

融合注意力机制的 OpenPose 人体跌倒检测算法

孟彩霞^{1,2)}, 薛洪秋³⁾, 石磊^{1,3)*}, 高宇飞³⁾, 卫琳³⁾

¹⁾ (郑州大学计算机与人工智能学院 郑州 450001)

²⁾ (铁道警察学院图像与网络侦查系 郑州 450053)

³⁾ (郑州大学网络空间安全学院 郑州 450002)

(shilei@zzu.edu.cn)

摘要: 人员密集场所跌倒事件易引发公共安全问题, 对人体跌倒进行实时监测和预警可降低安全风险. 针对现有基于姿态估计跌倒检测方法模型规模大、时效性差等问题, 提出一种融合注意力机制的 OpenPose 人体跌倒检测算法 DSC-OpenPose. 首先借鉴 DenseNet 稠密连接思想, 将每层与之前所有层在通道维度上直接连接, 实现特征复用, 减小模型参数规模; 然后在不同阶段之间添加坐标注意力机制, 获取特征图空间方向依赖和精确位置信息, 提高姿态估计精度; 最后提出一种基于人体外椭圆参数、头部高度、下肢高度共同识别跌倒行为的方法, 实现人体目标的跌倒检测. 实验结果表明, 在 COCO 数据集上, 所提算法在模型规模和精度之间取得了较好的平衡效果; 在 real fall(RF)数据集上, 所提跌倒检测算法的准确率达到 98.2%, 精度达到 96.6%, 检测速度达到 20.2 帧/s, 且模型规模较小, 满足嵌入式设备实时推理需求.

关键词: 人体姿态估计; 密集连接; 注意力机制; 跌倒检测

中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2024.20124

OpenPose Human Fall Detection Algorithm Based on Attention Mechanism

Meng Caixia^{1,2)}, Xue Hongqiu³⁾, Shi Lei^{1,3)*}, Gao Yufei³⁾, and Wei Lin³⁾

¹⁾ (School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou 450001)

²⁾ (Department of Image and Network Investigation Technology, Railway Police College, Zhengzhou 450053)

³⁾ (School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou 450002)

Abstract: Falls in crowded places are easy to cause safety problems, and real-time monitoring can reduce safety risks. Aiming at the problems of large scale and poor timeliness of existing fall detection methods based on pose estimation, this paper proposes an OpenPose human fall detection algorithm DSC OpenPose that integrates attention mechanism. Drawing on DenseNet's dense connection idea, each layer is directly connected with all previous layers in the channel dimension to achieve feature reuse and reduce the scale of model parameters. A method of identifying fall behavior based on the outer ellipse parameters, head height and lower limb height is proposed to realize the fall detection of human objects. The experimental results on COCO dataset show that compared with other algorithms, this algorithm achieves a good balance between model size and accuracy. At the same time, the fall detection method proposed in this paper has an accuracy of 98.2%, a precision of 96.6%, and a detection

收稿日期: 2023-01-17; 修回日期: 2024-01-17. 基金项目: 国家重点研发计划(2020YFB1712401-1); 国家自然科学基金青年科学基金(62006210); 2022 年度河南省重大科技专项(221100210100, 221100211200); 河南省研究生联合培养基地项目(YJS2023JD04); 河南省重点研发与推广专项(科技攻关)(232102210095); 南阳市协同创新重大专项(22XTCX12001); 河南省高校青年骨干教师培养计划(2020GGJS256); 河南省高校重点科研项目(24B520032); 郑州警察学院基科费项目(2023TJJBKY010, 2020TJJBKY006). 孟彩霞(1982—), 女, 博士研究生, 副教授, CCF 会员, 主要研究方向为目标检测; 薛洪秋(2000—), 男, 硕士研究生, 主要研究方向为目标检测; 石磊(1967—), 男, 博士, 教授, 博士生导师, 论文通信作者, 主要研究方向为人工智能; 高宇飞(1990—), 男, 博士, 讲师, 主要研究方向为深度学习; 卫琳(1968—), 女, 硕士, 副教授, 硕士生导师, 主要研究方向为深度学习.

speed of 20.2 frame/s on the real fall (RF) data set, and the model scale is small to meet the real-time reasoning needs of embedded devices.

Key words: human pose estimation; dense connection; attention mechanism; fall detection

根据世界卫生组织的统计报告, 跌倒是造成意外伤亡的第二大原因。在公共密集场所发生跌倒容易引发踩踏事故, 导致严重的安全事故。2014 年上海外滩和 2022 年韩国梨泰院发生的大规模踩踏事件均是由于跌倒引起的, 跌倒作为威胁生命安全的异常事件, 对其检测与预警一直是研究的重点。

目前, 跌倒检测方法分为 3 种: 基于可穿戴设备的跌倒检测方法、基于场景传感器的跌倒检测方法和基于计算机视觉的跌倒检测方法。基于可穿戴设备的技术利用陀螺仪和加速度计检测跌倒行为, 但设备需要随身佩戴且连接不够稳定, 对检测精度产生一定影响。基于场景传感器的技术对利用深度传感器、红外传感器、声学传感器和振动传感器收集到的传感器数据进行处理, 实现跌倒行为识别, 但传感器易受干扰, 不能应用于室外环境。基于计算机视觉的跌倒检测是一种更好的替代方案, 其利用人体骨骼序列提取和姿态估计的方法进行跌倒行为检测, 但在人员密集的公共场所检测的实时性和准确率还有待提升, 姿态估计模型规模较大, 且提取到的关键点和骨架信息精度不足, 导致跌倒检测存在一定误差。

本文提出一种参数规模小且融合坐标注意力(coordinate attention, CA)机制的人体姿态估计(densecoordinate openpose, DSC-OpenPose)算法; 它是基于改进的姿态估计网络, 结合人体姿态外椭圆、头部高度和下肢高度, 所设计的基于多层感知机(multi-layer perceptron, MLP)的跌倒检测方法。该方法面向复杂环境特征, 在常规 CPU 或嵌入式设备上运行, 准确率高、误报率低、速度快。

1 相关工作

目前, 大部分研究人员采用基于深度学习的姿态估计方法进行跌倒检测。Kang 等^[1]提出一种使用姿态估计和门控循环单元(gate recurrent unit, GRU)的跌倒检测系统, 利用 PoseNet 进行姿态估计, 将提取到的姿态信息输入到 GRU 中进行跌倒检测, 但检测的实时性还有待提升。Chen 等^[2]和

Nogas 等^[3]分别提出使用 OpenPose 的跌倒检测方法, 其中, Chen 等^[2]结合 OpenPose, 使用髋关节中心的下降速度、人体中心线角度和边界框的宽高比检测跌倒行为; Nogas 等^[3]使用融合 OpenPose 的长短期记忆网络(long short-term memory, LSTM)网络, 并与时空信息关联来判断跌倒的发生。但是, 上述方法中姿态估计提取到的关键点和骨架信息精度不足, 导致跌倒检测存在一定误差。

姿态估计分为单人姿态估计和多人姿态估计。单人姿态估计有回归方法和身体部位检测方法, 它们的检测精度较高, 但不适合复杂环境下的跌倒检测。多人姿态估计方法有自顶向下方法和自底向上方法。其中, 自顶向下方法有残差步数网络^[4]和 ViTPose++^[5], 它们是双阶段检测算法, 第 1 阶段通过目标检测方法识别出目标人体, 并标记人体区域矩形框; 第 2 阶段对已标注的区域进行关键点的检测。自顶向下的方法具有较好的检测性能, 但其检测精度和处理速度受到被检测人员数量的影响, 如 OpenPose^[6], PifPaf^[7], HigherHRNet^[8]和 PETR^[9]等。它们是一种端到端的检测算法, 先提取所有人体关键点, 再通过关键点之间的关联关系判断其属于哪个人体。Pishchulin 等^[10]提出一种基于区域卷积神经网络(regions with CNN features, R-CNN)框架的 DeepCut 检测器, 它与 OpenPose 算法一样, 先检测并标记图像中所有的候选关键点部位, 再利用整数线性规划方法将候选部位组合成最终的人体姿态, 其缺点是计算成本过高。Kreiss 等^[7]提出自底向上方法 PifPaf, 利用部分强度场(part intensity field, PIF)检测人体关键点位置, 之后利用部分关联场(part affinity fields, PAF)表示关节之间的关联信息, 在低分辨率和遮挡场景下优于 OpenPose 方法, 但其检测速度会受到被检测人数的影响; Cheng 等^[8]基于高分辨率网络(high-resolution net, HRNet)提出具有更高分辨率检测性能模型 HigherHRNet, 通过对热图反卷积解决检测过程中图像尺度变化的问题, 其缺点是模型复杂且参数量较多。OpenPose 等自底向上的方法适用于复杂环境下的多人姿态估计, 可以在较快检测速度下同时保证检测精度, 但仍存在模型较大, 参数冗余、检测耗时长等问题。

2 DSC-OpenPose 算法

为了满足复杂环境中跌倒检测实时性与准确性的需求,针对传统 OpenPose 结构中存在的层数过多导致的梯度消失、特征传递以及参数冗余问题,本文以 OpenPose 网络结构为基础,结合 DenseNet^[11]稠密连接思想,提出一种改进 OpenPose 人体姿态估计模型的跌倒检测算法 DSC-OpenPose,其结构如图 1 所示.首先,在确保网络中各个层之间最大程度的信息传递前提下,在每个阶段直接将所有卷积层进行连接,生成稠密连接块 DenseBlock,提高参数共享性,减少参数冗余和模型大小;为了确保前馈性,每层将之前所有层的输

入进行拼接,将输出的特征传递给之后的所有层;为了降低特征图的大小,减少通道数,使模型更加紧凑,进一步压缩模型,在相邻 DenseBlock 之间加入过渡层(Transition 层);为了解决传统 OpenPose 网络在复杂环境下检测精度低的问题,在不过多引入参数的前提下,在不同 DenseBlock 之间添加 CA 机制.通过 5 个阶段迭代微调,生成最终置信度图.

2.1 DenseBlock

DenseBlock 的内部结构如图 2 所示.假设通过卷积网络传递输入为 X_0 ,卷积网络由 n 层构成,各层实现一个非线性变换 $H_n(X)$,图 2 中对应 H_1, H_2, H_3, H_4 ,该变换可以是一系列过程的复合函数,如 ReLU、卷积过程、池化过程等^[12].

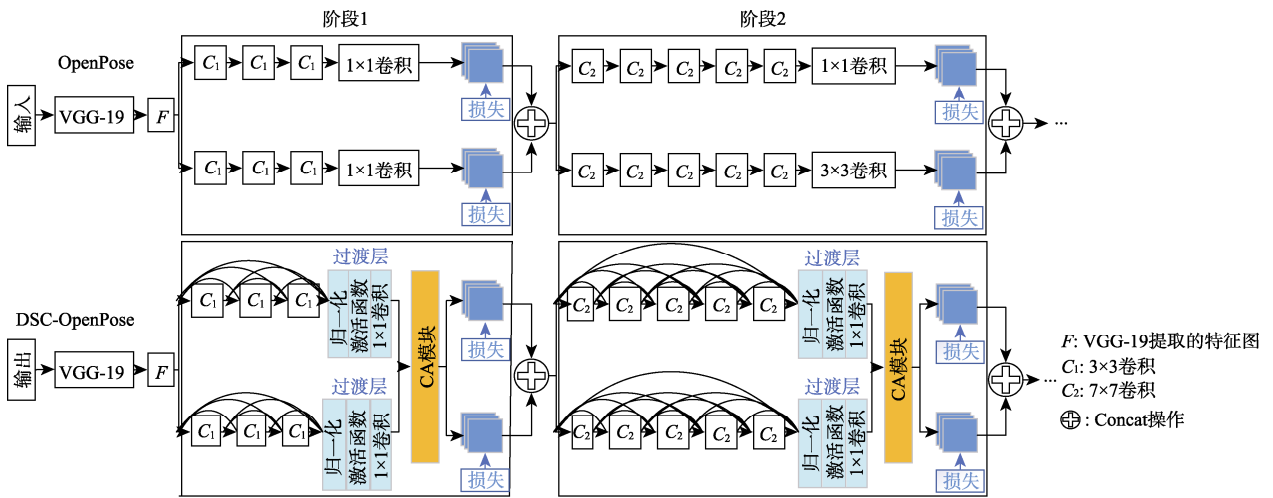


图 1 DSC-OpenPose 算法结构

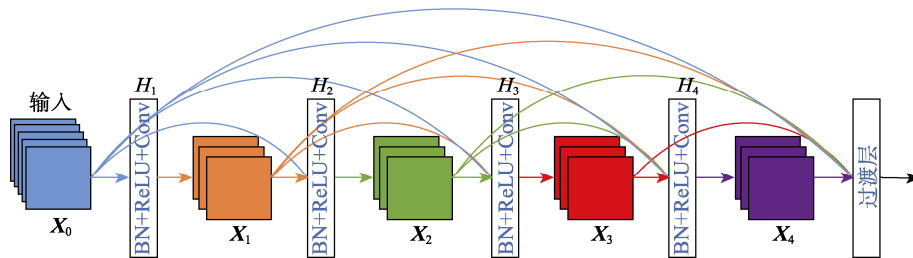


图 2 DenseBlock 内部结构

在 CNN 等传统的前馈网络中,将第 $n-1$ 层输出特征图传递给第 n 层,传递方式为

$$X_n = H_n(X_{n-1}).$$

传统 OpenPose 通过建立第 $n-1$ 层与第 n 层的跳跃连接来解决梯度消失问题,其特征传递方式为

$$X_n = H_n(X_{n-1}) + X_{n-1}.$$

本文提出的 DenseBlock 内部任意 2 个卷积块之间直接相连,其中,每层输入都由之前所有层的

输出通过 Concatenate 并联操作生成,每层学习到的特征图直接传递至之后所有层.特征传递方式为

$$X_n = H_n([X_0, X_1, \dots, X_{n-1}]).$$

其中, $[X_0, X_1, \dots, X_{n-1}]$ 分别表示在第 0 层~第 $n-1$ 层生成的特征图,在图 2 中对应 X_0, X_1, X_2, X_3, X_4 .

在 DenseBlock 中,每层的特征图大小相同,即 C, H 和 W 相同,因此各特征图可以在通道维度上连接.假设输入层的特征图通道数为 k_0 ,输出

通道数随层数 n 的增加不断叠加, 则第 n 层的输入通道数为 $k_0 + (n-1)k$; 其中, k 为网络增长率, 其实际含义为该层新提取出的特征^[13].

2.2 Transition 层

由于 DenseBlock 中的稠密连接特性存在较多的拼接操作, 导致其最终输出的特征图的通道数较多, 因此利用 Transition 层进行降维. Transition 层直接连接 2 个相邻的稠密块, 利用 1×1 的卷积块进行下采样, 在通道维度上降维, 结构为 BN+ReLU+ 1×1 Conv. 若一个 DenseBlock 中有 m 个特征图, 则通过过渡层生成 θm 个输出特征图, θ 表示压缩系数^[14]. 在 DSC-Openpose 中, 设置 $\theta = 0.5$, 表示通过过渡层的特征图数量减半.

2.3 CA 模块

本文采用柔性注意力机制中的 CA 机制. 通道注意力机制只考虑通过建模通道关系衡量每个通道的重要性, 忽略了位置信息, 但位置信息对于生成空间选择性的注意力图十分重要; 此外, 空间注意力机制通过减少输入特征的通道维度, 并利用卷积计算空间注意力增强位置信息, 但卷积操作会增加更多的参数. 在姿态估计过程中, 关键点位置信息对检测结果较为重要, 将位置信息有效地嵌入到 CA 机制中, 在不额外增加参数的同时实现对位置信息的关注, 使模型更准确地定位和识别需检测的目标.

图 3 所示为本文采用的 CA 模块结构图, 其中包

括坐标信息嵌入模块和 CA 生成模块^[15]. $C_1 \times H \times W$ 大小的特征图通过过渡层降维为 $C_2 \times H \times W$ 的特征图, 坐标信息嵌入模块利用全局池化将 X 和 Y 坐标分解为 2 个 1D 特征, 对于 2 个分支分别生成的特征图, 沿着水平和垂直方向分别用核大小为 $(H \times 1)$ 和 $(1 \times W)$ 的平均池化层对每个通道进行编码, 生成 z^h 和 z^w . CA 生成模块首先级联生成的 z^h 和 z^w 这 2 个特征图, 然后利用一个 1×1 的卷积对 F_1 进行变换, 变换公式为

$$f = \delta(F_1([z^h, z^w])), f \in C/r \times (H+W).$$

其中, f 表示空间信息在水平和垂直方向的中间特征图, 将 f 沿空间维度分为 f^h 和 f^w 这 2 个张量. 分别利用的 1×1 卷积 F^h 和 F^w 将 2 个张量的通道维度变换为输入信息的通道维度, 最后利用 Sigmoid 激活函数产生最后的注意力权重, 将所产生的权重与输入特征图相乘实现注意力加权操作^[16]. 公式为

$$\begin{cases} g^h = \sigma(F^h(f^h)) \\ g^w = \sigma(F^w(f^w)) \end{cases}.$$

注意力加权操作不仅捕捉每个通道的重要性, 同时通过空间方向特征的聚合捕获位置信息, 使感兴趣的关键点部位能被准确地捕获. CA 模块的输出将输入到下一个 DenseBlock 中, 公式为

$$y(i, j) = x(i, j) \times g^h(i) \times g^w(j).$$

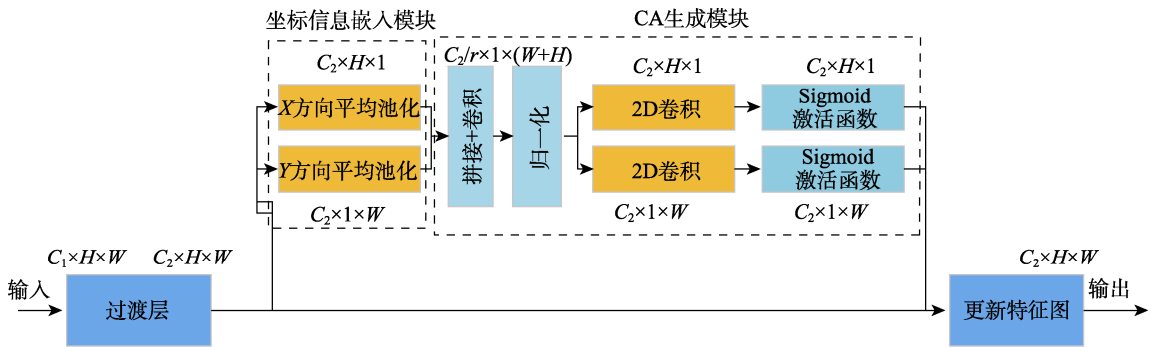


图 3 CA 模块

3 基于姿态估计的跌倒检测算法

3.1 跌倒指标

3.1.1 人体姿态外椭圆

已有的跌倒检测方法中, 大部分研究采用外部矩形来实现人体边界划分, 其在质心提取方面具有较好效果, 但是矩形特征和人体形状还存在一定差

异, 不能更准确地描述体形特征. 本文用一个更接近人体形状的外椭圆表示人体的姿态特征, 用椭圆的中心位置坐标 $O(x, y)$, 椭圆长轴与水平方向的夹角 θ , 椭圆长半轴的长度 e_l 和椭圆短半轴的长度 e_s 这 4 个参数表示.

根据人体体形特征^[17], 选取左眼、右眼、左肩、右肩、左肘、右肘、左踝和右踝这些人体关键点信

息, 利用最小二乘法进行椭圆拟合.

设待拟合的椭圆方程为

$$x^2 + Axy + By^2 + Cx + Dy + E = 0.$$

设 $P_i(x_i, y_i) (i=1, 2, \dots, N)$ 为提取到的 $(N \geq 5)$ 个关键点, 根据最小二乘法的原理, 所拟合的目标函数为

$$F(A, B, C, D, E) = \sum_{i=1}^N (x_i^2 + Ax_i y_i + AB y_i^2 + Cx_i + Dy_i + E)^2.$$

欲使 F 最小, 需要满足

$$\frac{\partial F}{\partial A} = \frac{\partial F}{\partial B} = \frac{\partial F}{\partial C} = \frac{\partial F}{\partial D} = \frac{\partial F}{\partial E} = 0.$$

由此可得方程

$$\begin{bmatrix} \sum x_i^2 y_i^2 & \sum x_i y_i^3 & \sum x_i^2 y_i & \sum x_i y_i^2 & \sum x_i y_i \\ \sum x_i y_i^3 & \sum y_i^4 & \sum x_i y_i^2 & \sum y_i^3 & \sum y_i^2 \\ \sum x_i^2 y_i & \sum x_i y_i^2 & \sum x_i^2 & \sum x_i y_i & \sum x_i \\ \sum x_i y_i^2 & \sum y_i^3 & \sum x_i y_i & \sum y_i^2 & \sum y_i \\ \sum x_i y_i & \sum y_i^2 & \sum x_i & \sum y_i & N \end{bmatrix} \begin{bmatrix} A \\ B \\ C \\ D \\ E \end{bmatrix} = - \begin{bmatrix} \sum x_i^3 y_i & \sum x_i^2 y_i^2 & \sum x_i^3 & \sum x_i^2 y_i & \sum x_i^2 \end{bmatrix}^T \quad (1)$$

求解式(1), 可得 A, B, C, D, E 的值. 根据椭圆的几何知识, 求得椭圆的 5 个参数为

$$\begin{cases} x = \frac{2BC - AD}{A^2 - 4B} \\ y = \frac{2D - AC}{A^2 - 4B} \\ e_1 = \frac{\sqrt{2(ACD - BC^2 - D^2 + 4BE - A^2E)}}{\sqrt{(A^2 - 4B)(B + 1 - \sqrt{A^2 + (1 - B)^2})}} \\ e_s = \frac{\sqrt{2(ACD - BC^2 - D^2 + 4BE - A^2E)}}{\sqrt{(A^2 - 4B)(B + 1 + \sqrt{A^2 + (1 - B)^2})}} \\ \theta = \arctan \sqrt{\frac{e_1^2 - e_s^2 B}{e_1^2 B - e_s^2}} \end{cases}$$

站立和跌倒情形下, 拟合后的椭圆及其参数如图 4 所示.

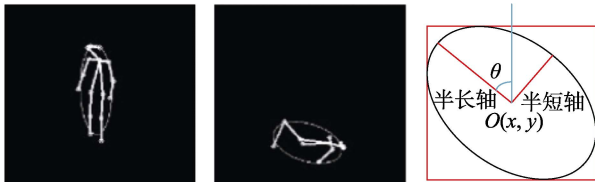


图 4 拟合椭圆及其参数

3.1.2 头部高度

当人体行走或站立时, 头部离地的高度约等

于人体高度, 当人体跌倒时, 头部接近地面, 其距离地面的高度值最小, 故头部距地面的高度变化可作为跌倒特征. 地面参考线为人体外矩形框下水平线 L , 2 点分别为 $R_1(x_1, y_1), R_2(x_2, y_2)$, 公式为

$$Ax + By + C = 0.$$

其中, $A = y_2 - y_1, B = x_2 - x_1, C = x_2 y_1 - x_1 y_2$.

为了保证检测的精确性, 取鼻子、左耳、左眼、右眼和右耳 5 个关键点的坐标, 计算其平均值, 记作头部中心点的坐标值, 公式为

$$\begin{cases} x_m = \frac{1}{5} \sum_{i=0}^4 x_i \\ y_m = \frac{1}{5} \sum_{i=0}^4 y_i \end{cases}$$

则头部距地面高度 h_1 可定义为

$$h_1 = \frac{|Ax_m + by_m + C|}{\sqrt{A^2 + B^2}}.$$

3.1.3 下肢高度

从站立状态到跌倒状态的过程中, 人体下肢高度会出现变化. 为了避免跌倒与蹲下、坐下等状态特征的重复性, 选取左膝(left knee, LK)、右膝(right knee, RK)、左踝(left ankle, LA)和右踝(right ankle, RA)关键点之间的高度特征作为下肢高度, 记作 h_2 . 在人的跌倒过程中, 下肢高度 h_2 减小, 可将其作为判断跌倒的特征之一. h_2 的计算公式为

$$h_2 = \frac{1}{2} (|y_{LA} - y_{LK}| + |y_{RA} - y_{RK}|).$$

其中, y_{LA} 表示 LA 高度; y_{LK} 表示 LK 高度, y_{RA} 表示 RA 高度, y_{RK} 表示 RK 高度.

3.2 MLP 分类器

基于以上跌倒指标, 对每个视频序列检测运动对象, 使用特征向量 F 的 7 个唯一特征识别每帧中对象的跌倒行为, 公式为

$$F = [h_1, h_2, x, y, \theta, a, b].$$

为了对跌倒行为和非跌倒行为进行分类, 将特征向量输入到 MLP 神经网络, 跌倒检测方法的结构如图 5 所示. MLP 网络分为输入层、隐藏层和输出层 3 层, 其中, 隐藏层利用缩放共轭梯度算法, 通过减少迭代次数压缩输入的动态范围, 使网络可以更好地执行, 该层的输出被传递到输出层, 输出层产生决定性的输出来识别跌倒行为.

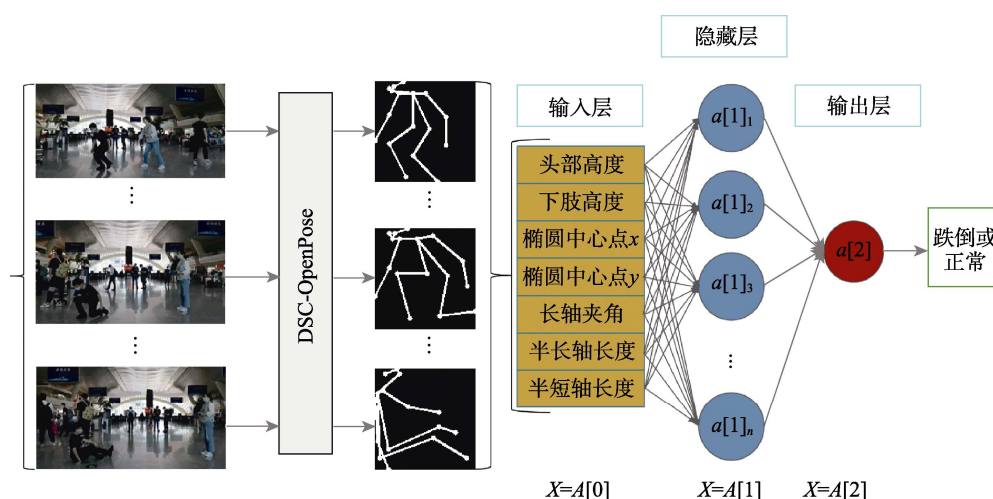


图 5 跌倒检测方法结构

4 实验与结果分析

4.1 数据集和实验配置

本文实验采用 COCO 2017 中用于关键点检测的数据集和 AI Challenger 数据集。在 COCO 2017 数据集中, 训练集样本数为 120 148 张, 总计 100 万个关键点; 测试集样本数为 39 635 张, 总计 40 万个关键点。在 AI Challenger 数据集中, 训练集样本数约为 20 万张, 总计 230 万个关键点; 测试集样本数为 4 万张, 总计 30 万个关键点。标签数据以 JSON 格式存取。单个人体骨骼关键点的标注如表 1 所示。

表 1 人体骨骼关键点详细介绍

序号	名称	序号	名称
0	鼻子	9	左腕
1	左眼	10	右腕
2	右眼	11	左髋
3	左耳	12	右髋
4	右耳	13	左膝
5	左肩	14	右膝
6	右肩	15	左踝
7	左肘	16	右踝
8	右肘		

跌倒检测实验使用自建的 RF(real fall)行为数据集, 数据来源为互联网图片, 以及道路监控、铁路监控的真实图片, 场景为火车站进出站口、过街天桥等公共场所。RF 数据集包括 1 400 多个视频和图片, 其中, 跌倒行为和正常行为比例为 8 : 2, 将数据集按 9 : 1 的比例划分为训练集和测试集。RF 数据集中, 视频图片数量、行为标签具体数量、不同场景数量和人员数量如图 6 所示。

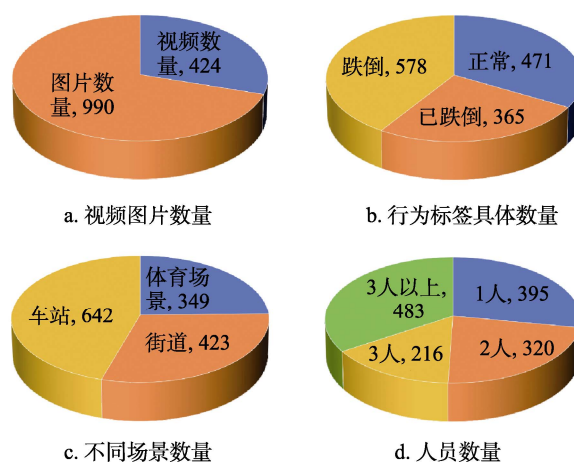


图 6 RF 数据集数量比例

针对数据集中未包含时序信息的图片数据, 部分场景图片如图 7 所示, 首先调整其尺寸, 与视频尺寸(256×256)匹配, 并将其扩展为平均帧数为 24 帧的视频对象, 在模型的训练过程中与数据集视频同时输入到模型进行处理。实验环境为 Ubuntu22.04LTS 操作系统, PyTorch 深度学习框架, Python 开发语言; 使用 Intel(R) Xeon(R) Silver 4210 CPU@2.20 GHz CPU, NVIDIA TESLA T4(16.0 GB) GPU 的计算机。

4.2 评价指标

因为人体关键点检测算法无法完全将预测的结果与真实值进行一一比较, 无法确定当前的预测结果是否出现了漏检或误检, 所以采用关键点检测评价指标(object keypoint similarity, OKS)计算预测值与真实关键点的相似程度^[18], 计算公式为

$$p_{OKS} = \frac{\sum_i \exp\left\{-d_{pi}^2 / 2S_p^2\sigma_i^2\right\} \delta(v_{pi}=1)}{\sum_i \delta(v_{pi}=1)}.$$



图 7 RF 数据集部分场景图

其中, p 表示 Groud Truth 中人体的 id; i 表示关键点的 id; d_{pi} 表示 Groud Truth 中人体关键点相对于预测的人体关键点的欧几里得距离; S_p 表示当前人体的尺度因子, 其值等于此人体在 Groud Truth 中所占面积的平方根, 即 $\sqrt{(x_2 - x_1)(y_2 - y_1)}$; σ_i 表示第 i 个关键点的归一化因子, 对数据集中所有 Groud Truth 求标准差所得, 反映关键点标注时的标准差, σ 越大, 表示该关键点越难标注; v_{pi} 表示第 p 个人体的第 i 个关键点是否可见; δ 表示将可见点选出并进行计算的函数。

基于关键点检测的特点, 利用 p_{OKS} 衡量整个算法的指标平均精确度(average precision, AP), 其计算方式为给定常数 t , 若当前的 $p_{OKS} > t$, 表明关键点被正确检测; 若 $p_{OKS} < t$, 表明未被成功检测或发生误检漏检。故对于所有 p_{OKS} 统计大于 t 的数量, 并计算其占有所有 p_{OKS} 的比值, 即为 AP 值, 公式为

$$AP^t = \sum_p \delta(p_{OKS} > t) / \sum_p 1.$$

利用 p_{OKS} 为 0.50, 0.75 和 0.90 (AP^{50} , AP^{75} 和 AP^{90}) 表示关键点不同要求的匹配精度^[19], 并利用其平均值(mean average precision, mAP)表示所有关键点的平均精确度。为了衡量算法的规模, 使用帧速

计算每秒检测的图片帧数。

4.3 模型对比实验

使用本文模型进行实验验证。训练过程中, 设置 400 个迭代周期, 初始学习率为 0.01, 采用阶梯衰减的方法对学习率进行处理^[20], 当迭代周期达到 100 时, 学习率下降至此时的 1/10; 当迭代周期达到 250 时, 学习率再次下降至当前的 1/10。反向传播过程采用梯度下降法, 其中设动量参数为 0.9, batch size 为 64。

将本文提出的姿态估计模型与自底向上的 OpenPose^[6], HigherHRNet-w32^[8], Lightweight OpenPose^[21]和 HRNet^[22]和模型在 COCO 关键点数据集(简称 COCO 数据集)和 AI Challenger 数据集上分别进行实验, 实验环境和配置相同, 5 种模型在 COCO 和 AI Challenger 数据集上的实验结果如表 2 所示。

为了验证稠密连接 Dense 机制对模型产生的影响, 对本文模型在参数量、运算速度及帧速指标上与对比算法进行比较, 结果如表 3 所示。

对比实验结果表明: (1) 在 COCO 数据集中, 与其他模型相比, 本文模型在 mAP 和 AP^{75} 上均有提升, 且在提升精度的同时并未增加检测周期, 检测时间明显优于 HigherHRNet-w32, OpenPose 和

表 2 5 种模型在 COCO 和 AI Challenger 数据集上的结果对比

%

模型	COCO				AI Challenger			
	mAP	AP^{50}	AP^{75}	AP^{90}	mAP	AP^{50}	AP^{75}	AP^{90}
OpenPose ^[6]	62.1	69.9	59.5	48.6	63.8	70.7	61.3	48.3
Lightweight OpenPose ^[21]	59.1	67.2	60.1	48.2	61.9	67.7	60.5	47.0
HRNet ^[22]	62.9	69.4	60.6	47.7	64.3	71.6	61.7	49.2
HigherHRNet-w32 ^[8]	63.6	72.9	60.9	47.9	65.1	71.9	61.4	48.9
本文	63.9	72.5	61.5	47.9	64.8	72.3	62.4	49.1

注. 粗体表示最优值。

表 3 5 种模型规模对比

模型	COCO			AI Challenger		
	参数量	GFLOPS	帧速/(帧·s ⁻¹)	参数量	GFLOPS	帧速/(帧·s ⁻¹)
OpenPose ^[6]	342M	33.0	12.1	364M	33.7	11.1
Lightweight OpenPose ^[21]	141M	8.3	23.2	144M	8.3	22.9
HRNet ^[22]	276M	21.5	14.4	290M	23.0	13.2
HigherHRNet-w32 ^[8]	286M	23.4	13.6	282M	22.8	13.5
本文	194M	22.3	20.6	199M	22.8	19.8

注. 粗体表示最优值.

HRNet 模型; 与具有近乎相同帧速的轻量化网络 Lightweight OpenPose 相比, 本文模型平均精度提升 4.8 个百分点, 同时在各阈值条件下均有较好的表现. (2) 在 AI Challenger 数据集中, 与 OpenPose 和 Lightweight OpenPose 模型相比, 本文模型在 mAP 上有一定提升; 与具有近似检测精度的 HigherHRNet-w32 模型相比, 本文模型参数量较少且具有较快检测速度. (3) 引入稠密连接机制和 CA 模块后, DSC-OpenPose 在精度和模型参数量之间取得较好的平衡效果.

从表 3 可以看出: (1) 与 OpenPose 和 HRNet 模型相比, 本文模型的参数量分别减少 43% 和 30%, 主要原因是采用 Dense 稠密连接强化参数共享性, 减少了参数存储量, 节省参数空间. (2) 在运算速度

方面, 本文模型通过反复的拼接过程, 将之前层的输出与当前层的输出拼接在一起传递给下一层, 每次的拼接过程都需要开辟新的内存保存拼接后的特征^[20], 会占用较多显存, 而内存的访问次数多且访问速度慢; 同时, 由于特征图和卷积过程计算量较大, 导致本文模型的运算速度无较大提升.

根据数据集特点, 选取鼻子(nose, N), 左肘(left elbow, LE), 右肘(right elbow, RE), 左膝(left knee, LK), 右膝(right knee, RK), 左踝(left ankle, LA)和右踝(right ankle, RA)关键点精度进行模型对比实验, 结果如表 4 所示. 可以看出, 在 COCO 和 AI Challenger 数据集中, 与其他模型相比, 本文模型在 LE, RK 关键点的精度有较大提升, 在其他关键点上也有较好检测效果.

表 4 5 种模型在 COCO 和 AI Challenger 数据集上关键点精度的结果对比

%

模型	COCO							AI Challenger						
	N	LE	RE	LK	RK	LA	RA	N	LE	RE	LK	RK	LA	RA
OpenPose ^[6]	81.6	77.6	78.2	74.4	71.8	64.6	65.4	81.7	78.5	77.9	74.6	72.3	65.4	64.9
Lightweight OpenPose ^[21]	80.8	77.0	77.1	73.1	70.6	62.2	62.9	79.8	77.6	76.8	73.9	72.6	63.2	62.2
HRNet ^[22]	82.0	78.2	78.0	76.1	72.2	64.7	66.3	80.2	78.9	77.1	75.0	73.1	65.1	64.4
HigherHRNet-w32 ^[8]	82.5	78.6	77.8	76.0	73.1	65.2	66.0	82.1	78.7	78.0	75.1	73.8	65.7	65.0
本文	82.4	79.6	77.9	77.0	74.5	65.5	66.5	81.7	80.6	79.8	74.5	75.1	66.0	64.5

注. 粗体表示最优值.

进一步, 验证 CA 机制模块对检测性能的影响. 本文在 COCO 和 AI Challenger 数据集上对 CA 模块进行消融实验, 结果如表 5 所示. 对 CA 模块消融实验的精度结果如表 6 所示.

本文采用原始 OpenPose 模型作为基线, 从表 5 可以看出, 添加 DenseBlock 后, 模型较基线在参数量方面有明显减少; 再添加 Transition 层后, 参数量再次下降, Transition 层中额外的 1×1 的下采样卷积层在通道维度上降维, 通过 Transition 层传递到下一个 DenseBlock 的通道数量减半, 大幅减少了参数量; 再添加 CA 模块后, mAP 指标优于基线及其他模型, 验证了 CA 模块在特征提取中的

表 5 COCO 和 AI Challenger 数据集上消融实验结果对比

模型	COCO		AI Challenger	
	mAP/%	参数量	mAP/%	参数量
基线	62.1	342M	63.8	364M
+DenseBlock	61.9	267M	64.0	295M
+DenseBlock+Transition	61.9	182M	63.9	188M
+CA	63.9	361M	64.8	379M
本文	63.9	194M	64.8	199M

注. 粗体表示最优值.

有效性. 实验结果表明, 本文模型具有更好的检测性能.

从表 6 可以看出,在添加注意力模块后,mAP 和部分关键点的精度均有一定提升;在 COCO 数据集中,mAP 提升 1.6 个百分点,在 AI Challenger 数据集中,mAP 提升 1.1 个百分点,在部分关键点精度上也具有较好表现.为了更直观地体现注意力模块的有效性,分别利用有/无注意力机制的姿态估计模型进行骨架提取,效果图如图 8 所示.

从图 8 可以看出,添加注意力模块后,提取到更为精确的姿态信息,生成了更准确的骨架图,表

明本文模型更聚焦于识别关键点的关键通道和位置,赋予这些通道和位置更高的权重,进而提高关键点的检测精度.通过有/无注意力机制模块的姿态估计实验,验证了注意力模块的有效性.

在跌倒检测部分,首先使用所提姿态估计模型对 RF 数据集进行关键点提取和骨架图生成,然后利用关键点和骨架信息生成跌倒检测特征,最终利用基于 MLP 分类器的跌倒检测模型进行实验.

表 6 注意力机制模块消融实验的结果对比

模型	COCO								AI Challenger								%
	mAP	N	LE	RE	LK	RK	LA	RA	mAP	N	LE	RE	LK	RK	LA	RA	
本文(有 CA 模块)	62.3	82.0	78.4	78.3	75.4	72.3	64.5	65.9	63.7	82.1	78.4	78.0	75.6	72.9	65.6	64.3	
本文(无 CA 模块)	63.9	82.4	79.6	77.9	77.0	74.5	65.5	66.5	64.8	81.7	80.6	79.8	74.5	75.1	66.0	63.5	

注.粗体表示最优值.



a. 无注意力机制



b. 有注意力机制

图 8 姿态估计对比图

使用十折交叉验证方法,将输入数据分为 10 组,其中,9 组用于进行训练,1 组用于进行验证,对 10 组的所有可能重复训练和测试.选用均方误差函数(mean-square error, MSE)作为损失函数,将每次迭代产生的误差平方的平均值最小化,并更新网络权重和偏差,仅使用此误差函数来确定训练的性能目标.为了验证其有效性,在相同实验环境下与其他基于姿态估计设计的跌倒检测模型^[1-2,23]以及基于光流的检测模型^[24]进行对比实验,综合考查准确率(accuracy, A)、精度(precision, P)、召回率(recall, R), F_1 分数(F_1 score)、帧速、模型总参数量指标.它们的计算公式分别为

$$A = \frac{TP+TN}{TP+FN+FP+TN},$$

$$P = \frac{TP}{TP+FP},$$

$$R = \frac{TP}{TP+FN},$$

$$F_1 = 2 \times \frac{P \times R}{P + R}.$$

其中, TP 表示跌倒类别检测为跌倒类别的数量; FP 表示非跌倒类别检测为跌倒类别的数量; FN 表示跌倒类别检测为非跌倒类别的数量; TN 表示非跌倒类别检测为非跌倒类别的数量.

5 种模型在 RF 数据集上的消融实验结果如表 7

表 7 5 种模型在 RF 数据集上的结果对比

模型	A/%	P/%	R/%	F_1 /%	帧速/(帧·s ⁻¹)	参数量
PoseNet+GRU ^[1]	97.2	96.1	97.6	96.8	17.6	302M
OpenPose+Conv+LSTM ^[2]	96.0	94.5	97.4	95.9	18.8	346M
3D LPN+CNN ^[23]	97.6	96.2	98.3	97.2	15.4	277M
VGG+Optical Flow ^[24]	97.4	96.0	97.4	96.6	19.1	211M
本文	98.2	96.6	98.0	97.2	20.2	219M

注.粗体表示最优值.

所示. 同时, 与单纯的位姿关键点作为分类器输入的方法以及基于外矩形的方法进行实验对比, 结果如表 8 所示.

表 8 在 RF 数据集上消融实验结果 %

模型	A	P	R	F ₁
+CNN	94.1	93.3	90.8	92.0
+MLP	95.0	94.2	92.4	93.2
+SVM	92.3	93.4	93.1	93.2
+外矩形	96.5	95.0	94.7	94.8
本文	98.2	96.6	98.0	97.2

注: 粗体表示最优值.

实验结果表明, 本文模型的准确率达到 98.2%, 精度达到 96.6%, 召回率达到 98%, 说明大部分跌

倒行为会被准确地识别; 同时与其他模型相比, 本文模型在准确率和精度上表现出更好的检测效果, 模型的总参数量比大部分对比模型更低, 同时检测速度更优, 更有利于实时检测的需求. 从表 8 可以看出, 与将姿态估计提取所有关键点直接输入到分类器(CNN, MLP, SVM)的算法以及基于人体外矩形的算法相比, 本文模型在各评价指标均具有较好表现, 跌倒检测精度更高, 说明该算法过滤掉了大部分与跌倒检测无关的关键点信息, 验证了其有效性.

本文算法的检测效果如图 9 所示, 其中, 黄色线表示提取到的骨架图, 绿色框表示正常状态, 红色框表示跌倒异常状态. 实验结果表明, 本文算法能较准确地识别跌倒行为.

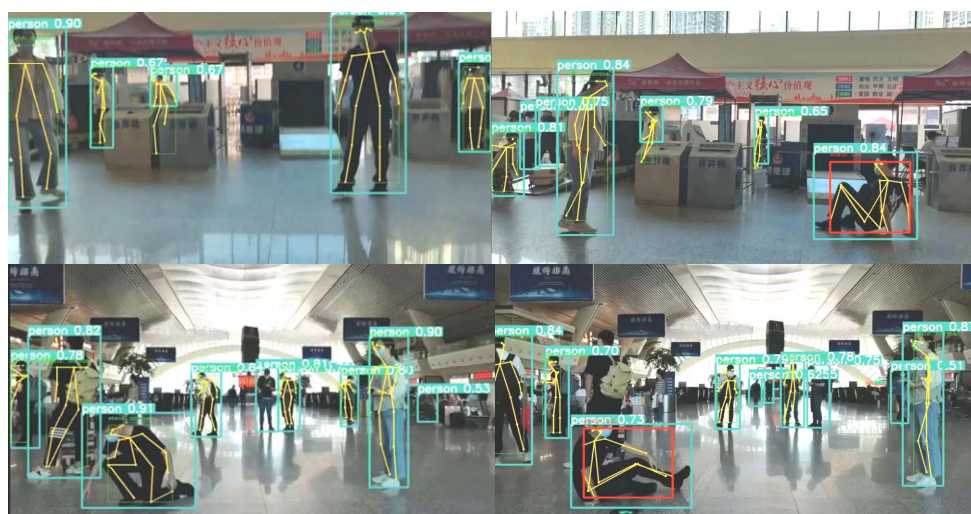


图 9 本文算法检测效果

5 结 语

本文提出融合注意力机制的 OpenPose 人体跌倒检测算法 DSC-OpenPose, 主要有 3 个方面的工作: (1) 提出利用串联的 DenseBlock 代替原始的 CNN 架构减少模型的参数量, 参数量较 OpenPose 算法减少 43%, 实现模型轻量化; (2) 引入 CA 机制捕获通道和位置信息精确定位关键点, 获得更准确的姿态估计结果; (3) 利用所得姿态估计结果设计 MLP 分类器识别跌倒行为, 实现了较高的检测精度和较快的检测速度. 在姿态估计部分, DSC-OpenPose 算法在 COCO 数据集上具有较好的检测精度; 在跌倒检测部分, 该算法在 RF 数据集上准确率达到 98.2%, 精度达到 96.6%, 具有较高的精确性和鲁棒性. 但是, 在较大规模数据集下, DSC-OpenPose 算法使用传统特征提取模型的效果较差, 同时在更复杂环境下, 对未提取到的关键点以及

错误识别的关键点处理能力不足. 未来工作的重点是降低显存占用、实现在遮挡复杂环境的准确估计.

参考文献(References):

- [1] Kang Y, Kang H, Kim J. Fall detection method based on pose estimation using GRU[M] // Lee R, Kim J B. Software Engineering, Artificial Intelligence, Networking and Parallel/ Distributed Computing. Heidelberg: Springer, 2021: 169-179
- [2] Chen W M, Jiang Z J, Guo H L, et al. Fall detection based on key points of human-skeleton using openpose[J]. Symmetry, 2020, 12(5): Article No.744
- [3] Nogas J, Khan S S, Mihailidis A. DeepFall: non-invasive fall detection with deep spatio-temporal convolutional autoencoders[J]. Journal of Healthcare Informatics Research, 2020, 4(1): 50-70
- [4] Cai Y H, Wang Z C, Luo Z X, et al. Learning delicate local representations for multi-person pose estimation[C] // Proceeding of the 16th European Conference on Computer Vision. Heidel-

- berg: Springer, 2020: 455-472
- [5] Xu Y F, Zhang J, Zhang Q M, *et al.* ViTPose++: vision transformer for generic body pose estimation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(2): 1212-1230
 - [6] Cao Z, Hidalgo G, Simon T, *et al.* OpenPose: realtime multi-person 2D pose estimation using part affinity fields[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, 43(1): 172-186
 - [7] Kreiss S, Bertoni L, Alahi A. PifPaf: composite fields for human pose estimation[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2019: 11969-11978
 - [8] Cheng B W, Xiao B, Wang J D, *et al.* HigherHRNet: scale-aware representation learning for bottom-up human pose estimation[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2020: 5386-5395
 - [9] Shi D H, Wei X, Li L Q, *et al.* End-to-end multi-person pose estimation with transformers[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2022: 11059-11068
 - [10] Pishchulin L, Insafutdinov E, Tang S Y, *et al.* DeepCut: joint subset partition and labeling for multi person pose estimation[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2016: 4929-4937
 - [11] Huang G, Liu Z, Van Der Maaten L, *et al.* Densely connected convolutional networks[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2017: 2261-2269
 - [12] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2018: 7132-7141
 - [13] Woo S, Park J, Lee J Y, *et al.* CBAM: convolutional block attention module[C] // *Proceedings of the 15th European Conference on Computer Vision*. Heidelberg: Springer, 2018: 3-19
 - [14] Hou Q B, Zhou D Q, Feng J S. Coordinate attention for efficient mobile network design[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2021: 13708-13717
 - [15] Zhang F, Zhu X T, Dai H B, *et al.* Distribution-aware coordinate representation for human pose estimation[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2020: 7093-7102
 - [16] Xue N, Wu T F, Xia G S, *et al.* Learning local-global contextual adaptation for multi-person pose estimation[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2022: 13055-13064
 - [17] Ma Wanyi, Zhang Deping. Human pose estimation based on multi-spectral attention and high resolution network[J]. *Journal of Computer-Aided Design & Computer Graphics*, 2022, 34(8): 1283-1292(in Chinese)
(马婉宜, 张德平. 基于多谱注意力高分辨率网络的人体姿态估计[J]. *计算机辅助设计与图形学学报*, 2022, 34(8): 1283-1292)
 - [18] Chen Lufei, Zhang Yong, Tang Yongzheng, *et al.* FP-Net: frontal pose estimation based on single human body image with arbitrary view[J]. *Journal of Computer-Aided Design & Computer Graphics*, 2022, 34(10): 1604-1612(in Chinese)
(陈路飞, 张勇, 唐永正, 等. FP-Net: 基于任意角度单幅人体图像的正面姿态估计[J]. *计算机辅助设计与图形学学报*, 2022, 34(10): 1604-1612)
 - [19] Li Xuanye, Hao Xingwei, Jia Jingong, *et al.* Human action recognition method based on multi-attention mechanism and spatiotemporal graph convolution networks[J]. *Journal of Computer-Aided Design & Computer Graphics*, 2021, 33(7): 1055-1063(in Chinese)
(李炫烨, 郝兴伟, 贾金公, 等. 结合多注意力机制与时空图卷积网络的人体动作识别方法[J]. *计算机辅助设计与图形学学报*, 2021, 33(7): 1055-1063)
 - [20] Zhao Baoqi, Yu Fei, Sun Junmei, *et al.* Glandular cell segmentation method combined with dense connective blocks and self-attention mechanism[J]. *Journal of Computer-Aided Design & Computer Graphics*, 2021, 33(7): 991-999(in Chinese)
(赵宝奇, 尉飞, 孙军梅, 等. 结合密集连接块和自注意力机制的腺体细胞分割方法[J]. *计算机辅助设计与图形学学报*, 2021, 33(7): 991-999)
 - [21] Osokin D. Real-time 2D multi-person pose estimation on CPU: Lightweight openpose[OL]. [2023-01-17]. <https://arxiv.org/abs/1811.12004>
 - [22] Sun K, Xiao B, Liu D, *et al.* Deep high-resolution representation learning for human pose estimation[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2019: 5686-5696
 - [23] Chen Z W, Wang Y Y, Yang W K. Video based fall detection using human poses[C] // *Proceedings of the 9th CCF Conference on Big Data*. Heidelberg: Springer, 2022: 283-296
 - [24] Menacho C, Ordoñez J. Fall detection based on CNN models implemented on a mobile robot[C] // *Proceedings of the 17th International Conference on Ubiquitous Robots*. Los Alamitos: IEEE Computer Society Press, 2020: 284-289