基于深度学习的人体姿态估计综述

江 云，刘述民*

江西理工大学软件工程学院 江西南昌

【摘要】人体姿态估计作为行为识别、行为检测的基础，是机器视觉领域的一个具有挑战的任务。近年来，随着深度学习的发展，基于深度学习的人体姿态估计算法展现出了非常优异的效果，并成为学者关注和研究的热点。本文首先将基于深度学习的人体姿态估计分为单人姿态估计、多人姿态估计两类；其次，分别介绍了近年来这两类人体姿态估计的发展，对比分析了各类算法的特性；再次，介绍了姿态估计常用数据集以及评价指标；最后，讨论了当前基于深度学习的人体姿态估计所面临的困难和挑战，并对未来发展趋势进行了展望。

【关键词】机器视觉；深度学习；人体姿态估计；关键点检测

【收稿日期】2025 年 3 月 15 日 【出刊日期】2025 年 4 月 16 日 【DOI】10.12208/j.aics.20250001

Overview on human pose estimation based on deep learning

*Yun Jiang, Shumin Liu**

School of Software Engineering, Jiangxi University of Science and Technology, Nanchang, Jiangxi

【**Abstract**】Human pose estimation, serving as the foundation for action recognition and detection, is a challenging task in the field of machine vision. With the rapid development of deep learning in recent years, deep learning-based human pose estimation algorithms have achieved remarkable performance and have become a focal point of academic research. This paper first classifies deep learning-based human pose estimation into two categories: single-person pose estimation and multi-person pose estimation. It then reviews the development of these two categories in recent years, providing a comparative analysis of the characteristics of various algorithms. Additionally, the paper introduces commonly used datasets and evaluation metrics for pose estimation. Finally, it discusses the challenges faced by current deep learning-based human pose estimation systems and offers an outlook on future research directions.

【**Keywords**】Machine vision; deep learning; Human pose estimation; Key point detection

1 引言

人体姿态估计是指从图像或视频中精确定位和识别人体关键点（如头部、肩膀、肘部、手腕、髋部、膝盖、脚踝等）坐标，并通过二维或三维坐标形式加以表示，以描述人体的姿态、动作或行为。在智能驾驶、智能安防、人机交互等领域，人体姿态估计有着十分重要的作用。目前，人体姿态估计相关技术已广泛应用于智能交通、智能驾驶、文娱体育、智能安防、健康监测、人机交互等领域。

早期的人体姿态估计研究主要采用基于手工特征的方法，如尺度不变特征变换（Scale-Invariant Feature Transform，SIFT）[1]、梯度方向直方图（Histogram of Oriented Gradients，HOG）[2]等。这类方法具有计算效率高、可解释性强和训练成本低的优点。然而，它们高度依赖特定场景的特征提取，对于背景复杂、光照变化大或存在遮挡的场景，其鲁棒性和适应性显著不足。此外，这类方法过度依赖人工设计与调优，不仅过程繁琐，还限制了模型的准确性，难以满足复杂场景下的应用需求。当前，随着深度学习的快速发展，基于深度学习的人体姿态估计算法得到业界的关注并成为研究热点。尽管已有研究对基于深度学习的人体姿态估计进行了广泛探讨，但现有综述多侧重于单一场景或特定算法类别的分析，缺乏对轻量化模型、多模态数据融合及实际应用挑战的系统性总结。本文将重点分析基于深度学习的人体姿态估计算法，从单人姿态估计、多人姿态估计两个方向对基于深度学习的姿态估计算法进行梳

*通讯作者：刘述民

理和分析，整理了相关数据集与评价指标，并对当前所面临的问题和未来发展趋势进行了阐述，为后续研究提供理论参考和实践指导。

2 单人姿态估计算法

单人姿态估计 (Single-Person Pose Estimation) 是计算机视觉中的一个基础性任务，旨在从静态图像或视频中提取出人体的各个关键点位置，并推断出人体的姿势。关键点通常包括头部、肩膀、肘部、膝盖等，构成一个骨架模型，反映了人体的整体姿态。

TOSHEV 等[3]开创性地提出了 DeepPose 单人姿态估计模型，首次将深度神经网络（DNN）应用于人体关键点检测，将人体姿态估计问题转化为回归问题，通过级联 DNN 网络直接回归关节点坐标，简化了算法流程，提升了精度和效率。直接回归关键点的算法其鲁棒性较差，而且无法读取人体关键点之间的空间关系和上下文关系。

Sun 等[4]提出了基于热力图积分回归的 Soft-argmax 方法，将人体关键点与热力图关联，利用归一化函数将热力图转化为概率分布，通过积分运算获得关键点坐标。然而，当出现多个峰值时，容易误选峰值，导致关键点回归的准确率和置信度下降。

WEI 等[5]提出了卷积姿态机（Convolutional Pose Machines，CPM）网络，通过多阶段序列卷积生成热力图以预测关键点信息，并捕捉变量之间的长期依赖关系，精确表示人体关键点。CPM 由图像特征计算模块和预测模块组成，通过特征计算生成热力图，与原图处理后的特征一起输入预测模块，循环此步骤直至输出最终结果。然而，多级预测模块在多场景下的检测效果不理想。

NEWELL 等[6]提出了基于 Hourglass 的单人姿态估计网络，通过全卷积神经网络处理输入图像，输出人体关键点位置，并通过多尺度特征捕捉空间位置信息。采用残差块有效融合多尺度特征，显著提高了姿态估计的准确性。然而，多次上采样和下采样导致计算开销大，处理高分辨率图像时性能不佳；需要存储多尺度特征图，训练时消耗较多显存。Hourglass 网络结构复杂，在小数据集上易过拟合。

SUN 等[7]开创性地提出了高分辨率网络（High-Resolution Representation Networks，HRNet），通过高低分辨率特征之间的密集连接，实现特征信息的深度融合，显著增强了信息表达能力，提高了人体姿态估计的准确性。HRNet 由基础网络（通常为 ResNet[8]）、多个不同分辨率的支网和特征融合模块构成。由于 HRNet 同时使用高低分辨率通道，其复杂结构和较大模型参数量对计算资源需求较高。此外，HRNet 也面临数据集不足导致的过拟合问题，且对小样本的泛化能力较弱。

Yu 等[9]提出了一种轻量级高分辨率网络 LiteHRNet，不仅保持了高分辨率输出，还通过随机交换模块和条件通道加权模块实现高效信息交互，显著提升关键点定位性能。LiteHRNet 在保持高分辨率的同时，精确定位关键点，并减少模型参数量和计算量。该网络采用模块化设计，能够根据需求进行调整和组合，适应多种应用。尽管 LiteHRNet 精度较高，但在复杂环境下，其精度不如更复杂的网络。

Wang 等[10]通过实验发现高分辨率网络存在冗余问题，提出了一种轻量级 LitePose 网络，通过融合反卷积头和大尺寸卷积核优化姿态估计性能，在显著减少模型参数量的同时，保持较高的姿态估计准确率。LitePose 不仅适用于算力有限的设备，还在姿态估计方面表现优异。然而，LitePose 在复杂场景下的表现不如大型姿态检测模型 HRNet。

Jian 等[11]提出 Graph-PCNN 网络，是一种基于现有热图回归方法并引入图卷积神经网络（GCN）的创新性两阶段单人姿态估计方法。第一阶段，利用现有的热图回归技术从图像中提取关键点坐标；第二阶段，通过引入图卷积神经网络（GCN），扩展并构建了图姿态精化模块（GPR），以考虑不同关键点之间的关系，从而更准确地优化关键点的位置。GPR 模块通过图卷积操作更新每个关键点的特征表示，利用相邻关键点的信息提高定位精度。

上述方法通常通过热力图与人体关键点的关联来获取关键点坐标。然而，热力图的计算和提取过程消耗大量资源，限制了其在算力受限平台上的部署。尽管一些研究提出了轻量化的热力图模型，能够在牺牲部分精度的情况下适配低算力环境，但模型性能往往难以达到预期效果。

Yang 等人[12]提出一种 TokenPose 网络，利用 Transformer 进行骨骼关键点估计和回归，通过自注意力机制实现 token 与输入图像的交互，从而精准回归人体关键点坐标。TokenPose 利用 Transformer 在网络的前几层关注全局上下文，最后收敛至单个关键点位置，并通过邻接关键点和对称关键点建立约束。TokenPose 从数据中提取关键点之间的约束关系，并将其编码为关键点 Tokens，通过向量结构记忆关键点之间的关系。

Li 等人[13]提出 Hourglass Tokenizer(HoT)网络，用

于高效地进行基于 Transformer 的三维人体姿态估计，旨在解决现有视频姿态 Transformer（VPT）在资源受限设备上的高计算成本问题，其核心思想是通过剪枝和恢复 Token 来提高计算效率。考虑到基于视频的 3D 人体姿态估计方法计算成本普遍较高，尤其处理长视频序列，HoT 框架通过引入 Token Pruning Cluster (TPC) 和 Token Recovering Attention (TRA)，在保持高精度的同时显著减少浮点运算次数（FLOPs），使这些模型在资源受限设备上的部署成为可能。

单人姿态估计算法大多使用热力图的方法进行关键点回归与识别，缺陷是对算力的要求较高。表 1 统计了当前主流单人姿态估计算法的相关性能。主流算法重点围绕热图积分和模型轻量化展开，并聚焦于提升模型的精度的同时降低模型的参数。

表 1 单人姿态估计算法性能对比

模型名称	算法特点	主要优缺点
DeepPose	将人体姿态估计问题转化为回归问题，通过级联的多个 DNN 网络，直接回归关节点的坐标值。	第一个端到端姿态估计算法，精度较高；鲁棒性较差，计算资源需求高。
Soft-argmax	基于热力图积分的回归方法，首次将人体关键点与热点图关联。	改善关键点回归的准确率和置信度；多个峰值时易出现错误选择峰值。
CPM	通过多阶段的序列卷积来生成热力图预测关键点信息。	精准标识图像中的人体关键点；多人场景下的检测效果并不好。
Hourglass	使用全卷积神经网络和对称的沙漏网络结构，多尺度特征捕捉人体关键点。	提高人体姿态估计的准确性；消耗大量计算资源，对小数据集不友好。
HRNet	结合高分辨率特征图和低分辨率特征图，实现特征信息的深度融合。	兼顾高低分辨率信息；网络复杂，需要大量的计算资源。
LiteHRNet	保持高分辨率输出，通过随机交换模块和条件通道加权模块实现高效信息交互。	模块化设计，模型可以根据不同的需求进行组合；模型精度降低了。
LitePose	通过融合反卷积头并利用大尺寸卷积核优化姿态估计性能。	显著减少模型参数量，同时保持较高姿态估计准确率；复杂场景下的检测精度较低
Graph-PCNN	引入图卷积神经网络（GCN），通过图姿态精化模块优化关键点位置。	提高定位精度；性能较差。
TokenPose	利用 Transformer 进行骨骼关键点的估计和回归，通过自注意力机制实现关键点坐标的精准回归。	关注全局上下文，最后收敛至单个关键点位置；注意力机制导致模型效果较差。
HourglassTokenizer	引入 Token Pruning Cluster (TPC) 和 Token Recovering Attention (TRA)，能够在保持高精度的同时显著减少浮点运算次数。	可以无缝集成到 Transformer 模块中，解决 Transformer 模型复杂的问题；3D 姿态估计，对数据集要求较高。

3 多人姿态估计算法

多人姿态估计 (Multi-Person Pose Estimation) 是计算机视觉中的一个重要任务，旨在从图像或视频中检测和识别图像中的多个人体，并为每个人体预测关键点的位置，推断出他们的姿势。单人姿态估计为多人场景奠定了基础，但其直接扩展面临人体重叠、遮挡及计算效率的严峻挑战。与单人姿态估计不同，多人姿态估计不仅需要识别和追踪多个人体，还需要解决人体之间的重叠、遮挡以及不同人的相对位置等问题。多人姿态估计以单人姿态估计算法为基础，可分为两类方法：自顶向下（Top-Down）和自底向上（Bottom-Up）。自顶向下的方法首先定位单个人，然后定位单人的关节，最后合并为整体的姿态。自底向上的方法则将图片或视频中的所有人视作一个整体提取关键点，最后通过聚类和分组将关键点与单人进行绑定，从而实现姿态估计。

自顶向下的方法在定位图像或视频中的人后，逐步细化到各个部位，最终计算出整体姿势。自顶向下的模型通常分为两个步骤，首先使用行人检测器（如 R-CNN[14]系列、YOLO[15]系列、SSD[16]等）定位人的位置，然后使用关键点检测模型进行关键点检测。

He 等[17]在行人检测器 Faster R-CNN [18]的基础上提出一种基于 Mask R-CNN 的多人姿态检测算法，增加了一个用于单人姿态估计的网络分支。通过这种联合优化方法，可以统一处理和优化行人检测和单人姿态估计任务。

上海交通大学研究团队[19]提出一种区域多人姿态估计框架 AlphaPose，即区域多人姿态检测（RMPE）框架。采用自顶向下的方法，主要包括对称空间变换网络（SSTN）、参数化姿态非极大值抑制（PP-NMS）和姿态引导区域框生成器（PGPG）。其中，SSTN 能在不精准的区域框中提取高质量的人体区域，PP-NMS 用

于解决冗余检测问题，PGPG 用于强化训练数据。

自顶向下的方法思路清晰，具有较高的精度，特别是在处理低分辨率人体图像时表现出色，但实时性较差，每次检测都需执行单人姿态估计，且随着检测人数的增加，时间复杂度呈线性增长，计算成本显著提高。此外，由于自顶向下的模型是多个模型的组合，需要对每个模型分开训练，无法实现端到端的训练。多人场景下，遮挡和密集人群会影响多人姿态估计算法的运行速度和准确率。相对而言，自底向上的方法首先检测图像中的所有关键点，再通过聚类和分组策略确定每个人与关键点的对应关系。此方法能够一次性检测所有关键点，效率较高，且模型相对较小。CHEN 等[20]提出了级联金字塔网络（Cascaded Pyramid Networks，CPN），分为 Global Net、Refine Net 两个阶段。其中，Global Net 阶段进行粗略的人体关键点检测，目标是快速大致定位关键点。Refine Net 阶段结合全局和局部特征，对 Global Net 未能准确检测的关键点进行回归，旨在精细调整初步检测到的关键点，特别是那些被遮挡或初步检测不准的关键点。Zhe 等[21]提出 OpenPose 框架，这是一个基于深度学习的人体姿势估计库，能从图像或视频中准确检测和估计人体关键点和姿势信息，目标是实现实时、多人、准确的姿势估计。OpenPose 通过非参数表示的部分亲和场（PAFs），首次提出了关联函数，用学习的方法将身体部位与图像中的个体关联起来。OpenPose 网络大致分为三个部分：第一部分为 VGG-19[22]特征提取网络，用以从输入图像中提取高层次的特征；第二部分为关键点提取网络，通过热图（heatmap）检测人体关键点，每个关键点对应一个特征图，表示该关键点在图像中出现的概率；第三部分为 PAFs，通过 PAFs 连接检测到的关键点，形成完整的人体姿态，目的是确定关键点之间的关联，以便更好地识别完整姿态。针对 OpenPose 使用热力图提取关键点而导致计算量大，Daniil 等[23]提出 Lightweight OpenPose 模型，选择轻量级 MobileNet[24]代替 VGG19 作为特征提取网络，并通过空洞卷积提升模型感受野，减少 MobileNet 网络结构深度不足的影响。

MAJI 等[25]提出的 YOLO-Pose 算法可实现目标检测与人体姿态估计。与自顶向下的方法不同，YOLO-Pose 是一种无热图联合检测的新方法，基于流行的 YOLO 目标检测框架进行二维多人姿态估计，一次推理即可定位所有人的姿态，无需多次前向传递。

与 YOLO-Pose 类似，YOLOv5[26]也可同时实现目标检测与人体姿态估计任务，不需要独立的目标检测算法和单独的人体姿态估计网络来定位关键点。在此基础上，YOLOv7[27]、YOLOv8[28]进一步提升了人体姿态估计精度。

方晓柯等[29]提出改进 YOLOv8 的人体姿态检测算法，核心在于使用可变形卷积 DCNV2[30]替换 C2F 模块中的卷积，增强了网络的特征提取能力。同时，使用加权双向金字塔 BiFPN[31]模块替换原模型中的特征融合模块，保留小目标信息的同时，融合更多浅层信息，提高识别准确度。

尽管基于 YOLO 系列的人体姿态估计算法在精度和速度上具优势，但需优先处理人体目标检测结果，再进行相应位置的关键点回归操作，若人体目标检测结果有误，会直接影响姿态估计的准确性。罗智杰等[32]提出一种改进 YOLOv8pose 的高效检测算法。该算法通过引入 RL_SEAM[33]模块优化关键点在遮挡情况下的检测效果，并结合 C2f-Context[34]机制增强上下文信息的利用，从而提升模型对复杂姿态的识别能力。此外，算法还利用 Pose_SA 轻量化检测头提升了模型在运动姿态识别方面的效果与效率，解决了 YOLO 模型在姿态估计中的若干问题。

多人姿态估计主要依赖行人识别和关键点检测两个模块，各算法的性能特点如表 2 所示。

4 数据集与评价指标

4.1 数据集

算法性能的评估离不开高质量数据集的支撑。以下从样本规模、场景多样性及标注粒度三个方面，对比分析主流数据集的特性及其适用场景。人体姿态估计数据集可分为单人数据集和多人数据集。其中，单人数据集有 Human3.6M[36]、LSP[37]、FLIC[38]、FreeMan[39]等；多人数据集有 COCO[40]、HiEve[41]、MPII[42]、AI Challenger[43]、CrowdPose[44]、PoseTrack[45]等，各数据集的相关特性如表 3 所示。Human3.6M 数据集由 11 位演员执行 15 种日常活动组成，定义了 24 个关键点，包含约 360 万张 3D 人体姿态数据，部分数据如图 1 中 a 图所示。LSP 数据集是一个体育姿势数据集，包含运动场景下的单人图像，定义了 14 个关节点，样本数约 2000 张，图像多与体育相关，姿势较复杂。FLIC 数据集来源于好莱坞电影片段，对截图中的单个人体关节点进行标注，不包含遮挡或清晰度低的图像。FreeMan 数据集由 8 部智能手机在 10 个不同场景、27 个真实场地同步录制，总计超过 1100 万帧的视频，涵盖不同照明条件，提供 2D 和 3D 人体关键点、SMPL 参数、边界框等注解信息。

表 2 多人姿态估计算法性能对比

模型	核心内容	主要优缺点
Mask R-CNN	在行人检测器基础上增加单人姿态估计的网络分支，实现联合优化。	对行人检测任务和单人姿态估计任务进行统一处理和优化；模型复杂，计算量大，对硬件资源要求较高，难以满足实时性要求。
AlphaPose	提出 RMPE 框架，通过 SSTN、PP-NMS 和 PGPG 优化姿态估计。	在人体候选框不准确的情况下进行姿态估计；计算复杂度较高，推理速度较慢。
Cascaded Pyramid Networks (CPN)	分为 Global Net 和 Refine Net 两个阶段，结合全局和局部特征信息进行精细化调整。	提高关键点检测的准确性；模型结构较为复杂，训练和推理的计算成本较。
OpenPose	通过 PAFs 连接检测到的关键点，形成完整的人体姿态。	实现实时、多人、准确的姿态估计；但对硬件资源需求较高，推理速度在复杂场景中可能受限。
Lightweight OpenPose	使用轻量级 MobileNet 和空洞卷积，提升模型感受野，降低计算量。	提高模型的实时性和准确性；复杂场景下失去一定精度。
YOLO-Pose	基于 YOLO 目标检测框架，实现无热图联合检测。	无需多次前向传递，定位所有人姿态；对人体检测结果要求较高。
YOLOv5	无需使用独立的目标检测算法和单独的人体姿态估计网络，省略关键点后处理步骤。	兼具高效性与准确性；在极复杂的姿态估计任务中精度不足。
YOLOv7	进一步提升精度，适用于实时应用。	更强的特征提取能力，更高效的姿态估计；模型复杂度增加。
YOLO-V8pose	改进了网络结构，提高模型的效率和运行速度。	实时性较好，泛化能力强；精度有所不足，模型复杂。
改进 YOLOv8	使用可变形卷积 DCNV2 和 BiFPN 模块，引入 Sim-Attention[35]注意力机制。	增强特征提取能力和识别准确度；使用了注意力机制，模型复杂度增加。

表 3 人体姿态估计数据集

数据集	样本数量/万	关键点数量	类型	场景
Human3.6M	360	24	单人	17 种场景如讨论、抽烟、玩手机等
LSP	0.2	14	单人	运动场景
FLIC	2	10	单人	电影截取片段
FreeMan	1100	17	单人	10 类生活场景如咖啡厅庭院
COCO	20	17	多人	网络图片，不包含具体场景
HiEve	4.9	14	多人	拥挤场景
MPII	2.5	16	单人/多人	日常生活场景
AL Challenger	30	15	多人	日常生活场景
CrowdPose	2	14	多人	来源于 MSCOCO、MPII 和 AI Challenger
PoseTrack	6.6	15	多人	对 MPII 的扩展

COCO 数据集由微软构建，图像来源于谷歌、Flicker 等，分为训练集、验证集和测试集，定义了 17 个关节点，包含 20 万张图像和 25 万个标注人体,部分数据如图 1 中 b 图所示。

HiEve 数据集在 YouTube 上收集了异常场景和事件的 32 个视频序列，长度共 33 分 18 秒，分为训练和测试集的 19 和 13 个视频，涵盖机场、餐厅、室内、监狱、商场、广场、学校、车站和街道 9 个不同场景。MPII 数据集来自 YouTube 的日常生活场景，手动筛选出包含人类的画面，包含 2.5 万张图像，部分图片如图 1 中 c 图所示，定义了 16 个关节点，标注了 4 万个目标。

AI Challenger 数据集通过网络爬取日常片段，包含训练集、验证集和测试集共 30 万张图像。

图 4 可以看出，大部分数据集在关键点的数量上差异不大，其中 Human3.6M 和 FreeMan 两个数据集含有最大量的数据集，COCO 和 MPII 数据集因为数据量适中而经常被应用于深度学习算法的性能对比中，其他小型数据集主要用于其他数据集的补充验证作用，如 LSP 数据集被使用在体育领域中。

a) Human3.6 b)COCO c)MPII

图 1 数据集示例

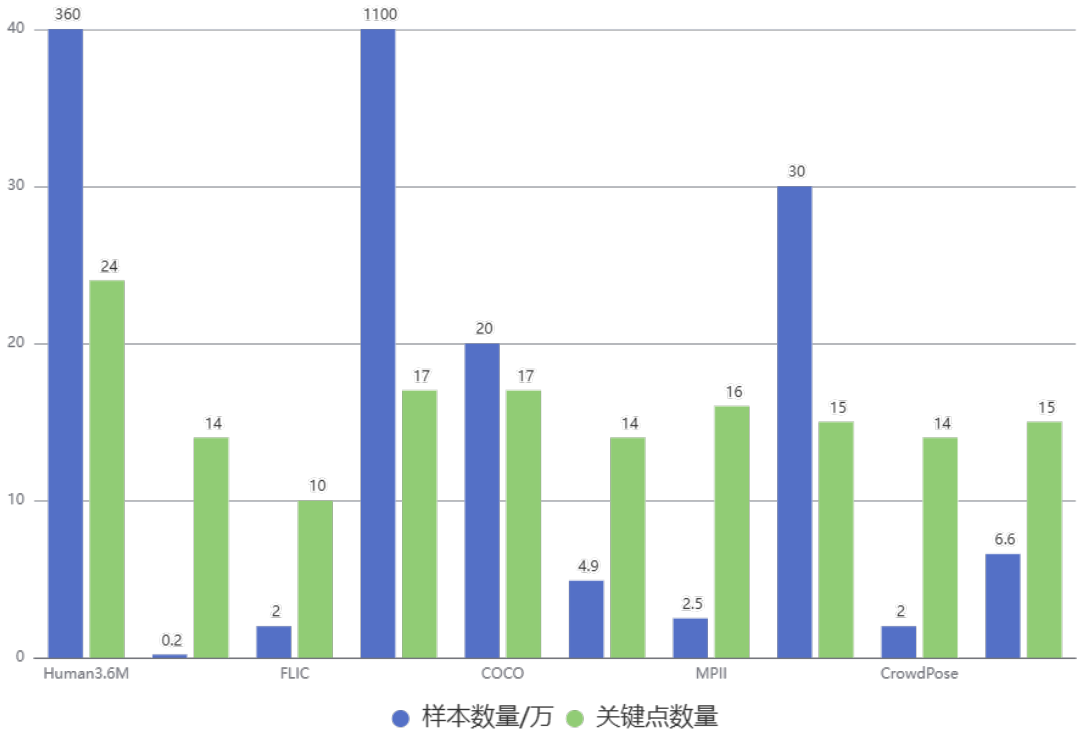

图 2 数据集对比图

4.2 评价指标

不同数据集因自身特点采用的评估指标也不同。常用的二维人体姿态估计指标主要有以下几种：

（1）关键点相似度(OKS)。计算关节点位置距离，检测关节点的相似度。

$$OKS = \frac{\sum_i \left[\exp\left(-\frac{d_i^2}{2s^2k_i^2}\right)\delta(v_i>0)\right]}{\sum_i \delta(v_i > 0)} \tag{1}$$

其中，i 为标注的关键点编号；d_i^2为检测出关键点位置与真实关键点位置的欧氏距离的平方；s^2为检测到的人体在图像中面积；k_i^2为归一化因子，表示标注关节点位移的标准差；v_i为可见关键点数量。

（2）平均精度（AP）。通过对预测值与真实值的匹配进行计算，评估每个关键点在数据集测试集上的平均准确率。

$$AP@s = \frac{\sum_p \delta(OKS > s)}{\sum_p 1} \tag{2}$$

其中，p 为人体检测框编号，AP@s为交并比，取值为 s 时对应 AP 值。

（3）部位正确估计百分比(PCP)。关节点正确估计的比例，用于评估人体关节点的定位精度。

（4）正确关键点比例(PCK)。用于衡量关键点预测准确性，当预测值与真实值距离小于阈值时为正确。

（5）平均关键点精度（APK）。将预测的人体姿态与真实姿态评估对比后，得出每个关节点定位准确的平均精度。

表 4 列出了部分姿态估计算法在 COCO2017 数据集的表现，AR 代表 OKS 阈值为 0.5 时的关键点平均召回率。

表 4 主流姿态估计算法在 COCO 数据集上的表现

Model	AP	AP50	AP75	AR
MASK R-CNN	63.1	87.3	68.7	-
AlphaPose	72.3	89.2	79.1	-
HRNet	76.3	90.8	82.9	-
CPN	73.0	91.7	80.9	79
OpenPose	61.8	84.9	67.5	66.5
Lightweight OpenPose	38.6	-	-	-
YoloV8Pose	50.4	80	54.1	57.8

5 总结与拓展

随着深度学习的快速发展，以及各领域人体行为识别日益发挥重要作用，人体姿态估计得到了快速发展，并产生一系列优秀成果和算法，并逐渐成为相关领域的研究热点，但是人体姿态估计仍存在一些问题和挑战。

（1）模型参数量较大，导致对计算能力的需求极高。目前的人体姿态估计算法通常采用复杂的模型结构，这些结构用于提取图像中的深层关键点信息，因此无法在低算力设备上直接部署，从而降低了算法的经济效益。此外，在将数据传回后端服务器进行关键点信息识别的过程中，可能还存在隐私泄露等风险。

（2）算法的检测精度与效率仍需提升。尽管现有算法已取得显著进展，但在实际人机交互、无人驾驶、视频监控等领域的广泛应用仍存在一定差距。因此，需要进一步简化网络结构并优化算法效率，同时可以引入注意力机制和图卷积网络（GCN）以增强网络的精度。

（3）算法易受环境因素影响。在实际应用中，人体姿态估计算法常常受到环境中杂物、光照和遮挡等因素的干扰。例如，针对广告牌等含有人体信息的杂物，算法难以有效区分。此外，在拥挤的人群中，由于重叠和遮挡，关键点的识别会受到干扰，容易导致漏检。另一方面，不同视角下的人体姿态也会影响关键点的识别，使得算法难以确定关键点所属的具体部位。

（4）数据集存在不足。虽然目前常用的数据集如 COCO、MPII 等已经具备相当规模，但样本分布依然不均，无法全面覆盖多样化的场景。现有数据集难以满足人体姿态变化的复杂性和多样性，尤其是在重叠人体和不同角度的人体关键点数据方面存在不足。因此，当前的数据集仍需进一步扩充以提升其质量。

自传统人体姿态估计方法逐步演进至深度学习方法，其模型与算法的性能持续提升，已在电影动画、无人驾驶、虚拟现实及智能监控等多个领域取得了显著的研究成果。基于图结构的传统方法为后续算法研究提供了宝贵的先验知识，而深度学习驱动的人体姿态估计方法无疑将引领未来的发展方向。在当前大量图像数据背景下，应充分挖掘并利用视频数据，以扩展人体姿态估计技术在更多领域的应用。作为计算机视觉众多任务的基础，二维姿态估计展现了极其广阔的研究前景。

参考文献

[1] Witkin A. Scale-space filtering: A new approach to multi-scale description[C]//ICASSP'84. IEEE international conference on acoustics, speech, and signal processing. IEEE, 1984, 9: 150-153.

[2] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). Ieee, 2005, 1: 886-893.

[3] Toshev, Alexander and Christian Szegedy. “DeepPose: Human Pose Estimation via Deep Neural Networks.”[J] 2014 IEEE Conference on Computer Vision and Pattern Recognition (2013): 1653-1660.

[4] Li J, Chen T, Shi R, et al. Localization with sampling-argmax[J]. Advances in Neural Information Processing Systems, 2021, 34: 27236-27248.

[5] Wei S E, Ramakrishna V, Kanade T, et al. Convolutional pose machines[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2016: 4724-4732.

[6] Xu T, Takano W. Graph stacked hourglass networks for 3d human pose estimation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 16105-16114.

[7] Wang J, Sun K, Cheng T, et al. Deep high-resolution representation learning for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2020, 43(10): 3349-3364.

[8] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[9] Yu C, Xiao B, Gao C, et al. Lite-hrnet: A lightweight high-resolution network[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 10440-10450.

[10] Wang Y, Li M, Cai H, et al. Lite pose: Efficient architecture design for 2d human pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 13126-13136.

[11] Wang J, Long X, Gao Y, et al. Graph-pcnn: Two stage human pose estimation with graph pose refinement[C]// Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. Springer International Publishing, 2020: 492-508.

[12] Li Y, Zhang S, Wang Z, et al. Tokenpose: Learning keypoint tokens for human pose estimation[C]//Proceedings of the IEEE/CVF International conference on computer vision. 2021: 11313-11322.

[13] Li W, Liu M, Liu H, et al. Hourglass Tokenizer for Efficient Transformer-Based 3D Human Pose Estimation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 604-613.

[14] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[15] Wang A, Chen H, Liu L, et al. Yolov10: Real-time end-to-end object detection[J]. arXiv preprint arXiv:2405.14458, 2024.

[16] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.

[17] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]// Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[18] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(6): 1137-1149.

[19] Fang H S, Li J, Tang H, et al. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(6): 7157-7173.

[20] Chen Y, Wang Z, Peng Y, et al. Cascaded pyramid network for multi-person pose estimation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7103-7112.

[21] Cao Z, Simon T, Wei S E, et al. Realtime multi-person 2d pose estimation using part affinity fields[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7291-7299.

[22] Simonyan K. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[23] Osokin D. Real-time 2d multi-person pose estimation on cpu: Lightweight openpose[J]. arXiv preprint arXiv:1811.12004, 2018.

[24] Howard A G. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[25] Maji D, Nagori S, Mathew M, et al. Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 2637-2646.

[26] Nguyen H C, Nguyen T H, Scherer R, et al. Unified end-to-end YOLOv5-HR-TCM framework for automatic 2D/3D human pose estimation for real-time applications[J]. Sensors, 2022, 22(14): 5419.

[27] 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[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 7464-7475.

[28] 傅裕,高树辉.改进 YOLOv8s-Pose 多人姿态估计轻量化模型研究[J/OL].计算机科学与探索,1-17[2025-01-16].

http://kns.cnki.net/kcms/detail/11.5602.TP.20240507.1148.002.html.

[29] 方晓柯,黄俊.基于 yolov8-pose 的人体姿态检测模型[J/OL].激光杂志,1-9[2025-01-17]. http://kns.cnki.net/kcms/detail/50.1085.tn.20240902.1533.007.html.

[30] Zhu X, Hu H, Lin S, et al. Deformable convnets v2: More deformable, better results[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 9308-9316.

[31] Doherty J, Gardiner B, Kerr E, et al. BiFPN-YOLO: One-stage object detection integrating Bi-Directional Feature Pyramid Networks[J]. Pattern Recognition, 2025, 160: 111209.

[32] 罗智杰,王泽宇,岑飘,等.基于改进 YOLOv8pose 的校园体测运动姿势识别研究[J].电子测量技术,2024,47(19): 24-33.

[33] Yu Z, Huang H, Chen W, et al. Yolo-facev2: A scale and occlusion aware face detector[J]. Pattern Recognition, 2024, 155: 110714.

[34] Wu T, Tang S, Zhang R, et al. A light-weight context guided network for semantic segmentation., 2020, 30[J]. DOI: https://doi. org/10.1109/TIP, 2020: 1169-1179.

[35] Yu H, Wan C, Liu M, et al. Real-Time Image Segmentation via Hybrid Convolutional-Transformer Architecture Search[J]. arXiv preprint arXiv:2403.10413, 2024.

[36] Ionescu C, Papava D, Olaru V, et al. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 36(7): 1325-1339.

[37] Johnson S, Everingham M. Clustered pose and nonlinear appearance models for human pose estimation[C]//bmvc. 2010, 2(4): 5.

[38] Sapp B, Taskar B. Modec: Multimodal decomposable models for human pose estimation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 3674-3681.

[39] Wang J, Yang F, Gou W, et al. Freeman: Towards benchmarking 3d human pose estimation in the wild[J]. arXiv preprint arXiv:2309.05073, 2023.

[40] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014: 740-755.

[41] Lin W, Liu H, Liu S, et al. HiEve: A large-scale benchmark for human-centric video analysis in complex events[J]. International Journal of Computer Vision, 2023, 131(11): 2994-3018.

[42] Andriluka M, Pishchulin L, Gehler P, et al. 2d human pose estimation: New benchmark and state of the art analysis[C]//Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. 2014: 3686-3693.

[43] Wu J, Zheng H, Zhao B, et al. Ai challenger: A large-scale dataset for going deeper in image understanding[J]. arXiv preprint arXiv:1711.06475, 2017.

[44] Li J, Wang C, Zhu H, et al. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 10863-10872.

[45] Andriluka M, Iqbal U, Insafutdinov E, et al. Posetrack: A benchmark for human pose estimation and tracking[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 5167-5176.