三维人体姿态估计中的多尺度时空特征融合

张宇[1)], 刘骊[1,2)*], 付晓东[1,2)], 刘利军[1,2)], 彭玮[1,2)]

[1)] (昆明理工大学信息工程与自动化学院 昆明 650500)
[2)] (昆明理工大学云南省计算机技术应用重点实验室 昆明 650500)
(ieall@kust.edu.cn)

摘 要: 针对视频输入的单人三维人体姿态估计中表征不精确、融合不充分、结果不平滑的问题, 提出三维人体姿态估计的多尺度时空特征融合方法. 首先在空域定义关节点、肢体和上/下身人体标记并通过位置嵌入表示人体的空间多尺度特征; 然后结合自注意力机制和多层感知机构建空间多尺度特征融合模块, 融合关节点、肢体和上/下身三个空间多尺度特征, 得到初步姿态特征序列; 最后建立时序多尺度编码进行时序特征融合获得最终姿态特征序列, 并通过时序解码, 优化生成细化的三维人体姿态. 在 Human3.6M 数据集上的实验结果表明, 所提方法的平均每关节位置 P-MPJPE 和速度误差 MPJVE 分别为 33.6 和 2.4, 较对比方法降低了 2.3%和 4.0%, 能够降低计算复杂度, 提高三维人体姿态估计精度, 生成准确、平滑的三维人体姿态估计结果. 此外, 在 HumanEva-I 数据集的测试结果表明, 所提方法也具有一定的泛化性.

关键词: 三维人体姿态估计; 多尺度特征; 自注意力机制; 时空特征融合; 时序编码
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2023-00041

Multi-Scale Spatial-Temporal Feature Fusion for 3D Human Pose Estimation

Zhang Yu[1)], Liu Li[1,2)*], Fu Xiaodong[1,2)], Liu Lijun[1,2)], and Peng Wei[1,2)]

[1)] (*Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming* 650500)
[2)] (*Yunnan Key Laboratory of Computer Technologies Application, Kunming University of Science and Technology, Kunming* 650500)

Abstract: To address the problem of inaccurate representation, inadequate fusion and unsmooth results in video-based single person three-dimensional human pose estimation, a multi-scale spatial-temporal feature fusion method is proposed. Firstly, the joint, limb and upper/lower body tokens were defined in spatial domain to represent the spatial multi-scale features of human body using positional embeddings. Secondly, the spatial multi-scale feature fusion module was constructed based on self-attention mechanism and multilayer perceptron to fuse joint, limb and upper/lower body features, obtaining initial pose feature sequence. Lastly, the temporal multi-scale encoding was established for temporal feature fusion to acquire final pose feature sequence, and optimize the generation of refined three-dimensional human pose through temporal decoding. Experimental results on Human3.6M dataset show that the mean per joint position error and joint velocity errors are 33.6 and 2.4 respectively, which reduce by 2.3% and 4.0%. The proposed method can improve three-dimensional human pose estimation accuracy and generate precise and smooth results while reducing computational cost. Furthermore, experimental results on HumanEva-I dataset show that the proposed method also has a certain degree of generalization ability.

收稿日期: 2023-04-21; 修回日期: 2023-12-15. **基金项目**: 国家自然科学基金(62262036, 61962030); 云南省中青年学术和技术带头人后备人才培养计划(202005AC160036). **张宇**(1999—), 男, 硕士研究生, CCF 学生会员, 主要研究方向为图形图像处理; **刘骊**(1979—), 女, 博士, 教授, 博士生导师, CCF 高级会员, 论文通信作者, 主要研究方向为计算机图形学与计算机视觉、图像处理; **付晓东**(1975—), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究方向为服务计算、决策理论与方法; **刘利军**(1978—), 男, 博士, 副教授, CCF 会员, 主要研究方向为图像处理、云计算、信息检索; **彭玮**(1980—), 女, 博士, 教授, 博士生导师, CCF 会员, 主要研究方向为机器学习、数据挖掘.

Key words: three-dimensional human pose estimation; multi-scale feature; self-attention mechanism; spatial-temporal feature fusion; temporal encoding

三维人体姿态估计[1]旨在从图像或视频中估计人体关节点在三维空间中的位置, 广泛应用于影视动画、虚拟现实、安防监控、自动驾驶、医疗辅助和步态分析中. 目前, 三维人体姿态估计方法按照输入数据的不同可分为 3 类: (1) 直接从输入图像到三维关节点坐标建立映射的端到端方法[2]; (2) 先由输入图像获得二维坐标, 再将二维坐标抬升到三维坐标的多阶段方法[3]; (3) 同时利用原始像素和附加二维关节点信息的联合方法[4].

近年来, 随着动作识别、人机交互等技术的不断发展, 对三维人体姿态估计结果提出了更高的要求. 区别于单幅图像, 基于单视角 RGB 视频输入的三维人体姿态估计[5]能提供更准确、直观的动作行为指导, 满足不同应用场景的现实需求, 成为计算机视觉和图形学领域的热点问题. 然而, 由于多帧输入, 使得三维人体姿态估计中的深度学习模型参数复杂、难以收敛且计算成本较高. 为解决上述问题, 很多研究工作通过二维姿态估计器获得高精度的二维关节点序列, 再将其作为多阶段方法的输入, 在降低计算复杂度的同时, 实现对输入端背景以及光照等无关噪声信息的过滤. 基于视频输入的多阶段三维人体姿态估计方法通常在精度上高于端到端方法, 且计算成本低于联合方法. 例如, Martinez 等[3]使用一个简单的深度前馈网络对二维姿态进行回归, 得到较为精确的三维人体姿态. Chen 等[6]根据骨长一致性, 提出双分支网络, 分别预测输入视频中人体骨骼向量的长度和方向, 作为二维到三维关节点的中间表示, 得到高精度的三维人体姿态. Bai 等[7]建立基于全连接层和图卷积的双链神经网络, 分别提取人体关节点的全局和局部依赖信息, 对不同姿态表现出良好的泛化能力. 然而, 上述方法仅考虑单一维度下人体姿态的部分尺度信息, 难以表示人体骨架在时空关系中的多尺度、多阶段分级特征, 导致在复杂姿态或遮挡场景下估计结果不平滑、精度不高.

在保证单帧姿态估计准确性的基础上, 基于视频输入的三维人体姿态估计还需考虑前后帧的姿态一致性以及帧间的平滑过渡. Tang 等[8]联合时域和频域对帧间姿态进行编码, 学习不同姿态频率间的短期相关性. Yu 等[9]用图表示时空结构并通过独立连接层对每个人体关节点独立学习局部关节特征得到三维姿态. Li 等[10]在抬升二维姿态时生成多个可能的潜在三维姿态, 并让神经网络自主学习最优假设, 该方法对遮挡姿态的估计结果有较高的精度并能生成平滑的视频, 但其计算复杂度较高, 网络推理速度慢. 此外, 不同的视频帧率和动作执行速率降低了通过单一尺度进行时序采样的估计精度和鲁棒性.

综上所述, 视频输入的三维人体姿态估计仍然存在以下难点: (1) 由于人体在运动过程中存在遮挡且姿势复杂, 难以对复杂动作和特定角度下的人体姿态进行精确表征; (2) 人体姿态特征不一, 且不同空间尺度下的运动语义信息跨度大, 导致融合不充分, 姿态估计精度较低; (3) 深度学习网络参数量庞大、计算成本高, 难以融合远距离的姿态时空特征, 生成稳定平滑的人体姿态估计结果.

针对上述难点, 本文以渐进方式对人体骨骼关节点间的多尺度空间关系进行建模; 同时在时域对人体骨骼序列进行时序多尺度编解码, 以获取深度信息并平滑空间域的人体姿态估计结果.

1 相关工作

端到端方法通过建立从图像到关节点的三维映射, 对人体姿态进行估计. Li 等[11]设计了一个多任务学习的深度卷积神经网络用于检测根节点坐标及其他关节点的相对位置, 从输入图像直接得到三维人体姿态. Pavlakos 等[12]采用将关节点表示为体积空间中的最大似然, 并通过卷积神经网络迭代逐步细化三维人体姿态估计结果, 大幅提高了三维人体姿态估计准确性. Li 等[13]提出一种具有残差对数似然估计的回归范式来建模输出的分布, 提高了人体姿态估计的准确性和计算效率. Zhou 等[14]利用 3 个关节热图来表示每个骨骼末端关节的相对深度信息, 并进行体积联合热图回归输出三维人体姿态. 在端到端方法的基础上, 联合方法加入二维人体姿态信息辅助神经网络进行学习. Zhou 等[15]通过在一个端到端深度神经网络中混合使用二维和三维注释进行弱监督迁移学习, 并引入三维几何约束进行正则化, 提高了在无三维标注的户外场景下三维人体姿态估计的准确性. Zhao 等[16]设计了一个集成主干和语义图卷积网络的端到端架构, 在主干网络中整合图像特征和二维关节点, 并用语义图卷积网络学习关节点的局

部和全局语义关系, 实现精确的三维人体姿态估计. 端到端和联合方法能够从单一图像中估计出正确的三维人体姿态, 但计算开销大, 且难以学习关节点间丰富的上下文语义信息.

多阶段方法中, 由于输入的二维骨骼在提供了人体约束先验信息的同时避免了因拍摄因素导致的图像畸变, 因此其估计误差低于上述方法, 且能较好地平衡计算效率和性能. Gong 等[17]通过将三维人体姿态进行旋转、平移和上下身的组合拼接生成新的二维-三维姿态对, 以丰富训练数据, 并开发基于运动学链式空间的判别模块, 以保证生成姿态的合理性, 解决了多阶段方法对新数据集泛化能力差的问题. 由于缺乏深度信息, 从二维人体姿态还原出正确的三维人体姿态具有极大的挑战性, 因此部分学者将二维抬升为三维视为一个多假设问题. 其中, Wehrbein 等[18]基于归一化流, 在三维人体姿态和带有潜在向量的二维人体姿态间构建双射, 重复采样潜在向量, 通过计算归一化流的逆路径生成多个假设姿态; 再通过判别器剔除不合理的姿态, 最小化最终得到的假设姿态与真值的误差. Wu 等[19]根据不同肢体关节的自由度差异, 提出一种运动约束感知网络, 并引入相对骨骼和绝对骨骼角度作为运动学约束条件, 提高了肢体关节的三维定位精度. 尽管上述方法的三维姿态估计精度较高, 但由于单个二维骨骼缺乏深度信息指导, 因此易生成不合理的错误姿态.

如何有效地利用时空特征是视频输入的三维人体姿态估计的重点工作之一, 如 Zheng 等[20]利用自注意力机制构建一个时空网络, 分别对空域的关节点和时域的二维人体姿态进行优化, 有效地增进了网络对长序列的时空关系感知能力. Cai 等[21]通过基于图卷积的局部-全局网络融合手势以及人体姿态的空间依赖和短序列的时间一致性, 学习图结构下的多尺度特征. Zhang 等[22]考虑不同关节的运动方式有所差异, 在时域上模拟每个关节点的运动轨迹, 并在空域让轨迹间进行信息交互, 最终融合三维人体姿态. Wang 等[23]提出注意增强的图卷积网络, 分别利用注意力机制和图拓扑结构对二维人体姿态的时空相关性进行表征. 针对视频估计结果不平滑问题, Zeng 等[24]通过对相邻帧的速度和加速度进行显式建模, 显著地降低了输出的抖动幅度. Li 等[25]利用跨步卷积聚合视频序列的局部上下文信息, 实现用多个二维人体姿态生成单个三维人体姿态. Zhan 等[26]输入相机外部参数将像素空间中的二维坐标转换为归一化三维空间中的三维射线, 同时使用时间卷积融合连续帧中提取的三维射线, 得到在绝对位置坐标系下的三维坐标. Shan 等[27]对输入序列中的关节点在时空域上的随机掩码进行预训练, 捕获二维时空相关性, 再将预训练好的模型装载至推理模型, 使用分离的时间空间编码器以及聚合器来提取三维时空特征, 并重建中间帧的三维人体姿态. 然而, 上述方法忽略了时间和空间上的多级分层特征, 且时空特征过度耦合, 导致计算成本较高.

针对上述问题, 区别于基于 Transformer 的时空域单一尺度特征[20], 以及仅在空域下基于图表示的多尺度特征[21,23], 本文分别在空域和时域引入多尺度分级特征表示, 以增强对人体特征的表征能力; 在空间上融合单帧的人体多尺度空间特征, 提升姿态估计精度; 在时域联合空间特征编码视频帧的时序特征, 实现时空特征的融合, 平滑输出结果, 并降低计算复杂度.

2　方法概述

本文提出的多尺度时空特征融合流程如图 1 所示, 以二维姿态序列为输入, 并在时空上解耦, 对二维人体姿态进行空间多尺度表征与融合; 然后学习姿态序列在不同尺度下的长短期时序关联, 融合多尺度时空特征; 最终输出精确、流畅的三维人体姿态估计结果. 主要工作如下:

(1) 通过空间多尺度表征模块, 提取 3 种空间尺度下的关节点、肢体和上/下身特征, 提高空间尺度下复杂运动人体姿态的表征能力.

(2) 构建空间多尺度特征融合模块, 融合 3 种空间尺度下的人体姿态特征, 降低关节点拟合的误差, 提高三维人体姿态估计精度.

(3) 提出一种基于长短期时序特征学习和融合的时序多尺度编码方法, 平滑序列中每帧姿态特征, 生成平滑且合理的三维人体姿态估计结果.

3　多尺度时空特征融合

由于人体运动姿态复杂多样, 因此仅通过单一网络难以估计出准确、平滑的三维人体姿态. 本文通过构建空间多尺度表征、空间多尺度特征融合和时序多尺度编码 3 个模块, 提升三维人体姿态估计精度并输出平滑结果.

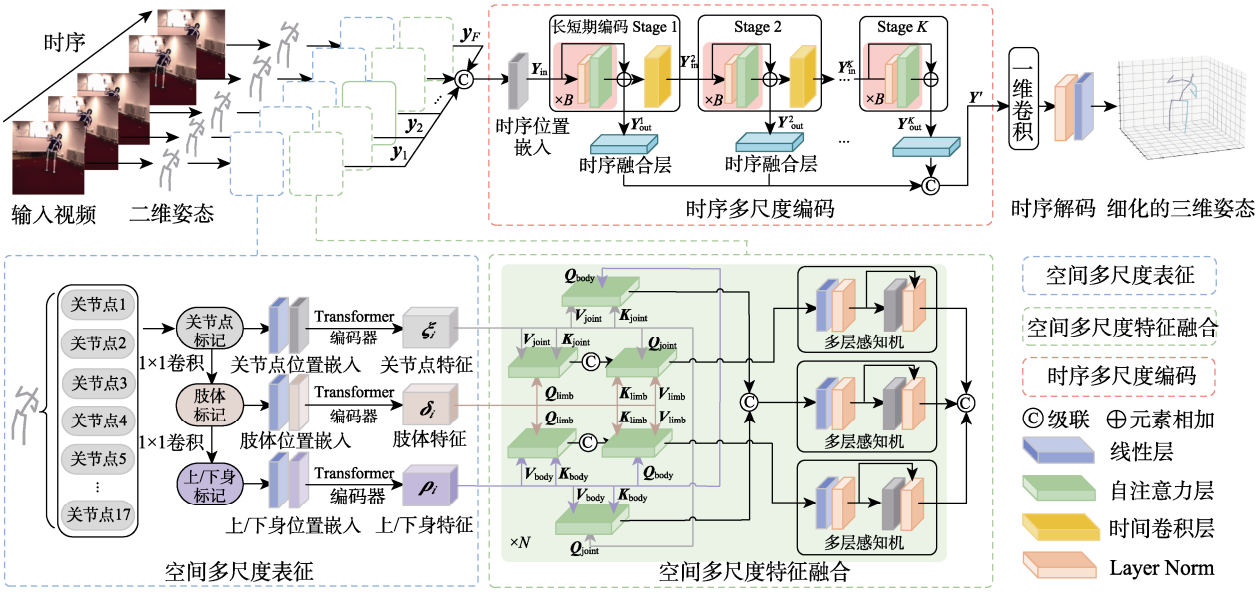


图 1 多尺度时空特征融合流程

为了对人体姿态进行空间多尺度表征，首先将运动视频 P 输入级联金字塔网络(cascaded pyramid network, CPN)[28]，得到二维人体姿态序列 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_F\}$. 视频中每一帧对应的二维姿态 $\boldsymbol{x}_i\in\mathbb{R}^{J\times 2}$，其中 F 为输入帧数，$i=[1,2,\cdots,F]$；J 是每个人体姿态所包含的关节点数.

3.1 空间多尺度表征

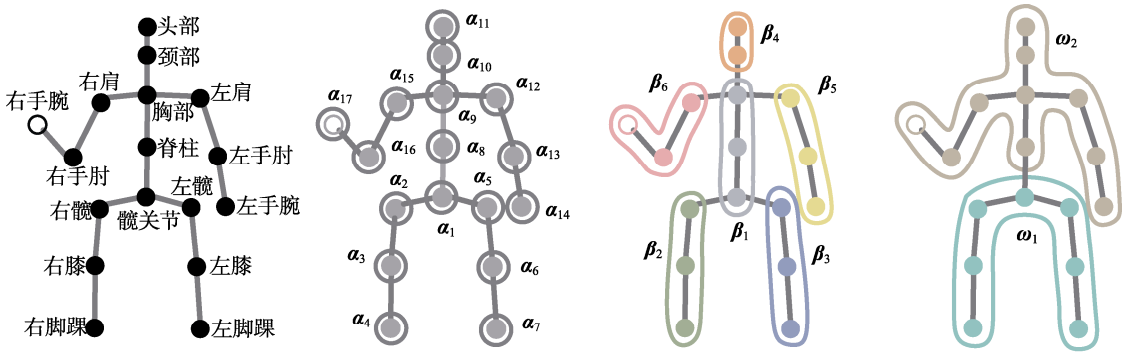
由于误差会随着人体拓扑结构积累[19]增大，从而降低姿态估计的精度. 而现有方法在表征人体姿态时往往未考虑人体骨骼内各关节点间的语义关联[26]，或仅在单一尺度下进行人体姿态表征[20]，忽略了人体不同部位的多尺度分级特征对人体关节点辅助定位的影响，因此难以精确地表征复杂姿态. 针对上述问题，区别于现有方法，本文在空间域内对人体姿态进行多尺度表征，将人体骨骼表述为 3 种尺度下的人体标记并分别提取对应特征，而且利用自注意力机制对同一尺度内各关节点分配不同注意力权重，对自由度高的关节点表征更精确. 如图 1 所示，首先将二维人体姿态的 J 个关节点作为关节点标记，并对关节点标记进行 1×1卷积，获得肢体标记、上/下身标记；然后将 3 种尺度下的人体标记分别线性投影至高维空间，并加入带有空间位置信息的嵌入矩阵；最后通过 Transformer 编码器[29]表示 3 种尺度下的人体姿态特征.

3.1.1 人体标记

本文基于人体骨架拓扑结构，聚合具有相似语义信息的关节点，将人体姿态描述为不同尺度下的人体标记. 如图 2a 所示，首先输入由 17 个关节点(包括非对称关节点头部、颈部、胸部、脊柱、髋关节以及对称关节点手肘、手腕、肩、髋、膝、脚踝)的二维坐标构成的二维人体姿态序列 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_F\}$，其中 $\boldsymbol{x}_i\in\mathbb{R}^{J\times 2}$. 通过上述关节点在二维和三维空间中抽象出的人体姿态，可在保证还原人物动作的前提下，最大限度地降低计算开销. 由于相邻节点只在同一人体部位中具有相似的运动语义信息，因此，按照运动语义种类进行标记划分，将人体姿态 $\boldsymbol{x}_i\in\mathbb{R}^{J\times 2}$ 中每个关节点的二维坐标定义为一个关节点标记 $\boldsymbol{\alpha}\in\mathbb{R}^{1\times 2}$，得到关节点标记集 $\boldsymbol{T}^i_{\text{joint}}=\{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_{17}\}$，如图 2b 所示. 再根据关节点连接关系将以上关节点标记划分为 6 个子集，对应人体的躯干、右下肢、左下肢、头部、左上肢和右上肢，并分别对每个子集进行 1×1卷积，得到关节点标记集的低层次特征抽象，即肢体标记集 $\boldsymbol{T}^i_{\text{limb}}=\{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\cdots,\boldsymbol{\beta}_6\}$，如图 2c 所示. 最后将 6 个肢体标记划分为 2 个子集，对应人体的下半身和上半身，并进行 1×1卷积得到上/下身标记集 $\boldsymbol{T}^i_{\text{body}}=\{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2\}$，该集合是肢体标记集的低层次特征抽象，如图 2d 所示.

3.1.2 标记位置嵌入

由于每个肢体和上/下身标记中都包含了多个人体关节点信息，因此将第 3.1.1 节中定义的肢体、上/下身标记集 $\boldsymbol{T}^i_{\text{joint}}\in\mathbb{R}^{J\times 2}$，$\boldsymbol{T}^i_{\text{limb}}\in\mathbb{R}^{6\times 2}$ 和 $\boldsymbol{T}^i_{\text{body}}\in\mathbb{R}^{2\times 2}$ 作为输入，通过标记位置嵌入投影至



a. 人体关节点　　b. 关节点标记集　　c. 肢体标记集　　d. 上/下身标记集

图 2　不同尺度的人体标记

更高的维度，以避免丢失过多语义信息. 此外，由于不同尺度下的人体各部分存在拓扑关系，但 Transformer 编码器在计算注意力分数时未考虑输序列的位置关系，因此会破坏人体拓扑结构并导致人体姿态估计结果不可信. 本文在文献[20]的基础上加入对应多个人体尺度标记的可学习空间位置嵌入矩阵，从而保留不同尺度下人体标记间的相对位置信息. 以关节点标记为例，其标记位置嵌入过程定义为

$$\boldsymbol{A}_{\text{joint}}^{i}=\text{Linear}_{c_1}\left(\boldsymbol{T}_{\text{joint}}^{i}\right)+\boldsymbol{E}_{\text{jpos}} \tag{1}$$

其中，$\text{Linear}_{c_1}(\cdot)$是线性投影层，将关节点标记投影至 c_1 维；$\boldsymbol{A}_{\text{joint}}^{i}\in\mathbb{R}^{J\times c_1}$ 为关节点矩阵；$\boldsymbol{E}_{\text{jpos}}$ 为关节点位置嵌入矩阵. 同理可得肢体矩阵 $\boldsymbol{A}_{\text{limb}}^{i}\in\mathbb{R}^{6\times c_2}$ 和上/下身矩阵 $\boldsymbol{A}_{\text{body}}^{i}\in\mathbb{R}^{2\times c_3}$，其中，$c_3>c_2>c_1$；再将 $\boldsymbol{A}_{\text{joint}}^{i}$，$\boldsymbol{A}_{\text{limb}}^{i}$，$\boldsymbol{A}_{\text{body}}^{i}$ 输入 Transformer 编码器，并加入卷积层进行特征提取，将特征维度统一至 c；最终得到第 i 帧 3 种空间尺度下的人体姿态特征表示：关节点特征 $\boldsymbol{\xi}_i\in\mathbb{R}^{J\times c}$，肢体特征 $\boldsymbol{\delta}_i\in\mathbb{R}^{6\times c}$ 以及上/下身特征 $\boldsymbol{\rho}_i\in\mathbb{R}^{2\times c}$.

3.2　空间多尺度特征融合

多尺度人体姿态特征包含了人体运动的局部细节与全局语义信息，融合不同尺度的姿态特征能弥补单一姿态特征对人体运动表征的限制，从而有效地降低人体关节点坐标的拟合误差，提升三维人体姿态估计精度. 然而，不同尺度的姿态特征存在异构的问题，而高层语义又呈现层层递进、相互关联的特点. 而大多方法如基于图卷积和图池化融合关节点间语义信息[21]，忽略了池化前后的关节点间语义关联，导致局部与全局特征融合不充分. 为此，本文结合自注意力机制和多层感知机构建空间多尺度特征融合模块，逐层融合 3 种空间尺度下的人体姿态特征，充分地融合空域从局部到全局的人体运动语义信息，得到初步姿态特征.

首先将 3 种空间尺度下的人体姿态特征 $\boldsymbol{\xi}_i$，$\boldsymbol{\delta}_i$ 和 $\boldsymbol{\rho}_i$ 进行线性映射，分别生成查询、键和值矩阵. 以关节点特征为例，它们的公式分别为

$$\begin{cases}\boldsymbol{Q}_{\text{joint}}=\boldsymbol{\xi}_i\cdot\boldsymbol{\gamma}_{\boldsymbol{Q}}\\ \boldsymbol{K}_{\text{joint}}=\boldsymbol{\xi}_i\cdot\boldsymbol{\gamma}_{\boldsymbol{K}}\\ \boldsymbol{V}_{\text{joint}}=\boldsymbol{\xi}_i\cdot\boldsymbol{\gamma}_{\boldsymbol{V}}\end{cases} \tag{2}$$

其中，$\boldsymbol{\gamma}_{\boldsymbol{Q}}\in\mathbb{R}^{c\times d}$，$\boldsymbol{\gamma}_{\boldsymbol{K}}\in\mathbb{R}^{c\times d}$，$\boldsymbol{\gamma}_{\boldsymbol{V}}\in\mathbb{R}^{c\times c}$ 为线性层，d 为映射的中间维度；$\boldsymbol{Q}_{\text{joint}}\in\mathbb{R}^{J\times d}$，$\boldsymbol{K}_{\text{joint}}\in\mathbb{R}^{J\times d}$，$\boldsymbol{V}_{\text{joint}}\in\mathbb{R}^{J\times c}$；同理可得肢体特征和上/下身特征查询、键和值矩阵，分别为 $\boldsymbol{Q}_{\text{limb}}\in\mathbb{R}^{6\times d}$，$\boldsymbol{K}_{\text{limb}}\in\mathbb{R}^{6\times d}$，$\boldsymbol{V}_{\text{limb}}\in\mathbb{R}^{6\times c}$ 和 $\boldsymbol{Q}_{\text{body}}\in\mathbb{R}^{2\times d}$，$\boldsymbol{K}_{\text{body}}\in\mathbb{R}^{2\times d}$，$\boldsymbol{V}_{\text{body}}\in\mathbb{R}^{2\times c}$.

然后，使用单一尺度的查询矩阵，显式地引导另一尺度的键-值矩阵对进行自注意力运算，在每 2 个尺度特征间双向计算得到 2 组特征并级联. 以关节点和肢体尺度为例，定义该过程为

$$\begin{cases}\boldsymbol{Z}_{\text{joint_limb}}=\text{Concat}\left(\text{SA}\left(\boldsymbol{Q}_{\text{joint}},\boldsymbol{K}_{\text{limb}},\boldsymbol{V}_{\text{limb}}\right),\right.\\ \quad\left.\text{SA}\left(\boldsymbol{Q}_{\text{limb}},\boldsymbol{K}_{\text{joint}},\boldsymbol{V}_{\text{joint}}\right)\right)\\ \text{SA}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\text{Softmax}\left(\boldsymbol{Q}\boldsymbol{K}^{\text{T}}/\sqrt{c}\right)\boldsymbol{V}\end{cases} \tag{3}$$

其中，$\boldsymbol{Z}_{\text{joint_limb}}$ 为计算过程的中间特征；$\text{Concat}(\cdot)$ 为级联操作；$\text{SA}(\cdot)$ 为自注意力计算；$\text{Softmax}(\cdot)$ 为归一化函数.

再将级联后的特征输入由线性层、LayerNorm 归一化以及多层感知机组成的前馈网络恢复维度，并精细化特征，得到融合了第 n 层的关节点尺度输出特征，$n=[1,2,\cdots,N]$. 其计算公式为

$$\boldsymbol{Z}_{\text{joint}}^{n}=\text{MLP}\left(\text{LN}\left(\text{Linear}\left(\boldsymbol{Z}_{\text{joint_limb}}\right)\right)\right)+\text{LN}\left(\text{Linear}\left(\boldsymbol{Z}_{\text{joint_limb}}\right)\right) \tag{4}$$

其中，$\text{Linear}(\cdot)$ 为线性映射；$\text{MLP}(\cdot)$ 为多层感知机；$\text{LN}(\cdot)$ 为 LayerNorm 归一化层.

最终将最后一层输出的 3 组特征级联，得到第 i 帧的初步人体姿态特征 $\boldsymbol{y}_i\in\mathbb{R}^{(J+6+2)\times 3}$.

3.3 时序多尺度编码

由于初步姿态特征未进行信息交互，因此导致输出结果不平滑帧间差异大. 现有方法在时域仅考虑单一时序尺度[20]或仅进行简单的序列池化处理[25]，未有效地利用时域上丰富的多尺度特征. 为更好地输出正确的人体姿态，不同于现有方法，本文在时域对初步姿态特征序列进行时序多尺度编码，逐阶段压缩姿态序列长度，学习并融合人体运动在不同时间跨度下的长短期时序信息，最后输出平滑且合理的三维姿态序列.

3.3.1 长短期编码

首先级联第 3.2 节中获得的所有初步姿态特征，组成初步三维姿态序列 $\boldsymbol{Y}_{\text{in}}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_F\}+\boldsymbol{E}_{\text{tpos}}$，以保留姿态在序列中的时序位置信息，其中，$\boldsymbol{E}_{\text{tpos}}$ 代表时序位置嵌入. 如图 1 所示，将初步三维姿态序列的长短期编码分为 K 个阶段. 以 81 帧输入为例，$\boldsymbol{Y}_{\text{in}}$ 作为全尺寸序列对应长期尺度时间片即 81 帧(约 16.2 s)，输入长短期编码 Stage1，通过 LayerNorm 以及自注意力层[29]对数据归一化处理，并学习该时序尺度下姿态特征的帧间时序关联. 在 Stage1 内重复该过程 B 次，并通过跳跃连接与原输入相加得到平滑的长期姿态序列

$$\begin{cases}\boldsymbol{Y}_{\text{out}}^{1}=\left(L_{\text{SA}}\right)^{B}+\boldsymbol{Y}_{\text{in}}\\ L_{\text{SA}}=\text{SA}\left(\text{LN}\left(\boldsymbol{Y}_{\text{in}}\right)\right)\end{cases} \tag{5}$$

其中，$\text{LN}(\cdot)$ 为 LayerNorm 归一化层；$\text{SA}(\cdot)$ 为自注意力计算；$\boldsymbol{Y}_{\text{out}}^{1}$ 是 Stage1 输出时序特征；B 是超参数. 随后利用一个时间卷积层融合冗余帧，将序列长度从 F 缩减至原来的 $1/\theta$ 对应较短期尺度时间片($\theta=3$ 时，F 从 81 帧缩减至 27 帧，约 5.4 s)，同时使用线性映射重塑序列维度 $c\to c\times\theta$. Stage1 的横向输出为

$$\boldsymbol{Y}_{\text{in}}^{2}=\text{Linear}\left(\text{TConv}_{\theta,\theta}\left(\boldsymbol{Y}_{\text{out}}^{1}\right)\right) \tag{6}$$

其中，$\boldsymbol{Y}_{\text{in}}^{2}\in\mathbb{R}^{F/\theta\times(J+6+2)\times c\times\theta}$ 且其直接作为 Stage2 的输入；$\text{Linear}(\cdot)$ 是线性映射；$\text{TConv}_{\theta,\theta}(\cdot)$ 代表卷积核大小为 θ、步长为 θ 的时间卷积层；$\boldsymbol{Y}_{\text{out}}^{K}$ 为经 $K-1$ 个阶段的计算，通过姿态序列 $\boldsymbol{Y}_{\text{in}}^{K}\in\mathbb{R}^{F/\theta^{K-1}\times(J+6+2)\times c\times\theta^{K-1}}$ 在 Stage K 中计算自注意力，得到平滑的短期姿态序列.

3.3.2 时序特征融合

在多个时序尺度下，将每个阶段纵向输出的时序特征 $\boldsymbol{Y}_{\text{out}}=\{\boldsymbol{Y}_{\text{out}}^{1},\boldsymbol{Y}_{\text{out}}^{2},\cdots,\boldsymbol{Y}_{\text{out}}^{K}\}$ 进行时序特征融合，以学习不同的时序特征. 首先用线性层将不同时序尺度的三维姿态序列维度统一至 D，并将序列长度上采样至 F. 融合了多尺度时序信息的最终姿态特征定义为

$$\boldsymbol{Y}'=\text{Concat}\left(\text{UpSample}_{u}\left(\boldsymbol{Y}_{\text{out}}^{k}\otimes\boldsymbol{\gamma}^{k}\right)\right) \tag{7}$$

其中，$k=[1,2,\cdots,K]$；$\text{UpSample}_{u}(\cdot)$ 表示上采样率 $u=\theta^{k}$ 的上采样操作；$\otimes$ 是矩阵乘法；$\boldsymbol{\gamma}^{k}\in\mathbb{R}^{F\times(J+6+2)\times\theta^{k-1}/c\times D}$ 表示线性映射矩阵，$D=2c$ 表示统一的特征维度；$\text{Concat}(\cdot)$ 表示级联操作. 姿态特征 $\boldsymbol{Y}'$ 再经过时序解码生成细化的三维人体姿态 $\boldsymbol{Y}=\text{Decode}(\boldsymbol{Y}')$. 其中，$\text{Decode}(\cdot)$ 表示时序解码，由一维卷积和线性层实现. 时序解码的维度变化过程为 $\boldsymbol{Y}'\in\mathbb{R}^{K\times F\times(J+6+2)\times D}\to\boldsymbol{Y}\in\mathbb{R}^{1\times(J\times 3)}$.

本文多尺度时空特征融合方法步骤如下.

Step1. 空间多尺度表征:

Step1.1. 将人体骨骼拓扑结构描述为如图 2 所示的 3 种空间尺度下的人体标记，通过 1×1 卷积得到 3 种人体标记集合 $\boldsymbol{T}_{\text{joint}}^{i}$，$\boldsymbol{T}_{\text{limb}}^{i}$，$\boldsymbol{T}_{\text{body}}^{i}$；

Step1.2. 通过式(1)标记嵌入将人体标记集合投影至高维空间，利用 Transformer 编码器表示 3 种空间尺度下的人体姿态特征，再统一特征维度 $\boldsymbol{\xi}_i\in\mathbb{R}^{J\times c}$，$\boldsymbol{\delta}_i\in\mathbb{R}^{6\times c}$，$\boldsymbol{\rho}_i\in\mathbb{R}^{2\times c}$.

Step2. 空间多尺度特征融合. 对 Step1.2 输出的空间多尺度人体姿态特征进行特征融合，通过重复式(2)~(4)融合关节点特征、肢体特征、上/下身特征，得到初步人体姿态特征 $\boldsymbol{y}_i$.

Step3. 时序多尺度编码:

Step3.1. 将初步人体姿态特征 $\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_F\}$ 级联，并加入位置嵌入矩阵 $\boldsymbol{E}_{\text{tpos}}$，得到初步三维人体姿态序列 $\boldsymbol{Y}_{\text{in}}$；通过交替使用式(5)(6)对该姿态序列进行长短期编码，获得多时序尺度下的姿态序列特征 $\boldsymbol{Y}_{\text{out}}=\{\boldsymbol{Y}_{\text{out}}^{1},\boldsymbol{Y}_{\text{out}}^{2},\cdots,\boldsymbol{Y}_{\text{out}}^{K}\}$；

Step3.2. 通过式(7)上采样和线性映射融合多尺度时序姿态序列特征 $\boldsymbol{Y}_{\text{out}}$，再经过时序解码最终输出细化的三维人体姿态 $\boldsymbol{Y}$.

3.4 损失函数

本文方法的各输入模型以端到端的方式进行

训练, 并通过平均每关节位置误差(mean per joint position error, MPJPE)损失函数

$$L_{\mathrm{MPJPE}}=\frac{1}{F\cdot J}\sum_{f=1}^{F}\sum_{j=1}^{J}\left\|\boldsymbol{Y}_{f,j}-\boldsymbol{G}_{f,j}\right\|_2,$$

计算人体姿态估计结果的三维坐标在欧几里得空间中与真值(ground truth, GT)的均方误差. 其中, F 和 J 分别为总帧数和关节点数; $\boldsymbol{Y}_{f,j}$ 为第 f 帧中第 j 个关节点的预测值; $\boldsymbol{G}_{f,j}$ 为第 f 帧中第 j 个关节点的真实值. 通过在训练过程中不断优化, 以降低损失并最优化各模型的性能.

4　实验结果及分析

4.1　数据集

本文方法各模型以 $\boldsymbol{X}\in\mathbb{R}^{F\times J\times 2}$ 为输入, 分别使用基准数据集 Human3.6M[30]和规模较小的数据集 HumanEva-I 作为实验训练集和测试集. 在 Human3.6M 进行训练与测试时, 取 $J=17$, 并使用 5 名对象(S_1, S_5, S_6, S_7, S_8)的数据作为实验训练集, 2 名对象(S_9, S_{11})的数据作为测试集, 训练集与测试集的划分均与对比方法保持一致. 此外, 本文还对 Human3.6M 数据集进行以下预处理: 将频率为 50 Hz 的原始视频下采样至 5 Hz, 以去除高帧率下的冗余动作, 提升相邻帧间的动作语义丰富度, 并在训练和测试阶段以水平翻转作为数据增强. 本文使用 Human3.6M 数据集预训练后的模型, 直接在 HumanEva-I 测试集(对象 S_1, S_2, S_3, 动作 Walk, Jog, Box)上以协议 2 进行评估, 未使用训练集训练, 以验证本文方法对小规模数据集的泛化性能.

4.2　评估指标

本文使用 2 个常见协议作为评估指标: 协议 1 即 MPJPE, 该指标计算估计的三维关节点到真值的平均欧几里得距离, 对于视频姿态估计来说, 误差在所有帧上计算平均值; 协议 2 即重建平均每关节位置误差(pose-processing mean per joint position error, P-MPJPE), 是使用普鲁克分析对三维人体姿态估计结果和真实值进行刚性变换后计算得到的平均位置误差. 此外, 在 Human3.6M 数据集中还使用了平均每关节速度误差(mean per joint velocity error, MPJVE)[31]作为衡量视频三维人体姿态估计结果流畅程度的评估指标, 上述所有评估指标的值越低, 代表方法效果越好.

4.3　实验设置

本文实验选用硬件平台为 Intel Core i9-9900k CPU @3.60 GHz, NVIDIA GeForce RTX 1080Ti GPU×2, 32 GB DDR4 2 666 MHz RAM. 本文方法基于 PyTorch 框架实现, 在 Human3.6M 数据集上使用 CPN 提取二维人体关节点, 随后采用 Adam 优化器以端到端方式进行训练. 初始学习率设定为 4×10^{-4}, 学习衰减率为 0.99, 共学习 60 个 epoch. 本文分别选取输入视频长度 $F=9$, 27, 81 进行实验, 设置 Batch_Size 分别为 512, 256, 64.

4.4　估计结果及对比分析

4.4.1　定量对比分析

(1) Human3.6M

表 1~表 3 所示为本文基于 Human3.6M 数据集的全部 15 个动作, 另以协议 1、协议 2 和 MPJVE 评估指标与其他现有方法估计结果进行定量对比分析, 并在测试集(S_9, S_{11})上进行测试, 对比方法的数据来自于原始论文及开源代码复现. 从表 1 可以看出, 本文方法以 81 帧输入时在多个动作的测试中均达到最优的表现, 并在 MPJPE 上与对比方法的对比降低了 0.3 mm. 使用更长的输入可从更多的时序信息中拟合更精准的三维人体姿态, 本文方法以 81 帧输入时的 MPJPE 低于部分以 243/351 帧输入的方法, 说明本文方法在利用较短的输入序列长度的同时仍能获得较好的性能.

从估计误差明显高于平均值的部分困难动作(如 Photo, Sit, SitDown, WalkDog 等)的结果来看, 本文方法以 81 帧输入时在 2 个协议中均达到了最优值, 说明本文方法对于困难姿态具备良好的表征能力, 能够准确地拟合关节点的三维坐标. 此外, 从表 1 可以看出, 本文方法以 81 帧输入时在协议 1 中的动作 Pose, Phone, Greet 和 Direction 也达到最优结果; 以 9 帧输入时的 MPJPE 不仅低于所有 1 帧/7 帧模型, 还在动作 Greet, Photo 和 SitDown 中优于部分以 243/351 帧输入的方法. 表 2 的结果验证了本文方法以 81 帧输入时在协议 2 中的大多数动作都达到了最优或次优结果, P-MPJPE 较对比方法降低了 2.3%. 而以 9 帧输入时在 Sit 和 SitDown 动作上超越了部分以 243/351 帧输入的方法, 以 27 帧输入时也在 4 个动作上得到次优结果. 以上结果表明本文方法较对比方法能以较短的输入获得更高的准确性. 从表 3 可看出, 本文方法与对比方法以 27 帧输入时的 MPJVE 相比降低了 0.2 mm(约 7.1%), 而以 81 帧输入时的 MPJVE 为 2.4 mm, 较对比方法(243 帧)降低了 4.0%, 证明本文方法能从有限的时序信息得到更平滑的三维人体姿态估计结果.

表 1 不同方法在 Human3.6M 数据集上协议 1 的 MPJPE 定量对比

方法	*F*	MPJPE/mm															
		Dir	Disc	Eat	Greet	Phone	Photo	Pose	Purch	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkT	平均
Pavllo 等[31]	1	47.1	50.6	49.0	51.8	53.6	61.4	49.4	47.4	59.3	67.4	52.4	49.5	55.3	39.5	42.7	51.8
	243	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
Cai 等[21]	7	44.6	47.4	45.6	48.8	50.8	59.0	47.2	43.9	57.9	61.9	49.7	46.6	51.3	37.1	39.4	48.8
Chen 等[6]	81	42.1	43.8	41.0	43.8	46.1	53.5	42.4	43.1	53.9	60.5	45.7	42.1	46.2	32.2	33.8	44.6
	243	41.4	43.5	40.1	42.9	46.6	51.9	41.7	42.3	53.9	60.2	45.4	**41.7**	46.0	31.5	32.7	44.1
Bai 等[7]	1	43.3	48.2	46.4	48.1	51.8	62.0	46.8	45.8	61.8	66.3	49.8	47.1	52.1	38.7	41.8	50.0
Zheng 等[20]	81	41.5	44.8	**39.8**	42.5	46.5	51.6	42.1	42.0	53.3	60.7	45.5	43.3	46.1	31.8	32.2	44.3
Wang 等[23]	96	39.9	43.9	41.0	43.4	46.1	53.5	42.1	40.9	55.2	57.7	45.3	42.2	45.5	30.6	31.3	43.9
Tang 等[8]	243	41.3	44.8	42.2	43.0	47.9	55.3	43.4	40.9	58.1	66.5	46.2	44.2	45.2	30.8	31.5	45.4
Yu 等[9]	243	41.3	44.3	40.8	41.8	45.9	54.1	42.1	41.5	57.8	62.9	45.0	42.8	45.9	**29.4**	**29.9**	44.4
Li 等[25]	351	40.3	**43.3**	40.2	42.3	45.6	52.3	41.8	**40.5**	55.9	60.6	**44.2**	43.0	44.2	30.0	30.2	43.7
Li 等[10]	81	41.0	45.2	41.2	43.1	45.6	52.7	42.2	42.5	54.4	61.3	45.1	42.8	46.9	31.4	33.1	44.6
本文	9	45.8	48.3	44.6	45.2	49.5	54.1	46.3	47.3	57.8	64.7	50.1	46.5	49.5	38.1	35.3	48.2
	27	41.0	45.9	42.5	42.8	46.9	51.8	43.0	45.6	53.1	60.3	47.7	43.1	45.3	33.7	32.4	45.0
	81	**38.8**	44.5	40.9	**41.3**	**45.2**	**50.5**	**41.2**	43.8	**51.6**	**57.1**	46.1	42.5	**43.2**	32.5	31.6	**43.4**

注. 粗体表示最优值, 下划线表示次优值.

表 2 不同方法在 Human3.6M 数据集上的协议 2 的 P-MPJPE 定量对比

方法	*F*	P-MPJPE/mm															
		Dir	Disc	Eat	Greet	Phone	Photo	Pose	Purch	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkT	平均
Pavllo 等[31]	1	36.0	38.7	38.0	41.7	40.1	45.9	37.1	35.4	46.8	53.4	41.4	36.9	43.1	30.3	34.8	40.0
	243	34.1	36.1	34.4	37.2	36.4	42.2	34.4	33.6	45.0	52.5	37.4	33.8	37.8	25.6	27.3	36.5
Cai 等[21]	7	35.7	37.8	36.9	40.7	39.6	45.2	37.4	34.5	46.9	50.1	40.5	36.1	41.0	29.6	33.2	39.0
Chen 等[6]	81	33.1	35.3	33.4	35.9	36.1	41.7	32.8	33.3	42.6	49.4	37.0	32.7	36.5	25.5	27.9	35.6
Bai 等[7]	1	34.1	37.8	36.7	38.9	40.1	47.0	35.5	34.6	49.5	53.2	40.1	36.0	41.3	30.1	34.2	39.3
Zheng 等[20]	81	32.5	34.8	32.6	34.6	35.3	39.5	**32.1**	32.0	42.8	48.5	34.8	32.4	35.3	24.5	26.0	34.6
Wang 等[23]	96	**31.4**	34.6	**32.2**	35.4	35.5	41.3	32.9	**31.0**	43.2	45.9	36.3	31.4	35.3	23.9	25.0	34.4
Tang 等[8]	243	31.6	35.6	34.4	35.0	36.7	42.5	33.1	**31.0**	46.5	52.3	37.1	33.5	35.4	24.4	25.0	35.6
Yu 等[9]	243	32.4	35.3	32.6	**34.2**	**35.0**	42.1	**32.1**	31.9	45.5	49.5	36.1	32.4	35.6	**23.5**	24.7	34.8
Li 等[25]	351	32.7	35.5	32.5	35.4	35.9	41.6	33.0	31.9	45.1	50.1	36.3	33.5	35.1	23.9	25.0	35.2
Li 等[10]	81	31.9	36.0	33.5	35.2	36.1	40.8	32.8	33.0	43.9	49.6	36.9	33.2	36.6	24.8	27.2	35.4
本文	9	35.3	37.1	39.5	37.5	39.7	42.7	38.7	37.4	44.3	49.6	40.9	35.7	38.7	27.3	31.9	38.4
	27	34.7	36.2	38.6	34.9	37.3	36.9	36.8	35.1	41.7	44.8	35.9	32.8	34.3	25.6	25.7	35.4
	81	32.8	**34.4**	35.6	34.7	35.5	**36.1**	33.5	31.7	**40.9**	**42.8**	**33.5**	**30.4**	**32.7**	25.4	**23.6**	**33.6**

注. 粗体表示最优值, 下划线表示次优值.

(2) HumanEva-I

为了验证本文方法对小数据集具有良好的泛化能力, 将 Human3.6M 数据集训练的模型直接通过 HumanEva-I 进行评估. 由于 HumanEva-I 数据集的 Box 动作难度较大且视频质量欠佳, 导致估计误差相对较高, 因此部分方法舍弃了该动作的实验数据, 但本文仍然保留该部分的数据并计算平均值, 以实现公平对比. 对比结果如表 4 所示, 可以看出, Cheng 等[32]的模型在 HumanEva-I 训练集上进行训练, 其在测试集的 Walk 和 Jog 动作上均取得较好表现. 而本文方法未在该数据集进行训练, 仍在对象 S_2 的 Walk 动作取得最优结果, 在 S_1-Walk, S_3-Jog, S_2-Box 以及 MPJPE 上取得次优结果, 优于以上使用 HumanEva-I 作为训练集的部分方法. 表 4 的对比结果说明, 本文方法不仅能够精确地表征人体, 而且具有较好的可迁移性, 在小样

表 3　不同方法在 Human3.6M 数据集上的 MPJVE 定量对比

方法	F	MPJVE/mm															
		Dir	Disc	Eat	Greet	Phone	Photo	Pose	Purch	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkT	平均
Zheng 等[20]	27	3.7	3.7	3.1	4.1	3.0	3.3	3.4	3.7	2.6	3.2	3.0	3.2	4.3	3.9	3.6	3.4
	81	3.2	3.4	2.6	3.6	2.6	3.0	2.9	3.2	2.6	3.3	2.7	2.8	3.8	3.2	2.9	3.1
Pavllo 等[31]	27	3.4	3.5	2.7	3.9	2.7	3.1	3.1	3.4	2.5	3.4	2.7	2.9	4.2	3.6	3.3	3.2
	243	3.0	3.1	2.2	3.4	2.3	2.7	2.7	3.1	2.1	2.9	2.3	2.4	3.7	3.1	2.8	2.8
Chen 等[6]	27	3.0	3.2	2.3	3.5	2.2	2.7	2.8	3.1	2.0	2.7	2.4	2.4	3.8	2.9	3.1	2.8
	243	2.7	**2.8**	**2.0**	3.1	**2.0**	2.4	**2.4**	2.8	**1.8**	2.4	2.0	**2.1**	3.4	**2.7**	2.4	2.5
本文	27	2.7	3.0	2.2	3.2	2.5	2.5	2.5	3.0	2.0	2.4	2.1	2.5	3.4	2.9	2.5	2.6
	81	**2.6**	**2.8**	**2.0**	**3.0**	2.2	**2.3**	**2.4**	**2.7**	**1.8**	**2.2**	**1.9**	2.3	**3.2**	**2.7**	**2.3**	**2.4**

注. 粗体表示最优值.

表 4　不同方法在 HumanEva-I 数据集上的协议 2 的定量对比　　mm

方法	Walk			Jog			Box			平均
	S_1	S_2	S_3	S_1	S_2	S_3	S_1	S_2	S_3	
Pavllo 等[31]	13.9	10.2	46.6	20.9	13.1	13.8	23.8	33.7	32.0	25.7
Cheng 等[32]	**10.6**	11.8	**19.3**	**15.8**	**11.5**	**12.2**				
Li 等[25]	14.0	10.0	32.8	19.5	13.6	14.2	22.4	21.6	22.5	18.9
Li 等[33]	12.8	9.7	26.5	16.0	12.2	12.7	**14.6**	**16.9**	**19.3**	**15.6**
Liu 等[34]*	13.1	9.8	26.8	16.9	12.8	13.3				
Zheng 等[20]	16.3	11.0	47.1	25.0	15.2	15.1				
本文	12.2	**9.4**	27.8	18.9	13.2	12.4	25.9	20.7	25.5	18.4

注. 粗体表示最优值, 下划线表示次优值, *表示使用该数据集训练, 空白表示原论文未提供该数据.

本量的数据集中仍能取得良好的表现. 以上所有对比方法的实验数据均来源于原论文.

4.4.2　定性对比分析

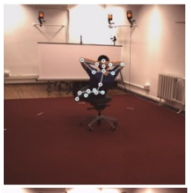
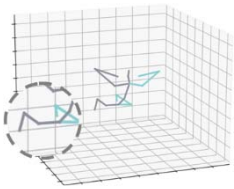
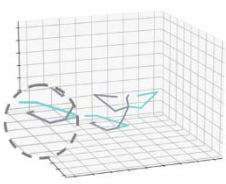
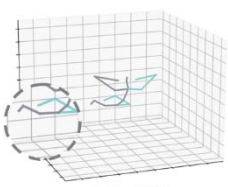
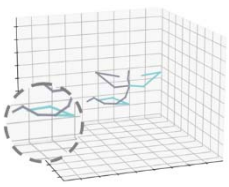
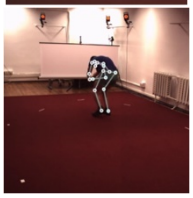
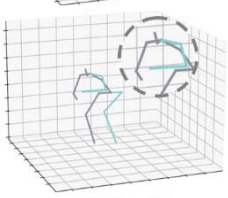
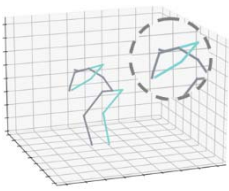
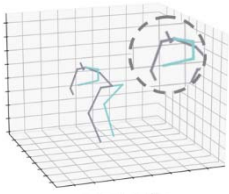
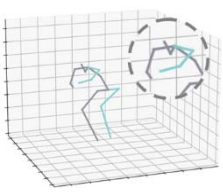
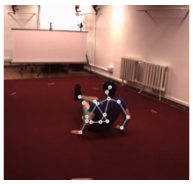
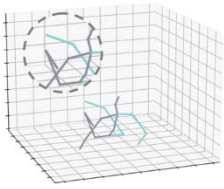
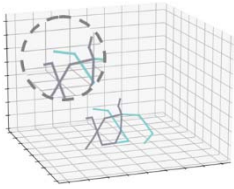
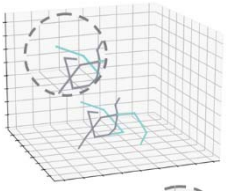
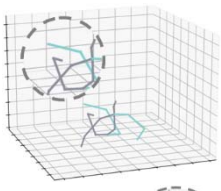
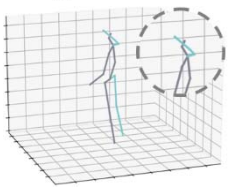
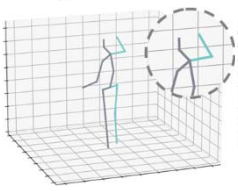
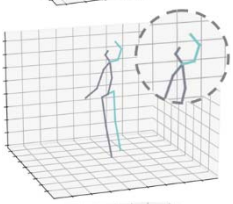
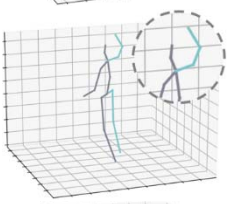
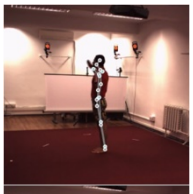
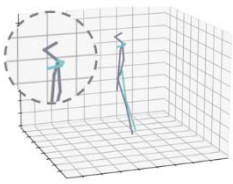
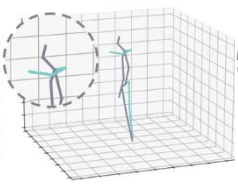
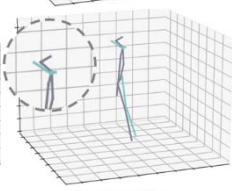
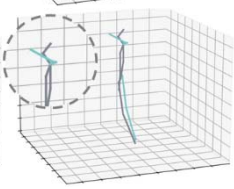
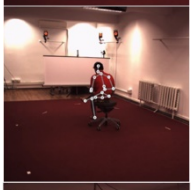
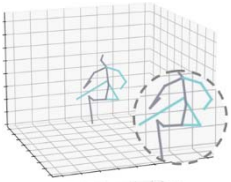
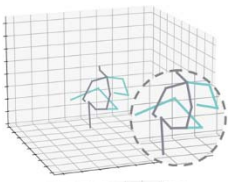
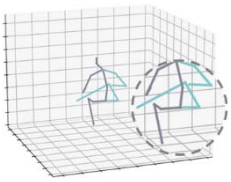
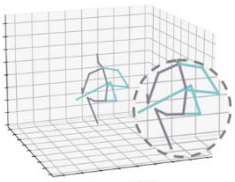
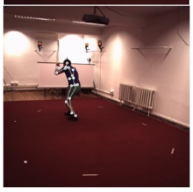
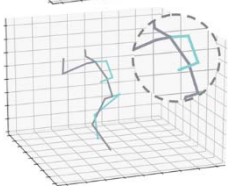
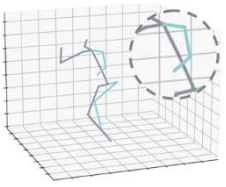
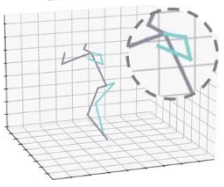
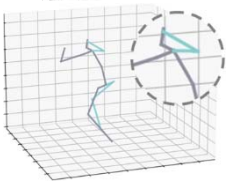
为更好地验证本文方法对复杂姿态表征的有效性, 表 5 给出了本文方法与对比方法在 Human3.6M 数据集上估计误差明显高于平均值的部分困难动作的可视化三维人体姿态估计结果. 从表 5 中的二维输入图像可以看出, 在执行动作 Sit 时, 受试者 S_{11} 背对摄像机且下半身存在严重遮挡情况, 关节点分布密集. 对比表 5 所示 GT, 表 5 第 3 列和第 4 列所示对比方法均不能正确估计人体下半身的姿态, 而第 5 列所示本文方法可以得到相对准确且合理的估计结果. 表 5 第 2 行的结果表明, 本文方法对 S_{11} 在 WalkDog 动作中的俯身角度(上下身距离)以及左手肘弯曲角度有着较为准确的还原. 从表 5 第 3 行的对比可以看出, 对比方法在左脚被自身遮挡的情况下估计出错误的左脚踝位置, 而本文得到的估计结果更接近 GT. 通过表 5 第 4 行的对比结果看出, 本文方法估计结果的受试者 S_9 右手腕位置以及肩部和手肘形成的夹角均比对比方法更准确. 表 5 第 5 行 S_9 执行 Greet 动作时几乎整个右半身均被遮挡, 从对比结果来看, 本文方法还原的右手臂位置更准确. 不适定姿态如表 5 第 6 行的结果显示, Zheng 等[20]的估计结果右手腕位置错误, Li 等[10]的估计结果右脚踝和右膝位置错误, 本文方法在上述关节点估计位置更接近 GT. 总之, 本文方法在严重遮挡(第 1, 3, 4 行)、关节点密集分布(第 1, 2, 4, 5 行)和不适定(第 2, 4, 6 行)的困难姿态时表现出更高的准确性. 然而, 从表 5 第 7 行可以看出, 由于受试者 S_{11} 执行 Photo 动作时整条右臂都被躯干遮挡, 且二维输入有较大误差, 导致本文方法与对比方法均估计出错误的右手臂位置.

为进一步验证本文方法对未见过数据的泛化性能以及估计结果的适用性, 图 3 展示了其对于非实验室环境下采集数据的可视化估计结果. 这些数据均为从互联网下载的人体运动视频(包括高难度的姿势, 存在背景、光源等干扰), 所选视频涵盖了各类困难动作, 如滑板速降(结果 1)、运动跑酷(结果 2)、各式舞蹈(结果 3)等, 所有动作均未在 Human3.6M 数据集中出现过.

4.4.3　计算复杂度对比分析

本文通过时空解耦的策略进行时空特征融合,

表 5 不同方法在 Human3.6M 数据集上的定性对比结果

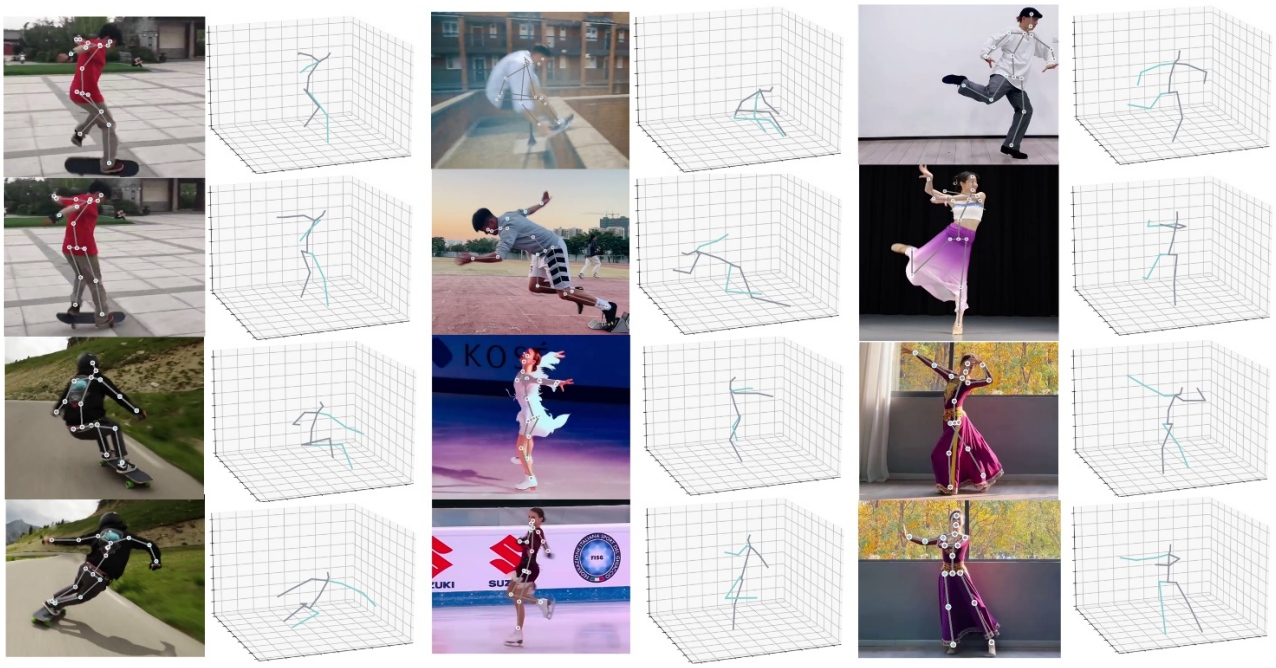
对象-动作	二维输入	Zheng 等[20]	Li 等[10]	本文	GT
S_{11}-Sit					
S_{11}-WalkDog					
S_{11}-SitDown					
S_9-Direction					
S_9-Greet					
S_9-Phone					
S_{11}-Photo					

不仅提高了人体姿态估计精度，还在一定程度上降低了模型的计算复杂度. 不同方法的计算复杂度对比如表 6 所示，它们均以 81 帧输入进行对比. 本文方法的总参数量(parameters, P)低于 Li 等[10]方法，但比 Zheng 等[20]方法有所增加. 这是由于本文加入了如肢体、上/下身位置嵌入矩阵等，增加了网络需要学习的参数量；而本文的时空特征融合策略以及超参数的选取，有效地降低了每秒浮点运算次数(floating point operations, FLOPs). 使用 NVIDIA GeForce RTX 2080Ti GPU 对不同方法的每秒处理帧数(frames per second, FPS)进行实验，结果显示本文方法测试速度最快. 综上，本文方法在降低了部分模型计算复杂度的同时，保持了更高的姿态估计精度.

4.5 消融分析

4.5.1 时空模型有效性分析

本文首先分析各模块的有效性以及对整体模型的性能影响，结果使用以 81 帧输入时的 MPJPE, P-MPJPE 以及 MPJVE 进行评估及分析，这 3 个指标衡量模型的精度以及输出视频平滑度，它们的



a. 定性结果 1　　b. 定性结果 2　　c. 定性结果 3

图 3　互联网运动视频中的定性结果

表 6　不同方法计算复杂度对比

方法	$10^{-6}\times P$	$10^{-9}\times$FLOPs	FPS	MPJPE/mm
Zheng 等[20]	9.60	1.358	270	44.3
Li 等[25]	4.06	0.392	112	45.4
Li 等[10]	19.84	3.140	116	44.5
本文	13.27	0.782	327	43.4

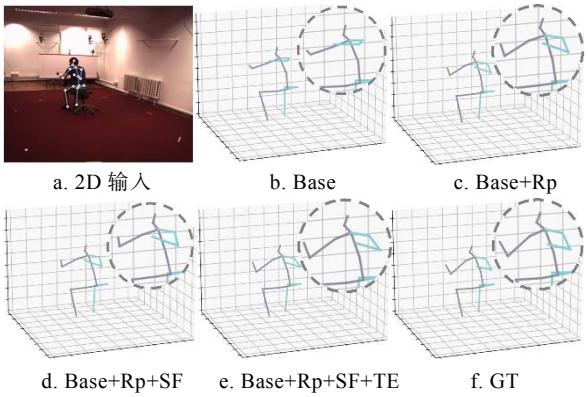
值越低，表示模型性能越好. 如表 7 所示，Base 为只使用关节点尺度特征、无特征融合、无时序编码的基准模型；Rp, SF 和 TE 分别表示空间多尺度表征、空间多尺度特征融合以及时序多尺度编码模块；其中，Base+Rp 模型的空间多尺度特征采取直接级联的方式进行融合. 从表 7 可以看出，在 Base 模型加入 Rp 后，P-MPJPE 降低了 1.7 mm(约 4.5%)；而再加入 SF 后，MPJPE 和 P-MPJPE 分别又降低了 1.7 mm 和 1.6 mm，证明 Rp 和 SF 可以有效地降低姿态估计误差；最后加入 TE，在进一步降低估计误差的同时，大幅降低了 MPJVE(约 11.3%)，有效地增加了三维人体姿态估计结果的平滑度.

表 7　各模块消融分析　　mm

模型	MPJPE	P-MPJPE	MPJVE
Base	47.2	38.0	2.95
Base+Rp	46.5	36.3	2.89
Base+Rp+SF	44.8	34.7	2.74
Base+Rp+SF+TE	**43.4**	**33.6**	**2.43**

注. 粗体表示最优值.

图 4 展示了加入不同模块后的以 81 帧输入时在 Human3.6M 数据集上的可视化姿态估计结果(受试者 S_{11} 执行 Eat 动作). 从图 4f 与图 4b，以及图 4c 对比的结果可以看出，加入 Rp 后，模型对人体姿态的表征能力有所提升；再增加 SF 和 TE 后的三维人体姿态估计结果更加接近 GT.



a. 2D 输入　　b. Base　　c. Base+Rp

d. Base+Rp+SF　　e. Base+Rp+SF+TE　　f. GT

图 4　消融分析定性结果

通过表 7 和图 4 的定量和定性结果对比，验证了本文方法的有效性，提出的 3 个技术要点在提升姿态估计结果的精度和平滑度上均有重要的作用，且结合构建的 3 个模块进行三维人体估计，能获得最优的效果.

4.5.2　空间多尺度有效性分析

针对不同的人体标记化策略进行消融实验分析，以 1-hop/2-hop 邻节点聚合的方式划分出包括

在 4 种尺度下对应人体标记数分别为 17, 13, 5 和 1 的策略, 以及在 3 种尺度下人体标记数分别为 17, 7 和 1 的策略. 表 8 展示了不同的空间多尺度表征策略的姿态估计结果, 使用以 81 帧为输入时的 MPJPE 进行评估分析, 从表 8 可以看出, 本文所选取的策略 4 达到最低的 MPJPE, 即性能最优. 策略 1 作为基准只使用单一关节点尺度表征人体姿态.

表 8 人体空间多尺度表征消融分析

策略	标记数	MPJPE/mm
1	17	44.8
2	17, 7, 1	44.2
3	17, 13, 5, 1	44.7
4	17, 6, 2	**43.4**

注. 粗体表示最优值.

为进一步验证人体空间多尺度特征对三维人体姿态估计结果精度的影响, 本文还进行了多尺度特征融合方式的消融实验, 结果如表 9 所示. 对比 2 种深度学习中常见的特征融合方式级联和对应元素相加, 本文提出的基于自注意力的人体空间多尺度特征融合方式大幅降低了生成结果的 MPJPE, 证明了该模块的有效性.

表 9 人体特征融合消融分析 mm

特征融合方式	MPJPE	P-MPJPE
级联	44.6	34.5
对应元素相加	46.7	35.9
本文	**43.4**	**33.6**

注. 粗体表示最优值.

4.5.3 时序多尺度有效性分析

本文利用人体姿态序列的多尺度时序特征学习深度信息并平滑姿态估计结果. 表 10 所示为不同输入序列长度的消融分析.

总体来看输入视频长度越长, MPJPE 和 MPJVE 越低, 代表模型的精度和输出视频的平滑度越好, 其中, $F=81$时模型达到最好效果. 受限于硬件性能, 本文未测试更长的输入序列. 由此说明时序信息能有效地用于模型获取深度信息以及对视频帧间进行平滑处理.

表 10 不同输入序列长度消融分析 mm

输入视频长度/帧	MPJPE	MPJVE
9	48.2	3.74
27	45.0	2.61
81	43.4	2.43

4.5.4 超参数消融分析

本文以 81 帧输入为例对超参数的消融分析如表 11 所示, 以更好地说明模型超参数对实验结果的影响. 从第 2 和第 3 行以及第 5 和第 6 行可以看出, 在其他参数一致的条件下, $B=3$时的 MPJPE 均低于 $B=2$ 与 $B=4$, 且取 $B=3$时的模型 P 以及 FLOPs 并未增加太多, 对计算的负担处于可接受范围内. 对比第 1 和第 3 行可知, 时序多尺度编码阶段数 $K=2$时, MPJPE 高于 $K=3$时, 说明模型获取的多尺度时序信息不饱和; 对比第 4 行和第 6 行可知, $K=4$时 MPJPE 明显高于 $K=3$时, 说明模型获取的多尺度时序信息冗余, 且参数量和浮点运算量也大幅提高, 增加了计算成本. 对比 $c=48$ 与 $c=32$时, 模型计算复杂度大幅上升, 而 MPJPE 无显著变化. 综上, 最终设置 $K=3$, $B=3$, $c=32$.

表 11 超参数消融分析

超参数			$10^{-6}\times P$	$10^{-9}\times$FLOPs	MPJPE/mm
K	B	c			
2	3	32	12.08	0.587	43.9
3	2	32	9.93	0.602	44.6
3	3	32	13.27	0.782	**43.4**
3	4	48	29.37	1.877	43.6
4	3	48	29.91	1.910	44.1
4	4	48	35.98	2.298	44.5

注. 粗体表示最优值.

5 结 语

针对视频输入的三维人体姿态估计存在姿态表征不精确、特征融合不充分、估计结果不平滑等问题, 本文通过构建多尺度表征、空间多尺度特征融合以及时序多尺度编码模块, 实现了人体姿态的多尺度时空特征融合. 空间多尺度表征模块提取 3 种空间尺度下的人体姿态特征, 提高了复杂人体姿态的表征能力; 空间多尺度特征融合模块融合空间多尺度人体姿态特征, 提升三维人体姿态估计精度; 针对时序输出不平滑、计算复杂度高的问题, 时序多尺度编码模块拟合输入序列中的多尺度时序信息, 对三维人体姿态特征进行平滑处理, 并在一定程度上降低计算复杂度; 最终生成平滑、精确的三维人体姿态估计结果.

本文方法能以相对较低的计算复杂度估计得到较精确且平滑的三维人体姿态序列, 但使用更长的输入(如 243 帧、351 帧)时仍受到算力限制. 此外, 本文方法对遮挡以及实验室环境下无法采集

的复杂人体姿态等场景的估计准确性仍有进一步提升的空间. 本文后续工作将重点围绕上述难点展开, 进一步提高在现实场景中的三维人体姿态估计精度, 并优化网络结构以降低计算复杂度.

参考文献(References):

[1] Wang J B, Tan S J, Zhen X T, *et al*. Deep 3D human pose estimation: a review[J]. Computer Vision and Image Understanding, 2021, 210: Article No.103225

[2] Garau N, Conci N. CapsulePose: a variational CapsNet for real-time end-to-end 3D human pose estimation[J]. Neurocomputing, 2023, 523: 81-91

[3] 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. Los Alamitos: IEEE Computer Society Press, 2017: 2659-2668

[4] Wan Yunchong, Song Yunpeng, Liu Ligang, *et al*. 3D human pose estimation based on volumetric joint coordinates[J]. Journal of Computer-Aided Design & Computer Graphics, 2022, 34(9): 1411-1419(in Chinese)
(万云翀, 宋云鹏, 刘利刚. 基于体素联合坐标的单人三维姿态估计[J]. 计算机辅助设计与图形学学报, 2022, 34(9): 1411-1419)

[5] Mehta D, Sridhar S, Sotnychenko O, *et al*. VNect: real-time 3D human pose estimation with a single RGB camera[J]. ACM Transactions on Graphics, 2017, 36(4): Article No.44

[6] Chen T L, Fang C, Shen X H, *et al*. Anatomy-aware 3D human pose estimation with bone-based pose decomposition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(1): 198-209

[7] Bai G H, Luo Y M, Pan X L, *et al*. Double chain networks for monocular 3D human pose estimation[J]. Image and Vision Computing, 2022, 123: Article No.104452

[8] Tang Z H, Hao Y B, Li J, *et al*. FTCM: frequency-temporal collaborative module for efficient 3D human pose estimation in video[OL]. [2023-04-21]. https://ieeexplore.ieee.org/document/10159259

[9] Yu B X B, Zhang Z, Liu Y X, *et al*. GLA-GCN: global-local adaptive graph convolutional network for 3D human pose estimation from monocular video[OL]. [2023-04-21]. https:// arxiv.org/abs/2307.05853

[10] Li W H, Liu H, Tang H, *et al*. MHFormer: multi-hypothesis transformer for 3D human pose estimation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 13137-13146

[11] Li S J, Chan A B. 3D human pose estimation from monocular images with deep convolutional neural network[C] //Proceedings of the 12th Asian Conference on Computer Vision. Heidelberg: Springer, 2014: 332-347

[12] Pavlakos G, Zhou X W, Derpanis K G, *et al*. Coarse-to-Fine volumetric prediction for single-image 3D human pose[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 1263-1272

[13] Li J F, Bian S Y, Zeng A L, *et al*. Human pose regression with residual log-likelihood estimation[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 11005-11014

[14] Zhou K, Han X G, Jiang N J, *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. Los Alamitos: IEEE Computer Society Press, 2019: 2344-2353

[15] Zhou X Y, Huang Q X, Sun X, *et al*. Towards 3D human pose estimation in the wild: a weakly-supervised approach[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 398-407

[16] Zhao L, Peng X, Tian Y, *et al*. Semantic graph convolutional networks for 3D human pose regression[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 3420-3430

[17] Gong K H, Zhang J F, Feng J S. PoseAug: a differentiable pose augmentation framework for 3D human pose estimation[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 8571-8580

[18] Wehrbein T, Rudolph M, Rosenhahn B, *et al*. Probabilistic monocular 3D human pose estimation with normalizing flows[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 11179-11188

[19] Wu L L, Yu Z B, Liu Y J, *et al*. Limb pose aware networks for monocular 3D pose estimation[J]. IEEE Transactions on Image Processing, 2022, 31: 906-917

[20] Zheng C, Zhu S J, Mendieta M, *et al*. 3D human pose estimation with spatial and temporal transformers[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 11636-11645

[21] Cai Y J, Ge L H, Liu J, *et al*. Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 2272-2281

[22] Zhang J L, Tu Z G, Yang J Y, *et al*. MixSTE: seq2seq mixed spatio-temporal encoder for 3D human pose estimation in video[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 13222-13232

[23] Wang T F, Zhang X X. Simplified-attention enhanced graph convolutional network for 3D human pose estimation[J]. Neurocomputing, 2022, 501: 231-243

[24] Zeng A L, Yang L, Ju X, *et al*. SmoothNet: a plug-and-play network for refining human poses in videos[C] //Proceedings of the 17th European Conference on Computer Vision. Heidelberg: Springer, 2022: 625-642

[25] Li W H, Liu H, Ding R W, *et al*. Exploiting temporal contexts with strided transformer for 3D human pose estimation[J].

IEEE Transactions on Multimedia, 2023, 25: 1282-1293

[26] Zhan Y, Li F H, Weng R L, *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. Los Alamitos: IEEE Computer Society Press, 2022: 13106-13115

[27] Shan W K, Liu Z H, Zhang X F, *et al*. P-STMO: pre-trained spatial temporal many-to-one model for 3D human pose estimation[C] //Proceedings of the 17th European Conference on Computer Vision. Heidelberg: Springer, 2022: 461-478

[28] Chen Y L, Wang Z C, Peng Y X, *et al*. Cascaded pyramid network for multi-person pose estimation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 7103-7112

[29] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need[C] //Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2017: 6000-6010

[30] 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, 2014, 36(7): 1325-1339

[31] Pavllo D, Feichtenhofer C, Grangier D, *et al*. 3D human pose estimation in video with temporal convolutions and semi-supervised training[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 7745-7754

[32] Cheng Y, Yang B, Wang B, *et al*. 3D human pose estimation using spatio-temporal networks with explicit occlusion training[C] //Proceedings of the 34th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020: 10631- 10638

[33] Li W W, Du R, Chen S D. Skeleton-Based spatio-temporal U-Network for 3D human pose estimation in Video[J]. Sensors, 2022, 22(7): Article No.2573

[34] Liu R X, Shen J, Wang H, *et al*. Attention mechanism exploits temporal contexts: real-time 3D human pose reconstruction[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 5063-5072