控制与决策
Control and Decision

一种基于多假设交互的三维人体姿态估计模型

胡楠, 张家豪, 魏晓彤, 朱宏博

引用本文:
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.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2024.1162

一种基于多假设交互的三维人体姿态估计模型

胡 楠[1†]，张家豪[1]，魏晓彤[2]，朱宏博[3]

(1. 沈阳建筑大学 电气与控制工程学院，沈阳 110168；
2. 中国医科大学 医学教育评价与改革研究院，沈阳 110122；
3. 沈阳理工大学 信息科学与工程学院，沈阳 110623)

摘 要: 近年来, 基于 Transformer 的方法在三维人体姿态估计任务中表现出色, 然而, 现有方法虽能通过全局自注意力机制有效建模关节间长程依赖关系, 但在肢体快速运动等场景下易产生局部运动轨迹预测偏差, 存在对局部运动特征建模不足问题. 鉴于此, 提出一种结合卷积神经网络 (CNN) 与混合注意力机制的 Transformer 架构模型, 通过加入卷积特征提取, 显著增强局部关节运动表征能力. 首先, 设计混合多假设生成模块, 兼顾效率的同时生成更丰富的假设信息, 有效弥补传统全局视角方法在捕捉局部依赖关系上的不足; 然后, 使用自假设精细化模块进一步挖掘数据中的多样化信息, 确保模型能够捕捉到更多细节; 最后, 通过跨假设交互模块充分融合不同假设间的特征信息, 增强模型的鲁棒性和精度. 实验结果表明, 该模型在数据集 Human3.6M 上的表现相较于基准模型 MHFormer 提升了 7.99%, 表明了所提出组件与整体架构在三维人体姿态估计领域的有效性.

关键词: 三维人体姿态估计；卷积神经网络；混合注意力机制；混合多假设；自假设精细；跨假设交互

中图分类号: TP391.4 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2024.1162

引用格式: 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.

A 3D human pose estimation model based on multiple hypothesis interaction

HU Nan[1†]，*ZHANG Jia-hao*[1]，*WEI Xiao-tong*[2]，*ZHU Hong-bo*[3]

(1. School of Electrical and Control Engineering， Shenyang Jianzhu University， Shenyang 110168， China；
2. Institute of Health Professions Education Assessment and Reform， China Medical University， Shenyang 110122， China； 3. School of Information Science and Engineering， Shenyang Ligong University， Shenyang 110623，China)

Abstract: In recent years, Transformer-based methods have shown excellent performance in the task of 3D human pose estimation. However, although existing methods can effectively model the long-range dependencies between joints through the global self-attention mechanism, they are prone to prediction biases of local motion trajectories in scenarios such as rapid limb movements, and there is a problem of insufficient modeling of local motion features. To address this issue, this paper proposes a Transformer architecture model that combines aconvolutional neural network (CNN) with a hybrid attention mechanism. By adding convolutional feature extraction, it significantly enhances the representation ability of local joint movements. A hybrid multi-hypothesis generation module (H-MHG) is designed to generate richer hypothesis information while taking efficiency into account, effectively making up for the deficiencies of traditional global perspective methods in capturing local dependencies. Subsequently, the self-hypothetical granular (SHG) module is used to further explore the diverse information in the data, ensuring that the model can capture more details. Finally, through the cross-hypothetical interaction (CHI) module, we fully integrate the feature information among different hypotheses, enhancing the robustness and accuracy of the model. Experimental results show that the performance of this model on the Human3.6M dataset is improved by 7.99% compared with the baseline model MHFormer, demonstrating the effectiveness of the proposed components and the overall architecture in the field of 3D human pose estimation.

收稿日期: 2024-09-30；**录用日期:** 2025-06-09.

基金项目: 国家自然科学基金项目 (72204267)；教育厅基本科研项目面上项目 (JYTMS20231576)；辽宁省教育厅基本科研一般项目 (LJ212410144025).

†通信作者. E-mail: hunan@sjzu.edu.cn.

Keywords: 3D human pose estimation; convolutional neural network; hybrid attention mechanism; hybrid multi-hypothesis; self-hypothetical granular; cross-hypothetical interaction

0 引 言

近年来, 3D 人体姿态估计在计算机视觉领域引起了广泛关注, 它能够在多种应用场景中 (如增强现实、医疗康复、人机交互等) 准确识别和分析人体姿态, 具有广泛的实用价值, 同时推动了深度学习和多模态融合等技术的进步, 是其他高级任务的基础[1]. 然而, 在实际应用中, 姿态估计面临着诸多挑战, 如视角变化、人体遮挡、多样化动作等问题, 这些因素都会影响姿态估计的准确性和鲁棒性[2-4]. 在 3D 人体姿态估计研究中, 主要有两大类方法: 端到端方法和两阶段方法. 端到端方法直接从图像或视频中估计出三维姿态[5], 这类方法可以充分利用深度学习模型的强大特征提取能力, 但在处理复杂背景和多样化动作时可能表现不够稳定. 相反, 两阶段方法首先估计二维关键点, 然后将其转换为三维姿态[6-7], 这种方法通常使用卷积神经网络 (convolutional neural network, CNN)[8] 和序列模型, 如长短期记忆网络 (long short-term memory, LSTM)[9] 或 Transformer, 以有效处理姿态估计中的时空特征[10-11]. 本文聚焦于两阶段的三维姿态估计方法, 探索其在高精度姿态预测中的应用潜力.

国内外研究围绕两阶段方法的核心挑战形成了互补的技术路径. PoseFormer[3] 开创性地构建时空双分支 Transformer, 通过空间自注意力解析关节拓扑、时间注意力捕捉运动轨迹, 为视频姿态估计奠定基础框架, 但其注意力机制导致长序列计算复杂度激增. 在此基础上的 HDFormer[12] 提出高阶骨骼关系建模, 通过关节-骨骼-超骨骼三级图注意力网络, 将复杂遮挡场景下的关节定位误差降低, 其 U 型 CNN-Transformer 混合架构使推理速度提升, 显著改善了实用性. HSTFormer[13] 则通过分层时空编码器实现多粒度特征提取, 其肢体级时间 Transformer 使快速肢体运动的预测精度提升, 但层级间特征融合的冗余计算仍限制其部署效率. Wang 等[14] 提出的轻量化跨模态框架, 通过姿态先验蒸馏策略在保持精度的同时压缩模型参数量, 为移动端部署提供了新思路. 值得关注的是, Li 等[7] 提出的 MHFormer 多假设生成机制方法通过一个二维关节点坐标预测多个三维人体姿态估计假设, 通过融合这些假设, 得出接近真实值的三维人体姿态估计结果. 然而, 上述方法在复杂运动场景下仍存在一些问题: 其一, 局部运动特征与全局姿态的关联问题; 其二, 模型针对长序列动作依赖的建模效率与精度难以平衡.

基于这些思考, 本文提出一种新的三维姿态估计方法, 旨在从时序和空间两个维度上全面分析和理解人体姿态变化. 工作主要从以下几个角度进行: 1) 引入混合多头自注意力机制: 解决姿态估计中的时序依赖性问题, 通过混合多头自注意力机制, 使模型能够有效捕捉输入序列中各个位置之间的依赖关系, 提高时序特征的建模能力. 2) 融入卷积特征提取器: 将卷积特征提取器融入模型中, 利用卷积操作在局部时间窗口内提取高效的特征表示, 增强模型对关键点细微运动的预测精度, 特别是局部特征的捕捉能力. 3) 假设优化模块: 针对姿态估计的全局建模需求, 建立一种假设优化模块. 通过自假设精细化和跨假设交互机制, 融合不同假设间的特征信息, 实现更高精度的三维姿态预测.

本文的组织结构如下: 第 1 部分描述所提出的模型架构和各个功能模块, 包括混合多头自注意力机制、卷积特征提取器和假设优化模块, 并解释这些模块如何共同作用以提高姿态估计的精度和鲁棒性; 第 2 部分在公开数据集上进行实验以评估所提出方法, 同时通过与主流方法的对比分析, 证明方法的性能优势; 第 3 部分总结全文工作, 讨论研究的主要贡献和发现, 并展望未来的研究方向.

1 模型架构

1.1 基于 CNN 与 Transformer 融合的多假设交互模型架构

图 1 展示了本文所研究的模型架构, 其中 (a) 为混合多假设产生模块, (b) 为自假设精细化模块, (c) 为跨假设交互模块. 该网络使用二维人体坐标序列作为网络的输入, 混合多假设产生模块 (hybrid multi-hypothesis generation module, H-MHG) 通过混合多头注意力机制生成多组初始假设, 解决全局依赖与局部特征的平衡问题; 自假设精细化模块 (self-hypothetical granular module, SHG) 通过并行注意力分支细化单假设内部的特征; 跨假设交互模块 (cross-hypothetical interaction Module, CHI) 则通过跨假设交互整合多组假设的优势, 形成互补的特征表达. 下面对上述内容展开详细介绍.

1.2 混合多假设生成模块

1) 混合多头注意力方法实现.

为了更好地符合人体运动学中肢体局部运动与整体姿态的耦合特性, 本文结合标准多头自注意力

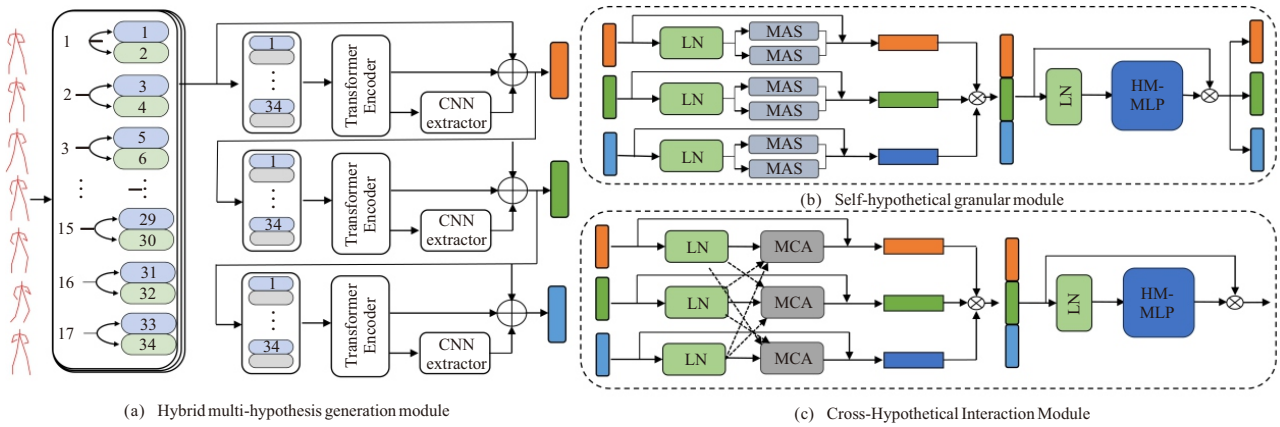

图1 基于 CNN 与 Transformer 融合的多假设交互模型架构

机制 (multi-head self-attention, MSA) 和 Nyström 方法的优点, 在保留全局依赖关系的同时, 通过局部地标点采样增强对快速运动关节的关注, 旨在提高注意力机制的表示能力, 同时减少计算复杂度, 如图2 所示. 多头自注意力机制通过多个注意力头并行处理输入数据, 捕捉输入序列中每个位置之间的全局依赖关系. 尽管其表达能力强大, 但计算复杂度较高, 尤其是在处理长序列时.

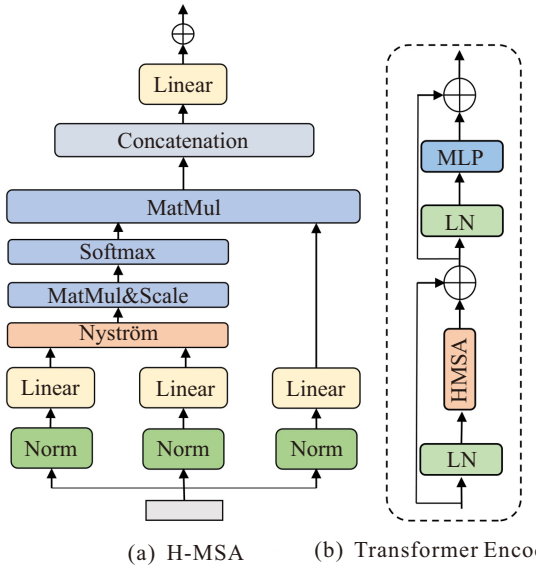

图2 混合多头自注意力机制模块结构

Nyström[15] 方法是一种用于近似求解大型矩阵特征值分解的技术, 通过选择矩阵的一个子集 (地标点) 进行低秩近似, 从而显著减少计算复杂度. 虽然其引入了近似, 但在保持较高准确度的前提下, 大幅度降低了计算开销.

在注意力机制中, Nyström 方法被用于近似计算注意力矩阵, 以减少计算复杂度[16]. 在计算注意力矩阵 $A=\mathrm{Softmax}(QK^{\mathrm{T}}/\sqrt{d_k})$时, 使用 Nyström 方法对 QK^{T}进行低秩近似. 使用采样的地标点计算 Q_m和K_m, 并通过这两个矩阵的乘积来近似QK^{T}后通过 Softmax 函数计算得到近似的注意力得分矩阵

$$\tilde{A}=\mathrm{Softmax}(QK_m^{\mathrm{T}}(Q_mK_m^{\mathrm{T}})^{-1}Q_mK^{\mathrm{T}}/\sqrt{d_k}). \quad (1)$$

通过这种方式, 能够将相似度矩阵 A 分解为 3 个子矩阵的乘积. 具体而言, 第 1 部分为$QK_m^{\mathrm{T}}\in\mathbb{R}^{N\times m}$, 第 2 部分为$\left(Q_mK_m^{\mathrm{T}}\right)^{-1}\in\mathbb{R}^{N\times m}$, 第 3 部分为$Q_mK^{\mathrm{T}}\in\mathbb{R}^{N\times m}$.

从计算复杂度的角度看, 标准注意力矩阵的计算复杂度为$O(N^2)$. 而在所提出的分解方法中, 第 1 部分的计算复杂度为$O(Nmd)$, 第 2 部分因涉及矩阵求逆操作, 其计算复杂度为$O(m^3)$, 第 3 部分的计算复杂度同样为$O(Nmd)$. 综合起来, 总计算复杂度为$O(Nmd+m^3)$. 考虑到m的值远小于N(即$m\ll N$), 该方法的计算复杂度可近似为$O(N\cdot m)$. 这种分解方式显著降低了计算复杂度, 为相关计算任务带来了更高的效率.

使用近似的注意力得分矩阵$\tilde{A}$对值矩阵V进行加权求和, 得到近似的输出矩阵

$$\tilde{Z}=\tilde{A}V. \quad (2)$$

2) 卷积特征提取器.

与混合多头自注意力模块不同, 卷积特征提取器的核心优势则体现在对局部特征的提取上[17-18]. 局部特征能够体现短时间内姿态变化的细微之处, 这些细节对于识别和预测关键点的微小运动起着至关重要的作用. 卷积神经网络在捕捉局部特征方面颇具优势, 特别是其卷积操作, 可在局部时间窗口内提取高效的特征表示.

在模型中, 卷积特征提取器由多个残差块 (residual blocks) 组成, 每个残差块包含两层一维卷积和批归一化 (batch normalization) 层, 并通过捷径连接 (shortcut connection) 实现残差学习, 在保留输入的原始信息的同时增强模型对局部特征的捕捉能力. 残差块的使用灵感来源于 ResNet[19] 架构, 其优势在于能够缓解深度网络中的梯度消失问题, 从而

实现更深层次的特征学习. 每个残差块后接一个最大池化层 (max pooling), 通过逐步减少特征图的时间维度, 卷积特征提取器能够有效地缩小数据的尺寸, 并增加感受野, 使得卷积层能够捕捉到更大范围内的局部特征, 从而增强模型对关键点细微运动的敏感度, 确保预测结果的精确性, 同时通过 CNN 的池化操作, 减少了特征的时间维度, 使得输入给假设 Transformer 的特征更加紧凑, 有效降低了计算量.

具体而言, 将每帧的二维关节点坐标以序列 $X \in \mathbb{R}^{N\times J\times 2}$ 序列的形式作为输入, 其中 N 代表视频帧数, J 代表关节点数量. 通过可学习的空间位置信量 E_{SPOS} 标记 2D 姿态关节点序列的输入顺序, 并将嵌入的特征输入 H-MHG 的编码器, 可表达为

$$X_0^m = \text{LN}(X^m) + E_{\text{SPOS}}, \tag{3}$$

$$X_l^{'m} = X_{l-1}^m + \text{HMSA}^m(\text{LN}(X_{l-1}^m)), \tag{4}$$

$$X_l^{''m} = X_l^{'m} + \text{MLP}^m(\text{LN}(X_l^{'m})), \tag{5}$$

$$X_L^m = \text{CNN}(X_l^{''m}) + X^m + \text{LN}(X_l^{''m}). \tag{6}$$

其中: LN(·) 为 LayerNorm 层, m 表示输出有 m 个不同的假设, $l \in [1, 2, \ldots, L]$ 为 HMHG 的层数索引. 最后的输出特征可以视为不同姿态假设的初始表示, 但需要进一步增强.

1.3　假设优化模块

在经过初步的特征提取后, 输入序列的特征表示进一步传递到假设优化模块中. 该模块由自假设精细与多假设交互两部分组成, 进一步对通过 H-MHG 提取的特征进行复杂的时序关系建模增强.

1) 自假设精细化模块.

为了更好地处理输入的二维姿态序列经过 H-MHG 模块后被映射成的特征向量, 提出一个并行的注意力机制的自假设精细化模块, 进一步捕捉数据中的多样化信息, 它可以被比喻为人类同时从不同角度观察同一对象. 例如, 一个人从左边观察, 另一个人从右边观察, 虽然他们看的是同一个对象, 但他们捕捉到的细节可能不同. 通过这样的方式将特征相加融合, 实现了更丰富的上下文建模, 同时捕获互补的特征交互模式更强的单分支表征能力, 每条路径都各自计算一组注意力分数, 并且计算是同步进行的, 如图 1(b) 所示.

首先将每个帧的编码假设特征 $\boldsymbol{X}_L^m$ 嵌入到高维特征 $\boldsymbol{Z}^m \in \mathbb{R}^{N\times C}$, 其中 C 为嵌入维数. 首先将不同假设的值馈送到几个并行 DualAttn 块中, 每个 DualAttn 块由两个 MSA 组成, 表示为

$$Z_l^{'m} = Z_{l-1}^m + \text{DualAttn}(\text{LN}(Z_{l-1}^m)). \tag{7}$$

其中: LN(·) 为 LayerNorm 层, l 为 SHR 各层的指数. 这样不同假设特征的信息可以通过自假设的方式传递, 从而实现各自假设的特征增强. 同时为了各个假设更加精练, 通过假设混合 HM-MLP 对多个假设特征进行连接, 然后再将特征均匀地划分为多个假设表示. 该过程可以表述为

$$Z_l^{'} = \text{Concat}(Z_l^{'1}, \ldots, Z_l^{'m}) \in R^{N\times C}, \tag{8}$$

$$\text{Concat}(Z_l^{'1}, \ldots, Z_l^{'m}) = Z_l^{'} + \text{HMLP}(\text{LN}(Z_l^{'})). \tag{9}$$

其中: Concat 为级联运算, HMLP 为假设混合 HM-MLP 的函数.

2) 跨假设交互模块.

由于 SHR 缺乏跨假设的连接, 限制了其相互作用的建模. 为了更好地实现跨假设通信中多假设间的相互关联, 需要将 SHG 中生成假设进行特征信息交互融合, 使网络能够自适应地寻找最优假设.

如图 1(c) 所示为跨假设交互模块, 该模块通过排列组合的方式融合假设交精细化模块输出的 m 个假设的特征信息, 主要由多假设交互注意力头 (MH-CA) 和 HM-MLP 组成, 多假设交互注意力头的网络结构图如图 3 所示, 每个分支通过聚合其他两个分支的信息, 拼接 3 个分支后通过 HM-MLP 处理, 其中多个假设被交替地视为 Query、Key 和 Value 以实现不同分支间的信息交互, 促进跨模态特征融合. 该过程可以表述为

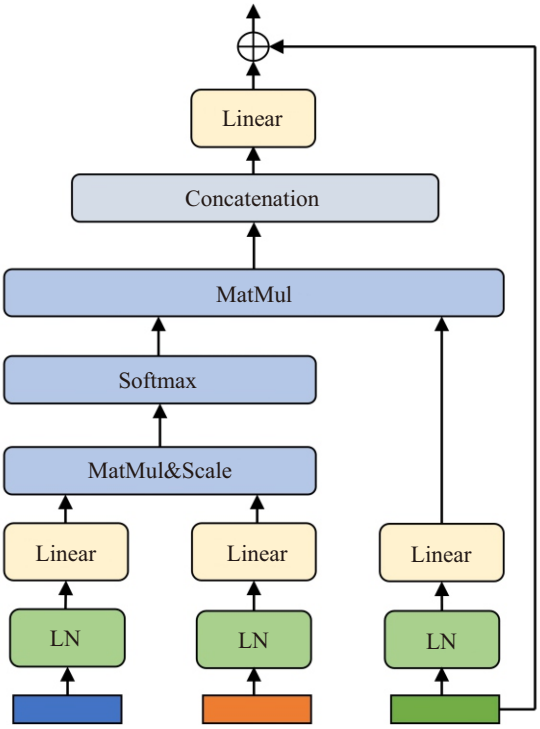

图3　多假设交互注意力头网络结构

$$Z_l^{'m} = Z_{l-1}^{m} + \mathrm{MCA}(\mathrm{LN}(Z_{l-1}^{m_1}, Z_{l-1}^{m_2}, Z_{l-1}^{m})). \tag{10}$$

其中: l为 CHI 层的指数, m_1和m_2为另外两个相应的假设. MCA 模块通过执行跨假设的注意力机制, 具体表现为: 以假设m_1的 Query 检索假设m_2的 Key 矩阵, 聚合假设m的 Value, 从而实现跨假设的特征交叉融合. CHI 中的假设混合 HM-MLP 与 SHR 中相似, 不同的是在最后一层不使用划分操作, 而是使多个假设最终聚合成单个假设表示.

总体而言, 混合多假设生成模块通过混合多头注意力机制和卷积特征提取生成多组初始假设, 解决全局依赖与局部特征的平衡问题. 然后进一步传递到假设 Transformer 模块中, 其中 SHG 模块通过并行注意力分支细化单假设内部的特征, CHI 模块则通过跨假设交互整合多组假设的优势, 形成互补的特征表达. 三者形成递进关系: H-MHG 提供多样化的初始假设, SHG 优化单假设的局部细节, CHI 融合多假设的全局信息, 最终实现高精度三维姿态估计.

1.4 回归预测

经过前述模块的特征提取后, 模型生成了多层次、多维度的特征表示. 具体而言, 为了提高姿态预测的精度, 模型会将来自 Transformer 模块和卷积特征提取器的特征进行展平和拼接, 形成一个高维度的特征向量. 这些特征表示综合了输入序列的全局和局部信息, 具有较高的表达能力. 随后, 在经过跨假设交互模块输出后, 模型通过一系列全连接层 (FC Layers) 对这些高维特征进行逐步降维和非线性变换. 首先, 特征向量通过一个大小为 2 048 的全连接层进行处理, 并应用 ReLU 激活函数以引入非线性特征; 然后, 经过若干层全连接网络, 逐步降维至所需的输出维度; 最后, 模型输出每个时间步上所有关节点的三维坐标, 完成姿态回归预测. 整个模型以端到端的方式进行训练, 采用均方误差 (mean squared error, MSE) 损失, 应用于最小化估计与地面真实姿态之间的误差.

2 实验结果及分析

2.1 数据集介绍

本文采用由卡内基梅隆大学与洪堡大学合作开发的 Human3.6M 数据集, 这是目前最大且应用最广泛的 3D 人体姿态估计数据集之一. 该数据集包含 11 名专业演员在实验室环境中执行的多种动作, 涵盖了如走路、跑步、坐下、打电话、吃饭、遛狗等 15 种不同场景, 丰富多样的动作为研究提供了极具价值的素材. 部分数据示例如图 4 所示. 数据集通过多个相机从不同角度记录这些动作, 确保了高质量的 3D 姿态标注和丰富的视角信息. 通过充分利用这一数据集, 研究能够有效捕捉人体动态变化的细微差别, 也为模型在训练过程中提供了更为全面的姿态信息, 显著提高了模型的泛化能力, 从而使其在实际应用中表现得更为可靠和精准. 采用 P-MPJPE (procrustes-aligned mean per joint position eError) 作为评估指标, 用于衡量模型预测的三维关节坐标与真实坐标之间的误差.

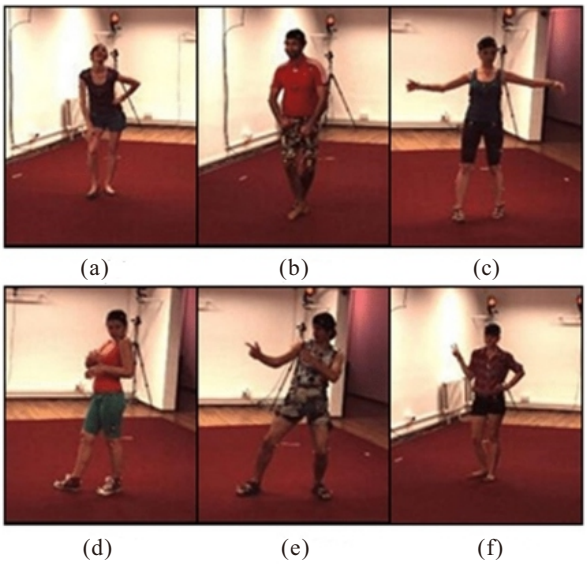

图4 Human3.6M 的部分数据

2.2 实验环境构建与参数

本文的实验研究搭建了基于 Python 3.8 和 PyTorch 2.2.2 框架的实验环境. 实验硬件平台配置为: Intel (R) Xeon (R) Platinum 8362 CPU @ 2.80 GHz 处理器, NVIDIA GeForce RTX 3090T (24 GB) 图形处理器, 32 GB 内存, Windows 10 操作系统. 软件环境的部分依赖库版本如下: NumPy 1.24.4: 用于提供高效的多维数组操作; Einops 0.8.0: 用于张量的重组和变换操作; Matplotlib 3.7.1: 用于数据可视化和结果展示; OpenCV-Python 4.10: 用于计算机视觉任务的处理.

2.3 基线与对比方法参数设置

实验参数设置如下: 卷积神经网络特征提取器包含 4 个残差块, 每个块的输入输出通道数设置为 512. Transformer 编码器部分, 使用 3 个编码器, 每个编码器具有 3 层深度, 嵌入维度为 512, 隐藏层维度为 1 024, 多头注意力头数为 8, Nyström 方法中使用的 landmarks 数量为 64. 同时, 为了更好地优化模型, 使用 Amsgrad 优化器, 初始学习率设置为 0.001, 使用权重衰减参数以防止过拟合.

2.4 消融实验

为了验证所提出的模型中每个组件和设计的影响, 进行广泛的消融实验. 首先评估不同模块对人体姿态估计性能的贡献, 在基准模型 MHFormer 的基础上进行消融实验. 本次实验的主要目的是分析 H-MHG、SHG 以及 H-MHG + SHG 这 3 个修改后的模型相较于 MHFormer 的性能变化, 并进一步探讨在这些组合方法的基础上结合 CNN 模块后模型的表现情况. 本次消融实验中, 从 MHFormer 基准模型出发, 逐步添加或修改其组件, 15 种不同场景动作的结果以及平均值如表 1 所示. 在 MHFormer的基础上引入 H-MHG 模块, 添加混合多头注意力机制, 由表 1 可以看出采用 H-MHG 相较于 MHFormer 提升 1.69%. 在 MHFormer 基础上, 替换部分原始模块以采用 SHG, 这一修改旨在通过并行的注意力机制的自假设精细化模块进一步捕捉数据中的多样化信息, 优化模型对姿态变化的感知能力, 可以看出采用 SHG 相较于 MHFormer 提升 2.66%. 将 H-MHG 和 SHG 这两种修改进行组合, 以其在两者的优势基础上, 进一步提升模型性能的 H-MHG + SHG 模型相较于 MHFormer 提升了 3.63%. 最后, 融合 CNN 模型, 最终的模型 H-MHFormer 相较于 MHFormer 提升了 7.99%.

表1 引入不同组件下的动作检测的 P-MPJPE(mm) 值的比较

	Dir.	Disc.	Eat	Great	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT	Avg.
MHFormer	44.2	41.8	40.0	44.8	38.9	43.4	44.0	40.9	37.5	36.1	38.7	41.6	40.6	42.0	44.5	41.3
H-MHG	43.3	41.1	39.4	44.0	38.1	42.9	43.3	40.2	37.1	35.9	38.1	40.9	39.9	40.9	43.7	40.6
SHG	43.0	40.9	39.2	43.7	38.0	42.7	42.8	40.0	37.0	35.4	37.9	40.6	39.4	40.0	42.6	40.2
H-MHG+SHG	43.0	40.9	39.1	43.5	37.8	42.6	42.7	39.9	37.0	35.3	37.7	40.5	39.2	39.8	42.4	39.8
Ours	40.6	38.6	37.4	41.1	36.0	40.5	40.0	38.4	35.3	33.7	35.9	38.0	37.7	37.5	39.8	38.0

在 3D 人体姿态估计任务中, 不同动作类别因运动模式、时序依赖性和关节交互复杂度的差异, 对模型的性能提出了多样化的挑战. 下面将从静态动作、动态动作和复杂交互动作 3 个维度, 结合消融实验数据展开深入分析.

静态动作不同组件下的消融实验结果如表 2 所示. 静态动作的误差主要来源于长时间姿态保持导致的关节位置模糊性, 以及遮挡场景下的局部特征丢失. 例如, 坐姿中骨盆与椅面的接触区域易产生遮挡噪声, 而等待动作中细微的肢体晃动可能导致时序预测的累积误差.

表2 静态动作下引入不同组件的 P-MPJPE(mm) 值

	MHFormer	H-MHG	SHG	H-MHG+SHG	Ours
SitD.	36.1	35.9	35.4	35.3	33.7
Wait	41.6	40.9	40.6	40.5	38.0

具体而言, 在 SitD.任务中, SHG 的下降相对降幅 1.94% 比 H-MHG 相对降幅 0.55% 更显著; Wait 任务中也是如此, 该现象表明静态场景下, 通过自假设细化来挖掘单假设内部潜在信息, 比多假设生成了更能有效应对姿态模糊性问题. 当引入 H-MHG + SHG 后相比基准方法下降幅度最大, 表明混合注意力生成的多假设为精细化处理提供了更优的初始表征. 在最后引入 CNN 局部特征提取后本文的模型的误差从基线 MHFormer 的 37.5 显著降至 33.7 (降幅 10.1%), Wait 静态场景下, 误差从 41.6 (MHFormer) 降至 38.0 (降幅 8.7%), 表明本文方法在遮挡条件下通过局部特征的构建提升了姿态恢复能力, 验证了模型在低运动幅度场景下的稳定性.

动态动作和复杂动作在不同组件下的消融实验结果如表 3 和表 4 所示. 动态动作要求模型捕捉运动轨迹的连续性和加速度变化规律, 复杂动作的核心挑战是需要同时建模肢体-物体交互关系和细粒度关节协作. 从表 3 的消融实验结果看, 本文模型在 WalkT 上的误差从基线 44.5 大幅降至 39.8 (降幅 10.6%), 验证了局部特征与全局特征的融合对于群

表3 动态动作下引入不同组件的 P-MPJPE(mm) 值

	MHFormer	H-MHG	SHG	H-MHG+SHG	Ours
Walk	42.0	40.9	40.0	39.8	37.5
WalkD	40.6	39.9	39.4	39.2	37.7
WalkT	44.5	43.7	42.6	42.4	39.8

表4 复杂动作下引入不同组件的 P-MPJPE(mm) 值

	MHFormer	H-MHG	SHG	H-MHG+SHG	Ours
Phone	43.4	42.9	42.7	42.6	40.5
Smoke	38.7	38.1	37.9	37.7	35.9

体交互场景拥有强建模能力. 从表 4 的消融实验结果看, 在 Smoke 场景下, 基线 MHFormer 的误差为 38.7, 本文模型则将其降低至 35.9, 误差降幅达到 7.2%. 与之形成鲜明对比的是, 其他模块与基线相比, 误差下降幅度极小. 这些数据有力地验证了, 通过对局部特征与全局特征进行有机融合, 能够切实提升三维姿态估计结果的准确性, 有效降低误差.

针对混合多假设生成模块中参数的影响进行消融实验, 以探究混合多假设生成模块中参数 (M、L、R、T) 对模型性能的影响. 其中: M 为假设数量, L 为 Transformer 层数, R 为 CNN 中 ResidualBlock 数量, T 为目标平均推理时间. 首先明确 ResidualBlock 数量在特定条件下如何影响 P-MPJPE 值, 进而确定使模型性能达到最优的参数配置. 通过控制变量 L 和 R 分别进行实验, 结果如表 5 和表 6 所示.

表5 L 固定条件下的消融实验 P-MPJPE(mm) 值比较

M	L	R	T	P-MPJPE
3	3	3	4.18	38.4
3	3	4	4.24	38.0
3	3	5	4.32	38.7

表6 R 固定条件下的消融实验 P-MPJPE(mm) 值比较

M	L	R	T	P-MPJPE
3	2	4	4.02	41.3
3	3	4	4.24	38.0
3	4	4	4.41	39.0

实验结果表明, 在 $M = 3$ 和 $L = 3$ 的情况下, 随着 R 从 3 增加到 4, 然后到 5, P-MPJPE 的值首先减小然后增大, 同时由于残差快的增加, 模型的耗时情况也在增加. 最优结果出现在 $R = 4$ 时, 此时 P-MPJPE 为 38.0 mm. 这表明在该配置下, 增加 ResidualBlock 的数量到 4 可以优化模型的性能, 但进一步增加到 5 反而导致性能下降.

在 $M = 3$ 和 $R = 4$ 的情况下, 随着 L 从 2 增加到 3, P-MPJPE 的值显著下降, 从 41.3 mm 降低到 38.0 mm, 表明适当增加网络深度能有效提升模型性能. 然而, 当 L 继续增加到 4 时, P-MPJPE 反而略微上升到 39.0 mm, 显示过多的 Transformer 层数不仅会导致性能轻微下降, 同时也会加重计算负担, 可能是由于模型过拟合或计算复杂度增大带来的负面影响.

基线方法 MHFormer 的目标平均推理时间为 2.72 ms, 而所提出方法是 4.24 ms. 从表 1 可以看出, 基线方法 MHFormer的平均误差值大于所提出方法的平均误差值, 基线方法 MHFormer 的平均误差值为 41.3, 所提出方法的平均误差值为 38.0, 表明所提出方法整体上表现更好, 误差值更低, 较低的平均误差意味着所提出方法能够提供高质量的姿态估计结果.

2.5 与其他先进方法性能对比

为了更好地体现本文方法的优势, 选择与人体姿态估计主流方法进行对比, 如图 5 所示. 图中横坐标表示不同动作, 纵坐标表示 P-MPJPE 值, P-MPJPE 值越小代表性能越好. 从整体结果上看, 本文方法取得了较优的识别结果, 且在各个动作识别上都取得了更好的效果.

与人体姿态估计主流方法在 Human3.6M 数据集的对比结果如表 7 所示. 可以看出, 本文方法在多个动作识别任务上都显现出显著的优势, 特别是在吃饭 (Eat)、坐着 (Sit) 和吸烟 (Smoke) 等动作上, P-MPJPE 达到了最低值, 较其他方法有明显改进. 这表明, 在这些动作的识别和预测上, 本文方法能够更精准地捕捉关键点的微小变化.

在 Eat 动作取得优势体现了结合卷积特征提取器后在捕捉进食过程中微小姿态变化方面的出色表

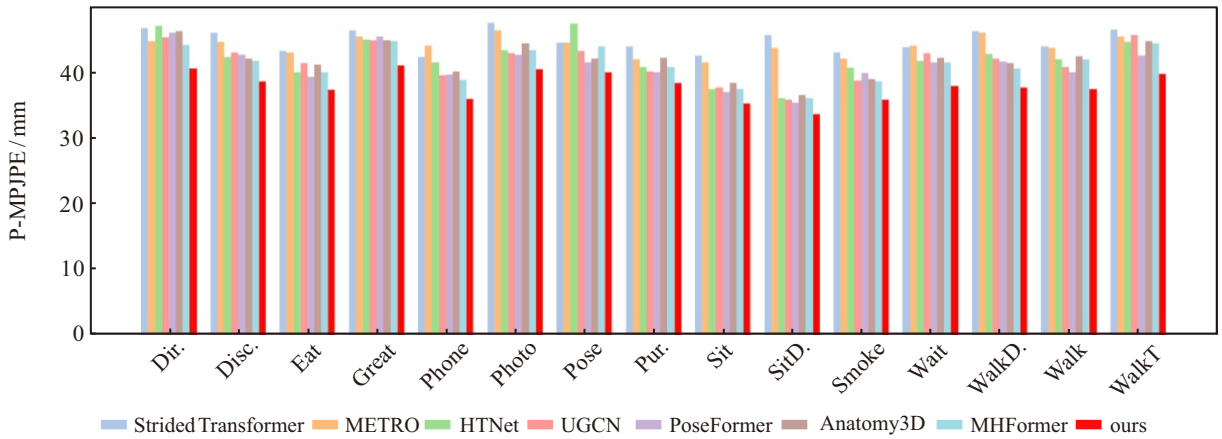

图5 在 Human3.6M 数据集下本文提出方法与主流人体姿态估计方法的 P-MPJPE(mm) 值比较结果

表7　在 Human3.6M 数据集下本文提出方法与主流人体姿态估计方法的 P-MPJPE(mm) 值比较结果

	Dir.	Disc.	Eat	Great	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT
Strided Transformer[20]	46.8	46.1	43.3	46.4	42.4	47.6	44.6	44.0	42.6	45.7	43.1	43.9	46.3	44.0	46.5
METRO[21]	44.8	44.7	43.1	45.5	44.1	46.4	44.6	42.0	41.6	43.7	42.1	44.1	46.1	43.7	45.5
HTNet[22]	47.1	42.4	40.0	45.0	41.5	43.4	47.5	40.9	37.5	36.1	40.7	41.8	42.8	42.0	44.7
UGCN[23]	45.4	43.1	41.4	44.9	39.6	42.9	43.3	40.2	37.7	35.9	38.8	42.9	42.1	40.9	45.7
PoseFormer[3]	46.1	42.7	39.4	45.5	39.7	42.7	41.5	40.0	37.0	35.4	39.9	41.6	41.7	40.0	42.6
Anatomy3D[24]	46.3	42.1	41.2	44.9	40.1	44.4	42.1	42.3	38.4	36.6	39.0	42.2	41.4	42.5	44.8
MHFormer[7]	44.2	41.8	40.0	44.8	38.9	43.4	44.0	40.9	37.5	36.1	38.7	41.6	40.6	42.0	44.5
Ours	40.6	38.6	37.4	41.1	36.0	40.5	40.0	38.4	35.3	33.7	35.9	38.0	37.7	37.5	39.8

现, 使得局部特征得以有效提取, 并为后续的姿态预测提供了更加精确的输入. 在 Sit 动作上误差的减少, 验证了模型在静态动作识别上的鲁棒性, 由于卷积特征提取器能够捕捉姿态局部特征的细节, 即便是较为静态的坐姿, 本文方法也能精准定位关键点, 减少预测误差. 而在 Smoke 动作结果上的这个差距进一步表明了卷积特征提取器和混合多头自注意力机制的有效结合能够在捕捉局部特征的同时, 通过自注意力模块识别全局依赖, 特别是在复杂动作 (如吸烟) 中的细微手部动作识别上, 表现尤为突出.

总体而言, 通过对表 7 的分析, 可以清楚地看到, 本文方法不仅在单个动作上表现优异, 而且在不同类型的动作上都展示出稳定的性能. 这一切都得益于模型结构中卷积特征提取器对局部特征的精细捕捉, 以及混合多头自注意力模块对全局依赖的有效建模. 通过这些优势, 本文方法不仅在静态姿态估计任务中表现出色, 也在复杂动态变化的姿态识别中展现了强大的能力.

3　结　论

本文针对单目视频三维人体姿态估计这一计算机视觉领域的核心挑战, 构建了融合卷积神经网络与混合注意力机制的 Transformer 架构深度学习模型. 针对现有方法在时序建模中存在的局部特征丢失和全局依赖关系建模不足的共性问题, 提出了通过跨模态特征融合机制, 将卷积操作的局部特征提取优势与 Transformer 的多头自注意力全局建模能力相结合, 同时利用假设间自假设细化模块和跨假设交互模块对多假设之间进行信息交互, 实现对三维人体姿态的估计. 实验在 Human3.6M 数据集上进行, 结果表明, 在算法耗时情况方面, 由于混合多头注意力机制与卷积特征提取局部特征设计引入了额外的计算复杂度, 本方法相较于基线方法存在劣势, 但混合注意力机制与卷积特征提取器的结合使模型在三维人体姿态估计中表现出良好的精度和鲁棒性, 不仅在简单静态场景中表现良好, 而且在动态场景中同样具有优越的性能, 验证了所提出方法的有效性.

参考文献 (References)

[1] 郝鹤菲, 张龙豪, 崔洪振, 等. 深度神经网络在人体姿态估计中的应用综述[J]. 计算机工程与应用, 2025, 61(9): 41-60.
(Hao H F, Zhang L H, Cui H Z, et al. Review of application of deep neural networks in human pose estimation[J]. Computer Engineering and Applications, 2025, 61(9): 41-60.)

[2] Hardy P, Kim H. Unsupervised multi-person 3D human pose estimation from 2D poses alone[C]. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, 2024: 4599-4603.

[3] Zheng C, Zhu S J, Mendieta M, et al. 3D human pose estimation with spatial and temporal transformers[C]. 2021 IEEE/CVF International Conference on Computer Vision. Montreal, 2021: 11636-11645.

[4] 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.)

[5] Yu Z B, Ni B B, Xu J W, et al. Towards alleviating the modeling ambiguity of unsupervised monocular 3D human pose estimation[C]. 2021 IEEE/CVF International Conference on Computer Vision. Montreal, 2021: 8631-8631.

[6] Zhao L, Zheng Y, Wang X, et al. PSTNet: Point spatio-temporal convolution on point cloud for 3D human pose estimation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, 2022: 7043-7052.

[7] Li W, 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. New Orleans, 2022: 13137-13146.

[8] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]. 2012 Advances in Neural Information Processing Systems. Lake Tahoe, 2012: 1097-1105.

[9] Hossain M R I, Little J J. Exploiting temporal information for 3D human pose estimation[M]. Computer Vision — ECCV 2018. Cham: Springer International Publishing, 2018: 69-86.

[10] Yang S, Quan Z B, Nie M, et al. TransPose: Keypoint localization via transformer[C]. 2021 IEEE/CVF International Conference on Computer Vision. Montreal, 2021: 11782-11792.

[11] 杨傲雷, 周应宏, 杨帮华, 等. 基于 Transformer 的三维人体姿态估计及其动作达成度评估[J]. 仪器仪表学报, 2024, 45(4): 136-144.
(Yang A L, Zhou Y H, Yang B H, et al. Transformer-based 3D Human pose estimation and action achievement evaluation[J]. Chinese Journal of Scientific Instrument, 2024, 45(4): 136-144.)

[12] Chen H, He J Y, Xiang W, et al. Hdformer: High-order directed transformer for 3d human pose estimation[J/OL]. 2023, arXiv: 2023.02.01825.

[13] Qian X, Tang Y, Zhang N, et al. Hstformer: Hierarchical spatial-temporal transformers for 3d human pose estimation[J/OL]. 2023, arXiv: 2023.01.07322.

[14] Wang C, Wang Y, Lin Z, et al. DistillPose: Lightweight 3D human pose estimation distilled from sparse 2D Poses[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, 2021: 13284-13293.

[15] Williams C K I, Seeger M. Using the nystrom method to speed up kernel machine[C]. Proceedings of the 14th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2023: 661-667.

[16] Xiong Y Y, Zeng Z P, Chakraborty R, et al. Nyströmformer: A nyström-based algorithm for approximating self-attention[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(16): 14138-14148.

[17] 王勇, 李邑灵, 苗夺谦, 等. 基于 Transformer-CNN 混合架构的跨模态融合抓取检测[J]. 控制与决策, 2024, 39(11): 3607-3616.
(Wang Y, Li Y L, Miao D Q, et al. Cross-modal interaction fusion grasping detection based on Transformer-CNN hybrid architecture[J]. Control and Decision, 2024, 39(11): 3607-3616.)

[18] 郭崇, 刘晟, 张文波, 等. 基于卷积混合注意力机制的多目标跟踪算法[J]. 控制与决策, 2025, 40(4): 1127-1135.
(Guo C, Liu S, Zhang W B, et al. Multi-target tracking algorithm based on convolutional hybrid attention mechanism[J]. Control and Decision, 2025, 40(4): 1127-1135.)

[19] 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.

[20] 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.

[21] Lin K, Wang L J, Liu Z C. End-to-end human pose and mesh reconstruction with transformers[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, 2021: 1954-1963.

[22] Cai J L, Liu H, Ding R W, et al. HTNet: Human Topology aware network for 3d Human pose estimation[C]. IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes Island, 2023: 1-5.

[23] Wang J B, Yan S J, Xiong Y J, et al. Motion guided 3D pose estimation from videos[C]. Computer Vision — ECCV 2020. Cham: Springer, 2020: 764-780.

[24] Chen T, Fang C, Shen X, et al. Anatomy-aware 3D human pose estimation with bone-based pose decomposition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(1): 198-209.

作者简介

胡楠 (1987−), 男, 副教授, 博士, 主要研究方向为目标检测与目标跟踪, E-mail: hunan@sjzu.edu.cn;

张家豪 (1999−), 男, 硕士生, 主要研究方向为姿态估计与目标跟踪, E-mail: 1521563447@qq.com;

魏晓彤 (1992−), 女, 助理研究员, 博士, 主要研究方向为医疗大数据与虚拟诊疗, E-mail: xtwei@cmu.edu.cn;

朱宏博 (1986−), 男, 副教授, 博士, 主要研究方向为医疗大数据与人工智能、边缘智能, E-mail: hombochu@sina.com.