目标, 在图 6 (b) 中正确检测目标遮挡将近 1/2 的 None 标签, 在图 6 (c) 中三个尺度差异较大的目标能正确检测并识别. 在图 5 中视频检测和图 6 中复杂环境的图像检测中, 分别展示了包含光线不足、小目标、多目标、遮挡、目标与背景色相似的例子, DS-YOLOv5 模型表现稳定, 可以适应复杂环境下的安全帽检测.

4　结论

本文针对复杂环境下安全帽佩戴情况的实时检测与识别任务, 以 YOLOv5 模型为基础网络, 提出结合 Deep SORT 多目标跟踪, 融合 BiFPN 和 Transformer 注意力机制的 DS-YOLOv5 安全帽检测与识别模型. 在 GDUT-HWD 数据集对比实验和消融实验结果表明, DS-YOLOv5 模型可以有效提升安全帽佩戴情况检测的准确性、鲁棒性和实时性. 而且, DS-YOLOv5 模型可以有效应对包括亮度变化、多目标等复杂环境的影响. 下一步, 我们将对该模型进行轻量化设计并移植到边缘化设备, 以便将其应用于生产和施工现场对安全帽佩戴情况进行实时检测与识别.

参　考　文　献

[1] Liu X H, Ye X N. Skin color detection and hu moments in helmet recognition research. *J East China Univ Sci Technol* (*Nat Sci Ed*), 2014, 40(3): 365
(刘晓慧, 叶西宁. 肤色检测和 Hu 矩在安全帽识别中的应用. 华东理工大学学报(自然科学版), 2014, 40(3): 365)

[2] Li K, Zhao X G, Bian J, et al. Automatic safety helmet wearing detection // 2017 *IEEE 7th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems* (*CYBER*). Honolulu, 2017: 617

[3] Zhang L Y, Wu W H, Niu H M, et al. Summary of application research on helmet detection algorithm based on deep learning. *Comput Eng Appl*, 2022, 58(16): 1
(张立艺, 武文红, 牛恒茂, 等. 深度学习中的安全帽检测算法应用研究综述. 计算机工程与应用, 2022, 58(16): 1)

[4] Yogameena B, Menaka K, Saravana Perumaal S. Deep learning-based helmet wear analysis of a motorcycle rider for intelligent surveillance system. *IET Intell Transp Syst*, 2019, 13(7): 1190

[5] Ferdous M, Masudul Ahsan S M. Multi-scale safety hardhat wearing detection using deep learning: A top-down and bottom-up module // 2021 *International Conference on Electrical, Communication, and Computer Engineering* (*ICECCE*). Kuala Lumpur, 2021: 1

[6] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector // *European Conference on Computer Vision*. Amsterdam, 2016: 21

[7] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection // 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). Las Vegas, 2016: 779

[8] Wu J X, Cai N, Chen W J, et al. Automatic detection of hardhats worn by construction personnel: A deep learning approach and benchmark dataset. *Autom Constr*, 2019, 106: 102894

[9] Xu X F, Zhao W F, Zou H Q, et al. Detection algorithm of safety helmet wear based on MobileNet-SSD. *Comput Eng*, 2021, 47(10): 298
(徐先峰, 赵万福, 邹浩泉, 等. 基于 MobileNet-SSD 的安全帽佩戴检测算法. 计算机工程, 2021, 47(10): 298)

[10] Redmon J, Farhadi A. Yolov3: An incremental improvement [J/OL]. *arXiv eprints* (2018-4-8) [2022-11-11]. https://arxiv.org/abs/1804.02767

[11] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal speed and accuracy of object detection [J/OL]. *arXiv* (2020-4-23) [2022-11-11]. https://arxiv.org/abs/2004.10934

[12] Ultralytics. yolov5 [DB/OL]. *Ultralytics* (2023) [2022-11-11] https://github.com/ultralytics/yolov5

[13] Zhao R, Liu H, Liu P L, et al. Research on safety helmet detection algorithm based on improved YOLOv5s. *J B Univ Aeronaut*

Astronaut, 2021: 1
(赵睿, 刘辉, 刘沛霖, 等. 基于改进 YOLOv5s 的安全帽检测算法. 北京航空航天大学学报, 2021: 1)

[14] Yue H, Huang X M, Lin M H, et al. Helmet-wearing detection based on improved YOLOv5. *Comput Mod*, 2022(6): 104
(岳衡, 黄晓明, 林明辉, 等. 基于改进 YOLOv5 的安全帽佩戴检测. 计算机与现代化, 2022(6): 104)

[15] Huang L Q, Jiang L W, Gao X F. Improved algorithm of YOLOv3 for real-time helmet wear detection in videos. *Mod Comput*, 2020(30): 32
(黄林泉, 蒋良卫, 高晓峰. 改进 YOLOv3 的实时性视频安全帽佩戴检测算法. 现代计算机, 2020(30): 32)

[16] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need // *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*. Long Beach, 2017: 6000

[17] Zhu X K, Lyu S C, Wang X, et al. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios [J/OL]. *arXiv* (2021-8-26) [2022-11-11]. https://arxiv.org/abs/2108.11539

[18] He K M, Zhang X Y, Ren S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans Pattern Anal Mach Intell*, 2015, 37(9): 1904

[19] Liu S, Qi L, Qin H F, et al. Path aggregation network for instance segmentation // 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, 2018: 8759

[20] Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric // 2017 *IEEE International Conference on Image Processing* (*ICIP*). Beijing, 2017: 3645

[21] Kalman R E. A new approach to linear filtering and prediction problems. *J Basic Eng*, 1960, 82(1): 35

[22] Zheng Z H, Wang P, Liu W, et al. Distance-IoU loss: Faster and better learning for bounding box regression. *Proc AAAI Conf Artif Intell*, 2020, 34(7): 12993

[23] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers // *European Conference on Computer Vision*. Glasgow, 2020: 213

[24] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [J/OL]. *arXiv* (2021-6-3) [2022-11-11]. https://arxiv.org/abs/2010.11929

[25] Tan M X, Pang R M, Le Q V. EfficientDet: scalable and efficient object detection // 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). Seattle, 2020: 10778

[26] Zheng L, Shen L Y, Tian L, et al. Scalable person re-identification: A benchmark // 2015 *IEEE International Conference on Computer Vision* (*ICCV*). Santiago, 2015: 1116

[27] Milan A, Leal-Taixé L, Reid I, et al. MOT16: A benchmark for multi-object tracking [J/OL]. *arXiv eprints* (2016-5-3) [2022-11-11]. https://ui.adsabs.harvard.edu/abs/2016arXiv160300831M

[28] Lv Z X, Wei X, Ma Z G. Improve the lightweight safety helmet detection method of YOLOX. *Comput Eng Appl*, 2022, 59(1): 61
(吕志轩, 魏霞, 马志钢. 改进 YOLOX 的轻量级安全帽检测方法. 计算机工程与应用, 2022, 59(1): 61)

[29] Wang J B, Wu Y X. Helmet wearing detection algorithm of improved YOLOV4-tiny, *Comput Eng Appl*, 2023, 59(4): 183
(王建波, 武友新. 改进 YOLOv4-tiny 的安全帽佩戴检测算法. 计算机工程与应用. 2023, 59(4): 183)