杨金金, 刘淑源, 朱家兵. 基于动态多粒度图卷积的三维人体姿态估计[J]. 智能计算机与应用,2026,16(8):137-144.
DOI:10.20169/j.issn.2095-2163.26011601

基于动态多粒度图卷积的三维人体姿态估计

杨金金[1], 刘淑源[1], 朱家兵[2,3]
(1 安徽理工大学 计算机科学与工程学院, 安徽 淮南 232001; 2 安徽省北斗卫星导航技术重点实验室, 安徽 淮南 232038;
3 淮南师范学院 电子工程学院, 安徽 淮南 232038)

摘 要: 针对现有基于 Transformer 的单目三维人体姿态估计网络对局部运动特征建模不足,以及传统图卷积网络因静态参数难以适应人体动作幅度动态变化的问题,本文提出了一种基于动态多粒度图卷积的三维姿态估计网络。首先该网络通过引入动态多粒度图卷积模块,不仅能对人体关节按照自由度分区建模,还能利用动态调制机制根据输入数据自适应生成调制参数,提升对局部关节运动表征的感知能力。在此基础上,本文引入了动态多粒度图 Transformer 模块,将上述图卷积与多头自注意力机制结合,实现了对关节局部特征与全局上下文信息的协同建模。最后引入双流时频特征融合 Transformer 模块,通过交叉注意力深度挖掘时域瞬时信息与频域全局趋势之间的互补性,弥补单一时域特征表达不充分的缺限。实验结果显示,本文方法在 Human3.6M 以及 MPI-INF-3DHP 数据集上的指标优于 PoseFormerV2 等主流方法,证明了本文所提方法的有效性。

关键词: Transformer;动态多粒度图卷积;动态调制机制;双流时频特征融合 Transformer;交叉注意力
中图分类号: TP391.41 **文献标志码**: A **文章编号**: 2095-2163(2026)08-0137-08

Dynamic multi-granularity graph convolution for 3D human pose estimation

YANG Jinjin[1], LIU Shuyuan[1], ZHU Jiabing[2,3]
(1 School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan 232001, Anhui, China;
2 Anhui Province Key Laboratory of Beidou Satellite Navigation Technology, Huainan 232038, Anhui, China;
3 School of Electronic Engineering, Huainan Normal University, Huainan 232038, Anhui, China)

Abstract: To address the limitations of existing Transformer-based monocular 3D human pose estimation networks in modeling local motion features, and the difficulty that traditional Graph Convolutional Networks (GCNs) have in adapting to dynamic human motion with static parameters, this paper proposes a 3D pose estimation network based on dynamic multi-granularity graph convolution. First, by introducing a dynamic multi-granularity graph convolution module, the network not only models human joints based on degree-of-freedom partitioning but also utilizes a dynamic modulation mechanism to adaptively generate modulation parameters based on input data. This enhances the perception capability for local joint motion representations. Building upon this, a dynamic multi-granularity graph transformer module is introduced. This module integrates the aforementioned graph convolution with multi-head self-attention mechanisms, achieving synergistic modeling of local joint features and global context information. Finally, a dual-stream time-frequency feature fusion transformer module uses cross-attention to explore the complementarity between instantaneous time-domain information and global frequency-domain trends, overcoming single time-domain feature limitations. Experimental results on Human 3.6M and MPI-INF-3DHP datasets show the method outperforms mainstream models like PoseFormerV2, verifying its effectiveness.

Key words: Transformer; dynamic multi-granularity graph convolution; dynamic modulation mechanism; dual-stream time-frequency feature fusion transformer; cross-attention

0 引 言

三维人体姿态估计是计算机视觉领域中的一项重要技术,其目标是从图像或视频序列中重建人体关节的三维空间坐标,从而获得更完整和准确的人体姿态信息。由于三维人体姿态包含人体运动几何

基金项目: 国家自然科学基金(62172183);安徽省科技重大专项(202003a05020031);安徽省重点研发计划(202004a05020023);安徽省科技创新平台重大科技项目(S202305a12020025);安徽省电子信息工程特色专业建设项目支持。
作者简介: 杨金金(2002—),女,硕士研究生。主要研究方向:计算机视觉; 刘淑源(2002—),男,硕士研究生。主要研究方向:计算机视觉。
通信作者: 朱家兵(1970—),男,博士,教授。主要研究方向:雷达信号处理。Email:zjb3617@163.com。
收稿日期: 2026-01-16。

及运动学等信息,故在智慧医疗、智能制造、人机交互[1]等方面应用广泛。

随着深度学习的发展,三维人体姿态估计也从传统方法转向了基于深度学习的方法。目前基于深度学习的三维人体姿态估计主要分为端到端方法[2]和两阶段方法[3]。端到端方法直接从二维图像或视频中推断出三维人体姿态,而不需要中间估计二维姿态表示[4]。相比于端到端方法,两阶段方法则采用了分而治之的策略,首先利用现有成熟的二维姿态估计模型检测二维关节坐标,然后将其提升至三维空间中。在多数场景下,由于现有二维姿态检测器性能出色,所以两阶段方法在人体姿态估计精度上普遍优于端到端方法。

近年来,基于自注意力机制的 Transformer 网络[5]受到了大量关注,无论是在自然语言处理领域[6]还是在计算机视觉领域[7]中,Transformer 都展现了其重要性。在三维人体姿态估计领域中,同样也有大量的工作利用了 Transformer 网络。Zheng 等[8]首次将一个完全基于 Transformer 的网络应用到三维人体姿态估计中,该模型以二维姿态序列为输入,通过时空 Transformer 模块全面建模每帧关节的空间关系以及帧与帧之间的时间相关性。Zhao 等[9]提出了 PoseFormerV2 模型,该模型在 PoseFormerV1 的基础上引入了频域表示,并将频域与时空注意力机制融合,有效地融合了时域和频域的特征,提高了三维关节点检测的鲁棒性。尽管基于 Transformer 的方法在长距离建模中表现出色,但 Transformer 架构通常会忽略人体本身具有的拓扑先验知识,在面对遮挡问题时无法利用人体的先验知识估计准确的关节位置。因此一些研究尝试将图卷积网络(GCN)引入三维人体姿态估计任务,并取得了显著的效果。Zhao 等[10]提出了一种语义图卷积结构(SemGCN),旨在通过学习邻接节点间的隐含关联与拓扑依赖来实现对人体结构语义特征的有效提取。Zou 等[11]提出调制图卷积网络(Modulated GCN),该网络引入权重调制和亲和度调制技术,解决了传统图卷积网络中不同关节之间共享特征变换的问题,但其共享同一个关系模式,在一定程度上限制了关节多样性的表达。Chen 等[12]提出了部件特殊关系共享策略(PSKS),该策略使得关节处理不同的邻居时可以根据邻居节点的特性学习对应的关系模式,能够更准确地捕捉关节间的内在联系和共性。但因其所有分组内部调制参数是静态的,难以自适应捕捉不同动作幅度的动态变化。尽管 GCN 能够捕捉人体关节拓扑结构,但由于其普遍使用共享卷积核机制,导致模型对所有关节同等建模,无法有效区分不同关节在运动幅度上的差异。这种对局部关节差异的忽视,限制了模型的精细化建模能力,使其在面对严重遮挡时,难以利用有效的结构先验推理出准确的关节位置。

综上所述,Transformer 架构和图卷积网络各有优点,将其有机融合能够更好地应用于三维人体姿态估计任务中。因此本文提出了基于动态多粒度图卷积的三维人体姿态估计网络,具体贡献如下:

(1)提出动态多粒度图卷积(Dynamic Multi-Granularity Graph Convolution,DMG-GCN),该网络通过对输入特征进行分组动态调制,突破了传统图卷积网络因共享静态卷积核而无法适应不同动作实例幅度差异的局限,能够更精细地捕捉局部关节的动态运动变化。

(2)设计动态多粒度图 Transformer 模块。该模块将多头自注意力机制(MSA)与 DMG-GCN 结合,有效实现了全局与局部空间特征的互补建模,显著提升了模型在复杂环境下的鲁棒性。

(3)构建双流时频特征融合 Transformer(Dual-stream Time-Frequency Feature Fusion Transformer,DTFF-Transformer)模块,利用交叉注意力机制深度挖掘时域瞬时细节与频域全局趋势之间的互补性,解决单一域特征表达不充分的问题。

1 方法

本文提出一种基于动态多粒度图卷积的三维人体姿态估计网络。该网络以 2D 姿态序列 X 为输入,首先采取序列的中心帧作为动态多粒度图 Transformer 模块的输入,表示为 X^{Time},将模块得到的输出进行时间位置嵌入(Time Position Embedding, TPE),表示为 Z^{Time};同时使用 DCT 变换和低通滤波器处理全序列,通过线性映射提取动作的全局频域特征,再将得到的频域特征进行频域位置嵌入(Frequency Position Embedding, FPE),表示为 Z^{Fre}。随后将 Z^{Time} 和 Z^{Fre} 并行输入到双流时频特征融合 Transformer 模块。最后融合的特征经过回归头输出最终的三维姿态。本文的模型如图 1 所示。

1.1 动态多粒度图 Transformer

1.1.1 整体架构

为了弥补 Transformer 在局部拓扑感知能力上的局限,并解决传统图卷积无法适应人体多变运动幅度的问题,本文提出了动态多粒度图 Transformer

模块，即在空间 Transformer 块中嵌入了一个动态多粒度图卷积。通过这种设计，模型可以在保持 MSA 全局建模优势的同时，有效弥补了其对局部骨架结构感知能力的不足，提升模型在复杂环境下的鲁棒性。

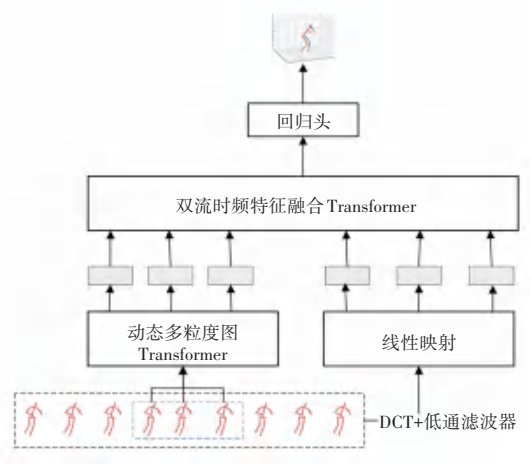


图 1　模型总体框架

Fig. 1　Overall architecture of the model

如图 2 所示，对于输入的二维姿态序列 $X \in R^{F\times J\times 2}$，模块首先对其采样，采样后的中心帧序列表示为 $X^{\text{Time}} \in R^{F'\times J\times 2}$，其中，$F$ 为原序列数量；F' 为中心帧数量；J 为关节点数量。而后将 X^{Time} 进行特征重塑，表示为 $X^{\text{Time}} \in R^{F'\times J\times C}$，其中 C 为通道数。将 X^{Time} 作为动态多粒度图 Transformer 的输入，首先通过多头自注意力机制进行处理，MSA 允许每个关节关注其他所有关节的信息，从而提取全局的空间依赖关系，计算公式如下：

$$Z = X^{\text{Time}} + \text{MSA}(\text{LN}(X^{\text{Time}})) \tag{1}$$

其中，LN(·)表示层归一化，MSA(·)表示 MSA 计算。尽管 MSA 能够捕捉长距离依赖，但无法利用人体固有的骨骼拓扑结构。为此本文在 MSA 后引入了 DMG-GCN 模块，该模块对 MSA 提取的全局特征进一步建模，分组捕捉局部关节的动态运动变化。计算公式如下：

$$Z' = Z + \text{DMG} - \text{GCN}(\text{LN}(Z)) \tag{2}$$

最后，特征通过 MLP 网络进行非线性变换，得到最终的输出 Z^{Time}：

$$Z^{\text{Time}} = Z' + \text{MLP}(\text{LN}(Z')) \tag{3}$$

其中 MLP(·)表示多层感知机。



图 2　动态多粒度图 Transformer

Fig. 2　Dynamic multi-granularity graph transformer

1.1.2　动态多粒度图卷积

针对传统图卷积无法适应人体多变运动幅度的问题，本文引入了动态多粒度图卷积模块。该模块将人体关节按照自由度分组，每组独立学习特征，使其能够自适应地处理不同类型的邻居关节，提高局部空间的建模能力。此外，其还引入动态通道调制（Dynamic Channel Modulation，DCM）机制，使得模型能够根据输入实例自适应调节特征权重。动态多粒度图卷积如图 3 所示。

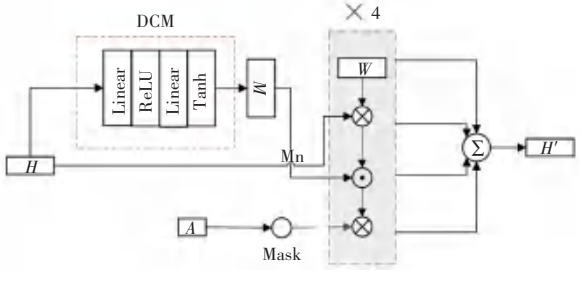


图 3　动态多粒度图卷积

Fig. 3　Dynamic multi-granularity graph convolution

对于传统图卷积，单层卷积层按照下式完成特征的变换与聚合：

$$\boldsymbol{H}' = \sigma(\boldsymbol{W}\boldsymbol{H}\hat{\boldsymbol{A}}) \tag{4}$$

其中，$\boldsymbol{H}' \in R^{D'\times J}$ 是更新后的特征矩阵；$\sigma(\cdot)$ 表示激活函数；$\boldsymbol{W} \in R^{D'\times D}$ 是权重矩阵；$\boldsymbol{H} \in R^{D\times J}$ 是输入特征矩阵；$\hat{\boldsymbol{A}}$ 是 $A \in \{0,1\}^{J\times J}$ 经过对称归一化操作的邻接矩阵，J 为节点个数；D 和 D' 表示特征维度。为了捕捉不同身体部位运动模式的差异，本文采用 Wang 等[13]提出的关节点划分策略，根据关节的运动自由度与解剖学位置，将人体骨骼拓扑划分为 3 个语义组：低自由度的躯干、中自由度的近端肢体以及高自由度的末端肢体。使网络能针对不同灵活度的邻接关节学习特定的特征变换模式。在此基础上，将矩阵分为 4 个子矩阵，定义如下：

$$A = \sum_{n=0}^{N-1} \boldsymbol{A}_n \tag{5}$$

其中，$\boldsymbol{N}$ 为子矩阵数量；$\boldsymbol{A}_0$ 为单位矩阵；$\boldsymbol{A}_1$、$\boldsymbol{A}_2$、$\boldsymbol{A}_3$ 分别为低、中、高自由度的 3 组矩阵。此时图卷积计算公式更新为：

$$\boldsymbol{H}^{'}=\sigma\left(\sum_{n=0}^{N-1}\boldsymbol{W}\boldsymbol{H}\hat{\boldsymbol{A}}_{n}\right) \tag{6}$$

其中，$\boldsymbol{W}\in R^{D^{'}\times D}$ 是权重矩阵；$\hat{\boldsymbol{A}}_{n}$ 是第 n 组归一化后的矩阵。为进一步解决静态特征变换矩阵 $\boldsymbol{W}$ 无法适应动态实例的问题，本文引入了 DCM 机制，该机制通过线性层和 ReLU 提取上下文特征，再由 tanh 激活函数处理生成调制权重，旨在根据当前的输入动态地调整特征的调制参数，从而实现更精细的特征学习。计算公式为：

$$\boldsymbol{M}=\tanh(W_{b}\cdot\mathrm{ReLU}(W_{a}\cdot H+b1)+b2) \tag{7}$$

其中，$\boldsymbol{M}\in R^{D^{'}\times J}$ 表示调制矩阵；W_{a}、W_{b} 分别为降维和升维的线性投影权重；ReLU(·)表示 ReLU 激活函数；$b1$、$b2$ 为归纳偏置；tanh(·)为激活函数。最终将调制系数控制在(-1,1)区间，使得模型不仅能增强或抑制特征，还能反转特征方向，从而灵活适应复杂的几何变换。生成的全局张量 $\boldsymbol{M}$ 随后根据语义分组解耦为 N 组，分别调制前面所述的 N 组区间的特征。最终 DMG-GCN 的公式更新为：

$$\boldsymbol{H}^{'}=\sigma\left(\sum_{n=0}^{N-1}(M_{n}\odot(\boldsymbol{W}\boldsymbol{H}))\hat{A}_{n}\right) \tag{8}$$

其中，$\boldsymbol{M}_{n}\in R^{D^{'}\times J}$ 为第 n 组的调制矩阵，$\odot$ 表示元素级相乘。

1.2 双流时频特征融合 Transformer

为克服单一域特征表达的局限性，并增强模型对复杂动作时空依赖的捕捉能力，本文设计了双流时频特征融合 Transformer，其中频域特征保留了动作的全局运动趋势，并且可以有效地抑制高频抖动噪声；而时域特征则提供局部运动的详细信息。将这两种特征进行交叉交互，能更充分地挖掘时域的瞬时信息和频域的全局信息之间的互补性，提升模型的性能，该模块如图 4 所示。

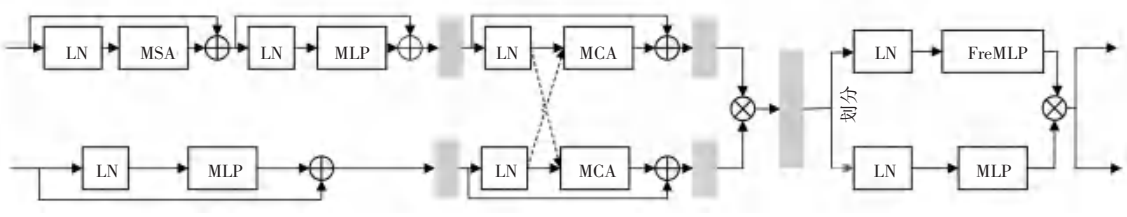


图 4 双流时频特征融合 Transformer

Fig. 4 Dual-stream time-frequency feature fusion Transformer

动态多粒度图 Transformer 输出带有空间信息的时域特征，将其重塑为 $Z^{\mathrm{Time}}\in R^{F^{'}\times(J\times C)}$，作为双流时频特征融合 Transformer 其中一支的输入，同时对原序列 $X\in R^{F\times J\times 2}$ 进行 DCT 变换，保留前 $F^{'}$ 个低频系数，将其通过线性投影到高维空间并进行重塑，得到的输出表示为 $Z^{\mathrm{Fre}}\in R^{F^{'}\times(J\times C)}$。将时域特征 Z^{Time} 和频域特征 Z^{Fre} 并行输入双流时频特征融合 Transformer 的两个分支。在时域分支，首先对时域特征进行 MSA 计算，捕捉动作序列中的时间相关性，随后将特征输入 MLP 进行深度提取，即对特征进行非线性变换，进一步提高特征的表达能力，时域特征的计算过程如下：

$$Z^{\mathrm{Time}}=Z^{\mathrm{Time}}+\mathrm{MSA}(\mathrm{LN}(Z^{\mathrm{Time}})) \tag{9}$$

$$Z^{\mathrm{Time}}=Z^{\mathrm{Time}}+\mathrm{MLP}(\mathrm{LN}(Z^{\mathrm{Time}})) \tag{10}$$

同时在频域分支中，考虑到频域特征本身包含全局信息，为了降低复杂度并保留全局信息，故而舍弃 MSA，直接采用 MLP 网络对频域特征进行映射与编码以保留全局运动趋势。该过程表示为：

$$Z^{\mathrm{Fre}}=Z^{\mathrm{Fre}}+\mathrm{MLP}(\mathrm{LN}(Z^{\mathrm{Fre}})) \tag{11}$$

随后将这两部分特征进行交叉注意力(MCA)计算，使时域特征与频域特征互为 Q、K 和 V，实现跨域信息的显式交互，从而利用频域的平滑性修正时域抖动，并借助时域细节补充高频信息。该过程定义为：

$$Z_{1}=Z^{\mathrm{Time}}+\mathrm{MCA}(\mathrm{LN}(Z^{\mathrm{Time}}),\mathrm{LN}(Z^{\mathrm{Fre}}),\mathrm{LN}(Z^{\mathrm{Fre}})) \tag{12}$$

$$Z_{2}=Z^{\mathrm{Fre}}+\mathrm{MCA}(\mathrm{LN}(Z^{\mathrm{Fre}}),\mathrm{LN}(Z^{\mathrm{Time}}),\mathrm{LN}(Z^{\mathrm{Time}})) \tag{13}$$

$$Z^{'}=\mathrm{Concat}(Z_{1},Z_{2}) \tag{14}$$

其中，MCA(Q,K,V)表示交叉注意力计算，Concat(·)表示级联操作。随后将融合后的特征划分为 2 部分，送入两个并行的分支。第一个分支采用频域感知机(FreMLP)模块，该模块能够在去除噪声的同时，有效恢复由低频滤波可能丢失的快速局部运动细节。第二个分支采用 MLP 对特征进行深度映射，旨在聚合 MCA 提取的特征，通过学习高阶非线性关系来增强特征的语义层次。这一过程定义为：

$$Z_1^{'}, Z_2^{'} = Z_{i,j,[0,C^{'}]}^{'}, Z_{i,j,[C^{'},k]}^{'} \tag{15}$$

$$Z_1^{'} = Z_1^{'} + \mathrm{FreMLP}(\mathrm{LN}(Z_1^{'})) \tag{16}$$

$$Z_2^{'} = Z_2^{'} + \mathrm{MLP}(\mathrm{LN}(Z_2^{'})) \tag{17}$$

其中，FreMLP(·)表示频域感知机，$Z_1^{'}$、$Z_2^{'}$ 表示划分出来的两部分特征。最后将两部分特征融合后划分，作为下一部分的输入，这一过程表示如下：

$$Z^{'} = \mathrm{Concat}(Z_1^{'}, Z_2^{'}) \tag{18}$$

$$Z_1, Z_2 = Z_{i,j,[0,C^{''}]}^{'}, Z_{i,j,[C^{''},k]}^{'} \tag{19}$$

双流时频特征融合 Transformer 利用了 MCA，有效地挖掘时域特征与频域特征之间的互补性，并结合 MLP 和 FreMLP 的非线性变换，实现时域与频域特征之间的深度交互，获得更鲁棒的姿态表示。

1.3　损失函数

本文使用平均每关节点误差(Mean Per Joint Position Error, MPJPE)[14]作为损失函数，计算每个关节点的三维姿态估计结果与真实姿态结果之间的平均欧几里得距离，公式如下：

$$L_{\mathrm{MPJPE}} = \frac{1}{N}\sum_{n=1}^{N} \| Y_{\mathrm{per}}^{n} - Y_{\mathrm{GT}}^{n} \| \tag{20}$$

其中，Y_{per}^{n} 表示第 n 个关节点的估计值，Y_{GT}^{n} 表示第 n 个关节点的真实值。

2　实验

本文通过对比实验和消融实验，验证模型的有效性. 实验环境配置如下：使用 Windows 11 操作系统，配备 Intel(R) Xeon(R) Gold 6430 CPU@2.10 GHz 处理器、NVIDIA RTX 5880 Ada Generation GPU 及 48 GB 内存；开发环境采用 PyTorch 1.13.0，Python 3.9 和 CUDA 11.7 搭建。

2.1　实验细节

2.1.1　数据集及评价指标

Human3.6M[15]数据集是一个包含 360 万帧图像的公开三维人体姿态库。该数据集通过 4 台摄像机在不同视角下录制，包含了 11 名参与人员在实验室环境下执行的 15 类日常行为动作，例如吃饭、跑步、行走、落座等。为与现有主流方法保持一致，本文采用标准的划分方式：将 S1、S5、S6、S7 和 S8 作为训练集，将 S9 和 S11 指定为测试集。并将 MPJPE 作为评价指标，MPJPE 用于计算估计姿态与真实姿态的欧几里得距离，单位为 mm。

MPI-INF-3DHP[16]数据集是一个更具挑战性的三维姿态数据集，包含受限的室内场景和复杂的户外场景，由 8 位演员在 14 个摄像机视角中完成 8 个动作表演，其中测试集包含 6 名演员表演的 7 个动作。并采用 MPJPE、150 mm 范围内的正确关键点的百分比(Percentage of Correct Keypoints, PCK)以及曲线下面积(Area Under Curve, AUC)作为评估指标，其中 PCK 和 AUC 的结果以%为单位。

2.1.2　实验超参数

在本次实验中，首先对原序列采样，选取 $F/9$ 帧作为动态多粒度图 Transformer 的输入，同时选取同等数量的低频系数作为 DTFF-Transformer 的频域输入，其中 F 为原序列长度。此外，本文的动态多粒度图 Transformer 和 DTFF-Transformer 网络均堆叠 4 层用于特征提取，多头自注意力头数设置为 8。在训练过程中，本文使用 AdamW 优化器更新网络参数，初始学习率设置为 0.001，权重衰减为 0.1，训练轮数为 80，批次大小为 1 024。

2.2　对比试验

在 Human3.6M 数据集中，本文使用级联金字塔网络(CPN)[17]的结果作为输入数据。而对于 MPI-INF-3DHP 数据集，本文则采用地面真实二维姿态标签作为输入。所有指标的最优结果用粗体标注。

表 1 展示了本文模型与当前主流方法在 Human3.6 M 数据集上不同动作的 MPJPE 的对比结果，所有方法均以 81 帧为输入，每个姿态的最优结果用加粗表示。可以看出，本文提出的方法在 Human3.6 M 数据集上取得了良好的效果，MPJPE 的值仅为 44.8 mm，优于 PoseFormerV2、DGFormer 等主流方法。相比于基准模型 PoseFormerV2，本文方法的 MPJPE 降低了 1.2 mm。对大幅度以及复杂动作(如拍照、坐下、摆姿势)和遮挡场景(如同行、坐着)能进行更准确地三维姿态预测。对比结果表明，本文方法在复杂动作下可以通过动态多粒度图卷积提升对局部特征的感知能力，使模型在应对复杂运动时表现出色；同时，凭借双流时频特征融合 Transformer 对时频特征的融合互补，模型在遮挡场景中也展现了极强的鲁棒性，这些数据证明了本文方法在处理人体动作幅度动态变化及遮挡场景下的有效性。

表 2 展示了本文模型与当前主流方法在 MPI-INF-3DHP 数据集上的对比结果。实验结果显示，本文方法在正确关键点百分比(PCK)，曲线下面积(AUC)和 MPJPE 评估指标上的值分别为 98%、79.5% 以及 25.7 mm，均优于 PoseFormerV2、P-STMO 等主流方法，验证了本文方法在 MPI-INF-3DHP 数据集上的优越性。

表 1 Human3.6M 数据集上的结果

Table 1 Results in Human3.6M dataset

方法	指路	讨论	吃饭	问候	打电话	拍照	摆姿势	购物	坐着	坐下	吸烟	等待	遛狗	走路	同行	平均误差
AMPose[18]	44.9	49.3	45.2	48.8	51.3	58.6	47.8	44.8	57.1	66.5	49.9	46.4	52.9	39.0	40.6	49.5
DG-Net[19]	41.5	46.6	41.0	**44.3**	47.1	54.1	44.2	**42.5**	54.9	58.8	46.9	**43.1**	46.9	32.6	35.6	45.3
GraFormer[20]	49.3	53.9	54.1	55.0	63.0	69.8	51.1	53.3	69.4	90.0	58.0	55.2	60.3	47.4	50.6	58.7
DGformer[21]	45.8	49.6	46.2	49.6	51.4	58.7	48.9	46.2	56.6	65.1	50.9	47.2	53.2	38.8	41.5	50.0
SemGCN[10]	48.2	60.8	51.8	64.0	64.6	53.6	51.1	67.4	88.7	**57.7**	73.2	65.6	48.9	64.8	51.9	60.8
Poseformerv2[9]	42.5	46.7	41.9	45.4	47.4	53.5	43.8	43.7	55.8	65.2	46.2	44.7	47.1	32.1	33.3	46.0
本文	**41.2**	**45.3**	**40.8**	44.6	**46.2**	**52.1**	**42.3**	42.6	**54.1**	63.2	**45.0**	44.9	**45.5**	**31.3**	**32.3**	**44.8**

表 2 MPI-INF-3DHP 数据集上的结果

Table 2 Results in MPI-INF-3DHP dataset

方法	PCK/%	AUC/%	MPJPE/mm
PoseFormerV1[8]	95.4	63.2	57.7
MixSTE[22]	94.4	66.5	54.9
AMPose[18]	**98.0**	79.1	30.8
MHFormer[23]	93.8	63.3	58.0
P-STMO[24]	97.9	78.8	32.2
PoseFormerV2[9]	97.9	78.8	27.8
本文	**98.0**	**79.5**	**25.7**

2.3 消融实验

为验证各个组件对模型的作用,本文开展了多组消融实验。在基准模型 PoseFormerV2 的基础上开展了消融实验,主要研究 DMG-GCN 以及双流时频特征融合 Transformer 模块对基准模型性能的影响。本次实验统一以 27 帧图片作为输入,从基准模型开始,首先在 PoseFormerV2 的基础上引入 DMG-GCN 模块,由表 3 可以看出加入 DMG-GCN 后模型较基准模型提升了 2.9%。而在 PoseFormerV2 基础上用双流时频特征融合 Transformer 模块时,模型误差为 46.7 mm,较基准模型提升了 2.5%。最后使用本文所提模型时,在 Human3.6M 数据集上的 MPJPE 达到了 46.1 mm,相比基准模型提升了 3.8%,这充分证明了本文模型的有效性。

表 3 引入不同组件在 Human3.6M 数据集上的 MPJPE 值比较

Table 3 Comparison of MPJPE on the Human3.6M dataset with different components

方法	指路	讨论	吃饭	问候	打电话	拍照	摆姿势	购物	坐着	坐下	吸烟	等待	遛狗	走路	同行	平均误差
PoseformerV2	43.9	49.0	44.2	46.4	50.3	56.2	44.6	43.6	58.2	67.4	49.2	45.5	49.7	34.0	36.2	47.9
+DMG-GCN	42.6	47.6	43.1	45.2	48.8	53.8	42.5	42.8	56.1	64.2	48.1	45.6	48.0	33.8	35.9	46.5
+DTFF-Transformer	42.8	47.9	43.3	45.4	49.1	55.3	43.4	43.2	55.4	65.1	47.6	45.4	48.4	33.2	35.4	46.7
本文	42.3	47.0	42.7	45.2	48.4	53.5	42.9	42.6	55.3	63.8	47.5	45.1	47.8	33.0	35.1	46.1

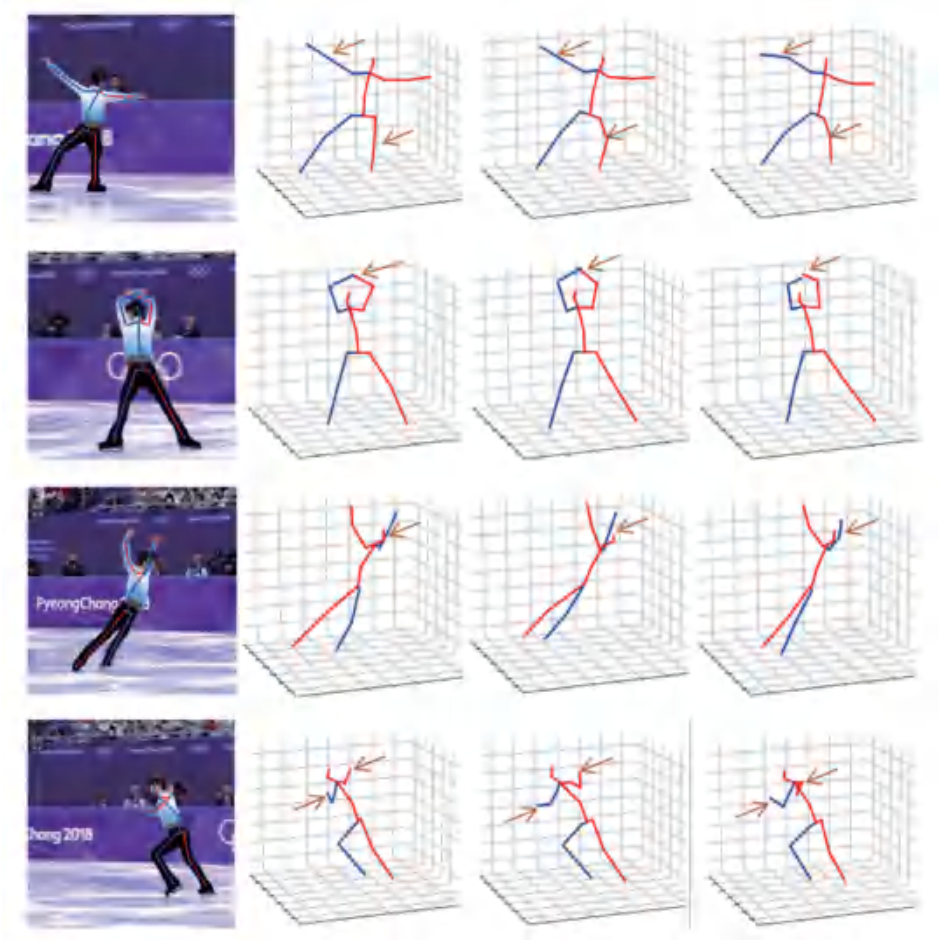
为进一步探索各组件在不同场景下的作用,本文对 Human3.6M 数据集中的具体动作进行了详细分析。从表 3 可以看出,在涉及到大幅度运动和复杂姿态的场景中,DMG-GCN 表现出了显著优势,例如在“坐下”这种场景下 MPJPE 由 67.4 mm 降至 64.2 mm,而在“拍照”这种场景下 MPJPE 由 56.2 mm 降至 53.8 mm。而在涉及到遮挡场景时,DTFF-Transformer 则表现出了优越性,该模块可以对时域和频域特征进行深度交互,利用频域特征的全局运动趋势推理出关节的合理位置。当同时加入 DMG-GCN 和 DTFF-Transformer 模块时,模型综合性能达到了最优,意味着两个模块结合,共同提升了模型在处理复杂时空依赖关系时的鲁棒性,使其在面对大部分场景时能保持良好的预测性能。

2.4 可视化结果

可视化实验中,为展示本文所提模型在三维人体姿态估计任务中的优势,本次实验统一采用 HRNet[25] 作为二维姿态检测器,将本文方法与当前主流方法 PoseFormerV2 与 MHFormer 在同一帧图片上进行定性比较,可视化结果如图 5 所示。

从图片上可以看出,本文方法在多个区域的预测均优于其他模型,尤其是在运动幅度较大的末端肢体(如手腕,脚踝等)上表现更突出,该效果得益于模型结合了 MSA 和 DMG-GCN,使得模型能够同时捕捉全局依赖关系和局部特征,并且能对不同自由度的肢体关节分区建模,从而有效捕捉不同部位间的运动学差异。此外,从第 4 行图片上可以看出,在发生遮挡时本文模型也能根据上下文预测出准确的姿态表示,这得益于本文提出的 DTFF-Transformer

模块，其能有效融合时域特征与频域特征，利用频域特征的全局运动趋势修正时域特征的高频抖动，帮助模型在特征缺失的情况下推理出合理的关节位置。综上所述，可视化结果充分验证了本文所提方法在处理复杂动作及遮挡场景下的优越性。



（a）输入　（b）本文　（c）PoseFormerV2　（d）MHFormer

图 5　可视化结果

Fig. 5　Visualization results

3　结束语

针对 Transformer 方法无法提取局部特征以及传统图卷积网络难以适配人体动作幅度动态变化的问题，本文提出一种融合动态多粒度图卷积与时频深度交互的三维人体姿态估计方法。首先引入 DMG-GCN 模块，该网络根据输入动态改变关节点的调制参数，使得模型能够差异化建模关节点的局部特征，提升模型在复杂姿态下的精度。随后引入多粒度图 Transformer 模块，将 DMG-GCN 模块与 MSA 串联，使模型能同时捕捉局部物理连接与全局语义依赖关系，有效缓解遮挡问题。最后引入 DTFF-Transformer 模块，利用 MCA 挖掘时域特征与频域特征的互补信息，利用频域的平滑性修正时域抖动，并借助时域细节补充高频信息，解决单一域特征表达不充分的局限。最后在 Human3.6M 和 MPI-INF-3DHP 数据集上的实验结果表明，本文所提出的方法能够有效降低关节预测误差，提升模型在复杂场景以及遮挡环境下的检测精度。

参考文献

[1] 樊虹岐，陈海军，吕丞干，等. 基于 ROS 的工业机器人及 AGV 的视觉协同控制系统[J]. 组合机床与自动化加工技术，2023(11)：121-127.

[2] Zhou Xingyi，Huang Qixing，Sun Xiao，et al. Towards 3D human pose estimation in the wild：A weakly-supervised approach[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway，NJ：IEEE，2017：398-407.

[3] 陈荣桂，贾振堂. 改进的三维人体姿态估计算法[J]. 计算机系统应用，2024，33(4)：187-193.

[4] 梁桉源，肖学中. 基于多层空间特征融合的三维人体姿态估计[J]. 计算机系统应用，2024，33(8)：250-256.

[5] Vaswani A，Shazeer N，Parmar N，et al. Attention is all you need

[C]//Proceedings of the Advances in Neural Information Processing Systems. NeurIPS,2017:5998-6008.
[6] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[C]//Proceedings of the Advances in Neural Information Processing Systems. NeurIPS,2020: 1877-1901.
[7] Liu Ze, Lin Yutong, Cao Yue, et al. Swin transformer: Hierarchical vision transformer using shifted windows [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2021: 10012-10022.
[8] Zheng Ce, Zhu Sijie, Mendieta M, et al. 3D Human pose estimation with spatial and temporal transformers [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2021: 11656-11665.
[9] Zhao Qitao, Zheng Ce, Liu Mengyuan, et al. Poseformerv2: Exploring frequency domain for efficient and robust 3D human pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023:8877-8886.
[10] Zhao Long, Peng Xi, Tian Yu, et al. Semantic graph convolutional networks for 3D human pose regression [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 3425-3435.
[11] Zou Zhiming, Tang Wei. Modulated graph convolutional network for 3D human pose estimation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2021: 11477-11487.
[12] Chen Lu, Liu Qiong. Relation-balanced graph convolutional network for 3D human pose estimation [J]. Image and Vision Computing, 2023, 140: 104841.
[13] Wang Jue, Huang Shaoli, Wang Xinchao, et al. Not all parts are created equal: 3D pose estimation by modeling bi-directional dependencies of body parts[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2019: 7771-7780.
[14] Liu Jiajie, Liu Mengyuan, Liu Hong, et al. TCPFormer: Learning temporal correlation with implicit pose proxy for 3D human pose estimation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. AAAI , 2025: 5478-5486.
[15] Ionescu C, Papava D, Olaru V, et al. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 36(7): 1325-1339.
[16] Mehta D, Rhodin H, Casas D, et al. Monocular 3D human pose estimation in the wild using improved cnn supervision [C]// Proceedings of 2017 International Conference on 3D Vision. Piscataway, NJ: IEEE, 2017: 506-516.
[17] Chen Yilun, Wang Zhicheng, Peng Yuxiang, et al. Cascaded pyramid network for multi-person pose estimation [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7103-7112.
[18] Lin Hongxin, Chiu Yunwei, Wu Peiyuan. AMPose: Alternately mixed global-local attention model for 3D human pose estimation [C]//Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2023: 1-5.
[19] Zhang Junhao, Wang Yali, Zhou Zhipeng, et al. Learning dynamical human-joint affinity for 3D pose estimation in videos [J]. IEEE Transactions on Image Processing, 2021, 30: 7914-7925.
[20] Zhao Weixi, Wang Weiqiang, Tian Yunjie. Graformer: Graph-oriented transformer for 3D pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 20438-20447.
[21] Chen Zhangmeng, Dai Ju, Bai Junxuan, et al. DGFormer: Dynamic graph transformer for 3D human pose estimation[J]. Pattern Recognition, 2024, 152: 110446.
[22] Zhang Jinlu, Tu Zhigang, Yang Jianyu, 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. Piscataway, NJ: IEEE, 2022: 13232-13242.
[23] Li Wenhao, Liu Hong, Tang Hao, et al. MHFormer: Multi-hypothesis transformer for 3D human pose estimation [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 13147 - 13156.
[24] Shan Wenkang, Liu Zhenghua, Zhang Xinfeng, et al. P-STMO: Pre-trained spatial temporal many-to-one model for 3D human pose estimation [C]//Proceedings of the European Conference on Computer Vision. Cham: Springer, 2022: 461-478.
[25] Sun Ke, Xiao Bin, Liu Dong, et al. Deep high-resolution representation learning for human pose estimation [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 5693-5703.