Learning Dynamic Relationships for 3D Human Motion Prediction

Qiongjie Cui Huaijiang Sun* Fei Yang
Nanjing University of Science and Technology, Nanjing, China
cuiqiongjie@njust.edu.cn, sunhuaijiang@njust.edu.cn, yangfei92516@163.com

Abstract

3D human motion prediction, i.e., forecasting future sequences from given historical poses, is a fundamental task for action analysis, human-computer interaction, machine intelligence. Recently, the state-of-the-art method assumes that the whole human motion sequence involves a fully-connected graph formed by links between each joint pair. Although encouraging performance has been made, due to the neglect of the inherent and meaningful characteristics of the natural connectivity of human joints, unexpected results may be produced. Moreover, such a complicated topology greatly increases the training difficulty. To tackle these issues, we propose a deep generative model based on graph networks and adversarial learning. Specifically, the skeleton pose is represented as a novel dynamic graph, in which natural connectivities of the joint pairs are exploited explicitly, and the links of geometrically separated joints can also be learned implicitly. Notably, in the proposed model, the natural connection strength is adaptively learned, whereas, in previous schemes, it was constant. Our approach is evaluated on two representations (i.e., angle-based, position-based) from various large-scale 3D skeleton benchmarks (e.g., H3.6M, CMU, 3DPW MoCap). Extensive experiments demonstrate that our approach achieves significant improvements against existing baselines in accuracy and visualization. Code will be available at https://github.com/cuiqiongjie/LDRGCN .

1. Introduction

Human motion prediction based on 3D skeleton data is committed to predicting future sequences from historical poses [24, 15]. Because of the potential in machine intelligence, autonomous vehicle, human-computer interaction, especially applications that require interaction with humans, it has been widely investigated and attracted considerable attention [14, 7, 24, 25, 21].

Traditional approaches have typically resorted to RNNs to model the human motion sequence [11, 18, 10]. However, RNN calculates the temporal context frame by frame, which may lead to the gradient vanishing or exploding, bringing their well-known training difficulty. Correspondingly, as described in the previous work [21, 14], recurrent models inevitably involve error accumulation and convergence to the mean pose. On the other hand, CNNs are also introduced to further extract multi-scale spatial correlations and have achieved remarkable performance [21]. Yet, skeleton sequence is essentially a non-euclidean data, while CNNs are theoretically only applicable standard 2D grid representation.

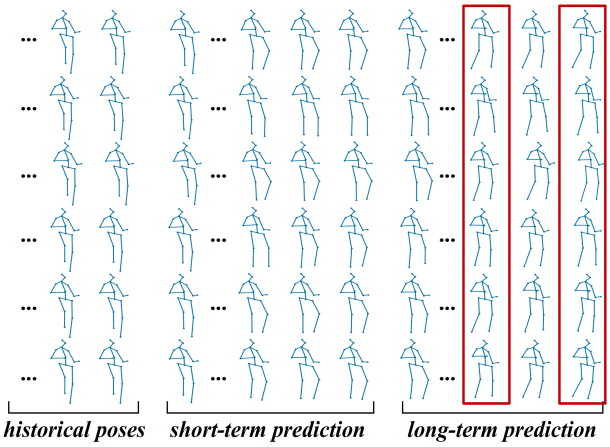

Figure 1. **Example result.** From top to bottom, we show the ground truth, the results of ConvSeqSeq, FC-GCN, FC-GCN 3D, Ours and Ours 3D. The red rectangle refers to a set of contrasting poses. The predictions of ours approach are indistinguishable from the ground truth in short-term prediction, and even for longer range prediction, the result are still semantically equivalent.

Recently, Graph Convolutional Networks (GCNs), a general form of conventional CNNs, with improved generalization and high interpretability, have received increasing attention and widely applied in many applications [6, 20, 33]. Researchers have also attempted to utilize GCNs to efficiently extract contextual information for forecasting human motion [23]. They suggest that the whole skeleton sequence serves as an implicit and unrestricted graph, and employs GCNs to learn these links between all joint pairs of the sequence. Although this fully-connected graph model (FC-GCN) has delivered impressive results, it cannot ex-

*Corresponding author

plicitly exploit the human skeleton structure. The hierarchical structure of the human body represents the topological relationship and indicates the inherent characteristics and strong dependencies of human joints. Ignoring such meaningful connectivities is equivalent to roughly viewing 3D skeleton data as a general format, thus generating unrealistic predictions.

To handle these aforementioned challenges, we propose a novel graph generative model to efficiently predict future poses from given historical movements. Specifically, we construct two parameterized graphs to learn the dynamic relationships among joints in 3D skeleton sequences: One is the connective graph that explicitly leverages the natural kinematic links of the human skeleton. Due to the heterogeneous information of human joints, in contrast to the fixed strengths in previous work, we innovatively parameterize the adjacent matrix $\boldsymbol{A}_p$ to learn these diverse patterns. Note that, for $\boldsymbol{A}_p$, only the relationship between physically connected parts is learnable, while the weights of other separated joints are always fixed; Another one is the global graph $\boldsymbol{Q}$. Except for natural connections, geometrically nonadjacent joints may potentially be interrelated. For example, during running, the movement of the left hand always shows a strong correlation with the right hand, rather than the left shoulder joint, which is connected to it. We solve this problem by a learnable global graph to learn these implicit connectivities along with the optimization process. Then, partially constrained $\boldsymbol{A}_p$ helps the flexible $\boldsymbol{Q}$ stabilize the training process, and $\boldsymbol{Q}$ assists $\boldsymbol{A}_p$ in capturing implicit relationships. Besides, inspired by [14], we further introduce a graph discriminator to distinguish the long sequence that spliced by input sequence and prediction or original future poses. Experiments empirically demonstrate that the adversarial regularization is indeed preserving the detail information and facilitating prediction visualizations.

The major contributions of this paper are summarized as: (1) We parameterize the adjacent matrix as the connective graph to learn the weights of natural connections instead of fixed ones, and propose a learnable global graph to capture implicit relationships, as shown in Figure 2. This data-driven method increases the flexibility of graph construction, making it more specific and applicable to the human kinetic structure. (2) Graph adversarial discriminator is introduced to further enhance the visualization of prediction. (3) On two skeleton representations of various large-scale benchmarks (*i.e.*, H3.6M [17], CMU [1], and 3DPW [28] MoCap datasets), extensive experiments show that our model surpasses the state-of-the-art methods in terms of visualization and precision in almost all scenarios.

2. Related Work

Human Motion Prediction. Typical methods formulate human motion prediction as a sequence-to-sequence

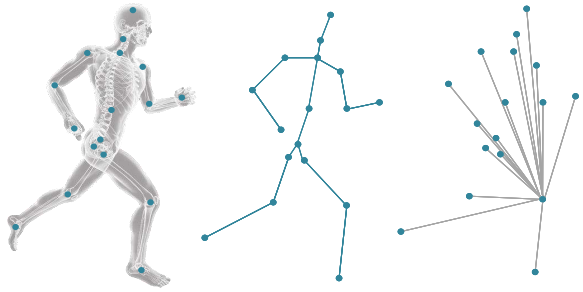

Figure 2. **Left:** human skeleton. **Middle:** connective graph where blue lines indicate the learnable strengths of natural connectivities. **Right:** global graph, and the grey line is the implicit relationships for right knee joint.

(seq2seq) learning problem [11, 24, 27]. Specifically, RNNs are proposed to capture the temporal information of human motion, achieving encouraging results [4, 25]. Fragkiadaki *et al*. [11] present two solutions: 3-layer long short-term memory (LSTM) network (*i.e*., LSTM-3LR) and *Encoder-Recurrent-Decoder* (*i.e*., ERD), where LSTM is utilized to extract long-term dependencies. Meanwhile, a structural RNN [18] is developed to semantically model the spatio-temporal structure of 3D skeleton sequence. The above two methods are action-specific models, and significant discontinuities between the predicted first frame and the last frame of the input sequence are often observed. Martinez *et al*. [24] introduce residual learning to recurrent model to produce a smooth prediction. Tang *et al*. [27] suggest using attention mechanism to capture the long-term temporal dependency to efficiently model human motion. However. RNN-based models often fall into the criticized problem of convergence to the static mean pose. Recently, researchers have proposed various variants of RNN, *e.g*., hierarchical motion recurrent model [22], Verso-Time Label Noise-RNN [12], and triangular-prism RNN [8]. Unfortunately, since RNN calculates the temporal context step by step, it still inevitably causes error accumulation.

Currently, generative adversarial networks (GANs) have shown impressive performance [2, 13, 21, 30]. Li *et al*. [21] propose a convolutional discriminator to model human motion sequences and generate realistic predicted poses. Gui *et al*. [14] present a novel framework, called adversarial geometry aware encoder-decoder (AGED), in which the discriminator distinguishes the concatenations of observed frames and prediction or ground truth.

Graph Convolutional Networks. As a generalization of CNNs, GCN is naturally suitable for data with specific graph structure, *e.g*., point cloud [30], social network [32], and 3D skeleton data [26, 23, 12]. Yan *et al*. [31] construct a spatial-temporal graph defined on the essential skeleton and the temporally consecutive poses for action recognition. This strategy leverages the strong natural dependencies among human joints; however, it reduces flexibility. Shi *et al*. [26] address this limitation by parameterizing an

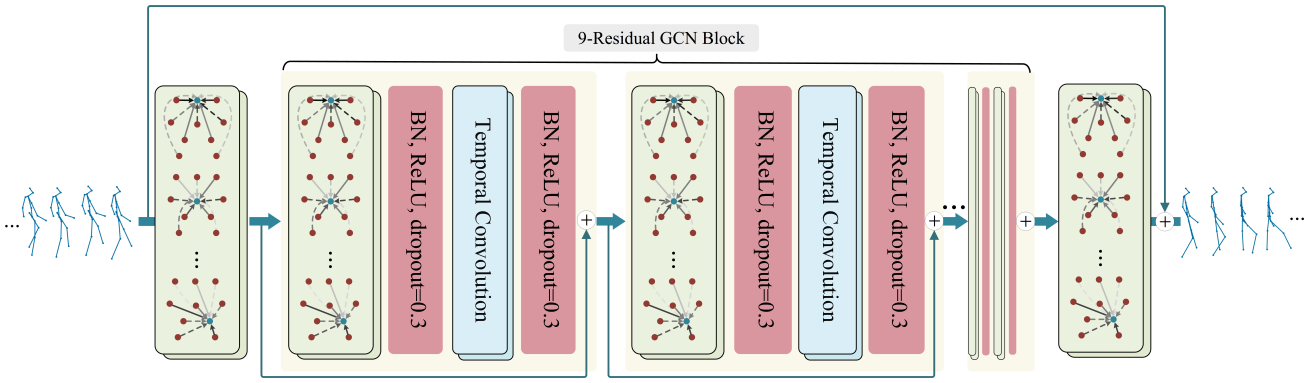

Figure 3. **Illustration of the proposed model.** Each GCN layer is followed by a temporal convolution to form GCN blocks to extract the spatial-temporal correlation hierarchically. In each block, solid lines and dashed lines respectively represent learnable natural connections and implicit relationships, wherein different gray depths indicate differential weights. The final model consists of 9 residual GCN blocks to learn the dynamic relationship of 3D skeleton sequences. Then, an observation of historical poses is fed into the network to predict the future sequence with an end-to-end manner. Note that the symbol $\oplus$ is residual learning or skip connections.

unrestricted matrix to adaptively learn the implicit connections, except for natural connections. For human motion prediction, Mao *et al*. [23] suggest that the whole motion sequence serves as an unconstrained topology, presenting impressive results. However, such construction is equivalent to treating the motion sequence cursorily as general data without meaningful natural connections of human joints, and unconstrained learning may lead to unstable training. Instead of the above solutions, we set the adjacency matrix as the model parameter, where the weights of the naturally connected parts are learnable in the full training, while the others are fixed at 0. The partially constrained adjacency matrix ensures flexibility and sufficient utilizing of inherent relationships of the human skeleton. Moreover, inspired by [26], in addition to natural connections, we also parameterize a global graph to learn the implicit connectivities of joints. With the above graph construction, our model can learn the dynamic relationship of the 3D skeleton sequence, so as to produce a high-fidelity prediction.

3. The Proposed Method

Following the previous works [14, 21, 9], in this paper, the 3D skeleton samples are obtained from motion capture (MoCap) technologies. A motion sequence consists of a series of consecutive frames (poses), wherein each frame records the angle or position information of each joint. Suppose an observed motion sequence is formulated as $\mathbf{X}_{1:T} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_t, ..., \mathbf{x}_T\}$ with $1 \leq t \leq T$, and each $\mathbf{x}_t \in R^{3N}$ from $\mathbf{X}$ represents a pose at time step t where N is joints number. Then, we denote $\tilde{\mathbf{Y}} = \{\tilde{\mathbf{y}}_{T+1}, \tilde{\mathbf{y}}_{T+2}, ..., \tilde{\mathbf{y}}_{T+\Delta t}\}$ be the prediction with Δt frames, and $\mathbf{Y} = \{\mathbf{y}_{+1}, \mathbf{y}_{T+2}, ..., \mathbf{y}_{T+\Delta t}\}$ be the corresponding ground truth of future motion. Our goal is to learning an optimal generator $\mathcal{P}_*$ that accurately map historical poses $\mathbf{X}$ to future sequence $\mathbf{Y}$. To this end, we propose to learn the dynamic relationships of skeleton sequences to minimize the difference between the prediction and ground truth.

3.1. Learning Dynamic Relationships

3D skeleton data, recording movement information of specific joint frame by frame from MoCap devices, which is essentially a sequential data and naturally suitable for recurrent neural networks (RNNs). However, due to limited abilities to extract spatial information and the inevitable error accumulation, RNN variants often produce unrealistic predictions. Additionally, conventional CNNs are also ignoring the kinematic dependencies of human joints.

Convolution on Graphs. Therefore, in this work, a novel dynamic GCN model is proposed to automatically learn the spatio-temporal relationships in MoCap sequence in order to efficiently predict future poses. Specifically, we present a skeleton-based pose as a undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, in which $\mathcal{V}$ is vertex set, *i.e*., joints set of human body. $\mathcal{E} = \{e_{ij} > 0 \,|\, i, j \in 1, 2, ..., N\}$ is edge set that v_i and v_j are naturally connected. Then, a motion sequence is formulated as $\mathcal{M} = \{\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_T\}$ with T frames. The typical operation of GCNs is formally expressed as:

$$\boldsymbol{F}^{(l+1)} = g(\boldsymbol{F}^{(l)}, \boldsymbol{A}) = \sigma[\hat{\boldsymbol{A}}\boldsymbol{F}^{(l)}\boldsymbol{W}^{(l)}], \;\; \hat{\boldsymbol{A}} = \tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{\frac{1}{2}}, \quad (1)$$

where $\boldsymbol{F}^{(l)} \in R^{N,S_l}$ and $\boldsymbol{F}^{(l+1)} \in R^{N,S_{l+1}}$ are input and output tensor at l-th layer, respectively. $\boldsymbol{W} \in R^{S_l,S_{l+1}}$ is learnable weight matrix, σ is activation function (*e.g*., ReLU). $\tilde{\boldsymbol{D}} \in R^{N,N}$ is diagonal degree matrix where $\tilde{\boldsymbol{D}}_{i,i} = \sum_j \tilde{\boldsymbol{A}}_{i,j}$. $\tilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I}$ where $\boldsymbol{A}$ is the adjacency matrix and $\boldsymbol{I}$ is identity matrix. Note that in the previous work [20, 31, 12], the weight of the adjacency matrix is consistently fixed with optimization process, *i.e*., $\boldsymbol{A} \in \{e_{i,j} = 1\}$. With the deepening of the network hierarchy, this constant representation may also partially model high-level features. However, this strategy is not an optimal expression. Intuitively, all the joints that connected with the vertex v_i contribute unequally to the movement pattern for v_i. For example, during walking, the shoulder joint is more dependent on the trunk than

the elbow joint.

Connective Graph. To solve this problem, we innovatively develop a connective graph parameterized by the adjacency matrix, formally as $\boldsymbol{A}_p \in R^{N,N}$, which represents the learnable connection strength of the natural links for the human skeleton. Instead of the fixed weight of the adjacency matrix in prior work [31, 12], we suggest that the interdependencies of natural connection in human skeletons are trainable rather than constant, and then we learn the weights of such natural relationships adaptively. With the optimization process, the $\boldsymbol{A}_p^{(l)} \in R^{N,N}$ at l-th layer gradually reach the optimal solution, and the relative importance between physically connected joint pairs can be automatically obtained, The operation can be simplified as:

$$\boldsymbol{F}^{(l+1)} = g(\boldsymbol{F}^{(l)}) = \sigma[(\boldsymbol{A}_p^{(l)} \circ \boldsymbol{M})\, \boldsymbol{F}^{(l)} W^{(l)}], \qquad (2)$$

where $\boldsymbol{A}_p$ is $N \times N$ learnable connective matrix, and $\boldsymbol{F}^{(l+1)} \in R^{N \times S_{l+1}}$ and $\boldsymbol{F}^{(l)} \in R^{N \times S_l}$ are input and output at l-th layer. $\boldsymbol{M} \in R^{N,N}$ is a fixed mask matrix, and the symbol $\circ$ indicates element-wise product. With the binary $\boldsymbol{M}$, for partially constrained $\boldsymbol{A}_p = \{e_{ij}\}$, throughout training, only the weights of the interconnected vertexes are optimized, while the separate parts are fixed at 0. In other words, $\boldsymbol{A}_p$ is introduced to learn the natural connection strength of the human skeleton. Unlike constant adjacency matrix, parameterized $\boldsymbol{A}_p$ can adaptively treat the connective relationships between joints, as shown in the middle of Figure 2. Besides, the fact that learnable A_p with a fixed topology means regularizing the proposed model based on prior knowledge, which can help the model converge to the global minimum faster. Note that, the trainable $\boldsymbol{A}_p$ is still initialized by the original adjacency matrix $\boldsymbol{A}$.

Global Graph. Up to now, the constructed graph is still manually designed from the kinematic structure of the human body. Even though the weights of natural connections are adaptively calculated, this configuration may fail in adequately modeling the spatial characteristics of the human skeleton. For instance, in the running, the left and the right leg always support each other, but there are no physical connectivities between them. Due to the separation of node relationships, $\boldsymbol{A}_p$ has a low ability to model such valuable information. To tackle the challenges, we further present a global graph to capture implicit but critical structural features that transcend natural connections. Particularly, we parameterize a matrix $\boldsymbol{Q} \in R^{N,N}$ initialized from the zeros matrix with the same size of $\boldsymbol{A}_p$ to adaptively learn the underlying relationships among all human joints. In contrast to $\boldsymbol{A}_p$, $\boldsymbol{Q}$ is flexible without any constrains, which means that it gradually achieves optimum along with the training process. Besides natural relations, the global $\boldsymbol{Q}$ can learn useful and implicit connection weights distributed in training samples. Finally, the updating formulation of the proposed model is expressed as:

$$\boldsymbol{F}^{(l+1)} = g(\boldsymbol{F}^{(l)}) = \sigma[(\boldsymbol{A}_p^{(l)} \circ \boldsymbol{M} + \boldsymbol{Q}^{(l)})\boldsymbol{F}^{(l)} \boldsymbol{W}^{(l)}], \qquad (3)$$

where $\boldsymbol{A}_p^{(l)}, \boldsymbol{Q}^{(l)} \in R^{N \times N}$ is optimal matrix at l-th layer with training process to jointly learn dynamic relationships of skeleton sequence. Such a construction brings several significant benefits:

(1) Learnable $\boldsymbol{A}_p$ adaptively extract the heterogeneous information of natural connections of human joints;

(2) Unconstrained $\boldsymbol{Q}$ improves flexibility;

(3) Partially restricted $\boldsymbol{A}_p$ ensures stable training for $\boldsymbol{Q}$;

(4) $\boldsymbol{Q}$ as a supplementary of $\boldsymbol{A}_p$ to learn the underlying topology;

(5) $\boldsymbol{A}_p$ and $\boldsymbol{Q}$ cooperate to learn dynamic relationships of skeleton sequences effectively.

Temporal modeling using convolutions The typical methods [24, 4, 27] are to use RNN to model the temporal information of human motion. However, RNN-based models inevitably accumulate errors and increases computational complexity. In contrast to RNN, TCN (*i.e.*, 1D convolution) is a feed-forward operation, demonstrating advantages in parameter number, parallelism, accuracy, and model complexity for modeling temporal patterns [3, 31]. Consequently, we employ TCN along the time dimension of human motion to extract temporal correlations.

3.2. Optimization

Due to the different characteristics of the position-based and angle-based skeleton sequence, we introduce the following loss functions to obtain better visualization and accuracy of predicted poses.

Content loss, to ensure that the predicted sequence is consistent with the global information of the experimental samples as much as possible, *i.e.*,

$$\mathcal{L}_{con} = \frac{1}{\Delta t} \sum_{i=T+1}^{T+\Delta t} \sum_{j=1}^{d} \left\| \mathbf{y}_{i,j} - \tilde{\mathbf{y}}_{i,j} \right\|_2, \qquad (4)$$

where $\mathbf{y}_{i,j}$ and $\tilde{\mathbf{y}}_{i,j}$ is the j-th joint of i-th frame for ground turth and the prediction, respectively. d, N are number of human joints and prediction length. Note that, for the two representations of 3D skeleton sequence, $\mathbf{y}_{i,j}$ and $\tilde{\mathbf{y}}_{i,j}$is angle or position information, respectively.

Gram matrix loss, is also introduced to preserve the consistency between prediction patterns and the original pose, and avoid convergence to mean pose, *i.e.*,

$$\mathcal{L}_{gram} = \frac{1}{\Delta t} \sum_{i=T+1t}^{T+\Delta t-1} \left\| H(\tilde{\mathbf{y}}^i, \tilde{\mathbf{y}}^{i+1}) - H(\mathbf{y}^i, \mathbf{y}^{i+1}) \right\|_2, \qquad (5)$$

where the gram matrix is defined as $H(\alpha, \beta) = [\alpha : \beta]\,[\alpha : \beta]^T$, and [:] indicates concatenation.

Bone length loss, enforces the bone length of each generated pose to approach the ground truth. Moreover, for 3D coordinates of skeleton sequence, a fixed bone length can force the predicted joint position to lie on a sphere with its parent joint as the origin and the bone length as the radius. This dramatically reduces the search space for joint movement and is conducive to the faster convergence of the network. For bone length $l_{i,j}$ of j-th joint in i-th pose and the

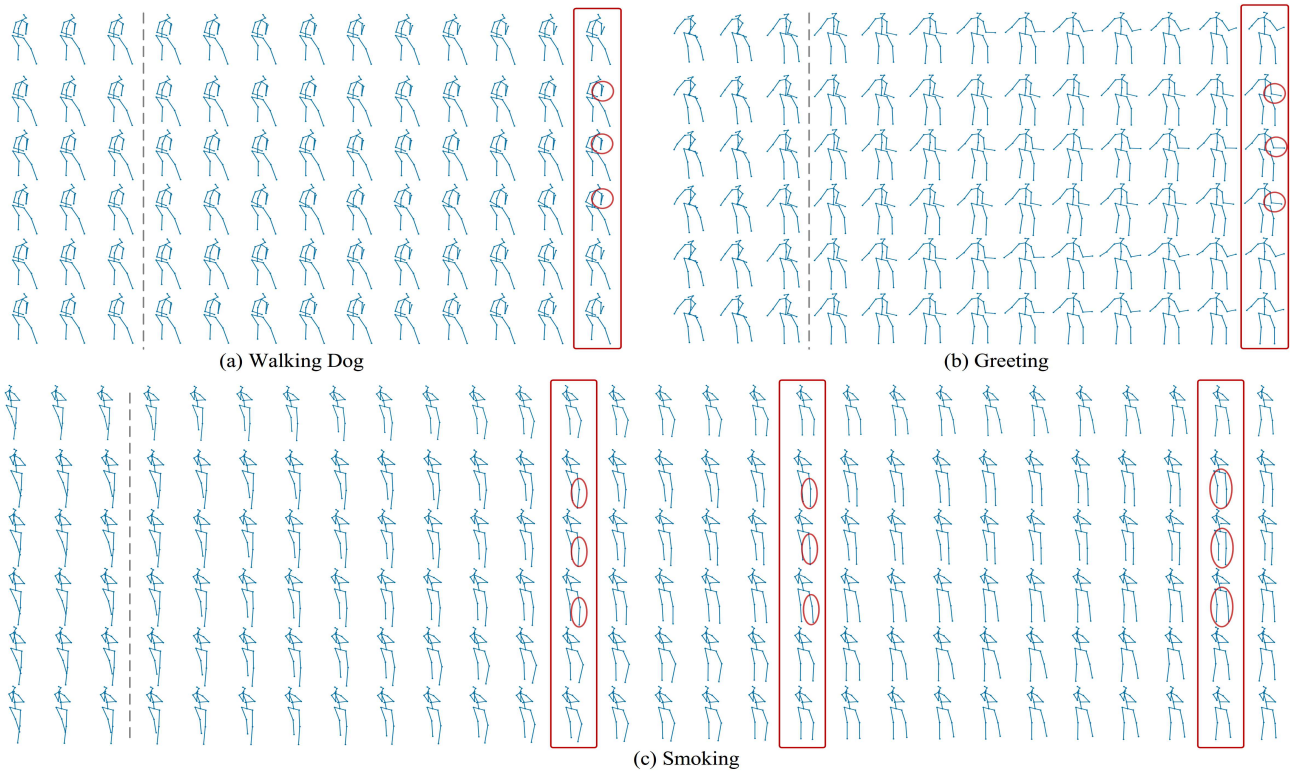

Figure 4. **Qualitative comparison.** The short-term predictions (400 ms) are shown in (a) walking dog, (b) greeting, and long-term prediction (1000 ms) in (c) smoking. In each sub-figure, the first row is ground truth; the second, third and fifth are the result of ConvSeq2Seq [21], FC-GCN [23], and our model based on angles; and the fourth and the bottom are the prediction of FC-GCN and our model based on 3D coordinate. In each row, the first 3 animations are observed poses, and the remainders are predicted frames, where the interval of each animation is 40 ms. The red rectangles refer to contrasting frames, and the circles are unreasonable parts. From the result, we observe that the proposed model produces more realistic visualization in all scenarios.

corresponding predicted part, this loss is denoted as:

$$\mathcal{L}_{bone} = \frac{1}{\Delta t}\sum_{i=T+\Delta t}^{T+N}\sum_{j=1}^{d}\left\|l_{i,j} - \tilde{l}_{i,j}\right\|_2 . \quad (6)$$

Recently, GANs [13, 2] have introduced into the variation of GCNs and achieved remarkable performance in many applications [30, 5, 32]. Motivation from these works, we develop adversarial learning for our generator $\mathcal{P}$ to further enhance the prediction visualization. In particular, following the formalism of WGAN-GP [16, 4], we design a graph discriminator $\mathcal{D}$ with gradient penalty, which shares the generator architecture but has fewer layers. Then, the **adversarial loss** can be expressed as:

$$\mathcal{L}_{\mathcal{D}} = \mathcal{D}([X:\mathcal{P}(X)]) - \mathcal{D}([X:Y]) + \lambda(\|\nabla_{\hat{x}}\mathcal{D}(\hat{x})\|_2 - 1)^2, \quad (7)$$

$$\mathcal{L}_{\mathcal{P}} = -\mathcal{D}([X:\mathcal{P}(X)]), \quad (8)$$

where $(\|\nabla_{\hat{x}}\mathcal{D}(\hat{x})\|^2 - 1)^2$ is the gradient penalty term, and $\hat{x} = \epsilon([X:Y]) + (1-\epsilon)([X:\mathcal{P}(X)])$ is a random sample with uniform distribution. Inspired by [14], the discriminator is introduced to distinguish the long sequence that is concatenated from the historical sequence and the prediction or ground truth, achieving better results. In all of our experiments, we set $\lambda = 5$.

In this paper, we exploit two final loss functions for both angle-based and position-based representations of human motion respectively:

Final loss for angle-based skeleton sequence. The following loss is used to optimize the proposed model based on joint angle representation, that is,

$$\mathcal{P}^* = \arg\min_{\mathcal{P}}\max_{\mathcal{D}\in D}\lambda_{con}\mathcal{L}_{con} + \lambda_{gram}\mathcal{L}_{gram} + \mathcal{L}_{\mathcal{P}} + \mathcal{L}_{\mathcal{D}}, \quad (9)$$

where $\lambda_{con} = 0.01$, $\lambda_{gram} = 0.001$ are hyperparameters to balance the importance of each loss term;

Final loss for position-based skeleton sequence. With optimum $\lambda_{con} = 0.01$, $\lambda_{bone} = 0.0005$, we present the final loss function for 3D coordinate skeleton sequence, *i.e.*,

$$\mathcal{P}^* = \arg\min_{\mathcal{P}}\max_{\mathcal{D}\in D}\lambda_{con}\mathcal{L}_{con} + \lambda_{bone}\mathcal{L}_{bone} + \mathcal{L}_{\mathcal{P}} + \mathcal{L}_{\mathcal{D}} \quad (10)$$

3.3. Implementation

As a primary component, each block consists of a proposed GCN layer and a TCN layer, with dropout rate of 0.3. Besides, each layer is followed by a batch normalization (BN) and ReLU activation function, as shown in Figure 3. We also add a residual connection in each block to stabilize the training process. Then, the final model consists of 9 residual dynamic GCN blocks. Due to more abstract representations for deeper layers, we gradually increase the number of output channels in the GCN layer, *i.e.*, 64, 64, 64, 128, 128, 128, 256, 256, 256. Skip connection is added to the input and output layers. We implement TCN with $k * 1$

	Walking					Eating					Smoking					Discussion					Directions				
millisecond (ms)	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
Residual sup.[24]	0.28	0.49	0.72	0.81	1.14	0.23	0.39	0.62	0.76	1.34	0.33	0.61	1.05	1.15	1.83	0.31	0.68	1.01	1.09	1.79	0.26	0.47	0.72	0.84	1.46
ConvSeqSeq[21]	0.33	0.54	0.68	0.73	0.92	0.22	0.36	0.58	0.71	1.24	0.26	0.49	0.96	0.92	1.62	0.32	0.67	0.94	1.01	1.86	0.39	0.60	0.80	0.91	1.45
AGED w/o adv [14]	0.28	0.42	0.66	0.73	0.73	0.22	0.35	0.61	0.74	**0.74**	0.30	0.55	0.98	0.98	**0.99**	0.30	0.63	0.97	1.06	1.06	0.26	0.46	0.71	0.81	1.32
AGED w/ adv [14]	0.22	0.36	0.55	0.67	0.91	0.17	0.28	0.51	0.64	0.93	0.27	0.43	0.82	0.84	1.21	0.27	0.56	0.76	0.83	1.30	**0.23**	**0.39**	0.63	0.69	1.21
FC-GCN [23]	0.18	0.31	0.49	**0.56**	0.79	**0.16**	0.29	0.50	**0.62**	1.05	0.22	0.41	0.86	**0.80**	1.13	0.20	0.51	0.77	0.85	0.85	0.26	0.45	0.71	0.79	1.07
Ours	**0.16**	**0.29**	**0.46**	0.57	**0.71**	**0.16**	**0.27**	**0.49**	0.64	0.97	**0.20**	**0.38**	**0.79**	0.82	1.08	**0.19**	**0.45**	**0.72**	**0.81**	**0.84**	0.29	0.43	**0.59**	**0.68**	**0.95**
	Greeting					Phoning					Posing					Purchase					Sitting				
millisecond (ms)	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
Residual sup.[24]	0.75	1.17	1.74	1.83	1.93	0.23	0.43	0.69	0.82	1.73	0.36	0.71	1.22	1.48	2.43	0.51	0.97	1.07	1.16	2.30	0.41	1.05	1.49	1.63	2.14
ConvSeqSeq[21]	0.51	0.82	1.21	1.38	1.72	0.59	1.13	1.51	1.65	1.81	0.29	0.60	1.12	1.37	2.65	0.63	0.91	1.19	1.29	2.52	0.39	0.61	1.02	1.18	2.67
AGED w/o adv [14]	0.61	0.95	1.44	1.61	1.81	0.23	0.42	0.61	0.79	1.77	0.34	0.70	1.19	1.40	2.01	0.46	0.89	1.06	1.11	1.89	0.46	0.87	1.23	1.51	2.11
AGED w/ adv [14]	0.56	0.81	1.30	1.46	2.12	**0.19**	**0.34**	**0.50**	**0.68**	1.41	0.31	0.58	1.12	1.34	1.78	0.46	0.78	1.01	**1.07**	1.77	0.41	0.76	1.05	1.19	1.72
FC-GCN [23]	0.36	0.60	0.95	1.13	1.43	0.53	1.02	1.35	1.48	2.08	0.19	**0.44**	1.01	1.24	1.54	**0.43**	0.65	1.05	1.13	1.73	0.29	0.45	0.80	**0.97**	1.47
Ours	**0.35**	**0.56**	**0.87**	**0.98**	**1.33**	0.43	0.54	0.63	0.78	**1.33**	**0.15**	**0.44**	**0.91**	**1.07**	**1.34**	**0.43**	**0.57**	**0.88**	1.08	**1.49**	**0.27**	**0.43**	**0.69**	1.01	**1.38**
	Sitting down					Taking photo					Waiting					Walking Dog					Walking together				
millisecond (ms)	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
Residual sup.[24]	0.39	0.81	1.40	1.62	2.72	0.24	0.51	0.90	1.05	1.51	0.28	0.53	1.02	1.14	2.34	0.56	0.91	1.26	1.40	1.86	0.31	0.58	0.87	0.91	1.42
ConvSeqSeq[21]	0.41	0.78	1.16	1.31	2.06	0.23	0.49	0.88	1.06	1.40	0.30	0.62	1.09	1.30	2.50	0.59	1.00	1.32	1.44	1.92	0.27	0.52	0.71	0.74	1.28
AGED w/o adv [14]	0.38	0.77	1.18	1.41	1.88	0.24	0.52	0.92	1.01	1.22	0.31	0.64	1.08	**1.12**	1.91	0.51	0.87	1.21	1.33	1.51	0.29	0.51	0.72	0.75	**1.08**
AGED w/ adv [14]	0.33	0.62	0.98	1.10	1.98	0.23	0.48	0.81	0.95	1.65	0.24	0.50	1.02	1.13	1.65	0.50	0.81	1.15	1.27	1.61	0.23	0.41	0.56	0.62	1.47
FC-GCN [23]	0.30	**0.61**	0.90	1.00	1.45	0.14	0.34	0.58	**0.70**	1.35	0.23	0.50	0.91	1.14	1.23	0.46	0.79	1.12	1.29	**1.31**	**0.15**	0.34	0.52	0.57	1.41
Ours	**0.29**	0.62	**0.87**	**0.93**	**1.42**	**0.13**	**0.33**	**0.54**	0.71	**1.20**	**0.21**	**0.48**	**0.84**	1.15	**1.21**	**0.45**	**0.68**	**0.93**	**1.14**	1.38	**0.15**	**0.33**	**0.49**	**0.54**	1.38

Table 1. Comparisons of angle error for short-term and long-term prediction on H3.6M dataset. The best results are highlighted in bold, and the second are underlined.

filter size, where the temporal dimension is $k = 9$. The discriminator has a similar structure with six layers. The unit number in the bottom layers of multi-layer perceptrons is set as 512, 246, 64, 1. Note that, inspired by AGED [14], our adversarial discriminator is introduced to distinguish the long sequence which is concatenated from historical poses and predictions or ground truth. The batch size is set to 32. We utilize Adam [19] to train our model, and the learning rate is initialized as 0.001 with a 0.98 decay per epoch.

4. Experiments

4.1. Datasets and Preprocessing

We use several action analysis benchmarks to verify the effectiveness of the proposed model:

H3.6M [17], is considered to be the largest and challenging human motion analysis dataset currently. It involves 15 complex action scenarios, including periodic (*e.g.*, walking) or aperiodic (*e.g.*, eating, smoking), performed by seven actors. Consistent with the solution of data preprocessing in [21, 23], we have removed global translation and rotation, and constant joints. Finally, each pose is represented as a skeleton of 17 joints. During training, we down-sample all sequences to 25 frames per second (fps) and re-expressed it as exponential mapping. Besides, skeleton sequences are normalized by subtracting the average pose of wholes datasets and then dividing into the standard deviation. Following the previous works [18, 21, 14], we use the subject-5 (S5) to test our model, and S11 is the validation set, and the remaining five subjects are training samples.

CMU MoCap [1]. We have also published experimental results on CMU MoCap dataset. As previous literature [24, 21, 23], 8 actions are selected as our samples, *e.g.*, walking, running, wash window. We used the same training/test split in their released code, the validation set is unavailable due to data limitations. Other pre-processing strategies are the same as those of H3.6M.

3DPW MoCap [29], is a recently released large-scale action analysis dataset, which contains 51k indoor or outdoor poses. In order to make a fair comparison, we adopt the partitioning solution of official training, testing and validation set. The frame rate of all sequences is 30fps.

4.2. Evaluation Criteria and Baselines

Criteria: Following the previous standard evaluation metric in [21, 24], we report the comparison results of the angle error between the ground truth and the prediction, *i.e.*,

$$E_{angle} = \frac{1}{\Delta t}\frac{1}{N}\sum_{i=T+1}^{T+\Delta t}\sum_{j=1}^{N}\left|\mathbf{y}_{i,j} - \tilde{\mathbf{y}}_{i,j}\right|, \tag{11}$$

where $\tilde{\mathbf{y}}_{i,j}$ is predicted angle in i-th frame of j-th joint, and $\mathbf{y}_{i,j}$ is the corresponding ground truth. Besides, we also evaluate 3D errors using Mean Per Joint Position Error (MPJPE) [17, 23] in millimeter, that is,

$$E_{3D} = \sqrt{\frac{1}{\Delta t}\frac{1}{N}\sum_{i=T+1}^{T+\Delta t}\sum_{j=1}^{N}\left\|\mathbf{p}_{i,j} - \tilde{\mathbf{p}}_{i,j}\right\|_2^2}, \tag{12}$$

where $\mathbf{p}_{i,j}$ and $\tilde{\mathbf{p}}_{i,j}$ are the position of the ground truth and the prediction, which can be measured either by converting the predicted angles to 3D space, or directly train on 3D coordinates of the skeleton sequence.

Baselines: To evaluate the effectiveness of the proposed model, five latest methods are used as the competitive methods, including recurrent model (Residual sup.) [24], feed-forward model (ConvSeqSeq) [21], GAN-based (AGED w/ or w/o adv) [14] and graph method (FC-GCN) [23]. In addition to evaluating the angle error, we also investigate the 3D error of the baseline methods exploiting strategy in [23] to transform the predicted angle into 3D Cartesian space. On the other hand, we take the position-based motion sequence as the input and output of baselines and our method, to statistics the comparison results of 3D error.

4.3. Results

Following the previous work [24, 25, 4, 27], in this paper, we focus on high-accuracy prediction in the next $400ms$

	Walking					Eating					Smoking					Discussion					Directions				
millisecond (ms)	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
ConvSeq2Seq [21]	21.8	37.5	55.9	63.0	92.1	13.3	24.5	48.6	60.0	87.7	15.4	25.5	39.3	44.5	67.5	23.6	43.6	68.4	74.9	134.4	26.7	43.3	59.0	72.4	132.2
ConvSeq2Seq 3D [21]	17.1	31.2	53.8	61.5	89.2	13.7	25.9	52.5	63.3	74.4	11.1	21.0	33.4	38.3	52.2	18.9	39.3	67.7	75.7	123.9	22.0	37.2	59.6	73.4	118.3
FC-GCN [23]	11.1	19.0	32.0	39.1	53.7	9.2	19.5	40.3	48.9	62.5	9.2	16.6	26.1	29.0	47.3	11.3	23.7	41.9	46.6	81.4	11.2	**23.2**	52.7	64.1	92.5
FC-GCN 3D [23]	**8.9**	15.7	29.2	33.4	50.9	8.8	18.9	39.4	47.2	57.1	**7.8**	14.9	25.3	28.7	44.3	9.8	22.1	39.6	44.1	78.5	12.6	24.4	48.2	58.4	89.1
Ours	9.7	17.7	28.3	32.2	51.3	10.2	17.4	38.7	49.3	56.6	8.9	14.1	25.9	26.7	**41.4**	**7.6**	23.4	36.6	**39.9**	69.5	**10.4**	24.1	44.7	51.3	78.8
Ours 3D	**8.9**	**14.9**	**25.4**	**29.9**	**45.8**	**7.6**	**15.9**	**37.2**	**41.7**	**53.8**	8.1	**13.4**	**24.8**	**24.9**	43.1	9.4	**20.3**	**35.2**	41.2	**67.4**	13.1	23.7	**44.5**	**50.9**	**78.3**

	Greeting					Phoning					Posing					Purchase					Sitting				
millisecond (ms)	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
ConvSeq2Seq [21]	30.4	58.6	110.0	122.8	198.9	22.4	38.4	65.0	75.4	133.2	22.4	42.1	87.3	106.1	187.3	28.4	53.8	82.1	93.1	142.4	24.7	50.0	88.6	100.4	182.3
ConvSeq2Seq 3D [21]	24.5	46.2	90.0	103.1	191.2	17.2	29.7	53.4	61.3	127.5	16.1	35.6	86.2	105.6	163.9	29.4	54.9	82.2	93.0	139.3	19.8	42.4	77.0	88.4	132.5
FC-GCN [23]	14.2	**27.7**	67.1	82.9	153.4	13.5	22.5	45.2	52.4	117.9	11.1	27.1	69.4	86.2	142.1	20.4	42.8	69.1	78.3	128.6	11.7	27.0	55.9	66.9	130.2
FC-GCN 3D [23]	14.5	30.5	74.2	89.0	148.4	11.5	20.2	37.9	43.2	94.3	9.4	23.9	66.2	82.9	143.5	19.6	38.5	64.4	**72.2**	127.2	10.7	24.6	50.6	62.0	119.8
Ours	13.4	31.2	69.3	86.1	133.2	11.7	18.3	**32.8**	44.1	87.9	**8.6**	**19.2**	59.4	84.2	141.7	18.2	39.1	63.2	75.2	121.4	9.8	25.2	48.9	59.4	**104.9**
Ours 3D	**9.6**	27.9	**66.3**	**78.8**	**129.7**	**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**	**9.2**	**23.1**	**47.2**	**57.7**	106.5

	Sitting down					Taking photo					Waiting					Walking dog					Walking together				
millisecond (ms)	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
ConvSeq2Seq [21]	23.9	39.9	74.6	89.8	189.3	18.4	32.1	60.3	72.5	156.4	24.9	50.2	101.6	120.0	221.5	56.4	94.9	136.1	156.3	234.1	21.1	38.5	61.0	70.4	156.3
ConvSeq2Seq 3D [21]	17.1	34.9	66.3	77.7	177.5	14.0	27.2	53.8	66.2	151.2	17.9	36.5	74.9	90.7	205.8	40.6	74.7	116.6	138.7	210.2	15.0	29.9	54.3	65.8	149.8
FC-GCN [23]	11.5	25.4	53.9	65.6	156.2	8.3	15.8	38.5	49.1	124.4	12.1	27.5	67.3	85.6	178.4	35.8	63.6	106.7	126.8	198.3	11.7	23.5	46.0	53.5	113.8
FC-GCN 3D [23]	11.4	27.6	56.4	67.6	163.9	6.8	15.2	38.2	49.6	125.7	9.5	22.0	57.5	73.9	157.2	32.2	58.0	102.2	122.7	185.4	8.9	18.4	35.3	44.3	102.4
Ours	10.8	24.2	49.7	61.4	146.1	**6.5**	14.3	32.3	46.7	117.9	**9.1**	21.5	50.9	68.7	144.2	26.5	**54.3**	94.7	119.2	168.3	10.3	20.6	34.9	45.3	98.7
Ours 3D	**9.3**	**21.4**	**46.3**	**59.3**	**144.6**	7.1	**13.8**	**29.6**	**44.2**	**116.4**	9.2	**17.6**	**47.2**	71.6	**127.3**	**25.3**	56.6	**87.9**	**99.4**	**143.2**	**8.2**	**18.1**	**31.2**	**39.4**	**79.2**

Table 2. Comparisons of 3D error on H3.6M dataset. For each method, we use two evaluation strategies: 1) train / test on angle-based samples, and then transfer the predicted angles to 3D position; 2) directly train / test on 3D coordinate sequence.

for short-term prediction (*i.e.*, 10 frames), and $1000ms$ for long-term prediction (*i.e.*, 25 frames). We evaluate the angle error and the 3D error on three benchmarks.

H3.6M: We first present qualitative comparison results on H3.6M dataset, as shown in Figure 4. For each subfigure, from the top to the bottom, we show the ground truth, and the prediction of ConvSeqSeq, FC-GCN, FC-GCN 3D, Ours, and Ours 3D. Note that FC-GCN 3D and Ours 3D are trained on 3D position-based skeleton sequences, while the others are based on angles. The Figure 4 (a) and (b) show the visualization of short-term prediction in "walking the dog" and "greeting" activities, and the Figure 4 (c) provides the long-term prediction results of "smoking" activity. The red rectangular indicates animations with distinct contrasts between the result of different methods, and the red circle or ellipse refer to unreasonable parts with ground truth. We observe that our method outperforms the competitive methods in both long-term and short-term prediction. Furthermore, the prediction of our method is closer to the ground truth than the baselines in almost all scenarios. This result evidences the superiority of our method.

We also further evaluate the angle error between the prediction and ground truth. The Table 1 shows a quantitative comparison for long-term and short-term prediction. We observe that the angle err or obtained by our method is smaller than that of the baseline methods in almost all cases. Such small errors are difficult to be detected by human eyes in human character animation. On the other hand, RNN-based Residual sup. and AGED gradually obtain larger errors due to the inevitable error accumulation problem and the extension of the prediction horizon. ConvSeq2Seq based on adversarial learning has achieved slightly better performance, but 2D convolution is not suitable for the 3D skeleton of non-euclidean human motion in essence. FC-GCN and our method are feedforward and can capture the connective relationship of skeleton sequence However, FC-GCN ignores the meaningful natural connection and regards human motion as a general data. Our method can not only explicitly learn the weights of natural connections, but also dynamically capture the implicit dependencies of skeleton sequences, thus achieving slightly better performance.

Angle-based representation is ambiguous, because poses with the same angle error may bring about differential distribution in 3D space. Moreover, the Euler angle cannot avoid the problem of gimbal lock. Therefore, to comprehensive verify our model, we also present the predicted 3D error using the following strategies: *First,* the predicted angles are converted into 3D position-based representation; *Second,* we directly train and test on 3D skeleton sequences. For example, in Table 2, we convert the predicted angle of FC-GCN into 3D space and then report the 3D error, while FC-GCN 3D directly takes the 3D coordinates of sequence, instead of the angles, as the input and output of the network. We observe that our method consistently surpasses the baselines (*i.e.*, ConvSeq2Seq, FC-GCN) under the conversion from predicted angles to the corresponding position-based result. When the 3D skeleton sequence is directly used to train and test for the competitive methods (*i.e.*, ConvSeq2Seq 3D, FC-GCN 3D) and Ours 3D, the proposed model also obtains realistic generalizations.

CMU and 3DPW MoCap: Similar to the above experimental strategy, we also investigate our method on CMU and 3DPW datasets with angles and 3D position as training samples respectively, as shown in Table 3, Table 4, and Table 5. The experimental results show that the proposed model substantially exceeds the baselines in both short-term and long-term prediction. These results once again confirm the effectiveness of our model for human motion prediction.

Comparison of training details with the state-of-the-art [23]. Previously, FC-GCN achieved the state-of-the-art results. However, FC-GCN is subject to a huge and unconstrained topology, which ignores the natural and meaning-

	Basketball					Basketball signal					Directing traffic					Jumping				
millisecond (ms)	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
ConvSeq2Seq [21]	0.37	0.62	1.07	1.18	1.95	0.32	0.59	1.04	1.24	1.96	0.25	0.56	0.89	1.00	2.04	0.39	0.60	1.36	1.56	2.01
FC-GCN [23]	0.33	0.52	0.89	1.06	1.71	**0.11**	0.20	0.41	0.53	1.00	0.15	0.32	0.52	0.60	2.00	0.31	0.49	1.23	1.39	1.80
Ours	**0.28**	**0.43**	**0.76**	**0.89**	**1.52**	0.12	**0.16**	**0.39**	**0.51**	**0.89**	**0.14**	**0.30**	**0.45**	**0.58**	**1.79**	**0.29**	**0.41**	**0.91**	**1.17**	**1.77**
	Running					Soccer					Walking					Wash window				
millisecond (ms)	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
ConvSeq2Seq [21]	**0.28**	0.41	0.52	**0.57**	**0.67**	0.26	0.44	0.75	0.87	1.56	0.35	0.44	**0.45**	**0.50**	0.78	0.30	0.47	0.80	1.01	1.39
FC-GCN [23]	0.33	0.55	0.73	0.74	0.95	0.18	0.29	0.61	0.71	1.40	0.33	0.45	0.49	0.53	0.61	0.22	0.33	0.57	0.75	1.20
Ours	0.31	**0.39**	**0.51**	0.68	0.87	**0.17**	**0.25**	**0.53**	**0.66**	**1.29**	**0.29**	**0.41**	0.54	0.61	**0.64**	**0.20**	**0.31**	**0.54**	**0.68**	**1.04**

Table 3. Quantitative comparisons of angle error for short-term and long-term prediction on 8 activities of the CMU MoCap dataset.

	Basketball					Basketball signal					Directing traffic					Jumping				
millisecond (ms)	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
ConvSeq2Seq 3D [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 3D [23]	14.0	25.4	49.6	61.4	106.1	3.5	**6.1**	11.7	15.2	53.9	7.4	**15.1**	31.7	42.2	152.4	16.9	34.4	76.3	96.8	164.6
Ours 3D	**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**
	Running					Soccer					Walking					Wash window				
millisecond (ms)	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
ConvSeq2Seq 3D [21]	**14.3**	**16.3**	**18.0**	**20.2**	**27.5**	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
FC-GCN 3D [23]	25.5	36.7	39.3	39.9	58.2	11.3	21.5	44.2	55.8	117.5	7.7	11.8	19.4	23.1	40.2	5.9	**11.9**	30.3	40.0	79.3
Ours 3D	15.2	19.7	23.3	35.8	47.4	**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**

Table 4. Quantitative comparisons of 3D error for short-term and long-term prediction on CMU MoCap dataset. In this case, the network directly take 3D-position based sequence as the input and output.

millisecond (ms)	200	400	600	800	1000
ConvSeq2Seq [21]	1.24	1.85	2.13	2.23	2.26
FC-GCN [23]	0.64	0.95	1.12	1.22	1.27
Ours	**0.57**	**0.72**	**1.07**	**1.18**	**1.25**
ConvSeq2Seq 3D [21]	71.6	124.9	155.4	174.7	187.5
FC-GCN 3D [23]	35.6	67.8	90.6	106.9	117.8
Ours 3D	**33.9**	**57.4**	**84.6**	**95.2**	**109.1**

Table 5. Quantitative comparisons of mean angle error and mean 3D error on whole testing set of 3DPW MoCap dataset.

	parameters	training time (minute / epoch)	testing time (millisecond / 25 frames)
FC-GCN [23]	$\approx 2.6\ M$	$\approx 4.1\ min$	$\approx 2.9\ ms$
Ours	$\approx$ **2.1** ***M***	$\approx$ **3.5** ***min***	$\approx$ **2.4** ***ms***

Table 6. Training details comparison with the state-of-the-art.

ful dependencies between human joints. Moreover, such a complicated graph structure also increase the model size. Therefore, we compare the training details (parameter number, training time, and testing time) between FC-GCN and our method with angle-based H3.6M dataset on NVIDIA 1070TI GPU. Our method converges at around 40 epoch, while FC-GCN is 50. Other results are shown in Table 6.

4.4. Ablation Analysis

We have run various ablation studies on H3.6M dataset to further explore the impact of the proposed modules. Specifically, we report the impact of (1) *residual connection* and (2) *adverarial discriminator*. In this paper, we propose the connective graph $\boldsymbol{A}_p$ and global graph $\boldsymbol{Q}$ to learn the dynamic relationships of skeleton sequence. Therefore, we also investigate using only (3) $\boldsymbol{A}_p$, $\boldsymbol{Q}$ or *their combination* respectively. The above results are shown in Table 7.

Our discriminator distinguishes the long sequence that concatenates the input sequence and the prediction or ground truth, as shown in Table 8. Therefore, we also analyze the influence of (4) *different inputs of discriminator*: (a) short sequence (25 frames): prediction and ground truth; (b) long sequence (50 frames): concatenation of input sequence and predicted or ground truth. The ablation studies evidence that the proposed components indeed benefit to the network to obtain more accurate predictions.

		mean angle error					mean 3D error				
resi	*adv*	80	160	320	400	1000	80	160	320	400	1000
yes		0.27	0.39	0.49	0.53	0.92	11.4	24.9	57.2	71.2	87.9
	yes	0.35	0.64	0.69	0.98	1.45	14.2	35.3	76.2	81.3	132.1
yes	*yes*	**0.24**	**0.37**	**0.45**	**0.52**	**0.89**	**10.4**	**22.2**	**56.7**	**64.2**	**81.3**
connective graph - $\boldsymbol{A}_p$	*global graph - $\boldsymbol{Q}$*										
yes		0.31	0.43	0.57	0.64	0.96	14.2	29.4	67.8	71.2	83.8
	yes	0.28	0.41	0.53	0.62	0.83	12.3	24.2	59.3	67.3	85.9
yes	*yes*	**0.24**	**0.37**	**0.45**	**0.52**	**0.89**	**10.4**	**22.2**	**56.7**	**64.2**	**81.3**

Table 7. **Top:** Impact of residual connection and adversarial leaning; **Bottom:**. Results on different graph construction.

	mean angle error					mean 3D error				
	80	160	320	400	1000	80	160	320	400	1000
(a) short sequence	0.50	0.64	0.71	0.82	0.99	11.2	24.1	57.8	68.7	89.3
(b) long sequence	**0.24**	**0.37**	**0.45**	**0.52**	**0.89**	**10.4**	**22.2**	**56.7**	**64.2**	**81.3**

Table 8. Influence of the different input of our discriminator.

5. Conclusion

In this paper, we have proposed a novel GCN approach to effectively forecast the future poses from given historical sequence. We parameterize the human structure through the learnable adjacency matrix and global graph. With the optimization process, the proposed model can not only capture the strength of the natural connection, but also adaptively extract the connectivities of geometrically separated joints. This data-driven method improves the flexibility of graph construction and ensures stable training, which is more suitable for human motion modeling. The final model exceeds current state-of-the-art performance on several large-scale human motion prediction benchmarks. In our future work, we will consider further exploration of combining bones and joints information.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NO. 61772272) and the Project of Science and Technology of Jiangsu Province of China under Grant BE2017031.

References

[1] CMU Graphics Lab: Carnegie-Mellon Motion Capture (Mocap) Database, http://mocap.cs.cmu.edu, 2003.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223, 2017.

[3] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *CoRR*, abs/1803.01271, 2018.

[4] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1499–149909, 2017.

[5] Aleksandar Bojchevski, Oleksandr Shchur, Daniel Zügner, and Stephan Günnemann. Netgan: Generating graphs via random walks. In *ICML*, 2018.

[6] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013.

[7] Judith Bütepage, Michael J. Black, Danica Kragic, and Hedvig Kjellström. Deep representation learning for human motion prediction and classification. In *CVPR*, pages 1591–1599, 2017.

[8] Hsu-kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1423–1432. IEEE, 2019.

[9] Qiongjie Cui, Beijia Chen, and Huaijiang Sun. Nonlocal low-rank regularization for human motion recovery based on similarity analysis. *Information Sciences*, 2019.

[10] Qiongjie Cui, Huaijiang Sun, Yupeng Li, and Yue Kong. A Deep Bi-directional Attention Network for Human Motion Recovery. In *IJCAI*, 2019.

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

[12] Pallabi Ghosh, Yi Yao, Larry S Davis, and Ajay Divakaran. Stacked spatio-temporal graph convolutional networks for action segmentation. *arXiv preprint arXiv:1811.10575*, 2018.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[14] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José M. F. Moura. Adversarial geometry-aware human motion prediction. In *ECCV*, 2018.

[15] Liang-Yan Gui, Yu-Xiong Wang, Deva Ramanan, and José M. F. Moura. Few-shot human motion prediction via meta-learning. In *ECCV*, 2018.

[16] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. In *NIPS*, 2017.

[17] 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.

[18] Ashesh Jain, Amir Roshan Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-RNN: Deep Learning on Spatio-Temporal Graphs. In *CVPR*, pages 5308–5317, 2015.

[19] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[20] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[21] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *CVPR*, 2018.

[22] Zhenguang Liu, Shuang Wu, Shuyuan Jin, Minghua Tang, Shijian Lu, Richard Zimmermann, and Li Chen Cheng. Towards Natural and Accurate Future Motion Prediction of Humans and Animals. In *CVPR*, 2019.

[23] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. 2019.

[24] Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *CVPR*, 2017.

[25] Alejandro Hernandez Ruiz, Juergen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. *arXiv preprint arXiv:1812.05478*, 2018.

[26] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, 2019.

[27] Yongyi Tang, Lin Ma, Wei Liu, and Wei-Shi Zheng. Long-term human motion prediction by modeling motion context and enhancing motion dynamics. In *IJCAI*, 2018.

[28] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018.

[29] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018.

[30] Huikai Wu, Junge Zhang, and Kaiqi Huang. Point cloud super resolution with adversarial residual graph networks. *ArXiv*, abs/1908.02111, 2019.

[31] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018.

[32] Jiaxuan You, Rex Ying, Xiang Ren, William L. Hamilton, and Jure Leskovec. Graphrnn: Generating realistic graphs with deep auto-regressive models. In *ICML*, 2018.

[33] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Graph neural networks: A review of methods and applications. *ArXiv*, abs/1812.08434, 2018.