doi: 10. 3969 / j. issn. 1674 - 1242. 2023. 04. 009

基于时频结合空间注意力网络的 3D 骨架人体运动预测

张禄钧，何志权
（广东省多媒体信息服务工程技术研究中心，深圳大学电子与信息工程学院，广东深圳 518060）

【摘要】以往基于时域和频域的图卷积网络模型在三维人体运动预测上显示出令人印象深刻的结果。然而，时域和频域是同一个人体动作信号在不同域的表现，该文同时结合人体位姿在时域和频域的序列进行编码，并在两个通道上对不同表现形式的关节信息通过注意力机制强化人体骨骼各节点之间的相互依赖关系。最后通过基于图的门控循环单元（G-GRU）对编码信息进行递归解码，输出预测的运动序列。该文在 Human 3.6M 和 CMU-MoCap 数据集上测试了所提的模型，实验证明，该文所提的模型能获得比以往模型更准确的预测结果。

【关键词】时域；频域；注意力机制；基于图的门控循环单元
【中图分类号】TP18　　【文献标志码】A　　文章编号：1674-1242（2023）04-0391-07

Time-Frequency Spatial Attention Network for 3D Skeleton Human Motion Prediction

ZHANG Lujun, HE Zhiquan
(Guangdong Engineering Research Center of Multimedia Information Service, School of Electronic and Information Engineering, Shenzhen University, Shenzhen, Guangdong 518060, China)

【**Abstract**】Previous works on graph convolutional networks based on temporal recurrent networks and frequency domains has shown impressive results in three-dimensional human motion prediction. However, the time domain and frequency domain are the manifestations of the same human action signal in different domains, and this paper encodes the observed movement sequence of the human body in the time and frequency domains in combination with the human posture in the time and frequency domains, and strengthens the interdependence between nodes of human bones through the attention mechanism of joint information of different manifestations in the two channels. Finally, the gated loop unit (G-GRU) based on the graph is used to recursively decode the encoded information and output the predicted motion sequence. We tested our model on the Human 3.6M and CMU-MoCap datasets, and experiments proved that our model can obtain more accurate predictions than previous methods.

【**Key words**】Time Domain; Frequency Domain; Attention Mechanism; A Graph-Based Gated Cycle Unit

收稿日期：2022-12-03。
基金项目：国家自然科学基金（61971290）。
作者简介：张禄钧（1998—），男，广东省茂名市人，硕士研究生，从事行为预测研究。
通信作者：何志权，男，副教授，硕士研究生导师，电话（Tel.）：18680358296，邮箱（E-mail）：zhiquan@szu.edu.cn。

0　引言

3D 人体运动预测旨在根据观测到的历史 3D 骨架姿态序列，预测人体在未来时刻的 3D 姿态和运动轨迹，是计算机视觉和机器学习领域的一个关键而又具有挑战性的问题。相比 2D 情形，3D 人体运动预测需要处理高维度的输入数据，并建模人体复杂的运动学和动力学约束关系。具体来说，该任务面临以下 3 个主要挑战。①捕捉人体运动的高维度。与 2D 情形相比，3D 骨架序列存在更多自由度，机器学习模型需要学习更复杂的时空关联模式。同时，真实运动中的遮挡也为准确建模带来了困难。②建模复杂的人体运动学和动力学约束。人体运动受复杂的生物力学约束，如各关节自由度的约束、左右半身的协调性的约束等。学习这些先验知识对产生逼真的预测至关重要。③产生连贯的长期预测。由于历史信息的快速衰减，模型很难捕捉长时间范围内的运动规律，典型的顺序模型产生的远期预测往往会出现“漂移”现象。

为应对上述挑战，研究者探索了各种创新性深度学习架构。首先，利用循环神经网络或图卷积网络建模时空数据。然后，采用注意力机制或残差连接学习远期依赖。最后，应用增强学习等方式引入运动学先验，另外，研究者在运动数据获取、力学参数测量及人体动力学模型建立等方面取得了许多重大进展，在一定程度上推动了 3D 人体运动预测的进展，但距离实际应用要求仍有一定差距。未来的研究可能会关注以下方向。①提高对长期规律的建模能力。②探索分层或模块化模型以捕捉不同粒度的时空模式。③设计更强大的框架，同时学习运动预测和运动生成。

基于 3D 骨架的人体运动预测通过给定的由关键关节和身体骨架组成的先验 3D 姿态序列预测未来的姿态序列，它在众多领域都起着至关重要的作用[1-4]。例如，基于 3D 骨架的人体运动预测对于自动驾驶[5]和行人跟踪[6-8]等智能交互至关重要。

近年来，神经网络在人体运动预测中得到了越来越多的认可。文献[9]提出了一种基于循环神经网络的编码器-循环-解码器（Encoder Recurrent Decoder，ERD）网络。由于姿态序列的不同身体部位之间的关系或约束，文献[10]提出了一种多尺度图神经网络来提取不同身体部位之间的特征关系。文献[11]提出了一个骨架网络（SkeNet）来分别学习不同身体组件的局部表示。文献[12]提出了一种简单的前馈深度网络进行运动预测，该网络同时考虑了人体关节之间的时间平滑度和空间依赖性。

循环神经网络（Recurrent Neural Network，RNN）也在运动预测领域取得了优秀的成果。例如，文献[13]捕获了底层时空图中的丰富交互，并提出了 Structural-RNN。除了基于 RNN 的方法，文献[14]提出了一种使用卷积网络的层次结构来捕获时空相关性的方法，文献[15]利用基于 Transformer 的架构来生成人体运动模型。

在以往的工作中，往往单独在时域或频域对动作序列进行预测，难以兼顾人体骨架运动序列在时间和空间上的依赖信息。为此，本文构建了一个以编码器-解码器为框架的 3D 骨架人体运动预测网络。本文的贡献如下。①同时结合时域和频域的信息对未来姿态进行预测。②提出了一个基于图卷积网络的空间注意力模块（Graph based Spatial Attention Model，GSAM），学习人体骨架的空间依赖性，以实现准确的预测。③采用参数量少的自控门注意力机制来强化人体骨骼各节点之间的相互依赖关系。

1　问题描述

在 3D 人体运动预测中，用关节组成的骨骼运动树来表示人体姿态是一种传统的方法。在解剖学中，人体包含数百个关节，但本文只关注运动捕捉系统记录的关节。假设 $X_{-T_p:0}=[X_{-T_p},\cdots,X_0]$ 表示长度为 T_p+1 的过去观察到的姿态，其中 $X_i\in\mathbb{R}^{N\times D}$ 表示 N 个节点的 $D=3$ 维数据在 i 时刻的运动姿态。姿态预测的任务是根据过去观测到的姿态 $X_{-T_p:0}$，产生未来的长度为 T_f 的姿态 $X_{1:T_f}=[X_1,\cdots,X_{T_f}]$。

准确预测人体未来的动作姿态序列的挑战在于人类行为的复杂性和人体的灵活性。因此，理解和预测人类运动对人类和机器来说都不是一项容易完成的任务。需要解决的两个关键问题是固有运动学问题和网络性能限制问题。人类运动预测的固有运动学问题是由人类运动的高度随机性、高维性和非线性引发的，这导致了未来人体姿态的高度不确定性。网络性能限制问题源于固有的网络缺陷，这些缺陷不可避免地涉及 RNN 在长时间预测上的错误积累，以及 2D CNN 的感受野局限性。

为了应对上述挑战，研究人员提出了不同的解决

方案。最初，研究人员专注于对人类运动序列进行建模，这取决于专门研究序列问题的基于深度 RNN 的架构。由于生成对抗性网络的快速发展，它们被用作一种新的人体运动预测学习算法。此外，考虑到关节之间的连接，图卷积网络是细胞神经网络的推广，用于对关节之间的相关性进行建模，用于人体运动预测。同时，研究人员还提出了其他细胞神经网络方法，如将注意力机制融入 RNN 中，以捕获人体动作的重要时刻和关键姿态，从而产生更准确的预测。

为进一步提高预测的长期依赖性，一些研究人员探索了强化学习方法和模仿学习方法中的思想。这些方法不仅可以预测短期运动，还可以捕捉人体运动的长期规律，产生更连贯的长序列预测。与此同时，一些研究人员提出将图神经网络与 RNN 或卷积网络相结合，以构建复杂的人体骨骼和肌肉之间的关系，进一步提高对高维度人体运动的建模能力。总体来说，随着深度学习等技术的发展，人类运动预测不断取得进展，但仍面临诸多挑战，需要使用更多创新的思路来建模运动学问题并解决网络性能限制问题，以产生更逼真和更连贯的人体动作预测。

2　相关技术介绍

2.1　注意力机制

深度学习中的注意力机制是一种模仿人类视觉和认知系统的方法，它允许神经网络在处理输入数据时集中注意力于相关的部分。通过引入注意力机制，神经网络能够自动地学习并选择性地关注输入中的重要信息，从而提高模型的性能和泛化能力。

注意力机制从本质上讲和人类的选择性注意力机制类似，两者的核心目标都是从众多信息中选出对当前任务目标更加关键的信息。在深度学习中，注意力机制通常应用于序列数据（如文本、语音或图像序列）的处理。其中，典型的注意力机制包括自注意力机制、空间注意力机制和时间注意力机制。这些注意力机制允许模型对输入序列的不同位置分配不同的权重，以便在处理每个序列元素时专注于最相关的部分。

在传统的预测方法中（如使用深度 CNN 模型预测未来动作序列），一般通过卷积核提取动作序列的局部信息。然而，每个局部信息的信息序列对未来动作序列的影响力是不同的，如何让模型知道序列中不同局部信息的重要性呢？为了解决这个问题，本文引入了注意力机制。

注意力机制本质上是一种权重分配技术，使模型能够根据权重关注重要的内容，抑制不重要的内容。Transformer[16]是最有效的注意力机制之一，另一个注意力机制是自门控操作[17]。与变压器相比，自门控通常需要较低的计算成本，可以利用各种特征信息计算门控权重。例如，SE-Net[17]通过沿空间轴聚合的信息动态计算元素的门控权重。本文所提的方法也采用了类似的策略进行全局信息建模。

2.2　离散余弦变换

离散余弦变换（Discrete Cosine Transform，DCT）是 Nasir Ahmed 于 1974 年提出的正交变换方法，被认为是对语音和图像信号进行变换的最佳方法之一。为了满足工程的需要，国内外许多学者花费了很大的精力去寻找或改进 DCT 的快速算法。由于近年来数字信号处理（Digital Signal Processing，DSP）芯片的发展迅速，加上专用集成电路具备设计优势，目前 DCT 在图像编码中的重要地位日益突出，成为 H.261、JPEG、MPEG 等国际上公用的编码标准的重要环节。在视频压缩中，最常用的变换方法也是 DCT，它被认为是性能接近 K-L 变换的准最佳变换。变换编码的主要特点有以下几个。①变换域的视频图像要比空间域的简单。②视频图像的相关性明显下降，信号的能量主要集中在少数几个变换系数上，采用量化和熵编码可有效压缩数据。③具有较强的抗干扰能力，传输过程中的误码对图像质量的影响远小于预测编码。DCT 等变换方法有快速算法，能实现实时视频压缩。

本文采用 DCT 对人体骨架运动的时间序列进行编码，这在以往的实验中被证明是有用的[18]。DCT 可以由下式计算。

$$D_{i,j}=\sqrt{\frac{2}{T}}\frac{1}{\sqrt{1+\delta_{i,0}}}\cos\left(\frac{\pi}{2T}(2j+1)i\right),$$

$$\delta_{i,j}=\begin{cases}1 & i=j\\ 0 & i\neq j\end{cases}$$

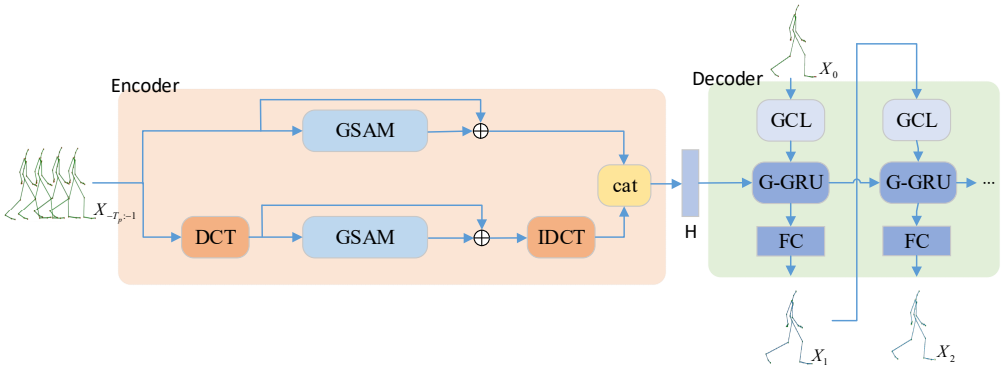
在频域进行信息编码和提取后，采用反余弦变换（Inverse Discrete Cosine Transform，IDCT）将运动信息转换为时域的位姿表示。

3　方法描述

3.1　整体网络架构

本方法的整体网络架构如图 1 所示，网络以编码

器-解码器为框架。编码器以已观测的位姿序列 $X_{-T_p:-1}$ 作为输入，在时域和频域分别对已观测的位姿序列进行特征提取，并利用自控门注意力机制捕捉人体各个关节之间的依赖关系，输出融合的隐藏层特征。解码器以观测到的最后一帧位姿 X_0 作为输入，递归输出预测的位姿序列 $X_{1:T_f}$ 。



⊕ 表示残差结构；cat 表示矩阵拼接方法。

图 1　整体网络架构

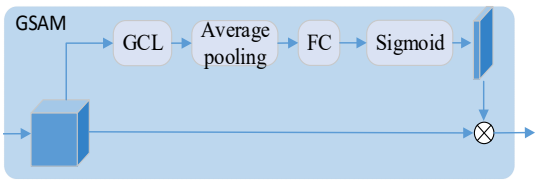
Fig. 1　Overall network architecture

3.1.1　编码器模块

考虑到人体不同关节之间在运动中具有相互依赖性，本文提出了一个基于 GCN 的空间注意力模块（GSAM），以提取人体不同关节之间在不同运动中的依赖信息，如图 2 所示。本文的方法通过多层的 GCN 层对已观测的位姿序列 $X \in \mathrm{R}^{N\times D}$ 进行特征提取，GCN 的特征提取可以总结为如下公式。

$$X^{(l+1)} = \sigma(AX^{l}W^{l})$$

式中，$A \in \mathrm{R}^{N\times N}$ 表示可训练的内建图邻接矩阵，W 表示参数调整矩阵。然后使用均值池化层沿空间维度聚合节点信息，根据 Sigmoid 函数计算每个节点在区间[0,1]的门控权重值，最后根据动态的关节权重聚合依赖节点之间的信息。



GCL 表示级联的 GCN 网络层（GC-Layer）；⊗ 表示矩阵的乘法操作。

图 2　基于 GCN 的空间注意力模块结构

Fig. 2　Spatial attention module structure based on GCN

3.1.2　解码器模块

解码器模块由多层的 GCN 网络、基于图的 GRU 模块（G-GRU）及全连接层构成。G-GRU 模块的功能是在图的引导下学习和更新隐藏的状态。关键是使用一个可训练的图来正则化状态，这些状态用于生成未来的位姿。设 $H(0) \in \mathrm{R}^{N\times D}$ 为 G-GRU 模块的初始状态。在时间 $t>0$ 处，G-GRU 有两个输入：初始状态 $H(t)$ 和 GCL 层的 3D 骨架的信息输出 $\mathrm{GCL}(X(t)) \in \mathrm{R}^{N\times D}$ 。G-GRU 的运算可以总结为下式。

$$r^{(t)} = \sigma(r_i(X^{(t)}) + r_h(H^{(t)}))$$

$$u^{(t)} = \sigma(u_i(X^{(t)}) + u_h(H^{(t)}))$$

$$c^{(t)} = \tanh(c_i(X^{(t)}) + r^{(t)} \times c_h(H^{(t)}))$$

$$H^{(t+1)} = u^{(t)} \times H^{(t)} + (1-u^{(t)}) \times c^{(t)}$$

式中，$r_i(\,)$、$r_h(\,)$、$u_i(\,)$、$u_h(\,)$、、$c_i(\,)$、$c_h(\,)$ 为可训练的线性层。G-GRU 的每个单元对当前输入的信息 $c^{(t)}$ 做选择性的添加，并且对当前时刻之前的历史信息 $H^{(t)}$ 做选择性的遗忘，生成下一帧的状态。

3.1.3　损失函数

我们在所有的输出上应用 L_1 作为损失函数。

$$L = \sum_{i=1}^{N} \left\| X_{1:T_f}^{i} - \hat{X}_{1:T_f}^{i} \right\|_1$$

3.2　实验

3.2.1　数据集

Human 3.6M[19]有 15 个不同的人体动作类别，由 7 个角色（S1、S5、S6、S7、S8、S9 和 S11）执行。每个位姿有 32 个关节，呈指数映射的形式。将它们转换为 3D 坐标和角度表示，并丢弃 10 个多余的关节。S5 和 S11 分别用于测试和验证，其余的角色用于训练。CMU-MoCap 有 8 个人体动作类别。每个位姿包含 38 个关节，也呈指数映射的形式，转换为 3D 坐标和角

度表示，保留 25 个关节，丢弃其他关节。训练数据集和测试数据集的划分与文献[12]相同。

3.2.2　评价指标

本文训练和测试位姿的三维坐标表示，展示三维坐标下的量测结果。使用平均每关节位置误差（Mean Per Joint Position Error，MPJPE）作为三维误差的评估指标。MPJPE 经常用作 3D 人体姿态估计算法的评价指标，指标的值越小，说明相应的 3D 人体姿态估计算法的性能越好。MPJPE 的计算公式如下。

$$\mathrm{MPJPE}(X,\hat{X})=\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\left\|p_t^n-\hat{p}_t^n\right\|_2$$

式中，N 是骨架中的关节数；T 是帧的长度；p 和 $\hat{p}$ 分别是预测的关节坐标和实际的关节坐标。对于一组帧，误差是所有帧的 MPJPE 的平均值。

3.2.3　方法对比

将本文所提方法与 Res.Sup[20]、DMGNN[10]、LTD[12]和 MSR[21]在 Human 3.6M 与 CMU-MoCap 这两个数据集上进行比较。其中，Res. Sup 是一种早期的基于 RNN 的方法；DMGNN 采用 GCN 提取特征，采用 RNN 进行解码；LTD 完全依赖 GCN，在频域内进行预测；MSR 是近年来通过多尺度强化预测的一种方法。

为了公平对比，本文沿用先前一贯使用的比较方法，即对比本文所提方法的结果在 Human 3.6M 与 CMU-MoCap 这两个数据集上的短时间预测结果（未来 80ms、160ms、320ms 和 400ms 4 个时间步）和长时间预测结果（未来 560ms 和 1 000ms 2 个时间步）。对比结果如表 1 和表 2 所示。在 Human 3.6M 数据集上的短时间预测中，本文所提方法在“Smoking”“Discussing”“Walking Dog”“Walking”中均有较大优势；在长时间预测中，本文所提方法在“Posing”“Walking Together”“Walking Dog”中有较大优势。根据所有动作的 MPJPE，本文所提方法在短时间预测和长时间预测方面都优于先前的方法。

表 1　Human 3.6M 数据集中各方法的短时间预测结果对比

Tab. 1　Comparison of short time prediction results of various methods in Human 3.6M dataset

运动时间/ms	Walking				Eating				Smoking				Discussing			
	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Res.Sup	29.4	50.8	76	81.5	16.8	30.6	56.9	68.7	23	42.6	70.1	82.7	32.9	61.2	90.9	96.2
DMGNN	17.3	30.7	54.6	65.2	11	21.4	36.2	43.9	9	17.6	32.1	40.3	17.3	34.8	61	69.8
LTD	12.3	23	39.8	46.1	8.4	16.9	33.2	40.7	7.9	16.2	31.9	38.9	12.5	27.4	58.5	71.7
MSR	12.2	22.7	38.6	45.2	8.4	17.1	33	40.4	8	16.3	31.3	38.2	12	26.8	57.1	69.7
本文所提方法	**11.4**	**21.6**	**38.3**	**44.4**	**8.5**	**16.7**	**32.8**	**39.0**	**7.7**	**15.1**	**29.9**	**35.3**	**11.3**	**25.1**	**55.8**	**67.7**
运动时间/ms	Directions				Greeting				Phoning				Posing			
	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Res.Sup	35.4	57.3	76.3	87.7	34.5	63.4	124.6	142.5	38	69.3	115	126.7	36.1	69.1	130.5	157.1
DMGNN	13.1	24.6	64.7	81.9	23.3	50.3	107.3	132.1	12.5	25.8	48.1	58.3	15.3	29.3	71.5	96.7
LTD	9	19.9	43.4	53.7	18.7	38.7	77.7	93.4	10.2	21	42.5	52.3	13.7	29.9	66.6	84.1
MSR	8.6	19.7	43.3	53.8	16.5	37	77.3	93.4	10.1	20.7	41.5	51.3	12.8	29.4	67	85
本文所提方法	**8.6**	**19.0**	**42.9**	**52.2**	**16.8**	**36.1**	**74.6**	**88.9**	**9.5**	**19.4**	**40.6**	**49.1**	**11.9**	**27.1**	**63.1**	**79.1**
运动时间/ms	Purchases				Sitting				Sitting Down				Taking Photo			
	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Res.Sup	36.3	60.3	86.5	95.9	42.6	81.4	134.7	151.8	47.3	86	145.8	168.9	26.1	47.6	81.4	94.7
DMGNN	21.4	38.7	75.7	92.7	11.9	25.1	44.6	50.2	15	32.9	77.1	93	13.6	29	46	58.8
LTD	15.6	32.8	65.7	79.3	10.6	21.9	46.3	57.9	16.1	31.1	61.5	75.5	9.9	20.9	45	56.6
MSR	14.8	32.4	66.1	79.6	10.5	22	46.3	57.8	16.1	31.6	62.5	76.8	9.9	21	44.6	56.3
本文所提方法	**13.6**	**29.7**	**62.0**	**74.2**	**9.7**	**19.9**	**44.2**	**54.7**	**14.5**	**28.4**	**58.8**	**71.9**	**9.3**	**19.4**	**43.3**	**53.7**

续表

运动时间/ms	Waiting				Walking Dog				Walking Together				Average			
	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Res.Sup	30.6	57.8	106.2	121.5	64.2	102.1	141.1	164.4	26.8	50.1	80.2	92.2	34.7	62	101.1	115.5
DMGNN	12.2	24.2	59.6	77.5	47.1	93.3	160.1	171.2	14.3	26.7	50.1	63.2	17	33.6	65.9	79.7
LTD	11.4	24	50.1	61.5	23.4	46.2	83.5	96	10.5	21	38.5	45.2	12.7	26.1	52.3	63.5
MSR	10.7	23.1	48.3	59.2	20.7	42.9	80.4	93.3	10.6	20.9	37.4	43.9	12.1	25.6	51.6	62.9
本文所提方法	**10.2**	**21.6**	**46.3**	**56.1**	**19.9**	**40.5**	**75.7**	**86.8**	**9.8**	**19.8**	**36.8**	**42.2**	**11.5**	**24.0**	**49.7**	**59.6**

注：粗体表示最好的结果。

表 2　Human 3.6M 数据集中各方法的长时间预测结果对比

Tab. 2　Comparison of long-term prediction results of various methods in Human 3.6M dataset

运动时间/ms	Walking		Eating		Smoking		Discussion		Directions		Greeting		Phoning		Posing	
	560	1 000	560	1 000	560	1 000	560	1 000	560	1 000	560	1 000	560	1 000	560	1 000
Res.Sup	81.7	100.7	79.9	100.2	94.8	137.4	121.3	161.7	110.1	152.5	156.1	166.5	141.2	131.5	194.7	240.2
DMGNN	73.4	95.8	58.1	86.7	50.9	72.2	81.9	138.3	110.1	115.8	152.5	157.7	78.9	98.6	163.9	310.1
LTD	54.1	59.8	53.4	77.8	50.7	72.6	91.6	121.5	71	101.8	115.4	148.8	69.2	103.1	114.5	173
MSR	52.7	63	52.5	77.1	49.5	71.6	88.6	117.6	71.2	100.6	116.3	147.2	68.3	104.4	116.3	174.3
本文所提方法	**53.2**	**60.8**	**53.4**	**76.3**	**48.3**	**68.7**	**90.0**	**119.0**	**71.7**	**100.4**	**113.4**	**145.6**	**68.0**	**102.3**	**110.9**	**166.7**

运动时间/ms	Purchases		Sitting		Sitting Down		Taking Photo		Waiting		Walking Dog		Walking Together		Average	
	560	1 000	560	1 000	560	1 000	560	1 000	560	1 000	560	1 000	560	1 000	560	1 000
Res.Sup	122.7	160.3	167.4	201.5	205.3	277.6	117	143.2	146.2	196.2	191.3	209	107.6	131.1	97.6	130.5
DMGNN	118.6	153.8	60.1	104.9	122.1	168.8	91.6	120.7	106	136.7	194	182.3	83.4	115.9	103	137.2
LTD	102	143.5	78.3	119.7	100	150.2	77.4	119.8	79.4	108.1	111.9	148.9	55	65.6	81.6	114.3
MSR	101.6	139.2	78.2	120	102.8	155.5	77.9	121.9	76.3	106.3	111.9	148.2	52.9	65.9	81.1	114.2
本文所提方法	**97.7**	**136.1**	**77.3**	**117.0**	**99.1**	**148.4**	**76.3**	**117.5**	**75.9**	**105.0**	**105.8**	**138.5**	**52.9**	**61.2**	**79.6**	**111.9**

注：粗体表示最好的结果。

表 3 展示了各方法在 CMU-MoCap 数据集上的短时间预测结果和长时间预测结果对比。由于篇幅限制，表中仅展示每个时间步的平均误差。结果表明，本文所提的方法在 CMU-MoCap 数据集上也有一定的优势。

表 3　CMU-MoCap 数据上各方法在所有时间步的预测结果对比

Tab. 3　Comparison of prediction results of various methods in CMU-MoCap dataset

运动时间/ms	80	160	320	400	560	1 000
Res.Sup	4	43	74.5	87.2	105.5	136.3
DMGNN	13.6	24.1	47	58.8	77.4	112.6
LTD	9.3	17.1	33	40.9	55.8	86.2
MSR	8.1	15.2	30.6	38.6	53.7	83
本文所提方法	**7.9**	**14.5**	**29.6**	**37.8**	**51.9**	**81.2**

注：粗体表示最好的结果。

4　结论

本文基于编码器-解码器框架提出了一个时频域结合的特征提取网络，对观测到的人体骨骼姿态序列进行编码，并通过沿空间轴的自控门注意力机制对运动中的人体骨骼的各个关节提取依赖信息，强化人体骨骼不同关节之间的相互依赖关系。将提取到的强化特征通过 G-GRU 模块进行递归解码，逐帧输出预测的动作序列。本文在 Human 3.6M 数据集上测试了各方法的预测性能，并在 CMU-Mocap 数据集上测试了各方法的泛化能力。大量的实验结果表明，本文所提方法在大多数情况下优于先前的方法。

参考文献

[1] KOPPULA H, SAXENA A. Learning spatio-temporal structure from rgb-d videos for human activity detection and anticipation[C]. International Conference on Machine Learning. PMLR, 2013: 792-800.

[2] KOPPULA H S, SAXENA A. Anticipating human activities using object affordances for reactive robotic response[J]. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 2015, 38(1): 14-29.

[3] GUI L Y, ZHANG K, WANG Y X, *et al*. Teaching robots to predict

human motion[C]. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018: 562-567.

[4] HUANG D A, KITANI K M. Action-reaction: forecasting the dynamics of human interaction[C]. European Conference on Computer Vision. Springer, Cham, 2014: 489-504.

[5] CHEN S, LIU B, FENG C, *et al*. 3D point cloud processing and learning for autonomous driving: impacting map creation, localization, and perception[J]. **IEEE Signal Processing Magazine**, 2020, 38(1): 68-86.

[6] ALAHI A, GOEL K, RAMANATHAN V, *et al*. Social lstm: human trajectory prediction in crowded spaces[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 961-971.

[7] GUPTA A, MARTINEZ J, LITTLE J J, *et al*. 3D pose from motion for cross-view action recognition via non-linear circulant temporal encoding[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 2601-2608.

[8] APRATIM B, MARIO F, BERNT S. Long-term on-board prediction of people in traffic scenes under uncertainty[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4194-4202.

[9] FRAGKIADAKI K, LEVINE S, FELSEN P, *et al*. Recurrent network models for human dynamics[C]. Proceedings of the IEEE International Conference on Computer Vision, 2015: 4346-4354.

[10] LI M, CHEN S, ZHAO Y, *et al*. Dynamic multiscale graph neural networks for 3D skeleton based human motion prediction[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 214-223.

[11] GUO X, CHOI J. Human motion prediction via learning local structure representations and temporal dependencies[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019: 33(1): 2580-2587.

[12] MAO W, LIU M, SALZMANN M, *et al*. Learning trajectory dependencies for human motion prediction[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 9489-9497.

[13] JAIN A, ZAMIR A R, SAVARESE S, *et al.* Structural-RNN: deep learning on spatio-temporal graphs[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 5308-5317.

[14] LI C, ZHANG Z, LEE W S, *et al*. Convolutional sequence to sequence model for human dynamics[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 5226-5234.

[15] AKSAN E, KAUFMANN M, CAO P, *et al*. A spatio-temporal transformer for 3D human motion prediction[C]. 2021 International Conference on 3D Vision (3DV), IEEE, 2021: 565-574.

[16] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need[J]. **Advances in Neural Information Processing Systems**, 2017, 30.

[17] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132-7141.

[18] MA T, NIE Y, LONG C, *et al*. Progressively generating better initial guesses towards next stages for high-quality human motion prediction[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 6437-6446.

[19] IONESCU C, PAPAVA D, OLARU V, *et al*. Human 3.6M: 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.

[20] MARTINEZ J, BLACK M J, ROMERO J. On human motion prediction using recurrent neural networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 2891-2900.

[21] DANG L, NIE Y, LONG C, *et al*. MSR-GCN: multi-scale residual graph convolution networks for human motion prediction[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 11467-11476.

[22] 王雨凡，凌芝，蒋子昂，等. 人体运动力学分析方法的回顾与展望[J]. **生物医学工程学进展**，2023，44（1）: 1-26.
WANG Yufan, LING Zhi, JIANG Ziang, *et al*. Review and prospect of analysis methods in human sports biomechanics[J]. **Progress in Biomedical Engineering**, 2023, 44(1): 1-26.

[23] 杜妍辰. 一种基于表面肌电信号映射人体下肢运动意图的方法[J]. **生物医学工程学进展**，2023，44（2）: 158-162.
DU Yanchen. A Method of Human Lower Limb Motion Intention Mapping Based on Surface Electromyography[J]. **Progress in Biomedical Engineering**, 2023, 44(2): 158-162.