Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2023, 12(2), 143-153
Published Online May 2023 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2023.122017

基于图注意力的单目3D人体姿态估计

朱志玮

华中师范大学，物理科学与技术学院，湖北 武汉

收稿日期：2023年4月24日；录用日期：2023年5月24日；发布日期：2023年5月31日

摘　要

人体姿态估计是计算机视觉领域的重要研究方向，如何抑制复杂背景、光照变化和遮挡等因素干扰，提高3D人体姿态的准确性和鲁棒性目前仍然是一个很大挑战。本文提出了一种基于深度学习的三维人体姿态估计算法，该算法充分利用人体骨骼节点的连接关系和对称关系构建了一种图注意力时间卷积网络，该网络可以充分利用单目视频中的时空信息，解读人体姿态随时间的变化。实验表明该算法在Human3.6M数据集上比传统方法预测准确率提高了约14.9%。

关键词

深度学习，图卷积，时序卷积，三维人体姿态估计

Graph Attention Based Monocular 3D Human Pose Estimation

Zhiwei Zhu

College of Physical Science and Technology, Central China Normal University, Wuhan Hubei

Received: Apr. 24th, 2023; accepted: May 24th, 2023; published: May 31st, 2023

Abstract

Human pose estimation is an important research area in computer vision. It remains a big challenge to perform 3D human pose estimation with high accuracy and robustness under the interference of such factors as complex background, lighting changes and occlusion. The paper proposed a deep learning-based 3D human pose estimation algorithm featuring a graph attention temporal convolutional network, which is constructed utilizing the connectivity and symmetry of human skeletal nodes. The network is capable to make full use of spatiotemporal information in monocular videos to interpret changes in human pose over time. Experiments show that the proposed algorithm improves prediction accuracy by about 14.9% compared to traditional methods on the Human3.6M dataset.

文章引用: 朱志玮. 基于图注意力的单目 3D 人体姿态估计[J]. 人工智能与机器人研究, 2023, 12(2): 143-153.
DOI: 10.12677/airr.2023.122017

Keywords

Deep Learning, Graph Convolution, Temporal Convolution, 3D Human Pose Estimation

1. 引言

2D 人体姿态估计主要关注人体关节在二维平面上的位置和连线关系，而 3D 人体姿态估计则是在三维空间中确定人体关节的位置和姿态。目前 2D 人体姿态估计相关研究已趋向成熟，其自底而上的多人姿态实时检测在精确性、实时性和鲁棒性上都达到非常高的水准。相比之下 3D 人体姿态估计面临着更多的挑战，因为单帧图像仅仅是三维物体的二维投影，自身缺乏深度信息。相关研究[1]总结提出了 2D 姿态预测 3D 姿态任务存在的两个最大的问题：

(1) 不适定问题(Ill-Posed Problem)，即一个二维姿态可能对应不同的三维姿态。

(2) 病态问题(Ill-Conditioned Problem)，即对异常检测点敏感，一个错误的二维检测点会极大影响三维检测准确度。

另一方面，深度学习算法依赖于大量的训练数据，而现有的 3D 姿态数据集非常少，且大都是依靠适合室内环境的动作捕捉(MOCAP)系统构建的，系统需要带有多个传感器和紧身衣裤的复杂装置，在室外环境使用是非常不方便的，因此目前的主流数据集基本都是在实验室环境下采集的，这势必会影响到算法在户外数据上的泛化性能。

为了解决不适定问题，本文引入扩张时序卷积模块来拟合时间信息。扩张时序卷积可以通过对输入 2D 姿态信息的跨步、距离、范围、关联等方面考虑，提高对姿态信息的抽取和重构能力，从而提高模型对不同三维姿态的区分能力。

为了解决病态问题，本文设计了一个嵌入连接矩阵和对称矩阵的图卷积注意力机制模块来拟合空间信息。该模块利用人体骨骼点之间的连接和对称关系构建邻接矩阵，图卷积网络通过局部聚合操作，将每个节点的邻居节点的信息进行聚合，并利用这些节点之间的关系进行信息的传递，以增加人体结构信息，减少误差累积。

基于以上两个模块，本文构建了一种基于深度学习的三维人体姿态估计算法：图注意力时空卷积网络(GA-TCN)，该算法灵活地结合了时空信息，相较其他解决方案有着更好的性能表现。

2. 相关工作

目前三维人体姿态估计的主要研究方向有两种：

(1) 基于回归的三维人体姿态估计：

Newell [2]提出了一个全新网络结构，称为“Stacked Hourglass Networks”。该网络由多个 hourglass 模块组成，每个 hourglass 模块由多个卷积层和池化层交替组成，以捕捉不同尺度的特征。此外，该方法还使用了一种自适应权重回归策略，可以有效地缓解姿态重心点偏移的问题。Chen [3]等人提出了一种基于层次分解的人体姿态估计方法。该方法通过分解人体姿态为不同部位的姿态，将整个姿态估计问题转化为一系列局部姿态回归问题。

直接回归方法的训练过程是直接将 2D 图像转换为 3D 关节坐标，这导致了其对噪声和非常规姿态的

鲁棒性较差。即使输入的 2D 图像相同，由于镜头的距离、光照和人体形变等因素的影响，得到的 3D 姿态也会有所不同，且模型通用性较差[4]：对于不同的人体形态和姿态，直接回归方法不具有很好的通用性。因为人体形态和姿态变化非常大，而该方法在训练时通常只学习到了一些特定的形态和姿态，因此在新的测试数据上可能表现不佳。

(2) 基于 2D-to-3D 提升的人体姿态：

由于目前 2D 人体姿态估计准确度非常高，从 2D 人体姿态出发推断 3D 人体姿态已成为一种流行的 3D 人体姿态估计解决方案。在第一阶段，使用现成的 2D 人体姿态估计模型来估计 2D 姿势。然后在第二阶段通过 2D 到 3D 提升获得 3D 姿态。

此外，Chen [5]等人提出了一种无监督提升网络，该网络基于提升－再投影－提升过程的闭包性和不变性提升特性，具有几何自一致性损失。闭合意味着一个经过提升得到的三维骨架，经过随机旋转和重投影后，得到的二维骨架将位于有效二维姿态的分布范围内。不变性是指以不同的角度投影三维骨架得到的二维姿态，经过重新提升后的三维骨架应该是相同的。这种方法可以在没有标记数据的情况下提高三维人体姿态估计的准确性。

随着序列方法(RNN、LSTM、Transformer 等)研究的成熟，有研究者尝试输入视频序列来预测 3D 姿态。Hossain 和 Little [6]提出了一种使用具有快捷连接的长短期记忆(LSTM)单元的循环神经网络，以利用人体姿势序列中的时间信息。他们的方法利用序列到序列网络中的过去事件来预测时间一致的 3D 姿态。但是之前的工作通常忽略了空间约束和时间相关性之间的互补性质，基于此 Pavllo [7]等人提出了一种可扩张时序卷积网络，用于从连续的 2D 序列中估计 2D 关键点上的 3D 姿态。

鉴于人体姿势可以表示为一个图，其中关节是节点，骨骼是边缘，图卷积网络(GCNs) [8]已被应用于 2D~3D 姿势提升问题，并显示出良好的性能。Ci [9]等提出了一种局部连通网络(Local Connected Network, LCN)，它利用全连通网络和 GCN 来编码局部联合邻域之间的关系。LCN 可以克服 GCN 权值共享方案对姿态估计模型表示能力的限制，以及结构矩阵缺乏支持自定义节点依赖的灵活性。Zhao [10]等人也解决了 GCN 中所有节点的卷积滤波器共享权矩阵的局限性。为了研究语义信息及其关系，提出了一种语义 GCN 算法，语义图卷积(SemGConv)操作用于学习边缘的通道权重。由于 SemGConv 和非局部层是交错的，因此可以捕获节点之间的局部关系和全局关系。

3. 模型框架

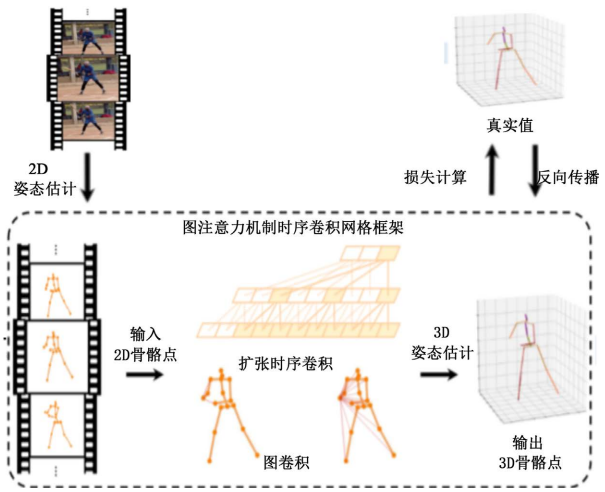


Figure 1. Framework for 3D pose recognition algorithm
图 1. 3D 姿态识别算法框架

本文提出的 3D 姿态检测算法的总体框架如图 1 所示，图中的箭头基于数据流动方向示意给出了算法流程。首先如图 1 左上方所示，它采用现有算法对视频图像进行 2D 姿态估计，获得人体二维骨骼坐标序列作为下一步 3D 姿态估计的输入；然后如图 1 中虚框中部所示，它采用本文提出的图注意力扩张时序卷积网络(GA-TCN)将 2D 姿态提升为 3D 姿态，输出人体三维骨骼坐标点。在训练阶段，如图 1 右上方所示，它以真实三维人体位姿作为监督计算损失，通过反向传播完成 GA-TCN 网络的参数训练。GA-TCN 网络的详细结构和原理见下节讨论。

4. 图注意力时序卷积网络

4.1. 图卷积网络与图注意力机制

对图 2(a)所示二维周期网格，例如图像像素，采用传统卷积神经网络就可对其进行特征提取等各种处理。但是对图 2(b)所示非规则网格，必须采用图卷积网络(Graph Convolutional Network, GCN)。在 GCN 中网格节点的拓扑关系复杂多变，需要通过连接矩阵进行定义。

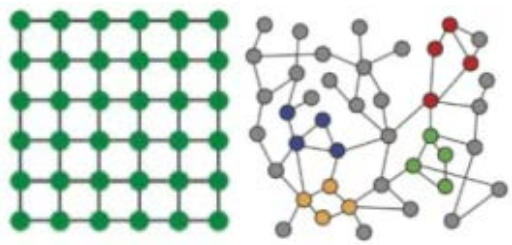


Figure 2. Periodical 2D grids and non-periodical 3D grids
图 2. 二维周期网格与非规则三维网格

在图卷积神经网络中，邻接矩阵能够提取节点之间的拓扑结构信息。它描述了每个节点与其他节点之间的连接关系，包括基于位置、图像特征等的表示方法。通过邻接矩阵，网络能够识别节点之间的依赖关系、拓扑结构及节点的位置信息等信息，从而更好地理解图像特征之间的关系，并捕捉人体关键点之间的空间和结构信息。

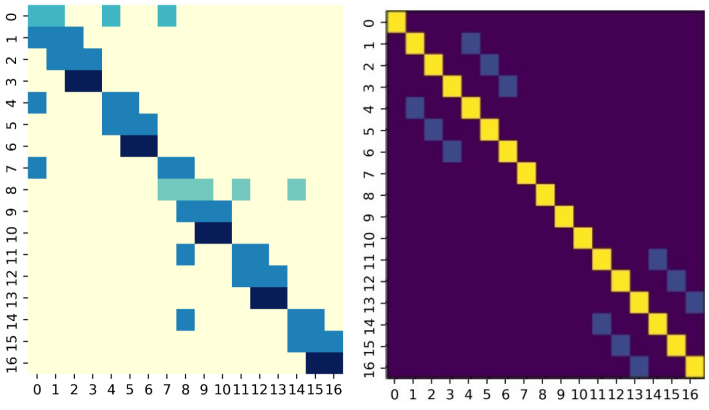


Figure 3. (a) Representation of the adjacency matrix $\tilde{A}_c$ as a heatmap; (b) Representation of the symmetric matrix $\tilde{A}_s$ as a heatmap
图 3. (a) 邻接矩阵 $\tilde{A}_c$ 热图表示；(b) 对称矩阵 $\tilde{A}_s$ 热图表示

图 3(a)为邻接矩阵的热图表示，其中边缘节点(3, 6, 10, 13, 16)分别为头、双脚、双手，这些节点在热图由深蓝色块标记，可以很清楚的看到相比其他节点，边缘节点仅有一条连接边。而在一般的图卷积任务中仅仅使用一阶邻域表示，这种表示在模拟以躯干为中心的人体时通常表现很差，因为这忽略了人体的对称性结构和人体四肢的运动学约束。一阶邻域表达将关节约束仅限于相邻的一个关节，这对于运动链末端的边缘关节缺乏约束，因此它们在空间中的位置不能被有效地定位。这也是重建误差的最大来源。因此本文引入人体结构相关知识，加入了基于人体骨骼对称结构的图卷积核邻接矩阵 $\tilde{A}_s$ 。

对于来自视频的 2D 姿态预测序列的输入，本论文将二维关键点视为代表人类骨骼的关节，用一个无向图表示骨架，其中关节是节点，关节之间的链接是边。本论文基于 Zhao [10]提出的 SemGCN 构造给定单帧的二位关键点骨架图，将骨骼 2D 姿态定义为 $graphG=(V,\ \mathrm{E})$，其中 V 是 N 个节点和 E 条边的集合。设 $X=\{x_1, x_2, \ldots, x_N \mid x_i \in \mathbb{R}^{1\times C}\}$ 是包含 C 个特征的节点的集合，在本文中 C = 2，即一个节点由其二维平面中的坐标值 (x,y) 表示。图的结构可以用一阶邻接矩阵 $\mathrm{A}\in\mathbb{R}^{N\times N}$ 表示节点之间存在的连接，用单位矩阵 I 表示自连接，$\tilde{A}=(A+\mathrm{I})$ 表示在邻接矩阵中添加自连接信息。给定第 l 层的节点特征，通过以下卷积得到后续层的输出特征如下所示：

$$X^{(l+1)}=\rho\left(M\odot\tilde{A}\right)X^{(l)}W \tag{1}$$

其中 $W\in\mathbb{R}^{C_l\times C_{l+1}}$ 是一个通过反向传播学习的矩阵，用于转换输出通道，$M\in\mathbb{R}^{N\times N}$ 是一个可学习的掩码矩阵，$\odot$ 是元素乘运算，ρ 是一个 Softmax 非线性激活函数，它将一个节点的特征归一化到图中相应的相邻节点。

通过对输出节点特征的通道引入一组掩码矩阵 $M_c\in\mathbb{R}^{N\times N}$，可将上式推广为：

$$X^{(l+1)}=\mathop{\Big\|}_{C_l}^{C_{l+1}}\rho\left(M_c\odot\tilde{A}\right)X^{(l)}w_c \tag{2}$$

其中 $\|$ 表示通道级联，w_c 是矩阵 W 的第 c 行。

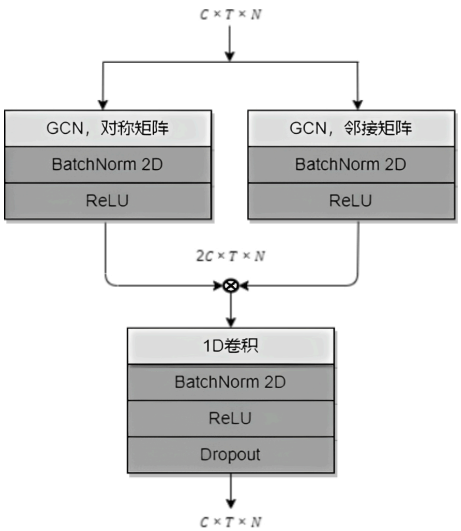


Figure 4. Graph attention mechanism module
图 4. 图注意力机制模块

用分别表示人体节点连接和对称关系的邻接矩阵和对称矩阵分别组成图卷积网络，再将两者相结合即构成图注意力机制模块，如图 4 所示，其中 C 为通道数，T 为接收域(即同时输入的视频帧数)，N 为关

节数量(17 个关节)。输入分别经过对称图卷积块和邻接图卷积块，然后将输出在通道维度上连接，最后通过一个一维卷积降维保持通道维度和输入维度一致。

经过图注意力机制对输入序列的每一帧数据进行处理后，添加了人体骨架连接信息和人体对称信息，后续进一步传输给扩张时序卷积层在时间维度上建模。

4.2. 扩张时序卷积

人体姿态是连续变化的，人体姿态的某一关节点在某一时刻可能无法准确提取，例如被遮挡，但是在下一时刻遮挡可能消失。换句话说，进一步利用视频动作的时间相关性，可以有效增加人体姿态估计的准确性和鲁棒性。为此本文引入了一个完全由卷积构成的具有残差连接的扩张时序卷积网络，它以图注意力机制处理后的 2D 姿势序列作为输入，最终重建出 3D 人体姿态序列。扩张时序卷积网络具有许多优点，例如可以克服 RNN [7]无法在时间上进行并行化处理的缺点；再如其输出和输入之间的梯度路径具有固定长度，不受序列长度影响，可以有效缓解影响 RNN [11]的消失和爆炸梯度问题；同时还提供了对于时域接收域精确控制，这在 3D 姿态估计任务中非常有效。

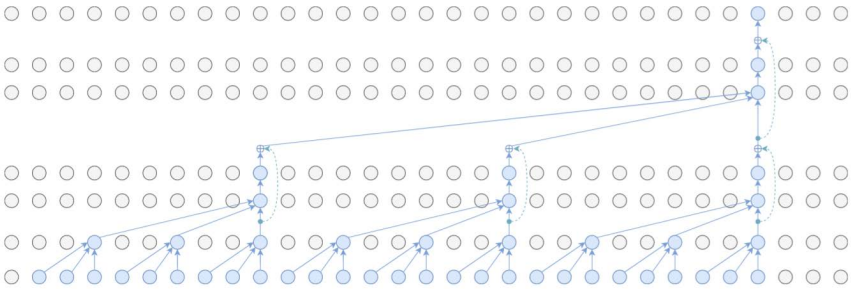


Figure 5. Graph Attention Mechanism Module
图 5. 扩张时序卷积网络核心结构

扩张时序卷积网络的核心结构如图 5 所示。根据其中的网络连接关系，给定一个第 l 层的长为 T 有 M 维特征的输入序列 $X^l \in R^{T\times M}$，以及卷积核，卷积运算 F 定义如下

$$F = \left(x *_d f\right) = \sum_{i=0}^{k-1} f(i) \cdot x_{s-d\cdot i} \tag{3}$$

其中 d 是扩张因子，k 是卷积核大小，$s - d\cdot i$ 表示过去的方向。扩张操作相当于在卷积核内部引入固定步长。当 $d = 1$ 时，扩张卷积退化为常规卷积。使用更大的扩张使得顶层的输出能够代表更大范围的输入，从而有效地扩展卷积层的接收域，有利于提取更长时间间隔的时间相关性。

5. 实验结果与分析

5.1. 数据集准备与评估指标

本文实验采用目前使用最广泛的室内人体姿态数据集 Human3.6M [12]。该数据集涉及从 4 个不同视角拍摄的 11 名专业演员(6 男 5 女)在室内进行的 17 项活动(例如吸烟、拍照、打电话)，共包含 360 万个 3D 人体姿势，以及由基于标记的准确 MoCap 系统捕获的 3D 地面实况注释，可同时获取 24 个身体部位的实时三维坐标。本文使用其中的 S1、S5、S6 和 S7 的图像进行训练，使用 S9 和 S11 的图像进行测试。实验对 GA-TCN 算法中的图注意力模块和扩张时序模块分别进行了消融实验，训练过程中使用 PyTorch 框架实现本论文的方法并进行训练与测试。对于 Human3.6M，使用批量大小 b = 128 的 Amsgrad 进行优化，并训练 50 个 epoch。学习率从 0.001 开始，然后在每个 epoch 中应用学习收缩因子 α = 0.95。每个 dropout

层均设置为 0.05。

本文采用平均关节点位置误差[13] (Mean Per Joint Position Error, MPJPE)和标准化平均关节位置误差[14] (Normalized Mean Per Joint Position Error, NMPJPE)作为模型的评价指标来评估其效果。其中 MPJPE 的计算公式如下：

$$MPJPE = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{J}\sum_{j=1}^{J}\left\|\mathbf{X}i,j-\hat{\mathbf{X}}i,j\right\|_2 \tag{4}$$

式中 N 表示输入视频帧数，J 表示关节数，Xi,j 表示第 i 个样本中第 j 个关节的真实三维坐标点，$\hat{\mathbf{X}}i,j$ 表示第 i 个样本中第 j 个关节的预测三维坐标点。

NMPJPE 通过计算预测关节位置与真实关节位置之间的欧氏距离来衡量算法的准确性，并对结果进行标准化以考虑不同人体模型的缩放和旋转。具体计算公式如下：

$$NMPJPE = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{J}\sum_{j=1}^{J}\frac{\|\mathbf{X}i,j-\hat{\mathbf{X}}i,j\|}{s} \tag{5}$$

其中 S 是一个标准化因子，通常是参考关节的长度或者整个人体模型的对角线长度。

5.2. 实验结果与讨论

在实验中本文采用的硬件设备和软件环境配置分别如表 1 和表 2 所示。

Table 1. Hardware equipment configuration
表 1. 硬件设备配置

硬件名称	硬件参数
处理器	Intel(R) Core(TM) i7-6700K CPU @ 4.00 GHz
内存容量	16 GB
操作系统	Windows 10 专业版 64 位
GPU	NVIDIA GeForce RTX 1060
显存	6 GB

Table 2. Software equipment configuration
表 2. 软件设备配置

环境名称	软件版本
Python	Python 3.7.13
Anaconda	Conda 22.9.0
CUDA	V11.4.100
Pytorch	1.12.0

本文首先比较了图注意力模块对模型的提升效果。在实验评估中本文以 VideoPose3D 提出的预测模型为基底，同时为了降低时域因素的干扰，将时序卷积的接收域统一规定为 27 帧。两个模型在训练过程中验证集的误差如图 4 所示，蓝色线段表示未引入图注意力机制的模型，橙色线段表示引入图注意力机制的模型。从图 4 中可以明显看出在引入图注意力后模型效果的提升。

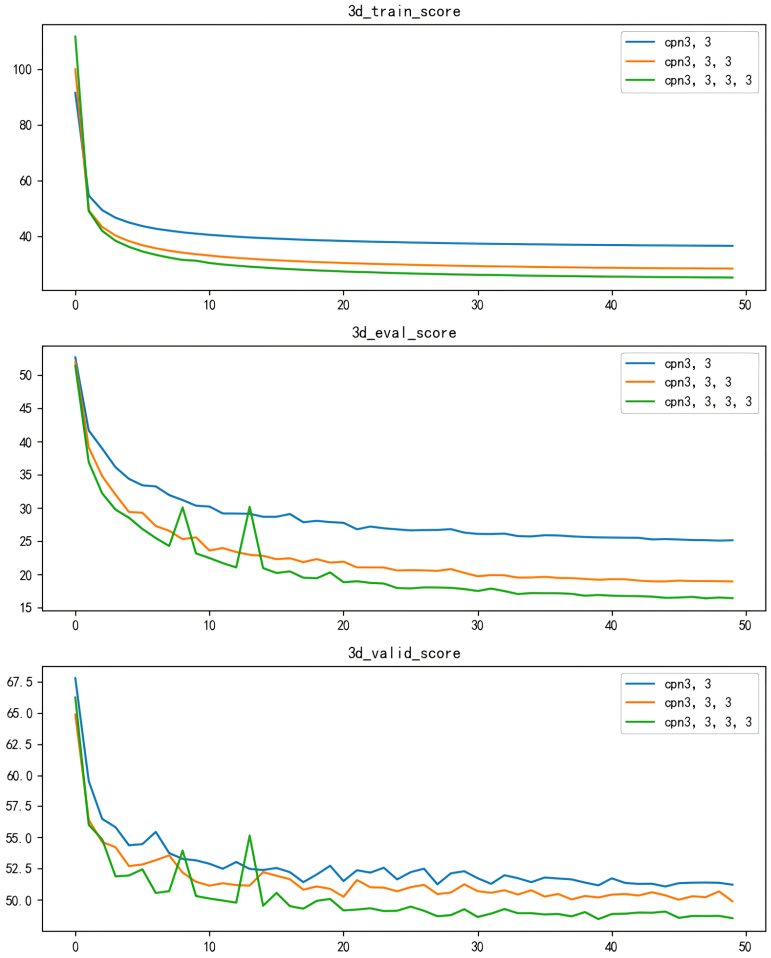
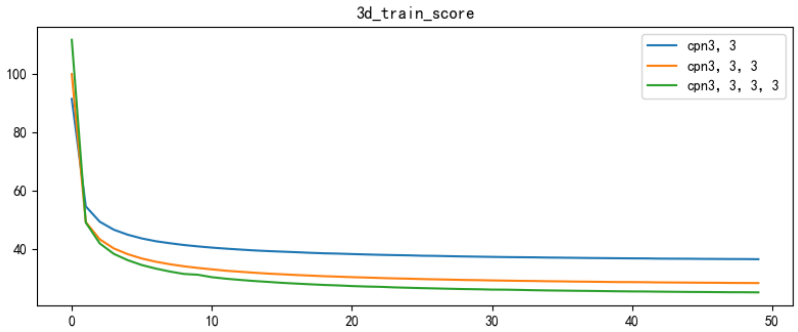


Figure 6. (top) Changes in loss during the training process; (middle) comparison of MPJPE; (bottom) comparison of NMPJPE
图 6. (上) 训练过程中损失变化；(中) MPJPE 对比；(下) NMPJPE 对比

图 6 说明本文提出的模型分别在时序卷积逐层增加的情况下，损失函数曲线出现明显的下降。在时序卷积扩展因子设定为[3, 3, 3, 3]的情况下，平均MPJPE为25.11 mm，相比于三层扩展因子[3, 3, 3]的28.30 mm 降低了 12.7%，相比于两层扩展因子[3, 3]的 36.46 mm 降低了 31.1%。这证明在卷积步长不变的基础上，加深时序卷积的叠加层数能有效地提升模型性能。



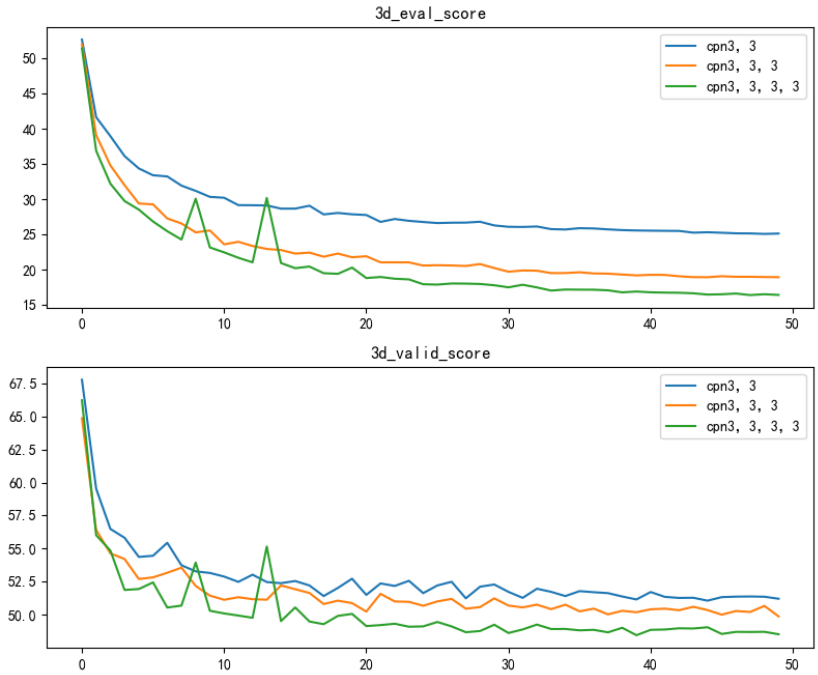


Figure 7. The decreasing curve of the loss function on the training set, validation set, and test set under different expansion factors
图 7 不同扩展因子下训练集、验证集和测试集的损失函数的下降曲线

图 7 进一步给出了引入图注意力机制后模型在各个动作上的误差表现(表中 GA-TCN，T = 27，without GA 那一行为不引入图注意力机制，接收域为 27 帧的模型，而后续各行为引入图注意力机制，但接收域不同)。从表 3 中可以看出，与其他 3D 姿态模型相比(表中前 10 行)，本文提出的模型在各种动作上的估算误差均有所下降。

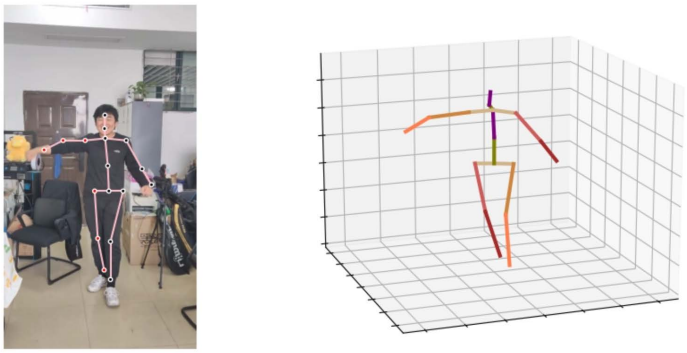
Table 3. Comparison of MPJPE among different models on various action validation sets
表 3. 各模型在不同动作验证集中的 MPJPE 对比

模型	吃饭	问候	电话	拍照	购物	坐着	坐下	抽烟	等待	行走	平均
Pavlakos et al. CVPR’17	66.7	69.1	72	77	68.3	83.7	96.5	71.7	65.8	59.1	71.9
Tekin et al. ICCV’17	60.2	61.2	79.4	78.3	81.6	70.1	107	69.3	70.3	51.8	69.7
Martinez et al. ICCV’17	58.1	59	69.5	78.4	58.1	74	94.6	62.3	59.1	49.5	62.9
Sun et al. ICCV’17	54.2	54.3	61.8	67.2	53.6	71.7	86.7	61.5	53.4	47.1	59.1
Fang et al. AAAI’18	57	57.1	66.6	73.3	55.7	72.8	88.6	60.3	57.7	47.5	60.4
Pavlakos et al. CVPR’18	54.4	52	59.4	65.3	52.9	65.8	71.1	56.6	52.9	44.7	56.2
Yang et al CVPR’18	50.4	57	62.1	65.4	52.7	69.2	85.2	57.4	58.4	60.1	58.6
Luvizon et al. CVPR’18	47.6	50.5	51.8	60.3	51.7	61.5	70.9	53.7	48.9	44.4	53.2
Hossain & Little ECCV’18	57.2	55.2	63.1	72.6	51.7	66.1	80.9	59	57.3	46.6	58.3
Lee et al. ECCV’18	47.8	52.6	50.1	75	43	55.8	73.9	54.1	55.6	43.3	52.8
GA − TCN, T = 27, without GA	49	51.8	53.6	61.4	47.4	59.3	67.4	52.4	49.5	39.5	52.7
GA − TCN, T = 27, CPN	45	47.5	50.7	56.9	45	58.3	65.3	49.1	45.8	36.1	49.6
GA − TCN, T = 91, CPN	43.2	46.1	49.2	56.5	43.8	56.7	64.8	47.8	44.5	33.4	48.2
GA − TCN, T = 105, CPN	42.9	46.1	48.3	57.3	44.6	56.7	65.5	47.4	44.9	33.5	48.3
GA − TCN, T = 125, CPN	43.2	45.5	49.2	56.7	43.5	56.4	65.4	47.3	44.1	33.4	48.1
GA − TCN, T = 125, GT	34.3	39.6	38.6	44.7	39.4	44	46.8	38.7	42.7	38.9	41.1

通过实验结果的对比分析可以发现，在引入图注意力机制之后，模型在测试集上的平均测量误差为49.6 mm，较改进前降低 5.8%。在扩展因子固定为三层的情况下，增加每次时序卷积的步幅，损失函数曲线略微下降。其中[5, 5, 5]和[3, 5, 7]的接收域分别为 125 帧和 105 帧，相比于[3, 3, 3]的 27 帧有较大的扩展。其结果也反映到模型性能上，在接收域扩展将近 3 倍的情况下，平均 MPJPE 仅仅降低了 2.7%。这表明相较于加深时序卷积的深度，在单个时序卷积上加大步幅对模型的性能提升较小。

表 3 给出了在不同模型在各动作上的误差对比，数据表明在不存在自遮挡的动作中(如问候、等待、行走等)，本模型较其他模型平均误差降低 8.9%；在存在自遮挡的动作中(如坐着、抽烟、电话等)，本模型较其他模型平均误差降低 11.5%验证了图注意力机制的有效性。实验证明本文构建的人体三维姿态估计模型具有较好的预测准确性，并在各项存在自遮挡的动作上提升较为显著。

图 6 给出了本文的一组 3D 人体姿态估计实验测试结果。其中在图 8(a)的视频图像中，标出了所检测到的人体 2D 姿态骨骼关节点，这些骨骼关节点对应的三维坐标如图 8(b)所示。在表 1 和表 2 所示配置下，文本所提出的 3D 人体姿态估计算法可以达到约每秒 14 帧的准实时检测速度。



(a) (b)

Figure 8. Model detection results
图 8. 模型检测结果

6. 结论

本文提出了一种基于深度学习的三维人体姿态估计算法，该算法充分利用人体骨骼节点的连接关系和对称关系构建了一种图注意力时间卷积网络，该网络可以充分利用单目视频中的时空信息，解读人体姿态随时间的变化，从而相较传统方法表现出更好的准确性和鲁棒性。该算法可以达到准实时检测速度，可以广泛应用于老人安保监测、VR、虚拟数字人和元宇宙等领域。

参考文献

[1] Moeslund, T. and Hilton, A. (2006) A Survey of Advances in Vision-Based Human Motion Capture and Analysis. *Computer Vision and Image Understanding*, **104**, 90-126. https://doi.org/10.1016/j.cviu.2006.08.002

[2] 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, 483-499. https://doi.org/10.1007/978-3-319-46484-8_29

[3] Chen, C.-H., Tyagi, A., Agrawal, A., *et al*. (2020) Unsupervised 3D Pose Estimation with Geometric Self-Supervision. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, USA, 15-20 June 2019. https://doi.org/10.1109/CVPR.2019.00586

[4] Zhou, X., Huang, Q., Sun, X., *et al*. (2017) Towards 3D Human Pose Estimation in the Wild: A Weakly-Supervised Approach. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, Italy, 22-29 October 2017. https://doi.org/10.1109/ICCV.2017.51

[5] Chen, C.-H. and Ramanan, D. (2017) 3D Human Pose Estimation = 2D Pose Estimation + Matching. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, USA, 21-26 July 2017. https://doi.org/10.1109/CVPR.2017.610

[6] Hossain, M.R.I., 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, Cham, 69-86. https://doi.org/10.1007/978-3-030-01249-6_5

[7] Pavllo, D., Feichtenhofer, C. and Grangier, D. (2020) 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, CA, USA, 15-20 June 2019. https://doi.org/10.1109/CVPR.2019.00794

[8] Wu, Z., Pan, S., Chen, F., *et al*. (2020) A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, **32**, 4-24.

[9] Ci, H., Wang, C., Ma, X., *et al*. (2020) Optimizing Network Structure for 3D Human Pose Estimation. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, Korea (South), 27 October - 2 November 2019. https://doi.org/10.1109/ICCV.2019.00235

[10] Zhao, L., Peng, X, Tian, Y., *et al*. (2020) Semantic Graph Convolutional Networks for 3D Human Pose Regression. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, USA, 15-20 June 2019. https://doi.org/10.1109/CVPR.2019.00354

[11] Wang, J., Tan, S., Zhen, X., et al. (2021) Deep 3D Human Pose Estimation: A Review. *Computer Vision and Image Understanding*, **210**, Article ID: 103225. https://doi.org/10.1016/j.cviu.2021.103225

[12] Ionescu, C., Papava, D., Olaru, V., *et al*. (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

[13] de La Gorce, M., Fleet, D.J. and Paragios, N. (2011) Model-Based 3D Hand Pose Estimation from Monocular Video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **33**, 1793-1805. https://doi.org/10.1109/TPAMI.2011.33

[14] Carreira, J., Agrawal, P., Fragkiadaki, K. and Malik, J. (2016) Human Pose Estimation with Iterative Error Feedback. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, NV, USA, 27-30 June 2016, 4733-4742. https://doi.org/10.1109/CVPR.2016.512