动态多尺度图神经网络在基于 3D 骨骼的人体运动预测中的应用（Part I）

Maosen Li[1], Siheng Chen[2], Yangheng Zhao[1], Ya Zhang[1], Yanfeng Wang[1] 以及 Qi Tian[1]

[1] 联合上海交通大学 Medianet 创新中心
[2] 三菱电子研究实验室

{maosen li, zhaoyangheng-sjtu, ya zhang, wangyanfeng} @sjtu.edu.cn, schen@merl.com, wywqtian@gmail.com

摘要

我们提出了新型的动态多尺度图神经网络（*DMGNN*）来预测基于三维骨骼的人体运动。*DMGNN* 的核心思想是利用多尺度图全面建模人体的内部关系进行运动特征学习。这个多尺度图在训练过程中是自适应的，并且是跨网络层的动态图。基于这个图，我们提出了一个多尺度图计算单元（*MGCU*）来提取各个尺度的特征，并进行跨尺度的特征融合。整个模型是动作类别无关的，并遵循一个编码器-解码器框架。编码器由一系列的 *MGCU* 组成，用于学习运动特征。解码器使用提出的基于图的门递归单元来生成未来的姿势。广泛的实验表明，所提出的 *DMGNN* 网络在 *Human 3.6M* 和 *CMU Mocap* 数据集上的短期和长期预测中都优于最先进的方法。我们进一步研究了学习的多尺度图的可解释性。这些代码可以从 *https://github.com/limaosen0/DMGNN* 下载。

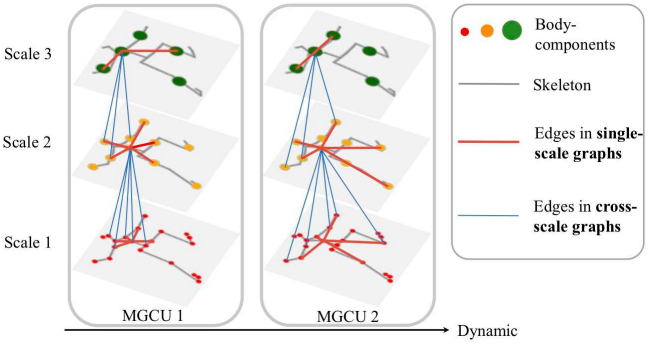

图 1: 在"Posing" 上的两种学习的多尺度图。我们显示了在单一尺度和跨尺度中与躯干相关的强关系。两个多尺度图从一个 MGCU 到另一个 MGCU 是动态的，分别捕捉了局部和远处的关系。

1 简介

基于三维骨架的人体运动预测基于人体骨骼给定过去的运动，预测未来的姿势。运动预测有助

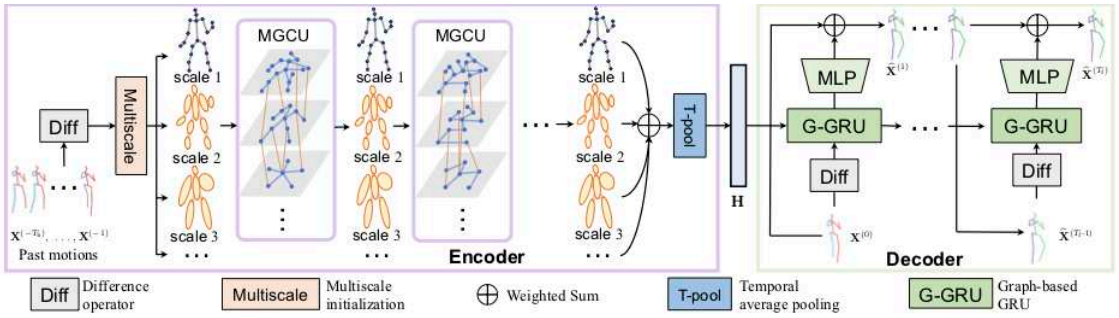

图 2: DMGNN 的架构。DMGNN 的架构，它使用编码器-解码器框架进行运动预测。在编码器中，级联多尺度图计算块（MGCU）利用动态多尺度图来提取时空特征。在解码器中，我们提出了一种基于图的 GRU（G-GRU）来预测姿势。

于机器理解人类的行为，引起了相当大的关注 [9, 20, 32, 5, 12, 2]。相关技术可广泛应用于许多计算机视觉和机器人场景，如人机交互 [24, 23, 17, 13]，自动驾驶 [6]，以及行人跟踪 [1, 15, 3]。

许多方法，包括传统的基于状态的方法 [25, 44, 38, 37, 36] 和基于深度网络的方法 [9, 32, 10, 7, 12, 14, 11, 33, 43]，已经被提出来实现有希望的运动预测。然而，大多数方法并没有明确地利用不同身体组件之间的关系或约束，而这些关系或约束携带着运动预测的关键信息。最近的一项工作 [31] 建立了跨身体关节的图，用于对关系建模；然而，这样的图仍然不足以反映身体关节的功能组。另一项工作 [43] 建立了预定义的结构来聚合身体-关节的特征，以表示固定的身体-部件，但这个模型只考虑了身体的物理约束，没有利用动作协调和关系。例如，”行走”这个动作往往是基于抽象的手臂和腿部的协作运动来理解，而不是手指和脚趾的详细位置。

为了对更全面的关系进行建模，我们提出了一种新的人体表示方法：多尺度图，其节点是不同尺度的身体部件，边缘是部件之间的配对关系。为了在多个尺度上对人体进行建模，多尺度图由两类子图组成：单尺度图，连接同一尺度上的身体部件；跨尺度图，连接跨越两个尺度的身体部件；见图 1。单尺度图共同提供了一个金字塔式的身体骨架表示。

每一个交叉尺度图是一个两段式图形，将一个单尺度图连接到另一个单尺度图。例如，一个粗尺度图中的”手臂”节点可以连接到细尺度图中的”手”和”肘”节点。这个多尺度图由预定义的物理连接初始化，并在训练中自适应地调整为运动敏感。总的来说，这种多尺度的表示方式为身体关系建模提供了一种新的势能。

基于多尺度图，我们提出了一种新的模型，称为动态多尺度图神经网络 (DMGNN)，它是动作类别无关的，并遵循一个编码器-解码器框架来学习用于预测的运动表示。编码器包含一个多尺度图计算单元（MGCU）的级联，其中每个单元与一个多尺度图相关联。一个 MGCU 包括两个关键组件：单尺度图形卷积块（SS-GCB），利用单尺度图形精确各个尺度的特征，以及跨尺度融合块（CS-FB），推断跨尺度图形将特征从一个尺度转换

到另一个尺度，实现跨尺度融合。多尺度图具有自适应和可训练的内置拓扑结构；由于拓扑结构从一个 MGCU 到另一个 MGCU 是变化的，所以它也是动态的；见图 1 中学习的动态多尺度图。值得注意的是，CS-FB 块中的跨尺度图是根据输入运动自适应构建的，并且反映了辨别性的运动模式，用于类别无关的预测。

至于解码器，我们采用了一个基于图的门控递归单元（G-GRU）来依次给定最后估计的姿势产生预测。G-GRU 利用可训练的图来进一步增强状态传播。我们还使用残差连接来稳定预测。为了学习更丰富的运动动态，我们引入差分算子来提取多阶运动差异作为位置、速度和加速度的代理。DMGNN 的架构如图 2 所示。

为了验证我们 DMGNNN 的优越性，我们在两个大规模的数据集上进行了广泛的实验：Human 3.6M[19] 和 CMU Mocap1。实验结果表明，我们的模型在短期和长期预测方面的效果和效率都优于大多数最先进的作品。本文的主要贡献如下：

- 我们提出了动态多尺度图神经网络 (DMGNN) 来提取多尺度的深度特征，并实现有效的运动预测。
- 我们提出了两个关键组件：一个多尺度图计算单元，利用多尺度图来提取和融合多尺度的特征，以及一个基于图的 GRU 来增强姿势生成的状态传播；以及—
- 我们进行了大量的实验，表明所提出的 DMGNNN 在两个大型数据集上的短期和长期运动预测方面优于大多数最先进的方法。我们进一步将学习到的图形可视化，以实现可解释性和合理性。

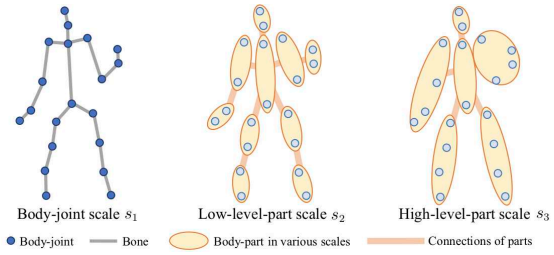

图 3: Human 3.6M 上的三种身体尺度。在 s1 中，我们考虑 20 个关节的非零指数图 [18]；在 s2 和 s3 中，我们分别考虑 10 个和 5 个部分。

2 相关工作

人的运动预测：为了预测运动，开发了一些传统方法，如隐藏马尔科夫模型 [25]、高斯过程 [44] 和随机森林 [25]。最近，深度网络正发挥着越来越关键的作用：一些基于递归网络的模型逐步生成了未来的姿势 [9, 20, 32, 41, 45, 11, 30, 12, 28]；一些前馈网络 [26, 31] 试图减少误差积累以实现稳定的预测；模仿学习算法也被提出 [42]。但这些方法很少考虑到不同尺度的关系，而这些关系又承载了人类行为理解的全面信息。在本工作中，我们构建动态多尺度图来捕捉丰富的多尺度关系，并提取灵活的语义用于运动预测。

图形深度学习：图，表达与非网格结构相关的数据，保存内部节点之间的依赖关系 [46, 40, 39]。许多研究集中在图表示学习和相对的应用上 [29, 8, 22, 16, 46, 35]。基于固定的图结构，以前的工作探索了根据图谱域 [8, 22] 或图顶点域 [16] 来传播节点特征。一些基于图的模型已经被用于基于骨架的动作识别 [46, 27, 34]，运动预测 [31] 和 3D 姿势估计 [47]；与之前的任何工作不同，我们的模型考虑了多尺度图和相应的操作。

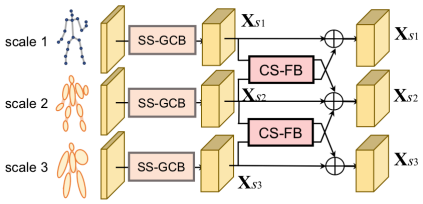

图 4: 一个 MGCU 使用单尺度图卷积块（SS-CB）跨尺度融合块（CS-FB）。

3 问题提出

假设历史上基于三维骨骼的姿势是 $\mathbb{X}_{-T_h:0} = [\mathbf{X}^{(-T_h)}, ..., \mathbf{X}^{(0)}] \in \mathbb{R}^{M\times(T_h+1)\times D_x}$，而未来的姿势是 $\mathbb{X}_{1:T_f} = [\mathbf{X}^{(1)}, ..., \mathbf{X}^{(T_f)}] \in \mathbb{R}^{M\times T_h\times D_{\mathbf{x}}}$，其中 $\mathbf{X}^{(t)} \in \mathbb{R}^{M\times D_{\mathbf{x}}}$，含有 M 个关节，$D_{\mathbb{X}}$ 为 3 个特征维度，描绘了 t 时刻的 3D 姿势。运动预测的目标是给定过去观察到的姿势，生成未来的姿势；在数学上，我们需要提出一个模型 $\mathcal{F}_{pred}(\cdot)$ 来预测 $\hat{\mathbb{X}}_{1:T_f} = \mathcal{F}_{pred}(\mathbb{X}_{-T_h:0})$，其中 $\hat{\mathbb{X}}_{1:T_f}$ 是接近目标 $\mathbb{X}_{1:T_f}$ 的预测的运动。

为了利用丰富的身体关系，我们将一个身体表示为一个跨多尺度身体组件的多尺度图。从理论上讲，我们可以使用任意数量的尺度。基于人类的天性，我们特别采用了 3 个尺度：身体-关节尺度、低级-部件尺度和高级-部件尺度。为了初始化多尺度体图，我们根据人性的先验，将空间上相邻的关节合并到更粗的尺度，见图 3。有了多尺度图，我们提出动态多尺度图神经网络（DMGNN），以端到端的方式预测未来的姿势。

4 关键部分

为了构建我们的动态多尺度图神经网络 (DMGNN)，我们考虑了三个基本组成部分：多尺度图计算单元 (MGCU)、基于图的 GRU(G-GRU) 和差分算子。

4.1 多尺度图计算单元 (MGCU)

MGCU 的功能是基于多尺度图提取和融合多个尺度的特征，并对其进行自适应和单独训练。一个 MGCU 包括两种类型的构建块：单尺度图卷积块，它利用单尺度图来提取每个尺度的特征；跨尺度融合块，它利用跨尺度图将特征从一个尺度转换到另一个尺度，并实现跨尺度的有效融合；见图 4。现在我们详细介绍一下各个块。

单尺度图卷积块（SS-GCB）。为了提取每个尺度的时空特征，我们提出了一种单尺度图卷积块（SS-GCB）。让单尺度图在尺度 s 处的可训练邻接矩阵为 $\mathbf{A}_s \in \mathbb{R}^{M_s\times M_s}$，其中 M_s 为主体成分的数量。$\mathbf{A}_s$ 首先由一个骨架图初始化，其节点为身体部件，边缘为物理连接，建模为物理约束的先验；见图 3。在训练过程中，$\mathbf{A}_s$ 中的每个元素都会进行自适应调整，以捕捉灵活的身体关系。

基于单尺度图，SS-GCB 通过两个步骤有效地提取深度特征。1）图卷积提取身体成分的空间特征；2）时空卷积提取运动序列的时空特征。设规模 s 的输入特征为 $\mathbf{X}_s \in \mathbb{R}^{M_s\times D_x}$，空间图卷积公式为

$$\mathbf{X}_{s,sp} = \mathrm{ReLU}(\mathbf{A}_s\mathbf{X}_s\mathbf{W}_s + \mathbf{X}_s\mathbf{U}_s) \in \mathbb{R}^{Ms\times D'_{\mathbf{x}}}, \quad (1)$$

其中 $\mathbf{W}_s, \mathbf{U}_s \in \mathbb{R}^{D_{\mathbf{x}}\times D'_{\mathbf{x}}}$ 为可训练参数。通过 (1)，我们从相关的身体成分中提取空间特征。每个 SS-GCB 中的 $\mathbf{A}_s$ 是单独训练的，并且在测试过程中保持固定。为了捕捉沿时间的运动，我们对特征序列进行了时间卷积。不同 SS-GCBs 中的单尺度图是动态的，显示出灵活的关系。需要注意的是，在不同尺度下提取的特征具有不同的维度，反映的

信息具有不同的接受场。

跨尺度融合块 (CS-FB)。为了实现跨尺度的信息扩散，我们提出了一种跨尺度融合块 (CS-FB)，它使用跨尺度图将一个尺度的特征转换到另一个尺度。跨尺度图是指将一个单尺度图中的节点与另一个单尺度图中的节点对应起来的二元图。例如，低级部分尺度 s_2 中的" 手臂" 节点的特征可以潜在地指导身体关节尺度 s_1 中的" 手" 节点的特征学习。我们旨在从数据中自适应地推断这种跨尺度图。这里我们以 s_1 到 s_2 的 CS-FB 为例进行介绍。

5 结论

Part II 待续……

参考文献

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Feifei Li, and Silvio Savarese. Sociallstm: Human trajectory prediction in crowded spaces. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 961–971, June 2016.

[2] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pages 1531–1540, June 2018.

[3] Apratim Bhattacharyya, Mario Fritz, and Bernt Schiele. Long-term on-board prediction of people in traffic scenes under uncertainty. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4194–4202, June 2018.

[4] Lon Bottou. Large-scale machine learning with stochastic gradient descent. In International Conference on Computational Statistics (COMPSTAT), pages 177–187, August 2010.

[5] Judith Butepage, Michael Black, Danica Kragic, and Hedvig Kjellstrom. Deep representation learning for human motion prediction and classification. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages

[6] Siheng Chen, Baoan Liu, Chen Feng, Carlos VallespiGonzalez, and Carl Wellington. 3d point cloud processing and learning for autonomous driving. IEEE Signal Processing Magazine Special Issue on Autonomous Driving, 2020.

[7] Hsukuang Chiu, Ehsan Adeli, Borui Wang, DeAn Huang, and Juan Niebles. Action-agnostic human pose forecasting. CoRR, abs/1810.09676, 2018.

[8] Michael Defferrard, Xavier Bresson, and Pierre Van-dergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In Advances in Neural Information Processing Systems (NeurIPS), pages 3844–3852, December 2016.

[9] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In The IEEE International Conference on Computer Vision (ICCV), pages 4346–4354, December 2015.

[10] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. Learning human motion models for long-term predictions. CoRR, abs/1704.02827, 2017.

[11] Anand Gopalakrishnan, Ankur Mali, Dan Kifer, Lee Giles, and Alexander Ororbia. A neural temporal model for human motion prediction. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 12116–12125, June 2019.

[12] Liangyan Gui, Yuxiong Wang, Xiaodan Liang, and Jose Moura. Adversarial geometry-aware human motion prediction. In The European Conference on Computer Vision (ECCV), pages 786–803, September 2018.

[13] Liangyan Gui, Kevin Zhang, Yuxiong Wang, Xiaodan Liang, Jose Moura, and Manuela Veloso. Teaching robots to predict human motion. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), October 2018.

[14] Xiao Guo and Jongmoo Choi. Human motion prediction via learning local structure representations and temporal dependencies. In AAAI Conference on Artificial Intelligence, February 2019.

[15] Ankur Gupta, Julieta Martinez, James Little, and Robert Woodham. 3d pose from motion for cross-view action recognition via nonlinear circulant temporal encoding. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2061–2068, June 2014.

[16] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In Advances in Neural Information Processing Systems (NeurIPS), pages 1024–1034, December 2017.

[17] Dean Huang and Kris Kitani. Action-reaction: Forecasting the dynamics of human interaction. In The European Conference on Computer Vision (ECCV), pages 489–504, July 2014.

[18] Du Huynh. Metrics for 3d rotations: Comparison and analysis. Journal of Mathematical Imaging and Vision, 35(2):155–164, October 2009.

[19] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 36(7):1325–1339, July 2014.

[20] Ashesh Jain, Amir Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5308–5317, June 2016.

[21] Diederik Kingma and Jimmylei Ba. Adam: A method for stochastic optimization. In International Conference on Learning Representations (ICLR), pages 1–15, May 2015.

[22] Thomas Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In International Conference on Learning Representations (ICLR), pages 1–14, April 2017.

[23] Hema Koppula and Ashutosh Saxena. Learning spatiotemporal structure from rgb-d videos for human activity detection and anticipation. In International Conference on Machine Learning (ICML), pages 792–800, June 2013.

[24] Hema Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 38(1):14–29, January 2016.

[25] Andreas Lehrmann, Peter Gehler, and Sebastian Nowozin. Efficient nonlinear markov models for human motion. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1314–1321, June 2014.

[26] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5226–5234, June 2018.

[27] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3595–3603, June 2019.

[28] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. CoRR, abs/1910.02212, 2019.

[29] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. In International Conference on Learning Representations (ICLR), pages 1–20, May 2016.

[30] Zhenguang Liu, Shuang Wu, Shuyuan Jin, Qi Liu, Shijian Lu, Roger Zimmermann, and Li Cheng. Towards natural and accurate future motion prediction of humans and animals. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 10004–10012, June 2019.

[31] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In The IEEE International Conference on Computer Vision (ICCV), October 2019.

[32] Julieta Martinez, Michael Black, and Javier Romero. On human motion prediction using recurrent neural networks. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4674–4683, July 2017.

[33] Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion. In British Machine Vision Converence (BMVC), pages 1–14, September 2018.

[34] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with directed graph neural networks. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 7912–7921, June 2019.

[35] Chenyang Si, Ya Jing, Wei Wang, Liang Wang, and Tieniu Tan. Skeleton-based action recognition with spatial reasoning and temporal stack learning. In The European Conference on Computer Vision (ECCV), pages 103–118, September 2018.

[36] Ilya Sutskever, Geoffrey Hinton, and Graham Taylor. The recurrent temporal restricted boltzmann machine. In Advances in Neural Information Processing Systems (NeurIPS), pages 1601–1608, December 2009.

[37] Graham Taylor and Geoffrey Hinton. Factored conditional restricted Boltzmann machines for modeling motion style. In International Conference on Machine Learning (ICML), pages 1025–1032, June 2009.

[38] Graham Taylor, Geoffrey Hinton, and Sam Roweis. Modeling human motion using binary latent variables. In Advances in Neural Information Processing Systems (NeurIPS), pages 1345–1352, December 2007.

[39] Diego Valsesia, Giulia Fracastoro, and Enrico Magli. Learning localized generative models for 3d point clouds via graph convolution. In International Conference on Learning Representations (ICLR), pages 1–15, May 2019.

[40] Nitika Verma, Edmond Boyer, and Jakob Verbeek. Feastnet: Feature-steered graph convolutions for 3d shape analysis. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2598–2606, June 2018.

[41] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In The IEEE International Conference on Computer Vision (ICCV), pages 3332–3341, October 2017.

[42] Borui Wang, Ehsan Adeli, Hsukuang Chiu, Dean Huang, and JuanCarlos Niebles. Imita-

tion learning for human pose prediction. In The IEEE International Conference on Computer Vision (ICCV), October 2019.

[43] He Wang, Edmond Ho, Hubert Shum, and Zhanxing Zhu. Spatio-temporal manifold learning for human motions via long-horizon modeling. IEEE Transactions on Visualization and Computer Graphics (TVCG), PP(99), August 2019.

[44] Jack Wang, Aaron Hertzmann, and David Fleet. Gaussian process dynamical models. In Advances in Neural Information Processing Systems (NeurIPS), pages 1441–1448, December 2006.

[45] Tianfan Xue, Jiajun Wu, Katherine Bouman, and Bill Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In Advances in Neural Information Processing Systems (NeurIPS), pages 91–99. December 2016.

[46] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In AAAI Conference on Artificial Intelligence (AAAI), pages 7444–7452, February 2018.

[47] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas. Semantic graph convolutional networks for 3d human pose regression. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3425–3435, June 2019.