软件学报 ISSN 1000-9825, CODEN RUXUEW
2025,36(5):2151−2166 [doi: 10.13328/j.cnki.jos.007200] [CSTR: 32375.14.jos.007200]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于平行多尺度时空图卷积网络的三维人体姿态估计算法*

杨红红[1,2], 刘泓希[1], 张玉梅[2,3], 吴晓军[1,2,3]

[1](现代教学技术教育部重点实验室 (陕西师范大学), 陕西 西安 710062)

[2](民歌智能计算与服务技术文化和旅游部重点实验室 (陕西师范大学), 陕西 西安 710062)

[3](陕西师范大学 计算机科学学院, 陕西 西安 710062)

通信作者: 张玉梅, E-mail: zym0910@snnu.edu.cn

摘　要: 针对基于图卷积神经网络 (GCN) 的人体姿态估计方法不能充分聚合关节点时空特征、限制判别性特征提取的问题, 构造基于平行多尺度时空图卷积的网络模型 (PMST-GNet), 提高三维人体姿态估计的性能. 该模型首先设计对角占优的时空注意力图卷积 (DDA-STGConv), 构建跨域时空邻接矩阵, 对骨架关节点信息进行基于自约束和注意力机制约束的建模, 增强节点间的信息交互; 然后, 通过设计图拓扑聚合函数构造不同的图拓扑结构, 以 DDA-STGConv 为基本单元构建平行多尺度子网络模块 (PM-SubGNet); 最后, 为了更好地提取骨架关节的上下文信息, 设计多尺度特征交叉融合模块 (MFEB), 实现平行子图网络之间多尺度信息的交互, 提高 GCN 的特征表示能力. 在主流 3D 姿态估计数据集 Human3.6M 和 MPI-INF-3DHP 数据集上的对比实验结果表明, 所提 PMST-GNet 模型在三维人体姿态估计中具有较好的效果, 优于 Sem-GCN、GraphSH、UGCN 等当前基于 GCN 网络的主流算法.

关键词: 三维人体姿态估计; 对角占优的时空注意力图卷积; 平行多尺度子网络; 多尺度特征交叉融合

中图法分类号: TP391

中文引用格式: 杨红红, 刘泓希, 张玉梅, 吴晓军. 基于平行多尺度时空图卷积网络的三维人体姿态估计算法. 软件学报, 2025, 36(5): 2151–2166. http://www.jos.org.cn/1000-9825/7200.htm

英文引用格式: Yang HH, Liu HX, Zhang YM, Wu XJ. Parallel Multi-scale Spatio-temporal Graph Convolutional Network for 3D Human Pose Estimation. Ruan Jian Xue Bao/Journal of Software, 2025, 36(5): 2151–2166 (in Chinese). http://www.jos.org.cn/1000-9825/7200.htm

Parallel Multi-scale Spatio-temporal Graph Convolutional Network for 3D Human Pose Estimation

YANG Hong-Hong[1,2], LIU Hong-Xi[1], ZHANG Yu-Mei[2,3], WU Xiao-Jun[1,2,3]

[1](Key Laboratory of Modern Teaching Technology (Shaanxi Normal University), Ministry of Education, Xi'an 710062, China)

[2](Key Laboratory of Intelligent Computing and Service Technology for Folk Song (Shaanxi Normal University), Ministry of Culture and Tourism, Xi'an 710062, China)

[3](School of Computer Science, Shaanxi Normal University, Xi'an 710062, China)

Abstract: As the human pose estimation (HPE) method based on graph convolutional network (GCN) cannot sufficiently aggregate spatiotemporal features of skeleton joints and restrict discriminative features extraction, in this paper, a parallel multi-scale spatio-temporal graph convolutional network (PMST-GNet) model is built to improve the performance of 3D HPE. Firstly, a diagonally dominant spatiotemporal attention graph convolutional layer (DDA-STGConv) is designed to construct a cross-domain spatiotemporal adjacency matrix and model the joint features based on self-constraint and attention mechanism constrain, therefore enhancing information interaction among nodes. Then, a graph topology aggregation function is devised to construct different graph topologies, and a parallel multi-scale sub-graph network module (PM-SubGNet) is constructed with DDA-STGConv as the basic unit. Finally, a multi-scale feature cross fusion

* 基金项目: 国家自然科学基金 (61907028, 11872036); 陕西省青年科技新星项目 (2021KJXX-91); 文化和旅游部重点实验室资助项目 (2023-02, 2022-13); 陕西省自然科学基金面上项目 (2024JC-YBMS-503)
收稿时间: 2022-11-14; 修改时间: 2023-07-20, 2024-01-08; 采用时间: 2024-04-07; jos 在线出版时间: 2024-06-20
CNKI 网络首发时间: 2024-06-21

block (MFEB) is designed, by which multi-scale information among PM-SubGNets can interact to improve the feature representation of GCN, therefore better extracting the context information of skeleton joints. The experimental results on the mainstream 3D HPE datasets Human3.6M and MPI-INF-3DHP show that the proposed PMST-GNet model has a good effect in 3D HPE and is superior to the current mainstream GCN-based algorithms such as Sem-GCN, GraphSH, and UGCN.

Key words: 3D human pose estimation (3D HPE); diagonally dominant spatio-temporal attention graph convolution; parallel multi-scale sub-graph network; multi-scale feature cross fusion

三维人体姿态估计 (3D human pose estimation, 3D HPE) 是计算机视觉研究领域的一个研究热点, 其在行为识别、异常行为检测、人机交互和智能视频分析等方面具有广泛的应用前景[1,2]. 3D HPE 的目的是从图像、视频中估计人体的 3D 关节坐标. 随着深度神经网络的快速发展, 基于深度神经网络的 3D 人体姿态估计研究取得了丰硕的成果, 但是, 由于遮挡、单视角 2D 到 3D 映射中固有的深度模糊性等不确定因素, 从单目视频图像中估计 3D 人体姿势仍然是一项具有挑战性的任务.

近年来, 基于深度神经网络框架的 3D 人体姿态估计方法主要分为两大主流: (1) 从 RGB 图像直接估计 3D 人体关节点坐标的方法; (2) 基于两阶段 2D-3D 人体姿态的估计方法. 前者直接从二维输入图像中回归 3D 人体关节坐标信息. 由于 2D 图像包含丰富的像素信息, 此类方法能够直接从图像中捕获目标原始信息, 但受图像噪声影响严重且受限于有限的三维标注数据[3,4]. 基于两阶段的 3D 人体姿态估计方法首先从原始图像中检测人体的 2D 关节点, 然后进行 2D 到 3D 的投影, 从而获得人体的 3D 关节点坐标信息[5,6]. 由于 2D 关节点提供了高度抽象的人体骨架信息, 因此, 此类方法充分利用二维骨骼数据进行 3D HPE, 其不仅有利于摆脱图像噪声的干扰, 而且得益于动捕设备提供的三维数据进行网络训练.

由于人体骨架的拓扑结构可以自然地建模为图结构, 而图卷积网络 (graph convolution network, GCN) 具有能够直接处理关节点拓扑图的能力, 其在基于两阶段的 2D-3D 的人体姿态估计研究中得到了广泛的关注. 基于 GCN 的 2D-3D 人体姿态估计算法将 2D 关节点作为图的节点, 将人体关节的自然连接作为图网络的边[7]. 尽管基于 GCN 的 2D-3D 人体姿态估计算法取得了较好的 3D 姿态估计结果, 但是, 此类方法仍然存在一定的局限性. 首先, 单视角 2D-3D 映射中固有的深度模糊性问题, 导致多个不同的三维姿势可以映射到同一个二维骨架, 造成 2D 向 3D 推理时的不确定性. 其次, 大部分基于 GCN 的 2D-3D 人体姿态估计算法采用文献 [8] 中的图卷积网络进行关节点特征提取, 其通过在单个尺度上构造节点 1 邻域的依赖关系聚合关节特征, 此类构造方法不利于远端关节点局部和全局信息的聚合. 由于其接收域限定为 1, 一定程度上削弱了网络特征表示的能力, 加之, 在 3D 人体姿态估计中涉及 2D 到 3D 特征的映射, 其在 2D 关节点估计中微小的误差都将会在三维空间中产生巨大的影响.

研究表明, 从时间和空间维度对骨架关节点信息进行建模, 构建基于 GCN 的时空图卷积网络可以有效捕获关节点序列复杂的空间结构特征和长时动态特性, 对于消除 3D HPE 中的遮挡和深度模糊性问题至关重要. 然而, 现有的算法, 如文献 [5−7], 大多采用空域与时域特征串联的框架, 此框架将空域和时域信息同等对待, 且在卷积过程中具有固定的感受野, 无法有效处理空域与时域特征的不均衡问题, 限制了判别性特征的提取. 因此, 如何有效的提取骨架关节点序列的时空相关性特征, 对于提升 3D HPE 的性能具有重要意义.

针对上述问题, 本文构造基于平行多尺度时空图卷积网络模型 (parallel multi-scale spatio-temporal graph convolution network model, PMST-GNet) 的三维人体姿态估计算法. 首先, 对时空图卷积进行改进, 设计对角占优的时空注意力图卷积 (diagonally dominant spatio-temporal attention graph convolution, DDA-STGConv), 构建跨域时空邻接矩阵, 从时间和空间维度出发对骨架关节点信息进行基于自约束和注意力机制约束的建模, 通过提升关节点之间信息的传递来增强节点的特征表达; 其次, 根据人体运动中关节点呈现的相似运动趋势, 设计图拓扑聚合函数, 构造不同尺度的图拓扑结构, 以 DDA-STGConv 为基本单元构建平行多尺度子网络模块 (parallel multi-scale sub-graph network, PM-SubGNet), 有效提取骨架关节点的局部和全局特征信息; 最后, 构造多尺度特征交叉融合模块 (multi-scale feature cross fusion block, MFEB), 实现平行子图网络之间多尺度信息的交互, 进一步提高 GCN 的特征表示能力.

综上所述, 本文的创新点主要有以下 3 点.

(1) 针对传统 GCN 在空域特征和时域特征提取过程中采用串联结构, 无法有效处理空域与时域特征不均衡的

问题, 本文设计一种平行多尺度时空图卷积网络模型 (PMST-GNet), 通过构建平行多尺度子网络模块 (PM-SubGNet), 有效地对骨架关节的时空信息进行建模.

(2) 由于传统的 GCN 在时空特征提取过程采用固定的图拓扑结构, 造成节点特征随网络深度加深产生同化的问题, 本文设计一种对角占优的时空注意力图卷积 (DDA-STGConv), 通过构建跨域时空邻接矩阵, 从时间和空间维度出发对骨架关节点信息进行基于自约束和注意力机制约束的建模, 提升关节点之间信息的传递, 增强节点的特征表达.

(3) 为了提高骨架关节点局部和全局信息的交互, 克服单尺度图卷积仅描述单个尺度中人体内部关节关联关系的局限性, 本文构建平行多尺度子网络模块 (PM-SubGNet), 通过设计图拓扑聚合函数, 构造不同尺度的图拓扑结构, 对不同身体部位之间的关联关系进行建模, 实现不同尺度节点特征的聚合.

1 时空图卷积神经网络

GCN 对人体骨架进行建模的核心思想是通过关节点之间的信息传递聚合节点特征, 其对于 2D 骨架序列数据而言, 构造以关节点为图节点 $\mathbf{V}=\{v_{ti}\mid t=1,\ldots,T;i=1,\ldots,N\}$, 以人体关节之间的自然连接 (空域) 和关节在时间维度上连接 (时域) 为图边 $\mathbf{E}=\{[e_{ij},e_{ti}]\mid t=1,\ldots,T;i,j=1,\ldots,N\}$ 的时空图 $\mathbf{G}=(\mathbf{V},\mathbf{E},\mathbf{A})$. 其中, $\mathbf{V}$ 表示由连续 T 帧, 每帧 N 个关节点坐标构成的图节点集, $\mathbf{E}$ 为由帧内及帧间关节点连通构成的图边集, $\mathbf{E}$ 由邻接矩阵 $\mathbf{A}=(\alpha_{ij})_{M\times M}$ 编码关节点之间的连接关系, 若关节点 i 和 j 之间具有物理连接, 则 $\alpha_{ij}=1$, 反之, $\alpha_{ij}=0$.

1.1 空域图卷积

空域图卷积 (spatial-graph convolution, S-GC) 通过对同一帧中的关节点进行卷积操作以对不同节点的特征进行聚合. 对于任意 t 时刻的图像而言, 存在 N 个人体骨架节点 $\mathbf{V}_t=\{v_{ti}|i=1,\ldots,N\}$, 其人体关节点连接形成的骨架空域边为 $\mathbf{E}_t=\{e_{ij}(v_{ti},v_{tj})\mid i,j=1,\ldots,N\}$, 则相应的空域图卷积可以表示为:

$$x_{v_i}^{(l+1)}=\sigma\left(\sum_{v_j\in B(v_i)}w(v_i,v_j)x_{v_j}^{(l)}\alpha_{ij}\right),i\in\{1,\ldots,N\},l\in\{1,\ldots,L\} \tag{1}$$

其中, $x_{v_j}^{(l)}$ 是节点 v_j 在第 l 层的输入特征, $x_{v_i}^{(l+1)}$ 为相应的输出, 同时也是 (l+1) 层的输入特征. $w(\cdot)$ 为权值向量, $\sigma(\cdot)$ 为激活函数, $B(v_i)$ 为节点 v_i 的采样邻域节点集, 表示为:

$$B(v_i)=\{v_j\mid d(v_i,v_j)\leqslant d'\} \tag{2}$$

其中, $d(v_i,v_j)$ 表示两个节点 v_i 和 v_j 之间的最短路径, d' 为预定义的最大采样距离. 大多数现有的 S-GC, 设置 $d'=1$, 即使用关节点的 1 阶邻域信息进行特征聚合.

对所有的关节点执行公式 (1) 中空域图卷积操作, 则公式 (1) 更新为:

$$\mathbf{X}_{\mathrm{S}}^{(l+1)}=\sigma(\mathbf{W}^{(l)}\mathbf{X}^{(l)}\mathbf{A}_{\mathrm{S}}) \tag{3}$$

其中, $\mathbf{X}^{(l)}\in\mathbb{R}^{D_l\times N}$ 和 $\mathbf{X}^{(l+1)}\in\mathbb{R}^{D_{l+1}\times N}$ 分别为 S-GC 第 l 层更新前和更新后的特征矩阵, D 为每个节点的特征维度, N 为节点数量. $\mathbf{W}^{(l)}\in\mathbb{R}^{D_{(l+1)}\times D_l}$ 为相应的权值矩阵, $\mathbf{A}_{\mathrm{S}}$ 为空域邻接矩阵.

1.2 时域图卷积

与空域图卷积类似, 时域图卷积 (temporal-graph convolution, T-GC) 通过在连续帧中对相同节点进行特征加权聚合实现骨架关节点时域信息的提取, 其表示为:

$$x_{v_{qi}}^{(l+1)}=\sigma\left(\sum_{v_{qi}\in B(v_{ti})}wx_{v_{qj}}^{(l)}\alpha_{ti,(t+1)i}\right),i\in\{1,\ldots,N\},l\in\{1,\ldots,L\} \tag{4}$$

其中, $B(v_{ti})=\left\{v_{qi}\middle||q-t|\leqslant\left\lfloor\frac{\Gamma}{2}\right\rfloor\right\}$ 为节点 v_i 在连续帧中的采样邻节点集, 其相应的时域图卷积核函数大小为 $\Gamma\times1$.

同样的, 对所有节点执行时域图卷积操作, 其相应的时域卷积层可以表示为:

$$\mathbf{X}_{\mathrm{T}}^{(l+1)}=\sigma(\mathbf{W}\mathbf{X}^{(l)}\mathbf{A}_{\mathrm{T}}) \tag{5}$$

其中, $\mathbf{A}_T$ 编码同一节点在第 t 帧中前向、后向的邻节点连接关系.

2 平行多尺度时空图卷积模型

上述空域和时域图卷积在骨架关节点信息提取过程中主要存在如下问题: (1) 在特征加权聚合中, 对所有层采用固定的邻接矩阵, 将造成节点特征随网络加深而逐渐被同化的问题, 不能充分挖掘图节点之间的差异性; (2) 上述图卷积操作在空域和时域上仅使用单尺度聚合关节点 1 阶邻域信息, 由于其感受野固定为 1, 限制了中心节点与远端关节点之间的信息聚合; (3) 现有的时空图卷积通过交替执行 S-GC 和 T-GC 聚合关节点的时空信息, 其将空域和时域信息同等对待, 不能充分挖掘关节点的时空关系特征, 无法有效处理空域与时域特征的不均衡问题, 因此, 其在上下文时空特征提取中存在一定的局限性. 针对上述问题, 本文设计平行多尺度时空图卷积网络模型 (PMST-GNet), 如图 1 所示, PMST-GNet 模型通过设计以对角占优的时空注意力图卷积为基本单元的平行多尺度子网络模块 (PM-SubGNet), 有效地对骨架关节的时空信息进行建模; 同时, 设计多尺度特征交叉融合模块 (MFEB), 加强不同尺度特征的交互, 提高模型性能.

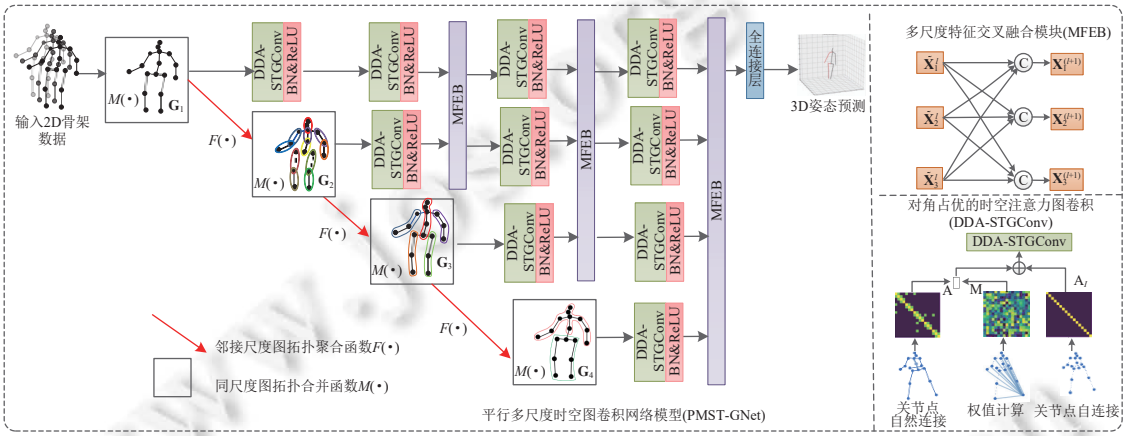

图 1 本文所提出的平行多尺度时空图卷积网络模型 (PMST-GNet)

2.1 对角占优的时空注意力图卷积

通过观察分析, 在图卷积网络特征更新过程中, 邻接矩阵 $\mathbf{A}$ 具有重要的作用, 如公式 (3) 和公式 (5) 所示, 其通过 $\mathbf{A}$ 实现特征从 $\mathbf{X}^{(l)}$ 到 $\mathbf{X}^{(l)}\mathbf{A}$ 的更新. 传统 GCN 方法基于人体自然骨架结构构造相应的 $\mathbf{A}$ (1 表示节点相连, 0 表示节点间不存在自然连接), 并且, 在所有的卷积层中都使用固定的 $\mathbf{A}$. 这就决定了所设计的网络模型不管使用什么结构, 采用多少个卷积层, 都保持固定的图拓扑结构, 其仅仅实现了从一个图映射到另一个图的作用. 但是, 对于不同的动作, 关节点之间的关系是不同的, 这种固定的图拓扑结构将造成节点特征随网络加深而逐渐被同化的问题[9]. 此外, 大多数现有的方法将时空信息分解为时域和空域两部分, 仅通过堆叠执行 S-GC 和 T-GC 达到融合骨架时空信息的目的, 然而, 当骨架关节点特征通过一系列如 S-GC 和 T-GC 的特征聚合操作后, 其不仅忽略了不同帧中关节之间的相互关系, 而且无法有效处理空域与时域特征的不均衡问题, 限制了判别性特征的提取.

针对上述问题, 本文提出对角占优的时空注意力图卷积 (DDA-STGConv), 所设计的 DDA-STGConv 从时间和空间维度出发对骨架关节点信息进行基于自约束和注意力机制约束的建模, 突出关节点自身对不同姿态的影响, 同时, 在时空域图卷积中引入注意力机制, 自适应的提取不同节点之间的动态关联性, 加强中心关节与远端关节的信息交互[10]. 此外, 本文设计了一个跨域的时空邻接矩阵, 其将 T 帧关节的帧内和帧间连接关系进行融合, 依靠 2D 卷积提取复杂的时空特征, 有效地对骨架关节的时空信息进行建模. 在此, 令 $\mathbf{X}^{(l)} \in \mathbb{R}^{D_l \times N}$ 和 $\mathbf{X}^{(l+1)} \in \mathbb{R}^{D_{l+1} \times N}$ 分别表示图卷积第 l 层的输入和输出, 其中, D 和 N 分别表示图卷积节点特征向量的维度和节点的个数, 所设计的 DDA-STGConv 的特点是构建了一个跨域的时空邻接矩阵 $\mathbf{A}_{ST}$, 其将 T 帧关节的帧内和帧间连接关系进行融合, 表示为:

$$\mathbf{X}^{(l+1)}=\sigma(\mathbf{W}^{(l)}\mathbf{X}^{(l)}(\mathbf{A}_{\mathrm{ST}}\odot\mathbf{M}_{\mathrm{att}}+\mathbf{A}_l)),l\in[1,\ldots,L]$$
$$\mathbf{A}_{\mathrm{ST}}=\begin{bmatrix}\mathbf{A}_{\mathrm{S}}^{1} & \mathbf{A}_{\mathrm{T}}^{12} & \cdots & \mathbf{A}_{\mathrm{T}}^{1i} & \cdots & \mathbf{A}_{\mathrm{T}}^{1T}\\ \mathbf{A}_{\mathrm{T}}^{21} & \mathbf{A}_{\mathrm{S}}^{2} & \cdots & \mathbf{A}_{\mathrm{T}}^{2i} & \cdots & \mathbf{A}_{\mathrm{T}}^{2T}\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ \mathbf{A}_{\mathrm{T}}^{j1} & \mathbf{A}_{\mathrm{T}}^{j2} & \cdots & \mathbf{A}_{\mathrm{S}}^{j} & \cdots & \mathbf{A}_{\mathrm{T}}^{jT}\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ \mathbf{A}_{\mathrm{T}}^{T1} & \mathbf{A}_{\mathrm{T}}^{T2} & \cdots & \mathbf{A}_{\mathrm{T}}^{Ti} & \cdots & \mathbf{A}_{\mathrm{S}}^{T}\end{bmatrix}\tag{6}$$
$$\mathbf{M}_{\mathrm{att}}=Softmax\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d}}\right)$$

其中, $\mathbf{A}_{\mathrm{ST}}$ 为由 T 帧骨架关节点帧内和帧间连接关系构成的时空域邻接矩阵, $\mathbf{M}_{\mathrm{att}}$ 为基于注意力计算的权值矩阵[11], $d=D_l$, $\mathbf{Q}$, $\mathbf{K}$ 分别代表当前节点特征与其图拓扑结构中邻域节点特征, 通过对输入特征 $\mathbf{X}^{(l)}\in\mathbb{R}^{D_l\times N_k}$ 进行线性变换获得相应的 $\mathbf{Q}=\mathbf{X}\mathbf{W}_{\mathbf{Q}}$, $\mathbf{K}=\mathbf{X}\mathbf{W}_{\mathbf{K}}$, $\mathbf{W}_{\mathbf{Q}}\in\mathbb{R}^{D_l\times D_l}$, $\mathbf{W}_{\mathbf{K}}\in\mathbb{R}^{D_l\times D_l}$. $\mathbf{A}_l\in\mathbb{R}^{N_s\times N_s}$ 为单位矩阵, N_s 为节点的数量, $\odot$ 代表矩阵的哈达玛积, 其中, N_k 代表多尺度中第 k 个尺度节点的个数, 其具体由第 2.2 节中多尺度结构划分决定.

由公式 (6) 可以得出, 通过所构造的 $\mathbf{A}_{\mathrm{ST}}$, 公式 (6) 将公式 (3) 和公式 (5) 中分开执行的时域图卷积和空域图卷积整合在一起. 其中, $\mathbf{A}_{\mathrm{ST}}$ 矩阵中主对角线的各元素表示同一帧中不同关节的自然连接构成的邻接矩阵, 称为空域邻接矩阵 $\mathbf{A}_{\mathrm{S}}^{j}\{j=1,\ldots,T\}$; 非对角矩阵的各元素表示不同帧中同一关节点在时间维度上的连接, 称为时域邻接矩阵 $\mathbf{A}_{\mathrm{T}}^{ji}\{i,j=1,\ldots,T\}$. 从公式 (6) 可以看出, DDA-STGConv 不仅依靠 $\mathbf{A}_{\mathrm{ST}}\odot\mathbf{M}_{\mathrm{att}}$ 自适应的提取不同节点之间的动态关联性, 在节点特征聚合中利用注意力机制衡量不同节点对当前节点的作用, 而且通过 $\mathbf{A}_l$ 突出关节点自身在特征聚合中对不同姿态的影响. 在此, 以 T=3 帧为例, 构建一个包含 3 帧信息的时空域邻接矩阵 $\mathbf{A}_{\mathrm{ST3}}$, 其中 $\mathbf{A}_{\mathrm{ST3}}=\begin{bmatrix}\mathbf{A}_{\mathrm{S}}^{1} & \mathbf{A}_{\mathrm{T}}^{12} & \mathbf{A}_{\mathrm{T}}^{13}\\ \mathbf{A}_{\mathrm{T}}^{21} & \mathbf{A}_{\mathrm{S}}^{2} & \mathbf{A}_{\mathrm{T}}^{23}\\ \mathbf{A}_{\mathrm{T}}^{31} & \mathbf{A}_{\mathrm{T}}^{32} & \mathbf{A}_{\mathrm{S}}^{3}\end{bmatrix}$, 根据 $\mathbf{A}_{\mathrm{ST}}$ 的定义, 其由 T 帧骨架关节点帧内和帧间连接关系构成, $\mathbf{A}_{\mathrm{ST}}$ 矩阵的主对角线的各元素 $\mathbf{A}_{\mathrm{S}}^{j}\{1\leqslant j\leqslant 3\}$ 表示同一帧中不同关节的自然连接构成的邻接矩阵, 称为空域邻接矩阵 $\mathbf{A}_{\mathrm{S}}^{j}$, j 为帧索引. $\mathbf{A}_{\mathrm{ST}}$ 矩阵的非对角的各元素表示不同帧中同一关节点在时间维度上的连接, 称为时域邻接矩阵 $\mathbf{A}_{\mathrm{T}}^{ji}\{1\leqslant i,j\leqslant 3\}$.

2.2 平行多尺度时空图卷积网络

现有的大多数基于 GCN 的 3D HPE 通过顺序堆叠多个图卷积层进行骨骼关节点特征的提取, 如文献 [5−7,12], 这种单通道串联模型仅对关节点特征进行从低层到高层的传播, 忽略了网络中间层特征在语义信息表征方面的优势, 因此, 受高分辨网络 HRNet 的启发[13], 本文设计平行多尺度时空图卷积网络模型 (PMST-GNet).

2.2.1 多尺度结构划分

由于人体运动是由人体部分关节的移动而形成的, 从而对于不同的动作, 其在同一个部件里的关节点具有相似的运动趋势和轨迹. 因此, 根据人体关节的运动趋势, 本文将人体关节点划分为如后文图 2 所示的 4 个尺度, k=1 由 17 个原始关节点构成; k=2 将原始关节按骨架层次性划分为 11 个子部件; k=3 通过将 k=2 尺度的 11 个部件关节进一步合并, 获得由人体左臂, 右臂, 躯干, 左腿和右腿关节组成的 5 个子部件; k=4 尺度在 k=3 的基础上将人体关节划分为由上半身和下半身关节组成的两个子部件. 在此, 虽然人体并不是刚体结构, 由局部到整体的划分并不能保证运动趋势完全一致, 但是在人们日常中表现出的动作以及不同动作中呈现出的姿态, 都会呈现一定的运动趋势, 对于不同的动作, 其在同一个部件里的关节点具有相似的运动趋势和轨迹, 比如, 在行走过程中, 上肢和下肢的多个关节点呈现出相似的运动趋势, 因此, 本文在多尺度划分过程中将运动趋势相似或运动邻域节点划分到一个部件中, 克服关节点物理连接对姿态估计的影响, 通过多尺度图呈现动态变化, 灵活地描述人体姿态的变化.

2.2.2 平行多尺度时空注意力图卷积模块

根据人体运动中关节点呈现的相似运动趋势, 本文所设计的 PMST-GNet 模型将人体骨架划分为由细到粗的 4 个尺度, 通过设计图拓扑聚合函数, 构造不同尺度的图拓扑结构, 实现不同尺度节点特征的聚合, 有效提取骨架关节点局部和全局的特征信息.

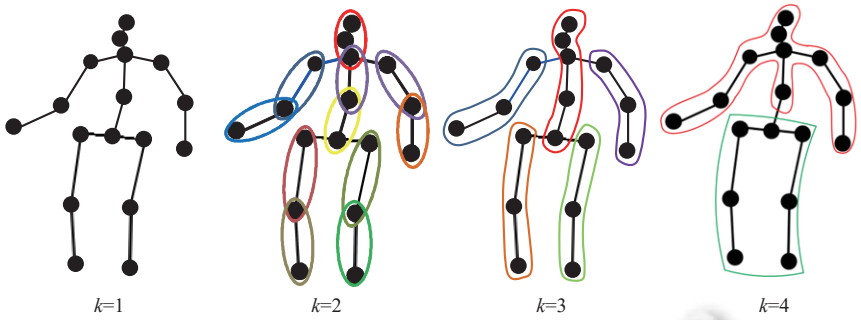

图 2 根据人体运动趋势划分的多尺度结构

● 图拓扑聚合函数: 对于任意尺度的节点 v_i, 以 v_i 为聚合中心, 以固定邻域 $B(v_i)$ 为半径对节点进行聚合, 使得每个聚类簇中心点代表 $B(v_i)$ 范围内的局部关节点, 由此聚类簇构成的子拓扑图结构表示为 $g_k^{v_i}=(\mathbf{V}_k^{v_i},\mathbf{E}_k^{v_i})$.

对于尺度 k 中所构成的 N_k 个子拓扑图 $g_1^{v_i}=(\mathbf{V}_1^{v_i},\mathbf{E}_1^{v_i}),i=1,\ldots,N_k$, 通过定义同尺度图拓扑合并函数 $M(\cdot)$, 构造 k 尺度的图拓扑结构:

$$\mathbf{G}_k=M\left(\sum_i^{N_k}g_k^{v_i}\right)=\left(\sum_i^{N_k}\cup\mathbf{V}_k^{v_i},\sum_i^{N_k}\cup\mathbf{E}_k^{i}\right) \tag{7}$$

对于尺度 k 中的任意两个以 v_i 和 v_j 为聚类中心构成的子拓扑图 $g_k^{v_i}=(\mathbf{V}_k^{v_i},\mathbf{E}_k^{v_i})$ 和 $g_k^{v_j}=(\mathbf{V}_k^{v_j},\mathbf{E}_k^{v_j})$, 定义邻接尺度图拓扑聚合函数 $F(\cdot)$, 通过对 k 尺度的子拓扑图进行融合, 构造 k+1 尺度的子图拓扑结构.

$$F(g_k^{v_i},g_k^{v_j})=(\mathbf{V}_k^{v_i}\cup\mathbf{V}_k^{v_j},\mathbf{E}_k^{v_i}\cup\mathbf{E}_k^{v_j}),B(v_i)=1 \tag{8}$$

其中, $B(v_i)=1$ 表示同一尺度子拓扑图合并的邻域范围. 对于以节点 v_i 为中心构成的聚类簇, 其对应的节点特征表示为 x_i, 在利用公式 (8) 进行子拓扑图聚合过程中, 对于两个子图中的重复节点, 其节点特征为 $g_k^{v_i}$ 和 $g_k^{v_j}$ 中相应节点特征的平均值.

● 平行多尺度子网络模块 (PM-SubGNet): 为了更好地提取骨架关节之间的上下文信息, 本文构造平行多尺度子网络模块 (PM-SubGNet), 如后文图 3 所示, PMST-GNet 由 4 个 PM-SubGNet 构成, 每一个子网络由多个 DDA-STGConv 层进行特征提取. 其中, 第 1 阶段的子网络以人体结构中的 $N_k=17$ 个原始关节点为聚类簇中心, $B(v_i)=0$ 为半径, 构造以自身节点, 自连接为边的 17 个子拓扑图 $g_1^{v_i}=(\mathbf{V}_1^{v_i},\mathbf{E}_1^{v_i}),i=1,\ldots,N_1$, $N_1=17$. 然后对该尺度的子拓扑图按照公式 (7) 执行同尺度图拓扑合并, 获得尺度 1 的图拓扑结构 $\mathbf{G}_1$. 第 2 阶段的子网络通过对尺度 1 的子拓扑图 $g_1^{v_i}=(\mathbf{V}_1^{v_i},\mathbf{E}_1^{v_i}),i=1,\ldots,N_1$ 按照公式 (8) 执行邻接尺度图拓扑聚合操作, 获得包含 11 个子部件的拓扑图 $g_2^{v_i}=(\mathbf{V}_2^{v_i},\mathbf{E}_2^{v_i}),i=1,\ldots,N_2$, $N_2=11$, 然后执行公式 (7), 获得尺度 2 的图拓扑结构 $\mathbf{G}_2$. 第 3 阶段的子网络通过将 k=2 尺度的 11 个部件关节子拓扑图按照公式 (8) 进一步合并, 获得由人体左臂、右臂、躯干、左腿和右腿关节组成的 5 个子部件的拓扑图 $g_3^{v_i}=(\mathbf{V}_3^{v_i},\mathbf{E}_3^{v_i}),i=1,\ldots,N_3$, $N_3=5$, 根据公式 (7), 获得尺度 3 的图拓扑结构 $\mathbf{G}_3$; 第 4 阶段的子网络在 k=3 尺度的基础上按照公式 (8) 进行节点聚合, 将人体关节划分为由上半身和下半身关节组成的两个子部件拓扑图 $g_4^{v_i}=(\mathbf{V}_4^{v_i},\mathbf{E}_4^{v_i}),i=1,\ldots,N_4$, $N_4=2$, 根据公式 (7), 获得尺度 4 的图拓扑结构 $\mathbf{G}_4$.

单尺度特征提取模块: 对于不同尺度的骨架关节, 首先根据每一尺度所建立的拓扑图 $\mathbf{G}_k$ 构造相应的时空域邻接矩阵 $(\mathbf{A}_{\mathrm{ST}})_k\in\mathbb{R}^{N_k\times N_k}$, N_k 为尺度 k 的节点个数. 然后, 利用所设计的 DDA-STGConv 图卷积提取相应尺度的时空域特征. 对于尺度 k 的第 l 层输入特征 $\mathbf{X}_k^{(l)}$, 根据公式 (6), 其相应的时空域图卷积可以表示为:

$$\tilde{\mathbf{X}}_k^{(l+1)}=\sigma\left(\mathbf{W}_k^{(l)}\mathbf{X}_k^{(l)}\left((\mathbf{A}_{\mathrm{ST}})_k\odot\mathbf{M}_k+\mathbf{A}_{I,k}\right)\right) \tag{9}$$

其中, $\tilde{\mathbf{X}}_k^{(l+1)}$ 表示尺度 k 的第 l 层输入特征 $\mathbf{X}_k^{(l)}$ 经过 DDA-STGConv 卷积后的输出.

2.2.3 多尺度特征交叉融合模块 (MFEB)

为了更好地提取骨架关节之间的上下文信息, 本文构造多尺度特征交叉融合模块, 将平行多尺度子网络模块中不同尺度的特征进行融合, 如图 4 所示, 对于第 l 层网络而言, $\{\tilde{\mathbf{X}}_1^{(l+1)},\tilde{\mathbf{X}}_2^{(l+1)},\ldots,\tilde{\mathbf{X}}_k^{(l+1)}\},k\in 1,2,3,4$ 表示平行多尺度

子网络中不同尺度的特征, 经过 MFEB 模块融合后的特征表示为:

$$\mathbf{X}_k^{(l+1)} = P(\{\tilde{\mathbf{X}}_k^{(l+1)} : k = 1,2,3,4\}) \tag{10}$$

其中, $P(\cdot)$ 为多尺度特征融合函数, 其通常为求和函数或者级联函数. 不管是求和函数还是级联函数, 在多尺度特征融合中仅仅是对特征进行了聚合, 没有考虑平行子图网络中多尺度特征之间的信息交互. 根据文献 [13], 加强不同尺度特征的交互对于提升网络模型特征的表达能力非常重要, 因此, 本文对公式 (10) 进行如下改进, 提出多尺度特征交叉融合模型:

$$\mathbf{X}_k^{(l+1)} = \overset{K}{\underset{i=1}{C}}at(P(\tilde{\mathbf{X}}_i^{(l+1)},k) : k = 1, \ldots, K) \tag{11}$$

$$P(\tilde{\mathbf{X}}_i^{(l+1)},k) = \begin{cases} \dfrac{conv_{3\times3}(\tilde{\mathbf{X}}_i^{(l+1)},k) + \tilde{\mathbf{X}}_k^{(l+1)}}{2}, \mathbf{G}_i = (\mathbf{V}_i \cap \mathbf{V}_k, \mathbf{E}_i \cap \mathbf{E}_i), i < k & \text{(12a)} \\ \tilde{\mathbf{X}}_i^{(l+1)}, \mathbf{G}_i = (\mathbf{V}_i \cup \mathbf{V}_k, \mathbf{E}_i \cup \mathbf{E}_i), i > k & \text{(12b)} \end{cases}$$

其中, $P(\tilde{\mathbf{X}}_i^{(l+1)},k)$ 为特征尺度转换函数, 其分别采用上下采样函数, 将特征 $\tilde{\mathbf{X}}_i^{(l+1)}$ 从尺度 i 转换到尺度 k. 如图 4 所示, 不同尺度的特征经过 $P(\tilde{\mathbf{X}}_i^{(l+1)},k)$ 后, 将获得的多尺度特征进行级联 $Cat(\cdot)$, 然后进行 1×1 的卷积映射以对特征大小进行调整. 在此, 当 $i < k$ 时, $P(\cdot)$ 为下采样函数, 其首先对两个尺度的图拓扑关节点进行如公式 (12a) 的合并, 然后对 $\tilde{\mathbf{X}}_i^{(l+1)}$ 执行 3×3 卷积操作, 最后, 对合并的关节特征求平均获得尺度 i 中节点转换到尺度 k 中所对应的节点特征; 当 $i > k$ 时, $P(\cdot)$ 为上采样函数, 其首先根据公式 (12b) 合并两个尺度的图拓扑关节点, 然后执行最近邻上采样操作, 通过对 i 尺度子图网络中节点特征进行复制, 获得对应 k 尺度子图网络中相应的节点特征.

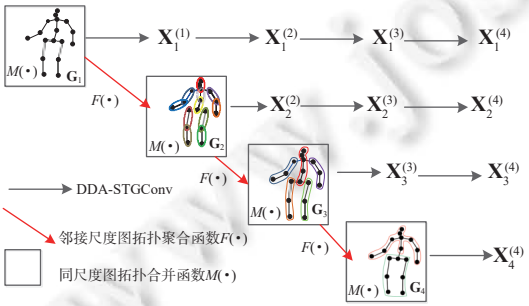

图 3 平行多尺度子网络模块 (PM-SubGNet)

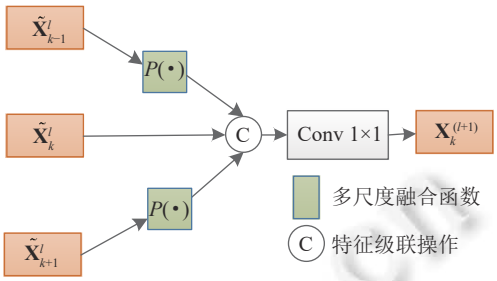

图 4 多尺度特征交叉融合模块 (MFEB)

2.3 整体网络结构

如前文图 1 所示, 本文所提出的平行多尺度时空图卷积网络模型 (PMST-GNet), 主要由以对角占优的时空注意力图卷积 (DDA-STGConv) 为基本单元的平行多尺度时空图卷积模块 (PM-SubGNet) 和多尺度特征交叉融合模块 (MFEB) 两部分堆叠组成. 首先, 将输入 2D 骨架关节信息进行预处理操作, 然后根据所设计的图拓扑聚合函数, 计算不同尺度的图拓扑结构, 以此构造平行多尺度子网络模块. 第 1 阶段的子网络模型由 4 个 DDA-STGConv 层组成, 每层包含 BN 层和 ReLU 层, 网络的第 2–4 阶段分别由 3, 2, 1 个 DDA-STGConv 层组成, 每层包含 BN 层和 ReLU 层. 每个 DDA-STGConv 层后都跟随一个 MFEB 模块, 实现多尺度特征的交叉融合. 最后, 使用第 1 阶段的高分别率特征进行 3D 关节点回归, 附加 1×1 全连接层, 调整输出维度, 进而预测 3D 关节点位置信息. 在此, 使用 ℓ_2 函数作为网络训练中的损失函数, 测量预测的 3D 关节点位置信息与真实值之间的误差.

3 实验与结果分析

为了测试本文所提出网络在 3D 人体姿态估计上的性能, 选用两个主流的 3D 数据集: Human3.6M[14]和 MPI-INF-3DHP[15]对所提出的网络性能进行分析. 首先, 为了验证本文网络中各模块对模型性能的影响, 在 Human3.6M 数据集上进行详细的消融实验. 然后, 将本文算法与其他主流的 3D 姿态估计算法进行比较分析.

3.1 数据集与评价指标

Human3.6M 数据集[14]是 3D 姿势估计中目前最大、使用最广泛的室内数据集, 包含由 11 个对象, 涉及 15 个日常动作 (如走路、打招呼、拍照、吸烟等) 组成的 360 万帧视频图像. 该数据集分为训练集 S1、S5、S6、S7、S8 和测试集 S9、S11 两部分, 其评价指标主要采用 MPJPE (mean per joint position error) 和 P-MPJPE (procrustes analysis MPJPE). MPJPE (Protocol #1) 是计算所有关节点的估计坐标与真实坐标之间欧氏距离的平均值, 单位为 mm, 其侧重于衡量误差结果的绝对性. P-MPJPE (Protocol #2) 是对关节点估计的结果进行刚性变换对齐后的 MPJPE, 其更侧重于衡量所估计姿态与真实姿态的误差和两者之间的相似性.

MPI-INF-3DHP[15]数据集是一个同时涉及室内及室外场景更大规模的数据集, 包含使用 14 个摄像头拍摄的 8 人, 涉及 8 个主题的 130 万帧视频图像. 该数据集的评价指标一般采用 MPJPE, 3D PCK (percentage of correct 3D keypoints) 和 AUC (area under curve), 其中, 3D PCK 衡量 3D 关节点的正确率, 当估计的关节点坐标与真实坐标之间的距离小于预定阈值 (通常阈值设置为 150 mm), 则认为估计的关节结果是正确的. AUC 为 ROC (receiver operating characteristic curve) 曲线下与坐标轴围成的面积, 其中, ROC 是以二分类方式 (分界值或决定阈), 以真阳性率为纵坐标, 假阳性率为横坐标绘制的曲线.

3.2 实验设置

本文实验在 Ubuntu 16.04, NVIDIA RTX 2080Ti GPU 上用 Python 3.7 在 PyTorch 平台上运行. 实验过程中, 网络使用批尺寸 (batch size) 为 256 的数据训练模型, 训练 120 代, 采用 Adam 对模型进行优化, 其初始学习率设置为 0.001, 收缩因子为 0.95, 遗忘因子为 0.05. 为了与其他主流算法进行公平对比, 本文按照文献 [5−7,9,16–18] 的方法采用 CPN (cascaded pyramid network)[19]对 Human3.6M 数据集进行 2D 姿态估计, 并将其作为网络的输入进行 3D 姿态估计; 同时, 遵循文献 [12,15,20–23] 的方法以 MPI-INF-3DHP 数据集的 2D 姿态真实值作为网络的输入.

3.3 实验结果分析

3.3.1 消融实验分析

本节通过消融实验验证所提出网络各模块对网络整体性能的影响, 所有消融实验都在 Human3.6M 数据集上执行, 采用 CPN 作为 2D 姿态估计检测的检测器, 并使用 Protocol #1 作为性能评价指标.

● 各模块性能分析: 为了分析网络各组成模块对模型性能的影响, 本文将 PMST-GNet 网络的每一部分从整体框架中剥离出来, 通过与未使用各模块的算法进行对比实验分析其对 3D 姿态估计模型的影响. 如表 1 所示, 使用 vanilla GCN 即文献 [8] 中提出的原始 GCN 代替 DDA-STGConv 作为主干网络的构成单元, 其 MPJPE 为 92.5 mm. 从表 1 中可以看出, 用本文所设计的 DDA-STGConv 为基本单元构建主干网络, 其 MPJPE 降低到 47.7 mm, 验证了 DDA-STGConv 卷积操作性能优于原始的 GCN 卷积, 这证明了从时间和空间维度出发对骨架关节点信息进行基于自约束和注意力机制约束的建模可以提升模型特征表示的能力, 从而增强网络的性能. 在此基础上, 进一步添加 MFEB 模块, 实现了平行子图网络之间多尺度信息的交互, 挖掘骨架关节之间的上下文信息, MPJPE 降低到 45.6 mm, 这一结果验证了加强不同尺度特征的交互有利于提升网络模型特征的表达能力. 如表 1 所示, 通过添加更多的模块到主干网络中, 本文所提出的网络模型性能得到稳步提高, 这进一步验证了本文网络设计的合理性.

表 1 网络各组成模块性能分析

方法	MPJPE (mm)
模型1 (vanilla GCN)	92.5
模型2 (DDA-STGConv)	47.7
模型3 (DDA-STGConv + MFEB)	45.6

对角占优的时空注意力图卷积 (DDA-STGConv) 对网络模型性能影响分析: 为了分析本文所设计的 DDA-STGConv 对网络性能的影响, 如表 2 所示, 使用 vanilla GCN 中的原始 GCN 代替 DDA-STGConv 模块, 其 PMST-GNet 模型中所有层使用固定的邻接矩阵 $\mathbf{A}$ 进行节点特征的聚合, 其 MPJPE 为 60.4 mm. 在此基础上, 本文构造时空域邻接

矩阵 $\mathbf{A}_{ST}$, 将连续 T 帧关节的连接关系融合到一个邻接矩阵中, 依靠 2D 卷积提取相应的时空特征, 其所构造的模型 2 获得 49.1 mm 的 MPJPE, 相比模型 1, MPJPE 下降了 11.3 mm. 在模型 2 的基础上, 进一步对 $\mathbf{A}_{ST}$ 进行改进, 基于注意力机制自适应的提取不同节点之间的动态关联性, 在节点特征聚合中利用所计算的权值矩阵 $\mathbf{M}_{att}$ 衡量不同节点对中心节点的作用, 其 MPJPE 下降到 46.5 mm. 此外, 由于不同的动作姿态产生主要依赖于关节点自身, 如“拍手, 打电话”等动作主要是由上肢胳膊位置关节点决定的, 因此, 关节点自身在特征聚合中具有重要的作用, 所以, 为了验证关节点自身对特征聚合的影响, 仅以 $\mathbf{A}_I$ 为邻接矩阵进行节点特征聚合, 其模型 4 获得 52.3 mm 的 MPJPE. 因此, 本文将模型 3 和模型 4 进行融合获得模型 5, 其不仅依靠 $\mathbf{A}_{ST}\odot\mathbf{M}_{att}$ 自适应的提取不同节点之间的动态关联性, 而且通过 $\mathbf{A}_I$ 突出关节点自身在特征聚合中的作用, 增强节点间的信息交互, 有效地对骨架关节的时空信息进行建模.

表 2　时空注意力图卷积对网络性能的影响

方法	MPJPE (mm)
模型1 (vanilla GCN)	60.4
模型2 ($\mathbf{A}_{ST}$)	49.1
模型3 ($\mathbf{A}_{ST}\odot\mathbf{M}_{att}$)	46.5
模型4 ($\mathbf{A}_I$)	52.3
模型5 ($\mathbf{A}_{ST}\odot\mathbf{M}_{att}+\mathbf{A}_I$)	45.6

平行多尺度子网络模型 (PM-SubGNet) 对网络性能的影响分析: 为了验证多尺度网络对模型 3D 姿态估计性能的影响, 本文使用文献 [7,8] 中提出的 vanilla GCN 和 SemGConv 代替 DDA-STGConv 作为主干网络的构成单元构造尺度为 1 的串行图卷积模型, 其实验结果如表 3 所示, 同时, 表 3 给出了以 DDA-STGConv 为主干网络构成单元构造不同尺度网络模型的性能对比结果, 其根据人体运动中关节点呈现的相似运动趋进行多尺度划分, 通过图拓扑聚合函数, 构造平行多尺度子网络模块 (PM-SubGNet), 然后利用多尺度特征交叉融合模块 (MFEB) 进行多尺度特征的融合. 因此, 去除 PMST-GNet 网络模型中的所有多尺度特征交叉融合模块, 仅以 k=1 尺度的关节点构造图拓扑结构 G_1. 通过 DDA-STGConv 卷积提取节点特征信息, 所构造的模型仅包含 k=1 尺度的特征信息, 其获得模型 3 的 MPJPE 为 47.78 mm, 模型 3 的 MPJPE 值优于模型 1 和模型 2. 在此基础上, 逐步引入 k=2 及 k=3 关节点构造的图拓扑结构, 同样通过 DDA-STGConv 卷积提取节点特征信息, 并利用 MFEB 模块对多尺度特征进行融合, 相应的网络模型分别表示为模型 4 和模型 5, 从表 3 可以看出, 模型 5 的 MPJPE 值优于模型 2 和模型 3, 达到了 45.91 mm. 显然, 组合两个尺度的特征模型比仅采用一个尺度的特征模型具有更好的性能, 这是因为 k=1 尺度缺乏对不同距离邻域关节点特征的描述. 通过对多尺度特征进行融合, 能对不同距离的邻域节点特征进行细粒度的刻画, 有助于网络性能的提升. 此外, 在模型 5 的基础上, 继续添加 k=4 (左臂、右臂、左腿、右腿和躯干) 和 k=5 (上半身和下半身) 的关节信息, 其实验结果如表 3 所示, 当将 k=4 和 k=5 尺度的关节点引入到网络中时, 其 MPJPE 降低的幅度小于引入 k=2 和 k=3 尺度的节点信息. 这是因为关节点的运动更多的受具有物理连接关节的影响, 即 k=1、k=2 和 k=3 尺度的邻域关节对运动的影响大于 k=4 和 k=5 尺度的邻域关节点. 此外, 如表 3 所示, 模型 1–3 为串行网络模型, 模型 4–7 为平行多尺度网络模型, 其模型 1–3 的 MPJPE 值均高于模型 4–7, 进一步证明, 平行时空图卷积网络模型优于串行网络模型, 多尺度特征提取有助于节点特征由粗到细的刻画, 从而提高平行网络模型的性能.

3.3.2　主流方法对比分析

● Human3.6M 数据集上的实验结果分析: 为了进一步验证本文所提出网络模型的有效性, 将所提出的算法与近年来主流的 3D 姿态估计算法在 Human3.6M 测试集上进行单目姿态估计对比. 表 4 给出了本文算法与其他主流算法[5–7,9,12,16–18,23–32]在两种不同评价标准下的对比结果. 如表 4 所示, 在输入 T=81 的情况下, 本文算法在 Protocol #1 下的平均 MPJPE 为 45.6 mm, 在 Protocol #2 下的平均 P-MPJPE 为 35.2 mm, 其性能优于具有相同接受域的大多数主流算法. 与基于时空卷积 GCN 的算法相比, 如文献 [7] 采用 vanilla GConv, 文献 [5] 采用 SemGConv

分别对 3D 人体姿态估计中的关节点进行图卷积操作, 本文所设计的网络获得更小的 MPJPE 值 (45.6 mm vs. 48.8 mm[7], 45.6 mm vs. 60.8 mm[5]). 尽管文献 [5,7] 中的模型根据人体骨架的结构信息设计了特殊的约束条件以满足 3D 人体姿态估计, 但其性能仍低于本文算法. 此外, 将本文算法与其他具有特殊涉及结构的基于 GCN 的 3D 人体姿态估计算法进行比较, 如文献 [33] 中提出的 UGCN、文献 [17] 中构造的 HGN、文献 [9] 中设计的 HPGCN 和文献 [24] 中设计的 GraphSH 网络, 本文算法展示出较好的性能. 进一步验证了本文网络设计的合理性. 在本文所提出的 PMST-GNet 网络中, 其不仅通过设计 DDA-STGConv 卷积层, 从时间和空间维度出发对骨架关节点信息进行基于自约束和注意力机制约束的建模, 增强节点间的信息交互, 有效地对骨架关节的时空信息进行建模. 而且, 以 DDA-STGConv 为基本单元, 通过设计图拓扑聚合函数, 构造不同尺度的图拓扑结构以建立平行多尺度子网络模块, 实现不同尺度节点特征的聚合, 有效提取骨架关节点局部和全局的特征信息. 此外, 构造 MFEB 模块, 实现平行子图网络之间多尺度信息的交互, 进一步提高网络特征表达的能力. 因此, 如表 4 所示, 本文所设计的网络表现出较好的 3D 姿态估计性能. 但是, 本文方法的 MPJPE 和 P-MPJPE 均大于文献 [28] 中的值, 这是因为本文方法是一种基于单视角单假设的 3D 姿态估计, 文献 [28] 通过构建多假设生成模型, 提取关键特征的依赖关系, 消除深度模糊和遮挡对姿态估计的影响. 这一现象激励本文在未来的工作中进一步在时空特征提取中引入多假设模型, 提高模型姿态估计的性能. 此外, 文献 [29] 是一种基于纯 Transformer 进行关节时空特征提取的网络模型, 其相比本文模型来说, 能更好利用 Transformer 机制捕捉关节时空域的全局上下文信息, 故而文献 [29] 的实验结果优于本文方法, 但是会带来模型复杂度的增加.

表 3　平行多尺度模型对网络性能的影响

尺度k	关节点数目 J					MPJPE (mm)
	17	11	7	5	2	
模型1 (vanilla GCN)	√	—	—	—	—	92.5
模型2 (SemGConv)	√	—	—	—	—	108.6
模型3 (k=1)	√	—	—	—	—	47.78
模型4 (k=1, 2)	√	√	—	—	—	46.63
模型5 (k=1, 2, 3)	√	√	√	—	—	45.91
模型6 (k=1, 2, 3, 4)	√	√	√	√	—	45.66
模型7 (k=1, 2, 3, 4, 5)	√	√	√	√	√	45.62

表 4　单视角模型在 Human3.6M 测试集上基于 Protocol #1 和 Protocol #2 评价指标对比结果 (mm)

Protocol	模型	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
#1	Martinez等人[25]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.5	62.3	59.1	65.1	49.4	52.4	62.9
	Zhao等人[5] (T=1)	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
	Zou等人[16]	49.0	54.5	52.3	53.6	59.2	71.6	49.6	49.8	66.0	75.5	55.1	53.8	58.5	40.9	45.4	55.6
	HPGCN[9]	50.6	57.3	50.1	56.0	59.3	68.4	53.8	54.3	65.3	69.1	56.6	54.5	63.2	50.5	46.2	57.3
	Liu等人[26]	46.3	52.2	47.3	50.7	55.5	67.1	49.2	46.0	60.4	71.1	51.1	50.1	54.5	40.3	43.7	52.4
	GraphSH [24] (T=64)	45.2	49.9	47.5	50.9	54.9	66.1	48.5	46.3	59.7	71.5	51.4	48.6	53.9	39.9	44.1	51.9
	HGN[17]	47.8	52.5	47.7	50.5	53.9	60.7	49.5	49.4	60.0	66.3	51.8	48.8	55.2	40.5	42.6	51.8
	Cai等人[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
	Zeng等人[27] (T=1)	43.1	50.4	43.9	45.3	46.1	57.0	46.3	47.6	56.3	61.5	47.7	47.4	53.5	35.4	37.3	47.9
	VideoPose 3D[12] (T=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
	Chen等人[18]	41.1	44.2	44.9	45.9	46.5	39.3	41.6	54.8	73.2	46.2	48.7	42.1	35.8	46.6	38.5	46.3
	GAST-Net[6] (T=81)	44.3	44.8	41.9	45.2	47.4	54.7	43.6	43.1	56.9	61.0	47.6	43.5	47.1	35.6	34.5	46.1
	PoseFormer[23]* (T=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
	MHFormer[28]* (T=81)	39.2	43.1	40.1	40.9	44.9	51.2	40.6	41.3	53.5	60.3	43.7	41.1	43.8	29.8	30.6	43.0

表 4 单视角模型在 Human3.6M 测试集上基于 Protocol #1 和 Protocol #2 评价指标对比结果 (mm)(续)

Protocol	模型	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
#1	P-STMO[29]* (*T*=81)	41.7	44.5	41.0	42.9	46.0	51.3	42.8	41.3	54.9	61.8	45.1	42.8	43.8	30.8	30.7	44.1
	Bai等人[30]	44.4	49.7	48.2	49.2	56.9	66.3	47.7	46.4	65.6	71.2	52.7	49.6	53.6	39.4	43.3	52.1
	Li等人[31]	47.9	50.0	47.1	51.3	51.2	59.5	48.7	46.9	56.0	61.9	51.1	48.9	54.3	40.0	42.9	50.5
	Han等人[32]	41.0	42.4	49.6	47.4	46.7	44.4	41.5	50.7	65.1	49.2	53.2	41.8	35.9	47.9	40.6	46.9
	Ours (*T*=81)	41.2	45.5	42.1	44.6	47.3	57.2	43.2	43.5	57.3	63.1	46.3	44.1	45.6	31.2	32.1	45.6
#2	Martinez等人[25]	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	69.4	49.2	45.0	49.5	38.0	43.1	47.7
	Zou等人[16]	38.6	42.8	41.8	43.4	44.6	52.9	37.5	38.6	53.3	60.0	44.4	40.9	46.9	32.2	37.9	43.7
	HPGCN[9]	37.5	41.1	38.2	42.2	40.9	46.9	39.0	37.9	48.6	53.1	42.1	38.3	44.3	37.7	32.9	41.6
	Liu等人[26]	35.9	40.0	38.0	41.5	42.5	51.4	37.8	36.0	48.6	56.6	41.8	38.3	42.7	31.7	36.2	41.2
	GraphSH[24] (*T*=64)	33.9	37.2	36.8	38.1	43.5	43.5	37.8	35.0	47.2	53.8	40.7	38.3	41.8	30.1	31.4	39.0
	HGN[17]	35.8	39.7	36.3	40.6	40.2	45.9	36.8	35.8	47.3	53.7	40.7	36.4	43.1	29.8	32.8	39.6
	Cai等人[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
	VideoPose 3D[12] (*T*=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
	Chen等人[18]	36.9	39.3	40.5	41.2	42.0	34.9	38.0	51.2	67.5	42.1	42.5	37.5	30.6	40.2	34.2	41.6
	GAST-Net[6] (*T*=81)	33.2	35.9	34.1	37.3	37.1	42.5	33.3	32.4	45.7	48.5	40.1	33.3	36.5	32.0	29.1	36.7
	P-STMO[29]* (*T*=81)	31.3	35.2	32.9	33.9	35.4	39.3	32.5	31.5	44.6	48.2	36.3	32.9	34.4	23.8	23.9	34.4
	Bai等人[30]	33.5	37.3	37.4	39.3	40.8	46.2	34.0	33.7	50.5	56.8	39.2	36.8	40.3	29.8	33.6	39.3
	Han等人[32]	50.2	54.4	50.1	52.0	53.6	46.7	50.1	61.5	65.1	52.2	57.1	49.2	41.2	54.5	46.6	52.8
	Ours (*T*=81)	31.1	34.8	33.6	34.8	37.2	42.5	33.1	33.7	44.5	49.9	35.9	31.8	35.9	23.7	25.9	35.2

注: *代表基于纯Transformer模型的方法

此外, 为了进一步分析本文所提出算法姿态估计的性能, 图 5 对比了本文算法与其他 6 种主流的人体姿态估计算法在 Human3.6M 测试集上的 MPJPE 分布情况. 以 55 mm MPJPE 值为阈值, 如果姿态估计的误差大于此阈值, 将会造成大的精度损失. 从图 5 可以看出, 本文算法的高误差比例较少, 其中, 本文算法姿态估计的 MPJPE 大多分布在 40–55 mm 之间, MPJPE 分布小于 40 mm 的 MPJPE 比例明显高于相应的对比算法, 如文献 [5−7,12,17,24]. 这一结果进一步证明本文算法具有较好的 3D 姿态估计性能.

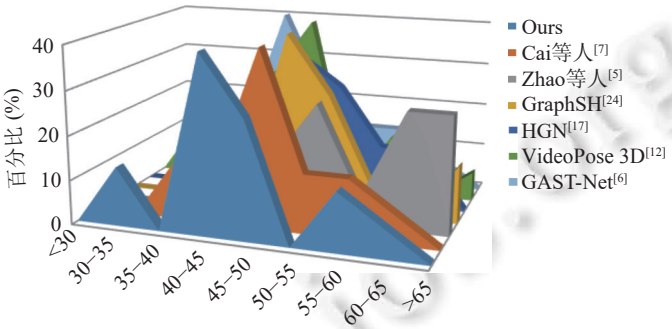

图 5 Human3.6M 测试集上 MPJPE 指标分布情况

● MPI-INF-3DHP 数据集上的实验结果分析: 为了进一步评估本文所提出网络的性能, 将所提出算法与 6 种主流的人体姿态估计算法在 MPI-INF-3DHP 数据集上进行对比. 文中采用 PCK、AUC 和 MPJPE 作为评估指标, 并将 2D 关节点的真实值作为输入, 如表 5 所示, 本文算法获得 88.9% 的 PCK、56.6% 的 AUC 和 73.2 mm 的 MPJPE. 虽然, 本文所提出的网络模型仅在 Human3.6M 数据集上进行训练, 并没有在 MPI-INF-3DHP 数据集上进行重新训练或微调, 但与其他 SOTA 方法相比, 所提出的方法仍然在 3 个评估指标中取得了较好的结果. 这些结果说明本文所提出的模型对于训练数据中没有出现过的动作依然具有较好的预测效果, 进一步证明本文所设计的网络具有良好的泛化能力, 在多个场景中具有较好的性能.

表 5 在 MPI-INF-3DHP 数据集上的对比结果

方法	PCK (%)↑	AUC (%)↑	MPJPE (mm)↓
Mehta等人[15] (T=1)	75.7	39.3	117.6
Chen等人[20] (T=243)	87.8	53.8	79.1
Pavllo等人[12] (T=243)	85.5	51.5	84.8
Lin等人[21] (T=25)	83.6	51.4	79.8
Li等人[22] (T=1)	81.2	46.1	99.7
Zheng等人[23] (T=9)	88.6	56.4	77.1
Li等人[31]	84.1	53.7	—
MHFormer[28]	93.8	63.3	58.0
Ours (T=81)	88.9	56.6	73.2

3.4 可视化结果

为了更直观地描述本文算法姿态估计的效果, 图 6 展示了本文算法在 Human3.6M 测试集 S9 和 S11 上与主流算法的部分 3D 姿态估计可视化结果. 图 6 每一列分别为来自 S9 和 S11 序列的 RGB 图片, 各对比方法 3D 姿态估计结果以及相应的真实 3D 姿态. 从图 6 展示的结果可以看出, 本文算法对于遮挡较为严重的动作, 如 SittingDown、Posing、Sitting 等动作, 依然能准确地估计 3D 姿态, 图中红色圆圈标记了本文方法与对比方法姿态估计差异较大的位置.

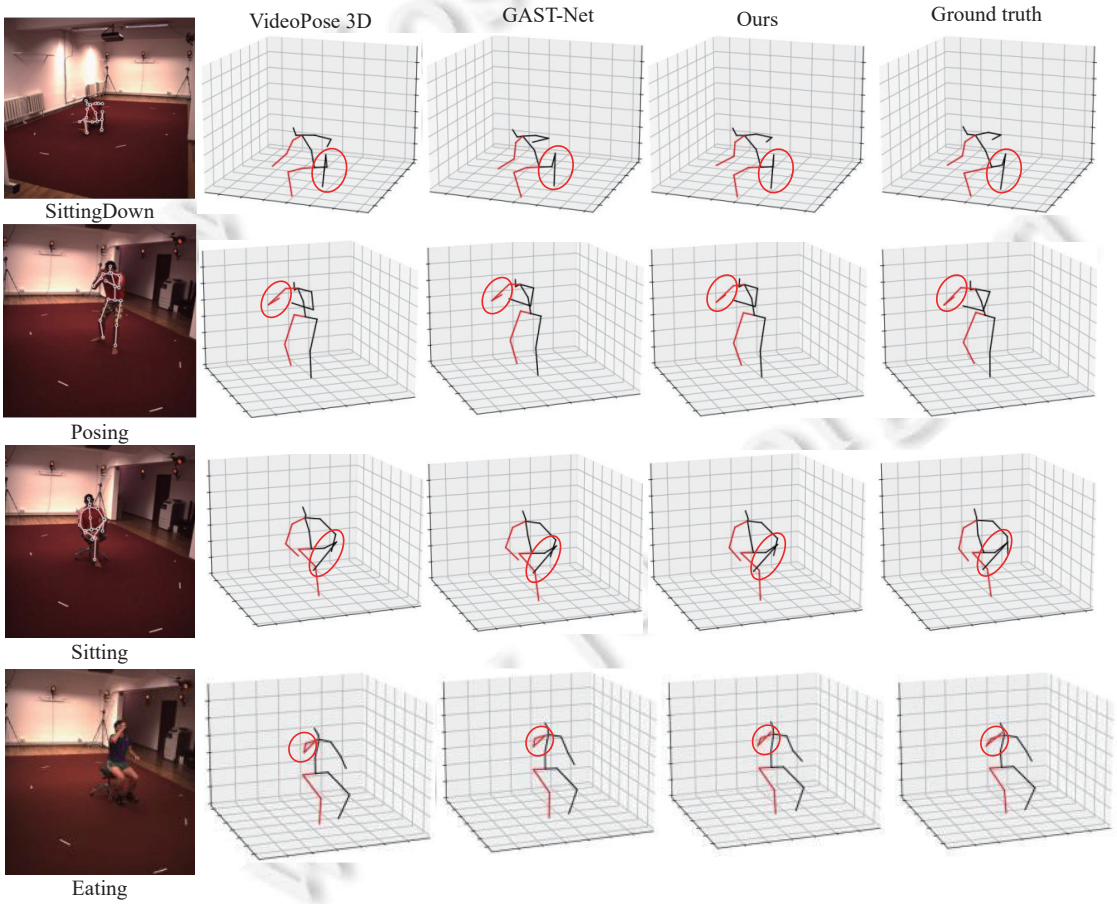

图 6 本文方法与对比方法在 Human3.6M 测试集上的可视化结果

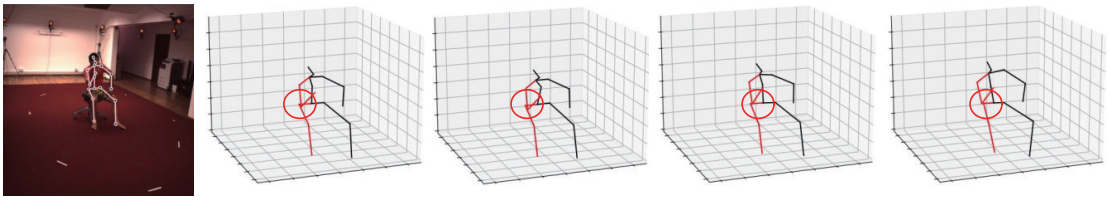

图 6　本文方法与对比方法在 Human3.6M 测试集上的可视化结果 (续)

3.5　本文方法的局限性及失败实验结果分析

虽然本文所提出的 PMST-GNet 在三维人体姿态估计中获得较好的实验效果, 但是, 本文所设计的网络模型依然遵循两阶段的 2D-3D 姿态估计的框架, 其性能高度依赖于 2D 姿态估计器的准确性, 其次, 本文所提出的模型是一种基于单目图像单视角的 2D-3D 姿态估计模型, 其不能很好地解决姿态估计中遮挡问题和深度模糊性问题造成的影响, 因此, 如何引入多假设估计理论和扩散模型消除上述 3 个因素对基于单目图像的 2D-3D 姿态估计的影响是本文下一步工作的重点[34]. 在此, 为了更好地说明 2D 姿态估计器产生的误差、遮挡问题及深度模糊性问题对本文实验的影响, 以 Human3.6M 中 SittingDown, Photo 和 WalkDog 为例, 列举了一些本文模型失败的案例, 如图 7 所示, 其中, 图 7 中用绿线突出显示上述 3 种原因.

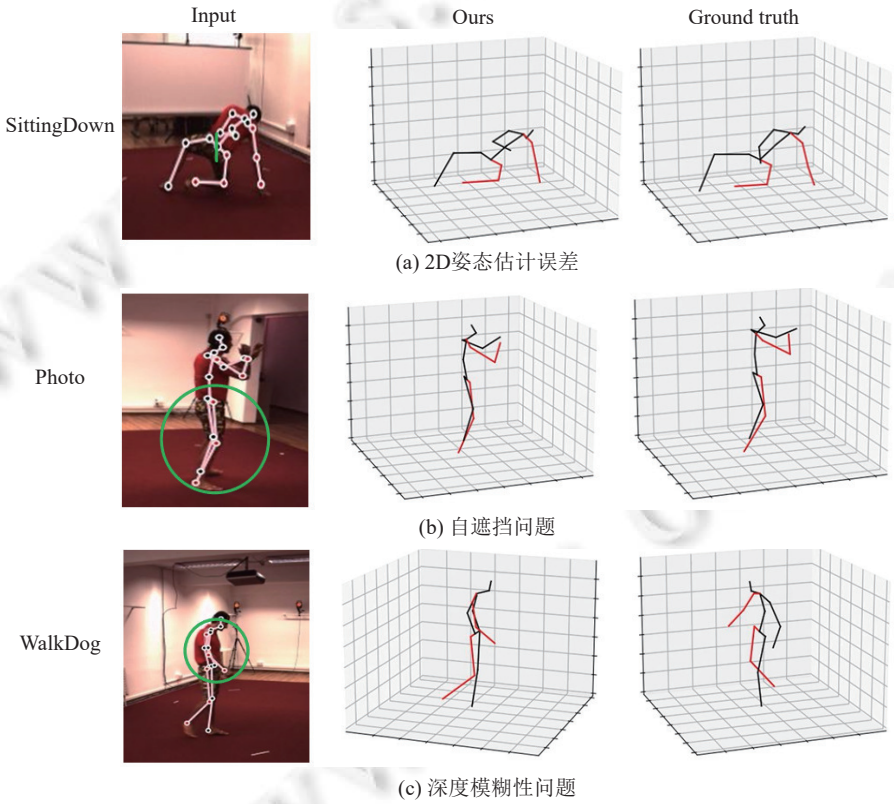

图 7　本文模型在 Human3.6M 上的失败示例

4　结　论

本文提出了一种基于平行时空多尺度卷积网络模型 (PMST-GNet) 的三维人体姿态估计算法. PMST-GNet 模

型构造了一个以 DDA-STGConv 为基本单元的平行多尺度网络, 通过构造跨域邻接矩阵同时捕捉空域和时域的动态判别特征, 利用注意力机制探索节点之间的关联关系, 增强节点间的信息交互. 此外, 通过设计图拓扑聚合函数, 构造基于不同图拓扑结构的平行多尺度子网络模块, 实现不同尺度节点特征的聚合, 有效提取骨架关节点局部和全局的特征信息. 同时, 设计多尺度特征交叉融合模块 (MFEB), 加强平行网络间多尺度信息的交互, 增强网络模型特征表示的能力. 在两大人体姿态估计数据集上与目前主流的方法进行对比, 实验结果表明所提出的网络模型获得较好的姿态估计结果. PMST-GNet 模型的灵活性可为行为识别, 运动预测及场景理解等领域的工作提供技术支持.

References:

[1] Zhang Y, Wen GZ, Mi SY, Zhang ML, Geng X. Overview on 2D human pose estimation based on deep learning. Ruan Jian Xue Bao/Journal of Software, 2022, 33(11): 4173–4191 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6390.htm [doi: 10.13328/j.cnki.jos.006390]

[2] Ding J, Shu XB, Huang P, Yao YZ, Song Y. Multimodal and multi-granularity graph convolutional networks for elderly daily activity recognition. Ruan Jian Xue Bao/Journal of Software, 2023, 34(5): 2350–2364 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6439.htm [doi: 10.13328/j.cnki.jos.006439]

[3] Yang HH, Liu HX, Zhang YM, Wu XJ. FMR-GNet: Forward mix-hop spatial-temporal residual graph network for 3D pose estimation. Chinese Journal of Electronics, 2024, 33(6): 1–14.

[4] Moon G, Lee KM. I2L-MeshNet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image. In: Proc. of the 16th European Conf. on Computer Vision (ECCV). Glasgow: Springer, 2020. 752–768. [doi: 10.1007/978-3-030-58571-6_44]

[5] Zhao L, Peng X, Tian Y, Kapadia M, Metaxas DN. Semantic graph convolutional networks for 3D human pose regression. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 3420–3430. [doi: 10.1109/CVPR.2019.00354]

[6] Liu JF, Rojas J, Li YH, Liang ZJ, Guan YS, Xi N, Zhu HF. A graph attention spatio-temporal convolutional network for 3D human pose estimation in video. In: Proc. of the 2021 IEEE Int'l Conf. on Robotics and Automation. Xi'an: IEEE, 2021. 3374–3380. [doi: 10.1109/ICRA48506.2021.9561605]

[7] Cai YJ, Ge LH, Liu J, Cai JF, Cham TJ, Yuan JS, Thalmann NM. Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Seoul: IEEE, 2019. 2272–2281. [doi: 10.1109/ICCV.2019.00236]

[8] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017.

[9] Wu YP, Kong DH, Wang SF, Li JH, Yin BC. HPGCN: Hierarchical poselet-guided graph convolutional network for 3D pose estimation. Neurocomputing, 2022, 487: 243–256. [doi: 10.1016/j.neucom.2021.11.007]

[10] Wang WG, Shen JB, Jia YD. Review of visual attention detection. Ruan Jian Xue Bao/Journal of Software, 2019, 30(2): 416–439 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5636.htm [doi: 10.13328/j.cnki.jos.005636]

[11] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

[12] Pavllo D, Feichtenhofer C, Grangier D, Auli M. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 7745–7754. [doi: 10.1109/CVPR.2019.00794]

[13] Sun K, Xiao B, Liu D, Wang JD. Deep high-resolution representation learning for human pose estimation. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5686–5696. [doi: 10.1109/CVPR.2019.00584]

[14] Ionescu C, Papava D, Olaru V, Sminchisescu C. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2014, 36(7): 1325–1339. [doi: 10.1109/TPAMI.2013.248]

[15] Mehta D, Rhodin H, Casas D, Fua P, Sotnychenko O, Xu WP, Theobalt C. Monocular 3D human pose estimation in the wild using improved CNN supervision. In: Proc. of the 2017 Int'l Conf. on 3D Vision. Qingdao: IEEE, 2017. 506–516. [doi: 10.1109/3DV.2017.00064]

[16] Zou ZM, Liu KK, Wang L, Tang W. High-order graph convolutional networks for 3D human pose estimation. In: Proc. of the 31st British Machine Vision Conf. BMVA Press, 2020.

[17] Li H, Shi BW, Dai WR, Chen YB, Wang BT, Sun Y, Guo M, Li CL, Zou JN, Xiong HK. Hierarchical graph networks for 3D human pose estimation. In: Proc. of the 32nd British Machine Vision Conf. (BMVC). BMVA Press, 2021. 387.

[18] Chen XP, Lin KY, Liu WT, Qian C, Lin L. Weakly-supervised discovery of geometry-aware representation for 3D human pose estimation. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 10887–10896. [doi: 10.1109/CVPR.2019.01115]

[19] Chen YL, Wang ZC, Peng YX, Zhang ZQ, Yu G, Sun J. Cascaded pyramid network for multi-person pose estimation. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7103–7112. [doi: 10.1109/CVPR.2018.00742]

[20] Chen TL, Fang C, Shen XH, Zhu YH, Chen ZL, Luo JB. Anatomy-aware 3D human pose estimation with bone-based pose decomposition. IEEE Trans. on Circuits and Systems for Video Technology, 2022, 32(1): 198–209. [doi: 10.1109/TCSVT.2021.3057267]

[21] Lin JH, Lee GH. Trajectory space factorization for deep video-based 3D human pose estimation. In: Proc. of the 30th British Machine Vision Conf. Cardiff: BMVA Press, 2019. 101.

[22] Li SC, Ke L, Pratama K, Tai YW, Tang CK, Cheng KT. Cascaded deep monocular 3D human pose estimation with evolutionary training data. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 6172–6182. [doi: 10.1109/CVPR42600.2020.00621]

[23] Zheng C, Zhu SJ, Mendieta M, Yang TJN, Chen C, Ding ZM. 3D Human pose estimation with spatial and temporal Transformers. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Montreal: IEEE, 2021. 11636–11645. [doi: 10.1109/ICCV48922.2021.01145]

[24] Xu TH, Takano W. Graph stacked hourglass networks for 3D human pose estimation. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 16100–16109. [doi: 10.1109/CVPR46437.2021.01584]

[25] Martinez J, Hossain R, Romero J, Little JJ. A simple yet effective baseline for 3D human pose estimation. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision (ICCV). Venice: IEEE, 2017. 2659–2668. [doi: 10.1109/ICCV.2017.288]

[26] Liu KK, Ding RQ, Zou ZM, Wang L, Tang W. A comprehensive study of weight sharing in graph networks for 3D human pose estimation. In: Proc. of the 16th European Conf. on Computer Vision 2020 (ECCV). Glasgow: Springer, 2020. 318–334. [doi: 10.1007/978-3-030-58607-2_19]

[27] Zeng AL, Sun X, Yang L, Zhao NX, Liu MH, Xu Q. Learning skeletal graph neural networks for hard 3D pose estimation. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Montreal: IEEE, 2021. 11416–11425. [doi: 10.1109/ICCV48922.2021.01124]

[28] Li WH, Liu H, Tang H, Wang PC, van Gool L. MHFormer: Multi-hypothesis transformer for 3D human pose estimation. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022. 13137–13146. [doi: 10.1109/CVPR52688.2022.01280]

[29] Shan WK, Liu ZH, Zhang XF, Wang SS, Ma SW, Gao W. P-STMO: Pre-trained spatial temporal many-to-one model for 3D human pose estimation. In: Proc. of the 17th European Conf. on Computer Vision (ECCV). Tel Aviv: Springer, 2022. 461–478. [doi: 10.1007/978-3-031-20065-6_27]

[30] Bai GH, Luo YM, Pan XL, Wang J, Guo JM. Real-time 3D human pose estimation without skeletal a priori structures. Image and Vision Computing, 2023, 132: 104649. [doi: 10.1016/j.imavis.2023.104649]

[31] Li H, Shi BW, Dai WR, Zheng HW, Wang BT, Sun Y, Guo M, Li CL, Zou JN, Xiong HK. Pose-oriented Transformer with uncertainty-guided refinement for 2D-to-3D human pose estimation. In: Proc. of the 37th AAAI Conf. on Artificial Intelligence. Washington: AAAI, 2023. 1296–1304. [doi: 10.1609/aaai.v37i1.25213]

[32] Han CC, Yu X, Gao CX, Sang N, Yang Y. Single image based 3D human pose estimation via uncertainty learning. Pattern Recognition, 2022, 132: 108934. [doi: 10.1016/j.patcog.2022.108934]

[33] Wang JB, Yan SJ, Xiong YJ, Lin DH. Motion guided 3D pose estimation from videos. In: Proc. of the 16th European Conf. on Computer Vision (ECCV). Glasgow: Springer, 2020. 764–780. [doi: 10.1007/978-3-030-58601-0_45]

[34] Yang HH, Liu HX, Zhang YM, Wu XJ. Hierarchical parallel multi-scale graph network for 3D human pose estimation. Applied Soft Computing, 2023, 140: 110267 [doi: 10.1016/j.asoc.2023.110267]

附中文参考文献:

[1] 张宇, 温光照, 米思娅, 张敏灵, 耿新. 基于深度学习的二维人体姿态估计综述. 软件学报, 2022, 33(11): 4173–4191. http://www.jos.

org.cn/1000-9825/6390.htm [doi: 10.13328/j.cnki.jos.006390]
[2] 丁静, 舒祥波, 黄捧, 姚亚洲, 宋砚. 基于多模态多粒度图卷积网络的老年人日常行为识别. 软件学报, 2023, 34(5): 2350–2364. http://www.jos.org.cn/1000-9825/6439.htm [doi: 10.13328/j.cnki.jos.006439]
[10] 王文冠, 沈建冰, 贾云得. 视觉注意力检测综述. 软件学报, 2019, 30(2): 416–439. http://www.jos.org.cn/1000-9825/5636.htm [doi: 10.13328/j.cnki.jos.005636]

杨红红(1988－), 女, 博士, 副研究员, 主要研究领域为人工智能, 深度学习, 计算机视觉.

张玉梅(1977－), 女, 博士, 教授, 主要研究领域为信号处理与分析.

刘泓希(2001－), 女, 硕士生, 主要研究领域为知识工程, 智能教学系统.

吴晓军(1970－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为模式识别, 智能系统, 复杂系统.