综述：基于深度学习的三维人体姿态估计技术

姓名：沈韵泓　学院：工程师学院　学号：22460037

1 问题引入

人体姿态估计（Human Pose Estimation，HPE）旨在从输入的图像或视频数据中定位人体部位、并构建人体表示。其预测结果提供了人体的几何与运动信息，在人机交互、运动分析、增强现实与虚拟现实等领域得到广泛应用。

基于带有 2D 姿态标注的数据进行 2D-HPE 是较为容易实现的。传统的 2D-HPE 技术依赖手工特征提取，这类工作将人体描述为骨架模型以实现对关键点的二维或空间位置估计。

相较之下，3D-HPE 要困难得多：尽管我们可以通过运动捕捉系统在受控的实验室环境中收集 3D 姿态标注信息，但在野外环境中仍存在较大的局限性、难以实现准确标注。此外，由于在 3D 场景投影至 2D 图像过程中的深度信息丢失，基于单目输入的 3D-HPE 面临着“深度模糊”的挑战；而在多视图场景中，“深度模糊”和“自遮挡”问题均可通过三角互证法（triangulation）进行处理，但又引入了多视点关联的挑战。

随着机器学习算法的发展，3D-HPE 的解决方案朝着基于卷积神经网络、乃至基于多种深度学习模型混合的方向演进。各种基于概率性多假设及基于扩散模型的方法不断涌现，研究内容也从单人场景拓展至多人复杂场景。尽管上述方法取得了显著的进展与卓越的性能，深度模糊、遮挡、训练数据不足等挑战仍有待克服。

我们将在第 2 节中从多个角度对现有基于深度学习的 3D-HPE 方法进行分类梳理，在第 3 节中对 3D-HPE 常用的数据集与评估指标进行介绍，最后在第 4 节中对现有方法机及其挑战进行总结、并从自身出发对该领域研究的未来发展进行展望。

2 方法总结

3D-HPE 可以通过预测人体关键点的深度信息、提供比 2D-HPE 更加精确的姿态信息。如图 2.1 所示，我们可以根据需要识别的目标数量，将 3D-HPE 分为单人三维姿态识别和多人三维姿态识别两大类。即便预测精度已随深度学习技术的快速发展得到大幅提升，单人任务的执行效率与自遮挡问题、多人任务中的相互遮挡问题，仍待解决。此外，缺少经专业标

注的训练数据也为相关研究带来了极大的挑战。本节将详细介绍该领域的研究进展。

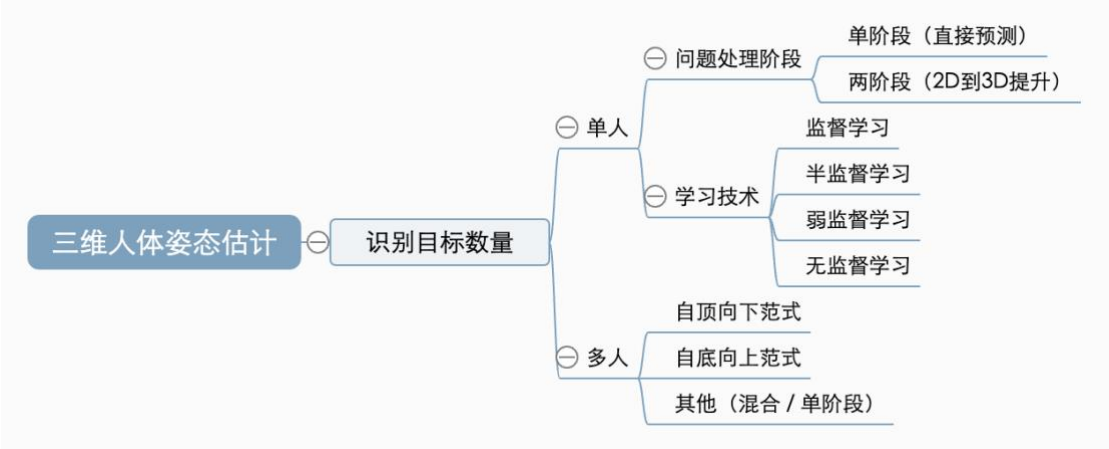


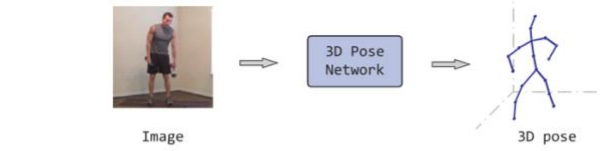
图 2.1 三维人体姿态估计方法分类

2.1 单人三维人体姿态估计

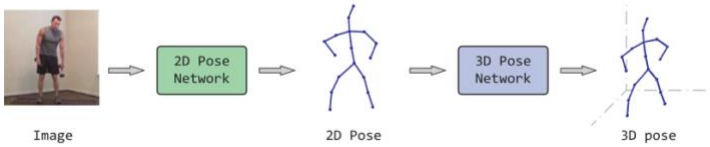
我们可以从多个维度对单人 3D-HPE 技术进一步细分。在本节中，我们将从“问题解决阶段”和“学习策略”两个维度对现有的单人 3D-HPE 技术进行分类介绍。

2.1.1 根据问题处理阶段划分

我们可以按照“问题解决阶段”将现有的单人 3D-HPE 方法划分为单阶段（直接预测）和两阶段（2D 到 3D 提升）两种范式，其基本处理流程如图 2.2 所示。



(a) The direct estimation method



(b) The 2D to 3D lifting method

图 2.2 两种经典的单人三维人体姿态估计范式[1]

（一） 单阶段（直接预测）范式

单阶段范式旨在通过坐标回归预测、“图像-姿态”间最近邻匹配或、在姿态类别组上进行分类等方法直接从 2D 输入数据中估计 3D 姿态。Li 和 Chan[2]通过基于多任务学习框架，同步对人体关键点检测器和姿态回归预测器进行训练。为弥补基于回归方法的不足，Sun 等

人[3]基于骨架型人体表示提出了结果感知回归方法，并提出一种用于量化 3D 骨骼间长距离交互的组合损失。

单阶段范式的一个显著优势在于：实际应用中的 2D 姿态估计结果往往存在误差、将显著影响后续的 3D 姿态估计阶段，而单阶段范式的预测最终结果与中间 2D 估计无关。然而，缺少对中间估计结果的监督也使得单阶段方法容易受背景和环境光照变化影响。此外，由于单阶段方法的训练严重依赖于标注数据，野外环境下大规模标注数据集的缺失使得单阶段方法的泛化性能往往较差。从解空间角度考虑，直接基于图像进行 3D-HPE 是一个高度非线性的过程、全局最优解的搜索空间大，这将导致回归模型必须具有较大的参数量。

（二）两阶段（2D 到 3D 提升）范式

得益于 2D-HPE 技术的高可用性与轻量性，基于 2D 中间结果的两阶段范式已经成为 3D-HPE 的主流解决方案。该范式在第一阶段采用现成的 2D-HPE 模型基于输入图像生成 2D 姿态特征或 2D 姿态估计结果，并以此作为中间输入、在第二阶段的 2D 到 3D 提升中重建 3D 姿态。

一些研究工作采用 2D 姿态作为中间结果。Martinez 等人[4]在早年提出使用全连接残差网络实现基于 2D 关节位置回归预测 3D 关节位置，在当时获得了最优结果。Liu 等人[5]则采用基于注意力机制的时间卷积神经网络捕捉长期时间以来关系，并通过多尺度扩张卷积结构适应变长输入序列。Zhan 等人[6]提出的 Ray3D 则开创性地将输入从 2D 像素转换为规范化坐标系中的 3D 射线，缓解由相机内参和俯仰角变化导致的误差。

另一些研究工作则采用捕捉了 2D 姿态上下文信息的 2D 姿态感知特征作为中间结果。Tekin 等人[7]和 Zhou 等人[8]采用 2D 热图作为中间表示，以提取关键点的深度信息、降低 2D 到 3D 提升过程中的不确定性。Kim 等人[9]则利用 2D 姿态关节置信度来创建伪标签，通过置信度计算每个关节的加权平均以生成 3D 伪真实值、用于自监督学习。Kundu 等人[10]和 Usman 等人[11]则将 2D 热图与 2D 姿态置信度相结合，通过高置信度关节来校正低置信度、不确定关节的位置，增强整体预测准确性。

由于 2D-HPE 技术提供的 2D 关节估计能良好的与 3D 姿态在空间上对齐，两阶段范式在图像失真（如，背景、任务服装变化等）情形下具有鲁棒性。此外，两阶段范式能够利用现有的大规模 2D 姿态数据集，减轻了对小型 3D 数据集上产生过拟合的风险。然而，2D 到 3D 提升阶段也引入了一些新的挑战：如何处理变长数据以及不同数量的视图、如何处理相机校准问题仍待解决。

2.1.2 根据学习策略划分

目前的 3D-HPE 方法大多遵循监督学习范式，需要大量数据以确保模型的鲁棒性和泛化能力。然而，获取 3D 姿态数据既昂贵又耗时，通常需要多视图设置或动作捕捉系统，在野外场景中显得不切实际。为应对野外 3D 姿态数据稀缺的困境，研究者们采用了半监督、弱监督等学习技术。

（一） 半监督学习

半监督学习（Semi-Supervised Learning，SSL）旨在通过少量标注数据和大量的未标注数据共同训练模型。其中，标注数据提供了明确的监督信号，而大量未标注数据则弥补了标记数据不足的挑战。Pavllo 等人[^e]提出了一种基于循环一致性的 SSL 技术，他们在未标注数据上应用了 2D 到 3D 提升技术，并将最终的 3D 姿态映射回 2D 空间。Mitra 等人\[^e\]引入了一种基于多视图一致性度量的 SSL 方法、使模型无需再对 2D 姿态这一中间结果进行预测，同时还通过有限监督信号促进了网络对姿态特定特征的学习。

（二） 弱监督学习

弱监督学习（Weakly Supervised Learning，WSL）涉及使用不完全标注的数据（如，标签包含噪声、不完整或不准确），通过开发对标签缺陷具有鲁棒性的模型来解决标签质量差的问题。Chen 等人[[12]]提出的演绎弱监督学习（Deductive WSL）方法使用了未校准的摄像机和多视图 2D 人体姿态数据。通过学习深度和摄像机姿态的潜在表示、结合演绎推理，从不同视图推断人体姿态。Qiu 等人[[13]]则提出了一种基于预训练的 WSL 方法，最初在 2D 数据集上进行预训练、随后在 3D 数据集上进行微调，支持在无标注的情景下提取弱 3D 信息。

（三） 无监督学习

无监督学习旨在挖掘无标注数据中的隐藏模式。Xu 和 Takano [14]提出了一种无监督的师生网络模型。其中，教师网络将来自不同视图的 2D 姿态对齐到 3D 姿态，利用循环一致架构确保旋转不变性。学生网络也遵循循环一致架构，以确保输入视图中的姿态估计一致性并具有旋转等变性，通过几何自监督增强其训练。Chai 等人[15]提出了一种无监督的域适应框架，通过全局位置对齐模块弥合 2D-HPE 预训练所用的数据集与需要进行 3D-HPE 的目标数据集之间的差异。

2.2 多人三维人体姿态估计

多人三维人体姿态估计通常基于单目图像输入开展。其经典范式与多人二维人体姿态估

计一致，可被大致划分为自顶向下、自底向上两类（如图 2.3 所示）。在通常情况下，自顶向下范式拥有更高的准确性，而自底向上范式则具有更高的处理效率。为同时利用两者的优势，一些整合了两种范式的混合方法得到提出。近年来，不同于以上的两阶段范式，一些单阶段的解决方案也得到探究。

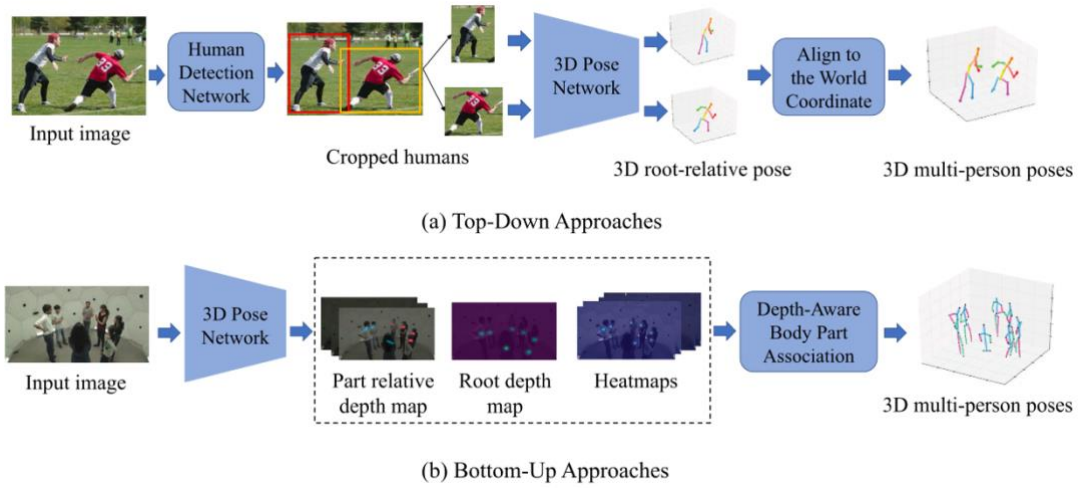


图 2.3 两种经典的多人三维人体姿态估计范式[16]

2.2.1 自顶向下范式

自顶向下范式首先识别输入图像中每个人体的包围盒（bounding box），以分辨不同个体。随后，针对每个检测到的人体，利用神经网络预测其绝对根节点（即人体的中心关节点）坐标与相对于根节点的三维姿态。并最终基于每个个体的绝对根节点坐标及其根节点相对姿态，将所有姿态对齐至世界坐标系。

LCR-Net[17]提出了“定位-分类-回归”的处理流程，即：先定位每个人体的候选区域、生成潜在的姿态，随后使用回归器优化姿态估计提案。该方法在环境可控的数据集上表现良好，但在野外数据集中不能很好地泛化。为了解决这个问题，该方法的研究人员又提出了通过数据合成以实现数据增强、进而提升预测性能的 LCR-Net++[18]。

AlphaPose[19]框架预测了包括面部、身体、手和脚在内的全身多人 3D 人体姿态。该框架通过对称积分关键点回归模块实现对人体关键点的快速精确定位，通过参数姿态非最大抑制模块消除冗余包围盒，通过姿态感知身份嵌入模块实现联合姿态估计和跟踪。

多人场景往往面临着比单人更严重和复杂的遮挡问题。为应对这一挑战，Wu 等人[20]提出了一种基于图的多人多视图 3D-HPE 方法，利用 GNN 提高了信息传递效率。该方法利用多视图匹配图模块关联粗略的跨视图姿态，随后通过中心细化图模块进一步优化结果。其单次学习开销显著低于典型的监督学习模型。

此外，选择合适的表征形式可以有效提升估计精确度。Zhang 等人[21]提出的 Voxeltrack 模型使用了体素（voxel）表示，能确定每个体素是否包含特定的身体关键点。Benzine 等人[22]提出的 PandaNet 模型则引入了一种低分辨率的锚点（anchor）表示，通过姿态感知锚点选择策略丢弃模糊锚点以解决重叠问题。

自顶向下的多人 3D-HPE 方法具有极高的精确度，适用于对准确度要求较高的场景。然而，此类方法严重依赖于人体检测方法和单人 HPE 方法的准确性[23]，且无法检测到大部分被遮挡的个体。且仅在包围盒中进行预测可能会忽略场景中的全局信息，包围盒深度的错误估计可能导致预测的人体骨架被放置在重叠的空间位置上。此外，它们的计算复杂性随着人数的增加而增加，不适合需要实时处理的拥挤场景。

2.2.2 自底向上范式

自底向上范式不需要进行人检测，可以同时对多人姿态进行估计。该类方法一次性生成所有身体关键点和深度图，允许在存在强遮挡的情形下正常进行姿态推断[22]。随后，根据根节点深度与其余节点的相对深度将关键点分配给每一个人、组装成完整的人体骨架。

因此，如何正确分配属于不同个体的关键点是自底向上范式面临的一个关键挑战：Zanfir 等人[24] 将该问题表述为 0-1 整数规划问题（Binary Integer Program，BIP），通过肢体评分模块来估计关键点的候选邻接点，随后求解 BIP 以将关键点组装成骨架。Fabbri 等人[25]提出了一种基于距离的启发式方法，从检测到的具有最高置信度的关键点开始，选择具有最小欧氏距离的关键点进行连接。

自底向上方法面临的另一个仍是遮挡。为了应对这一挑战，Metha 等人[26] 提出通过遮挡鲁棒姿态图（Occlusion-Robust Pose-Map，ORPM），通过将冗余纳入位置图以增强关键点关联，实现在强部分遮挡下的全身姿态估计。Zhen 等人[27] 则利用图像中的深度线索估计个体的绝对位置，通过推理人物间遮挡和骨长约束将关键点分配给正确个体。

自底向上方法具有线性的计算和时间复杂度。然而，目前的大多数自底向上方法以单一尺度处理所有个体，使其对多个人物间的尺度变化敏感，但难以检测较小个体的关键点。

2.2.3 其他方法

为了充分利用两种范式的优势，Cheng 等人[28]提出了基于互补双网络的方法：首先通过自顶向下网络在每个包围盒内估计关节热图，再通过自底向上网络结合热图来处理尺度变化，最终输入集成网络以获得最终的 3D 姿态。该方法在存在包围盒错误预测与多尺度变化的场景下，展现出比任一单独范式更强的鲁棒性。

Jin 等人[29]认为当前的各种两阶段方法存在冗余计算和高计算成本的问题、实时处理效率不足，因此引入了一种单阶段的解耦回归模型。该方法提出在图像平面中解耦表示 2D 姿态，并通过 2D 姿态特征和深度回归预测深度信息和个体的尺度信息。此外，Wang 等人[30]也提出了一种基于分布感知，递归增强关键点估计的单阶段模型。

3 三维人体姿态估计评估

3.1 常用数据集

数据集对于基于深度学习的 3D-HPE 方法来说是不可或缺的。无论是单一模型的训练，还是对多种方法的公平比较都离不开数据集。表 3.1 中列出了近期 3D-HPE 数据集的摘要信息，其中最常用的部分将在本节中进行更详细的介绍。

表 3.1 现有的 3D-HPE 数据集

数据集	总帧数	单人	多人	单视图	多视图	描述
Human3.6M[31]	3.6M	○		○	○	
3DPW[32]	51k	○	○	○		
MPI-INF-3DHP[33]	2k	○		○	○	户外数据
HumanEva[34]	40k	○		○	○	
CMU-Panoptic[35]	1.5M	○	○	○	○	
MuCo-3DHP[26]	8k		○	○		遮挡场景
SURREAL[36]	6M	○		○		
UP-3D[37]	8k	○		○		
TotalCapture[38]	1.9M	○		○		
MuPoTS-3D[26]	8k	○	○	○	○	户外数据
AMASS[39]	9M	○		○	○	户外数据
GTA-IM[40]	1M	○		○		游戏引擎
Occlusion-Person[41]	73k	○		○	○	遮挡场景

Human3.6M[31]是 3D-HPE 研究领域规模最大、使用最广泛的数据集，包含来自四个视角的共计 360 万帧 RGB 和红外图像数据。该数据集记录了 12 名受试者（六男五女）在室内执行的 15 种不同日常活动（如，吃饭、走路、打电话等）。

MPI-INF3DHP[33]是首个使用无标记动作捕捉系统、而非传统标记系统收集的数据集，提供了超过 2 千帧视频数据。该数据集记录了来自 8 位受试者在室内和室外执行 8 种不同活动（如，走路、坐着、进行复杂锻炼等）的 13 个数据人体关键点标注信息。

HumanEva[34]是多视图 3D-HPE 数据集，分为 HumanEva-I 和 HumanEva-II 两部分。其中，HumanEva-I 体量更大、包含从七个视角捕获的超过 4 万帧数据，涉及的 6 种不同活动（如，走路、拳击、慢跑、投掷-接球）由 4 位受试者执行。相较之下，HumanEva-II 规模较小、但具有更高质量的动作捕捉质量，仅包含 2 千多帧数据、由 2 位受试者执行。

3.2 评估指标

平均关节位置误差（Mean Per Joint Position Error，MPJPE）是 3D-HPE 研究中使用最为广泛的指标，量化了估计 3D 关节坐标与其真实坐标之间的平均欧氏距离：

$$MPJPE = \frac{1}{N}\sum_{i=1}^{N} |J_i - J_i^*|_2 \qquad (3-1)$$

其中，N 是骨架模型中的关节总数；J_i、J_i^* 分别为第 i 个关节点 3D 坐标的真实值与估计值，单位为毫米（mm）。

MPJPE 提供了关节 3D 位置的绝对误差，但为忽略全局变换影响、聚焦于姿态结构本身的准确性，研究人员又提出了 MPJPE 的两种变体。其中，**归一化平均关节位置误差**（Normalized MPJPE，N-MPJPE）在计算前将预测骨架放缩到与真实骨架一致；**普氏对齐平均关节位置误差**（Procrustes Aligned MPJPE，PA-MPJPE）也称“重建误差”，则在计算前基于 Procrustes 分析对预测骨架进行最佳比例的刚体变换（包括旋转、平移和缩放）。两者的计算公式定义与公式（3-1）中一致。

平均关节空间角误差（Mean Per Joint Angle Error，MPJAE）则侧重于量化估计 3D 关节与真实关节间的空间角度误差，在三维空间中的计算如下：

$$\text{MPJAE} = \frac{1}{3N}\sum_{i=1}^{3N} |(r_i - r_i^*)\ mod \pm 180°| \qquad (3-2)$$

其中，r_i、r_i^* 分别为第 i 个关节点空间角的真实值与估计值，单位为度（°）。

平均关节定位误差（Mean Per Joint Localization Error，MPJLE）是一种比 MPJPE 和 MPJAE 更加敏感和鲁棒的量化指标，支持通过阈值参数 t 以调整容差水平：

$$MPJLE = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{|l_i - l_i^*|_2 \geq t} \quad (3-3)$$

3D 正确关节百分比（3D Percentage of Correct Keypoints，3D-PCK）是二维评估指标的扩展版本，统计了预测关节点中与实际关节坐标间欧氏距离小于指定阈值T（一般为 155 毫米）的比例：

$$3D - PCK = \frac{1}{N}\sum_{p=1}^{N} \delta\left(\frac{d_p^i}{d_p} < T\right) \quad (3-4)$$

其中，d_p 为第 p 个人的归一化因子、一般取头部直径，d_p^i 为第 i 个关节点预测值与真实值间的三维欧氏距离。

4 本人对问题的理解

在前面的内容中，我们首先简要介绍了 3D-HPE 任务，随后分别对基于深度学习的 3D-HPE 技术、数据集和评估标准进行归类介绍。即便研究人员已为应对 3D-HPE 中的挑战付出大量努力、并在多方面取得了突破性进展，但 3D-HPE 场景的多样性仍使得该方面研究充满挑战。下面我们将分点概述：

处理速度 在算法部署阶段，处理速度是一个需要着重考虑的维度。即便当前的大部分方法能在 GPU 上达到实时处理，但多数应用场景涉及的还是手机等边缘计算平台、而智能手机采用的 ARM 处理器和 GPU 间存在显著的性能差距。如何优化算法速度，在边缘计算平台上实现对 3D 场景的实时高效处理仍待解决。

可控性和动画性 在虚拟现实的相关工作中，“可控性”指精确控制模型的姿态、动作轨迹和面部表情，使其能在虚拟环境中执行预期行为；“动画性”则指模型能够生动流畅地执行指定动作与表情。如何使 3D-HPE 在连续帧中提取到更加流畅的动作、更加细致的表情，使得开发人员能够便捷地创建真实生动、富有表现力的数字角色，增强其在虚拟世界中的互动性和吸引力仍需进一步探究。

对抗攻击 被多数 3D-HPE 方法采用的深度神经网络结构易受对抗攻击（Adversarial Attacks），各类人眼不可察觉的噪声可能会显著降低 3D-HPE 的性能，而当前的研究工作很少能顾及这一点。构建鲁棒的、能够防御对抗攻击的 HPE 模型，能够使其在实际应用中更加安全可靠、推动姿态估计的相关技术落地。

参考文献

[1] Liu Y, Qiu C, Zhang Z. Deep learning for 3D human pose estimation and mesh recovery: A survey[J]. Neurocomputing, 2024: 128049.

[2] Li S, Chan A B. 3d human pose estimation from monocular images with deep convolutional neural network[C]. Computer Vision--ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part II 12. Springer International Publishing, 2015: 332-347.

[3] Sun X, Shang J, Liang S, et al. Compositional human pose regression[C]. Proceedings of the IEEE international conference on computer vision. 2017: 2602-2611.

[4] Martinez J, Hossain R, Romero J, et al. A simple yet effective baseline for 3d human pose estimation[C]. Proceedings of the IEEE international conference on computer vision. 2017: 2640-2649.

[5] Liu R, Shen J, Wang H, et al. Enhanced 3D human pose estimation from videos by using attention-based neural network with dilated convolutions[J]. International Journal of Computer Vision, 2021, 129: 1596-1615.

[6] Zhan Y, Li F, Weng R, et al. Ray3D: ray-based 3D human pose estimation for monocular absolute 3D localization[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 13116-13125.

[7] Tekin B, Katircioglu I, Salzmann M, et al. Structured prediction of 3d human pose with deep neural networks[J]. arXiv preprint arXiv:1605.05180, 2016.

[8] Zhou K, Han X, Jiang N, et al. Hemlets pose: Learning part-centric heatmap triplets for accurate 3d human pose estimation[C]. Proceedings of the IEEE/CVF international conference on computer vision. 2019: 2344-2353.

[9] Kim H W, Lee G H, Nam W J, et al. MHCanonNet: Multi-Hypothesis Canonical lifting Network for self-supervised 3D human pose estimation in the wild video[J]. Pattern Recognition, 2024, 145: 109908.

[10] Kundu J N, Seth S, YM P, et al. Uncertainty-aware adaptation for self-supervised 3d human pose estimation[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 20448-20459.

[11] Usman B, Tagliasacchi A, Saenko K, et al. Metapose: Fast 3d pose from multiple views without 3d supervision[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 6759-6770.

[12] Chen X, Wei P, Lin L. Deductive learning for weakly-supervised 3d human pose estimation via uncalibrated cameras[C]. Proceedings of the AAAI conference on artificial intelligence. 2021, 35(2): 1089-1096.

[13] Qiu Z, Qiu K, Fu J, et al. Weakly-supervised pre-training for 3D human pose estimation via perspective knowledge[J]. Pattern Recognition, 2023, 139: 109497.

[14] 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.

[15] Chai W, Jiang Z, Hwang J N, et al. Global adaptation meets local generalization: Unsupervised domain adaptation for 3d human pose estimation[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 14655-14665.

[16] Zheng C, Wu W, Chen C, et al. Deep learning-based human pose estimation: A survey[J]. ACM Computing Surveys, 2023, 56(1): 1-37.

[17] Rogez G, Weinzaepfel P, Schmid C. Lcr-net: Localization-classification-regression for human pose[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 3433-3441.

[18] Rogez G, Weinzaepfel P, Schmid C. Lcr-net++: Multi-person 2d and 3d pose detection in natural images[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 42(5): 1146-1161.

[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] Wu S, Jin S, Liu W, et al. Graph-based 3d multi-person pose estimation using multi-view images[C]. Proceedings of the IEEE/CVF international conference on computer vision. 2021: 11148-11157.

[21] Zhang Y, Wang C, Wang X, et al. Voxeltrack: Multi-person 3d human pose estimation and tracking in the wild[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(2): 2613-2626.

[22] Benzine A, Chabot F, Luvison B, et al. Pandanet: Anchor-based single-shot multi-person 3d pose estimation[C].

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 6856-6865.

[23] Han C, Yu X, Gao C, et al. Single image based 3D human pose estimation via uncertainty learning[J]. Pattern Recognition, 2022, 132: 108934.

[24] Zanfir A, Marinoiu E, Zanfir M, et al. Deep network for the integrated 3d sensing of multiple people in natural images[J]. Advances in neural information processing systems, 2018, 31.

[25] Fabbri M, Lanzi F, Calderara S, et al. Compressed volumetric heatmaps for multi-person 3d pose estimation[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 7204-7213.

[26] Mehta D, Sotnychenko O, Mueller F, et al. Single-shot multi-person 3d pose estimation from monocular rgb[C]. 2018 International Conference on 3D Vision (3DV). IEEE, 2018: 120-130.

[27] Zhen J, Fang Q, Sun J, et al. Smap: Single-shot multi-person absolute 3d pose estimation[C]. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16. Springer International Publishing, 2020: 550-566.

[28] Cheng Y, Wang B, Tan R T. Dual networks based 3d multi-person pose estimation from monocular video[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(2): 1636-1651.

[29] Jin L, Xu C, Wang X, et al. Single-stage is enough: Multi-person absolute 3D pose estimation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 13086-13095.

[30] Wang Z, Nie X, Qu X, et al. Distribution-aware single-stage models for multi-person 3D pose estimation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 13096-13105.

[31] 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.

[32] Von Marcard T, Henschel R, Black M J, et al. Recovering accurate 3d human pose in the wild using imus and a moving camera[C]. Proceedings of the European conference on computer vision (ECCV). 2018: 601-617.

[33] Mehta D, Rhodin H, Casas D, et al. Monocular 3d human pose estimation in the wild using improved cnn supervision[C]. 2017 international conference on 3D vision (3DV). IEEE, 2017: 506-516.

[34] Sigal L, Balan A O, Black M J. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion[J]. International journal of computer vision, 2010, 87(1): 4-27.

[35] Joo H, Liu H, Tan L, et al. Panoptic studio: A massively multiview system for social motion capture[C]. Proceedings of the IEEE international conference on computer vision. 2015: 3334-3342.

[36] Varol G, Romero J, Martin X, et al. Learning from synthetic humans[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 109-117.

[37] Lassner C, Romero J, Kiefel M, et al. Unite the people: Closing the loop between 3d and 2d human representations[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 6050-6059.

[38] Trumble M, Gilbert A, Malleson C, et al. Total capture: 3D human pose estimation fusing video and inertial sensors[C]. BMVC. 2017, 2(5): 1-13.

[39] Mahmood N, Ghorbani N, Troje N F, et al. AMASS: Archive of motion capture as surface shapes[C]. Proceedings of the IEEE/CVF international conference on computer vision. 2019: 5442-5451.

[40] Cao Z, Gao H, Mangalam K, et al. Long-term human motion prediction with scene context[C]. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. Springer International Publishing, 2020: 387-404.

[41] Zhang Z, Wang C, Qiu W, et al. Adafuse: Adaptive multiview fusion for accurate human pose estimation in the wild[J]. International Journal of Computer Vision, 2021, 129: 703-718.