Collaborative Multi-dynamic Pattern Modeling for Human Motion Prediction

Jin Tang†*, Jin Zhang†, Rui Ding, Baoxuan Gu, Jianqin Yin*, *Member, IEEE*

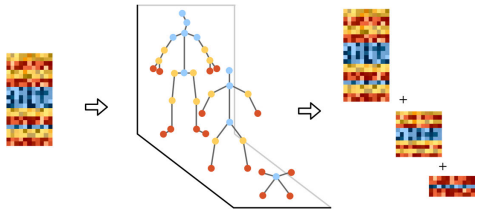
***Abstract*—The dynamic information of the joints, such as the movement amplitude, is critical for forecasting precise human joint trajectories. Existing methods adopt global modeling in which all joints are treated as a whole to extract features for movement coordination. Though global modeling can exploit hidden relationships between joints, it also inevitably introduces undesired trajectory dependencies, which weakens the dynamic information effects and simplifies the constraints and kinetics model of joints. Therefore, we propose a dynamic pattern-based collaborative modeling framework (DPnet) that contains a keyframe enhanced module (KEM) and multi-channel feature extractor blocks (MFE-block). The KEM tackles the discontinuity between the last frame of observation and the first predicted one by duplicating the decisive frame. The MFE-block utilizes a multi-channel graph structure to enrich the dynamic information effects and recessive constraints of joints. To distinguish the dynamic information of each joint, we calculate the movement amplitude of the joints and propose three dynamic patterns, including active, inactive, and static patterns. We also propose a dynamic pattern-guided feature extractor (DP-FE) to alleviate the trajectory dependencies between joints with different dynamic patterns. We evaluate our approach on three standard benchmark datasets, including H3.6M [8], CMU-Mocap [44], and 3DPW [45]. Our approach achieves impressive results in both short-term and long-term predictions, confirming its effectiveness and efficiency.**

***Index Terms*—collaborative modeling, multi-graph structure, multi-dynamic pattern, human motion prediction.**

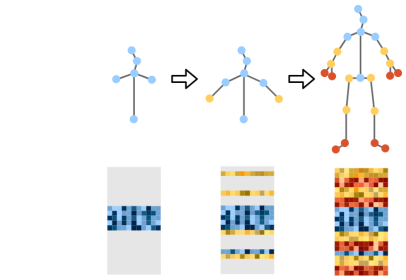
I. Introduction

Human motion prediction is a classic task in the field of computer vision. The task is to predict future human motion sequences based on the observations of past sequences, which can be applied to service robots [1], [2], virtual reality [3], [4], and other fields. The biggest challenge is to reasonably refine the pattern of human dynamic features and generate the following rational and natural human poses as possible.

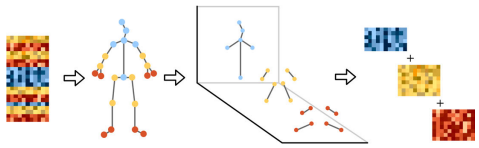
For learning the spatial correlation of joints, many encoder and decoder strategies have been proposed in previous works. Those strategies focus on adjusting the order and changing the scales of joints. Liu et al. [5] represented a new joint order which concentrated upper and lower limb joints to learn the relationship of adjacent joints. Li et al. [6] processed human joint trajectories into three scales: joint level, body part level, and limb level, as shown in Fig. 1(a). Although each scale modeling adopted different GCN [46] streams to model the corresponding spatial correlation, it still involved all the joints as global modeling. Cai et al. [7] proposed to use RNN to recursively fill the human body pose starting from body center to limb extremity (Fig. 1(b)). In this way, the network can effectively deal with joint features within the corresponding scale, meaning static torso joint features are first generated and then active extremity joint features. However, the mentioned works focus more on the body's physical constraints and treat all joints as a whole to model the movement without explicitly exploiting the movement coordination and relations at the same movement amplitude pattern joints, like distal joints. For example, the movement amplitude of the ankles is much larger than the head in the action of "Running" as



(a) Multiscale encoder strategy.



(b) Progressive transformer decoder.



(c) Our pattern-guided encoder.

Fig. 1. Encoder-decoder strategies in human motion prediction. (a) A multiscale encoder is designed to model the global spatial correlation. (b) A progressive transformer decoder is designed based on global modeling. (c) Our collaborative modeling of joints with similar dynamic information alleviates the involvement and refines recessive correlation.

Jin Tang, Jin Zhang, Baoxuan Gu, 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: tangjin@bupt.edu.cn, jinzhang@bupt.edu.cn, gbx@bupt.edu.cn, jqyin@bupt.edu.cn. Rui Ding is with information engineering college of capital normal university. Email: 5758@cnu.edu.cn.

*Corresponding authors.

†These authors contributed equally to this work.

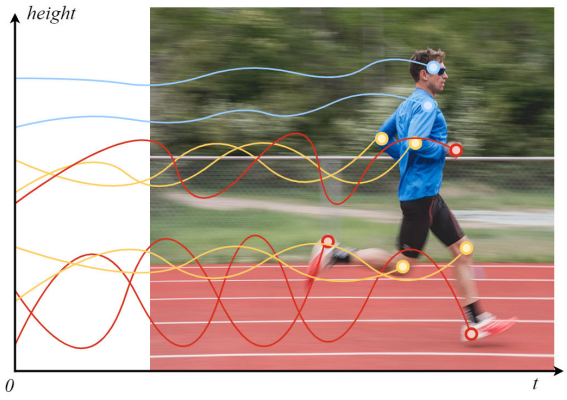
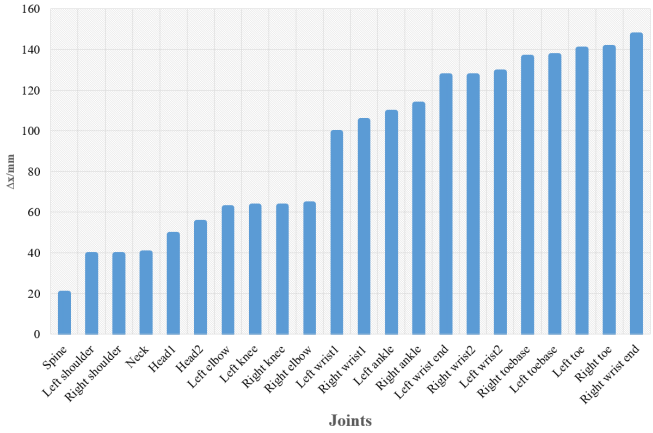
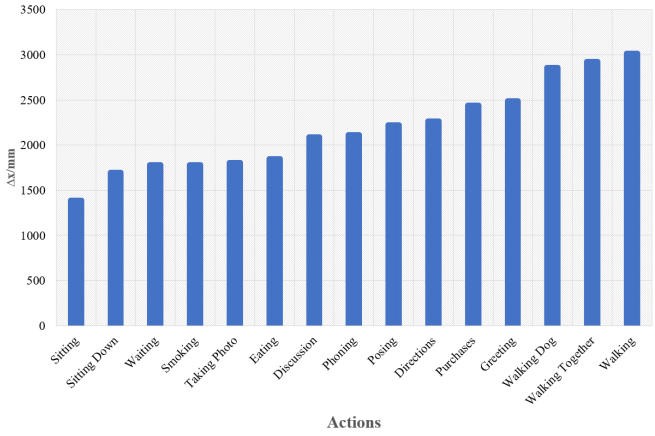


Fig. 2. Illustration of our motivation. In "Running" action, the joints from torso, limbs, and limb extremities present different movement amplitudes, leading to their different contribution to human motion prediction.



(a) Mean movement amplitude of joints.



(b) Mean movement amplitude of actions.

Fig. 3. Movement amplitude statistics on H3.6M dataset. Calculates Δx to indicate motion trajectory amplitude. The active joints mainly distribute at wrists and toes compared with static joints that distribute at torso. The active actions mainly consist of big moves like "Walking" compared with small moves like "Sitting." This suggests that active joints contribute more motion patterns in big moves.

shown in Fig. 2. Though the global modeling can exploit hidden relationships between ankle and head, it also inevitably introduces undesired trajectory constraints from the head to the ankle, which weakens the dynamic information effects and simplifies kinetics model of joints. On the contrary, collaborative multi-dynamic pattern modeling can explicitly exploit the relations at the same dynamic pattern joints, like the relation between the ankle and wrist.

In a joint movement, the movement amplitude is often affected by the kinematic connection. Since the movement amplitude of the limb extremity joints is usually larger than that of the torso joints, the corresponding trajectories of the active joints often contain more dynamic information and have a greater impact on the prediction. As mentioned, global modeling may introduce mutual involvement constraints and relations, ignoring the unique dynamic information of each joint. To quantify the dynamic information of various joints in the human body, we calculate the movement amplitude to differentiate the dynamic patterns of the joints in the H3.6M dataset [8]. That is, calculate the displacement of the joints as Δx between adjacent frames. As shown in Fig. 3, the ordinate indicates the mean trajectory movement amplitude of the corresponding joint or action category. In Fig. 3(a), the large movement amplitude joints like "Toe" of the human body are distributed at the extremities as expected, while the small movement amplitude joints like "Neck" are all distributed near the torso. In Fig. 3(b), sequences from the big-move actions like "Walking Dog" has obviously larger movement amplitude than some small-move actions like "Smoking." This motivates our multi-dynamic patterns to distinguish the diverse dynamic information of joints. As shown in Fig. 4, the joints are distinguished into different dynamic patterns, namely active, inactive and static patterns. Meanwhile, three graphs are constructed based on the joint dynamic patterns to exploit the relations or constraints at the same dynamic pattern. Finally, our collaborative multi-dynamic pattern graph convolutional network is constructed, which models collaborative movements within the pattern corresponding joints of a motion sequence effectively.

For further utilizing the key frame information, many works focused on enhancing the key frame temporal features. Martinez et al. [9] proposed to use the residual connection at the end of the RNN to introduce the positional information of the last observed frame, which effectively solved the pose discontinuity problem between the observed sequence and the predicted sequence. Lebailly et al. [10] used 1D convolutional layers with different kernel sizes to extract temporal features of cropped sub-sequences. Among them, the convolutional layer with a smaller kernel is dedicated to extracting features of recent frames, placing more emphasis on the recent frames than the older ones. Consequently, the information from the latest observed frames helps the model deduce the future poses.

Generally, the latest observed frames often contain a more explicit inertial pattern. For example, when we predict a motion from 10 minutes observed sequence, the last few observed frames apparently have a higher value than the beginning ones. Since the prediction comes from the inference of the last observed frame, we enhance the features of the last

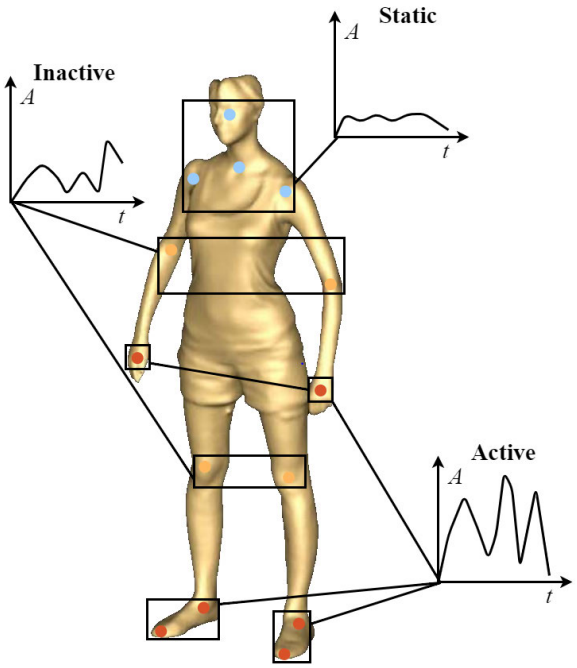


Fig. 4. Schematic of joint dynamic patterns. The joints are divided into red active pattern, orange inactive pattern, and blue static pattern according to motion amplitude as shown in the figure. Different movement amplitudes of joint trajectories in different dynamic patterns.

few observed frames, particularly the last observed frame, to keep the continuity between the last frame of observation and the first predicted one.

The main contributions are summarized as follows: 1) To distinguish the dynamic information of each joint, we propose three dynamic patterns based on the joints' movement amplitude, including active, inactive, and static patterns. 2) To promote finer feature extraction, we propose a dynamic pattern-based collaborative modeling framework that contains a keyframe enhanced module (KEM) and multi-channel feature extractor blocks (MFE-block). In MFE-block, we use the multi-graph structure to alleviate the trajectory dependencies between joints with different dynamic patterns.

II. Related Work

Human motion prediction. A human motion sequence can be regarded as serial frames of human poses. Thus many works [7], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21] implement RNN to model the temporal sequence effectively. Fragkiadaki et al. [11] early proposed a sequence to sequence RNN to synthesize human future poses temporally. Ghosh et al. [18] proposed a 3-layer LSTM network, which effectively exploited the temporal features and reduced the accumulation of correlated error. Al-aqel et al. [20] implemented an attention mechanism to RNN for focusing on keyframes. Compared with temporal feature refining by RNN, many CNN-based models [5], [22], [23], [24], [25], [26] show effectiveness in learning spatio-temporal coupling features by enlarging the convolution receptive field to a proper size. TrajectoryNet [5] constructed multiple trajectory blocks to extract the dynamic features of a motion sequence. Cui et al. [23] adopted temporal convolution to increase the receptive field of the network on the time dimension, and the adversarial training strategy effectively improved the long-term modeling ability. In recent works [6], [27], [28], [29], [30], [31], [32], [33], [34], the utilization of GCN [46] has become famous since the coordinates and joint connections can constitute the nodes and edges of a graph structure. Mao et al. [31] proposed to model joint-wise trajectories by graph convolutional layers to strengthen the spatial correlation. Li et al. [32] successfully calculated graph spectrum attention to learn the rich spectral representation of a motion sequence.

Multi-level graph modeling. In recent years, many works [34], [35], [36], [37], [38], [39], [40] from both human motion prediction and skeletal action recognition have proved the effectiveness of modeling human motion patterns in multi-level by cropping or down-sampling human pose sequences. Yan et al. [40] proposed ST-GCN for modeling spatial-temporal correlations adaptively by first constructing a joint-level graph and connecting frame-level nodes afterward. Dang et al. [34] proposed an hourglass model to down-sample human pose into multiple spatial scales, which helps extract features in multiple granularities. For enriching input modalities, Song et al. [37] split the human body into five parts according to kinematic connection. Each part was down-sampled and modeled respectively to compute spatial attention weights, which effectively provided an explanation for the classification results. While existing works fail to implement differential modeling by joint-wise trajectory characteristics, making it difficult for the network to perceive the pattern of the active joints, which holds more significance when synthesizing or recognizing an action.

Temporal feature augmentation. Many works [9], [10], [30], [41] have emphasized that recently observed frames are often more valuable for human motion prediction. Mao et al. [30] narrowed the output length of the predicting sequence and proposed to recurrently learn human dynamics by absorbing previous output sequences to inherit the newly represented pattern. Compared with enhancing the features of the latest frames of an input sequence, Tang et al. [41] achieved human velocity prediction and directly pointed out that the last frame is the most decisive one for deducing future poses. Therefore, properly enhancing the recently observed information can effectively improve the pattern consistency of the network input and output sequences.

III. Methodology

A. Overview

Human motion prediction task is to generate the future pose sequence under the observation sequence. Since there are ambiguities in human posture represented in angle space according to [31] and [10], our study mainly discusses the methodology in 3D coordinate space. In Table I, we first enumerate some necessary notations and variables in this paper to help readers better understand this work. Composed by frames of human pose, the observed sequence and the predicted sequence can be denoted by $\boldsymbol{O} = \{P_1, P_2, ..., P_T\}$ and $\widehat{\boldsymbol{S}} = \left\{\widehat{P_{T+1}}, \widehat{P_{T+2}}, ..., \widehat{P_{T+L}}\right\}$, where T and L represent the

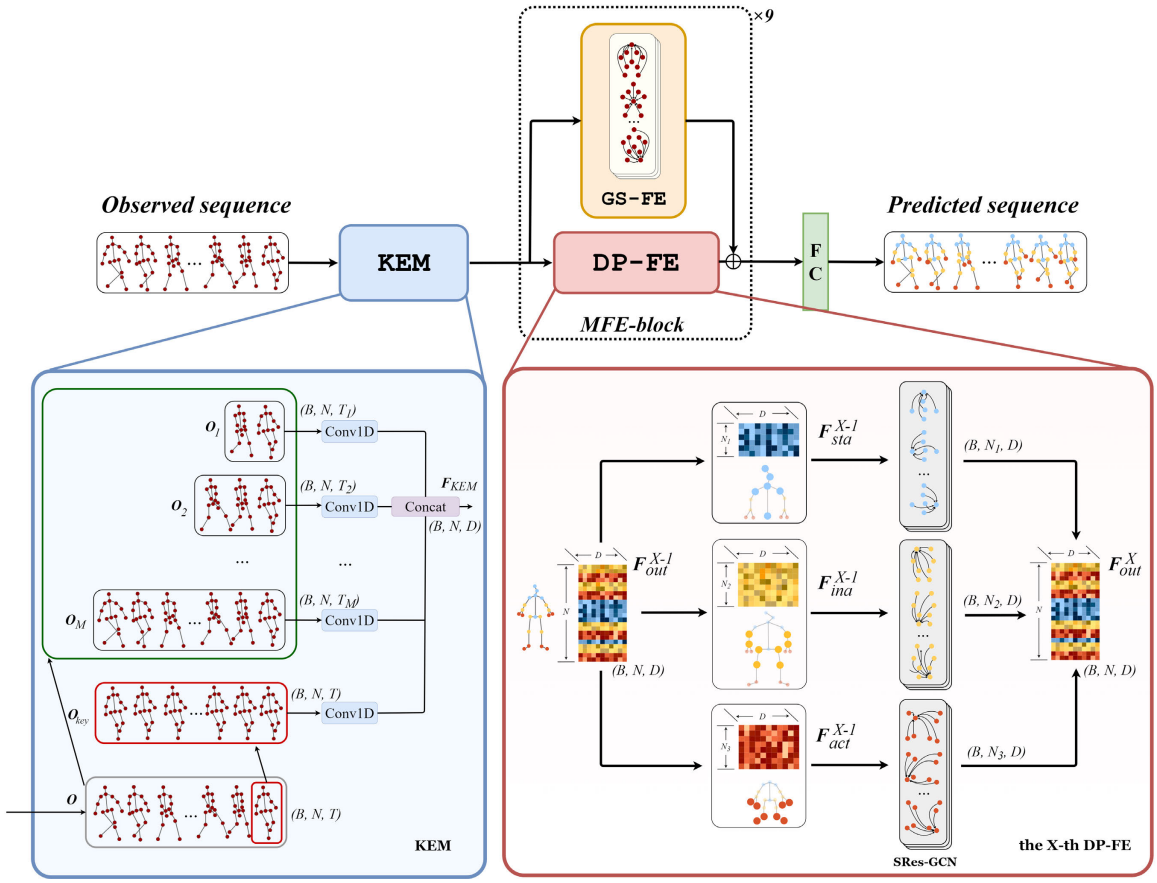


Fig. 5. Architecture. The framework is mainly constructed by KEM and MFE-blocks. The KEM enhances the temporal features by encoding the observation to different length sub-sequences. The DP-FE is adapted to congregate trajectories within the same dynamic pattern and modeled by symmetric residual graph convolution blocks. Moreover, a global branch is utilized over the DP-FE for introducing global spatial features to the next DP-FE and the output. "B" denotes the batch size.

number of observed frames and predicted frames respectively. The ground truth sequence can be correspondingly denoted as $\boldsymbol{S} = \{P_{T+1}, P_{T+2}, ...P_{T+L}\}$. The i-th frame of the observed frame P_i is composed of N joint coordinate tuples, which is $\boldsymbol{P}_i = \{J_{1,i}, J_{2,i}, ..., J_{N,i}\}$. Meanwhile, the k-th joint tuple can be denoted as $\boldsymbol{J}_{k,i} = \{x_{k,i}, y_{k,i}, z_{k,i}\}$ in 3D coordinate space. Consequently, the shapes of matrix $\boldsymbol{O}$ and $\widehat{\boldsymbol{S}}$ are $(T, N, 3)$ and $(L, N, 3)$. The trajectory of a joint along an axis in 3D coordinate space during T frames can be represented by a matrix of $(T, 1, 1)$.

Our dynamic pattern-based collaborative modeling framework includes two key components, as shown in Fig. 5, including the keyframe enhanced module (KEM) and multi-channel feature extractor blocks (MFE-block). KEM extracts the sequence's temporal features at the first stage, encoding hidden features by temporal dimension and enhancing the latest frame significance of the observation sequence. According to the mentioned joint-wise static, inactive and active patterns, we build three graph structures to exploit the joints' relations and constraints at the same dynamic pattern respectively. Based on the mentioned three graph structures, we propose a novel module called dynamic pattern-guided features extractor (DP-FE), which is a three-channel GCN block for extracting the hidden features at the same dynamic pattern. Finally, a fully connected layer gives the final prediction results. The network details are discussed below.

B. Keyframe Enhanced Module (KEM)

To generate natural-looking poses, especially keeping the continuity of the first predicted frame. We adopt a keyframe-enhanced module (KEM) to enhance the features of the keyframes inspired by [10]. Lebailly et al. [10] mentioned the latest frames of a human motion sequence often play the dominant role in the prediction task and also contains more inertia pattern in the motion. Following [10], we firstly clipper the input sequence into M sub-sequences, and the m-th sub-sequence noted as $\boldsymbol{O}_m = \{P_{T-T_m+1}, P_{T-T_m+2}, ..., P_T\}$ (Framed in green in Fig. 5), T_m represents the number of frames in sub-sequence $\boldsymbol{O}_m$. Therefore, these sub-sequences can be described as $[\boldsymbol{O}_1, \boldsymbol{O}_2, ..., \boldsymbol{O}_M]$ Each sub-sequence segment will be encoded by a 1D convolutional layer as Eq. (1).

$$\boldsymbol{F}_m = Cov1(\boldsymbol{O}_m) \tag{1}$$

According to V-CMU [41], the position information contained in the last frame of a motion sequence is the most decisive one. To enhance features of the last observed frames, we created a new sequence by duplicating the last observed frame T times, noted as $\boldsymbol{O}_{key}$ (Framed in red in Fig. 5). Then, features of this

TABLE I
NOTATIONS AND DEFINITIONS.

Number	Notations	Explanations
1	$\boldsymbol{O}$	The observed motion sequence, the shape is $T \times N \times 3$.
2	$\boldsymbol{P}_i$	The i-th frame of the observed position sequence.
3	T	The number of observed frames.
4	$\widehat{\boldsymbol{S}}$	The predicted motion sequence.
5	$\widehat{\boldsymbol{P}_{T+l}}$	The predicted l-th frame of the position sequence.
6	L	The number of predicted frames.
7	$\boldsymbol{J}_{k,i}$	The k-th joint tensor in the frame i.
8	$x_{k,i}, y_{k,i}, z_{k,i}$	The 3D coordinate of the k-th joint in the i-th frame.
9	N	The number of human skeletal joints used.
10	$\boldsymbol{S}$	The corresponding ground truth of the prediction.
11	M	The number of clipped sub-sequences in KEM.
12	T_m	The number of frames in the m-th sub-sequence.
13	$\boldsymbol{O}_m$	The m-th sub-sequence, preserving the recent T_m frames.
14	$\boldsymbol{O}_{key}$	The key sequence created by duplicating the last observed frame.
15	$\boldsymbol{F}_m$	The output encoder of 1D convolutional layer from $\boldsymbol{O_m}$
16	$\boldsymbol{F}_{key}$	The output encoder of 1D convolutional layer from $\boldsymbol{O_{key}}$
17	$\boldsymbol{F}_{KEM}$	The temporal reinforced features generated by KEM.
18	X	The serial number of MFE-block.
19	$G_g^X(\cdot)$	The graph convolutional function of GS-FE branch in the X-th MFE-block.
20	$G_d^X(\cdot)$	The graph convolutional function of DP-FE branch in the X-th MFE-block.
21	$\boldsymbol{F}_{out}^X$	The features generated by the X-th MFE-block.
22	$\boldsymbol{F}_{pa}^X$	The features of different pattern pa in the X-th MFE-block. $pa \in \{sta, ina, act\}$
23	sta, ina, act	Static, inactive, active pattern of human joint trajectories.
24	N_1, N_2, N_3	The number of joints in static, inactive, active pattern.

copied sequence $\boldsymbol{F}_{key}$ were encoded by a 1D convolutional layer. Finally, the output encoders of M subsequences and the $\boldsymbol{O}_{key}$ are concatenated together along the time dimension as the temporal features of the joint, which can be illustrated by Eq. (2):

$$\boldsymbol{F}_{KEM} = Concat(\boldsymbol{F}_1, \boldsymbol{F}_2, ..., \boldsymbol{F}_M, \boldsymbol{F}_{key}) \tag{2}$$

Compared with the traditional GCN encoder, which expands the time dimension as the channels to a fixed length, our KEM encodes temporal features more efficiently in a controllable manner and correspondingly enhances the feature of the last few frames, especially the latest frame in the sequence.

C. Multi-channel Feature Extractor block (MFE-block)

In Fig. 5, nine MFE-blocks are stacked together to acquire more deep information. Each MFE-block has two branches: the dynamic pattern-guided feature extractor (DP-FE) branch and the global spatial feature extractor (GS-FE) branch. Specifically, given the X-th MFE-block, the corresponding graph convolutional function of DP-FE and GS-FE branches can be presented as $G_d^X(\cdot)$, $G_g^X(\cdot)$ respectively. The output of the X-th MFE-block is presented as $\boldsymbol{F}_{out}^X$, which can be computed as Eq. (3).

$$\boldsymbol{F}_{out}^X = \begin{cases} G_g^X\left(\boldsymbol{F}_{out}^{X-1}\right) + G_d^X\left(\boldsymbol{F}_{out}^{X-1}\right), & 2 \le X \le 9 \\ G_g^X\left(\boldsymbol{F}_{KEM}\right) + G_d^X\left(\boldsymbol{F}_{KEM}\right), & X = 1 \end{cases} \tag{3}$$

Dynamic Pattern-guided Feature Extractor (DP-FE) branch. Our collaborative multi-dynamic pattern modeling strategy is shown in Fig. 5. As mentioned in the section Introduction, the calculation of movement amplitude of different joint trajectories shows that each joint trajectory is affected by the human body's kinematic connection. In relative coordinates, joints that close to the center of the body trunk are often carrying weak dynamic pattern, yet the motion pattern of the body extremity joints are often active. Therefore, collaborative modeling of joints with similar patterns helps the network refine recessive correlation on different limbs. We achieve the corresponding forward modeling of joint trajectories at different dynamic patterns by connecting several DP-FE branches. In the DP-FE branch, the joints are

congregated into the following three patterns according to the kinematic structure and movement amplitude: (1) static pattern, (2) inactive pattern, and (3) active pattern. Therefore, the input features of the X-th MFE-block $\boldsymbol{F}_{out}^{X-1}$ can be described as three patterns features in Eq. (4):

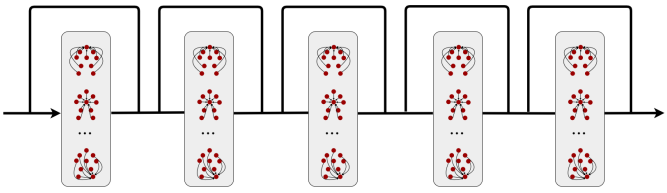
$$\boldsymbol{F}_{out}^{X-1} = \left\{\boldsymbol{F}_{sta}^{X-1}, \boldsymbol{F}_{ina}^{X-1}, \boldsymbol{F}_{act}^{X-1}\right\} \tag{4}$$

Then, the graph convolutional layers are adopted to learn the joint trajectory pattern in a three-channel structure.

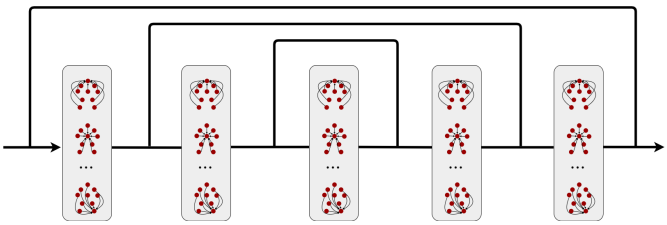
In graph convolutional layers, we make use of a symmetric residual graph convolution structure for modeling of human dynamics efficiently. Following the notations in [31], when modeling a joint set as a fully-connected graph of $N_i\,(i = 1, 2, 3)$ nodes (N_1, N_2, N_3 represents the number of joints in static, inactive, active pattern, respectively), a graph convolutional layer can be formulated as Eq. (5):

$$\boldsymbol{H}^{(p+1)} = \sigma\left(\boldsymbol{A}^{(p)}\boldsymbol{H}^{(p)}\boldsymbol{W}^{(p)}\right) \tag{5}$$

where $\boldsymbol{W}^{(p)}$ represents the learnable weights of the p-th graph convolutional layer. Noted in the X-th MFE-block, the first graph convolutional layer $\boldsymbol{H}^{(1)} = \boldsymbol{F}_{pa}^{X}$, $pa \in \{sta, ina, act\}$. And a set of learnable parameters $\boldsymbol{A}^{(p)}$ is utilized to learn the strength of the edges in the graph $\boldsymbol{H}^{(p)}$ as the adjacency matrix. Each layer is followed with an activation function $\sigma(\cdot)$ and a dropout option. To this end, each graph convolutional layer models both hidden features and graph connectivity of input joint nodes though the joints are not directly connected in the human body. When stacking graph convolutional layers, we use symmetric residual connections instead of equidistant residual connections as in Fig. 6. Compared with the traditional equidistant connection, the symmetric residual connection adopted in our work provides a closer distance between output and input, and introduces shallower dynamic features to the end of the module. Consequently, the following DP-FE branches can inherit the initial features by the first symmetric residual connection of the previous DP-FE in each channel, enriching the multi-granularity features.



(a) Equidistant residual connection.



(b) Symmetric residual connection. (Noted as "SRes-GCN" in Fig. 5)

Fig. 6. Residual graph convolution block. Compared with traditional equidistant residual connections, our symmetric residual connection transmits richer features from different depths and delivers the initial features of a block.

In the training phase, the learnable weights of the graph convolution layers can simulate the motion dynamic pattern. Therefore, a multi-channel graph structure is used to model the trajectory features with different levels of dynamic respectively, which reduces the complexity of the motion pattern to be simulated by the network compared with the undifferentiated modeling methods. On the other hand, our collaborative modeling uses a smaller adjacency matrix, and makes it easier to establish spatial correlation within the same dynamic pattern. Finally, the three-channel features are restored to match the global features according to the joint order before the previous pattern congregating to ensure the spatial consistency of the subsequent feature coupling.

Global spatial feature extractor (GS-FE) branch. Another problem is that the multi-channel modeling cannot obtain the global spatial connection relations of all adjacent joints. This results in the deficiency of global spatial information as the network deepens. Therefore, our network uses an additional branch to introduce global spatial features to DP-FE branch. In the X-th MFE-block, the preliminary global spatial feature is denoted as $\boldsymbol{F}_{out}^{X-1}$, and this branch adopts three graph convolutional layers to produce shallow global spatial features, which is represented as $G_g^X(\boldsymbol{F}_{out}^{X-1})$. In this way, the utilization of shallow features introduces global spatial features and helps avoid gradient disappearance. Consequently, as the network goes deeper, the global branch prevents weak coupling when restoring trajectory features from the different dynamic patterns.

D. Fully connected layer

For the output of stacked MFE-blocks $\boldsymbol{F}_{out}^{X}$, we apply a full connection to produce the predicted motion sequence $\widehat{\boldsymbol{S}}$, noted $FC(\cdot)$ is a fully connected operator, which can be described in Eq. (6):

$$\widehat{\boldsymbol{S}} = FC(\boldsymbol{F}_{out}^{X}), X = 9 \tag{6}$$

IV. Experiments

A. Baselines

We compare our model with recent effective state-of-the-art methods in 3D coordinate space, including RNN-based [7], [9], CNN-based [5], [22] and GCN-based [6], [10], [31], [34]. Specifically, we compare the effectiveness of our multi-dynamic patterns with [5], [7], [34]. Liu et al. [5] adopted a closely spaced joint arrangement in the input sequence to refine the spatial feature. Cai et al. [7] utilized a transformer-based method to decode the joint spatial features from torso to limb extremities. Dang et al. [34] extracted the spatial features by multiple levels and implemented intermediate supervisions for multi-level modeling. We compare the effectiveness of our graph convolution with [6], [10], [31], [34] to evaluate the overall graph modeling performance.

B. Implementation Details

We implement our model with Pytorch [42] on NVIDIA 2080ti GPU. Each graph convolutional layer in our residual block is followed with Leaky-Relu [47] activation and dropout

TABLE II
SHORT-TERM RESULTS ON H3.6M. RESULTS AT 80MS, 160MS, 320MS, AND 400MS IN THE FUTURE ARE SHOWN. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

	Walking				Eating				Smoking				Discussion			
Milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Res sup [9]	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
Seq2Seq [22]	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 [31]	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
DMGNN [6]	9.3	15.1	28.6	35.2	8.5	15.4	37.2	46.8	8.5	14.4	27.1	30.4	10.2	20.8	39.7	46.3
Traj [5]	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
TIM [10]	9.3	15.9	30.1	34.1	8.4	18.5	38.1	46.6	6.9	13.8	24.6	29.1	8.8	21.3	40.2	45.5
LPJP [7]	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
MSR [34]	8.7	15.5	**28.3**	**32.3**	**8.3**	**17.7**	**36.3**	43.6	7.5	15.4	27.3	31.4	9.3	22.1	40.6	45.6
DPnet	**7.3**	15.2	30.1	32.6	8.6	18.3	36.4	**43.5**	6.9	13.5	24.3	28.7	8.2	20.1	**38.2**	**43.0**
	Directions				Greeting				Phoning				Posing			
Milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Res sup [9]	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
Seq2Seq [22]	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 [31]	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
DMGNN [6]	12.9	26.2	48.8	58.0	14.3	29.6	74.5	87.8	11.2	18.6	37.1	45.8	9.0	23.6	67.3	84.2
Traj [5]	**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
TIM [10]	11.0	22.3	48.4	59.3	13.7	29.1	72.6	88.9	11.5	19.8	38.5	44.4	7.5	22.3	64.8	80.8
LPJP [7]	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
MSR [34]	11.4	22.0	45.9	**56.2**	13.5	**26.5**	68.8	86.2	11.8	20.6	37.6	41.8	8.5	21.7	**61.1**	**76.3**
DPnet	10.1	**21.0**	**45.8**	56.7	12.7	27.1	65.6	82.9	**10.2**	**17.4**	**35.7**	**41.3**	7.4	21.9	63.5	78.8
	Purchases				Sitting				Sitting Down				Taking Photo			
Milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Res sup [9]	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
Seq2Seq [22]	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 [31]	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
DMGNN [6]	19.8	37.7	62.8	74.3	10.5	24.3	49.8	61.9	12.8	28.4	55.2	69.1	8.2	15.6	38.9	53.7
Traj [5]	**17.1**	**36.1**	64.3	75.1	**9.0**	**22.0**	49.4	62.6	10.7	28.8	55.1	62.9	**5.4**	**13.4**	36.2	**47.0**
TIM [10]	19.0	39.2	65.9	74.6	9.3	22.3	**45.3**	**56.0**	11.3	28.0	54.8	64.8	6.4	15.6	41.4	53.5
LPJP [7]	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
MSR [34]	18.9	38.7	64.5	72.5	11.3	26.6	56.2	69.3	11.0	28.2	56.2	66.8	6.6	15.7	40.5	52.9
DPnet	17.8	37.0	62.1	**65.6**	9.1	23.0	48.1	62.8	**9.7**	**24.2**	**49.7**	**62.0**	5.7	14.4	**35.6**	47.9
	Waiting				Walking Dog				Walking Together				Average			
Milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Res sup [9]	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
Seq2Seq [22]	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 [31]	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
DMGNN [6]	9.0	21.4	56.7	72.8	30.4	57.2	105.6	120.8	8.6	19.0	35.7	45.2	12.2	24.5	51.0	62.1
Traj [5]	**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
TIM [10]	9.2	21.7	55.9	72.1	29.3	56.4	99.6	119.4	8.9	18.6	35.5	44.3	11.4	24.3	50.4	60.9
LPJP [7]	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
MSR [34]	8.9	20.8	53.6	69.7	24.4	53.6	95.8	**110.6**	8.7	18.6	35.4	45.7	11.2	24.2	49.8	60.0
DPnet	8.4	**20.5**	53.6	69.1	25.7	51.8	**94.9**	112.3	**8.3**	18.8	35.6	44.8	10.3	**22.9**	**47.9**	**58.1**

option. Our model is trained with Adam optimizer [43]. In the training phase, the batch size and learning rate are respectively set to 16 and 0.0005.

Our experiments are carried out in 3D coordinate space. Thus we use mean per joint position error (MPJPE) [8] as the loss function to train our model, as illustrated in Eq. (7).

$$loss = MPJPE = \frac{1}{L*N}\sum_{i=1}^{L}\sum_{k=1}^{N}\|\widehat{J_{k,i}} - J_{k,i}\|^2 \quad (7)$$

MPJPE is widely utilized as the evaluation matrix in human motion prediction tasks [10], [31], [34]. It calculates the average Euclidean distance error between the predicted joint and the corresponding ground truth joint of the sequence. And we also evaluate our model by MPJPE in the following analysis.

To be consistent with the literature, we evaluate our results on both short-term and long-term predictions. The length of input frames is set to 10, and the output lengths are respectively set to 10 (400ms) and 25 (1000ms) on both H3.6M dataset [8] and CMU-Mocap dataset [44]. And the output length on 3DPW [45] is 30 (1000ms).

C. Datasets

H3.6M. H3.6M [8] is the most commonly used dataset on human motion prediction, containing 3.6 million motion sequences on 15 activities performed by 7 professional subjects. There are 32 joints in each pose, and 22 are used in our training phase following [31]. Specifically, subject 1, 6, 7, 8, 9 are used as training data, and subject 5 is used as testing data. Each sequence is down-sampled by 2.

CMU-Mocap. CMU-Mocap [44] provides 2,235 human motion sequences. Each pose contains 38 joints. Following baselines in [31], we preserve 25 joints to train our model, and 8 actions are selected from 5 categories: "locomotion", "physical activities & sports", "common behaviors and expressions" and "communication gestures and signals." The same dataset splits for training and testing are also the same with [31].

TABLE III
LONG-TERM RESULTS ON H3.6M. RESULTS AT 560MS, AND 1000MS IN THE FUTURE ARE SHOWN. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

	Walking		Eating		Smoking		Discussion		Directions		Greeting		Phoning		Posing	
Milliseconds	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000
Res sup [9]	73.8	86.7	101.3	119.7	85.0	118.5	120.7	147.6	-	-	-	-	-	-	-	-
Seq2Seq [22]	59.2	71.3	66.5	85.4	42.0	67.9	84.1	116.9	-	-	-	-	-	-	-	-
LTD [31]	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
Traj [5]	**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
TIM [10]	39.6	46.9	56.9	**68.6**	33.5	61.7	68.5	97.0	80.1	105.7	97.4	90.6	64.2	**111.5**	107.8	218.7
MSR [34]	42.1	**43.5**	57.1	71.5	35.1	62.3	75.6	113.4	**78.6**	**101.7**	100.2	95.2	63.6	113.8	**103.1**	219.6
DPnet	40.5	48.6	**56.5**	69.6	32.8	59.9	**66.3**	**96.7**	80.2	103.5	93.7	85.6	**61.4**	113.9	105.9	**205.6**
	Purchases		Sitting		Sitting Down		Taking Photo		Waiting		Walking Dog		Walking Together		Average	
Milliseconds	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000
Res sup [9]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Seq2Seq [22]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
LTD [31]	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
Traj [5]	**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
TIM [10]	93.8	131.6	**68.4**	106.8	87.4	144.7	79.1	97.8	98.1	170.7	**135.9**	175.1	56.4	78.0	77.8	113.7
MSR [34]	86.5	125.3	83.0	**103.8**	83.2	146.1	72.5	95.8	100.7	**164.4**	144.5	193.7	**55.6**	84.5	78.8	115.6
DPnet	94.2	123.2	72.5	106.9	84.6	131.0	74.0	**83.2**	96.7	167.7	136.7	174.9	59.8	78.1	**77.0**	**109.9**

TABLE IV
RESULTS ON CMU-MOCAP. RESULTS AT 80MS, 160MS, 320MS, 400MS, AND 1000MS IN THE FUTURE ARE SHOWN. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

	Basketball					Basketball Signal					Directing Traffic				
Milliseconds	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
Res sup [9]	18.4	33.8	59.5	70.5	106.7	12.7	23.8	40.3	46.7	77.5	15.2	29.6	55.1	66.1	127.1
LTD [31]	14.0	25.4	49.6	61.4	106.1	3.5	6.1	11.7	15.2	53.3	7.4	15.1	31.7	42.2	152.4
DMGNN [6]	13.6	24.9	49.4	62.0	105.7	3.3	5.9	13.1	15.6	55.5	7.6	14.5	30.9	41.6	148.3
Traj [5]	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**
LPJP [7]	11.6	21.7	44.4	57.3	90.9	2.6	4.9	12.7	18.7	75.8	6.2	12.7	29.1	39.6	149.1
TIM [10]	12.7	22.6	44.6	55.6	102.0	3.0	5.6	11.6	15.5	57.0	7.1	14.1	31.1	41.4	138.3
MSR [34]	12.6	22.9	44.5	54.5	**89.7**	2.7	4.8	11.1	14.7	**49.1**	6.9	14.5	30.1	39.7	117.7
DPnet	**10.7**	**17.8**	**38.4**	**49.5**	98.4	2.6	4.4	10.0	13.4	61.2	5.9	11.8	26.6	33.5	143.3
	Jumping					Running					Soccer				
Milliseconds	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
Res sup [9]	36.0	68.7	125.0	145.5	195.5	**15.6**	19.4	31.2	36.2	43.3	20.3	39.5	71.3	84.0	129.6
LTD [31]	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
DMGNN [6]	16.6	34.0	74.6	95.8	162.4	25.1	38.3	39.5	39.9	59.7	11.9	21.4	44.5	56.1	115.8
Traj [5]	**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
LPJP [7]	12.9	**27.6**	73.5	92.2	176.6	23.5	34.2	35.2	36.1	43.1	9.2	18.4	39.2	49.5	**93.9**
TIM [9]	14.8	31.1	71.2	91.3	163.5	24.5	37.0	39.9	41.9	62.6	11.2	22.1	45.1	58.1	122.1
MSR [34]	15.1	30.6	73.2	95.3	**160.5**	20.4	26.4	26.9	28.0	**34.1**	8.4	**15.9**	36.0	**46.1**	108.9
DPnet	12.4	28.3	**70.2**	**89.2**	166.1	16.7	**18.4**	**19.6**	**25.1**	40.1	9.0	17.1	**35.8**	48.7	115.0
	Walking					Washing Window					Average				
Milliseconds	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
Res sup [9]	8.2	13.7	21.9	24.5	32.2	8.4	15.8	29.3	**35.4**	**61.1**	16.8	30.5	54.2	63.6	96.6
LTD [31]	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
DMGNN [6]	8.3	12.4	21.9	23.6	41.0	5.8	11.5	29.7	39.3	76.8	11.5	20.3	38.0	46.7	95.5
Traj [5]	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
LPJP [7]	6.7	10.7	21.7	27.5	37.4	5.4	11.3	29.2	39.6	79.1	9.8	17.6	35.7	45.1	93.2
TIM [10]	7.1	11.1	19.9	22.8	39.3	5.9	12.3	32.1	42.6	80.4	10.8	19.5	36.9	46.2	95.7
MSR [34]	6.5	10.5	18.0	21.6	34.9	5.3	11.3	29.8	39.7	82.2	9.7	17.1	33.7	42.5	**84.6**
DPnet	**5.8**	**9.0**	**17.2**	**21.4**	**34.1**	**4.5**	9.8	**27.3**	36.7	72.1	8.4	**14.5**	**30.6**	**39.7**	91.3

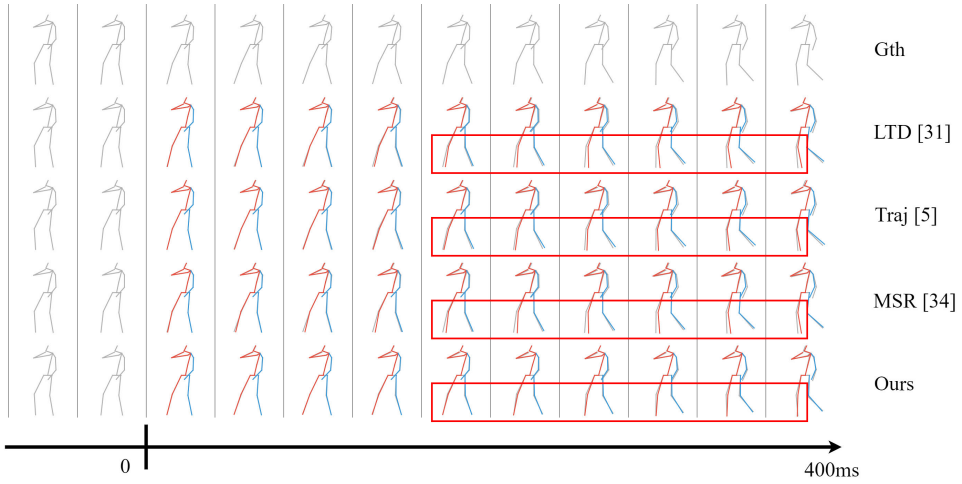
3DPW. 3DPW [45] consists of indoor and outdoor actions such as shopping, doing sports, and hugging, including 60 sequences and more than 51k frames. Each pose is composed of 24 joints. We use the official split sets for experiments for fair comparison.

D. Results

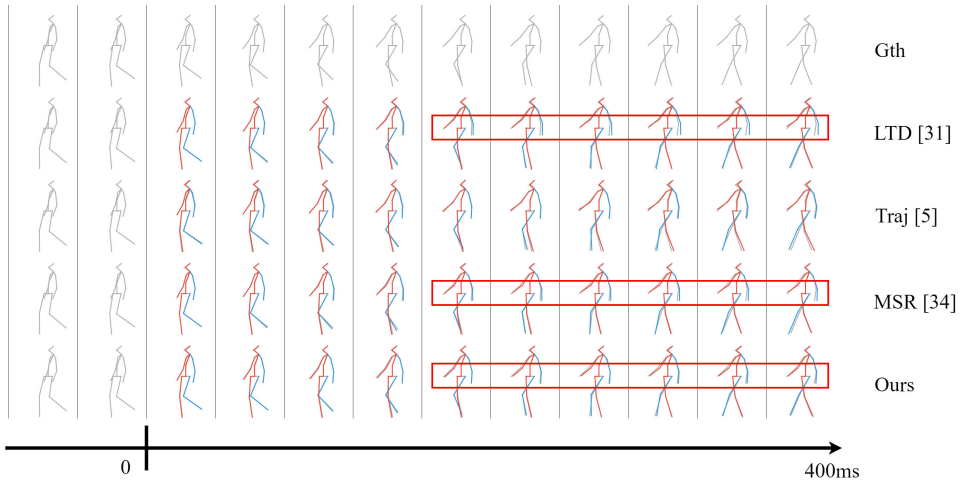
Results on H3.6M. Our results for short-term and long-term prediction MPJPE on H3.6M dataset are shown in Table II and Table III. Significantly, our method gives much better prediction results than the other 3 GCN-based methods. Specifically, TIM [10] performs better than LTD [31], and MSR [34] performs better than TIM [10], but our method is the best on all prediction time steps. And our method also outperforms the RNN-based [7].

To explain the effectiveness of our KEM in detail, TIM [10] conducts multi-channel modeling on different temporal observation lengths, placing more significance on the latest observed frames than LTD [31]. Our KEM introduces an additional channel for modeling a new sequence composed by copying the last input frame. Thus, the gain comes from placing more significance on the last frame. More discussion will be given in Ablation Study.

To explain the effectiveness of our DP-FE in detail, the



(a) Phoning.



(b) Walking.

Fig. 7. Quality comparison on H3.6M dataset. The 1st line represents the ground truth. The 2nd to the 4th lines represent the baselines. The last line represents the DPnet results. The results indicate that our method generates more realistic joint trajectories on body extremities.

main idea of MSR [34] is to predict human motion in multiple spatial scales. The fine to coarse scales are designed according to human kinematic connections. However, each spatial scale still contains both active and static joint trajectories, making the corresponding GCN blocks harder to estimate the dynamic pattern than our DP-FE. Therefore, our pattern collaboration benefits the graph convolutional layer to model the dynamic pattern within different dynamic scales. Cai et al. [7] also proposed to model human dynamics from torso trajectory to extremity trajectory progressively. Its transformer-based network converts the traditional temporal prediction task to a spatial inpainting task in the decoding phase, but the recurrent modeling suffers from losing global spatial features compared with our global branch that introduces global spatial features.

Unfortunately, our short-term prediction on 80ms has a slight gap with Traj [5], though other predictions of DPnet achieve lower MPJPE, including 560ms and 1000ms. From methodology, it can be concluded that CNN-based TrajectoryNet [5] achieves better local perception and shorter temporal channels than DPnet, which benefits the most recent prediction.

In Fig. 7, we give our examples of quality comparison on H3.6M dataset [8]. In the "Phoning" action, the subject maintains the motion of holding the phone with his right arm, walks forward with his legs and moves his left arm in coordination. At this time, the accuracy of our prediction results in the extremity of the leg and left arm is better than other methods. In the "Walking" action, the torso remains relatively static, and the limbs swing in coordination. Our method achieves the closest prediction results to the ground truth, verifying the importance of collaboration modeling of active joints.

Results on CMU-Mocap and 3DPW. Our prediction results on CMU-Mocap and 3DPW datasets are shown in Table IV and Table V. Our method still outperforms most GCN-based methods. Yet Traj [5] shows lower MPJPE again on 80ms prediction in CMU-Mocap, which confirms our analysis of weakness in H3.6M dataset [8]. It is worth mentioning that our DPnet achieves the best accuracy for "Walking" and "Running" actions on all time steps in Table IV. That explains

TABLE V
RESULTS ON 3DPW. RESULTS AT 200MS, 400MS, 600MS, 800MS, AND 1000MS IN THE FUTURE ARE SHOWN. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Milliseconds	200	400	600	800	1000
Ressup [9]	113.9	173.1	191.9	201.1	210.1
Seq2Seq [22]	71.6	124.9	155.4	174.7	187.5
LTD [31]	35.6	67.8	90.6	106.9	117.8
DMGNN [6]	37.3	70.1	94.5	109.7	123.6
DPnet	**29.5**	**58.0**	**84.7**	**103.1**	**109.3**

TABLE VI
ABLATION STUDY ON KEM. RESULTS AT 80MS, 160MS, 320MS, AND 400MS IN THE FUTURE ARE SHOWN. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Milliseconds	80	160	320	400
w/o lf	11.1	24.3	50.1	60.6
w/o trimming	10.8	24.1	49.7	60.0
DPnet	**10.3**	**22.9**	**47.9**	**58.1**

the effectiveness of DP-FE for modeling the regular dynamic pattern of persistent movements.

E. Ablation Study

In this section, we carry out ablation experiments on our main proposed components, and discuss the effectiveness of different experiment settings.

Effectiveness of KEM. The ablation structures are shown in Fig. 8. Firstly, we remove the duplicate channel in the KEM to analyze the importance of the last frame, noted as "w/o lf" (Fig. 8(b)). As shown in Table VI, the MPJPE is higher than the origin KEM, confirming the dominance of the last frame. Secondly, we cancel the recent frame trimming to analyze the effectiveness of temporal focusing, noted as "w/o trimming" (Fig. 8(c)). It means the input length of each stream is 10. It is found that the observation of total sequence in each stream disturbs the key information of a motion sequence, though "w/o trimming" keeps the coping stream. Consequently, the results support the effectiveness of the recent frames, which is also emphasized in TIM [10].

Effectiveness of GS-FE. To verify the effectiveness of our global spatial stream, we conduct an ablation experiment on removing our proposed global spatial feature extractor (GS-FE) in Fig. 5. The results of "w/o GS-FE" denotes the structure of removing the residual GS-FE of each DP-FE, which are shown in Table VII. The result of "w/o GS-FE" on each prediction length shows lower accuracy than DPnet that using GS-FE. This confirms that the global spatial features help reinforce the human dynamic as the network goes deeper. The GS-FE focuses on modeling the global joint trajectories and the DP-FE focuses on modeling the pattern-wise joint trajectories. Hence, the collaborative modeling strategy allows greater depth of our DPnet for sufficient extracting human dynamics.

TABLE VII
ABLATION STUDY ON GS-FE. RESULTS AT 80MS, 160MS, 320MS, AND 400MS IN THE FUTURE ARE SHOWN. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Milliseconds	80	160	320	400
w/o GS-FE	12.0	26.0	53.2	63.8
DPnet	**10.3**	**22.9**	**47.9**	**58.1**

TABLE VIII
ABLATION EXPERIMENTS ON DIFFERENT DYNAMIC PATTERN COLLABORATIONS IN DP-FE. RESULTS AT 80MS, 160MS, 320MS, AND 400MS IN THE FUTURE ARE SHOWN. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Milliseconds	80	160	320	400
Sta+Ina, Act	10.6	23.6	49.0	59.2
Ina+Act, Sta	10.7	23.5	49.3	59.7
Sta+Act, Ina	10.6	23.2	48.7	59.1
Sma, Lar	10.9	23.8	49.7	60.0
DPnet	**10.3**	**22.9**	**47.9**	**58.1**

Effectiveness of DP-FE. To verify the effectiveness of our collaboration modeling strategy, we conduct experiments on different dynamic scales. The results of "Sta+Ina, Act", "Ina+Act, Sta" and "Sta+Act, Ina" denote the structure of fusing three streams of DP-FE in pairs for forwarding modeling. And "Sma, Lar" denotes splitting human joints into two dynamic levels based on their motion amplitude: small range level and large range level. And the results are shown in Table VIII. Results verify the effectiveness of our three-stream modeling strategy compared with other settings. It can be inferred that the joints from Ina have a unique dynamic pattern, and the dynamic features are deranged when combing Ina with other levels. Moreover, when splitting features evenly into Sta level and Act level according to the joint distances to the human torso as "Sma, Lar", the network performs the worst accuracy. The results again confirmed the necessity of distinguishing the Ina pattern. Therefore, it is necessary to build the dynamic pattern of each dynamic level in three streams.

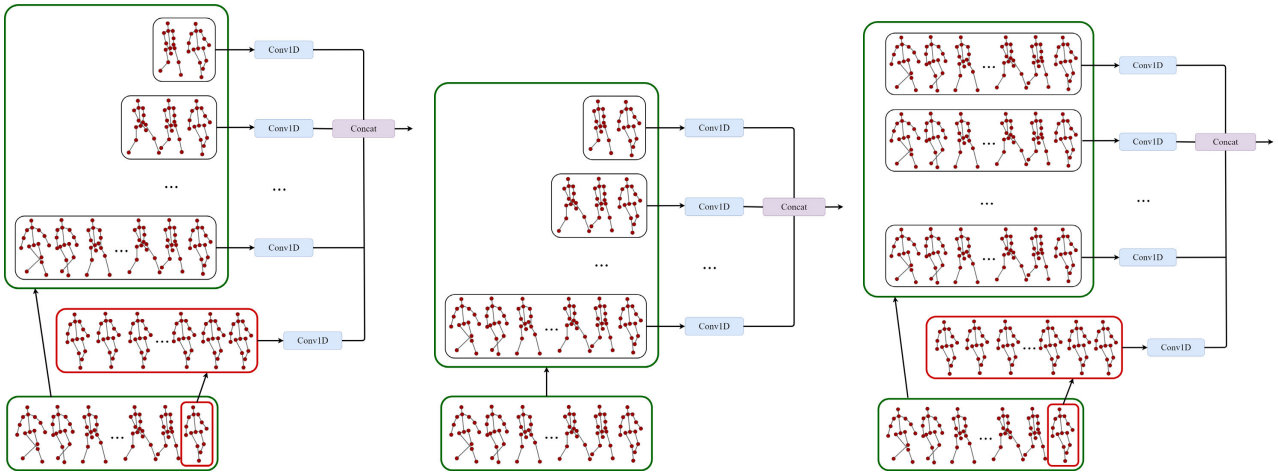
V. CONCLUSION

In this paper, we propose a new human body representation based on our dynamic patterns and the dynamic pattern-based collaborative modeling framework to predict future 3D poses. We use Dynamic Pattern-guided Feature Extractor (DP-FE) to congregate joint-wise trajectory features, modeling more comprehensive relations, especially the recessive spatial relations of indirectly connected joints. Moreover, KEM is proposed to enhance the latest observed information, advancing the inertial pattern of the temporal features. Experiment results on three common benchmark datasets prove that the dynamic pattern collaboration modeling and temporal reinforcing strategies lower the human motion prediction error.

Our future work will focus on optimizing the DP-EF to autonomously distinguish human dynamic patterns from trajectory features. Compared with fixing the same dynamic pattern of each channel, adaptive exploiting different patterns of spatial features may improve the network flexibility on different datasets.

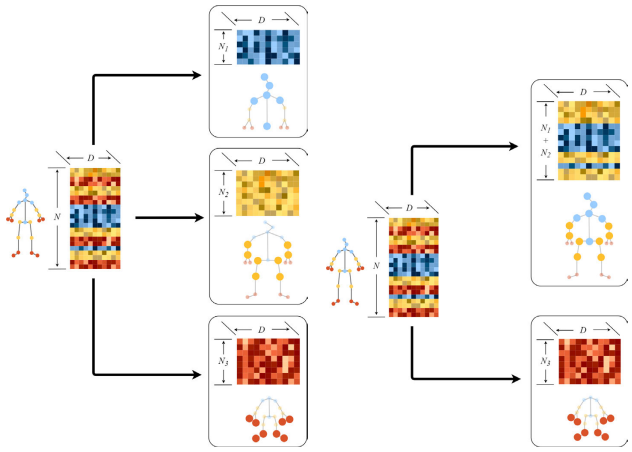
ACKNOWLEDGMENT

This work was supported partly by the National Natural Science Foundation of China (Grant No. 62173045, 61673192), and partly by the Fundamental Research Funds for the Central Universities (Grant No. 2020XD-A04-2).



(a) KEM. (b) w/o lf. (c) w/o trimming.

Fig. 8. Ablation study on KEM. "w/o lf" removes the key sequence $\boldsymbol{O}_{key}$. And "w/o trimming" preserves all frames of the sub-sequences.



(a) Sta, Ina, Act. (DPnet). (b) Sta+Ina, Act.

(c) Ina+Act, Sta. (d) Sta+Act, Ina.

Fig. 9. Different dynamic pattern collaborations in DP-FE. Compared with "Sta, Ina, Act" utilized in DPnet, the other three strategies fuse the dynamic patterns in pairs to conduct two-channel modeling in DP-FE.

References

[1] R. Li, H. Wang, and Z. Liu, "Survey on mapping human hand motion to robotic hands for teleoperation," IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 5, pp. 2647–2665, 2021. I

[2] K. Qian, X. Ma, X. Dai, and F. Fang, "Robotic etiquette: Socially acceptable navigation of service robots with human motion pattern learning and prediction," Journal of Bionic Engineering, vol. 7, no. 2, pp. 150–160, 2010. I

[3] T. Andrews, B. Searcy, and B. Wallace, "Using virtual reality and motion capture as tools for human factors engineering at nasa marshall space flight center," in International Conference on Applied Human Factors and Ergonomics. Springer, 2019, pp. 399–408. I

[4] G. F. Welch, "History: The use of the kalman filter for human motion tracking in virtual reality," Presence, vol. 18, no. 1, pp. 72–91, 2009. I

[5] X. Liu, J. Yin, J. Liu, P. Ding, J. Liu, and H. Liu, "Trajectorycnn: a new spatio-temporal feature learning network for human motion prediction," IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 6, pp. 2133–2146, 2020. I, II, IV-A, IV-A, IV-A, II, III, IV, IV-D, IV-D, IV-D

[6] 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 214–223. I, II, IV-A, IV-A, II, IV, V

[7] Y. Cai, L. Huang, Y. Wang, T.-J. Cham, J. Cai, J. Yuan, J. Liu, X. Yang, Y. Zhu, X. Shen et al., "Learning progressive joint propagation for human motion prediction," in European Conference on Computer Vision. Springer, 2020, pp. 226–242. I, II, IV-A, IV-A, IV-A, II, IV, IV-D, IV-D

[8] 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 transactions on pattern analysis and machine intelligence, vol. 36, no. 7, pp. 1325–1339, 2013. (document), I, IV-B, IV-B, IV-C, IV-D, IV-D

[9] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2891–2900. I, II, IV-A, II, III, IV, V

[10] T. Lebailly, S. Kiciroglu, M. Salzmann, P. Fua, and W. Wang, "Motion prediction using temporal inception module," in Proceedings of the Asian Conference on Computer Vision, 2020. I, II, III-A, III-B, III-B, III-B, IV-A, IV-A, II, IV-B, III, IV, IV-D, IV-D, IV-D, IV-E

[11] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 4346–4354. II

[12] A. Gopalakrishnan, A. Mali, D. Kifer, L. Giles, and A. G. Ororbia, "A neural temporal model for human motion prediction," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12116–12125. II

[13] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-rnn: Deep learning on spatio-temporal graphs," in Proceedings of the ieee conference on computer vision and pattern recognition, 2016, pp. 5308–5317. II

[14] H. Wang, J. Dong, B. Cheng, and J. Feng, "Pvred: a position-velocity recurrent encoder-decoder for human motion prediction," IEEE Transactions on Image Processing, vol. 30, pp. 6096–6106, 2021. II

[15] B. Wang, E. Adeli, H.-k. Chiu, D.-A. Huang, and J. C. Niebles, "Imitation learning for human pose prediction," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7124–7133. II

[16] E. Barsoum, J. Kender, and Z. Liu, "Hp-gan: Probabilistic 3d human motion prediction via gan," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 1418–1427. II

[17] J. Butepage, M. J. Black, D. Kragic, and H. Kjellstrom, "Deep representation learning for human motion prediction and classification," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6158–6166. II

[18] P. Ghosh, J. Song, E. Aksan, and O. Hilliges, "Learning human motion models for long-term predictions," in 2017 International Conference on 3D Vision (3DV). IEEE, 2017, pp. 458–466. II

[19] H.-k. Chiu, E. Adeli, B. Wang, D.-A. Huang, and J. C. Niebles, "Action agnostic human pose forecasting," in 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2019, pp. 1423–1432. II

[20] A. F. Al-aqel and M. A. Khan, "Attention mechanism for human motion prediction," in 2020 3rd International Conference on Computer Applications & Information Security (ICCAIS). IEEE, 2020, pp. 1–6. II

[21] L.-Y. Gui, Y.-X. Wang, X. Liang, and J. M. Moura, "Adversarial geometry-aware human motion prediction," in Proceedings of the european conference on computer vision (ECCV), 2018, pp. 786–803. II

[22] C. Li, Z. Zhang, W. S. Lee, and G. H. Lee, "Convolutional sequence to sequence model for human dynamics," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5226–5234. II, IV-A, II, III, V

[23] Q. Cui, H. Sun, Y. Kong, X. Zhang, and Y. Li, "Efficient human motion prediction using temporal convolutional generative adversarial network," Information Sciences, vol. 545, pp. 427–447, 2021. II

[24] J. Tang, J. Zhang, and J. Yin, "Temporal consistency two-stream cnn for human motion prediction," Neurocomputing, vol. 468, pp. 245–256, 2022. II

[25] P. Ding and J. Yin, "Towards more realistic human motion prediction with attention to motion coordination," IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 9, pp. 5846–5858, 2022. II

[26] X. Shu, L. Zhang, G.-J. Qi, W. Liu, and J. Tang, "Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 6, pp. 3300–3315, 2021. II

[27] E. Aksan, M. Kaufmann, and O. Hilliges, "Structured prediction helps 3d human motion modelling," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7144–7153. II

[28] Q. Cui and H. Sun, "Towards accurate 3d human motion prediction from incomplete observations," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4801–4810. II

[29] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 6, pp. 3316–3333, 2021. II

[30] W. Mao, M. Liu, and M. Salzmann, "History repeats itself: Human motion prediction via motion attention," in European Conference on Computer Vision. Springer, 2020, pp. 474–489. II

[31] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning trajectory dependencies for human motion prediction," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9489–9497. II, III-A, III-C, IV-A, IV-A, II, IV-B, IV-C, IV-C, IV-C, III, IV, IV-D, IV-D, V

[32] M. Li, S. Chen, Z. Liu, Z. Zhang, L. Xie, Q. Tian, and Y. Zhang, "Skeleton graph scattering networks for 3d skeleton-based human motion prediction," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 854–864. II

[33] T. Ma, Y. Nie, C. Long, Q. Zhang, and G. Li, "Progressively generating better initial guesses towards next stages for high-quality human motion prediction," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 6437–6446. II

[34] L. Dang, Y. Nie, C. Long, Q. Zhang, and G. Li, "Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11467–11476. II, IV-A, IV-A, IV-A, IV-A, II, IV-B, III, IV, IV-D, IV-D

[35] H. Zhou, C. Guo, H. Zhang, and Y. Wang, "Learning multiscale correlations for human motion prediction," in 2021 IEEE International Conference on Development and Learning (ICDL). IEEE, 2021, pp. 1–7. II

[36] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," IEEE Transactions on Image Processing, vol. 29, pp. 9532–9545, 2020. II

[37] Y. Song, Z. Zhang, C. Shan, and L Wang, "Stronger, Faster and More Explainable: A Graph Convolutional Baseline for Skeleton-based Action Recognition," In Proceedings of the 28th ACM International Conference on Multimedia (MM '20). Association for Computing Machinery, 2020, pp. 1625–1633. II

[38] J. Vig, "A multiscale visualization of attention in the transformer model," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2019, pp. 37–42. II

[39] D. A. White, W. J. Arrighi, J. Kudo, and S. E. Watts, "Multiscale topology optimization using neural network surrogate models," Computer Methods in Applied Mechanics and Engineering, vol. 346, pp. 1118–1135, 2019. II

[40] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in Thirty-second AAAI conference on artificial intelligence, 2018. II

[41] J. Tang, J. Liu, and J. Yin, "A hierarchical static-dynamic encoder-decoder structure for 3d human motion prediction with residual cnns," Mathematical Problems in Engineering, vol. 2020, 2020. II, III-B

[42] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga et al., "Pytorch: An imperative style, high-performance deep learning library," Advances in neural information processing systems, vol. 32, 2019. IV-B

[43] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017. IV-B

[44] CMU. (2003)Graphics lab motion capture database. [Online]. Available: http://mocap.cs.cmu.edu/ (document), IV-B, IV-C

[45] T. Von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. PonsMoll, "Recovering accurate 3d human pose in the wild using imus and a moving camera," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 601–617. (document), IV-B, IV-C

[46] J. Bruna, W. Zaremba, A. Szlam, and Y. Le-Cun. "Spectral networks and locally connected networks on graphs," ICLR, 2014. I, II

[47] L. Andrew, Y. Awni, and Y. Andrew, "Rectifier nonlinearities improve neural network acoustic models," Proceedings of the 30th International Conference on Machine Learning, 2013. IV-B

Jin Tang received the Ph.D. degree from Beijing Institute of Technology, Beijing, China, in 2007. She currently is a Assistant Professor with Artificial Intelligence School, Beijing University of Posts and Telecommunications, Beijing, China. Her research interests include signal processing, pattern recognition, and deep learning.

Jin Zhang received the bachelor's degree from Beijing University of Post and Telecommunications, Beijing, China, in 2022. He currently is a researcher of Lenovo Research. His research interests include human motion prediction and image synthesis in computer vision.

Rui Ding received the Ph.D. degree from Beijing Institute of Technology, Beijing, China, in 2009. He currently is a Assistant Professor with Information Engineering College, Capital Normal University, Beijing, China. His research interests include signal processing, Embedded system, and deep learning.

BaoXuan Gu received the B.S. degree from the Beijing University of Post and Telecommunications, Beijing, China, in 2022, where he is currently pursuing the M.S. degree with the School of Artificial Intelligence. His research interest is human motion prediction.

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.