文章编号:2096-1472(2024)01-0074-05
DOI:10.19644/j.cnki.issn2096-1472.2024.001.017

基于改进 YOLOv5s-pose 的多人人体姿态估计

蒋锦华,庄丽萍,陈 锦,姚洪泽,蔡志明

(福建工程学院,福建 福州 350118)

✉ 1422881869@qq.com; 1504879132@qq.com; 464525151@qq.com; 894130700@qq.com; caizm@fjut.edu.cn

摘 要:为了提高多人人体姿态检测的准确率,本研究采用 YOLOv5s 模型用于多人人体姿态检测并对模型进行改进。首先,引入坐标注意力(Coordinate Attention)模块改进骨干网络,将注意力资源分配给关键区域,降低复杂环境中的背景干扰,增强模型对多人目标的精准定位能力。其次,使用双向特征金字塔网络改进 YOLOv5s 的特征融合网络,增强网络的信息表达能力。实验结果表明:在多人人体姿态 MS COCO2017 验证集上,经改进的 YOLOv5s 算法的检测平均精度高达 61.9%,相比原始 YOLOv5s 网络,平均精度提升了 1.5%。由此可见,改进后的网络能更加精准、有效地检测多人人体姿态。

关键词:多人人体姿态检测;YOLOv5s;双向特征金字塔网络;检测精度

中图分类号:TP183 **文献标志码**:A

Multi-person Pose Estimation Based on Improved YOLOv5s-pose

JIANG Jinhua, ZHUANG Liping, CHEN Jin, YAO Hongze, CAI Zhiming

(*Fujian University of Technology, Fuzhou 350118, China*)

✉ 1422881869@qq.com; 1504879132@qq.com; 464525151@qq.com; 894130700@qq.com; caizm@fjut.edu.cn

Abstract: In order to improve the accuracy of multi-person pose detection, this paper proposes to use and improve the YOLOv5s model for multi-person pose detection. Firstly, the Coordinate Attention module is introduced to improve the backbone network by allocating attention resources to key areas, reducing background interference in complex environments, and enhancing the model's precise localization ability for multi-person targets. Secondly, a bidirectional feature pyramid network is used to improve the feature fusion network of YOLOv5s and enhance the network's information expression ability. The experimental results show that on the MS COCO2017 validation set for multi-person poses, the improved YOLOv5s algorithm achieves an average detection precision of 61.9% and the average accuracy increases by 1.5%, compared to the original YOLOv5s network. It can be seen that the improved network can more accurately and effectively detect multiple human body postures.

Key words: multi-person pose detection; YOLOv5s; bidirectional feature pyramid network; detection precision

0 引言(Introduction)

深度学习技术在图像分割、目标检测等方向取得的一系列突破,促进了多人人体姿态目标检测算法的进步。

目前,基于深度学习的多人人体姿态检测方法分为双阶段目标检测算法和单阶段检测算法,其中针对双阶段目标检测,如 PAPANDREOU 等[1]在第一阶段采用 FasterR-CNN 检测人体;在第二阶段采用 ResNet 预测每个关键点的热力图和偏移量,通过融合得到关键点的精确位置。CAO 等[2]建立了一个 OpenPose 检测器,加快了人体关键点的检测速度。CHENG 等[3]提出一种尺度感知的高分辨率网络(HigherHRNet),通过生成高分辨率热图来更精确地定位人体关键点。在单阶段检测算法中,NIE 等[4]首次提出了单阶段的多人姿态估计网络,可以直接预测每个人的位置和关键点。MCNALLY 等[5]提出一个密集基于锚的单阶段检测框架,同时检测关键点对象和姿态对象,能够更加快速地得出检测结果。

收稿日期:2023-04-09
基金项目:福建工程学院科研启动基金(GY-Z21064)

单阶段检测与双阶段检测相比,检测速度更快,但准确率相对较低。本研究在单阶段检测的基础上,进一步提高人体姿态的检测准确率。在单阶段检测算法中 YOLO 系列的 YOLOv5s 网络具有计算高效、实时性强等特点,更适合在实际场景中检测多人人体姿态。因此,本文选择 YOLOv5s 网络作为基准模型。

1 YOLOv5s 目标检测模型(YOLOv5s object detection model)

1.1 YOLOv5s 网络结构

YOLOv5s 的结构由四个部分组成:输入端、Backbone(骨干网络)、Neck(特征金字塔)和 Head(目标检测)。输入端调节输入图片的尺寸大小,Backbone 部分进行特征提取,Neck 部分通过将特征与位置信息融合使模型获得更丰富的特征信息,Head 部分进行最终的预测输出。

Backbone 部分使用 Darknet53 作为特征提取网络,主要由 Focus 网络和 CSPNet 结构组成。Focus 结构对特征图进行切片操作,使特征图的长和宽都缩小了 1/2,减少了算法的计算量,加快了计算速度[6]。CSPNet 结构用于提取输入图像特征信息,将梯度变化完全集成到特征图中,在减少模型参数量的同时,兼顾了推理速度和准确率,能够更好地解决其他大型卷积神经网络中的梯度信息冗余问题[7]。

Neck 部分使用特征金字塔+路径聚合(FPN+PAN)结合的 PANet 结构,特征金字塔的结构为自上而下,该结构使顶层特征图享受来自底层的特征信息,从而提升对较大目标的检测效果[8]。

Head 的主体是 4 个 Detect 检测器,基于网格的锚框在不同尺度的特征图,检测器分别用来检测大、中、小目标。YOLOv5s 网络结构如图 1 所示。

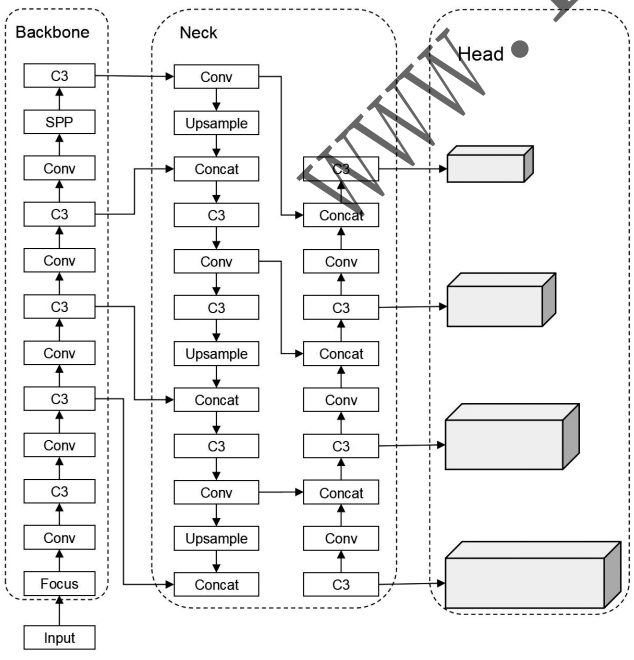


图 1　YOLOv5s 网络结构

Fig. 1 YOLOv5s network structure

1.2 人体关键点检测过程

检测基于 YOLOv5s 目标检测框架进行多人人体姿态估计,该模型能够在一次前向传递中联合检测多个人体边界框及其相应的 2D 姿态。对于输入的图像,将一个人的所有关键点与它对应的目标框联系起来,存储其整个 2D 姿态和边界框。使用 CSP-darknet53 作为主干网络,对人体关键点进行特征提取,生成不同尺度的特征图。使用 PANet 融合主干网络输出的不同尺度特征,生成四个不同尺度的检测头。每个检测头分别用于预测框和关键点。

2 改进后的结构与分析(Improved structure and analysis)

为了提高 YOLOv5s 在多人人体姿态中的检测精度,本研究对模型做出以下改进:(1)对主干网络进行改进,在主干网络卷积层后的每一层中都加入 CA 注意力模块,再输出给 C3 模块,通过将位置坐标信息嵌入信道注意力中,使移动网络能够大范围关注检测目标,同时避免产生大量的计算,提高了对人体目标的定位精度,使其能够抑制其他无用特征,进而更多地关注人体关键点这一特征信息。(2)对 Neck 部分进行改进,在 Neck 部分的特征连接层模块中加入一个可学习的权重参数,形成一种简单且高效的加权双向特征金字塔网络(BiFPN),将 YOLOv5s 原始的 Concat 替换成新构建的 Bifpn_Concat,提升网络对不同尺度目标的特征融合能力。改进后的 YOLOv5s-pose 网络结构如图 2 所示。

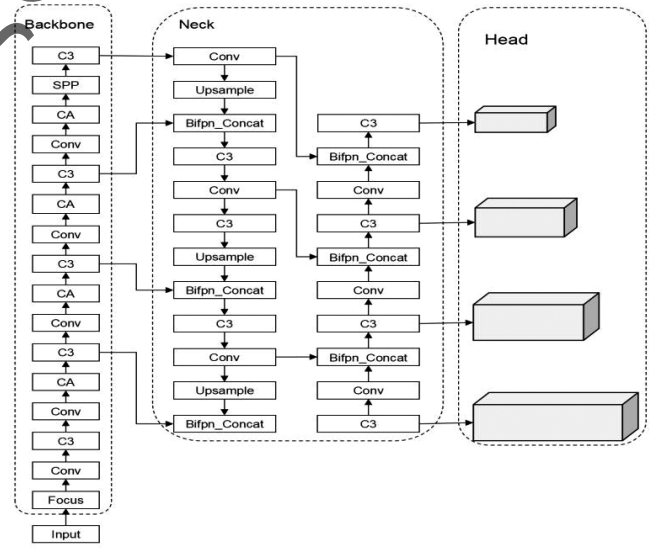


图 2　改进后的 YOLOv5s-pose 网络结构

Fig. 2 Improved YOLOv5s-pose network structure

2.1 加入 CANet 模块

由于待检测数据集为多人人体目标,存在人体目标被遮挡、实际场景中大小不一且分布疏密不均等问题,为了进一步增强网络对待检测目标的特征提取能力且不影响检测的实时性,又引入了一种更加高效且轻量级的注意力模块,称之为"坐标注意力机制",即 CANet[9]。

CANet是对输入特征图进行水平方向和垂直方向上的平均池化，其本质是空间注意力，在通道注意力当中嵌入位置信息后，赋予空间中不同位置处不同的权重系数。从空间上来看，类似于从两个方向对网络进行建模，对不同的特征进行一个融合再提取的过程。CA的整体结构如图3所示。

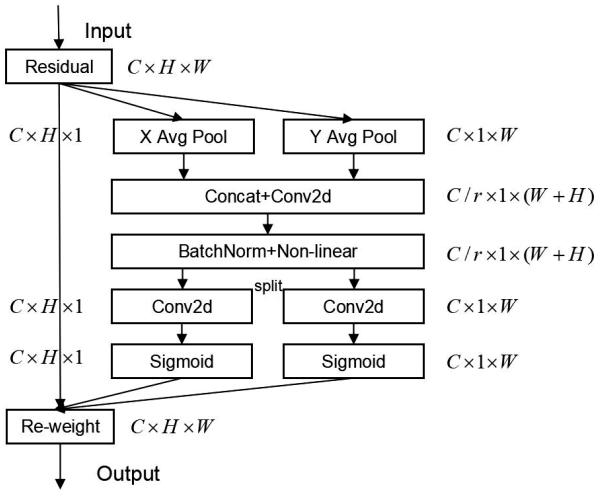


图3 CA注意力机制模块结构

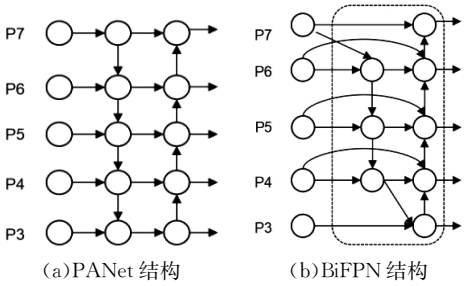
Fig. 3 CA attention mechanism module structure

由图3可知，将输入特征图分别按 X 轴和 Y 轴方向进行池化，对每个通道进行编码，产生 $C\times H\times 1$ 和 $C\times 1\times W$ 形状的特征图。将所提取到的特征图按空间维度进行拼接，再通过卷积和Sigmoid激活函数得到坐标注意力。通过这种方式所产生的一对感知特征图可以使CA注意力能够在一个通道内捕获长距离的依赖关系，并且有助于保留精确的位置信息，使网络能够更加准确地定位对象。

2.2 BiFPN——加权双向特征金字塔多尺度特征融合

BiFPN是在PANet的基础上改进而来的。双向特征金字塔结构(BiFPN)运用双向融合思想，重新构造了自顶向下和自底向上的路线，对不同尺度的特征信息进行融合，通过上采样和下采样统一特征分辨率尺度，并且在同一尺度的特征图之间建立双向连接，在一定程度上解决了特征信息遗失的问题[10]。

原始的YOLOv5s网络中，PANet作为YOLOv5s的特征融合网络，其结构如图4(a)所示，虽然可以实现浅层信息的传递和高层特征图强语义信息的融合，但是浅层和高层两部分融合采用的相加运算并没有相关的权重设计，而且只有一边输入没有特征融合的节点，冗余节点对特征融合的作用甚微，增加了额外的参数和计算量。基于以上问题，本文基于BiFPN结构修改YOLOv5s网络的Neck部分，将Neck部分中PANet的节点连接方式做出部分改变，减少了对网络特征融合贡献度较小的不必要连接，增加了输入节点和输出节点处于同一层时二者的连接，BiFPN结构节点连接方式如图4(b)所示。在YOLOv5s网络的特征融合部分中，将负责特征信息融合的张量拼接操作结合加权双向特征金子塔(BiFPN)，结合后的张量拼接操作记作Bifpn_Concat，其结构如图5所示。



(a)PANet结构 (b)BiFPN结构

图4 特征网络结构

Fig. 4 Feature network structure

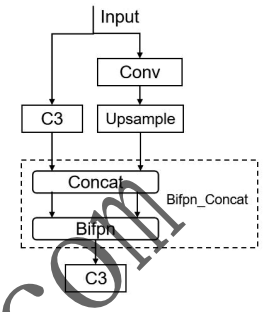


图5 Bifpn_Concat模块结构

Fig. 5 Bifpn_Concat module structure

BiFPN使用加权特征融合的方式为每个特征添加一个额外的权重。使网络可以不断调整权重，确定每个输入特征对输出特征的重要性；快速归一方法如公式(1)所示，用来约束每个权重的大小，使权重大小保持在0～1，提高模型在GPU上的运算速度。

$$Output=\sum_i \frac{\omega_i}{\sum_j \omega_j+\varepsilon}\cdot I_i \tag{1}$$

其中：ω_i 代表可以学习的权重大小；I_i 表示输入特征；$\varepsilon=10^{-4}$，是一个很小的值，用来避免数值不稳定。

3 实验结果与分析(Experimental results and analysis)

3.1 实验环境与数据集

训练使用AutoDL云端服务器，在云端服务器上的实验环境配置如下：操作系统为Ubuntu 18.04、PyTorch 1.9.0框架、CUDA11.1、Python3.8，使用RTX3090显卡一块，*batch-size* 设置为64，*epoch* 为300。训练和测试时将数据集中的图像尺寸固定为640×640，学习率为0.01。

本次训练的数据集为公共数据集MS COCO2017，MS COCO关键点检测数据集是目前主流的二维人体姿态估计数据集之一，它包含20万张以上的图像和25万个带有关键点注释的人体实例，每个实例最多包含17个人体关键点，并对这些关键点进行了标注。在Train2017(约57 000张图像，包含150 000个人体实例)数据集上进行网络模型的训练，在Val2017数据集上进行网络模型的验证和测试。

3.2 评价指标

对于COCO数据集，采用官方指定关节点相似度 *OKS*(Object Keypoint Similarity)为模型性能评价的度量方法。*OKS* 定义了不同人体关键点之间的相似性，值为0～1，越接近

1,说明预测得到的人体关节点与数据集标注的真实值越相似,预测效果越好,OKS 的公式如下:

$$OKS=\frac{\sum_i \exp(-d_i^2/2s^2k_i^2)\delta(v_i>0)}{\sum_i \delta(v_i>0)} \tag{2}$$

其中:i 表示关节点的类型,d_i 表示检测出来的关键点与其相应的标签值之间的欧氏距离,s 表示目标比例,v_i 表示真实值的可见性标志,δ 函数表示当关键点被标注时才纳入计算,k_i 表示控制衰减的每个关键点的常数。

本文选用 AP、AP_{75}、AP_M、AP_L、AR 为评价指标,AP 为 $OKS=0.50,0.55,\cdots,0.95$ 时,每种检测类型的准确率,用于预测关键点的平均精度值。AP_{75} 为在 $OKS=0.75$ 时关键点的准确率。AP_M 为中等目标检测的 AP 值,AP_L 为大目标检测的 AP 值,AR 表示 $OKS=0.50,0.55,\cdots,0.95$ 这 10 个阈值上的平均查全率。准确率 P、查全率 R 的具体计算如公式(3)和公式(4)所示。平均精度值 AP 的计算如公式(5)所示:

$$P=\frac{TP}{TP+FP} \tag{3}$$

$$R=\frac{TR}{TR+FN} \tag{4}$$

$$AP=\int_0^1 P\mathrm{d}R \tag{5}$$

其中:TP 为正样本被正确识别为正样本的数量,FP 为负样本被错误识别为正样本的数量,FN 为正样本被错误识别为负样本的数量,N 为目标的类别数。AP 的意义是 P-R 曲线所包围的面积。

3.3 结果与分析

本实验比较了原始的 YOLOv5s 网络和改进后的 YOLOv5s-CA 在人体姿态检测中的各项精度指标,融合 CA 模块前后检测结果对比见表 1。从表 1 中可以看出,在多人人体姿态检测中,改进后的 YOLOv5s-CA 算法相比于原始的 YOLOv5s 网络,AP 指标提升了 0.6%,AP_{75} 指标提升了 1.1%,中等人体目标的准确率 AP_M 提升了 0.3%,大目标的准确率提升了 0.8%。由此说明:在 YOLOv5s 的主干特征提取网络中加入 CA 坐标注意力机制模块能提升人体姿态关键点的检测精度。

表 1　融合 CA 模块前后检测结果对比

Tab.1 Comparison of test results before and after CA module fusion

算法名称	AP	AP_{75}	AP_M	AP_L	AR
YOLOv5s-pose	0.568	0.604	0.470	0.709	0.640
YOLOv5s-pose-CA	0.574	0.615	0.473	0.717	0.643

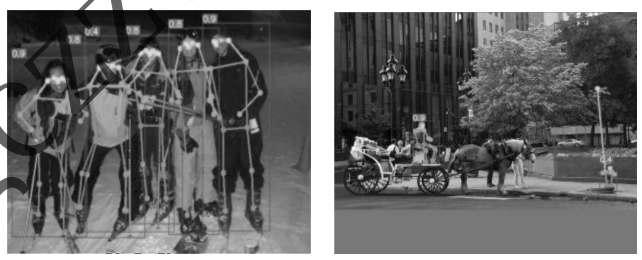
在上述改进的基础上,本研究引入双向特征金字塔网络(BiFPN),进一步改进 YOLOv5s 网络的颈部(Neck)结构,将改进后的检测效果与原始的 YOLOv5s 进行对比,结果如表 2 所示。从表 2 中的数据可以发现,平均精度 AP 提升了 0.5%,AP_{75} 指标提升了 1.5%,中等人体目标的准确率 AP_M 提升了 0.8%;在改进 CA 注意力机制的基础上,AP_{75} 指标的检测精度又提升了 0.4%。实验证明,在 YOLOv5s 的网络模型中融入 BiFPN 模块以改进原来的 PANet,可以进一步提升人体目标和人体关键点的检测精度。

表 2　融合 CA 模块和 BiFPN 模块后与原始 YOLOv5s 网络检测性能对比

Tab.2 Comparison of the detection performance of the CA module and BiFPN module with the original YOLOv5s network

算法名称	AP	AP_{75}	AP_M	AP_L	AR
YOLOv5s-pose	0.568	0.604	0.470	0.709	0.640
YOLOv5s-pose-CA-BiFPN	0.576	0.619	0.478	0.717	0.647

图 6 与图 7 分别为 YOLOv5s-pose 算法改进前和改进后的检测效果对照,对比图 6(a)与图 7(a)的检测效果可以看出,改进后的算法应用于多人人体姿态关键点检测,检测结果更准确。如图 6(a)所示,原始的 YOLOv5s-pose 网络存在关键点对应人体姿态错乱现象,如图 7(a)所示,改进后的 YOLOv5s-pose 网络在关键点及对应人体姿态目标上的检测精度有所提升。从图 6(b)中的检测效果可以看出,当人体目标被遮挡,只有部分人体部位显示时,改进前的 YOLOv5s-pose 网络存在人体目标漏检的情况,对于小目标人体,关键点及人体姿态检测精度较低,检测效果较差。对比图 6(b)与图 7(b)的检测效果可以看出,改进后的 YOLOv5s-pose 网络能够较好地检测出被遮挡部分的人体目标及其对应的人体姿态。此外,对于检测小目标的人体姿态,改进后的网络的检测效果更佳。



(a)YOLOv5s-pose 图片 1 检测效果　(b)YOLOv5s-pose 图片 2 检测效果

图 6　YOLOv5s-pose 网络多人人体姿态检测效果图

Fig. 6 YOLOv5s-pose network multi-person human pose detection rendering

(a)改进 YOLOv5s-pose 图片 1 检测效果　(b)改进 YOLOv5s-pose 图片 2 检测效果

图 7　改进 YOLOv5s-pose 网络多人人体姿态检测效果图

Fig. 7 Improved YOLOv5s-pose network multi-person human pose detection rendering

4 结论(Conclusion)

本文提出了一种基于 YOLOv5s 网络改进的多人人体姿态检测算法,为了提升模型在 COCO 人体姿态数据集中的检测精度,对模型进行了改进。首先在主干网络中通过融入 CA 坐标注意力机制模块提升主干网络部分对人体关键点的特征提取能力,其次基于 BiFPN 结构在 Neck 部分对原始网络进行

优化改进，增强了不同目标尺度的融合度，进一步提升了模型对多人人体姿态目标的检测能力。实验结果表明，本文提出的方法在人体姿态 COCO 数据集中的检测能力更强，在面对部分人体被遮挡时，算法的鲁棒性更好，检测精度更高。

参考文献(References)

[1] PAPANDREOU G，ZHU T，KANAZAWA N，et al. Towards accurate multi-person pose estimation in the wild[C]//IEEE. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway：IEEE，2017：3711-3719.

[2] CAO Z，HIDALGO G，SIMON T，et al. OpenPose：real-time multi-person 2D pose estimation using part affinity fields[J]. IEEE transactions on pattern analysis and machine intelligence，2021，43(1)：172-186.

[3] CHENG B W，XIAO B，WANG J D，et al. HigherHRNet：scale-aware representation learning for bottom-up human pose estimation[C]//IEEE. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway：IEEE，2020：5385-5394.

[4] NIE X C，FENG J S，ZHANG J F，et al. Single-stage multi-person pose machines[C]//IEEE. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Piscataway：IEEE，2019：6950-6959.

[5] MCNALLY W，VATS K，WONG A，et al. Rethinking keypoint representations：modeling keypoints and poses as objects for multi-person human pose estimation[C]//AVIDAN S，BROSTOW G，CISSE M，et al. Proceedings of the 17th European Conference on Computer Vision. Cham：Springer，2022：37-54.

[6] 吕禾丰，陆华才. 基于 YOLOv5 算法的交通标志识别技术研究[J]. 电子测量与仪器学报，2021，35(10)：137-144.

[7] 邓天民，谭思奇，蒲龙忠. 基于改进 YOLOv5s 的交通信号灯识别方法[J]. 计算机工程，2022，48(9)：55-62.

[8] LIU S，QI L，QIN H F，et al. Path aggregation network for instance segmentation[C]//IEEE. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway：IEEE，2018：8759-8768.

[9] HOU Q B，ZHOU D Q，FENG J S. Coordinate attention for efficient mobile network design[C]//IEEE. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway：IEEE，2021：13708-13717.

[10] TAN M X，PANG R M，LE Q V. EfficientDet：scalable and efficient object detection[C]//IEEE. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway：IEEE，2020：10778-10787.

作者简介：

蒋锦华(1998-)，女，硕士生。研究领域：计算机视觉，人体姿态识别。

庄丽萍(1998-)，女，硕士生。研究领域：深度学习，人体姿态识别。

陈　锦(1998-)，男，硕士生。研究领域：计算机视觉，深度学习。

姚洪泽(2001-)，男，硕士生。研究领域：计算机视觉，深度学习。

蔡志明(1977-)，男，博士，教授。研究领域：计算机视觉与图像处理。本文通信作者。

(上接第 67 页)

参考文献(References)

[1] WANG C Y，BOCHKOVSKIY A，LIAO H Y M. YOLOv7：Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[DB/OL]. (2022-07-06) [2023-05-22]. https://arxiv.org/abs/2207.02696.

[2] 李昌夏，加文浩，黄政龙，等. 基于 YOLOv5 的实时抽烟检测研究[J]. 电脑知识与技术，2022，18(8)：100-102.

[3] 邢予权. 基于手势识别的监控场景下抽烟检测[D]. 杭州：浙江工业大学，2020.

[4] 李倩. 基于深度学习的烟支检测技术研究与应用[D]. 西安：西安邮电大学，2020.

[5] 陈睿龙，罗磊，蔡志平，等. 基于深度学习的实时吸烟检测算法[J]. 计算机科学与探索，2021，15(2)：327-337.

[6] 刘婧，杨旭，刘董经典，等. 基于人体关节点的多人吸烟动作识别算法[J]. 计算机工程与应用，2021，57(1)：234-241.

[7] 向凯. 高速行驶车辆的实时检测识别方法研究[D]. 成都：电子科技大学，2020.

[8] 方路平，何杭江，周国民. 目标检测算法研究综述[J]. 计算机工程与应用，2018，54(13)：11-18，33.

[9] 孙月莹，陈俊霖，张胜茂，等. 基于改进 YOLOv7 的毛虾捕捞渔船作业目标检测与计数方法[J]. 农业工程学报，2023，39(10)：151-162.

[10] 成浪，敬超. 基于改进 YOLOv7 的 X 线图像旋转目标检测[J]. 图学学报，2023，44(2)：324-334.

[11] 涂宙霖，陈涵深. 基于 YOLOv7 与 Jetson Orin 的路面破损检测系统的设计与实现[J]. 电脑知识与技术，2023，19(9)：50-52.

[12] 乔羽. 基于 Mask R-CNN 泳池中溺水行为检测系统的设计与实现[D]. 青岛：青岛大学，2019.

作者简介：

孙　冰(2001-)，女，本科生。研究领域：计算机视觉。

李　好(2002-)，男，本科生。研究领域：计算机视觉。

黄鑫凯(2002-)，男，本科生。研究领域：计算机视觉。

任长宁(1978-)，男，硕士，副教授。研究领域：人工智能与数据分析。

邹启杰(1978-)，女，博士，副教授。研究领域：计算机视觉。