Motion Prediction using Trajectory Cues

Zhenguang Liu
Zhejiang University
Hangzhou, Zhejiang, China
liuzhenguang2008@gmail.com

Pengxiang Su
Jilin University
Changchun, Jilin, China
supx19@mails.jlu.edu.cn

Shuang Wu
Nanyang Technological University
50 Nanyang Ave, Singapore
wushuang@outlook.sg

Xuanjing Shen
Jilin University
Changchun, Jilin, China
xjshen@jlu.edu.cn

Haipeng Chen
Jilin University
Changchun, Jilin, China
chenhp@jlu.edu.cn

Yanbin Hao
University of Science and Technology of China
Hefei, Anhui, China
haoyanbin@hotmail.com

Meng Wang
Hefei University of Technology
Hefei, Anhui, China
eric.mengwang@gmail.com

Abstract

Predicting human motion from a historical pose sequence is at the core of many applications in computer vision. Current state-of-the-art methods concentrate on learning motion contexts in the pose space, however, the high dimensionality and complex nature of human pose invoke inherent difficulties in extracting such contexts. In this paper, we instead advocate to model motion contexts in the joint trajectory space, as the trajectory of a joint is smooth, vectorial, and gives sufficient information to the model. Moreover, most existing methods consider only the dependencies between skeletal connected joints, disregarding prior knowledge and the hidden connections between geometrically separated joints. Motivated by this, we present a semi-constrained graph to explicitly encode skeletal connections and prior knowledge, while adaptively learn implicit dependencies between joints.

We also explore the applications of our approach to a range of objects including human, fish, and mouse. Surprisingly, our method sets the new state-of-the-art performance on 4 different benchmark datasets, a remarkable highlight is that it achieves a 19.1% accuracy improvement over current state-of-the-art in average. To facilitate future research, we have released our code at `https://github.com/Pose-Group/MPT`.

1. Introduction

The ability for machines to anticipate and model human motion dynamics is very much coveted [29] in a wide range of applications such as autonomous driving, human tracking, and regulating the response of a robot when interacting with humans. As a result, future motion prediction has attracted considerable attention in the past decade [9, 41, 46, 6, 39].

Whereas existing methods achieve accurate prediction for a few immediate future frames, it is still difficult to expect accurate and natural forecasting in the long-term since the information hidden in the conscious activity of a human is complex and high-dimensional [15]. To tackle the challenge, we seek to reduce the complexity of motion context modeling at the base level, *i.e. representation space level*, and capture long range dependencies to yield accurate and natural prediction on both short-term and long-term.

Fundamentally, human motion prediction aims to learn a mapping function that bridges the historical skeleton pose sequence to the future pose sequence. Poineering approaches adopt Gaussian Processes [40], Hidden Markov Models [20], and Restricted Boltzmann Machine [38], to predict future human skeleton poses. Unfortunately, these models impose strong assumptions such as Gaussian distributions on the motion dynamics, leading to unsatisfactory results.

Recent approaches explored using different sorts of deep neural networks to address the issue [24, 16, 36, 23, 19, 4, 45, 1, 35]. One line of work [8, 12, 13, 22] utilized recur-

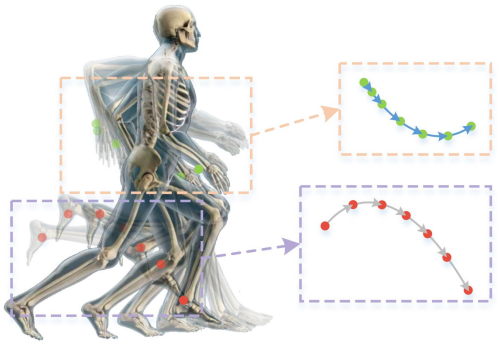

Figure 1. The movement trajectories of the right-wrist and left-ankle joints during running.

rent neural networks (RNNs) and various RNN variants to model motion contexts. Another line of work [7, 32, 43] built upon the success of graph convolutional networks (GCNs) to better characterize the spatial connections between joints. There are also other efforts that adopt Generative Adversarial Networks (GANs) [2, 18] or consider using multiple networks to learn skeletal structures and temporal dynamics [13].

A major shortcoming of prior works is that they tend to learn motion contexts in the intuitive and direct *pose space*. A pose technically translates to the configuration of all joints. Therefore, modeling motion contexts in the pose space implicitly incorporates all the joints, putting motion prediction task on an unnecessarily high-dimensional manifold. A redeeming feature that we may capitalize upon is that forces exerted upon joints during movement generally vary gradually or linearly. Consequently, the movement trajectories of individual joints tend to be smooth, which we may observe from an illustrated example of the *right-wrist* and *left-foot* movements during a running motion. These facts and [32] motivate us to consider modeling motion contexts in the *joint trajectory space*, leveraging the smooth trajectory of a joint to predict its future. A pioneering work [32] converts a joint trajectory to the frequency domain and represents it as discrete cosine transform coefficients. In contrast, we cast the trajectory as the joint position and its (first-order) velocity. This has two key advantages. First, it avoids spectral decomposition (in frequency domain) and thus is not subject to any information loss. Second, by incorporating the velocity, we have a complete characterization of the trajectory configuration space, which is consistent with the Lagrangian formulation of dynamical systems. In addition, we decompose the pose into individual joint trajectories, leveraging the smoothness of the trajectory to predict its future. Compared to inputs set in a structured pose configuration space such FC-GCN [32], SDMTL [26], our proposed trajectories representation has the crucial advantage of being smooth and low dimensional.

Another severe limitation of existing works is that they consider only the connectivity between adjacent joints while ignoring the movement coordination between geometrically separated joints. Dissevering these additional cues results in insufficient context modeling and inaccurate prediction. To address this issue, [32] incorporates dense connections between each pair of joints, [7] engages in a dynamic graph, and [22] adopts a multiscale graph to model the relations. However, the problem is still not efficiently addressed and useful prior knowledge, such as *limb mirror symmetry tendency (e.g., symmetry tendency between two arms) and cross sides synchronization tendency (e.g., synchronization tendency between left arm and right leg)*, are ignored. In this paper, we propose a new graph convolutional network that uses a semi-constrained graph to explicitly encode skeletal connection and useful prior knowledge, while adaptively learn flexible connections between joints. We would like to highlight that the proposed convolutional network has an edge in adopting efficient matrix operations and maintaining constraints that facilitate the training.

Interestingly, most existing methods typically focus on 3D *human* motion prediction. In this paper, we explore applying our approach to a range of objects including human, fish, and mouse. Extensive experiments are conducted on large benchmark datasets including H3.6M and CMU MoCap, and on animal datasets that involve motions of fish and mouse. Empirically, our approach outperforms state-of-the-art methods by a large margin (more than 19.1% accuracy gain) in both short-term and long-term motion predictions. Our code is released, hoping to inspire future research.

Contributions To summarize, the key contributions of this paper are: 1) A new motion representation is proposed, which models motion contexts in the trajectory space instead of the traditional pose space. 2) A semi-constrained graph convolution network is presented to comprehensively learn the relationships between joints, which simultaneously considers skeletal connection, prior knowledge, and implicit dependencies between joints. 3) Our method sets the new state-of-the-art, is applicable to a range of objects, and provides more interesting insights overall.

2. Related Work

Human Motion Prediction Traditional methods tackle the human motion prediction task by utilizing shallow models such as Gaussian Processes [40], Hidden Markov Models [20], and Restricted Boltzmann Machine [38]. With the success of deep learning in various fields [47, 10, 42, 25, 44, 28, 11], and the availability of large-scale public datasets including Human3.6M [14] and CMU MoCap [5], various deep learning methods have been proposed recently to address this problem, which can be roughly classified into three categories: RNN, GCN, and

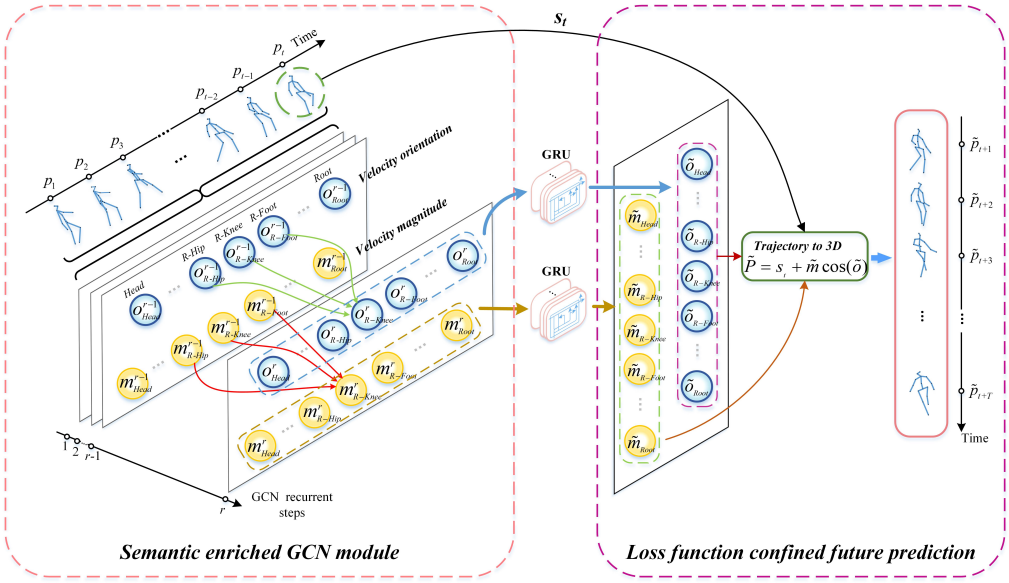

Figure 2. The architecture of the motion context modeling network, which includes a graph semantic enriched GCN module, a GRU layer, and a pose reconstruction block. The GCN module encodes skeletal and known prior connections between joints, and learns implict connections. The GRU layer deals with the sequence data, and pose reconstruction block converts the prediction results to pose space.

GAN based approaches. For instance, [8] proposes an Encoder-Recurrent-Decoder architecture that relies on long short-term memory (LSTM) layers to forecast future human pose. [30] simultaneously models local contexts for individual frames and global contexts for the motion sequence with a hierarchical recurrent network. GAN based approaches [2, 18] try to predict multiple future sequences, while [13] advocates to learn local spatial structure and temporal dynamics using two different models. [32] employs Discrete Cosine Transform to encode temporal information and a feed-forward network to encode the dynamical information. [7] designs a generative model based on Graph Convolutional Network (GCN) and Adversarial learning. [22] builds a dynamic multiscale graph network to extract features at individual scales and fuse features across different scales. [3] introduces a transformer-based architecture to capture the spatial correlations and the temporal smoothness of human motion.

Structural Connection Modeling Human motion is a coordinated movement involving multiple joints. Recently, a set of models attempt to encode spatial dependencies or physical constraints between joints, which contain useful information for prediction. [15] proposes a spatio-temporal graph to explicitly model the structural information of human pose. [30] charactrerizes the pose as a kinematic tree based on the representation of Lie algebra to explicitly model the anatomical constraints. [43] divides human joints into several body parts and constructs a graph to capture joint dependencies. [7] and [32] design novel GCN architectures for capturing spatial dependencies via treating a pose as a generic graph. [22] develops a novel representation for human body, characterizing a body at multiple scales to capture more comprehensive correlations. [3] applies a global attention mechanism and a progressive decoding strategy to extract the long-range structural relations among the joints.

3. Our approach

Problem Definition Presented with a historical pose sequence $P_{0:t} = \langle p_0, p_1, \cdots, p_t \rangle$, we are interested in predicting its future pose sequence $\langle \tilde{p}_{t+1}, \tilde{p}_{t+2}, \cdots, \tilde{p}_{t+T} \rangle$. A pose p_i can be conveniently considered as the 3D coordinates of all body joints.

Method Overview The proposed method MPT (Motion Prediction leveraging Trajectory cues) consists of two key components. (1) MPT casts the historical trajectory of a joint j as its frame-wise velocities and its final (last observed) position. (2) Trajectory cues are then fed into a novel motion context modeling network for future trajectory prediction, which considers rich semantic dependencies between joints. In what follows, we will elaborate the two components, respectively.

3.1. Trajectory Representation

Conventionally, the human posture is described as the 3D coordinates or angles of all joints, then a recurrent neural network is engaged to absorb the historical pose sequence and output the future sequence. This characterizes the pose

of each frame statically and all joints are mixed together, bringing inherent difficulties in extracting motion dynamics. In contrast, a joint trajectory directly conveys temporal motion dynamics of per joint [32], which naturally reduces the complexity of motion context modeling at the base level. Inspired by these facts, we represent the pose sequence in the joint trajectory space.

Formally, given the historical pose sequence $\langle p_0, p_1, \cdots, p_t \rangle$, the trajectory of a joint j can be formulated as:

$$\Gamma = (\upsilon_1, \upsilon_2, \cdots, \upsilon_t, \mathbf{s}_t), \tag{1}$$

where $\upsilon_i \in \mathbb{R}^3$ denotes the position displacement of j between the adjacent i^{th} and $i-1^{th}$ frames, and $\mathbf{s}_t \in \mathbb{R}^3$ is the position of j in the t^{th} frame (the last observed frame). Put differently, $\{\upsilon_i\}_{i=1}^t$ model the frame-wise velocities while $\mathbf{s}_t$ describes the final (last observed) position of j. We further decompose velocity υ_i into velocity magnitude $m_i \in \mathbb{R}$ and velocity orientation $\mathbf{o}_i \in \mathbb{R}^3$. Finally, we arrive at the formulation:

$$\Gamma = (\{m_i\}_{i=1}^t, \{\mathbf{o}_i\}_{i=1}^t, \mathbf{s}_t). \tag{2}$$

Overall, there exist n joints in the human skeleton and the n joints are represented by n historical trajectories.

The proposed trajectory representation in Eq. (2) has the following benefits. (1) Using Eq. (2), we can easily restore the entire joint trajectory with no information loss. Meanwhile, explicitly modeling of velocities and position of a joint leads to a richer motion context for predicting its future. (2) Mathematically, in this problem the Lagrangian corresponds to the product space of joint position and joint velocity, and learning dynamical evolution amounts to solving the Euler-Lagrange equation. Position $\mathbf{s}_t$ corresponds to *potential energy* while velocities υ_i corresponds to *kinetic energy*. By incorporating them, we have a complete characterization of the trajectory configuration space, which is consistent with the Lagrangian formulation of dynamical systems. Empirically, the representation also translates to significantly better performance compared to conventional models.

3.2. Semantic Enriched GCN For Motion Context Modeling and Pose Sequence Prediction

Up to this point, we have discussed reducing the motion prediction problem to extrapolating the trajectories of all joints. However, it is crucial to take into account the interdependence and interaction among these joints when we consider motion. To tackle the challenge, we model the human body as a semi-constrained graph. In particular, to adequately describe the rich spatial dependencies between joints, we explicitly consider three types of joint connections.

(1) Skeletal The natural skeletal connection between joints is obviously meaningful in motion context modeling. We model such connections using the skeletal adjacency matrix A_s. **(2) Motion Prior Knowledge** Most existing methods tend to consider merely the skeletal connections. However, geometrically separated joints may also show strong correlations [7, 32]. For example, the two arms always coordinate each other when clapping, walking, and swimming. Ignoring these valuable prior knowledge may lead to severe performance degradation. Therefore, we explicitly encode these useful prior knowledge in a semantic adjacency matrix A_p. More specifically, in A_p we encode connections between two arms and two legs respectively in consideration of *mirror symmetry tendency*, and connections between a arm (e.g., left arm) and a leg (e.g., right leg) in opposite sides regarding *synchronization tendency*. It is easy to see that the model is extendable to other prior knowledge. **(3) Learnt** Besides fixed connections encoded in A_s and A_p, we parameterize a trainable matrix A_f, which is adaptively tuned to learn flexible and implicit dependencies between joints, providing important complementary connections.

Further, the connection strengths between joints are learned during training instead of being constant, which are captured by a weight matrix W [32]. The diagonal elements in the skeletal adjacency matrix A_s is set to 1 to take account of self-adjacency.

Typically, the operation of a general graph convolutional layer is given by:

$$X^{r+1} = \sigma(\hat{A} X^r M^r) \tag{3}$$

where $X^r \in \mathbb{R}^{n \times l_r}$ and $X^{r+1} \in \mathbb{R}^{n \times l_{r+1}}$ are the features of the r^{th} and $r+1^{th}$ layers, respectively. n is the number of nodes in the graph, which translates to the number of joints in this problem. l_r is the length of joint features at the r^{th} layer. $\sigma(\cdot)$ is an activation function, *e.g.*, ReLU. Matrix $M^{\mathrm{r}} \in \mathbb{R}^{l_r \times l_{l+1}}$ is network parameter (transformation matrix). Filter matrix $\hat{A}$ is computed based on the adjacency matrix A by $\hat{A} = \tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2}$, with $\tilde{A} = A + I$ and $\tilde{D} \in \mathbb{R}^{n \times n}$ being the degree matrix, $\tilde{D}_{i,i} = \sum_j \tilde{A}_{i,j}$.

Similarly, one layer graph convolution in our GCN module is formulated as:

$$X^{r+1} = \sigma\big((A_s + A_p + A_f) \circ W^r X^r M^r\big) \tag{4}$$

where $A_s \in \mathbb{R}^{n \times n}$ and $A_p \in \mathbb{R}^{n \times n}$ encode skeletal connections and prior knowledge connections respectively. Trainable A_f captures implicit joint dependencies. Symbol $\circ$ denotes element-wise product and $W^r \in \mathbb{R}^{n \times n}$ is the trainable connection weight matrix.

Benefits. A_f adaptively extract flexible and implicit connections between joints, while fixed A_s and A_p constrains the training. A_f, A_s, and A_p complement each other, cap-

turing rich joint dependencies. W^r enables learnable connection weights instead of constant ones.

Upon graph convolution, the rich dependencies between joints are considered. Mathematically, the trajectory features of a joint j is updated by incorporating trajectory features of other joints that are correlated to j. The updated frame-wise trajectory features are then passed through a GRU layer, as shown in Fig. 2, to output future trajectories in the form of frame-wise velocities (namely velocity magnitudes $\{m_i\}_{i=t+1}^{t+T}$ and orientations $\{\mathbf{o}_i\}_{i=t+1}^{t+T}$) for all joints. Finally, a simple pose reconstruction block is used to restore 3D poses from the predicted future frame-wise velocities.

Loss functions. We use *weighted trajectory loss* and *bone length loss* to acquire accurate motion prediction. Trajectory loss ensures that the predicted trajectory is consistent with the ground truth. Existing methods such as [32, 7] adopted equal weights for all joints in each predicted frame. This fails to attend to the **spatial aspect** that different joints engage differently in motion and the **temporal aspect** that later predictions rely on earlier predictions. Therefore, we assign higher weights to the joints possessing wider motion ranges and to earlier frames in the prediction. Formally,

$$\mathcal{L}_{Traj} = \frac{1}{n \cdot T} \sum_{p=t+1}^{t+T} \sum_{k=1}^{n} \left\| (\tilde{J}_k^p - J_k^p) \circ \lambda_k^p \right\|_2 \quad (5)$$

where J_k^p denotes the ground truth of the k^{th} joint in the p^{th} frame, while $\tilde{J}_k^p$ denotes the corresponding estimation. J_k^p is represented in the trajectory space by velocity of the k^{th} joint from $p-1^{th}$ frame to p^{th} frame. λ_k^p is the associated weight. Specifically, the spatial weights are designed following kinematic chain configurations while temporal weights decay as prediction goes futher.

Bone length loss enforces the bone length invariance across frames, which can be formulated as:

$$\mathcal{L}_{Bone} = \frac{1}{n \cdot T} \sum_{p=t+1}^{t+T} \sum_{b=1}^{n-1} \left\| (L_b^p - \tilde{L}_b^p) \circ \lambda_b^p \right\|_2 \quad (6)$$

where $\tilde{L}_b^p$ and L_b^p is the estimated and ground truth bone lengths of the b^{th} bone in the p^{th} frame. λ_b^p is the associated weight.

4. Experiments

In this section, we evaluate the proposed method on large benchmark datasets of three distinct articulate objects, namely human, mouse, and fish. We seek to answer the following research questions.

- **RQ1:** How is the proposed method comparing to state-of-the-art motion prediction approaches?
- **RQ2:** How much do different components of MPT contribute to its performance?
- **RQ3:** What interesting insights and findings can we obtain from the empirical results?

Next, we first present the experimental settings, followed by answering the above research questions one by one.

4.1. Experimental Settings

Datasets For human motion prediction, the large benchmark motion capture datasets Human3.6M (H3.6M) [14] and CMU MoCap [5] are engaged. For animal motion prediction, we utilize the public datasets of [30].

Human 3.6M H3.6M dataset is the most widely used and largest public dataset for evaluating human motion prediction methods. It contains 3.6 million 3D poses and videos for 7 subjects, each subject performs 15 different actions, such as eating, sitting, and purchases. Following the data processing schema of prior works [8, 30, 15], we downsample the motion sequence to 25 frames per second (FPS), use 6 subjects (S1, S6, S7, S8, S9, S11) for training, and test with subject 5 (S5).

CMU MoCap The CMU MoCap dataset contains 3D skeletal motion data of 40 objects under multiple infrared cameras. We adopt the same training/test split strategy as [21, 7]. For fair comparison, the sequences are also downsampled to 25 FPS.

Fish and Mouse Datasets The two datasets of [30] contain eight 3D fish pose sequences (50 FPS) and four 3D mouse pose sequences (25 FPS), respectively. In general, the sequence lengths vary from 298 frames to 15,387 frames. We follow [30] for data preprocessing.

Parameter Settings We implemented our methods on PyTorch [34] and experimented on a Nvidia GeForce Titan V GPU. The size of the convolution kernel for semantic enriched GCN is 25×25. The hidden unit size of GRUs is 128. The Adam Optimizer [17] is employed with an initial learning rate of 0.001 which decays by 10% every 10 epochs. Batch size is set to 16 and the gradient clipping is used at a threshold of 5 and trained for 50 epochs. We utilize $t = 10$ $(400ms)$ historical frames as inputs to predict future $T = 25$ $(1,000ms)$ frames in training. In the loss functions of Eqs.(5) & (6), we assign gradually decreasing temporal weights to the predicted frames. The spatial weights for different joints are computed based on their spatial moving ranges, where joints undergoing wider range of motions are assigned higher weights.

4.2. Comparison with Existing Motion Prediction Methods (RQ1)

Human motion prediction We first compare our method with the state-of-the-art approaches on the H3.6M and CMU datasets. The performance of all approaches are evaluated using the widely adopted metric MPJPE (Mean Per Joint Position Error) in millimeter [32, 7, 30], *i.e.*, the

Table 1. Comparisons of position error (in millimeter) for short-term and long-term predictions on H3.6m dataset. Our method consistently outperformance other methods.

	Directions						Eating						Greeting					
Millisecond(ms)	80	160	320	400	560	1,000	80	160	320	400	560	1,000	80	160	320	400	560	1,000
LSTM3LR [8]	36.9	52.1	88.3	102.6	117.6	132.4	34.9	46.8	75.3	83.9	112.7	126.1	27.1	61.8	84.2	98.5	109.7	173.5
Res-GRU [33]	21.6	41.3	72.1	84.1	101.1	129.1	16.8	31.5	53.5	61.7	74.9	98.0	31.2	58.4	96.3	108.8	126.1	153.9
ConSeq2Seq [21]	13.5	29.0	57.6	69.7	86.6	115.8	11.0	22.4	40.7	48.4	61.3	87.1	22.0	45.0	82.0	96.0	116.9	147.3
HMR [30]	23.3	25.0	47.2	61.5	80.9	116.9	9.2	13.9	34.6	47.1	61.3	84.8	12.9	31.9	55.6	82.5	104.3	123.2
FC-GCN [32]	12.6	24.4	48.2	58.4	72.2	105.8	8.8	18.9	39.4	47.2	50.0	74.1	14.5	30.5	74.2	89.0	103.7	140.9
LDR [7]	13.1	23.7	44.5	50.9	—	78.3	7.6	15.9	37.2	41.7	—	53.8	9.6	27.9	66.3	78.8	—	129.7
TrajNet [27]	9.7	22.3	50.2	61.7	84.7	104.2	8.5	18.4	37.0	44.8	59.2	71.5	12.6	28.1	67.3	80.1	91.4	84.3
SDMTL[26]	9.8	23.4	53.8	67.0	88.3	107.9	8.2	16.4	33.8	42.4	53.9	68.8	11.7	25.3	61.9	75.0	88.7	89.0
HRI [31]	7.4	18.4	44.5	56.5	73.9	106.5	6.4	14.0	28.7	36.2	50.0	75.7	13.7	30.1	63.8	78.1	101.9	138.8
Our	**5.6**	**13.1**	**35.9**	**40.4**	**62.7**	**75.1**	**5.3**	**11.4**	**24.5**	**32.9**	**43.6**	**51.4**	**7.3**	**19.6**	**49.3**	**62.7**	**78.1**	**80.3**
	Sitting						Sitting Down						Taking Photo					
Millisecond(ms)	80	160	320	400	560	1,000	80	160	320	400	560	1,000	80	160	320	400	560	1,000
LSTM3LR [8]	34.1	57.1	95.2	111.8	127.4	169.2	37.3	63.3	89.1	121.5	146.6	199.7	25.4	47.9	71.6	74.6	97.3	156.5
Res-GRU [33]	23.8	44.7	78.0	91.2	113.7	152.6	31.7	58.3	96.7	112.0	138.8	187.4	21.9	41.4	74.0	87.6	110.6	153.9
ConSeq2Seq [21]	13.5	27.0	52.0	63.1	82.4	120.7	20.7	40.6	70.4	82.7	106.5	150.3	12.7	26.0	52.1	63.6	84.4	128.1
HMR [30]	12.6	25.6	44.7	60.7	76.4	118.4	9.6	18.6	41.1	57.7	101.7	148.3	7.9	19.0	31.5	57.3	83.5	108.5
FC-GCN [32]	10.7	24.6	50.6	62.0	76.4	115.7	11.4	27.6	56.4	67.6	96.2	142.2	6.8	15.2	38.2	49.6	72.5	116.3
LDR [7]	9.2	23.1	47.2	57.7	—	106.5	9.3	21.4	46.3	59.3	—	144.6	7.1	13.8	29.6	44.2	—	116.4
TrajNet [27]	9.0	22.0	49.4	62.6	81.0	116.3	10.7	28.8	55.1	62.9	79.8	123.8	5.4	13.4	36.2	47.0	73.0	86.6
SDMTL[26]	8.7	22.2	52.2	65.5	83.9	115.5	9.3	23.8	50.6	60.9	77.7	118.9	6.0	14.0	36.1	47.0	67.1	91.1
HRI [31]	9.3	20.1	44.3	56.0	76.4	115.9	14.9	30.7	59.1	72.0	97.0	143.6	8.3	18.4	40.7	51.5	72.1	115.9
Our	**7.2**	**16.4**	**40.7**	**49.8**	**73.2**	**98.5**	**7.6**	**16.9**	**38.6**	**57.1**	**68.2**	**113.0**	**4.8**	**10.6**	**24.8**	**36.3**	**58.9**	**78.9**
	Phoning						Posing						Purchases					
Millisecond(ms)	80	160	320	400	560	1,000	80	160	320	400	560	1,000	80	160	320	400	560	1,000
LSTM3LR [8]	30.1	54.6	68.4	89.3	106.9	131.1	35.1	70.3	129.6	157.5	164.3	179.4	39.0	68.5	88.2	104.4	116.2	143.1
Res-GRU [33]	21.1	38.9	66.0	76.4	94.0	126.4	29.3	56.1	98.3	114.3	140.3	183.2	28.7	52.4	86.9	100.7	122.1	154.0
ConSeq2Seq [21]	13.5	26.6	49.9	59.9	77.1	114.0	16.9	36.7	75.7	92.9	122.5	187.4	20.3	41.8	76.5	89.9	111.3	151.5
HMR [30]	12.5	21.3	39.3	58.6	71.3	112.8	13.6	23.5	62.5	114.1	126.3	143.6	15.3	30.6	64.7	73.9	97.5	122.7
FC-GCN [32]	11.5	20.2	37.9	43.2	67.8	105.1	9.4	23.9	66.2	82.9	107.6	175.0	19.6	38.5	64.4	72.2	98.3	139.3
LDR [7]	10.4	14.3	33.1	39.7	—	**85.8**	8.7	21.1	58.3	81.9	—	133.7	16.2	36.1	62.8	76.2	—	112.6
TrajNet [27]	10.7	18.8	37.0	43.1	62.3	113.5	6.9	21.3	62.9	78.8	111.6	210.9	17.1	36.1	64.3	75.1	84.5	115.5
SDMTL [26]	10.5	18.5	37.2	43.1	60.8	112.3	6.8	20.5	64.0	82.4	107.2	204.7	18.4	38.8	61.1	68.2	80.9	113.6
HRI [31]	8.6	18.3	39.0	49.2	67.4	105.0	10.2	24.2	58.5	75.8	107.6	178.2	13.0	29.2	60.4	73.9	95.6	134.2
Our	**6.8**	**10.6**	**27.2**	**33.1**	**54.6**	97.8	**5.3**	**15.8**	**53.8**	**62.9**	**92.0**	**108.4**	**9.7**	**24.2**	**56.9**	**62.8**	**75.9**	**107.6**
	Waiting						Walking Dog						Average					
Millisecond(ms)	80	160	320	400	560	1,000	80	160	320	400	560	1,000	80	160	320	400	560	1,000
LSTM3LR [8]	31.3	57.4	100.5	120.5	122.8	159.3	47.2	81.4	123.9	136.2	153.5	185.3	36.4	60.7	95.4	111.8	125.7	157.7
Res-GRU [33]	23.8	44.2	75.8	87.7	105.4	135.4	36.4	64.8	99.1	110.6	128.7	164.5	25.0	46.2	77.0	88.3	106.3	136.6
ConSeq2Seq [21]	14.6	29.7	58.1	69.7	87.3	117.7	27.7	53,6	90.7	103.3	122.4	162.4	16.6	33.3	61.4	72.7	90.7	124.2
HMR [30]	12.8	24.5	45.2	85.1	87.5	121.9	30.1	41.4	78.4	100.1	134.7	157.4	13.3	23.2	44.7	63.8	86.1	116.2
FC-GCN [32]	9.5	22.0	57.5	73.9	73.4	107.5	32.2	58.0	102.2	122.7	105.8	142.2	12.1	25.0	51.0	61.3	78.3	114.0
LDR [7]	9.2	17.6	47.2	71.6	—	127.3	25.3	56.6	87.9	99.4	—	143.2	10.7	22.5	45.1	55.8	—	97.8
TrajNet [27]	8.2	21.0	53.4	68.9	92.9	165.9	23.6	52.0	98.1	116.9	141.1	181.3	10.2	23.2	49.3	59.7	77.7	110.6
SDMTL [26]	7.5	19.0	46.8	58.3	81.4	159.2	21.0	54.9	100.4	119.8	137.7	181.5	9.8	22.7	48.0	58.2	74.5	110.7
HRI [31]	8.7	19.2	43.4	54.9	74.5	108.2	20.1	40.3	73.3	86.3	108.2	146.9	10.4	22.6	47.1	58.3	77.3	112.1
Our	**6.2**	**14.2**	**38.9**	**44.3**	**63.6**	**95.7**	**16.4**	**33.3**	**63.7**	**68.4**	**96.3**	**138.7**	**8.3**	**18.8**	**39.0**	**47.9**	**65.3**	**96.4**

spatial distance between ground truth and prediction. Following the literature convention [33, 37], we evaluate our method on both short-term (< 400ms) and long-term (400-1,000ms) predictions.

The performance of different models on the H3.6M dataset is evaluated on all kinds of actions, including "Directions", "Eating", "Greeting", "Purchases", "Sitting Down", "Walking Dog", etc. A total of 10 methods are compared, including LSTM3LR [8], Res-GRU [33], ConSeq-Seq [21], HMR [30], FC-GCN [32], LDR [7], TrajNet [27], SDMTL [26], HRI [31], our MPT model. We present the results of 11 various actions and the overall average results for all actions.

The short-term and long-term prediction results are presented in Table 1. The first observation is that our MPT outperforms state-of-the-art methods by a large margin and keeps delivering the best performance on different actions. Surprisingly, compared to the current state-of-the-art, our method achieves a remarkable 17.5% accuracy improvement in average. It is noteworthy that our method consistently achieves the best results for both short-term and long-term predictions. Our second observation is that more complex actions such as "Walking Dog" and "Purchase" are harder to be predicted, leading to performance decay for all methods.

In addition to quantitative comparisons, we further visualize the prediction results of state-of-the-art methods. Fig. 3 demonstrates the prediction results for "Eating" and "Talking Photo", where the first 3 frames (in blue) in each line are historical frames and the subsequent frames in red

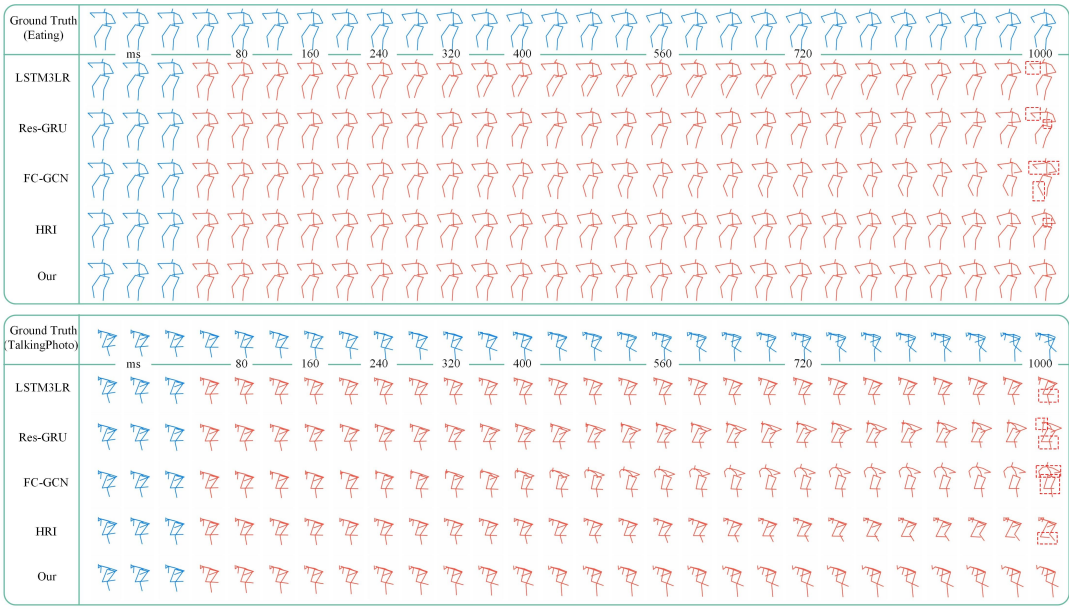

Figure 3. The visual results comparison on H3.6M. The first line demonstrates ground truth, the second to 6^{th} lines present the results of compared methods. The figures show results on "Eating" and "Talking Photo". Better viewed in color.

are future motion predictions. Specifically, we can observe that in the "Talking Photo" action, our model successfully captured the descending trend of the right leg while other methods did not. As a result, their predicted positions of the right leg are significantly distinct from the ground truth. In the "Eating" action, LSTM3LR tends to predict freezed motion in the long term, Res-GRU yields wrong movements of the two hands, FC-GCN has problems in predicting the position of the right leg, while HRI has errors in predicting the position of the left hand. In contrast, the movements of the hands and head are more coordinated and smooth in our results, which are also more consistent with the ground truth. Similar results are observed for other actions. Motivated readers may refer to `https://github.com/Pose-Group/MPT` for more visualized results. The empirical evidences reveal that our model can better discover subtle movement trends and achieve more accurate forecasts. Overall, the predicted pose sequences generated from our model are closer to the ground truth.

Furthermore, we conducted extensive experiments on the CMU MoCap dataset. The results are shown in Table 2. Consistently, we found that our model significantly outperforms existing methods for both short-term and long-term predictions. For example, on the "Basketball Signal" and "Soccer" actions, our model achieves an average of 20.8% and 22.6% improvements over state-of-the-art method, respectively. This reconfirms the effectiveness of working in trajectory space for motion prediction.

Animal Motion prediction We use the fish and mouse datasets to further evaluate our model and other methods. Whereas the human datasets pose the challenge of having to model multiple kinematic chains, the fish and mouse datasets raise different issues, *i.e.,* 1) long kinematic chain of 21 joints for fish; 2) animals exhibit faster and highly stochastic movements than humans; 3) relatively smaller datasets for training. The quantitative results are reported in Table 3. Empirical results suggest that mouse is more easier to be predicted than fish, which may due to the fact that fish is more active in action and the fish contain 21 joints which is significant longer than the 5 joints of mouse. Moreover, our method is shown to consistently outperform state-of-the-art methods on the animal datasets. Specifically, the proposed method achives a 24.1% accuracy improvement over HMR on Fish dataset, and 13.8% on Mouse dataset. This validates the effectiveness and generalizability of the method.

4.3. Ablation Study (RQ2)

We further study the influence of individual components in the proposed framework through the following ablation studies. Experiments are performed on the H3.6M dataset, with empirical results reported in Table 4. In the table, the motion prediction accuracies for 80, 160, 320, 400, 560 and 1,000 ms are presented.

First, to verify the impact of the proposed trajectory representation in modeling temporal motion contexts, we replace it with the conventional pose sequence representation (adopting 3D coordinates of joints). As presented in the second and fifth lines of Table 4, empirical results reveal

Table 2. Prediction results (in MPJPE) on CMU-MoCap dataset.

	Basketball						Basketball Signal						Directing Traffic						Jumping					
Millisecond(ms)	80	160	320	400	560	1,000	80	160	320	400	560	1,000	80	160	320	400	560	1,000	80	160	320	400	560	1,000
Res-GRU [33]	18.4	33.8	59.5	70.5	—	106.7	12.7	23.8	40.3	46.7	—	77.5	15.2	29.6	55.1	66.1	—	127.1	36.0	68.7	125.0	145.5	—	195.5
ConSeq2Seq [21]	16.7	30.5	53.8	64.3	—	91.5	8.4	16.2	30.8	37.8	—	76.5	10.6	20.3	38.7	48.4	—	115.5	22.4	44.0	87.5	106.3	—	162.6
FC-GCN [32]	14.0	25.4	49.6	61.4	77.4	106.1	3.5	6.1	11.7	15.2	25.3	53.9	7.4	15.1	31.7	42.2	70.3	152.4	16.9	34.4	76.3	96.8	131.4	164.6
LPJP [3]	11.6	21.7	44.4	57.3	—	90.9	2.6	4.9	12.7	18.7	—	75.8	6.2	12.7	29.1	39.6	—	149.1	12.9	27.6	73.5	92.2	—	176.6
LDR [7]	13.1	22.0	37.2	55.8	—	97.7	3.4	6.2	11.2	13.8	—	47.3	6.8	16.3	27.9	38.9	—	131.8	13.2	32.7	65.1	91.3	—	153.5
SDMTL [26]	10.9	20.2	40.9	50.8	66.1	110.2	2.9	6.2	16.4	23.1	37.4	71.6	5.1	10.9	23.2	30.2	46.1	105.5	11.1	24.6	65.7	90.3	130.9	191.2
Our	**10.3**	**16.2**	**32.9**	**41.2**	**55.1**	**84.7**	**2.4**	**4.8**	**9.5**	**10.9**	**18.7**	**40.2**	**4.1**	**9.3**	**18.2**	**27.4**	**38.7**	**85.9**	**9.2**	**20.6**	**53.4**	**81.5**	**111.3**	**139.6**
	Soccer						Walking						Wash window						Average					
Millisecond(ms)	80	160	320	400	560	1,000	80	160	320	400	560	1,000	80	160	320	400	560	1,000	80	160	320	400	560	1,000
Res-GRU [33]	20.3	39.5	71.3	84.0	—	129.6	8.2	13.7	21.9	24.5	—	32.2	8.4	15.8	29.3	35.4	—	61.1	16.9	30.5	54.2	63.6	—	96.5
ConSeq2Seq [21]	12.1	21.8	41.9	52.9	—	94.6	7.6	12.5	23.0	27.5	—	49.8	8.2	15.9	32.1	39.9	—	58.9	12.5	22.2	40.7	49.7	—	84.6
FC-GCN [32]	11.3	21.5	44.2	55.8	82.6	117.5	7.7	11.8	19.4	23.1	27.2	40.2	5.9	11.9	30.3	40.0	53.0	79.3	11.5	20.4	37.8	46.8	62.9	96.5
LPJP [3]	9.2	18.4	39.2	49.5	—	93.9	6.7	10.7	21.7	27.5	—	**37.4**	5.4	11.3	29.2	39.6	—	79.1	9.8	17.6	35.7	45.1	—	93.2
LDR [7]	10.3	21.1	42.7	50.9	—	91.4	7.1	10.4	17.8	20.7	—	37.5	5.8	12.3	27.8	38.2	—	56.6	9.4	18.8	31.6	43.2	—	82.9
SDMTL [26]	8.1	16.5	36.6	50.6	77.0	140.7	6.1	9.0	17.5	20.0	26.3	51.9	4.6	10.1	29.6	39.2	50.9	82.4	8.0	14.5	31.9	41.9	59.4	102.7
Our	**6.6**	**13.2**	**28.4**	**40.8**	**66.5**	**82.4**	**5.3**	**7.8**	**14.1**	**16.7**	**22.0**	45.6	**4.2**	**8.0**	**22.9**	**30.6**	**44.1**	**53.3**	**6.6**	**12.4**	**26.8**	**36.3**	**51.8**	**79.8**

Table 3. Evaluation (in MPJPE) on Fish and Mouse datasets.

	Fish						Mouse					
Time (ms)	80	160	320	400	560	1,000	80	160	320	400	560	1,000
ERD[8]	215.4	274.6	374.7	415.4	458.7	598.3	7.2	8.5	10.2	11.0	13.6	18.0
LSTM3LR [8]	160.2	199.2	291.4	326.1	450.1	736.2	7.1	8.8	10.4	11.3	13.5	18.5
Res-GRU [33]	75.0	138.0	290.0	372.6	477.7	810.9	4.4	6.5	9.5	11.1	13.6	19.2
HMR [30]	60.4	112.2	308.6	398.7	485.5	706.8	3.2	6.0	8.4	10.7	13.3	20.4
Our	**54.2**	**96.3**	**220.5**	**271.1**	**422.5**	**604.3**	**2.8**	**5.4**	**6.9**	**9.3**	**12.5**	**17.6**

Table 4. Impact of GCN module and trajectory representation.

Trajectory Representation	*GCN Module*	*Fixed Connection*	80	160	320	400	560	1,000
	✓	✓	29.9	38.6	47.5	65.7	96.2	115.1
✓		✓	11.3	24.3	43.9	59.6	82.6	111.7
✓	✓		9.9	22.7	42.1	55.6	77.3	105.2
✓	✓	✓	**8.3**	**18.8**	**39.0**	**47.9**	**65.3**	**96.4**

that using our trajectory representation to encode temporal dynamics significantly boost accuracy for both short-term and long-term predictions. Specifically, using the trajectory representation achieves more accuracy gain on short-term prediction than on long-term prediction.

Next, we evaluate the impact of explicit relations, namely using only the skeletal connection and prior knowledge while removing the adapatively learning of hidden joint connections. The results in the second line of Table 4 show that when implicit relation matrix is removed, the accuracies of the short-term forecast and long-term prediction are significantly affected. However, the performance drop is relatively smaller than that of replacing trajectory representation.

We also explore using merely the semantic enriched implicit relations. Towards this aim, we directly remove the explicit relation matrixes from the architecture. The results are demonstrated in Table 4. We see that utilizing merely the adapatively learning of hidden joint connections while removing *skeletal connection* and *prior knowledge* (fixed connections) lead to slight performance drop.

The results of these ablation experiments show the contribution of each module that constitutes our method: 1) the trajectory representation contributes to better encoding of the temporal dynamic information and plays an crucial role in motion prediction. 2) The semantic enriched GCN module captures useful dependencies between joints, which is also importent for generating accurate predictions. 3) The learnt hidden connections contribute more than the fixed joint connections.

4.4. Discussions (RQ3)

Experiments on 4 different benchmark datasets suggest that representing 3D skeleton motion sequence in trajectory space achieves significantly improved accuracy over representations in conventional pose space. Meanwhile, the generated visualization results are more natural and exhibit better inter-frame continuity.

Interestingly, a point that attracts our attention is: for long-term prediction, we find that although our proposed method still outperforms state-of-the-art approach, but as the prediction horizon goes deeper, the advantage of trajectory representation decreases. We plan to dive deeper into this phenomenon and come up with new solutions.

5. Conclusion

In this paper, we have proposed a trajectory representation consisting of position and frame-wise velocities, where position corresponds to *potential energy* and velocities correspond to *kinetic energy*. We further engage in a semi-constrained graph to model the *constraints*. These components together formulate a complete characterization of the trajectory configuration space and ultimately facilitate learning the Euler-Lagrange equation, *i.e.* modeling motion context. Extensive experiments confirm that our method significantly surpasses existing work on 4 different benchmark datasets. Interestingly, the method can also be generalized to fish and mouse datasets.

6. Acknowledgments

Corresponding authors: Pengxiang Su, Shuang Wu, Xuanjing Shen, Haipeng Chen.

This research is supported by the National Key Research and Development Program of China under Grant No.2020AAA0140004, the Natural Science Foundation of Zhejiang Province, China (Grant No. LQ19F020001), the National Natural Science Foundation of China (No. 61902348 & No. 61876070), and the Key R&D Program of Zhejiang Province (No. 2021C01104).

References

[1] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. *2019 IEEE/CVF International Conference on Computer Vision*, pages 7143–7152, 2019.

[2] Emad Barsoum, John Kender, and Zicheng Liu. HP-GAN: probabilistic 3d human motion prediction via GAN. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1418–1427, 2018.

[3] Yujun Cai, Lin Huang, Yiwei Wang, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Xu Yang, Yiheng Zhu, Xiaohui Shen, Ding Liu, Jing Liu, and Nadia Magnenat-Thalmann. Learning progressive joint propagation for human motion prediction. *Computer Vision - ECCV 2020 - 16th European Conference*, 12352:226–242, 2020.

[4] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. *Computer Vision - ECCV 2020 - 16th European Conference*, 12346:387–404, 2020.

[5] CMU Graphics Lab: Carnegie-Mellon Motion Capture (Mo cap) Database. `http://mocap.cs.cmu.edu`. 2003.

[6] Enric Corona, Albert Pumarola, Guillem Alenyà, and Francesc Moreno-Noguer. Context-aware human motion prediction. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6990–6999, 2020.

[7] Qiongjie Cui, Huaijiang Sun, and Fei Yang. Learning dynamic relationships for 3d human motion prediction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6518–6526, 2020.

[8] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. *IEEE International Conference on Computer Vision*, pages 4346–4354, 2015.

[9] Zeyu Fu, Federico Angelini, Jonathon A. Chambers, and Syed Mohsen Naqvi. Multi-level cooperative fusion of gm-phd filters for online multiple human tracking. *IEEE Transactions on Multimedia*, 21(9):2277–2291, 2019.

[10] Feng Fuli, He Xiangnan, Tang Jie, and Chua Tat-Seng. Graph adversarial training: Dynamically regularizing based on graph structure. *IEEE Transactions on Knowledge and Data Engineering*, 33:2493–2504, 2021.

[11] Di Gai, Xuanjing Shen, Haipeng Chen, and Pengxiang Su. Multi-focus image fusion method based on two stage of convolutional neural network. *Signal Process.*, 176:107681, 2020.

[12] Anand Gopalakrishnan, Ankur Arjun Mali, Dan Kifer, C. Lee Giles, and Alexander G. Ororbia II. A neural temporal model for human motion prediction. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12116–12125, 2019.

[13] Xiao Guo and Jongmoo Choi. Human motion prediction via learning local structure representations and temporal dependencies. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, pages 2580–2587, 2019.

[14] 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*, 36(7):1325–1339, 2013.

[15] Ashesh Jain, Amir Roshan Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5308–5317, 2016.

[16] Boyuan Jiang, Mengmeng Wang, Weihao Gan, Wei Wu, and Junjie Yan. STM: spatiotemporal and motion encoding for action recognition. *2019 IEEE/CVF International Conference on Computer Vision*, pages 2000–2009, 2019.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

[18] Jogendra Nath Kundu, Maharshi Gor, and R. Venkatesh Babu. Bihmp-gan: Bidirectional 3d human motion prediction GAN. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, pages 8553–8560, 2019.

[19] Inwoong Lee, Doyoung Kim, Seoungyoon Kang, and Sanghoon Lee. Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks. *IEEE International Conference on Computer Vision*, pages 1012–1020, 2017.

[20] Andreas M. Lehrmann, Peter V. Gehler, and Sebastian Nowozin. Efficient nonlinear markov models for human motion. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1314–1321, 2014.

[21] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5226–5234, 2018.

[22] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 211–220, 2020.

[23] Tianhong Li, Lijie Fan, Mingmin Zhao, Yingcheng Liu, and Dina Katabi. Making the invisible visible: Action recognition through walls and occlusions. *2019 IEEE/CVF International Conference on Computer Vision*, pages 872–881, 2019.

[24] Xiaodan Liang, Lisa Lee, Wei Dai, and Eric P. Xing. Dual motion GAN for future-flow embedded video prediction. *IEEE International Conference on Computer Vision*, pages 1762–1770, 2017.

[25] Anan Liu, Yuting Su, Weizhi Nie, and Mohan S. Kankanhalli. Hierarchical clustering multi-task learning for joint human action grouping and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39:102–114, 2017.

[26] Xiaoli Liu and Jianqin Yin. SDMTL: semi-decoupled multi-grained trajectory learning for 3d human motion prediction. *CoRR*, abs/2010.05133, 2020.

[27] Xiaoli Liu, Jianqin Yin, Jin Li, Pengxiang Ding, Jun Liu, and Huaping Liu. Trajectorycnn: A new spatio-temporal feature learning network for human motion prediction. *IEEE Trans. Circuits Syst. Video Technol.*, 31:2133–2146, 2021.

[28] Zhenguang Liu, Haoming Chen, Runyang Feng, Shuang Wu, Shouling Ji, Bailin Yang, and Xun Wang. Deep dual consecutive network for human pose estimation. *2021 IEEE/CVF*

International Conference on Computer Vision, pages 525–534, 2021.

[29] Zhenguang Liu, Kedi Lyu, Shuang Wu, Haipeng Chen, Yanbin Hao, and Shouling Ji. Aggregated multi-gans for controlled 3d human motion prediction. *Thirty-Fifth AAAI Conference on Artificial Intelligence*, pages 2225–2232, 2021.

[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. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10004–10012, 2019.

[31] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. *Computer Vision - ECCV 2020 - 16th European Conference*, 12359:474–489, 2020.

[32] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. *IEEE/CVF International Conference on Computer Vision*, pages 9489–9497, 2019.

[33] Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4674–4683, 2017.

[34] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, and Zachary DeVito. Automatic differentiation in pytorch. *NIPS-W*, 2017.

[35] Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion. *British Machine Vision Conference 2018*, page 299, 2018.

[36] Yemin Shi, Yonghong Tian, Yaowei Wang, Wei Zeng, and Tiejun Huang. Learning long-term dependencies for action recognition with a biologically-inspired deep network. *IEEE International Conference on Computer Vision*, pages 716–725, 2017.

[37] Yongyi Tang, Lin Ma, Wei Liu, and Wei-Shi Zheng. Long-term human motion prediction by modeling motion context and enhancing motion dynamics. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 935–941, 2018.

[38] Graham W. Taylor, Geoffrey E. Hinton, and Sam T. Roweis. Modeling human motion using binary latent variables. *Advances in Neural Information Processing Systems 19*, pages 1345–1352, 2006.

[39] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. *CoRR*, abs/2012.05522, 2020.

[40] Jack M. Wang, David J. Fleet, and Aaron Hertzmann. Gaussian process dynamical models. *Advances in Neural Information Processing Systems*, pages 1441–1448, 2005.

[41] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. *Computer Vision - ECCV 2018 - 15th European Conference*, pages 466–481, 2018.

[42] Yang Xun, Zhou Peicheng, and Wang Meng. Person reidentification via structural deep metric learning. *IEEE Trans. Neural Networks Learn. Syst.*, 30:2987–2998, 2019.

[43] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 7444–7452, 2018.

[44] Wei Yinwei, Wang Xiang, Nie Liqiang, He Xiangnan, Hong Richang, and Chua Tat-Seng. MMGCN: multi-modal graph convolution network for personalized recommendation of micro-video. *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1437–1445, 2019.

[45] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. *Computer Vision - ECCV 2020 - 16th European Conference*, 12354:346–364, 2020.

[46] Jason Y. Zhang, Panna Felsen, Angjoo Kanazawa, and Jitendra Malik. Predicting 3d human dynamics from video. *2019 IEEE/CVF International Conference on Computer Vision*, pages 7113–7122, 2019.

[47] Lei Zhu, Xu Lu, Zhiyong Cheng, Jingjing Li, and Huaxiang Zhang. Deep collaborative multi-view hashing for large-scale image search. *IEEE Transactions on Image Processing*, 29:4643–4655, 2020.