Towards more realistic human motion prediction with attention to motion coordination

Pengxiang Ding, Jianqin Yin, *Member, IEEE*

arXiv:2404.03584v1 [cs.CV] 4 Apr 2024

Abstract—Joint relation modeling is a curial component in human motion prediction. Most existing methods rely on skeletal-based graphs to build the joint relations, where local interactive relations between joint pairs are well learned. However, the motion coordination, a global joint relation reflecting the simultaneous cooperation of all joints, is usually weakened because it is learned from part to whole progressively and asynchronously. Thus, the final predicted motions usually appear unrealistic. To tackle this issue, we learn a medium, called coordination attractor (CA), from the spatiotemporal features of motion to characterize the global motion features, which is subsequently used to build new relative joint relations. Through the CA, all joints are related simultaneously, and thus the motion coordination of all joints can be better learned. Based on this, we further propose a novel joint relation modeling module, Comprehensive Joint Relation Extractor (CJRE), to combine this motion coordination with the local interactions between joint pairs in a unified manner. Additionally, we also present a Multi-timescale Dynamics Extractor (MTDE) to extract enriched dynamics from the raw position information for effective prediction. Extensive experiments show that the proposed framework outperforms state-of-the-art methods in both short- and long-term predictions on H3.6M, CMU-Mocap, and 3DPW.

Index Terms—human motion prediction, joint relation modeling, motion coordination, enriched dynamics

I. INTRODUCTION

Human motion prediction aims to generate future skeleton sequences from given past observed ones. It is a crucial research field since it can help machines respond rapidly to unknown situations in the future. Thus this technique has attracted much attention in many scenarios, such as human-robot interaction [1], [2], [3], [4], autonomous driving [5], and pedestrian tracking [6], [7].

Predicting human motion is a challenging task because human motion is highly dynamic, non-linear, and more stochastic over time. Traditional works [8], [9] extend classical sequential models [10], [11] to motion prediction and achieved good performance in simple actions. Recently, those learning-based methods [12], [13], [14], [15], [16], [17], [18], [19], [20], [21] have proven to be more effective in the complicated motions due to their better nonlinear fitting ability. They are mainly divided into two categories: RNN-based methods [12], [13], [14], [15], [16], [17] and Feed-forward methods[18], [19], [20], [22], [21], [23]. RNN-based methods usually suffer from some inherent problems existing in RNN, e.g., the

Pengxiang Ding and Jianqin Yin are with the School of Artificial Intelligence of Beijing University of Posts and Telecommunications, No.10 Xitucheng Road, Haidian District, Beijing 100876, China. E-mail: dingpx2015@bupt.edu.cn, jqyin@bupt.edu.cn.
Corresponding author: Jianqin Yin.

prediction discontinuities and error-accumulation. Differently, feed-forward methods, including CNNs [19], [18] and GCNs [20], [22], [21], [23], can help alleviate those problems because the whole prediction process is not recursive. Besides, compared with RNN-based methods, feed-forward methods can model spatial and temporal relations simultaneously and exploit richer motion features. Thus, recent SOTA approaches mainly adopt feed-forward neural networks, and our work also falls under this category.
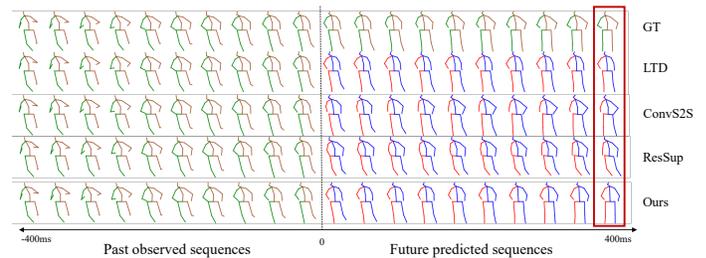

Fig. 1. Qualitative results of short-term predictions of motion "discussion" on H3.6M. From top to bottom, we show the ground truth, the results of LTD [22], ConvS2S [18], ResSup [14] and our approach. Compared with the result of our approach, the predicted motions of other works have the same problem: the limbs are uncoordinated which makes the predicted motion appear unrealistic.

Although promising results are achieved in previous works, there is still a universal problem in previous works: they mainly focus on the local interactive relations between joint pairs but overlook the motion coordination, a specific global joint relation that encodes the simultaneous cooperation of all joints. In prior works, joint relations are usually modeled according to skeletal structure [19], [20], [17], [16] or dynamic graphs [23], [22], [21]. They can exploit rich local joint relations between both spatial-connected and spatial-separated joint pairs. However, global joint relations are usually learned by fusing different body components' local motion features. In this way, the learned global joint relations can't reflect the simultaneous cooperation of all joints, and thus the final predicted motion usually appears unrealistic, e.g., the arms and legs are uncoordinated, as is shown in Fig. 1.

Therefore, in this paper, we propose to model motion coordination by exploring the simultaneous cooperation of all joints. To this end, we design a medium, called Coordination Attractor (CA), to correlate all joints simultaneously in an indirect way. In particular, the CA is learned by calculating the feature aggregation of all joints to represent the global motion features, which is next used to generate new relative joint features by subtracting raw features of each joint. In

this way, all joints are correlated through the CA and generate new relative joint relations. Especially, the new relative joint relations enhance the cooperative relations of joints by reducing the motion commonality of joints to exploit more pure motion coordination. And then, the resulting relative joint features are next used to calculate joints similarities to generate final joint relations. In this way, all joints are correlated simultaneous in an indirect way, and thus the motion coordination can be modeled explicitly. Furthermore, the global motion coordination of all joints is not used alone but is combined with the local interactions between joint pairs to exploit richer joint relations for more accurate and realistic predictions.

Additionally, it is beneficial to exploit the enriched motion dynamics for effective prediction. As is well known, the skeleton sequences only include the position information of joints, which is insufficient to convey the dynamics of motion. Previous works [26], [27] tended to introduce extra velocity information to represent the motion dynamics via the two-stream architecture. However, the velocity only extracts the dynamics from neighbor frames, and thus more fine-grained dynamics is hard to capture simply from velocity. Therefore, in this paper, we propose to enrich the joint representation by extracting more diverse dynamics existing in different timescales.

Based on the above two aspects, we present our framework to generate more realistic and accurate human motions. Given observed motion sequences, we first learn enriched dynamics from raw position information adaptively through Multitimescale Dynamics Extractor (MTDE). Next, we introduce the Comprehensive Joint Relation Extractor (CJRE), including a Local Interaction Extractor (LIE), a Global Coordination Extractor (GCE), and an Adaptive Feature Fusing Module (AFFM). The GCE is designed to present the global coordination of all joints, and the LIE is used to encode the local interactions between joint pairs. The above different joint relations are adaptively aggregated in the Adaptive Feature Fusing module (AFFM). Especially, we utilize the lateral connections between CJRE blocks for more fine-grained motion features. In this way, our proposed framework can capture more comprehensive joint relations and generate more diverse motion features for realistic and accurate predictions.

The main contributions of this paper are summarized as follows.

- We present to model motion coordination, a specific global joint relation that encodes the simultaneous cooperation of all joints, to enhance the realness of predicted motion. This motion coordination is further combined with the local interactions between joint pairs in a unified joint relation modeling module, CJRE, to extract richer joint relations for more realistic and accurate predictions.
- We also put forward an MTDE module to extract enriched dynamics from the raw input data for effective prediction.
- Our proposed framework outperforms most state-of-the-art methods for short and long-term motion prediction on three standard benchmark datasets: H3.6M, CMU-Mocap, and 3DPW.

II. RELATED WORK

Skeleton-based motion prediction has attracted increasing attention recently. Recent works using neural networks [12], [28], [29], [30], [31], [22], [13], [14], [15], [18], [19], [16], [17], [20], [23], [21], [32], [33], [34] have significantly outperformed traditional approaches [8], [9].

Human motion prediction. RNNs [13], [14], [15] are first used to predict human motion for their ability on sequence modeling. The first attempt was made by Fragkiadaki et al. [13], who proposed an Encoder-Recurrent-Decoder (ERD) model to combine encoder and decoder with recurrent layers. They encoded the skeleton in each frame to a feature vector and built temporal correlation recursively. Julieta et al. [14] introduced a residual architecture to predict velocities and achieved better performance. Chen et al. [12] propose to modify the RNN structure using a novel diffusion convolutional recurrent predictor to model spatialtemporal motion features for better prediction. However, these works all suffer from discontinuities between the observed poses and the predicted future ones. Though Gui et al. [15] proposed to generate a smooth and realistic sequence through adversarial training, it is hard to alleviate error-accumulation in a long-time horizon inherent to the RNNs scheme. A feedforward network was widely adopted to help alleviate those above questions because its prediction was not recursive and thus could avoid error-accumulation. Li et al. [18] introduced a convolutional sequence-to-sequence model that encodes the skeleton sequence as a matrix whose columns represent the pose at every time step. However, their spatiotemporal modeling is still limited by the convolutional filters' size. Recently, [19], [22] were proposed to consider global spatial and temporal features simultaneously. They all transformed temporal space to trajectory space to take the global temporal information into account. It contributes to capturing richer temporal correlation and thus achieved state-of-the-art results. In this paper, we follow this scheme but use different methods to model the spatial relations of joints.

Joint relation modeling. Previous work mainly focused on skeletal constraints to model correlations between joints. Jain et al. [16] first introduced a Structural-RNN model to explicitly model structural information relying on high-level spatiotemporal graphs. However, the graph is designed according to kinetic structure and is not flexible for different motions. Recently, some dynamic graph structures [22], [32], [20], [35] were developed to model more flexible joint relations. Mao et al. [22] used an adaptive graph to model motion, but it is still unreliable because the graph is initialized randomly without structure prior. Cai et al. [32] further combined kinematic structure with dynamic graph structure. Li et al. [20] used stacked GCNs to build the interaction of different scales structure in each layer to model the correlation of both neighbor and distant joints. However, those above learned joint relations mainly refer to the local interactions between joint pairs without considering global motion coordination of all joints, which usually makes the final predicted motion appear unnatural or unrealistic. Therefore, in this paper, we aim to model more comprehensive joint relations, including both global motion coordination of all joints and local interactions
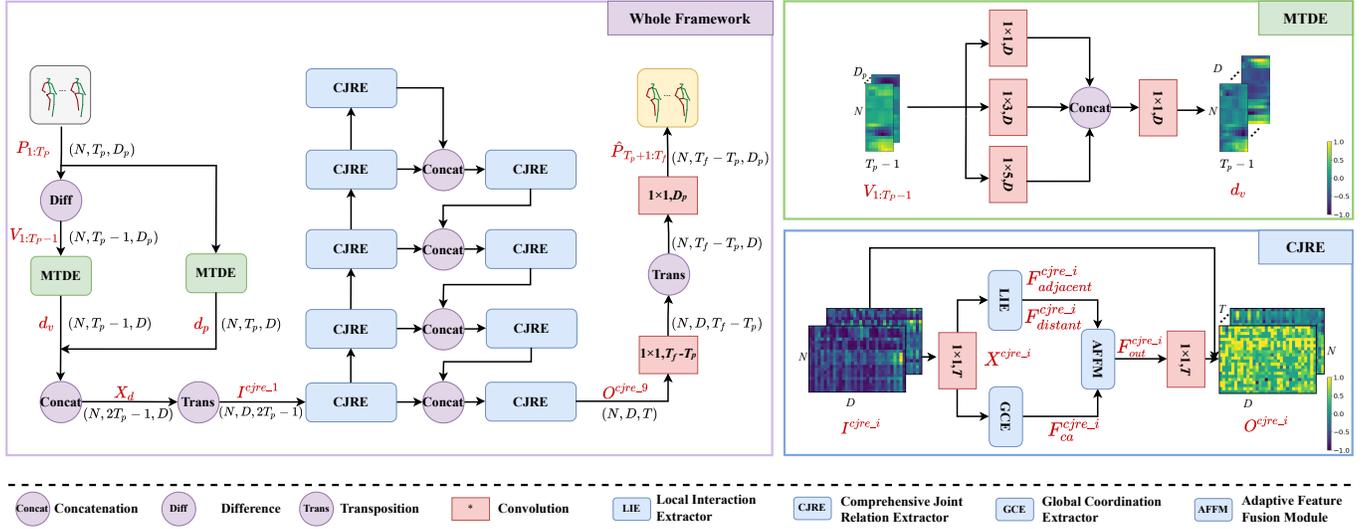
Fig. 2. The left panel describes the whole framework of our proposed framework and the right two panels represent the details of MTDE and CJRE. Based on the two-stream architecture, the MTDE module is used to extract the enriched motion dynamics. The CJRE module is adapted to encode the global coordination of all joints and local interactions between joint pairs through GCE and LIE, respectively. We here denote I^{cjre_i} and O^{cjre_i} as input and output of the ith CJRE module. AFFM is introduced to fuse features according to the channel-wise attention mechanism. The whole CJRE is built based on the bottleneck architecture of ResNet [24] for efficiency. Especially, lateral connections are used to offer fine-grained motion features inspired by U-Net [25]. At last, two 1×1 convolutions are successively used to transform temporal and spatial dimensions to get final prediction results.

between joint pairs.

Motion dynamics of skeleton sequences. The raw skeleton sequences are insufficient to convey the dynamics of motion because they only represent each joint's position information at each time step. Many attempts [26], [36], [27] proposed to extract enriched dynamic representation from raw input data. They tended to rely on the two-stream architecture to introduce velocity information. A drawback of these methods is that they only extract the dynamics from neighbor frames. Though Li et al. [26] enlarged the time horizon by convolution operation, it is still insufficient because different timescales encode different dynamics. Therefore, in this paper, we propose to extract the enriched through multi-timescale convolutions from raw motion sequences.

III. OUR METHOD

The proposed framework aims to generate more realistic and accurate human motions for effective prediction. The overall architecture is shown in Fig. 2. It mainly includes two components, Multi-timescale Dynamics Extractor (MTDE) and Comprehensive Joint Relation Extractor (CJRE). The MTDE extracts multi-timescale temporal information to achieve richer motion dynamics for prediction. The CJRE mines comprehensive joint relations to model the spatiotemporal evolution of human motion. There exist several components in the CJRE. The global Coordination Extractor (GCE) and Local Interaction Extractor (LIE) are proposed to model the global coordination of all joints and local interactions between joint pairs separately. The Adaptive Feature Fusion module (AFFM) is introduced to fuse different joint relations according to the

channel-wise attention mechanism. Especially, the lateral connections between CJRE blocks are designed to get more fine-grained motion features for more accurate predictions. Finally, two 1×1 convolutions are successively used to transform temporal and spatial dimensions to get final prediction.

A. Problem formulation

We denote the historical 3D skeleton-based poses as $P_{1:T_p} = [p_1, \cdots, p_{T_p}] \in \mathbb{R}^{N \times T_p \times D_p}$ and future poses as $P_{T_p+1:T_f} = [p_{T_p+1}, \cdots, p_{T_f}] \in \mathbb{R}^{N \times (T_f-T_p) \times D_p}$, where $p_t \in \mathbb{R}^{N \times D_p}$ represents the 3D pose at time t with N joints. The $D_p = 3$ depicts the dimension of joint coordinates. Our goal is to generate predicted poses, $\hat{P}_{T_p+1:T_f} = M(P_{1:T_p})$ through the proposed framework $M(\cdot)$.

B. Multi-timescale Dynamics Extractor (MTDE)

Motion dynamics contains more motion cues compared with the position information in the raw motion sequences. It encodes the evolution of motion and thus is helpful to anticipate future motion trends. However, most previous works didn't make use of this modality information. They tended to introduce the velocity as another input branch to enrich the input features and extract dynamics. Although it makes sense to some extent, it is insufficient only to use velocity to represent motion dynamics because more detailed and fine-grained dynamics couldn't be captured simply from the velocity. We take Fig. 3 as an example. In this motion sequence, the duration time of the head movement is about four frames, while the left foot is ten frames. It shows two important details. First, the motion dynamics of different joints in a motion

sequence usually are various. This observation reminds us that it is unsuitable to treat all joints equally, like the operation in calculating velocity. Second, unlike the velocity which only encodes the dynamics of adjacent frames, dynamics existing in different temporal scales could offer more diverse motion cues for accurate prediction.
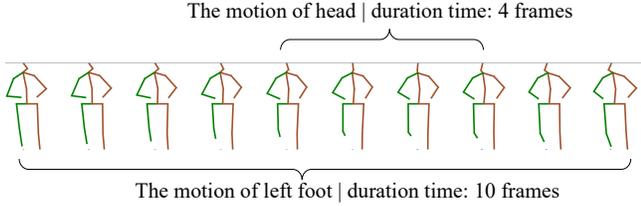

The motion of head | duration time: 4 frames

The motion of left foot | duration time: 10 frames

Fig. 3. The obeservation of motion. It shows the duration time of the head is four frames, while the left foot is ten frames.

To address these issues, we propose two improvements based on the two-stream architecture in the Multi-timescale Dynamics Extractor (MTDE). Firstly, we extend the feature dimensions based on the raw joint position and velocity. In this way, we can extract more dynamic motion cues in high dimensional space, which bare position and velocity in 3D space can't provide. Secondly, considering the dynamics of different joints are different even if in the same motion, we here apply multiple temporal convolutions to model more diverse motion dynamics existing in different temporal scales. Notably, we here only learn intra-joint dynamics to extract the motion dynamics without the interference of other joints. In this way, our proposed MTDE could extract richer motion dynamics to enrich intra-joint features, which is also beneficial for later inter-joint relation modeling.

As is shown in Fig. 2, for the input $P_{1:T_p} \in \mathbb{R}^{N \times T_p \times D_p}$, we first get the velocity of raw motion sequence $V_{1:T_p-1} = [v_1, \cdots, v_{T_p-1}] \in \mathbb{R}^{N \times (T_p-1) \times D_p}$ by calculating the difference between adjacent frames of $P_{1:T_p}$. These two different branches $P_{1:T_p}$ and $V_{1:T_p-1}$ are all connected to a MTDE module to encode enriched dynamics through multi-timescale temporal convolutions respectively. Formally,

$$
\begin{aligned}
v_t &= (p_{t+1} - p_t), t = 1, ..., T_p - 1 \\
d_{k_i}^p &= \sigma(W_{k_i}^p * P_{1:T_p}), i = 1, 2, 3 \\
d_{k_i}^v &= \sigma(W_{k_i}^t * V_{1:T_p-1}), i = 1, 2, 3 \\
d_p &= W_p * Concat([d_{k_1}^p, d_{k_2}^p, d_{k_3}^p]), \\
d_v &= W_v * Concat([d_{k_1}^v, d_{k_2}^v, d_{k_3}^v]), \\
X_d &= d_p \oplus d_v
\end{aligned}
\tag{1}
$$

Here, the $\sigma(\cdot)$ is the activation function. $W_{k_i}^p$ and $W_{k_i}^v$ indicate the different $1 \times k_i$ temporal convolution kernel with different timescale k_i (the size of output channel is D). $Concat(\cdot)$ is the concatenation operation along the channel. W_p and W_v denote the 1×1 convolution kernels used to fuse multi-timescale dynamics. $*$ and $\oplus$ denote the convolution operator and concatenation along temporal dimensions. For different input branches $P_{1:T_p}$ and $V_{1:T_p-1}$, $d_p \in \mathbb{R}^{N \times T_p \times D}$

and $d_v \in \mathbb{R}^{N \times (T_p-1) \times D}$ encode enriched dynamics existing in different timescales through fusing the features from different temporal convolutions respectively. To make use of different features of two branches, we synthesize d_p and d_v along temporal dimensions to get the representation $X_d \in \mathbb{R}^{N \times (2T_p-1) \times D}$. This operation enables the model to capture more detailed and fine-grained motion cues for later prediction.

C. Global Coordination Extractor (GCE)

The global coordination of all joints plays an essential role in human motion. It describes the mutual constraints of all joints during motion and thus could offer richer motion cues to predict human motion. However, previous works mainly focused on modeling local interactions of joint pairs, and thus the global coordination wasn't exploited well. Besides, there exist two major observations as to the global coordination of all joints. First, the global coordination essentially reflects the relative relations of all joints without global motion trends, which is not explored in previous works. Second, the learned global relations in most previous works are predefined and fixed, which is insufficient to represent the diversity of global coordination, such as balance, inertia, etc.

As to the above two observations, we propose the global Coordination Extractor (GCE) to model global coordination of all joints. It mainly contains two important components: Feature Normalization Unit and Multi-head Self-attention Unit. In the Feature Normalization Unit, we aim to extract relative joint motion representation without the interference of global motion trends for later global coordination relation modeling. In the Multi-head Self-attention Unit, multiple relation graphs of joints are built by calculating the self-attention of the learned relative joint representation. Notably, we learn the multiple relation graphs to combine different relative relations to extract richer joint relations. Next, we will illustrate more details in Fig. 4. Notably, in the following part, all of the variables are simplified by removing superscript "$cjre\_i$". For example, we use X represents the ith CJRE module's variable X^{cjre_i}.

1) Definition of global motion trends: As we discussed above, the global motion trends in this paper can be regarded as that the trajectory of global coordination center. Notably, we here learn it in high-dimensional trajectory space instead of in 3D space. The reason is that it is better to exploit more temporal consistency directly on trajectory space than 3D space, as illustrated in TrajCNN[19]. To better introduce the trajectory of global coordination center, we first define it in the 3D space mathematically. Specifically, we denote the pose sequence as $P_{1:T_p} = [p_1, \cdots, p_{T_p}] \in \mathbb{R}^{N \times T_p \times D_p}$ where p_i represents the pose of ith frame. The global coordination center cc_i of each pose can be regarded as the collective effect of all joints. Notably, considering that a skeleton is a non-rigid object, we use nonlinear transformation $nt(\cdot)$ to achieve cc_i, which is formulated by formula 2.

$$
\begin{aligned}
p_i &= [j_1, ..., j_N] \in \mathbb{R}^{N \times D_p} \\
W &= [w_1, ..., w_N] \in \mathbb{R}^{1 \times N} \\
cc_i &= nt(p_i) = \sigma(W \times p_i) \in \mathbb{R}^{1 \times D_p}
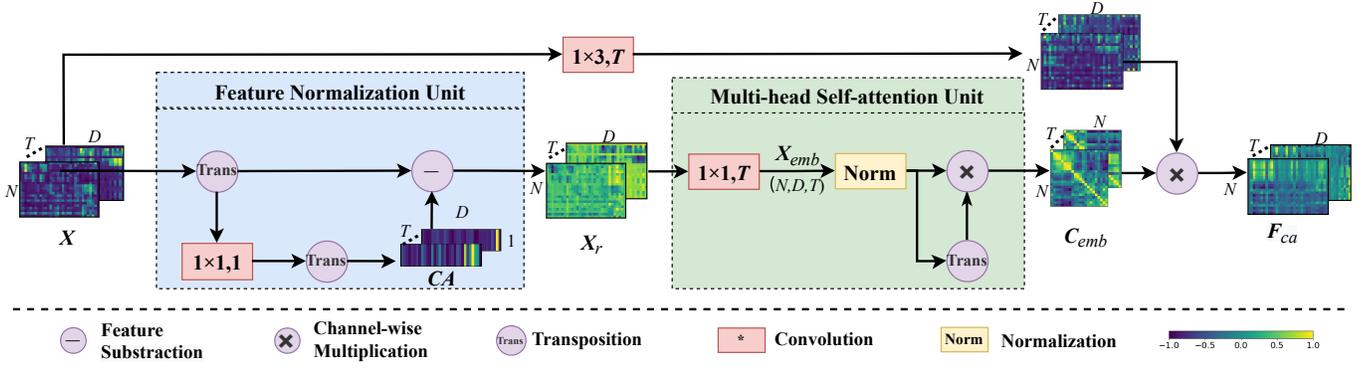\end{aligned}
\tag{2}
$$

Fig. 4. The overall architecture of GCE. It mainly contains two parts. The Feature Normalization Unit is designed to extract relative joint motion representation without the interference of global motion trends for later global coordination relation modeling. The Multi-head Self-attention Unit is proposed to generate multiple relation graphs of joints to extract richer global coordination. (To simplify the representation, X and F_{ca} are used to represent X^{cjre_i} and $F_{ca}^{cjre\_i}$, respectively.)

where $\sigma(\cdot)$ is the activation function and W is the weight matrix. In this way, the trajectory of global coordination center of all frames can be represented as $CC = nt(P'_{1:T_p}) = [cc_1, \cdots, cc_{T_p}] \in \mathbb{R}^{1 \times D_p \times T_p}$. (Notably, we here use $P'_{1:T_p} \in \mathbb{R}^{N \times D_p \times T_p}$ through transposing the dimension D_p and T_p of $P_{1:T_p}$ to keep the dimension consistency.)

Next, we extend the definition from 3D space to trajectory space. Concretely, we first get the features $X = f(P_{1:T_p}) \in \mathbb{R}^{N \times D \times T}$ through the feature extractor f. Here D and T represent the size of new features dimensions on trajectory space. Similarly, we can use the nonlinear transformation $nt(\cdot)$ to get the trajectory features of the global coordination center $CA = nt(f(P_{1:T_p})) \in \mathbb{R}^{1 \times D \times T}$. Notably, we name the trajectory features of the global coordination center in trajectory space as Coordination Attractor CA to distinguish it with the trajectory of the global coordination center CC in 3D space.

Furthermore, we illustrate the equivalent form of nonlinear transformation nt. Formally,

$$
\begin{aligned}
CA &= nt(X) \\
&= Trans(\sigma(W_{ca} * Trans(X)))
\end{aligned}
\tag{3}
$$

The $\sigma(\cdot)$ is the activation function. W_{ca} is a 1×1 convolution kernel (the size of output channel is 1) and $*$ denotes the convolution operator. We use $Trans(\cdot)$ to transpose the joint dimension and the temporal dimension. Because the size of the input channel is N and the size of the output channel is 1, the output of the convolution operator is the global response of the N joint features. In other words, the convolution operator computes the average weight of N feature maps, which is the equivalent form of the nonlinear transformation $nt(\cdot)$ listed on formula 2. We also use another $Trans(\cdot)$ to transpose the joint dimension and the temporal dimension back.

2) Feature Normalization Unit: This module aims to generate relative joint representation without global motion trends for later coordination modeling, as is illustrated in Fig 4. Specifically, we first learn the Coordination Attractor (CA) to characterize the global motion features according to formula 3. Subsequently, CA is used as a medium to build a new

relative joint representation $X_r \in \mathbb{R}^{N \times D \times T}$ indirectly through conducting feature subtraction. Formally,

$$
X_r = X - CA
\tag{4}
$$

In this way, the effect of global motion trends can be removed and the global coordination of all joints can be better modelled.

3) Multi-head Self-attention Unit: We aim to generate multiple joint relations which reflect global coordination by measuring the similarities of new relative features of each joint. Specifically, the whole process are shown as follows:

$$
\begin{aligned}
X_{emb} &= \sigma(W_{emb} * X_r) = \{x_t^{emb}\}_{t=1}^T \\
C_{emb} &= \{c_t^{emb}\}_{t=1}^T \\
c_t^{emb} &= SelfAtt(x_t^{emb})
\end{aligned}
\tag{5}
$$

The $\sigma(\cdot)$ is the activation function. $SelfAtt(\cdot)$ is the function to calculate the self-attention of all joints in each feature map. W_{emb} is a 1×1 convolution kernel (the size of output channel is T) and $*$ denotes the convolution operator. $x_t^{emb} \in \mathbb{R}^{N \times D}$ and $c_t^{emb} \in \mathbb{R}^{N \times N}$ are the feature map of $X_{emb} \in \mathbb{R}^{N \times D \times T}$ and $C_{emb} \in \mathbb{R}^{N \times N \times T}$ respectively. For the new relative joint features X_r, we first use W_{emb} to learn a specific embedding for relation modeling. Next, we aim to calculate the relative joint relations C_{emb}. Notably, we calculate the relative relations on each feature map because each feature map encodes specific spatiotemporal features and should focus on different joint relations. We take x_t^{emb} as an example. For x_t^{emb}, each row vector represents the features of one joint. Therefore, we can calculate the cosine similarity between all row vector pairs to illustrate the correlation between joint pairs. The reasons why we choose cosine similarity are: (1) this metric contains angle information that corresponds to the physical relations of joints; (2) the value is limited into $[-1, 1]$, which avoids the violent variance. In this way, we could generate diverse global coordination through multiple relations graphs.

The last step is to calculate the coordinated motion features according to the joint relations C_{emb}. Considering each feature map of C_{emb} represents one specific coordination, we apply channel-wise multiplication to make use of different relation graphs. Specifically, 1×3 convolution W_{intra} is used to extract intra-joint features and then combine with the guidance of C^{emb} to get the final features $F_{ca} \in \mathbb{R}^{N \times D \times T}$.

$$F_{ca} = C_{emb} \odot (\sigma(W_{intra} * X)) \qquad (6)$$

where $\odot$ represents channel-wise multiplication.
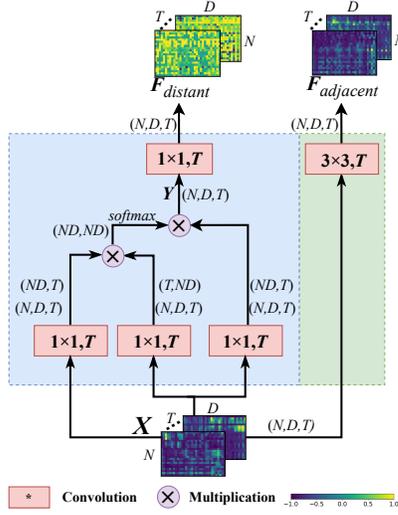
D. Local Interaction Extractor (LIE)

Fig. 5. The implementation of Local Interaction Extractor (LIE). The left is the path using a Non-local block without residual connection to learn the relations between distant joint pairs. The right is the the path with convolutions to learn the relations between adjacent joint pairs. (To simplify the representation, X, $F_{adjacent}$ and $F_{distant}$ are used to represent X^{cjre_i}, $F_{adjacent}^{cjre\_i}$, $F_{distant}^{cjre\_i}$ respectively.)

Local Interaction Extractor (LIE) is used to learn local interactions between joint pairs, including adjacent and distant joints. The local connection via bones brings spatial correlation for adjacent joints. For distant joints, some joints may have a strong correlation even if they are not directly connected, e.g., left hand and right hand are tightly correlated during "eating". Therefore, these two relations are equally important for effective prediction.

As is shown in Fig. 5, given an input $X \in \mathbb{R}^{N \times D \times T}$ which is the same as GCE, there exist two main paths to separately learn the relations between adjacent joint pairs and distant joint pairs.

For the adjacent joints, we here use the 3×3 convolution to combine the co-occurrence features of adjacent joints. In this way, local interactions between adjacent joints can be built.

$$F_{adjacent} = \sigma(W_{adjacent} * X) \qquad (7)$$

For the distant joints, considering the spatially separated joints are often far away from each other, we here adopt the self-attention mechanism used in the Non-local [37] block to

exploit their relations without the limitation of their arrangements in the feature maps.

Specifically, the input $X = [q_1, ..., q_{N \times D}] \in \mathbb{R}^{N \times D \times T}$, where $q_i \in \mathbb{R}^T$ represents the features of i_{th} pixel in all feature maps. To encode the relations between i and different position j, we calculate the similarity of the q_i and q_j as the response of position j to position i. In this way, the features of distant joints are correlated and the sum of all other position j's response form the resulting response y_i. Formally,

$$\begin{aligned} g(q_i) &= W_g q_i \\ f(q_i, q_j) &= softmax(\theta(q_i)^T \varphi(q_j)) \\ y_i &= \frac{1}{C(X)} \sum_{\forall j} f(q_i, q_j) g(q_j) \end{aligned} \qquad (8)$$

where θ, φ, g are 1×1 convolutions and $i, j \in [1, ..., N \times D]$. $C(X)$ is the normalization parameter. θ and φ are used to learn the similarity between different positions. g is used to learn an embedding representation of raw input X. The result y_i is calculated by the weighted sum of other positions' effect and finally constitutes the output $Y = [y_1, ..., y_{N \times D}] \in \mathbb{R}^{N \times D \times T}$. In this way, the new feature Y gets the relations of spatially separated joints without limiting their arrangements in the feature maps. And thus it can be used to calculate the final features of distant joints through the convolution kernel $W_{distant}$ for further feature extraction.

$$F_{distant} = \sigma(W_{distant} * Y) \qquad (9)$$

E. Adaptive Feature Fusing Module (AFFM)

The different motions will have a respective preference for local interactions between joint pairs and global coordination of all joints. For example, dynamic actions like 'walking' may pay more attention to the local interactions between joint pairs of which the movement are more obvious while static actions like 'sitting' may focus on the entire structure of all joints. Thus, we aim to measure the importance of different features in AFFM module.

Fig. 6. The implementations of Adaptive Feature Fusing Module (AFFM). The features learnt from previous block are fused with channel attention machanism. (To simplify the representation, $F_{adjacent}$, $F_{distant}$, F_{ca}, F_{out} are used to represent $F_{adjacent}^{cjre\_i}$, $F_{distant}^{cjre\_i}$, $F_{ca}^{cjre\_i}$ and $F_{out}^{cjre\_i}$ respectively.)

Specifically, as is shown in Fig. 6, we first combine different features extracted from GCE and LIE along the channel

TABLE I
SHORT-TERM PREDICTION OF 8 SUB-SEQUENCES PER ACTIONON ON H3.6M. WHERE "MS" DENOTES "MILLISECONDS".

motion	Walking				Eating				Smoking				Discussion			
time(ms)	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
ResSup [14]	23.8	40.4	62.9	70.9	17.6	34.7	71.9	87.7	19.7	36.6	61.8	73.9	31.7	61.3	96.0	103.5
ConvS2S [18]	17.1	31.2	53.8	61.5	13.7	25.9	52.5	63.3	11.1	21.0	33.4	38.3	18.9	39.3	67.7	75.7
LTD [22]	8.9	15.7	29.2	33.4	8.8	18.9	39.4	47.2	7.8	14.9	25.3	28.7	9.8	22.1	39.6	**44.1**
LPJP [32]	7.9	14.5	29.1	34.5	8.4	18.1	37.4	45.3	6.8	13.2	24.1	**27.5**	8.3	21.7	43.9	48.0
TrajCNN [19]	8.2	14.9	30.0	35.4	8.5	18.4	37.0	44.8	6.3	**12.8**	**23.7**	27.8	7.5	**20.0**	41.3	47.8
Ours	**7.2**	**13.7**	**25.6**	**31.0**	**7.7**	**16.7**	**35.8**	**44.2**	**6.3**	13.3	24.5	29.7	**7.5**	20.3	**38.7**	44.7

motion	Direction				Greeting				Phoning				Posing			
time(ms)	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
ResSup [14]	36.5	56.4	81.5	97.3	37.9	74.1	139.0	158.8	25.6	44.4	74.0	84.2	27.9	54.7	131.3	160.8
ConvS2S [18]	22.0	37.2	59.6	73.4	24.5	46.2	90.0	103.1	17.2	29.7	53.4	61.3	16.1	35.6	86.2	105.6
LTD [22]	12.6	24.4	48.2	58.4	14.5	30.5	74.2	89.0	11.5	20.2	37.9	43.2	9.4	23.9	66.2	82.9
LPJP [32]	11.1	22.7	48.0	58.4	13.2	28.0	64.5	77.9	10.8	19.6	37.6	46.8	8.3	22.8	65.6	81.8
TrajCNN [19]	9.7	22.3	50.2	61.7	12.6	28.1	67.3	80.1	10.7	18.8	37.0	43.1	6.9	21.3	62.9	78.8
Ours	**9.3**	**21.1**	**45.0**	**55.0**	**11.2**	**23.9**	**63.4**	**79.6**	10.2	**18.5**	**34.3**	**38.5**	**6.8**	20.5	**60.6**	**76.6**

motion	Purchasing				Sitting				Sitting down				Taking photo			
time(ms)	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
ResSup [14]	40.8	71.8	104.2	109.8	34.5	69.9	126.3	141.6	28.6	55.3	101.6	118.9	23.6	47.4	94.0	112.7
ConvS2S [18]	29.4	54.9	82.2	93.0	19.8	42.4	77.0	88.4	17.1	34.9	66.3	77.7	14.0	27.2	53.8	66.2
LTD [22]	19.6	38.5	64.4	72.2	10.7	24.6	50.6	62.0	11.4	27.6	56.4	67.6	6.8	15.2	38.2	49.6
LPJP [32]	18.5	38.1	**61.8**	**69.6**	9.5	23.9	49.8	61.8	11.2	29.9	59.8	68.4	6.3	14.5	38.8	49.4
TrajCNN [19]	17.1	**36.1**	64.3	75.1	9.0	23.0	49.4	62.6	10.7	28.8	55.1	62.9	**5.4**	**13.4**	**36.2**	**47.0**
Ours	**17.1**	38.0	65.0	73.0	**7.8**	**19.9**	**44.9**	**56.4**	9.2	**23.7**	**47.7**	**59.4**	5.6	14.3	37.6	48.9

motion	Waiting				Walking dog				Walking Together				Average			
time(ms)	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
ResSup [14]	29.5	60.5	119.9	140.6	60.5	101.9	160.8	188.3	23.5	45.0	71.3	82.8	30.8	57.0	99.8	115.5
ConvS2S [18]	17.9	36.5	74.9	90.7	40.6	74.7	116.6	138.7	15.0	29.9	54.3	65.8	19.6	37.8	68.1	80.2
LTD [22]	9.5	22.0	57.5	73.9	32.2	58.0	102.2	122.7	8.9	18.4	35.3	44.3	12.1	25.0	51.0	61.3
LPJP [32]	8.4	21.5	53.9	69.8	22.9	50.4	100.8	119.8	8.7	18.3	34.2	44.1	10.7	23.8	50.0	60.2
TrajCNN [19]	8.2	21.0	53.4	68.9	23.6	52.0	98.1	116.9	8.5	18.5	33.9	43.4	10.2	23.2	49.3	59.7
Ours	**7.7**	**18.8**	**48.0**	**64.7**	22.0	49.2	**90.9**	110.0	**7.8**	**17.3**	**32.1**	43.3	**9.6**	**22.0**	**46.2**	**57.0**

dimension. The next global average pooling squeeze each feature map as a scalar. In this way, each feature map can be encoded into a global feature. Then to learn the importance ratio between channels, we use the nonlinear transformation by two 1×1 convolution layers and use a sigmoid layer to get the final importance ratio. Formally,

$$F_{concat} = Concat([F_{distant}, F_{ca}, F_{adjacent}])$$
$$ratio = Sigmoid(W_{a2} * \sigma(W_{a1} * GAP(F_{concat}))) \quad (10)$$
$$F_{out} = ratio \odot F_{concat}$$

where $Concat(\cdot)$ is the concatenation operation along the channel. GAP is the global average pooling. W_{a1} and W_{a2} indicate the different convolution kernel to be used to nolinear transformation. The $\sigma(\cdot)$ is the activation function. The $Sigmoid(\cdot)$ function is used to limit the value of $ratio$ in $[0, 1]$. The $\odot$ represents channel-wise multiplication. At last, channel attention mechanism can be used to conduct channel-wise multiplication between ratio and concatenated features F_{concat} to reform new features. In this way, those features with higher important ratios can be enhanced and the later prediction could pay more attention to those enhanced features, which is beneficial to more precise prediction.

F. Loss Function

Following [19], [22], we make use of the Mean Per Joint Position Error (MPJPE). In particular, for one training sample, the loss is as follows:

$$L = \frac{1}{N \times (T_f - T_p)} \sum_{i=T_p+1}^{T_f} \sum_{j=1}^{N} \| P_{i,j} - \hat{P}_{i,j} \|_2 \quad (11)$$

where $\hat{P}_{i,j} \in R^3$, representing the 3D coordinates of the jth joint of the ith human pose, is the predicted result and $P_{i,j} \in R^3$ is the ground truth.

IV. EXPERIMENTS

In this section, we first introduce the datasets used in our experiments and the implementation details of our work. Then, we compare our method with baselines. Next, we carry out some experiments to analyze the contributions of the proposed method and visualize the predictive performance of our model.

A. Datasets and Implementation Details

Human3.6M [38] is the most widely used benchmark for motion prediction. It involves 15 actions performed by professionals, and each human pose involves a 32-joint skeleton. Following [22], [19], we compute the joint's 3D coordinates by applying forward kinematics and down-sample the motion sequence to 25 frames per second. After removing the global rotation, translation and constant 3D coordinates of each human pose, there remains 22 joints. We test our method on subject 5(S5). Considering [32] and [27] show the setting of 8 random sub-sequences may lead to high variance, we test our model with two different division methods in previous works. One is 8 random sub-sequences per action on subject 5 (S5) and another is 256 sub-sequences per action.

3DPW [39] The 3D Pose in the Wild dataset(3DPW) [39] consists of challenging indoor and outdoor actions. The dataset consists of various activities such as shopping, doing sports, and hugging, including 60 sequences and more than 51k frames. For a fair comparison, we evaluate the whole test sets.

TABLE II
LONG-TERM PREDICTION OF 8 SUB-SEQUENCES PER ACTIONON ON H3.6M. WHERE "MS" DENOTES "MILLISECONDS".

motion	Walking		Eating		Smoking		Discussion		Directions		Greeting		Phoning		Posing	
time(ms)	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000
LTD [22]	42.2	51.3	**56.5**	**68.6**	32.3	60.5	**70.4**	103.5	85.8	109.3	91.8	87.4	65.0	113.6	113.4	220.6
TrajCNN [19]	37.9	46.4	59.2	71.5	32.7	58.7	75.4	**103.0**	**84.7**	104.2	91.4	**84.3**	62.3	113.5	111.6	**210.9**
Ours	**35.5**	**42.7**	57.3	70.3	**30.9**	**55.0**	74.3	105.7	89.7	**103.5**	**91.1**	90.5	**59.1**	**110.5**	**107.3**	211.9
motion	Purchases		Sitting		Sitting down		Taking photo		Waiting		Walking Dog		Walking Together		Average	
time(ms)	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000
LTD [22]	94.3	130.4	79.6	114.9	82.6	140.1	68.9	87.1	100.9	167.6	136.6	174.3	**57.0**	85.0	78.5	114.3
TrajCNN [19]	84.5	**115.5**	81.0	116.3	79.8	**123.8**	**73.0**	**86.6**	92.9	165.9	141.1	181.3	57.6	**77.3**	77.7	110.6
Ours	**82.1**	117.6	**73.1**	**105.1**	**78.0**	126.1	75.9	88.9	**85.9**	**154.4**	**130.2**	**170.7**	57.1	82.2	**75.1**	**109.0**

TABLE III
SHORT-TERM PREDICTION OF 256 SUB-SEQUENCES PER ACTIONON ON H3.6M. WHERE "MS" DENOTES "MILLISECONDS".

motion	Walking				Eating				Smoking				Discussion			
time(ms)	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
ResSup [14]	23.2	40.9	61.0	66.1	16.8	31.5	53.5	61.7	18.9	34.7	57.5	65.4	25.7	47.8	80.0	91.7
ConvS2S [18]	17.1	33.5	56.3	63.6	11.0	22.4	40.7	48.4	11.6	22.8	41.3	48.9	17.1	34.5	64.8	77.6
LTD [22]	11.1	21.4	37.3	42.9	7.0	14.8	29.8	37.3	7.5	15.5	30.7	37.5	10.8	24.0	52.7	65.8
HRI [27]	10.0	19.5	**34.2**	**39.8**	6.4	14.0	28.7	36.2	7.0	14.9	29.9	36.4	10.2	23.4	52.1	65.4
Ours	**9.4**	**18.9**	34.5	41.3	**5.6**	**13.0**	**27.5**	**35.2**	**6.2**	**13.7**	**28.3**	**35.4**	**8.8**	**21.8**	**50.5**	**63.8**
motion	Direction				Greeting				Phoning				Posing			
time(ms)	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
ResSup [14]	21.6	41.3	72.1	84.1	31.2	58.4	96.3	108.8	21.1	38.9	66.0	76.4	29.3	56.1	98.3	114.3
ConvS2S [18]	13.5	29.0	57.6	69.7	22.0	45.0	82.0	96.0	13.5	26.6	49.9	59.9	16.9	36.7	75.7	92.9
LTD [22]	8.0	18.8	43.7	54.9	14.8	31.4	65.3	79.7	9.3	19.1	39.8	49.7	10.9	25.1	59.1	75.9
HRI [27]	7.4	18.4	44.5	56.5	13.7	30.1	63.8	78.1	8.5	18.3	39.0	49.2	10.2	24.2	58.5	75.8
Ours	**6.3**	**16.9**	**42.1**	**54.0**	**11.9**	**27.9**	**61.3**	**76.1**	**7.6**	**17.1**	**37.4**	**47.6**	**8.4**	**21.9**	**54.8**	**71.7**
motion	Purchasing				Sitting				Sitting down				Taking photo			
time(ms)	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
ResSup [14]	28.7	52.4	86.9	100.7	23.8	44.7	78.0	91.2	31.7	58.3	96.7	112.0	21.9	41.4	74.0	87.6
ConvS2S [18]	20.3	41.8	76.5	89.9	13.5	27.0	52.0	63.1	20.7	40.6	70.4	82.7	12.7	26.0	52.1	63.6
LTD [22]	13.9	30.3	62.2	75.9	9.8	20.5	44.2	55.9	15.6	31.4	59.1	71.7	8.9	18.9	41.0	51.7
HRI [27]	13.0	29.2	**60.4**	**73.9**	9.3	20.1	44.3	56.0	14.9	30.7	59.1	72.0	8.3	18.4	40.7	51.5
Ours	**11.3**	**27.6**	60.6	74.9	**8.0**	**18.1**	**41.1**	**52.7**	**13.0**	**28.5**	**56.5**	**70.2**	**7.4**	**17.0**	**39.3**	**50.2**
motion	Waiting				Walking dog				Walking Together				Average			
time(ms)	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
ResSup [14]	23.8	44.2	75.8	87.7	36.4	64.8	99.1	110.6	20.4	37.1	59.4	67.3	25.0	46.2	77.0	88.3
ConvS2S [18]	14.6	29.7	58.1	69.7	27.7	53.6	90.7	103.3	15.3	30.4	53.1	61.2	16.6	33.3	61.4	72.7
LTD [22]	9.2	19.5	43.3	54.4	20.9	40.7	73.6	86.6	9.6	19.4	36.5	44.0	11.2	23.4	47.9	58.9
HRI [27]	8.7	19.2	43.4	54.9	20.1	40.3	73.3	86.3	8.9	18.4	35.1	**41.9**	10.4	22.6	47.1	58.3
Ours	**7.4**	**17.0**	**41.0**	**52.3**	**17.4**	**37.0**	**71.2**	**85.0**	**8.0**	**17.1**	**33.7**	41.9	**9.1**	**21.0**	**45.3**	**56.8**

CMU-Mocap [40] The CMU mocap dataset mainly includes five categories, naming "human interaction", "interaction with environment", "locomotion", "physical activities & sports" and "situations & scenarios". Be consistent with [22], [19], we select 8 detailed actions: "basketball", "basketball signal", "directing traffic", "jumping", "running", "soccer", "walking" and "washing window". We evaluate our model with the same approach as we do for H3.6M.

Network Setting. We take three timescales: 1, 3, and 5 frames around the target frame in MTDE. The size of the high-level dimension D and T is 32 and 64. To get enough receptive field, we adapt nine stacked CJRE modules with the lateral connection.

Training. All training is conducted on the pytorch platform with one 2080Ti GPU. We use Adam [41] optimizer with an initial learning rate of 0.0005. We use a weight decay of 0.96 and set the learning rate as 0.0001 at the epoch 4. The batch size is set to 16.

B. Comparison with baselines

Here we show the prediction performance for both short-term and long-term motion prediction on Human3.6M (H3.6M), CMU-Mocap and 3DPW. We quantitatively evaluate various methods by the mean per joint position error (MPJPE)

between the generated motions and ground truths. To be consistent with the literature [19], [22], we report our results for short-term ($< 500ms$) and long-term ($> 500ms$) predictions. For all datasets, we are given 10 frames (400 milliseconds) to predict the future 10 frames (400 milliseconds) for short-term prediction and to predict the future 25 frames (1 second) for long-term prediction.

1) Results on H3.6M: TABLE I provides the short-term prediction of 8 sub-sequences per action on H3.6M for the 15 activities and the average results. Note that our method outperforms all the baselines on average and almost all motions, which indicates our proposed framework's effectiveness. These possible reasons are twofold: (1) Our model extracts better motion features with our proposed joint relation modeling. These features contain both global coordination of all joints and local interactions between joint pairs and thus could offer more reliable guidance for effective prediction. (2) Our model extracts enriched motion dynamics in MTDE, which provide more motion cues for later motion prediction. Thus, our model generally outperforms the listed baselines in almost all actions. Specifically, for those motions that need the upper body and lower body to cooperate, e.g., "Walking dog", "Phoning" and "Sitting down", our method outperforms most, which reflects the efficacy of our proposed global coordination of all joints.

TABLE IV
LONG-TERM PREDICTION OF 256 SUB-SEQUENCES PER ACTION ON H3.6M. WHERE "MS" DENOTES "MILLISECONDS".

motion	Walking				Eating				Smoking				Discussion			
time(ms)	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000
ResSup [14]	71.6	72.5	76.0	79.1	74.9	85.9	93.8	98.0	78.1	88.6	96.6	102.1	109.5	122.0	128.6	131.8
ConvS2S [18]	72.2	77.2	80.9	82.3	61.3	72.8	81.8	87.1	60.0	69.6	77.2	81.7	98.1	112.9	123.0	129.3
LTD [22]	51.8	56.2	58.9	60.9	50.0	61.1	69.6	74.1	51.3	60.8	68.7	73.6	87.6	103.2	113.1	118.6
HRI [27]	**47.4**	**52.1**	**55.5**	**58.1**	50.0	61.4	70.6	75.7	**47.6**	**56.6**	**64.4**	**69.5**	86.6	102.2	113.2	119.8
Ours	47.9	52.5	56.2	59.6	**47.6**	**59.2**	**68.1**	**73.2**	48.4	57.7	64.9	69.6	**85.3**	**100.5**	**110.0**	**115.3**
motion	Direction				Greeting				Phoning				Posing			
time(ms)	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
ResSup [14]	101.1	114.5	124.5	129.1	126.1	138.8	150.3	153.9	94.0	107.7	119.1	126.4	140.3	159.8	173.2	183.2
ConvS2S [18]	86.6	99.8	109.9	115.8	116.9	130.7	142.7	147.3	77.1	92.1	105.5	114.0	122.5	148.8	171.8	187.4
LTD [22]	76.1	91.0	102.8	108.8	104.3	120.9	134.6	140.2	68.7	84.0	97.2	105.1	109.9	136.8	158.3	171.7
HRI [27]	73.9	88.2	100.1	106.5	101.9	118.4	132.7	138.8	67.4	82.9	96.5	**105.0**	107.6	136.8	161.4	178.2
Ours	**72.8**	**87.3**	**98.4**	**104.6**	**99.4**	**115.8**	**128.7**	**134.6**	**66.2**	**82.2**	**96.2**	105.1	**105.3**	**113.4**	**115.9**	**170.4**
motion	Purchasing				Sitting				Sitting down				Taking photo			
time(ms)	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000
ResSup [14]	112.1	137.2	148.0	154.0	113.7	130.5	144.4	152.6	138.8	159.0	176.1	187.4	110.6	128.9	143.7	153.9
ConvS2S [18]	111.3	129.1	143.1	151.5	82.4	98.8	112.4	120.7	106.5	125.1	139.8	150.3	84.4	102.4	117.7	128.1
LTD [22]	99.4	114.9	127.9	135.9	78.5	95.7	110.0	118.8	99.5	118.5	133.6	144.1	76.8	95.3	110.3	120.2
HRI [27]	**95.6**	**110.9**	125.0	134.2	76.4	93.1	107.0	115.9	97.0	116.1	132.1	143.6	**72.1**	90.4	105.5	115.9
Ours	96.2	111.5	**124.3**	**131.6**	**75.1**	**92.0**	**105.6**	**113.8**	**94.9**	**114.6**	**131.1**	**142.5**	72.3	**90.3**	**105.0**	**113.8**
motion	Waiting				Walking dog				Walking Together				Average			
time(ms)	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000
ResSup [14]	105.4	117.3	128.1	135.4	128.7	141.1	155.3	164.5	80.2	87.3	92.8	98.2	106.3	119.4	130.0	136.6
ConvS2S [18]	87.3	100.3	110.7	117.7	122.4	133.8	151.1	162.4	72.0	77.7	82.9	87.4	90.7	104.7	116.7	124.2
LTD [22]	75.1	88.7	99.5	106.9	105.8	118.7	132.8	142.2	58.0	63.6	67.0	69.6	79.5	94.0	105.6	112.7
HRI [27]	74.5	89.0	100.3	108.2	108.2	120.6	135.9	146.9	**52.7**	**57.8**	62.0	64.9	77.3	91.8	104.1	112.1
Ours	**69.7**	**83.4**	**95.0**	**102.4**	**102.9**	**115.2**	**131.6**	**141.6**	53.1	58.1	**61.3**	**63.8**	**75.8**	**90.3**	**102.2**	**109.5**

TABLE V
SHORT-TERM AND LONG-TERM PREDICTION ON CMU-MOCAP.

motion	basketball					baskeball Signal					Directing Traffic				
time (ms)	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
LTD [22]	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
TrajCNN [19]	11.1	19.7	43.9	56.8	114.1	**1.8**	3.5	**9.1**	13.0	**49.6**	**5.5**	10.9	23.7	**31.3**	**105.9**
Ours	**11.1**	**19.5**	**42.8**	**55.7**	113.1	1.9	**3.5**	9.3	**13.0**	57.5	5.8	11.7	25.6	33.4	139.0
motion	Jumping					Running					Soccer				
time (ms)	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
LTD [22]	16.9	34.4	76.3	96.8	164.6	25.5	36.7	39.3	39.9	58.2	11.3	21.5	44.2	55.8	117.5
TrajCNN [19]	12.2	28.8	**72.1**	94.6	166.0	17.1	24.4	28.4	32.8	49.2	**8.1**	**17.6**	40.9	51.3	126.5
Ours	**11.4**	**28.0**	72.7	**94.1**	**155.3**	**16.4**	**20.1**	**22.9**	**27.6**	**41.9**	8.6	18.3	**39.1**	**48.4**	**103.6**
motion	Walking					Wash Window					Average				
time (ms)	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
LTD [22]	7.7	11.8	19.4	23.1	40.2	5.9	11.9	30.3	40.0	79.3	11.5	20.4	37.8	46.8	96.5
TrajCNN [19]	6.5	10.3	19.4	23.7	41.6	**4.5**	**9.7**	29.9	41.5	89.9	8.3	15.6	33.4	43.1	92.8
Ours	**5.9**	**9.0**	**17.4**	**21.1**	**38.8**	4.6	10.0	**28.6**	**39.0**	**73.1**	**8.2**	**15.1**	**32.3**	**41.5**	**90.3**

These motions all need the whole body to participate in, and thus global coordination could offer more reliable guidance. Besides, the result on 320ms and 400ms increase most, which shows our method is good at capturing temporal continuity for long-term prediction compared with other methods.

In TABLE II, we compare our model with other baselines for long-term prediction of 8 sub-sequences per action on H3.6M. With the uncertainly of motion increasing, our method still obtains competitive performances on almost all motions. In TABLE III and TABLE IV, we also show the short-term and long-term prediction of 256 sub-sequences per action on H3.6M for the 15 activities and the average results. We can find that with the number of test data increasing, our method still outperforms all baselines in almost all motions. It proves that our method has great generalization in different situations, whether the samples are big or not. The above observations all demonstrate the advantages of our proposed enriched dynamics and comprehensive joint relation modeling.

2) Results on CMU Mocap and 3DPW: TABLE V reports the MPJPE for short-term and long-term prediction on CMU-Mocap and TABLE VI reports the results on 3DPW. Essentially, the conclusions remain unchanged: our method

TABLE VI
SHORT AND LONG-TERM PREDICTION ON 3DPW.

time (ms)	200	400	600	800	1000
LTD [22]	36.0	69.0	91.0	107.6	118.6
Ours	**34.7**	**66.7**	**85.6**	**98.0**	**108.4**

consistently outperforms the baselines for both short-term and long-term prediction.

C. Ablation study

In this section, we conduct several ablation experiments to testify the effectiveness of different components in our proposed framework. The models are trained on the H3.6M training set and evaluated on the test dataset. The comparison results are shown in TABLE VII, VIII, IX, X, XI.

1) Multi-timescale dynamics: The multi-timescale dynamics could offer enriched motion cues than position or velocity information provided by the raw input motion sequences. And we design the MTDE module to encode this information in our work. As is shown in Table VII, we design two extra experiments. "TS" means that we only adopt two-stream input
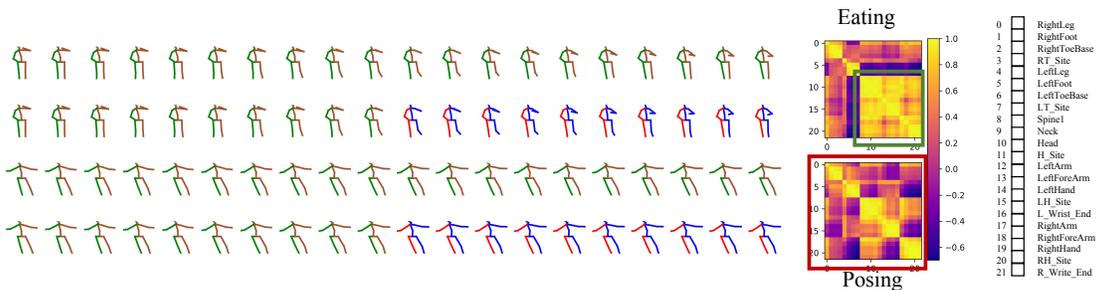
Fig. 7. The middle panel is the visualizations of one feature map of C_{emb}, of which the size is $[22 \times 22]$. The right panel is the corresponding prediction results. The left panel is the annotation of the feature map. From top to bottom, we show the ground truth and our approach of two different motions.

and "TS+FE" means that we intorduce bare 1*3 convolution without extra designs. Our final scheme is "TS+FE+MT" which adds the multiple temporal convolutions compared with "TS+FE". We can see "TS+FE" is superior to "TS", which indicates that more motion cues are useful. Besides, after using "TS+FE+MT", we could get a fairly good promotion by adopting the MTDE module in all time horizons. Significantly, the results of 320ms and 400ms increase a lot, showing that enriched dynamics could offer more meaningful guidance for middle or long-term prediction.

TABLE VII
RESULTS OF ABLATION EXPERIMENTS ON MTDE

MTDE	80	160	320	400
TS	10.2	22.9	49.5	60.6
TS+FE	9.8	22.6	48.0	58.4
TS+FE+MT	**9.6**	**22.0**	**46.3**	**57.0**

2) Global coordination of all joints: The global coordination of all joints could reflect the entirety and coordination of the predicted motion and thus is one vital joint relation. In our work, we propose the GCE module to model this global coordination. We here illustrate the effectiveness of several special designs in this module. In the Feature Normalization (FNU), relative joint motion representation is a vital design used to remove the effect of global motion trends because the global coordination essentially reflects the mutual constraints of all joints. Here, we denote "RJ" to represent the usage of relative joint motion representations. In the Multi-head Self-attention Unit (MSU), we generate multiple relation graphs to extract richer motion coordination. Here we use "MR" to represent the usage of multiple relation graphs. Besides, In the MSU, we also choose cosine similarity to calculate the joint relations in our paper and illustrate the advantage of this choice. To prove the effectiveness, we design an experiment with the softmax function as a comparison. Here "Sim_c" and "Sim_s" represent the usage of cosine similarity or softmax.

(1) By adopting the relative joint representation, the final performance outperforms 0.1, 0.3, 1.1, 1.4 for 80ms, 160ms, 320ms, 400ms, respectively. This proves that the relative joint representation without the global motion trends is more suitable for modeling joints' global coordination.

(2) We can find the results are better with the usage of multiple relation graphs. It demonstrates that combining

TABLE VIII
RESULTS OF ABLATION EXPERIMENTS ON GCE

RJP	MR	Sim_c	Sim_s	80	160	320	400
✓	✓		✓	10.2	23.4	49.5	60.6
	✓	✓		9.7	22.3	47.4	58.4
✓		✓		9.8	22.1	46.7	57.7
✓	✓	✓		**9.6**	**22.0**	**46.3**	**57.0**

different relations is beneficial to get richer and more diverse coordination.

(3) The cosine similarity is better compared with the softmax function used in computing self-attention. It arises from two aspects. First, it avoids violent differences in the softmax function because cosine similarity limits the value in $[-1, 1]$. Second, it has the angle information to represent both orientation and intensity of correlation, while softmax only represents the intensity of correlation.

TABLE IX
RESULTS OF ABLATION EXPERIMENTS ON GCE, LIE AND AFFM

GCE	LIE	$AFFM$	80	160	320	400
✓		✓	9.7	22.6	48.3	58.9
	✓	✓	10.1	23.1	49.2	60.0
✓	✓		9.6	22.4	46.8	57.4
✓	✓	✓	**9.6**	**22.0**	**46.3**	**57.0**

3) Importance of different joint relations: The global co-ordination of all joints and local interactions between joint pairs are complementary and crucial joint relations. In TABLE IX, we could easily find that these two joint relations can promote each other, and adopting both of them can achieve better improvement. Notably, the method with a single GCE outperforms the one with a single LIE. This observation demonstrates that global coordination could extract more cues and offer more effective guidance than local interactions in prediction.

Furthermore, we also conduct the experiments in TABLE X to verify the performance of different combination method of GCE and LIE. We denote the "Parallel" as putting the GCE and LIE in the parallel path and "Serial" means that local interaction is conditioned on global coordination. We can see that if we set the LIE behind the GCE, we won't get equivalent performance compared with putting the GCE and LIE in parallel. The main reason is the output features of GCE

(a) Sitting 0°

(b) Sitting 30°
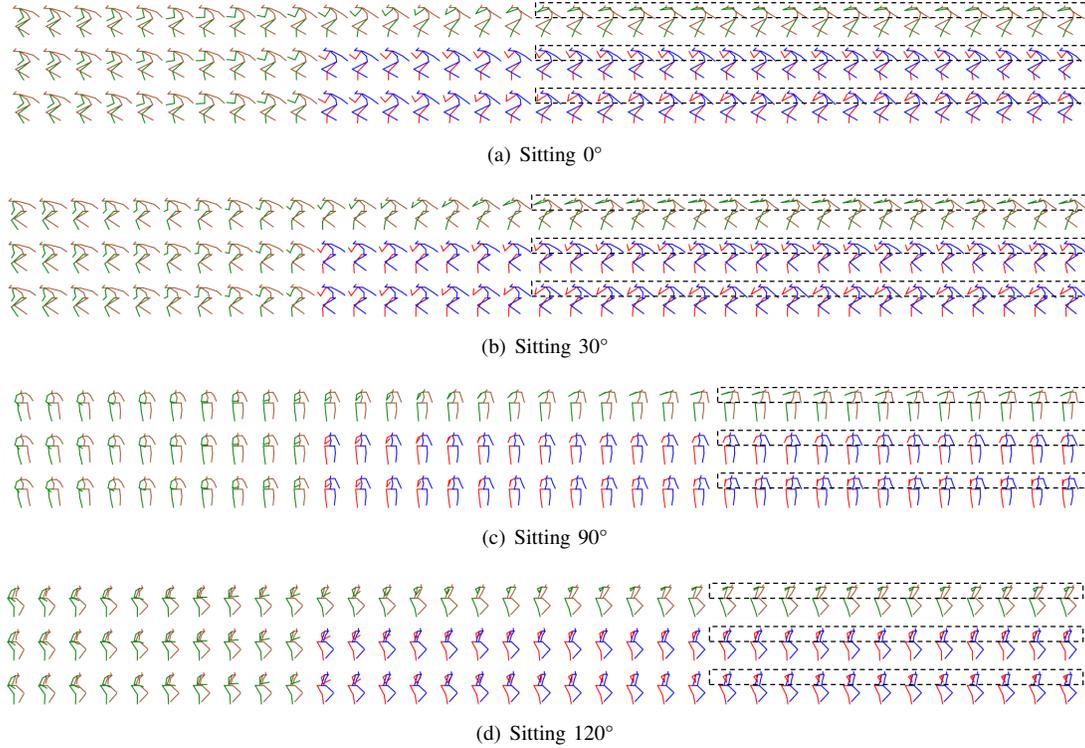
(c) Sitting 90°

(d) Sitting 120°

Fig. 8. Qualitative comparison of multi-view long-term predictions on H3.6M. From top to bottom, we show the ground truth, the results of LTD [22], and our approach. The highlight illustrates the difference between three sequences.
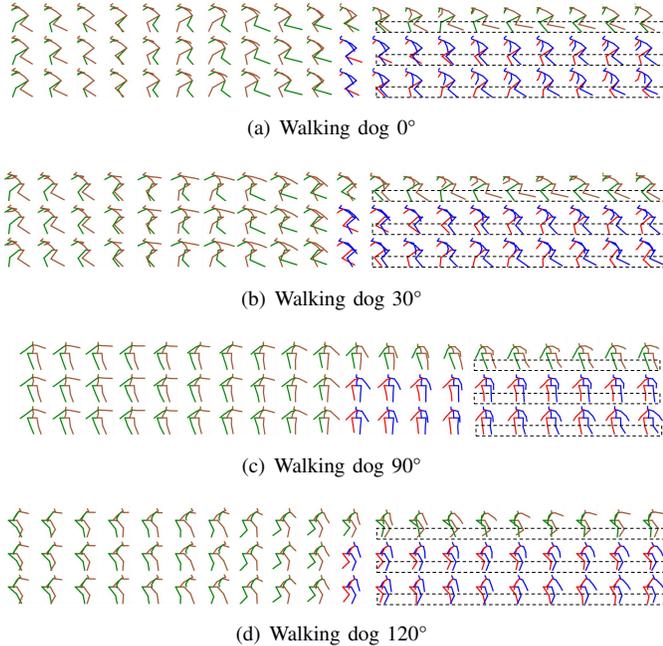

(a) Walking dog 0°

(b) Walking dog 30°

(c) Walking dog 90°

(d) Walking dog 120°

Fig. 9. Qualitative comparison of multi-view short-term predictions on H3.6M. From top to bottom, we show the ground truth, the results of LTD [22], and our approach. The highlight illustrates the difference between three sequences. The results evidence that our approach generates high-quality predictions.

are not suitable to act as the input of LIE. More concretely, the output features of GCE represent the joint features relative to the CA. Thus, the adjacent pixels in the new relative feature maps lost the adjacent relations existing in the raw skeletal structure, which is the prerequisite of the LIE module.

TABLE X
RESULTS OF ON THE COMBINATION METHOD OF GCE AND LIE

combination method	80	160	320	400
Serial	10.5	23.5	49.6	60.4
Parallel(Ours)	**9.6**	**22.0**	**46.3**	**57.0**

Besides, we consider that different motions may have different preferences for those above two joint relations and thus design the AFFM module. As is shown in TABLE IX, AAFM improves the results by 0.4 on average. It reflects that fusing the motion features achieved from different joint relations enhances the whole performance. Notably, the limited improvement of AFFM reflects that the joint coordination modeling is the key design to improve the final prediction performance.

To demonstrate how LIE and GCE are weighted in AFFM module, we here offer the relative weight of different features in different motions. As is shown in TableXI, $w_{distant}$, $w_{adjacent}$, w_{ca} represent the relative weight of $F_{distant}$, $F_{adjacent}$, F_{ca} respectively. From the results, we can find the ratio of global motion coordination w_{ca} varies a little, which means that the global motion coordination is of almost equal importance in all kinds of motions. Besides, we can also find $w_{distant}$ and $w_{adjacent}$ are usually contrary. For those motions with little movement like "smoking", "sitting down" and "posing", the importance of $F_{distant}$ is higher. For those motions with more movement like "walking", "walking

(a) Running 0°

(b) Running 30°

(c) Running 90°

(d) Running 120°

Fig. 10. Qualitative comparison of multi-view short-term and long-term predictions on CMU-Mocap. From top to bottom, we show the ground truth, the results of LTD [22], and our approach. The highlight illustrates the difference between three sequences.

TABLE XI
RELATIVE WEIGHT OF DIFFERENT FEATURES OF DIFFERENT MOTIONS IN AFFM

$Motions$	$w_{distant}$	$w_{adjacent}$	w_{ca}
walking	0.26	0.40	0.34
eating	0.32	0.36	0.33
smoking	0.35	0.33	0.31
discussion	0.31	0.36	0.33
directions	0.31	0.36	0.33
greeting	0.30	0.36	0.34
phoning	0.31	0.37	0.32
posing	0.35	0.33	0.33
purchases	0.31	0.36	0.33
sitting	0.34	0.34	0.32
sittingdown	0.35	0.33	0.32
takingphoto	0.33	0.34	0.33
waiting	0.29	0.37	0.34
walkingdog	0.28	0.38	0.34
walkingtogether	0.27	0.39	0.34

dog" and "walking together", the importance of $F_{adjacent}$ is higher. It demonstrates that dynamic motions may focus on the movement of local bodies while static motions pay more attention to the overall structure of the human body.

D. Qualitative analysis

1) Visulization of different motions on H3.6M and CMU Mocap: As is shown in Fig. 8, Fig. 9, and Fig. 10, we list more qualitative visualizations of different datasets. We can find that our method outperforms LTD[22] in both short-term and long-term prediction. In detail, compared with LTD[22], it is easier for our model to capture motion tendency and make more precise predictions. For example, in Fig. 9, we can see that the movement of the left leg are well learned from four perspectives in our prediction for the motion "Walking dog" while not in LTD[22]. And in the motion "Sitting", our model is more precise in predicting upper limb movements from multiple viewpoints. It demonstrates that our methods can better capture the mutual constraints of different body parts, and thus the resulting predicted motion appears more natural and realistic. Besides, for those motions like "Running", our results usually outperform other methods, which shows the enriched dynamics can offer more useful motion cues for motion prediction.

2) Visualization of global coordination: In Fig. 7, we show the one feature map of C_{emb}, which reflects the global coordination of all joints. The value increases with the color become brighter. For the motion "Eating", there exist high correlations on the upper body because the joints on the upper body need to coordinate to finish the motion. While for motion "Posing" there exist fewer correlations between different body parts because this motion is more static than "Eating". Thus, the coordination only occurs in the local body parts. This observation demonstrates that our proposed GCE extract reliable global coordination for effective prediction.

V. CONCLUSION

In this paper, we focus on more realistic human motion prediction with attention to motion coordination. To this end,

we propose the CJRE to explore richer joint relation modeling, mainly including GCE and LIE. The former presents the global coordination of all joints and the latter encodes local interactions between joint pairs. Besides, we also extract enriched motion dynamics of raw skeleton data through the MTDE to exploit more motion cues for effective prediction. Experimental results on three benchmark datasets suggest that our proposed framework is able to improve the coordination of predicted motions with lower errors to generate more realistic actions.

REFERENCES

[1] Y. Ji, Y. Yang, F. Shen, H. Shen, and X. Li, "A survey of human action analysis in HRI applications," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 30, no. 7, pp. 2114–2128, 2020.

[2] L. Chen, J. Lu, Z. Song, and J. Zhou, "Recurrent semantic preserving generation for action prediction," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 31, no. 1, pp. 231–245, 2021.

[3] H. S. Koppula and A. Saxena, "Anticipating human activities for reactive robotic response," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 14–29, 2016.

[4] J. Bütepage, H. Kjellström, and D. Kragic, "Anticipating many futures: Online human motion prediction and generation for human-robot interaction," in *IEEE Int. Conf. Robot. Automat*, 2018, pp. 1–9.

[5] B. Paden, M. Čáp, S. Z. Yong, D. Yershov, and E. Frazzoli, "A survey of motion planning and control techniques for self-driving urban vehicles," *IEEE Trans. Intell. Veh.*, vol. 1, no. 1, pp. 33–55, 2016.

[6] I. Hasan, F. Setti, T. Tsesmelis, V. Belagiannis, S. Amin, A. D. Bue, M. Cristani, and F. Galasso, "Forecasting people trajectories and head poses by jointly reasoningon tracklets and vislets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1267–1278, 2021.

[7] A. Bhattacharyya, M. Fritz, and B. Schiele, "Long-term on-board prediction of people in traffic scenes under uncertainty," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 4194–4202.

[8] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 283–298, 2008.

[9] A. M. Lehrmann, P. V. Gehler, and S. Nowozin, "Efficient nonlinear markov models for human motion," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 1314–1321.

[10] A. Lehrmann, P. V. Gehler, and S. Nowozin, "Efficient nonlinear markov models for human motion," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 1314–1321.

[11] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 283–298, 2008.

[12] Q. Men, E. S. L. Ho, H. P. H. Shum, and H. Leung, "A quadruple diffusion convolutional recurrent network for human motion prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[13] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *Int. Conf. Comput. Vis.*, 2015, pp. 4346–4354.

[14] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 4674–4683.

[15] Y. Liang, X. Yu, X. Liang, and J. Moura, "Adversarial geometry-aware human motion prediction," in *Eur. Conf. Comput. Vis.*, 2018, pp. 823–842.

[16] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-rnn: Deep learning on spatio-temporal graphs," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 5308–5317.

[17] E. Aksan, M. Kaufmann, and O. Hilliges, "Structured prediction helps 3d human motion modelling," in *Int. Conf. Comput. Vis.*, 2019, pp. 7143–7152.

[18] C. Li, Z. Zhang, W. S. Lee, and G. H. Lee, "Convolutional sequence to sequence model for human dynamics," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 5226–5234.

[19] X. Liu, J. Yin, J. Liu, P. Ding, J. Liu, and H. Liub, "Trajectorycnn: a new spatio-temporal feature learning network for human motion prediction," *IEEE Trans. Circuit Syst. Video Technol.*, pp. 1–1, 2020.

[20] M. Li, S. Chen, Y. Zhao, Y. Zhang, Y. Wang, and Q. Tian, "Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 211–220.

[21] Q. Cui, H. Sun, and F. Yang, "Learning dynamic relationships for 3d human motion prediction," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 6519–6527.

[22] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning trajectory dependencies for human motion prediction," in *Int. Conf. Comput. Vis.*, 2019, pp. 9488–9496.

[23] W. Mao, M. Liu, and M. Salzmann, "History repeats itself: Human motion prediction via motion attention," in *Eur. Conf. Comput. Vis.*, 2020, pp. 474–489.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 770–778.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, vol. 9351, 2015, pp. 234–241.

[26] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," in *IJCAI*, 2018, pp. 786–792.

[27] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Adv. Neural Inform. Process. Syst.*, 2014, pp. 568–576.

[28] X. Guo and J. Choi, "Human motion prediction via learning local structure representations and temporal dependencies," in *AAAI Conf. Artif. Intell.*, 2019, pp. 2580–2587.

[29] H. Chiu, E. Adeli, B. Wang, D. Huang, and J. Niebles, "Action-agnostic human pose forecasting," in *IEEE Winter Conf. Appl. Comput. Vis.*, 2019, pp. 1423–1432.

[30] D. Pavllo, C. Feichtenhofer, M. Auli, and D.Grangier, "Modeling human motion with quaternion-based neural networks," *Int. J. Comput. Vis.*, vol. 128, no. 4, pp. 855–872, 2020.

[31] A. H. Ruiz, J. Gall, and F. Moreno, "Human motion prediction via spatio-temporal inpainting," in *Int. Conf. Comput. Vis.*, 2019, pp. 7133–7142.

[32] Y. Cai, L. Huang, Y. Wang, T. Cham, J. Cai, J. Yuan, J. Liu, X. Yang, Y. Zhu, XShen *et al.*, "Learning progressive joint propagation for human motion prediction," in *Eur. Conf. Comput. Vis.*, 2020, pp. 226–242.

[33] J. Butepage, M. Black, D. Kragic, and H. Kjellstrom, "Deep representation learning for human motion prediction and classification," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 6158–6166.

[34] B. Wang, E. Adeli, H. Chiu, D. Huang, and J. Niebles, "Imitation learning for human pose prediction," in *Int. Conf. Comput. Vis.*, 2019, pp. 7124–7133.

[35] Y. Cai, L. Ge, J. Liu, J. Cai, T. Cham, J. Yuan, and N. Thalmann, "Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks," in *Int. Conf. Comput. Vis.*, 2019, pp. 2272–2281.

[36] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI Conf. Artif. Intell.*, 2018, pp. 7444–7452.

[37] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 7794–7803.

[38] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, 2014.

[39] T. Marcard, R. Henschel, M. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3d human pose in the wild using imus and a moving camera," in *Eur. Conf. Comput. Vis.*, 2018, pp. 614–631.

[40] CMU.(2003)Graphics lab motion capture database. [Online]. Available: http://mocap.cs.cmu.edu/

[41] A. Paszke, S.Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVitoand Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

Pengxiang Ding received the B.S. degree from the Beijing University of Post and Telecommunications, Beijing, China, in 2019, where he is currently pursuing the M.S. degree with the School of Artificial Intelligence. His research interests include motion prediction and action recognition in computer vision.

Jianqin Yin (Member, IEEE) received the Ph.D. degree from Shandong University, Jinan, China, in 2013. She currently is a Professor with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China. Her research interests include service robot, pattern recognition, machine learning, and image processing.