面向复杂工业场景的人体姿态快速估计方法

张泽辉[1†]，吴富龙[1]，李帆雅[1]，徐晓滨[1]，章振杰[1]，邵海滨[2]

(1. 杭州电子科技大学 自动化学院 中国-奥地利人工智能与先进制造“一带一路”联合实验室，杭州 310018；
2. 上海交通大学 电子信息与电气工程学院，上海 200240)

摘　要: 近年来, 以人体姿态信息为基础的行为识别技术正逐步应用到工人行为安全检测中. 然而, 在复杂工业场景下, 遮挡和算力受限等问题使得现有的基于计算机视觉的人体姿态估计方法难以同时满足高精度和低复杂度的要求. 因此, 本文结合量化自编码器和轻量化的 ResNeSt 网络, 提出了一种面向复杂工业场景的人体姿态快速估计方法. 特别地, 本文提出了一种循环权重迁移训练方法, 通过在不同尺寸的骨干网络模型之间迁移权重参数, 以保证姿态估计的精度. 实验结果表明, 所提方法能够在复杂工业场景中准确地估计出人体姿态, 相较于原始方法, 模型计算量减少了 4 倍, 为工业领域的实时姿态估计提供了一种高效、低资源消耗的解决方法.

关键词: 人体姿态估计；复杂工业场景；计算机视觉；量化自编码器；迁移学习

中图分类号: TP391.41　　**文献标志码:** A

DOI: 10.13195/j.kzyjc.2025.0884

引用格式: 张泽辉, 吴富龙, 李帆雅, 等. 面向复杂工业场景的人体姿态快速估计方法 [J]. 控制与决策, xxxx, x(x): xxxx-xxxx.

Fast human pose estimation method for complex industrial environments

ZHANG Ze-hui[1†]，*WU Fu-long2*[1]，*LI Fan-ya*[1]，*XU Xiao-bin*[1]，*ZHANG Zhen-jie*[1]，*SHAO Hai-bin*[2]

(1. China-Austria Belt and Road Joint Laboratory on Artificial Intelligence and Advanced Manufacturing，Hangzhou Dianzi University，Hangzhou 310018，China；2. School of Electronic Information and Electrical Engineering，Shanghai Jiao Tong University，Shanghai 200240，China)

Abstract: In recent years, behavior recognition technology based on human pose information has been increasingly applied to worker safety monitoring. However, in complex industrial settings, challenges such as occlusion and limited compu-tational resources have hindered existing computer vision-based human pose estimation methods from simultaneously achieving high accuracy and low computational complexity. Thus, this paper proposes a novel approach for rapid human pose estimation in complex industrial scenarios by integrating a quantization autoencoder with a lightweight ResNeSt network. Furthermore, a cyclic weight transfer training method is proposed, which enhances estimation ac-curacy by transferring weight parameters across backbone networks of varying sizes. The experimental results demonstrate that the proposed method achieves accurate human pose estimation in complex industrial environments, reducing the computational cost of the original model by a factor of four, thereby providing an efficient and re-source-effective solution for real-time pose estimation in industrial applications.

Keywords: human pose estimation；complex industrial environments；computer vision；quantization autoencoder；transfer learning

0　引　言

在工业 5.0 背景下, 企业应将“以人为本”作为核心发展原则, 将工人福祉置于生产流程的优先位置. 其中, 保障工人安全作为最基本的需求, 不仅是企业社会责任的底线, 更是实现可持续发展的关键. 根据统计数据显示, 2023 年化工行业发生的较大及重大事故中, 83% 的事故直接由人员不安全行为引发[1]. 人员行为安全在工业安全生产中扮演着至关重

收稿日期: 2025-08-27；**录用日期:** 2026-01-14.
基金项目: 浙江省尖兵领雁科技项目 (2025C04005), 衢州市科技计划项目 (2024K154), 国家自然科学基金项目 (52401376), 浙江省自然科学基金资助项目 (LTGG24F030004), 国家水运安全工程技术研究中心开放基金资助项目 (A202501).
责任编委: 程龙.
†通信作者. E-mail: zhangtianxia918@163.com.

要的角色, 尤其是在复杂的工业场景中 (高危化学品操作、受限空间作业等), 作业人员的任何一项违规操作或疏忽大意的行为, 都可能引发严重的安全事故, 造成无法挽回的损失. 传统的人工监控方式存在效率低、易疏漏等缺点, 难以满足工业 5.0 背景下安全生产的要求[2-4]. 因此, 许多企业开始引入智能化模型和现代监控设备, 建立以计算机视觉为核心的智能人员行为监控系统[5-7].

近年来, 行为识别技术开始初步应用于工业安全领域, 通过实时监控与分析作业人员行为, 能够及时发现不安全操作或违规行为, 减少人为失误带来的风险[8-10]. 人体姿态估计是行为识别的基础, 其通过提取人体关键点信息, 实时捕捉并分析作业人员的动作轨迹和姿态特征, 从而识别作业过程中存在的不规范动作或潜在危险行为. 尤其是在复杂工业场景中, 如何快速、准确地获取作业人员的姿态数据, 是实现对工人行为进行实时分析、风险评估与智能决策的前提条件. 目前主流的人体姿态估计方法主要有基于穿戴传感器的人体姿态估计方法和基于计算机视觉的人体姿态估计方法.

(1) 基于穿戴传感器的人体姿态估计方法

基于穿戴传感器的人体姿态估计方法通过加速度计、陀螺仪、压力传感器等可穿戴设备采集人体运动数据, 并对其进行处理与分析以估计人体姿态[11,12]. 例如, Digo 等人[13] 利用多组三轴惯性测量单元实现了上肢运动学的实时估计, 可用于识别典型工业手势; Li 等人[14] 基于可穿戴惯性传感器网络, 采用扩展卡尔曼滤波融合多轴传感器数据, 并结合零速度更新与航向校准策略, 有效降低了姿态估计误差. 然而, 此类方法在工业场景中仍存在一定局限性. 一方面, 传感器的安装受空间条件和人体活动范围限制, 难以全面捕捉复杂动作细节; 另一方面, 环境振动、电磁干扰等因素易导致数据不稳定. 此外, 复杂工业作业中的行为识别往往需要结合环境信息与操作语境, 而单纯依赖穿戴传感器难以满足这一需求. 因此, 基于传感器的人体姿态估计方法在复杂工业环境中的应用仍面临挑战[15].

(2) 基于计算机视觉的人体姿态估计方法

近年来, 基于深度学习的人体姿态估计方法在精度和鲁棒性方面取得了显著进展, 已成为行为识别与安全监测系统的重要基础[17,18]. 随着卷积神经网络和 Transformer 结构在计算机视觉领域的广泛应用, 研究者开始探索其在人体姿态建模中的优势. 胡楠等人[16] 提出了一种融合卷积神经网络与 Transformer 的多假设交互三维人体姿态估计模型, 通过引入混合注意力机制和多假设特征交互, 有效增强了对局部关节运动与全局姿态关系的建模能力. 在姿态信息的高层语义应用方面, 宋忱等人[17] 基于骨架序列构建了多语义动态图卷积网络, 实现了对人体行为时序特征的有效建模, 验证了姿态数据在动作理解与行为识别任务中的重要作用. 进一步地, 陈博等人[18] 将姿态估计结果与多模态信息相结合, 用于关节角度预测与运动分析, 体现了人体姿态在实际工程与康复应用中的实用价值.

尽管上述研究在姿态估计和行为识别方面取得了一定成果, 但在复杂工业环境中, 仍面临遮挡严重、光照变化频繁以及边缘计算资源受限等问题, 现有方法难以同时兼顾高精度与低延迟. 需要指出的是, 人体姿态估计并非行为识别的终点, 而是智能安全监测系统的基础感知层, 其输出精度与实时性直接影响后续行为识别与风险评估模块的性能. 因此, 本文围绕复杂工业场景, 重点研究实时高精度的人体姿态估计方法, 为作业行为识别与安全风险评估提供可靠的数据基础.

要实现复杂工业场景下精准高效的行为识别, 快速且准确的人体姿态估计是关键[19]. 然而, 现有的主流人体姿态估计算法, 尽管在 COCO、MPII 等基准数据集上取得了较高的准确率, 但在实际场景的应用中仍面临一些挑战. 首先, 这些模型通常计算量大、内存占用高, 难以满足工业实时性. 其次, 当部分关节受到遮挡时, 模型的精度往往会大幅度下降. 因此, 在复杂工业场景中实现精确且高效的行为识别时, 主要面临着模型计算量大、内存需求高以及遮挡等挑战. 如何在保证实时性和准确性的同时, 减少遮挡对姿态估计的负面影响, 仍然是当前人体姿态估计技术在工业领域应用中的重点难题[20]. 尤其是在控制与决策场景中, 快速姿态估计不仅能够支撑对工人行为的实时监测, 还能为后续的风险预测、生产调度与安全决策提供高质量数据支撑, 从而形成“感知—分析—决策”的闭环. 因此, 本文提出了一系列创新的优化方法, 旨在解决复杂工业环境中的遮挡问题和实时推理需求, 从而提高人体姿态估计的准确性和效率.

本文的主要研究工作如下:

1) 为实现复杂工业场景中高效的实时姿态估计, 本文采用轻量化的 ResNeSt 模型作为骨干网络, 构建了一种人体姿态快速估计模型 FastPose, 将计算量减少了 4 倍.

2) 针对姿态估计模型的精度下降问题, 本文提出了循环权重迁移训练方法 CyclicTrain, 该方法通

过在不同尺寸的 ResNeSt 模型之间进行权重迁移与微调, 逐步优化小模型的性能.

3) 本文构建了一个涵盖多种工业场景的自制数据集, 包括不同角度、动态遮挡及人员姿态变化的样本, 用于验证本文所提方法的有效性. 实验结果表明, 所提方法能够在复杂工业场景下准确估计人体姿态, 与原始方法相比, 模型计算量减少了 4 倍.

1 预备知识

1.1 量化编码器

量化自编码器由编码器、密码本和解码器三部分构成, 通过联合优化实现对人体姿态的高效压缩与重建[21].

编码器将原始姿态$G \in \mathbb{R}^{K\times D}$($K$为关节点数, D为坐标维度) 映射为M个连续的 token 特征:

$$T = (t_1, t_2, \ldots, t_M) = f_e(G). \tag{1}$$

其中$f_e(\cdot)$为基于 MLP-Mixer 的多层感知结构. 具体而言, 各关节点坐标首先通过线性投影提升特征维度, 随后经由 MLP-Mixer 模块进行跨关节特征融合, 最终通过线性层生成M个 token 特征. 每个 token 对应于人体姿态中的局部子结构, 这种冗余设计能够提高模型对遮挡情况的鲁棒性.

密码本定义为$\mathrm{C} = (c_1, \ldots, c_V)^T \in \mathbb{R}^{V\times N}$, 其中$V$为密码本条目数, N为嵌入维度. 通过最近邻搜索将连续 token 特征离散化:

$$q(t_i = v \mid G) = \begin{cases} 1 \text{ if } v = \arg\min_j \|t_i - c_j\|_2 \\ 0 \text{ otherwise} \end{cases} \tag{2}$$

其中$q(t_i)$表示第i个 token 对应的密码本索引. 量化后的 token 特征为$c_{q(t_1)}, c_{q(t_1)}, \ldots, c_{q(t_M)}$.

解码器$f_d(\cdot)$将量化后的 token 重建为原始姿态:

$$\hat{G} = f_d(c_{q(t_1)}, c_{q(t_2)}, \ldots, c_{q(t_M)}). \tag{3}$$

其结构与编码器对称, 但采用更浅的 MLP-Mixer 层以减少计算量.

编码器、密码本与解码器通过以下损失函数联合优化:

$$l_{pct} = smooth_{L1}(\hat{G}, G) + \beta \sum_{i=1}^{M} \|t_i - sg[c_{q(t_i)}]\|_2^2. \tag{4}$$

其中第一项为姿态重建损失 (平滑 L1 损失), 第二项为密码本对齐损失, $sg[\cdot]$表示梯度截断操作, β为超参数.

综合而言, 量化编码器通过离散化特征表示实现了姿态信息的高效压缩与重建, 显著减少了特征冗余并提升了对遮挡的鲁棒性.

1.2 迁移学习

迁移学习 (Transfer Learning) 是一种机器学习方法, 它将从一个或多个源任务中学习到的知识迁移到目标任务中, 以解决目标任务中的新问题. 该方法利用源任务与目标任务之间的相关性, 缓解目标任务中数据不足或训练困难的问题[22], 突破了传统机器学习中样本独立同分布的假设.

在迁移学习的中, 领域 D 由特征空间 X 和边际概率分布$P(x)$组成, 公式如下:

$$D = \{x, P(X)\} \tag{5}$$

其中, X表示样本集合.

由标签 Y 和目标预测函数$f(x)$组成, 公式如下:

$$T = \{Y, f(x)\} \tag{6}$$

其中, T是目标任务.

迁移学习的优势在于能够充分利用源领域的先验知识, 降低目标领域的学习难度, 提高模型训练效率和泛化性能. 该方法已广泛应用于图像识别、语音识别和自然语言处理等任务中, 尤其在目标任务数据有限的情况下表现出显著优势[23,24].

2 人体姿态快速估计方法

2.1 工作流程

整体模型由四个主要模块构成: (1) 骨干网络: 采用轻量化的 ResNeSt 结构进行初步视觉特征提取; (2) 特征提取网络: 进一步整合多层语义信息, 提取高分辨率姿态特征; (3) 量化编码器: 将提取到的姿态特征映射为离散 token 表示, 以实现特征压缩与姿态重建; (4) 解码器模块: 基于量化 token 还原人体关键点坐标, 实现最终姿态估计.

这四个模块构成了从图像输入到关键点预测的完整处理流程, 其关系如图 1 所示. 基于量化自编码器构建人体姿态估计模型, 该模型通过骨干网络、特征提取网络、量化密码本和解码器的协同工作, 从输入图像中估计出完整的人体姿态.

此外, 在模型设计中, 选择量化自编码器作为核心模块, 主要基于以下三个原因: (1) 通过将连续的姿态特征离散化为 token 表示, 能够有效减少冗余信息, 提取最关键的姿态特征; (2) 量化自编码器采用多 token 冗余设计, 每个 token 对应人体姿态的局部子结构 (如肢体组合), 即使部分关键点受到遮挡, 其他 token 仍能保留有效信息, 从而显著提高模型在复杂工业场景中的鲁棒性; (3) 解码器采用轻量化的 MLP-Mixer 结构, 相比传统解码器计算量更低, 有助于满足工业场景对实时性的严格要求. 随后采用轻量化的 ResNeSt 模型作为骨干网络来降低计算量.

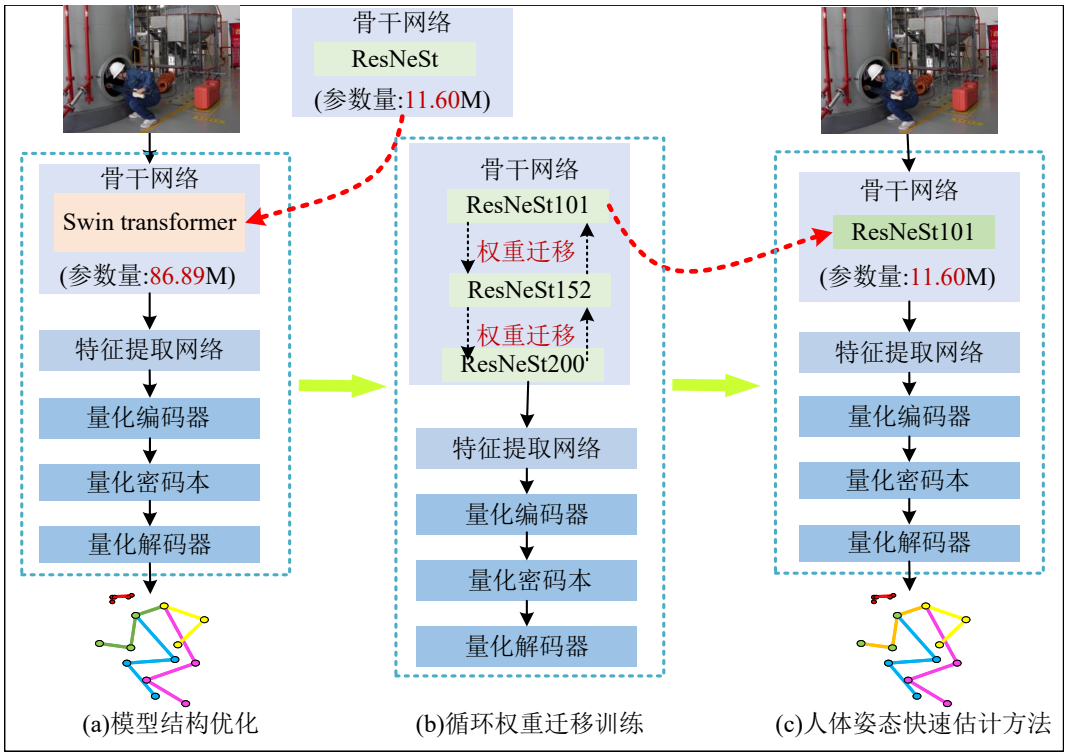


图1 面向复杂工业场景的人体姿态快速估计方法流程

接着, 以 ResNeSt101、ResNeSt152 和 ResNeSt200 作为骨干网络, 分别构建人体姿态估计模型, 并在这三个模型之间循环迁移权重参数, 优化小尺寸模型 (ResNeSt101) 的精度. 最后, 经过轻量化处理和循环权重迁移训练后的 ResNeSt101 模型作为骨干网络, 与其他组件共同构建人体姿态快速估计模型, 实现复杂工业场景下快速且准确的人体姿态估计.

2.2 模型结构优化

以 Swin Transformer 作为骨干网络的人体姿态估计模型, 虽然人体估计精度高, 但存在复杂工业场景应用运行速度慢、内存占用高等问题. 鉴于此, 本文采用轻量化模型 ResNeSt(Split-Attention Networks) 对人体姿态估计模型进行结构优化.

ResNeSt 是一种基于 CNN 的新型架构, 其核心优势在于引入了 Split-Attention 模块, 这个模块结合了多路径网络布局和通道注意力机制. 相比传统的 ResNet, ResNeSt 能够自适应地关注不同特征通道, 通过多路径特征融合增强模型对局部特征的表征能力. 这一特性使得 ResNeSt 在处理复杂工业场景中的遮挡、光照变化、复杂背景等问题时具有更强的鲁棒性. 特别是在遮挡情况下, Split-Attention 机制能够动态调整特征权重, 使模型更好地关注未被遮挡的关键点区域, 从而提高姿态估计的准确性.

具体而言, Split-Attention 模块首先将输入特征图划分为多个“卡片组”, 每个组内的特征图会通过一系列的卷积操作进行变换, 对于第 k 个卡片组, 其表示为:

$$\hat{U}^k = \sum_{j=R(k-1)+1}^{Rk} U_j. \tag{7}$$

其中, R为每个卡片组内子组数量, U_j表示第j个子组.

接着, ResNeSt 通过全局平均池化计算全局上下文信息, 来为每个通道分配一个注意力权重, 这一操作通过以下公式完成:

$$s_k^c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \hat{U}_k^c(i,j). \tag{8}$$

其中, s_k^c表示卡组k中通道c的全局上下文信息, $\hat{U}_k^c$表示卡片组k中通道c在位置(i,j)的特征值, H和W是特征图的空间维度.

然后, 使用这些全局上下文信息计算每个通道的软注意力权重, 并对每个卡片组内的特征图进行加权求和, 具体公式如下:

$$V_k^c = \sum_{i=1}^{R} a_k^i(c) U_{R(k-1)+i}. \tag{9}$$

$U_{R(k-1)+i} $

综上所述, ResNeSt 通过 Split-Attention 机制在保持较低计算复杂度的同时, 显著提升了模型在复杂工业场景中的精度和鲁棒性. 具体模型参数信息如表 1 所示. 从表中数据可以看出, 采用 ResNeSt 作为骨干网络显著降低了模型的内存占用、参数量和计算量, 为实时姿态估计提供了有力支持.

表1 两种骨干网络构建的姿态估计模型参数对比

骨干网络	内存占用	参数量	计算量
Swin Transformer	851M	215.03M	15.26G
ResNeSt	446M	64.44M	3.61G

2.3 循环权重迁移训练

模型结构的优化有效减少了对计算资源的需求, 但也带来了一定的精度损失. 为提高人体姿态快速估计模型的精度, 本文引入知识迁移技术, 提出了一种循环权重迁移训练方法 (Cyclic Weight Transfer Training, CyclicTrain). 该方法核心思想是利用大尺寸模型的学习成果为小尺寸模型提供知识学习方向, 提升小尺寸模型的表征能力和泛化性能, 并降低训练难度, 减少计算成本, 具体流程如图 2 和表 2 所示.

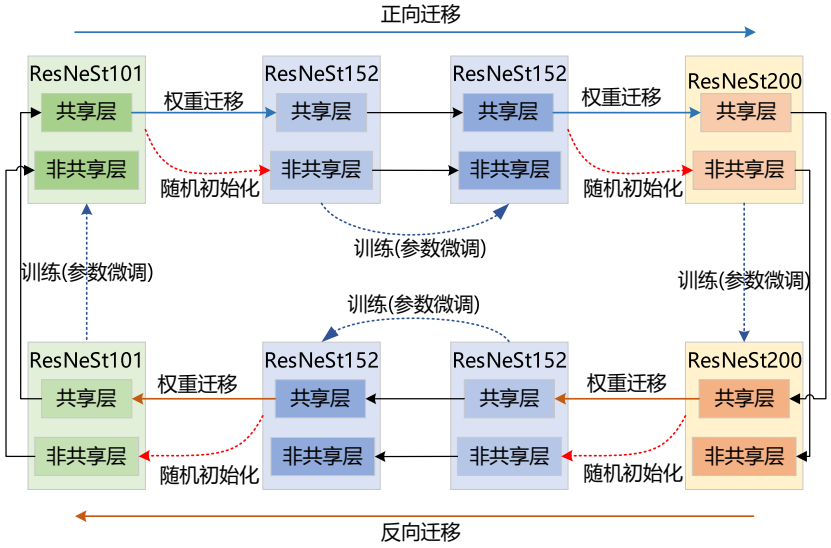


图2 CyclicTrain 工作流程

表2 CyclicTrain 方法伪代码

算法1 循环权重迁移训练CyclicTrain

输入: 训练数据集D_{train}, 模型集合$\{M_0, M_1, M_2, \}$, 迭代次数T

输出: 优化后的模型集合$\{M_0, M_1, M_2, \}$

1:在数据集D_{train}上训练M_0至收敛
2: for $t = 1$至T do
3: //正向迁移阶段
4: for 模型层级$k = 1$至2 do
5: if M_k与M_{k-1}存在共享层 then
6: M_k.共享层参数←M_{k-1}.权重参数
7: end if
8: 随机(或线性投影)初始化M_k新增层参数
9: 在D_{train}上训练微调M_k
10: end for
11: //反向迁移阶段
12: for 模型层级$k = 2$降至1 do
13: if M_{k-1}与M_k存在共享层 then
14: M_{k-1}.共享层参数←M_k.权重参数
15: end if
16: 随机(或线性投影)初始化M_{k-1}差异层参数
17: 在D_{train}上训练微调M_{k-1}
18: end for
19: end for

CyclicTrain 训练流程分为正向迁移和反向迁移两个阶段. 在正向迁移阶段, 首先使用较小的 ResNeSt101 模型进行初步训练, 使模型迅速收敛并学习基础姿态特征. 随后, 将训练完毕的小尺寸模型的权重迁移至容量更大的模型 (如 ResNeSt152、ResNeSt200). 对于大尺寸模型中新增的网络层 (如更深的残差块或更宽的通道), 采用随机初始化或线性投影以实现参数匹配, 对于结构一致的共享层, 则直接继承小模型的权重参数. 大尺寸模型在权重迁移后通过微调进一步优化, 重点调整新增层以捕捉更细粒度的姿态特征. 该阶段的目标是逐步放大模型容量, 以便更好地捕捉复杂姿态变化.

在反向迁移阶段, 将训练完成的大尺寸模型 (例如 ResNeSt200、ResNeSt152) 的权重参数回传至小尺寸模型 (如 ResNeSt101). 在此过程中, 仅共享结构相同的层的权重, 而对于差异部分则采用随机初始化进行训练. 经过多次循环迭代, 反向迁移可以显著提升了小尺寸模型的泛化能力, 使得姿态估计模型在不牺牲精度的前提下提高运行效率. 综上, 通过将 CyclicTrain 方法运用到 ResNeSt 构建的姿态估计模型中, 可以显著提高小尺寸模型在人体姿态估计任务中的精度. 正向与反向迁移的循环训练过程不断优化姿态估计模型在工业环境中的适应能力, 并在

图3 部分工业场景数据集

精度和计算效率之间取得了有效平衡.

3 实验与分析

3.1 实验环境及实现

实验环境为 Windows 10, Python 3.8, PyTorch 1.13.0 和 CUDA 11.7. PyTorch 用于构建深度神经网络模型. 实验数据采用自定义的工业场景数据集, 该数据集包含多种工业环境下的 2D 人体姿态图像, 其中训练集包含 1015 张图像, 测试集包含 285 张图像, 部分图像如图 3 所示. 数据集涵盖了多种工业场景和多样化的关节配置, 具有较强的代表性.

本实验采用基于 OKS(Object Keypoint Similarity) 的 AP、AP^{50} 和 AP^{75} 为评价指标[25]. 其中, OKS 用于量化预测关键点与真实关键点之间的相似程度, 其计算公式如下:

$$OKS_p = \frac{\sum_i \exp\left(-d_{p^i}^2/2s_p^2\sigma_i^2\right)\delta\left(v_{p^i}=1\right)}{\sum_i \delta\left(v_{p^i}>0\right)}. \quad (10)$$

其中, p表示第p个人, i表示人体第i个关键点, $v_{p^i}=0$表示人体关键点存在遮挡且未标注, $v_{p^i}=1$表示人体关键点不存在遮挡且标注, $v_{p^i}=2$表示人体关键点存在遮挡且标注, s_p^2表示第 p 个目标的边界框面积, σ_i^2表示各关键点的标准偏差调节不同关键点的权重, δ表示可见性, d_{p^i}表示预测关键点和真实关键点的欧氏距离. OKS 计算得到每个人物实例的相似度后, 可以在阈值 s 不同取值下计算 AP^s 值, 其计算公式如下:

$$AP^s = \frac{\sum_p \delta\left(OKS_p > s\right)}{\sum_p 1}. \quad (11)$$

其中, AP 表示平均精度, 反映了模型在不同阈值下的综合性能表现.

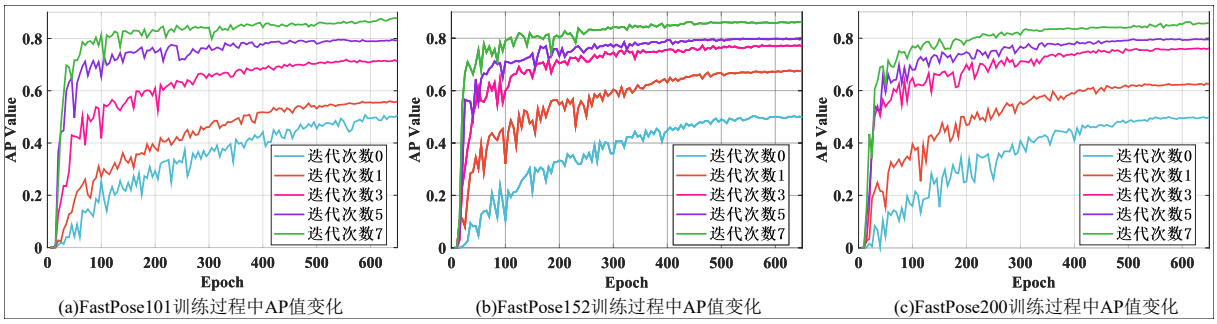
在模型训练过程中, 本文将初始学习率设置为 1×10^{-3}, 动量为 0.9, 权重衰减为 1×10^{-4}, 并在自定义工业数据集上进行了 650 个 epoch 的训练. 此外, 本文还采用一些数据增强方法, 包括随机缩放、旋转、翻转等.

3.2 实验 1

本节主要对比不同尺寸的 ResNeSt 模型作为骨干网络在人体姿态估计任务中的表现, 并探讨了 CyclicTrain 方法对模型性能的提升作用. 具体而言, 本实验分别以 ResNeSt101、ResNeSt152 和 ResNeSt200 作为骨干网络构建人体姿态快速估计模型, 并将其分别命名为 FastPose101、FastPose152 与 FastPose200. 所有模型均在相同的自定义工业数据集上进行了 650 个 epoch 的完整训练周期.

实验分为两部分: 一是采用常规训练方法直接训练; 二是采用 CyclicTrain 方法, 在不同尺寸模型间循环进行权重迁移与微调.

图 4 展示了 FastPose 系列姿态估计模型在训练过程中的平均精度 (AP) 随训练轮次 (Epoch) 的变化曲线. 蓝色曲线表示采用常规训练方法的 AP 值变化, 其他曲线则代表在循环权重迁移训练过程中, 不同迭代次数下 AP 值的变化趋势. 以 FastPose101 为例, 由图 4(a) 可知, 常规训练方法的 AP 值在初始阶段有所提升, 但随后增长速度减缓, 最终稳定在约 0.5 的水平波动, 表明在没有任何权重迁移的情况下, 模型的性能提升缓慢, 且最终达到的精度较低. 相比之下, 采用 CyclicTrain 方法时, 随着迭代次数的增加, 模型的 AP 值显著提高, 增长速度逐渐加快. 特别是在进行多次权重迁移和微调后, 模型的 AP 值在训练初期便表现出快速上升的趋势, 最终稳



(a)FastPose101训练过程中AP值变化 (b)FastPose152训练过程中AP值变化 (c)FastPose200训练过程中AP值变化

图4 FastPose 系列人体姿态估计模型的训练实验曲线

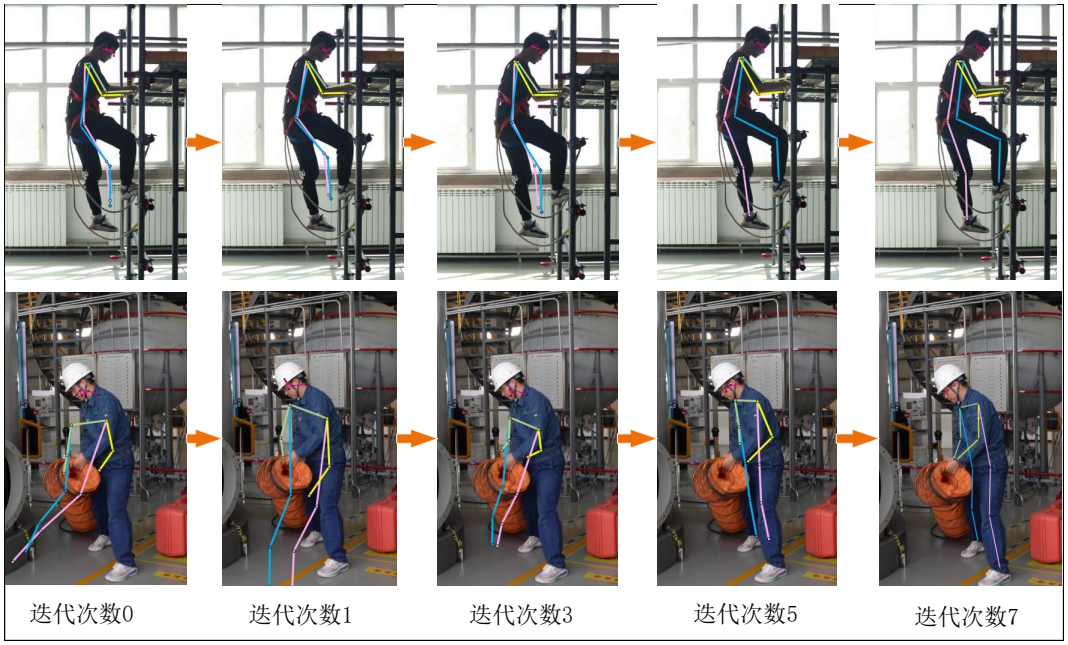


图5 CyclicTrain 过程中模型精度提升效果

定在约 0.88 的位置, 明显优于常规训练. 这一结果验证了 CyclicTrain 方法在提升模型精度方面的有效性. 图 4(b) 和图 4(c) 分别展示了 FastPose152 和 FastPose200 在训练过程中的 AP 值变化, 其表现与 FastPose101 类似.

为进一步验证 CyclicTrain 过程中模型精度的提升, 本文使用 FastPose101 模型进行了多次循环权重迁移的迭代训练. 在每次迭代后, 使用相同的图片进行推理分析, 图 5 展示了这一过程中模型推理结果的变化情况, 随着迭代次数的增加, 模型推理出的人体关键点位置逐渐逼近实际位置, 直观地反映出模型在训练过程中对姿态估计的精确度不断提升.

表 3 展示了 FastPose 系列姿态估计模型在常规训练和 CyclicTrain 下的 AP、AP50、AP75 和计算量等性能指标对比. 根据表 3 所示, 采用 CyclicTrain 方法的姿态估计模型在各项指标上均显著优于常规训练模型. 以 FastPose101 模型为例, CyclicTrain 使其平均精度 (AP) 从 0.5393 提升至 0.8788, 提升幅度达 62.9%. 同样, AP50 和 AP75 也分别提高了 3.6% 和 83.8%. 这一结果表明, CyclicTrain 方法在提高模型精度方面具有显著优势, 尤其在小尺寸模型上表现突出. 此外, FastPose152 和 FastPose200 在采用 CyclicTrain 方法训练后, AP、AP50 和 AP75 等精度指标也有显著提升, 进一步证明该方法在不同模型规模下均能有效提高精度.

表3 不同训练方法下的姿态估计实验对比结果

姿态估计模型	训练方法	AP	AP^{50}	AP^{75}	计算量(Flops)
PCT	—	0.8689	1	0.9864	15.26G
FastPose101	常规训练	0.5393	0.9649	0.5391	3.61G
FastPose101	CyclicTrain	**0.8846**	**1**	0.9901	**3.61G**
FastPose152	常规训练	0.5033	0.9413	0.4411	4.84G
FastPose152	CyclicTrain	0.8621	1	0.9860	4.84G
FastPose200	常规训练	0.4995	0.9514	0.4579	6.01G
FastPose200	CyclicTrain	0.8428	1	0.9861	6.01G

在计算量方面, FastPose101 的每秒浮点运算次数 (Flops) 仅为 3.61G, FastPose200 为 6.01G, 而 PCT 模型的 Flops 高达 15.26G. 值得注意的是, 经过 CyclicTrain 方法训练后的 FastPose101, 其精度甚至略高于 PCT 模型. 这表明, 通过 CyclicTrain 方法, 可以在保持小模型较低计算成本的同时, 达到甚至超

表4 不同姿态估计模型的实验结果

姿态估计模型	AP	AP^{50}	AP^{75}	计算量(Flops)
FastPose101	**0.8846**	**1**	**0.9901**	**3.61G**
ViTPose-base[26]	0.8258	0.9802	0.9358	18.54G
PEM-SCNet101[27]	0.8577	0.9906	0.9554	8.49G
PEM-ResNet101[28]	0.8116	0.9898	0.9347	7.69G
RTMPose-l[29]	0.8711	0.9901	0.9689	4.16G
PoseBH-VitBase[30]	0.8399	1	0.9536	18.5G

过大模型的精度表现.

上述实验结果验证了 CyclicTrain 方法在不同尺寸 FastPose 模型中的有效性. 该方法通过正向迁移和反向迁移的循环迭代, 逐步优化小尺寸模型的初始化, 使其能够快速收敛, 并充分利用从大尺寸模型中学到的细粒度姿态特征, 从而提升其在复杂工业场景中的精度和泛化能力. 在资源受限的环境中, 首先使用小尺寸模型进行初步训练, 再通过 CyclicTrain 方法逐步优化. 该方法在保持小模型较低计算量的同时显著提升精度, 使轻量化模型在实时推理任务中发挥更大作用.

3.3 实验 2

本节对基于自定义工业数据集训练的多种人体姿态估计模型进行了性能比较, 重点分析了 FastPose101 与其他主流模型之间的差异. 具体对比结果如表 4 所示, 其中 PEM-SCNet101 与 PEM-ResNet101 分别表示以 SCNet101 与 ResNet101 为骨干网络构建的姿态估计模型.

从表 4 的实验结果可以看出, 不同模型在该工业数据集上的性能表现存在较为显著的差异. 总体而言, FastPose101 在各项指标上均表现出最优性能. 其 AP 达到 0.8846, AP50 与 AP75 分别为 1 与 0.9901, 在保证高精度的同时, 计算量仅为 3.61G, 显著低于 ViTPose-base(18.54G) 与 PEM-SCNet101 (8.49G) 等模型, 展现出较高的计算效率与轻量化优势. 相较而言, RTMPose-l 虽在部分指标上表现较好, 但整体精度仍略低于 FastPose101; ViTPose-base 在部分高阈值指标上表现较优, 但其计算量过大, 难以满足工业场景对实时性和资源受限环境的要求. PEM-SCNet101、PEM-ResNet101 和 PoseBH 等模型在精度与计算复杂度方面均不及 FastPose101, 其在复杂工业任务中的适用性相对有限.

综上所述, FastPose101 在精度与计算效率之间实现了良好的平衡, 能够在资源受限的工业环境中保持高精度和实时性, 因此在复杂工业场景下具有显著的应用优势.

3.4 消融实验

本节旨在验证本文所提出方法中各关键模块的有效性, 包括 ResNeSt 骨干网络、量化编码器以及 CyclicTrain 训练策略. 为此, 设计了一系列消融实验, 以对比分析不同模块对模型性能的具体影响. 所有实验均在自定义工业数据集上进行, 训练配置与第 3.1 节保持一致. 实验结果如表 5 所示.

表5 不同模块组合对模型性能的影响

骨干网络	量化编码器	训练策略	AP	计算量(Flops)
ResNet101	√	CyclicTrain	0.8116	3.52G
ResNeSt101	×	CyclicTrain	0.5765	3.45G
ResNeSt101	√	常规训练	0.5393	3.61G
ResNeSt101	√	CyclicTrain	**0.8846**	3.61G

表 5 所示, 不同模块组合对模型性能产生了显著影响. 首先, 在相同量化编码器与训练策略条件下, 将骨干网络由 ResNet101 替换为 ResNeSt101 后, 模型的 AP 由 0.8116 提升至 0.8846, 表明 ResNeSt 中引入的分组通道注意力机制能够增强关键人体区域的特征表达能力, 从而在工业场景中获得更高的关键点定位精度. 其次, 在相同骨干网络与 CyclicTrain 策略下, 移除量化编码器后模型性能明显下降 (AP 由 0.8846 降至 0.5765). 这说明量化编码器在减少特征冗余、保持姿态结构一致性方面具有重要作用, 尤其在存在遮挡和复杂背景时, 对模型的精度提升贡献显著. 最后, 将 CyclicTrain 替换为常规训练策略后, 模型 AP 从 0.8846 降至 0.5393, 性能大幅下降, 表明 CyclicTrain 有效缓解了小模型特征表达能力受限的问题, 通过跨尺度迁移训练显著提升了模型的泛化能力和关键点预测稳定性.

综上所述, 骨干网络、量化编码器与 CyclicTrain 训练策略均对模型性能提升具有不可替代的作用. 其中, 量化编码器与 CyclicTrain 对性能的提升最为显著, 而 ResNeSt 骨干网络则进一步增强了模型的鲁棒性和特征表达能力.

3.5 鲁棒性分析

为全面评估 FastPose101 在复杂工业环境中的适应能力, 本节从遮挡与光照变化两个方面对模型的鲁棒性进行分析.

3.5.1 遮挡场景下的性能分析

在实际工业作业中, 操作人员的身体关键点常因设备、工具或其他人员的遮挡而导致部分信息缺失. 为定量评估模型的抗遮挡能力, 将测试集按照遮挡比例划分为无遮挡 (0%)、轻度遮挡 (0–20%)、中度遮挡 (20–40%) 和严重遮挡 (>40%) 四类, 实验结果如表 6 所示.

表6 不同遮挡程度下各方法的 AP 值对比

模型	无遮挡	轻度遮挡	中度遮挡	严重遮挡
FastPose101	0.9783	0.9022	0.8773	0.8074

从表 6 可以看出, FastPose101 在不同遮挡等级下均保持较高的 AP 值, 表现出优越的鲁棒性. 尤其是在中度遮挡条件下, AP 仅下降约 10.3%, 即使在严重遮挡条件下, AP 仍保持在 0.8074, 表明模型能够较好地应对关键点缺失带来的信息干扰.

这种鲁棒性主要得益于以下两方面机制: (1) ResNeSt 骨干网络的 Split-Attention 机制能够自适应地聚焦于可见区域特征, 从而降低遮挡区域带来的噪声干扰; (2) 量化编码器的多 token 冗余设计使得模型在部分关键点丢失的情况下, 仍可利用剩余 token 的局部结构信息实现姿态结构的可靠重建. 因此, FastPose101 在复杂遮挡环境下依然能够稳定输出高质量姿态估计结果.

3.5.2 光照变化下的性能分析

工业作业场景多为室内环境, 整体光照条件相对稳定, 但局部区域可能存在一定亮度波动, 如设备反光、阴影或监控摄像头自动曝光引起的光照变化. 为验证模型在不同光照条件下的鲁棒性, 本文通过调整图像亮度分别模拟了轻微变暗 (亮度降低 20%) 与轻微过曝 (亮度提升 20%) 两种情况, 实验结果如表 7 所示.

表7 不同模拟光照下各方法的 AP 值对比

模型	原始光照	轻微变暗	轻微过曝
FastPose101	0.8846	0.8818	0.8855

从表 7 可以看出, FastPose101 在光照变化下的性能几乎保持稳定, AP 仅在轻微变暗条件下降约 0.3%, 在轻微过曝条件下甚至略有提升, 表现出较强的光照鲁棒性. 这一结果表明, ResNeSt 骨干网络提取的多尺度特征具有较强的亮度不变性, 可在亮度变化时保持稳定的特征响应. 同时, 量化编码器对特征的离散化表示进一步增强了模型对低对比度区域的结构敏感性, 从而有效抵抗光照扰动带来的信息损失.

综上所述, 通过遮挡与光照变化两方面的分析结果表明, FastPose101 在复杂工业场景中具备出色的鲁棒性与稳定性. 该模型能够在不同视觉干扰条件下保持较高精度, 为其在实际工业安全监测系统中的应用奠定了可靠基础.

4 结论及未来工作

本文提出了一种基于 ResNeSt101 骨干网络构建的人体姿态快速估计模型 FastPose101, 并结合循环权重迁移训练方法, 显著提升了复杂工业环境中人体姿态估计的精度与计算效率. 实验结果表明, 所提出的模型在复杂背景、遮挡等情况下表现优异, 且相较于基于 Swin Transformer 骨干网络构建的姿态估计模型, 具有更低的计算量, 能够适应实时性要求较高的工业场景. 未来的研究将集中于优化模型在极端复杂场景下的表现, 探索多模态信息融合, 提升模型在遮挡和光照变化中的鲁棒性, 并进一步优化实时推理性能, 以更好地满足工业应用需求. 此外, 我们已在初步实验中尝试将本文的人体姿态估计模型与行为识别模块进行融合测试, 用于识别工人违规操作与潜在危险动作. 初步结果表明, 本文模型输出的姿态关键点特征能够有效支持高层行为识别任务. 未来, 我们将进一步基于本文提出的姿态估计结果, 构建面向工业安全的行为识别与风险评估系统, 实现从姿态感知到安全决策的闭环应用.

参考文献 (References)

[1] 危化监管一司. 2023 年全国化工事故分析报告[R]. 北京: 危化监管一司, 2024.
(Wei Hua Supervision Department I. 2023 National Chemical Accident Analysis Report[R]. 2024.)

[2] Park J, Kim H. A case study to address the limitation of accident scenario identifications with respect to diverse manual responses[J]. Reliability Engineering & System Safety, 2024, 251: 110406.

[3] Khan M, Khalid R, Anjum S, et al. Tag and IoT based safety hook monitoring for prevention of falls from height[J]. Automation in Construction, 2022, 136: 104153.

[4] Misra S, Roy C, Sauter T, et al. Industrial Internet of Things for safety management applications: A survey[J]. IEEE Access, 2022, 10: 83415-83439.

[5] 朱红蕾, 卫鹏娟, 徐志刚. 基于骨架的人体异常行为识别与检测研究进展[J]. 控制与决策, 2024, 39(8): 2484-2501.
(Zhu H L, Wei P J, Xu Z G. Research progress on skeleton-based human abnormal behavior recognition and detection[J]. Control and Decision, 2024, 39(8): 2484-2501.)

[6] 南静, 宁传峰, 建中华, 等. 基于随机配置网络的轻量级人体行为识别模型[J]. 控制与决策, 2023, 38(6): 1541-1550.
(Nan J, Ning C F, Jian Z H, et al. A lightweight model for human activity recognition using stochastic configuration networks[J]. Control and Decision, 2023, 38(6): 1541-1550.)

[7] Cheng J P, Wong P K, Luo H, et al. Vision-based monitoring of site safety compliance based on worker re-identification and personal protective equipment

classification[J]. Automation in Construction, 2022, 139: 104312.

[8] Yang X J, Weng C Y, Jiao L, et al. ATD-GCN: A human activity recognition approach for human-robot collaboration based on adaptive skeleton tree-decomposition[J]. Robotics and Computer-Integrated Manufacturing, 2025, 95: 103019.

[9] Liu T Y, Weng C Y, Jiao L, et al. Toward fast 3D human activity recognition: A refined feature based on minimum joint freedom model (Mint)[J]. Journal of Manufacturing Systems, 2023, 66: 127-141.

[10] 张晓平, 纪佳慧, 王力, 等. 基于视频的人体异常行为识别与检测方法综述[J]. 控制与决策, 2022, 37(1): 14-27.
(Zhang X P, Ji J H, Wang L, et al. Overview of video based human abnormal behavior recognition and detection methods[J]. Control and Decision, 2022, 37(1): 14-27.)

[11] Wang X M, Yu H L, Kold S, et al. Wearable sensors for activity monitoring and motion control: A review[J]. Biomimetic Intelligence and Robotics, 2023, 3(1): 100089.

[12] Sadeghi S, Soltanmohammadlou N, Nasirzadeh F. Applications of wireless sensor networks to improve occupational safety and health in underground mines[J]. Journal of Safety Research, 2022, 83: 8-25.

[13] Digo E, Gastaldi L, Antonelli M, et al. Real-time estimation of upper limbs kinematics with IMUs during typical industrial gestures[J]. Procedia Computer Science, 2022, 200: 1041-1047.

[14] Li J, Liu X F, Wang Z L, et al. Real-time human motion capture based on wearable inertial sensor networks[J]. IEEE Internet of Things Journal, 2022, 9(11): 8953-8966.

[15] Wang B C, Song C, Li X Y, et al. A deep learning-enabled visual-inertial fusion method for human pose estimation in occluded human-robot collaborative assembly scenarios[J]. Robotics and Computer-Integrated Manufacturing, 2025, 93: 102906.

[16] 胡楠, 张家豪, 魏晓彤, 等. 一种基于多假设交互的三维人体姿态估计模型[J]. 控制与决策, 2025, 40(12): 3704-3712.
(Hu N, Zhang J H, Wei X T, et al. A 3D human pose estimation model based on multiple hypothesis interaction[J]. Control and Decision, 2025, 40(12): 3704-3712.)

[17] 宋忱, 钱惠敏, 吴大伟. 面向骨架行为识别的多语义动态图卷积网络[J]. 控制与决策, DOI: 10.13195/j.kzyjc.2025.0793.
((Song C, Qian H, Wu D. Multi-semantic dynamic graph convolutional networks for skeleton-based action recognition[J]. Control and Decision, DOI: 10.13195/j.kzyjc.2025.0793.)

[18] 陈博, 王斌, 周袁, 等. 基于多模态数据融合的康复机器人关节角度预测方法[J]. 控制与决策, DOI: 10.13195/j.kzyjc.2025.0504.
((Chen B, Wang B, Zhou Y, et al. Prediction of rehabilitation robot's joint angle based on multi-modal data fusion method[J]. Control and Decision, DOI: 10.13195/j.kzyjc.2025.0504.)

[19] Liu T Y, Weng C Y, Huang J, et al. A lightweight future skeleton generation network(FSGN) based on spatio-temporal encoding and decoding[J]. Knowledge-Based Systems, 2024, 306: 112717.

[20] Boldo M, De Marchi M, Martini E, et al. Real-time multi-camera 3D human pose estimation at the edge for industrial applications[J]. Expert Systems with Applications, 2024, 252: 124089.

[21] Geng Z G, Wang C Y, Wei Y X, et al. Human pose as compositional tokens[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, 2023: 660-671.

[22] Zhao Z H, Alzubaidi L, Zhang J L, et al. A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations[J]. Expert Systems with Applications, 2024, 242: 122807.

[23] Yu X, Wang J, Hong Q Q, et al. Transfer learning for medical images analyses: A survey[J]. Neurocomputing, 2022, 489: 230-254.

[24] 赵健程, 冯良骏, 岳嘉祺, 等. 从零样本学习理论模型到工业应用——动机、演变与挑战[J]. 控制与决策, 2024, 39(9): 2833-2857.
(Zhao J C, Feng L J, Yue J Q, et al. From zero-shot learning theoretical model to its industrial application: Motivation, evolution and challenges[J]. Control and Decision, 2024, 39(9): 2833-2857.)

[25] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context[C]. Computer Vision – ECCV 2014. Cham: Springer, 2014: 740-755.

[26] Xu Y F, Zhang J, Zhang Q M, et al. ViTPose: Simple vision transformer baselines for human pose estimation[J/OL]. 2022, arXiv: 2204.12484.

[27] Liu J J, Hou Q B, Cheng M M, et al. Improving convolutional networks with self-calibrated convolutions[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 10093-10102.

[28] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 770-778.

[29] Jiang T, Lu P, Zhang L, et al. RTMPose: Real-time multi-person pose estimation based on MMPose[J/OL]. 2023, arXiv: 2303.07399.

[30] Jeong U, Freer J, Baek S, et al. PoseBH: Prototypical multi-dataset training beyond human pose estimation[C]. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, 2025: 13466-13476.

作者简介

张泽辉 (1994−), 男, 副研究员, 博士, 主要研究方向为计算机视觉、故障诊断, E-mail: zhangtianxia918@163.com;

吴富龙 (1997−), 男, 硕士生, 主要研究方向为计算机视觉, E-mail: 232060333@hdu.edu.cn;

李帆雅 (2004−), 女, 本科生, 主要研究方向为计算机视觉, E-mail: lify_st@163.com;

徐晓滨 (1980−), 男, 教授, 博士, 主要研究方向为智能信息融合与证据推理, E-mail: xuxiaobin1980@hdu.edu.cn;

章振杰 (1989−), 男, 副研究员, 博士, 主要研究方向为图模型、信息融合, E-mail: zhangzhenjie1017@163.com;

邵海滨 (19xx−), 男, 副研究员, 博士, 主要研究方向为具身群体智能、集群无人系统, E-mail: shore@sjtu.edu.cn.