Computer Science and Application 计算机科学与应用, 2026, 16(2), 201-213
Published Online February 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.162052

Hans汉斯

基于近邻卷积Transformer的视频序列3D人体姿态估计方法

潘帅杰[1]，智　宇[1,2]，陈　昂[1,2*]

[1]温州大学计算机与人工智能学院，元宇宙与人工智能研究中心，浙江　温州
[2]温州大学元宇宙与人工智能研究院，浙江　温州

收稿日期：2026年1月1日；录用日期：2026年1月29日；发布日期：2026年2月9日

摘　要

近年来，基于Transformer的方法在单目三维人体姿态估计领域取得了显著进展，其强大的自注意力机制能够有效建模全局特征与长程依赖关系。然而，现有方法大多侧重于构建全局的时空依赖，其交互机制缺乏对局部时空结构(特别是相邻帧之间强相关性)的显式归纳偏置。这可能导致模型对近邻帧间紧密而具结构性的时序关联挖掘不足。为此，本文提出一种新颖的注意力架构——近邻卷积Transformer (NCFormer)，它通过近邻帧卷积与轴向多层感知机显式地建模近邻帧间的依赖关系。具体而言，NCFormer包含三个核心组件：(1) 用于捕获全局时空依赖的多头自注意力模块；(2) 近邻卷积模块，利用时间方向的卷积核提取近邻帧关系；(3) 轴向多层感知机，该模块旨在对时间和空间维度进行独立的特征变换，避免跨维度信息的无差别混合，使模型能够更专注地学习各维度特有的模式。在两个广泛使用的三维人体姿态估计基准数据集——Human3.6M和MPI-INF-3DHP上进行的实验表明，NCFormer在多种评估设定下均取得了具有高度竞争力的性能。

关键词

三维人体姿态估计，时空Transformer，卷积，轴向多层感知机

3D Human Pose Estimation Method in Video Sequences Based on Neighbor Convolution Transformer

Shuaijie Pan[1], Yu Zhi[1,2], Ang Chen[1,2*]

[1]Research Center for Metaverse and Artificial Intelligence, College of Computing Science and Artificial Intelligence, Wenzhou University, Wenzhou Zhejiang
[2]Institute of Metaverse and Artificial Intelligence, Wenzhou University, Wenzhou Zhejiang

[*]通讯作者。

文章引用: 潘帅杰, 智宇, 陈昂. 基于近邻卷积 Transformer 的视频序列 3D 人体姿态估计方法[J]. 计算机科学与应用, 2026, 16(2): 201-213. DOI: 10.12677/csa.2026.162052

Received: January 1, 2026; accepted: January 29, 2026; published: February 9, 2026

Abstract

In recent years, Transformer-based methods have achieved significant progress in the field of monocular 3D human pose estimation, owing to their powerful self-attention mechanism that effectively capture global representations and long-range dependencies. However, most existing approaches predominantly focus on constructing global spatiotemporal dependencies, and their interaction mechanisms lack explicit inductive bias toward local spatiotemporal structures, particularly the strong correlations between adjacent frames. This may lead to insufficient exploitation of the close and structured temporal relationships among neighboring frames. To address this, this paper proposes a novel attention architecture—the Neighbor Convolution Transformer (NCFormer)—which explicitly models dependencies between neighboring frames through neighbor-frame convolution and axial multi-layer perceptrons. Specifically, NCFormer consists of three core components: (1) A multi-head self-attention module for capturing global spatiotemporal dependencies; (2) A neighbor convolution module, which employs temporal convolution kernels to extract relationships among neighboring frames; and (3) An axial multi-layer perceptron, designed to perform independent feature transformations along the temporal and spatial dimensions, thereby avoiding undifferentiated mixing of cross-dimensional information and enabling the model to focus more on learning dimension-specific patterns. Experiments conducted on two widely used benchmark datasets for 3D human pose estimation—Human3.6M and MPI-INF-3DHP—demonstrate that NCFormer achieves highly competitive performance across various evaluation settings.

Keywords

3D Human Pose Estimation, Spatiotemporal Transformer, Convolution, Axial Multi-Layer Perceptron

1. 引言

单目三维人体姿态估计[1]是计算机视觉领域的一项基本任务，旨在从单幅图像或视频中准确检测人体的关键解剖点，以重建完整精确的三维人体姿态模型。这种能力对于包括人类行为分析、人机交互和虚拟/扩增实境在内的广泛应用至关重要[2]-[5]。由于其广泛的实用性，单目三维人体姿态估计一直是学术研究和产业发展的核心焦点。

在大规模三维姿态数据集(如 Human3.6M [6]和 MPI-INF-3DHP [7])和 GPU 技术进步的推动下，三维姿态估计的深度学习方法已经取得了重大进展。目前，主流方法[8]通常采用“提升策略”，首先使用先进的二维人体姿态检测算法[9] [10]从图像中提取二维关键点坐标。然后，使用专门设计的模型将这些坐标映射到三维空间，以重建人体三维姿态。然而，仅依赖二维关键点会丢弃固有的图像深度信息，导致深度模糊性和结构不确定性。

为应对上述挑战，研究者转向了端到端的解决方案，其方法演进主要体现在架构创新上：基于时序卷积网络(TCN)的方法[11] [12]更擅长建模局部时序模式，但在捕捉复杂长程依赖上仍面临感受野限制；基于图卷积网络(GCN)的方法[13] [14]虽能利用人体关节的拓扑结构进行空间推理，但其在时序维度上的

建模能力通常较为薄弱或依赖额外设计；近期兴起的基于 Transformer 的方法，凭借其全局自注意力机制，在统一建模长距离时空依赖上展现出显著优势，已成为主流[15] [16]。

然而，标准 Transformer 的自注意力机制虽然能够根据内容动态加权，但其全局交互模式缺乏对局部时空结构(特别是相邻帧之间强相关性)的显式归纳偏置与约束。这可能导致模型在捕捉运动连续性与姿态平滑性等关键时序模式时效率不足，而这些恰恰是保证姿态序列时域一致性的基础。

为此，本文提出近邻卷积 Transformer (NCFormer)，其架构如图 1 所示。其核心是在保留全局建模能力的基础上，创新性地引入轴向卷积来构建一个专注于局部邻域建模的并行机制。具体而言，模型通过三个协同的组件工作：一个多头自注意力模块捕获全局上下文；一个近邻帧卷积模块利用多尺度轴向卷积核聚焦并增强局部帧关系；一个轴向多层感知机对时空特征进行解耦与增强。该设计显式增强了对局部近邻帧的依赖关系建模，从而有效弥补了全局注意力在捕捉精细时序结构上的不足。

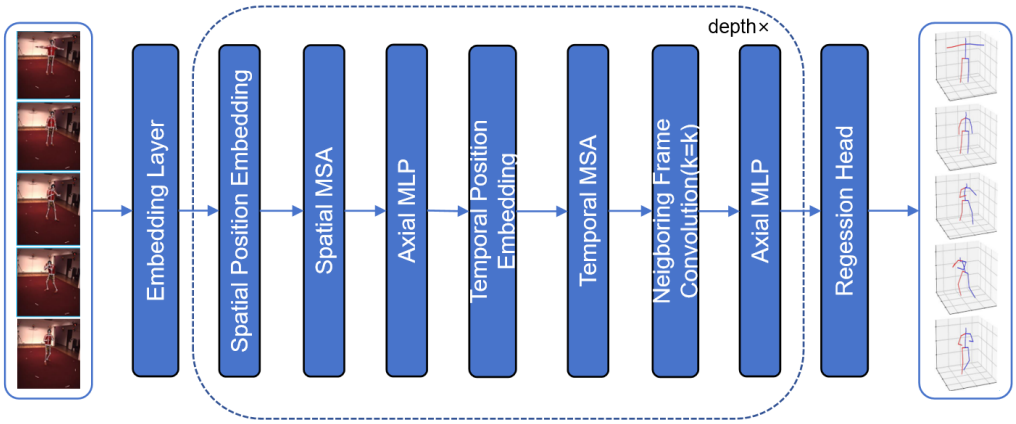

Figure 1. NCFormer model architecture diagram
图 1. NCFormer 模型架构图

2. 相关工作

目前的单目三维人体姿态估计方法主要由两种范式组成：从图像直接对联合坐标进行三维回归[17]，以及利用二维姿态检测器[11]-[16]进行二维到三维提升，然后进行三维重建。直接方法通过端到端学习保留空间信息，但通常需要大量的标记数据和计算资源。相比之下，二维到三维方法降低了建模复杂性，增强了稳健性，通常比直接回归达到更高的准确性。然而，这两种范式都面临着持续的挑战，包括二维到三维模糊性、遮挡、深度不确定性、外观变化和复杂的运动动力学。

为应对这些挑战，研究者先后探索了时序卷积网络(TCNs)与图卷积网络(GCNs)。例如，Pavllo 等人[12]利用扩张时序卷积捕捉长程依赖；Liu 等人[11]引入多尺度时序卷积并结合注意力机制；Cai 等人[13]则构建时空图，并通过层次化图卷积提取全局特征。然而，TCN 依赖固定卷积核，难以灵活建模多尺度动态模式；GCN 通常基于静态图结构，无法适应关节间随时间变化的依赖关系，限制了二者对复杂时空动态的表达能力。

因此，研究焦点逐渐转向基于 Transformer 的方法。其核心自注意力机制通过全局交互动态建模长程依赖，为统一表征时空信息提供了强大框架。相关研究例如，PoseFormerV2 [8]探索在频域中进行高效且鲁棒的特征表示；P-STMO [15]借助掩码机制进行时空预训练，增强模型对含噪声或遮挡的二维姿态输入的鲁棒性；MHFormer [16]则通过多假设 Transformer 建模姿态估计中固有的歧义性，以回归出置信度更高的三维姿态。这些方法共同体现了 Transformer 在处理长序列和挖掘全局上下文方面的卓越能力。

然而，标准 Transformer 的全局注意力缺乏对近邻帧之间强局部关联的显式归纳偏置，可能导致局部运动连续性被全局上下文稀释，影响序列平滑性与细节捕捉。因此，如何在保持全局建模能力的同时有效加强近邻帧依赖的局部建模，成为关键问题，亦是本文工作的出发点。

为了解决这个问题，我们提出了 NCFormer，其架构如图 1 所示，该方法在继承 Transformer 全局建模能力的基础上，引入卷积，构建了一个轻量高效的近邻帧卷积模块。该模块通过两次卷积操作，沿时间维度显式地构建近邻帧之间的依赖关系。此外，通过轴向多层感知机对时间、空间与通道维度进行解耦式建模，以抑制跨维度间的特征混杂，从而增强局部表征的判别性与融合效能。

3. 方法

本文提出的三维人体姿态估计方法整体框架如图 1 所示，图中箭头指示了前向传播的数据流。首先，利用二维关键点检测器从输入视频中提取二维姿态序列。该序列经过嵌入层(Embedding Layer)映射至高维空间，再编码坐标信息(Spatial Position Embedding)。随后，通过空间多头自注意力(Multi-Head Self-Attention)建模关节间的空间依赖，并由轴向多层感知机(Axial Multi-Layer Perceptron)对时间、空间与通道维度进行解耦的独立特征变换。

接下来，将时间信息嵌入特征中，通过时间多头自注意力建立长程时间依赖。之后，近邻帧卷积(Neighboring k Frame Convolution)显式聚合近邻 k 帧的局部时空信息。在此基础上，轴向多层感知机在时间维度执行独立的解耦变换时：由于特征已融合多帧上下文，该变换能够超越单帧空间结构的局限，在近邻帧信息支撑下更有效地建模帧间连续的局部运动模式，从而实现对时间动态的结构化细化表征。k 近邻卷积以及轴向多层感知机的详细结构分别在 3.3 和 3.4 节详细讨论。

3.1. 位置编码

标准 Transformer 中的自注意力机制本质上是置换等价的，不具备对输入序列顺序的感知能力。为使其适用于具有明确时空结构的人体姿态序列，必须显式地向模型中注入关于关节空间位置与帧时序位置的先验知识。

为此，我们参考 PoseFomer [18]的思路，分别引入可学习的位置编码。空间位置编码(Spatial Position Embedding)用于区分同一帧内不同的关节点。设人体骨架关节总数为 J，我们为每个关节 j 分配一个可学习的嵌入向量 $pe_j^{sp} \in \mathbb{R}^C$，嵌入向量集合 $PE^{sp} \in \mathbb{R}^{J\times C}$。将该嵌入向量嵌入输入，可使编码与不同关节绑定，从而帮助模型识别不同关节的语义。

时间位置编码(Temporal Position Embedding)则用于区分视频序列中的不同时刻。对于长度为 F 的序列，我们为每一帧 f 分配一个可学习的嵌入向量 $pe_t^{tmp} \in \mathbb{R}^C$，嵌入向量合集 $PE^{tmp} \in \mathbb{R}^{F\times J}$。将该嵌入向量嵌入输入，从而帮助模型感知动作的时间演进顺序与阶段变化。

3.2. 多头自注意力

多头自注意力机制是 Transformer 架构的核心思想，对于给定的输入特征 $X \in \mathbb{R}^{B\times F\times J\times C}$，其中 B 为批大小，F 为帧数，J 为关节数，C 为特征通道维度。多头注意力的计算首先通过一个线性投影层 $Linear_{C\to 3C}$ ($C \to 3C$ 表示通过线性投影最后维度长度发生的变化)生成查询 Q (Query)、键 K (Key)和值 V (Value)：

$$QKV = Linear_{C\to 3C}(X),\ QKC \in \mathbb{R}^{B\times F\times J\times C} \tag{1}$$

随后将特征重塑并分割为 H 个头(head)：

$$QKV \in \mathbb{R}^{B\times F\times J\times 3C} \to Q,K,V \in \mathbb{R}^{3\times B\times H\times F\times J\times C_h} \tag{2}$$

其中 H 为头数，$C_h = C/H$ 为每个头的特征维度。在空间注意力的情况下，注意力在每一帧内部的所有关节之间进行，用于学习关节点之间的空间依赖关系，对于每个头 h，注意力权重矩阵通过缩放点积计算：

$$Attention_{spatial}^{(h)} = Softmax\left(\frac{Q^{(h)}\left(K^{(h)}\right)^{T}}{\sqrt{C_h}}\right) \in \mathbb{R}^{J\times J} \tag{3}$$

在时间注意力的情况下，注意力针对每个独立的关节在所有帧之间进行，用于捕捉关节在时间维度上的运动轨迹。计算前需将 F 与 J 维度进行转置，使时间维度成为计算注意力的主体：

$$Attention_{temporal}^{(h)} = Softmax\left(\frac{Q^{(h)}\left(K^{(h)}\right)^{T}}{\sqrt{C_h}}\right) \in \mathbb{R}^{F\times F} \tag{4}$$

空间注意力权重矩阵，建立了不同关节之间的空间依赖关系，使模型能够学习人体关节之间的相关性，而时间注意力权重矩阵则建立不同时间步之间的全局依赖关系，使模型能够对人体姿态的运动轨迹进行连贯建模。最后加权求和得到多头注意力机制的输出：

$$MSA(X) = Attention^{(h)}V^{(h)} \tag{5}$$

3.3. k 近邻卷积

为克服标准 Transformer 在时间维度缺乏显式局部建模能力的局限，本文提出一个具有强归纳偏置的 k 近邻时间卷积模块，其结构如图 2 所示。该模块并非通用卷积，而是专为在时间轴上捕获局部帧间依赖而设计，其核心思想是向模型注入一个关键先验：在时间序列中，相邻帧之间的相关性远高于远距离帧，且此类局部依赖具有连续平滑的结构化特征。

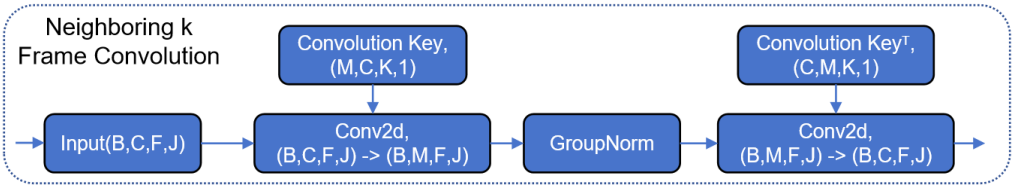

Figure 2. Workflow diagram of k neighboring convolution
图 2. k 近邻卷积工作流程图

对于给给定输入特征张量 $X \in \mathbb{R}^{B\times C\times F\times J}$，其中 B 为批大小，C 为通道数，F 为帧数(时间维度)，J 为关节数(空间维度)。在下面的表述中 $\otimes$ 代表卷积操作。该模块的计算过程可有以下公示链完整刻画：

$$Z^{(1)} = X \otimes K,\ K \in \mathbb{R}^{M\times C\times k\times 1} \tag{6}$$

$$Z^{(2)} = GN\left(Z^{(1)}\right) \tag{7}$$

$$Z^{(3)} = X \otimes K^{\mathrm{T}},\ K^{\mathrm{T}} \in \mathbb{R}^{C\times M\times k\times 1} \tag{8}$$

3.3.1. 时间局部性偏置：时间卷积核设计

我们通过一个一维时间卷积核，显式地注入了时间局部性的强归纳偏置。在实现上，我们使用一个二维卷积核并将其空间宽度约束为 1，从而在数学上等价于对每个关节的时序独立进行一维卷积。设输入特征为 $X \in \mathbb{R}^{B\times C\times F\times J}$，其中 F 为帧数，J 为关节数。我们定义的卷积核参数为 $K \in \mathbb{R}^{M\times C\times k\times 1}$，该设计的关键在于其形状 $(k,1)$ 所编码的硬性约束：k(时间维度)：决定了局部时间邻域的宽度，即模型能够观察当

前帧前后共 k 帧的信息；1 (空间维度)：强制卷积核在空间维度的感受野为 1，这意味着卷积操作不会在同一时间点上跨越不同的关节进行混合。

因此，该卷积操作在物理上等价于对 J 个关节中的每一个，独立地在其时间序列 $\mathbb{R}^{B\times C\times F}$ 上应用一个长度为 k 的一维卷积。这完美契合了人体运动在短时窗口内连续、平滑的先验知识。

3.3.2. 时间局部性偏置：轴向卷积核设计

该模块采用参数共享的两次卷积结构，注入了对称重建的归纳偏置：首先，通过前向卷积将输入特征从 C 维投影至 M 维，并在局部时间窗口内进行信息融合；随后，应用组归一化对通道分组处理，以提升模型对输入噪声的鲁棒性、稳定中间层分布，从而加速收敛；最后，利用同一卷积核的转置进行对称重建，将特征从 M 维重建为 C 维。该设计强制前向投影与重建过程共享同一组基，促使模型学习到一组紧凑且具解释性的时间局部特征基。

3.3.3. 在 NCFormer 中的角色：与全局注意力的偏置协同

在 NCFormer 架构中，此模块被实例化为近邻帧卷积。它提供了一种与全局多头自注意力互补且正交的归纳偏置，专门负责捕捉帧间运动的微观平滑性与连续性。这种设计使得 NCFormer 能够同时利用两种强大的模型范式：既拥有 Transformer 全局建模的灵活性，又具备卷积网络局部建模的效率与结构性。k 近邻时间卷积模块确保运动轨迹的局部细节得以精确维护，从而生成更自然、抖动更少的三维姿态序列。

3.4. 轴向多层感知机

为在时空维度上实现解耦的特征变换，同时避免不同维度间的信息干扰，我们设计了一个轴向多层感知机模块，其结构如图 3 所示。该模块的核心设计思想是：对输入特征的时间、空间和通道维度进行独立的线性变换，从而在保持维度特异性的同时增强特征表示，同时提供了更强的结构化归纳偏置。

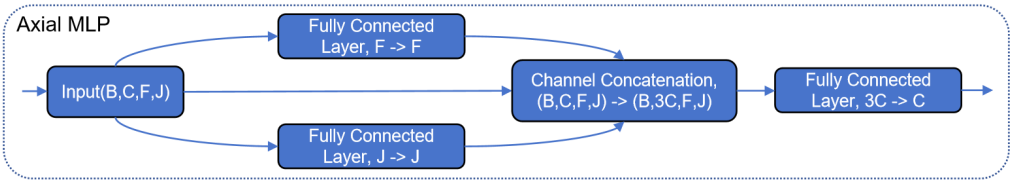

Figure 3. Workflow diagram of axial MLP
图 3. 轴向多层感知机工作流程图

3.4.1. 维度分离变换机制

维度分离变换机制是轴向 MLP 与传统 MLP 的核心区别。该机制通过引入强结构归纳偏置，改变了特征交互的基本范式：传统 MLP 作为通用函数逼近器，其交互是无结构且无差别的；而轴向 MLP 通过对时间轴与空间轴进行分离处理，显式注入了时空局部性与维度分离性的先验知识。这使得模型能够分别学习不同维度的特定模式(如时间上的运动平滑性、空间上的关节拓扑关系)，避免无关维度信息的干扰，从而显著降低学习难度。

因此，轴向 MLP 并非简单的多层感知机变体，而是一种专为高维结构化数据(如图像、视频、姿态序列)设计的特征变换层。它将 Transformer 中“注意力应具备结构性”的思想引入前馈网络，其核心优势在于通过维度解耦，在保持全局感受野的同时，实现更高效、更贴合视觉数据本质的表示学习。

给定输入特征 $X\in\mathbb{R}^{B\times C\times F\times J}$，其中 B 为批大小，C 为通道数，F 为帧数，J 为关节数。轴向 MLP 分别沿两个维度执行独立的线性变换，在下面的公式中 $Linear$ 表示线性变换，$T(a,b)$ 表示交换 a 和 b 维度：

1) 时间维度变换：

$$X_F = Linear\left(X^{T(2,3)}\right)^{T(2,3)} \tag{9}$$

2) 空间维度变换：

$$X_J = Linear\left(X\right) \tag{10}$$

3.4.2. 特征融合与重构

将空间维度变换 X_F、时间维度变换 X_J、初始输入特征 X 三个分支的输出沿通道维度拼接后，通过融合层进行降维重构，在下面的公式中，$Concat$ 表示通道维度拼接，$Linear_{a\to c}$ 表示线性变换，将最后维度的数据由长度 a 变化至长度 b：

$$Y = Linear_{3C\to C}\left(Concat\left(X_F, X_J, X\right)\right) \tag{11}$$

轴向 MLP 的设计体现了以下关键优势：轴向 MLP 的设计具有以下关键优势：首先，其维度解耦偏置通过对时间与空间维度的独立处理，避免了跨维度信息的无差别混合，从而使模型能够更专注地学习各维度特有的模式；其次，三个分支的明确分工赋予了模型良好的结构可解释性——时间分支专注于时序动态，空间分支关注关节间关系，恒等分支则保留原始信息；最后，该模块与注意力机制具备天然的互补性：当与近邻卷积模块协同工作时，轴向 MLP 的时间分支能够进一步细化由卷积模块提取的局部时间特征，其空间分支则增强空间结构的表示，二者形成“注意力建立依赖，MLP 精炼特征”的高效协作模式。

3.5. 损失函数

NCFormer 的损失函数 L 定义为：

$$L = L_w + \lambda_t L_t + \lambda_m L_m \tag{12}$$

其中 L_w 为 Zhang 等人[19]提出的一种加权平均关节点位置误差，L_t 为 Hossain 等人[20]提出的一种带有时间平滑性约束的损失函数，L_m 为 MPJVE 损失，λ_t 和 λ_m 为权重系数，用于调控各损失函数对结果的影响。L_w 的数学公式可以表示为：

$$L_w = \frac{1}{J}\sum_{j=1}^{J}\left(W_j \times \frac{1}{F}\sum_{f=1}^{F}\left\|p_{j,f} - gt_{j,f}\right\|_2^2\right) \tag{13}$$

其中 F 为输入帧序列长度，J 为关节数量，W_j 是分配给第 j 个关节的权重，$W \in \mathbb{R}^J$ 是一个权重向量，$p_{j,f}$ 和 $gt_{j,f}$ 分别是预测和真实的第 f 帧第 j 个关节的三维坐标。其核心思想在于，在计算所有关节的平均位置误差时，为不同关节分配不同的权重，以反映它们在人体姿态估计中的重要程度。比如骨盆、脊柱等根关节的位置误差，会直接导致整个骨架的全局偏移。

L_m 和 L_t 的数学公式可以表达为：

$$L_m = \frac{1}{J}\sum_{i=1}^{J}\left(\frac{1}{F}\sum_{f=1}^{F}\left\|p_{j,f} - gt_{j,f}\right\|_2\right) \tag{14}$$

$$L_t = \alpha \times L_m^2 + \beta \times \left\|\nabla_t\right\|_2^2 \tag{15}$$

$$\left\|\nabla_t\right\|_2^2 = \frac{1}{J\left(F-1\right)}\sum_{j=1}^{J}\sum_{f=2}^{F}\left(\eta\left\|p_{j,f}^{TH} - p_{j,f-1}^{TH}\right\|_2^2 + \rho\left\|p_{j,f}^{LL} - p_{j,f-1}^{LL}\right\|_2^2 + \tau\left\|p_{j,f}^{LA} - p_{j,f-1}^{LA}\right\|_2^2\right) \tag{16}$$

其中 $p_{j,f}^{TH}$ 、 $p_{j,f}^{LL}$ 和 $p_{j,f}^{LA}$ 分别表示预测的躯干头部、肢体腿部和肢体手臂集合的关节的 3D 位置，α 和 β 为权重系数，α 用于调控预测与真实坐标损失的权重，而 β 则用于调控预测前后帧坐标差异损失的权重，确保预测序列在时间上平滑。η 、 ρ 和 τ 则用于调控躯干头部、肢体腿部和肢体手臂集合的权重。

4. 实验

4.1. 数据集以及评估协议

我们在两个具有挑战性的三维人体姿态估计基准数据集 Human3.6M [6]和 MPI-INF-3DHP [7]上对所提方法进行了全面评估。

Human3.6M 是该领域最具代表性和广泛使用的室内数据集之一。为与主流研究保持一致[15] [16]，我们使用受试者 S1、S5、S6、S7 和 S8 进行训练，并在受试者 S9 和 S11 上进行评估。性能评估采用广泛使用的指标，包括平均关节点位置误差(MPJPE)、重建误差(P-MPJPE)以及平均关节点速度误差(MPJVE)。

MPI-INF-3DHP 是另一个广泛使用的高质量基准数据集，其特点在于同时涵盖室内和室外场景。为全面评估模型在该数据集上的性能，我们遵循先前工作中建立的标准评估协议[12]，采用多种评价指标——包括 MPJPE、正确关节点百分比(PCK)和曲线下面积(AUC)。

4.2. 补充细节

我们基于 PyTorch 框架实现了所提出的 NCFormer 模型，所有实验均在一张 GeForce RTX 4090D GPU 上完成。为评估方法的有效性，我们使用由预训练的二维姿态检测器生成的二维关键点[9]或真值二维关键点作为输入。模型使用 Adam 优化器[21]进行训练，批大小设置为 1024，丢弃率为 0.1，并采用 GELU 作为激活函数。初始学习率设为 0.00004，衰减因子为 0.99。模型共训练了 200 个 epochs。

4.3. Human3.6M 数据集评估

Table 1. Performance indicator comparison on Human3.6M Protocol 1 based on 2D key points from the CPN
表 1. 基于 CPN 2D 关键点在 Human3.6M 协议 1 上的性能指标对比

模型	ECCV20 SRNET [22]	ICCV21 PoseFormer [18]	ECCV22 STMO [15]	CVPR22 MHFormer [16]	ICCV23 GLA-GCN [23]	IVC24 STAFormer [24]	Ours
方向指示	46.6	41.5	38.9	39.2	41.3	38.0	39.5
交谈讨论	47.1	44.8	42.7	43.1	44.3	41.8	41.7
进食	43.9	39.8	40.4	40.1	40.8	39.5	36.6
打招呼	41.6	42.5	41.1	40.9	41.8	39.8	40.0
通电话	45.8	46.5	45.6	44.9	45.9	44.8	42.8
拍照	49.6	51.6	49.7	51.2	54.1	48.9	48.3
摆姿势	46.5	42.1	40.9	40.6	42.1	39.8	41.0
购物	40.0	42.0	39.9	41.3	41.5	39.0	39.6
坐下	53.4	53.3	55.5	53.5	57.8	53.7	52.2
坐下的过程	61.1	60.7	59.4	60.3	62.9	57.2	57.9
吸烟	46.1	45.5	44.9	43.7	45.0	43.4	42.2
等待	42.6	43.3	42.2	41.1	42.8	41.3	40.7
遛狗	43.1	46.1	42.7	43.8	45.9	41.9	40.9
行走	31.5	31.8	29.4	29.8	29.4	28.6	27.5
并肩行走	32.6	32.2	29.4	30.6	29.9	28.7	28.3
平均	44.8	44.3	42.8	43.0	44.4	41.7	41.3

我们在 Human3.6M 数据集上对提出的 NCFormer 和多种先进的方法进行了全面比较。如表 1 和表 2 所示，以 CPN 检测到的二维关键点作为输入，我们分别使用 MPJPE 和 P-MPJPE 指标评估姿态估计精度。表 3 则进一步通过 MPJVE (Mean Per Joint Velocity Error，平均关节速度误差)衡量预测姿态的时间平滑性与一致性。

Table 2. Performance indicator comparison on Human3.6M Protocol 2 based on 2D key points from the CPN
表 2. 基于 CPN 2D 关键点在 Human3.6M 协议 2 上的性能指标对比

模型	ICCV21 PoseFormer [18]	ECCV22 P-STMO [15]	PR23 MHFormer++ [25]	ICCV23 GLA-GCN [23]	IVC24 STAFormer [24]	Ours
方向指示	32.5	31.3	31.6	32.4	31.5	31.5
交谈讨论	34.8	35.2	34.8	35.3	34.7	41.7
进食	32.6	32.9	32.2	32.6	32.5	36.6
打招呼	34.6	33.9	33.2	34.2	33.5	32.1
通电话	35.3	35.4	34.7	35.0	34.9	33.3
拍照	39.5	39.3	39.7	42.1	39.5	38.1
摆姿势	32.1	32.5	33.0	32.1	31.8	31.5
购物	32.0	31.5	31.0	31.9	31.1	30.1
坐下	32.0	44.6	43.5	45.5	43.8	42.3
坐下的过程	48.5	48.2	49.6	49.5	47.5	46.5
吸烟	34.8	36.3	36.1	36.1	35.8	34.3
等待	32.4	32.9	32.4	32.4	32.5	31.5
遛狗	35.3	34.4	33.8	35.6	33.9	32.5
行走	24.5	23.8	23.9	23.5	23.4	22.0
并肩行走	26.0	23.9	24.7	24.7	23.9	23.2
平均	34.6	34.4	34.2	34.8	34.0	32.9

为保障评估的公平性与全面性，我们遵循两种广泛采用的评估协议。在协议 1 (基于 CPN 二维关键点)下，NCFormer 取得了最优的 MPJPE 结果，平均误差为 41.3 mm。在协议 2 (使用相同 CPN 关键点)中，我们的方法同样获得最佳性能，平均 P-MPJPE 为 32.9 mm。在两种协议下，NCFormer 在绝大多数动作类别上均取得领先或接近最优的表现，这验证了其在全局结构与局部细节建模之间的有效平衡，以及良好的泛化能力。

Table 3. MPJVE performance indicator comparison on Human3.6M based on 2D key points from the CPN
表 3. 基于 CPN 2D 关键点在 Human3.6M 上的 MPJVE 性能指标对比

模型	ICCV21 PoseFormer [18]	CVPR22 MHFormer [16]	CVPR22 MixSTE [19]	PR23 MHFormer++ [25]	Ours
平均	2.5	2.4	2.3	2.3	2.1

在时序一致性方面，NCFormer 实现了 2.1 mm 的 MPJVE，显著优其他方法。这一结果表明，我们的方法能够生成具有优越平滑性和时间一致性的运动序列，从动态建模角度体现了其有效捕捉近邻帧依赖的优势。综上，实验说明 NCFormer 能够基于准确的二维观测恢复出高精度的三维姿态，进一步印证了其模型表达能力与泛化性能。

4.4. MPI-INF-3DHP 数据集评估

表 4 总结了 MPI-INF-3DHP 数据集的实验结果。我们的方法取得 PCK 98.2、AUC 75.7，MPJPE 32.8

mm 的优异性能。这些结果证明了我们方法在空间位置预测方面的高准确性，以及在复杂环境和多样化动作类别下强大的泛化能力。值得注意的是，MPI-INF-3DHP 数据集涵盖了广泛的室内和室外场景。尽管存在这种变异性，NCFormer 始终产生稳定和高质量的预测，突出了其在面对挑战性现实世界条件时的鲁棒性和适应性。上述结果进一步证实了所提方法的有效性与实际应用潜力。

Table 4. Performance comparison on MPI-INF-3DHP database
表 4. MPI-INF-3DHP 数据集性能对比

模型	ICCV21 PoseFormer [18]	CVPR22 MHFormer [16]	CVPR22 MixSTE [19]	PR23 MHFormer++ [25]	Ours
PCK↑	88.6	93.8	94.4	94.8	98.2
AUC↑	56.4	63.3	66.5	65.8	75.7
MPJPE↓ (mm)	77.1	58.0	54.9	54.0	32.8

4.5. 消融实验

4.5.1. 组件分析

为了评估我们提出模型中每个组件的有效性，我们在 Human3.6M 数据集上进行了系统的消融实验，并以 MPJPE 作为评价指标。实验以仅包含位置编码的模型作为基线，逐步添加近邻卷积模块和轴向多层感知机模块，结果如表 5 所示。

实验表明，当仅使用位置编码作为基线时，MPJPE 为 47.2 mm。在加入近邻卷积模块后，MPJPE 下降至 42.3 mm，性能显著提升，说明该模块能有效捕捉近邻帧间的局部依赖关系。进一步引入轴向多层感知机模块后，MPJPE 进一步降低至 41.3 mm，表明该模块通过对时空维度进行解耦建模，进一步增强了特征的判别力与泛化能力。最终，完整模型(包含所有模块)取得了最优性能。

Table 5. Component ablation experiment
表 5. 组件消融实验

模型	位置编码	k 近邻卷积	轴向多层感知机	MPJPE↓ (mm)	P-MPJPE↓ (mm)	MPJVE↓ (mm/s)
	√	×	×	42.5	34.0	2.53
	√	√	×	42.3	33.5	2.25
	√	×	√	41.6	33.3	2.16
Ours	√	√	√	41.3	32.9	2.10

上述结果验证了近邻卷积模块与轴向多层感知机模块各自的有效性，同时体现了二者在建模时空依赖时的协同作用，共同提升了三维姿态估计的精度与鲁棒性。

4.5.2. 超参数分析

我们在 Human3.6M 数据集上对所提方法进行了系统的超参数消融实验，实验结果总结于表 6。实验依次考察了模型深度(5、6、7)、嵌入维度(256、512、640)、卷积核大小(5、7、9)以及输入帧长度(81、243、351)对性能的影响。结果表明，深度为 6 时模型表现最优，优于深度 5 的较浅结构与深度 7 的较深结构；嵌入维度设为 512 时性能最高，优于 256 与 640 的设定；卷积核大小为 7 时效果最佳，优于 5 和 9；而输入帧数为 243 时预测误差最低，优于 81 和 351 帧的配置。

此外，我们分析了各超参数对模型复杂度的影响。其中，嵌入维度与模型深度对参数量和单帧推理计算

量(FLOPs)的影响最为显著：嵌入维度因自注意力机制的计算复杂度而带来平方级增长，深度则线性增加模型复杂度。相比之下，卷积核大小(因采用轴向卷积而结构稀疏)与输入帧数对整体复杂度的影响相对较小。

基于上述结果，我们确定了模型深度 6、嵌入维度 512、卷积核大小 7、输入帧数 243 为最优超参数组合。该配置在所有测试中取得了最低的 MPJPE (41.3 mm)，表明其能在模型复杂度与表征能力之间实现有效平衡，从而提升姿态估计的准确性。

Table 6. Hyperparameter ablation experiment
表 6. 超参数消融实验

深度	嵌入维度	k	Frame	Parameters↓ (M)	FLOPs↓ (M)	MPJPE↓ (mm)
5	512	7	243	24.75	525.6	41.9
6	512	7	243	29.67	630.7	41.3
7	512	7	243	34.59	735.8	41.5
6	256	7	243	7.74	168.3	42.4
6	640	7	243	46.1	972.3	41.5
6	512	5	243	28.1	577.3	41.8
6	512	9	243	31.2	684.2	41.4
6	512	7	81	29.3	605.3	42.1
6	512	7	351	30.1	647.7	41.5

4.6. 定性结果

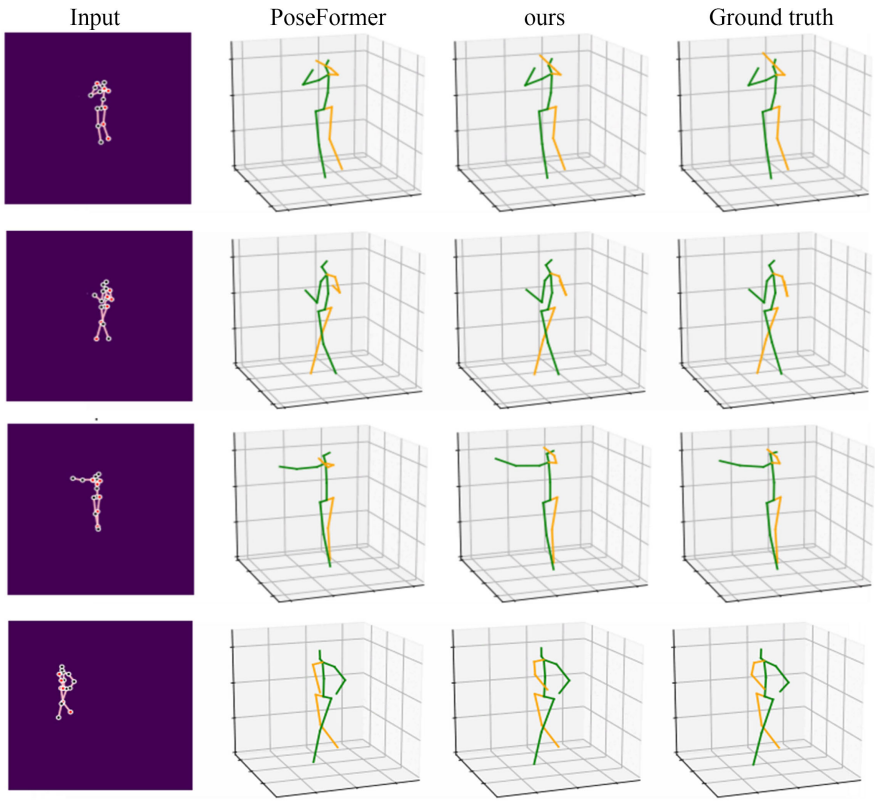

Figure 4. 3D pose reconstruction comparison (NCFormer vs. PoseFormer) on Human3.6M database
图 4. NCFormer 与 PoseFormer 在 Human3.6M 数据集上的三维姿态重建对比

我们进一步通过定性可视化来验证所提 NCFormer 模型在三维人体姿态估计上的有效性，从 Human3.6M 测试集中随机选取若干样本，将 NCFormer 的三维姿态重建结果与图 4 中有代表性的 Transformer 方法 PoseFormer 结果进行对比。可视化结果清晰表明，在所有动作类别中，NCFormer 预测的三维关节位置均更接近真实姿态。相较于其他模型，本方法在肢体长度一致性、关节连接平滑度以及整体结构合理性方面均表现更优。上述可视化结果表明，NCFormer 通过有效建模近邻帧依赖关系，显著提升了姿态估计的准确性与鲁棒性。

5. 结论

本文提出了一种新颖的近邻卷积 Transformer (NCFormer)，用于单目视频中的三维人体姿态估计。NCFormer 的核心创新在于其近邻帧依赖捕捉机制：一方面，通过轴向卷积显式地聚合局部近邻帧的时空特征，以捕捉运动连续性与短程依赖；另一方面，通过轴向多层感知机对时间与空间维度进行解耦建模，避免跨维度信息干扰，从而增强特征的结构化表示能力。

实验结果表明，NCFormer 在姿态估计精度与运动时序平滑性之间取得了卓越的平衡。在 Human3.6M 和 MPI-INF-3DHP 两个权威基准上，NCFormer 在 MPJPE、P-MPJPE 和 MPJVE 等多个关键指标上达到当前先进方法的性能。定性与定量分析共同验证，我们的方法能够生成更准确、更自然且时间一致性更强的三维姿态序列。

致　谢

我们感谢相关领域研究者公开的代码资源，特别是 PoseFormer [18]的工作为我们提供了重要参考。本研究使用的 Human3.6M [6]和 MPI-INF-3DHP [7]数据集为实验评估提供了基础，在此一并表示感谢。

基金项目

本课题受到“温州大学元宇宙与人工智能研究院”的“重大课题及项目产业化专项资金”(编号：2023103)的资助。

参考文献

[1] Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., *et al.* (2023) Deep Learning-Based Human Pose Estimation: A Survey. *ACM Computing Surveys*, **56**, 1-37. https://doi.org/10.1145/3603618

[2] Li, C., Huang, Q., Mao, Y., Li, X. and Wu, J. (2024) Multi-Granular Spatial-Temporal Synchronous Graph Convolutional Network for Robust Action Recognition. *Expert Systems with Applications*, **257**, Article ID: 124980. https://doi.org/10.1016/j.eswa.2024.124980

[3] Liu, M., Liu, H. and Chen, C. (2017) Enhanced Skeleton Visualization for View Invariant Human Action Recognition. *Pattern Recognition*, **68**, 346-362. https://doi.org/10.1016/j.patcog.2017.02.030

[4] Gong, J., Fan, Z., Ke, Q., Rahmani, H. and Liu, J. (2022) Meta Agent Teaming Active Learning for Pose Estimation. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 11069-11079. https://doi.org/10.1109/cvpr52688.2022.01080

[5] Yoon, J.S., Liu, L., Golyanik, V., Sarkar, K., Park, H.S. and Theobalt, C. (2021) Pose-Guided Human Animation from a Single Image in the Wild. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 15034-15043. https://doi.org/10.1109/cvpr46437.2021.01479

[6] Ionescu, C., Papava, D., Olaru, V. and Sminchisescu, C. (2014) Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **36**, 1325-1339. https://doi.org/10.1109/tpami.2013.248

[7] Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., *et al.* (2017) Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision. 2017 *International Conference on 3D Vision* (*3DV*), Qingdao, 10-12 October 2017, 506-516. https://doi.org/10.1109/3dv.2017.00064

[8] Zhao, Q., Zheng, C., Liu, M., Wang, P. and Chen, C. (2023) PoseFormerV2: Exploring Frequency Domain for Efficient

and Robust 3D Human Pose Estimation. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 8877-8886. https://doi.org/10.1109/cvpr52729.2023.00857

[9] Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G. and Sun, J. (2018) Cascaded Pyramid Network for Multi-Person Pose Estimation. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 7103-7112. https://doi.org/10.1109/cvpr.2018.00742

[10] Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., *et al.* (2021) Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**, 3349-3364. https://doi.org/10.1109/tpami.2020.2983686

[11] Liu, R., Shen, J., Wang, H., Chen, C., Cheung, S. and Asari, V. (2020) Attention Mechanism Exploits Temporal Contexts: Real-Time 3D Human Pose Reconstruction. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 5063-5072. https://doi.org/10.1109/cvpr42600.2020.00511

[12] Pavllo, D., Feichtenhofer, C., Grangier, D. and Auli, M. (2019) 3D Human Pose Estimation in Video with Temporal Convolutions and Semi-Supervised Training. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 7745-7754. https://doi.org/10.1109/cvpr.2019.00794

[13] Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T., Yuan, J., *et al.* (2019) Exploiting Spatial-Temporal Relationships for 3D Pose Estimation via Graph Convolutional Networks. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 2272-2281. https://doi.org/10.1109/iccv.2019.00236

[14] Jia, R., Yang, H., Zhao, L., Wu, X. and Zhang, Y. (2023) MPA-GNet: Multi-Scale Parallel Adaptive Graph Network for 3D Human Pose Estimation. *The Visual Computer*, **40**, 5883-5899. https://doi.org/10.1007/s00371-023-03142-z

[15] Shan, W., Liu, Z., Zhang, X., Wang, S., Ma, S. and Gao, W. (2022) P-STMO: Pre-Trained Spatial Temporal Many-To-One Model for 3D Human Pose Estimation. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M. and Hassner, T., Eds., *Computer Vision—ECCV* 2022, Springer, 461-478. https://doi.org/10.1007/978-3-031-20065-6_27

[16] Li, W., Liu, H., Tang, H., Wang, P. and Van Gool, L. (2022) MHFormer: Multi-Hypothesis Transformer for 3D Human Pose Estimation. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 13137-13146. https://doi.org/10.1109/cvpr52688.2022.01280

[17] Newell, A., Yang, K. and Deng, J. (2016) Stacked Hourglass Networks for Human Pose Estimation. In: Leibe, B., Matas, J., Sebe, N. and Welling, M., Eds., *Computer Vision—ECCV* 2016, Springer, 483-499. https://doi.org/10.1007/978-3-319-46484-8_29

[18] Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C. and Ding, Z. (2021) 3D Human Pose Estimation with Spatial and Temporal Transformers. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 11636-11645. https://doi.org/10.1109/iccv48922.2021.01145

[19] Zhang, J., Tu, Z., Yang, J., Chen, Y. and Yuan, J. (2022) MixSTE: Seq2seq Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in Video. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 13222-13232. https://doi.org/10.1109/cvpr52688.2022.01288

[20] Hossain, M.R.I. and Little, J.J. (2018) Exploiting Temporal Information for 3D Human Pose Estimation. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., Eds., *Computer Vision—ECCV* 2018, Springer, 69-86. https://doi.org/10.1007/978-3-030-01249-6_5

[21] Kingma, D.P. and Ba, J. (2015) Adam: A Method for Stochastic Optimization. arXiv: 1412.6980.

[22] Zeng, A., Sun, X., Huang, F., Liu, M., Xu, Q. and Lin, S. (2020) SRNet: Improving Generalization in 3D Human Pose Estimation with a Split-And-Recombine Approach. In: Vedaldi, A., Bischof, H., Brox, T. and Frahm, J.M., Eds., *Computer Vision—ECCV* 2020, Springer, 507-523. https://doi.org/10.1007/978-3-030-58568-6_30

[23] Yu, B.X.B., Zhang, Z., Liu, Y., Zhong, S., Liu, Y. and Chen, C.W. (2023) GLA-GCN: Global-Local Adaptive Graph Convolutional Network for 3D Human Pose Estimation from Monocular Video. 2023 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Paris, 1-6 October 2023, 8784-8795. https://doi.org/10.1109/iccv51070.2023.00810

[24] Hao, F., Zhong, F., Yu, H., Hu, J. and Yang, Y. (2024) STAFFormer: Spatio-Temporal Adaptive Fusion Transformer for Efficient 3D Human Pose Estimation. *Image and Vision Computing*, **149**, Article ID: 105142. https://doi.org/10.1016/j.imavis.2024.105142

[25] Li, W., Liu, H., Tang, H. and Wang, P. (2023) Multi-Hypothesis Representation Learning for Transformer-Based 3D Human Pose Estimation. *Pattern Recognition*, **141**, Article ID: 109631. https://doi.org/10.1016/j.patcog.2023.109631