TrajectoryCNN: A New Spatio-Temporal Feature Learning Network for Human Motion Prediction

Xiaoli Liu, Jianqin Yin, *Member, IEEE*, Jin Liu, Pengxiang Ding, Jun Liu, *Member, IEEE*, and Huaping Liu, *Senior Member, IEEE*

***Abstract*— Human motion prediction is an increasingly interesting topic in computer vision and robotics. In this paper, we propose a new end-to-end feedforward network, TrajectoryCNN, to predict future poses. Compared with the most existing methods, we introduce a new trajectory space and focus on modeling motion dynamics of the input sequence with coupled spatio-temporal features, dynamic local-global features, and global temporal co-occurrence features in the new space. Specifically, the coupled spatio-temporal features describe the spatial and temporal structural information hidden in a natural human motion sequence, which can be easily mined using CNN by simultaneously covering the spatial and temporal dimensions of the sequence with the convolutional filters. The dynamic local-global features encode different correlations among joint trajectories of human motion (i.e. strong correlations among joint trajectories of one part and weak correlations among joint trajectories of different parts), which can be captured by stacking multiple residual trajectory blocks and incorporating our skeletal representation. The global temporal co-occurrence features represent different importance of different input poses to mine the motion dynamics for predicting future poses, which can be obtained automatically by learning free parameters for each pose with our TrajectoryCNN. Finally, we predict future poses with the captured motion dynamic features in a non-recursive manner. Extensive experiments show that our method achieves state-of-the-art performance on five benchmarks (e.g. Human3.6M, CMU-Mocap, 3DPW, G3D, and FNTU), which demonstrates the effectiveness of our proposed method. The code is available at https://github.com/lily2lab/TrajectoryCNN.git.**

***Index Terms*— Human motion prediction, spatio-temporal feature learning, CNN, skeleton.**

I. Introduction

HUMAN activity analysis has been an important topic in computer vision due to the undeniable significance in a number of applications, ranging from the biomechanics of human movement [1], video surveillance [2], [3], to human-machine interaction [4], [5], and service robotics [4], [5]. Among many problems in human activity analysis, how to predict future human motion based on the currently observed poses is of great importance to enable automated systems or robots to seamlessly interact with people [6], [7]. For instance, a service robot can provide immediate support to an elderly person to avoid danger if the system is able to predict the person is likely to fall. In this paper, as is shown in Fig. 1, we focus on the problem of human motion prediction which aims to predict the future poses based on the observed poses.

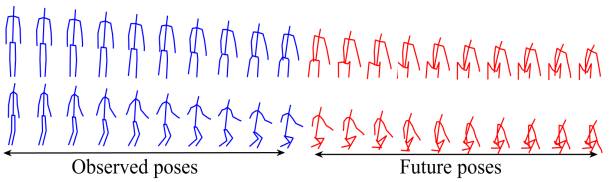


Fig. 1. Human motion prediction. The blue poses are the observed poses, and the red poses are the future poses.

Coupled spatio-temporal modeling plays a key role to predict future poses [8]–[10]. In general, previous spatio-temporal modeling used two major types of methods, RNN (Recurrent Neural Network) [11]–[15] and CNN (Convolutional Neural Network) [8]–[10], [16]. (1) **RNN models** [11], [12], [14] are especially powerful in processing short-term temporal information while having an inherent weakness in spatial modeling. For example, Martinez *et al.* [11] built their RNN model to predict future poses entirely based on GRU (Gated Recurrent Unit). This method mainly focused on capturing temporal information and ignored a part of the spatial structure of the human body. (2) **CNN models** [9], [10], [16] have been successfully used to predict human poses but can not capture the coupled spatio-temporal structural information well. In previous research, Xu *et al.* [10] and Liu *et al.* [9], [10] proposed CNN-based methods and processed the spatial and temporal information separately for predicting future poses. However, when a human pose changes across multiple frames, the spatial and temporal information is intrinsically coupled. Separately processing of the spatial and temporal information inevitably breaks the natural state of the information when representing the pose changes, and therefore is ineffective to predict complicated motions.

Another important aspect of human motion prediction is the global temporal co-occurrence modeling. Different frames have different contributions for predicting human motion.

Manuscript received February 20, 2020; revised June 23, 2020 and August 10, 2020; accepted August 27, 2020. Date of publication September 3, 2020; date of current version June 4, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61673192, in part by the Fundamental Research Funds for the Central Universities under Grant 2020XD-A04-1 and Grant 2019RC27, and in part by the BUPT Excellent Ph.D. Students Foundation under Grant CX2019111. This article was recommended by Associate Editor J. M. Martinez. *(Corresponding authors: Jianqin Yin; Jun Liu.)*

Xiaoli Liu, Jianqin Yin, and Pengxiang Ding are with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: liuxiaoli134@bupt.edu.cn; jqyin@bupt.edu.cn).

Jin Liu is with the School of Modern Post, Beijing University of Posts and Telecommunications, Beijing 100876, China.

Jun Liu is with the the Department of Mechanical Engineering, City University of Hong Kong, Hong Kong (email: Jun.Liu@cityu.edu.hk).

Huaping Liu is with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China.

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2020.3021409

For example, for a complex motion such as "hug", the frames that contain the basic motions "grab" and "touch" are important than other frames to characterize the complex motion. Therefore, to better mine motion dynamics for predicting complex motion, it is important to measure the importance of each pose in the motion sequence. In this paper, we name the different importance of all input poses as global temporal co-occurrence relationships. However, most of the existing works focused on modeling the global temporal information of the input poses and ignored the co-occurrence relationships among them [9], [10], [17]. They modeled the information of different temporal scales in different temporal levels and shared weight in each temporal level, making it difficult to learn free parameters for each pose to capture the global temporal co-occurrence relationships of the input poses. Butepage *et al.* [8] proposed a temporal Encoder to model diffident temporal scale information. Due to the local shared mechanism of convolution, different subsequences share the same weights, and thus it is difficult to learn free parameters for each pose to capture the global temporal co-occurrence relationships of the input poses. Some works focused on modeling the spatial co-occurrence information among joints of the human body [18], [19] but ignored the temporal co-occurrence modeling of the input poses. For example, Li *et al.* [18] and Zhu *et al.* [19] modeled the spatial co-occurrence relationships among joints to describe actions. However, this method lacked the modeling of global temporal co-occurrence information and therefore can not to predict pose changes in a long period.

To address the aforementioned limitations, we introduce a new trajectory space and model the motion dynamics of human motion in the new space. Moreover, we propose a new end-to-end spatio-temporal feature learning network, TrajectoryCNN, to achieve trajectory space transformation, capture motion dynamics, and predict future poses simultaneously. The proposed TrajectoryCNN can automatically transform the human motion sequence from the pose space to the trajectory space. In the trajectory space, the global temporal information and dynamic local-global correlations among joint trajectories of human motion can be easily captured since each element in the trajectory space encodes the global joint trajectory information of human movement. Different from prior works, we capture the motion dynamics by simultaneously encoding the coupled spatio-temporal features, dynamic local-global features, and global temporal co-occurrence information from a sequence of poses. Specifically, the coupled spatio-temporal features can be effectively captured with our encoder by covering multiple joint trajectories using filters in the trajectory space. Based on our proposed network incorporating our specially designed skeletal representation, the dynamic local-global features are extracted to model the strong correlations among joint trajectories of the same part, and the weak correlations among joint trajectories of different parts. Using our TrajectoryCNN and our skeletal representation, the global temporal co-occurrence features can be easily captured by learning free parameters for each pose. Finally, future poses can be predicted in a non-recursive manner using the captured motion dynamics.

The main contributions of this research are highlighted as follows:

- A new trajectory space is introduced and in the new space, we model the motion dynamics of the input sequence by simultaneously capturing the coupled spatio-temporal structure, dynamic local-global correlations, and the global temporal co-occurrence relationships. Compared with prior methods, we can model the different correlations among joint trajectories with the dynamic local-global correlations and measure the different importance of different input poses for mining motion dynamics with the global temporal co-occurrence relationships, which is important to achieve more accurate predictions.
- A new simple but effective end-to-end feedforward network, TrajectoryCNN, is proposed to simultaneously achieve trajectory space transformation, model motion dynamics, and predict future poses by combining our skeletal representation. The proposed network automatically transfers the human motion sequence to the trajectory space, so as to mine the motion dynamics in this space for predicting human motion.
- Experimental results show that our proposed method achieves state-of-the-art performance on five benchmark datasets, demonstrating the effectiveness of our method.

The remainder of this work is organized as follows: Section II summarises the related literature on human motion prediction. Section III describes the proposed TrajectoryCNN to predict future human motion. Section IV reports experimental results on five benchmark datasets both quantitatively and qualitatively. Section V briefly concludes our paper.

II. Related Work

In order to accurately predict the human motion, researchers have done significant investigations. We review these works from two aspects: the spatio-temporal modeling and the long-term temporal modeling of human motion.

A. The Spatio-Temporal Modeling

The spatio-temporal modeling is key to the sequence learning (including human motion prediction) [11], [20]–[22]. A typical methods of spatio-temporal modeling for predicting huamn motion sequence are built with RNNs [11], [17], [23]–[25]. Due to the inherent weakness of RNNs, these models can not capture the spatial features of the human body and long-term temporal dependencies well. For example, Chiu *et al.* [23] modeled the latent hierarchical structure of human motion by capturing the temporal dependencies with different temporal scales hierarchically using LSTM cells but did not capture the spatial structure of the human body well. Martinez *et al.* [11] proposed a residual architecture to model the velocities of the human motion sequence using GRUs. But the author only focused on short-term temporal modeling and ignored modeling the long-term temporal dependencies and spatial structure of the human body. To alleviate these limitations hidden in RNNs, some literature is proposed [13], [14], [26], [27]. Jain *et al.* [13] proposed a structural-RNN model to encode the high-level spatio-temporal structure of

the human motion sequence by combining LSTMs and fully connected (FC) layers. Guo *et al.* [27] modeled the local structure of the human body using FC layers and captured the long-term temporal dependencies with GRUs, but ignored capturing the interactions among different limbs.

Another type of spatio-temporal modeling for predicting human motion is based on feedforward networks [16], [28]–[31]. There are two schemes to model the spatio-temporal information: modeling the spatial and temporal information separately, modeling the spatial and temporal information equally. For example, Cho *et al.* [28] modeled the spatial and temporal information of previous frames separately. In detail, they first modeled the spatial information using VGG-16 [32], and then proposed T-CNN (Temporal Convolutional Neural Network) to model the temporal information of previous frames by using convolution across temporal dimension. Li *et al.* [16] modeled the spatial and temporal information using CNN. The authors modeled the local characteristic of the human body depending on a large convolutional kernel, which is not effective enough [32]–[34]. Moreover, this model focused on modeling the spatial correlations among joints of the human body in the pose space and can not capture the dynamic correlations among joints of the human body when the body changes over the whole temporal context. Therefore, their model was not effective to capture the motion dynamics for predicting human motion. Mao *et al.* [31] first modeled the temporal information of the human motion sequence using DCT (Discrete Cosine Transform) and then captured the spatial dependencies of joint trajectories using GCNs (Graph Convolutional Networks), and also achieved state-of-the-art performance. But their temporal modeling relied on manual features and their model was not built end to end, which was not flexible enough to capture the motion dynamics for predicting human motion.

B. The Long-Term Temporal Modeling of Human Motion

Most of the existing models for human motion prediction built with RNN (such as LSTM, GRU, etc) are inherently hard to capture the long-term temporal dependencies of previous frames using its recurrent unit [11], [12].

Other methods have been reported to model the long-term dynamics of previous poses hierarchically [9], [10], [17], [23]. Because of the shared weight mechanism in each temporal level, these models can not learn free parameters for each frame. Therefore, it is difficult to mine the temporal co-occurrence relationships of all input frames. For example, Xu *et al.* [10] and Liu *et al.* [9] captured the temporal information of adjacent frames using CMU (Cascade Multiplicative Unit) [10] and modeled the global temporal information using CMUs hierarchically. In each temporal level, the CMUs shared parameters. Therefore, it was difficult to capture the temporal co-occurrence information in a given full temporal context.

Moreover, other researchers modeled the global temporal information of all history poses [8], [29], [35]. Butepage *et al.* [8] captured the temporal information of all input poses using multiple FCs. Li *et al.* [35] modeled the long-term temporal information of previous poses by mapping the input sequence into deep features using autoencoder [36]. Butepage *et al.* [8] and Li *et al.* [35] treated the spatial and temporal information equally and ignored the difference among spatial and temporal dimensions so that their model can not capture the coupled spatio-temporal structural information and the global temporal co-occurrence relationships of the input poses well.

Recently, some researchers modeled temporal information using dilated convolutions [29]. For example, Pavllo *et al.* [29] proposed a QuaterNet to predict future human motion by modeling the long-term temporal information of the input poses hierarchically using dilated convolutions.

Although great success has been made in the long-term temporal modeling, little study has been done to analyze the global temporal co-occurrence information of the previous sequence. Therefore, these models discussed above did not measure the different contributions of different input poses for describing the complex human motion sequence. This motivates us to design a new model that enables the network to get a global response from all input poses to mine the global temporal co-occurrence features of the human motion sequence.

III. Methodology

A. Problem Formulation

Human motion sequence can be represented by a group trajectories of a set of 3D joints. Given an input human motion sequence $S = \{p_1, p_2, \cdots, p_{N_{t_i}}\}$ with length N_{t_i} and its corresponding future human motion sequence $\widehat{S} = \{\widehat{p}_1, \widehat{p}_2, \cdots, \widehat{p}_{N_{t_o}}\}$ with length N_{t_o}, where $p_{i_1} = \{J_{i_1 k_1}\}_{k_1=1}^{N_j}$ and $\widehat{p}_{i_2} = \{\widehat{J}_{i_2 k_2}\}_{k_2=1}^{N_j}$ are the i_1-th pose of S and i_2-th pose of $\widehat{S}$, respectively, $J_{i_1 k_1} = (x_{i_1 k_1}, y_{i_1 k_1}, z_{i_1 k_1})$ is the k_1-th joint of the i_1-th pose of S, $\widehat{J}_{i_2 k_2} = (x_{i_2 k_2}, y_{i_2 k_2}, z_{i_2 k_2})$ is the k_2-th joint of the i_2-th pose of $\widehat{S}$, N_j is the number of joints. Then human motion prediction can be formulated as a mapping $S \rightarrow \widehat{S}$ from the previous human motion sequence S to the future human motion sequence $\widehat{S}$.

Most of the existing methods were commonly proposed based on the Encoder-Decoder framework [11], [37], [38]. The encoder was usually used to encode the previous poses into an intermediate representation that represents the motion dynamics of the previous poses, and the decoder was used to restore the spatial and temporal information of the future poses. Following this framework, in this paper, we propose a new network, TrajectoryCNN, to model the motion dynamics of the input sequence and predict future motion sequence end to end. We model the motion dynamic law of human motion in a new trajectory space by encoding the coupled spatio-temporal information and global temporal co-occurrence features of previous poses and also modeling the different correlations among joint trajectories.

B. Definition

We introduce five concepts related to the prediction of human motion as follows.

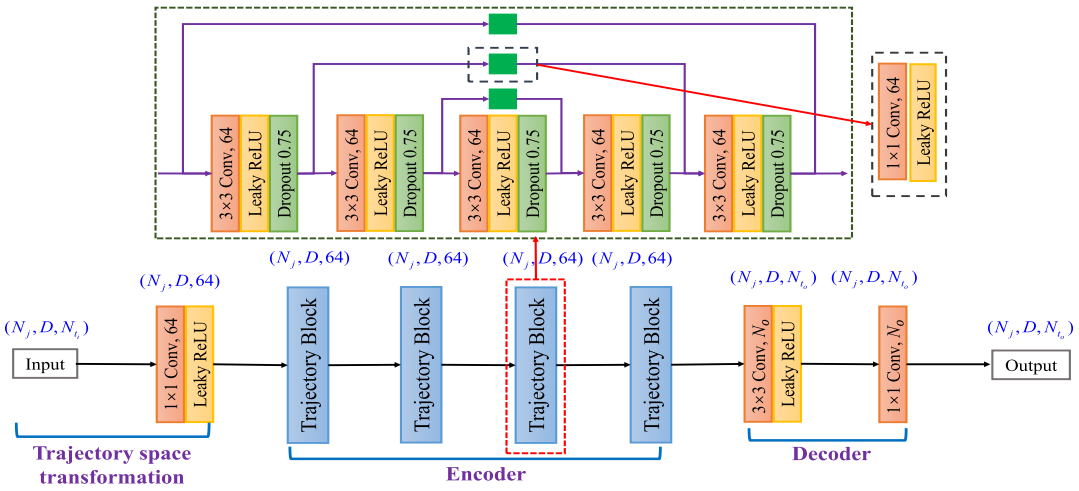


Fig. 2. Overall architecture of our proposed TrajectoryCNN. Convolutional layers are in orange, such as (3 × 3 Conv, 64), where "3 × 3 Conv" denotes convolutional operation with 3 × 3 filter, and 64 denotes the number of output channels. (N_j, D, N_{t_i}) denotes the shape of the output tensor.

- **Term** 1 **(Coupled spatio-temporal structure):** the evolution of human poses is an organic whole, including spatial and temporal dimensions. For representing human motion, the spatial information is closely related to the temporal information. Therefore, the human motion sequence has a coupled spatial-temporal structure. In essence, the prediction of human motion is a problem of coupled spatio-temporal modeling.
- **Term** 2 **(Dynamic local-global correlations):** human movement is constrained by the skeletal bone of the human body. The constraints among joints of one bone are stronger than that among joints of different bones. Moreover, in the process of human motion, the state of the human body will change over time. Therefore, for a human motion sequence, different joint trajectories have different correlations. The human body is usually divided into five parts according to the human anatomy, including two arms, two legs, and trunk [5]. Because joint trajectories are a set of joints evolving with time. In this paper, we name the correlations among joint trajectories of the same part as dynamic local correlations since joint trajectories of the same part cover a local area of the human motion sequence; we name the correlations among joint trajectories of different parts as dynamic global correlations since joint trajectories of different parts may cover the whole sequence of the human motion. Finally, we name the different correlations among different joint trajectories of human motion as dynamic local-global correlations.
- **Term** 3 **(Global temporal co-occurrence relationships):** in a complex human motion sequence, different poses have different contributions to mine motion dynamics of observed poses for predicting future poses. We name the relationships of these poses as temporal co-occurrence relationships. To better mine the motion dynamics, it is important to capture the relationships of these poses. In this paper, we model the temporal co-occurrence relationships of all input poses, and therefore we name them as global temporal co-occurrence relationships.
- **Term** 4 **(Pose space):** the human pose can be represented by the positions of a group of joints, therefore, the space that includes all poses is considered as the pose space.
- **Term** 5 **(Trajectory space):** in the trajectory space, each point represents the trajectory information of a special point evolving with time. Since a human motion sequence can be considered as a group of joint points evolving with time, the human motion sequence S in the trajectory space can be defined as equation 1.

$$(x'_{k_1}, y'_{k_1}, z'_{k_1}) = f(x_{i_1 k_1}, y_{i_1 k_1}, z_{i_1 k_1}) \tag{1}$$

where $i_1 = 1, 2, \cdots, N_{t_i}$, $k_1 = 1, 2, \cdots, N_j$, N_j is the number of joints and N_{t_i} is the length of the input sequence S. $(x'_{k_1}, y'_{k_1}, z'_{k_1})$ denotes the k_1-th point in the trajectory space, corresponding to the k_1-th joint trajectory from the sequence S in the pose space. $f(\cdot)$ denotes the trajectory space transformation from the pose space to the trajectory space, which can be built with one 1×1 convolutional layer and our skeletal representation shown in Fig. 4 and Section III-C.

C. Architecture of TrajectoryCNN

In this section, a new spatio-temporal convolutional network, TrajectoryCNN, is proposed to predict future poses as is shown in Fig. 2, which mainly consists of three parts: trajectory space transformation, Encoder, and Decoder.

Skeletal Representation: we first describe the skeletal representation used in our paper. As is shown in Fig. 3, given a human motion sequence $S = \{p_1, p_2, \cdots, p_{N_{t_i}}\}$ with length N_{t_i}, the poses in sequence S can be represented by a tensor $X = [p_1; p_2; \cdots; p_l; \cdots; p_{N_{t_i}}]$, where the joints are set as the width, the coordinates (e.g. x, y, and z) are set as the height, and the frames are set as the channel. Here, as is shown in equation 2, p_l is the l-th pose of sequence S. The shape of the input tensor is $N_j \times D \times N_{t_i}$, where N_j is the number of joints and D is the dimension of each joint. To conveniently capture the dynamic local-global features of joint trajectories, consistent with [9], as is shown in Fig. 3, the joints of

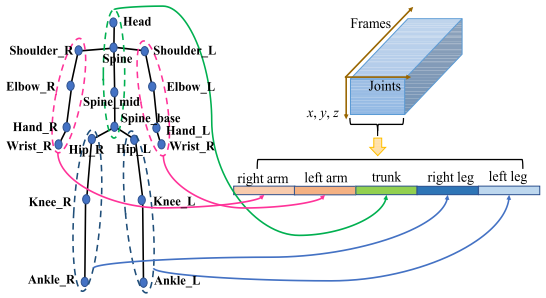


Fig. 3. Skeletal representation. The left one denotes the skeleton of the human body, and the right one denotes the skeletal representation of the input sequence. The joints of the same part are placed in the adjacent areas.

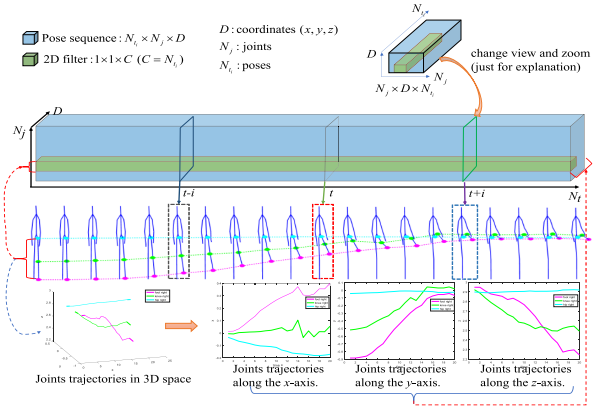


Fig. 4. Trajectory space transformation. Taking the joints of the left leg as an example, from top to bottom, we show a diagram of trajectory space transformation using 2D convolution, the corresponding human motion sequence, and the joint trajectories in 3D space and along each axis, where the filter covers one joint trajectory along one axis across all time-steps. Note: the channel of 2D filter (i.e. C) is usually ignored, and we show it explicitly to explain the modeling of joint trajectories.

the same part are placed in the connected positions. Finally, the organized order of different parts of the human body is right arm, left arm, trunk, right leg, and left leg. The detailed joint annotations of the skeletal representation refer to Fig. 11(d).

$$p_l = \begin{bmatrix} x_{l,1} & y_{l,1} & z_{l,1} \\ x_{l,2} & y_{l,2} & z_{l,2} \\ \vdots & \vdots & \vdots \\ x_{l,N_j} & y_{l,N_j} & z_{l,N_j} \end{bmatrix}, \quad l = 1, 2, \cdots, N_{t_i} \tag{2}$$

Trajectory Space Transformation: according to the definition of the trajectory space, each point in the trajectory space represents the trajectory information of a special point over time. Therefore, compared with the original pose space, the trajectory space contains richer trajectory information, which can easily model the global temporal information of the trajectory of the point. To better mine the motion dynamics of joint trajectories of human movement, the original motion sequence is converted into the trajectory space by equation 1. Meanwhile, two findings appear naturally: (1) due to the physical constraint of the human body, different joints have different trajectories; (2) as is shown at the bottom of Fig. 4, joint trajectories along different axes are different. Therefore, during trajectory space transformation, each joint trajectory and its trajectory along each axis can be treated separately.

In this paper, as is shown in Fig. 2, the trajectory space transformation $f(\cdot)$ in equation 1 is built with one 1×1 convolutional layer. As is shown in Fig. 4, when the temporal dimension is specified as the channel, the 1×1 filter only covers one joint trajectory along one axis, which can distinguish: (1) one joint trajectory from other joint trajectories; (2) one joint trajectory along one axis from other axes. With the specially designed skeletal representation and the 1×1 convolutional layer, we can get the global response from all input poses, and each element in the output feature map of this layer encodes the information of one joint trajectory along one axis. Therefore, the output feature maps of the 1×1 convolutional layer encode the information of joint trajectories of human movement, so that we can achieve trajectory space transformation.

Encoder: the encoder aims to mine the motion dynamics of the human motion sequence in the trajectory space. Therefore, the human motion sequence will be encoded into a latent representation T, which can be formulated as equation 3.

$$T = \phi(x'_j, y'_j, z'_j), \quad j = 1, 2, \cdots, N_j \tag{3}$$

where $\phi(\cdot)$ denotes the encoder.

As is shown in Fig. 2, our encoder mainly consists of four trajectory blocks. Each trajectory block mainly consists of five 2D convolutional layers, and each convolutional layer is followed by a Leaky ReLU and Dropout layer to improve the performance of the network and also avoid overfitting.

The motion dynamics can be captured from these perspectives:

1) Coupled Spatio-Temporal Features: We capture the coupled spatio-temporal features to model the coupled spatio-temporal structure of human motion. As is shown in Fig. 2, the width, the height, and the channel of the feature maps in the encoder represent the joints, coordinates, and trajectories of joints, respectively. In our encoder, the filter size is set to 3×3, covering multiple joint trajectories. Therefore, the coupled spatio-temporal features of the input sequence can be easily modeled in the encoder from the joints, coordinates, and trajectories.

2) Dynamic Local-Global Features: We capture the dynamic local-global features to model the dynamic local-global correlations among joint trajectories of human movement, the dynamic local-global features can be captured from two folds:

a) Trajectory block: As is shown in Fig. 2, a new backbone layer, trajectory block, is proposed to build our encoder, mainly including five convolutional layers. The filer size of these convolutional layers is set to 3×3, covering multiple joint trajectories of one part, which enables the network to model the strong correlations among local joint trajectories of human movement by combining our skeletal representation. The lower layer usually captures fine-grained features (including dynamic local features), while the deeper layer captures coarse-grained features (including dynamic global features). In the trajectory block, the deeper layer is connected with the lower layer. The residual connections in the trajectory block

have two advantages: (i) obtain point-level features. The 1×1 convolution in the residual connections enables the network to obtain the point-level features from the lower layer since the 1×1 filter covers one joint trajectory along one axis; (ii) enhance coarse-grained features. The residual connections provide a shortcut from the lower layer to the deeper layer, enabling the network to enhance the coarse-grained features by element-wisely adding the point-level features.

b) Stacking multiple trajectory blocks: Multiple trajectory blocks are stacked to enlarge the receptive filed to capture a larger spatial context such as dynamic global features. In this way, the dynamic local-global features can be captured from local to global. In this paper, the filter size is set to 3, and the stride is 1. Therefore, the receptive field of the first convolutional layer in the first trajectory block is 3 (cover 3 joints), and the receptive field of each layer will be increased by 2. The receptive field of the second layer in the third trajectory block is 25, which is large enough to cover all joints of the human body (in this paper, excluding the unchanged joints and repeated joints, 25 joints need to be predicted at most). Therefore, 3 trajectory blocks are needed to capture the global relationships among joint trajectories of human movement. Moreover, we empirically show that adding another trajectory block can further improve the final performance, and there is no need to stack more trajectory blocks because the performance is no longer improved.

Finally, the shape of the output tensor in the encoder is ($N_j \times D \times 64$), denoting the learned dynamic features of the input sequence. Specifically, each element of the output tensor denotes the spatio-temporal features of a joint trajectory along one axis, including the correlations among different dimensions, the correlations among joint trajectories of the same part, and the global features of all joints.

3) Global Temporal Co-Occurrence Features: We capture the global temporal co-occurrence features to describe the global temporal co-occurrence relationships of all input poses. The key to modeling global temporal co-occurrence relationships of the input sequence is to learn free parameters for each pose. As is shown in Fig. 4, the filters cover the whole temporal axis, and thus we can learn free parameters for each pose with these filters. Therefore, these filters encode the global temporal co-occurrence relationships of the input poses, and different filters encode different global temporal co-occurrence relationships. Therefore, during trajectory space transformation, the global temporal co-occurrence relationships of the input poses have been encoded with multiple group parameters. The following layers of the network (i.e. our Encoder) can automatically explore the global temporal co-occurrence relationships according to the specific input poses.

Decoder: the decoder aims to reconstruct the spatial and temporal information of the future poses from the latent representation T learned by the encoder, which can be formulated as equation 4.

$$\widehat{p}_i = \psi(T), \quad i = 1, 2, \cdots, N_{t_o} \tag{4}$$

where N_{t_o} is the number of future poses, $\widehat{p}_i$ is the i-th future pose, $\psi(\cdot)$ denotes the decoder.

As is shown in Fig. 2, our decoder mainly consists of two convolutional layers. As the discussion above, the correlations among joint trajectories of the same part are stronger than that of different parts. Therefore, one 3×3 convolutional layer is first applied to reconstruct the spatio-temporal information of future poses. In this way, the coordinates of future poses can be predicted with the dynamic features of local joints. Finally, one convolutional layer with a 1×1 convolutional filter that covers one joint trajectory along each axis is applied to further recover the spatial details of future poses, smooth the trajectory of future poses and also enhance the predictive performance.

IV. Experiments

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

A. Datasets and Implementation Details

Datasets: we use Human3.6M, CMU mocap dataset, 3D Pose in the Wild dataset, G3D and FNTU datasets to evaluate the performance of our model. (1) Human3.6M: Human3.6M (H3.6M) [39] is a commonly used dataset for human motion prediction. The dataset consists of 15 actions performed by 7 professional actors. (2) CMU mocap dataset (CMU-Mocap) [1]: CMU-Mocap dataset mainly includes five categories, naming "human interaction", "interaction with environment", "locomotion", "physical activities & sports" and "situations & scenarios". We adopt the same training and testing sets for evaluation. Finally, eight actions are selected for experiments, including running, walking, and so on. (3) 3D Pose in the Wild dataset (3DPW) [40]: 3DPW is a new dataset with accurate 3D poses in the wild. 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 use the official split sets for experiments. (4) G3D: G3D [41] is a gaming dataset collected with Microsoft Kinect. The dataset consists of 20 actions performed by 10 subjects in a controlled indoor environment. Each people performs several times and each sequence may contain multiple actions. In experiments, we process this dataset following the experimental setting in [9]. (5) FNTU: FNTU [9] is a dataset collected from NTU RGB+D [42], i.e. each sequence is clipped from the video in NTU RGB+D. The dataset consists of 18102 sequences. Among them, 12001 sequences for training, and the rest for testing. More details can be found in [9].

Implementation Details: following the setting and processing of [31], all experiments are carried out in 3D coordinate space. In experiments, all models are implemented by TensorFlow. The channels of convolutional layers in the trajectory space transformation and encoder are set to 64, and the

[1] http://mocap.cs.cmu.edu/

TABLE I
SHORT-TERM PREDICTION ON H3.6M

Milliseconds	Walking				Eating				Smoking				Discussion			
	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Residual sup [11]	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
convSeq2Seq [16]	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
LearnTrajDep [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**
TrajectoryCNN (Ours)	**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
Milliseconds	Directions				Greeting				Phoning				Posing			
	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Residual sup [11]	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
convSeq2Seq [16]	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
LearnTrajDep [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
TrajectoryCNN (Ours)	**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**
Milliseconds	Purchases				Sitting				Sitting Down				Taking Photo			
	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Residual sup [11]	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
convSeq2Seq [16]	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
LearnTrajDep [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
TrajectoryCNN (Ours)	**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**
Milliseconds	Waiting				Walking Dog				Walking Together				Average			
	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Residual sup [11]	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
convSeq2Seq [16]	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
LearnTrajDep [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
TrajectoryCNN (Ours)	**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**

channels of convolutional layers in decoder are set to 10 or 25 for short-term or long-term prediction, respectively (i.e. equal to the number of future poses). Since the coordinate value can be negative, positive, and zero, Leaky ReLU is selected as our activation function. Following [31], we repeat the last frame to align the future poses. To train our model, we use MPJPE (Mean Per Joints Position Error) proposed in [39] as our loss as is shown in equation 5. All models are trained with Adam optimizer, and the learning rate is initialized to 0.0001. We use MPJPE [39] in millimeter on H3.6M, CMU-Mocap and 3DPW as our metrics, and use MSE (Mean Squared Error) and MAE (Mean Absolute Error) [9] in meter for per sequence on G3D and FNTU.[2] All experimental settings are consistent with the baselines.

$$l = \frac{1}{N_j * N_o} \sum_{n=1}^{N_o} \sum_{j=1}^{N_j} \left\| \widehat{J}_{n,j} - J_{n,j} \right\|^2 \tag{5}$$

where N_j is the number of joints, N_o is the number of future poses, $J_{n,j}$ is the groundtruth one, and $\widehat{J}_{n,j}$ is the j-th predictive joint at the n-th time-step.

B. Baselines

We compare our method with several recent works for predicting human motion with 3D coordinate data, i.e Residual sup [11], convSeq2Seq [16], LearnTrajDep [31], PredCNN′ [9], [10] and PISEP^2 [9]. (1) Residual sup [11] is an RNN model for human motion prediction. (2) convSeq2Seq [16], PredCNN′ [9], [10] and PISEP^2 [9] are three feedforward models built with CNN for human motiom prediction. (3) LearnTrajDep [31] is built based on DCT and GCN, and is currently the state-of-the-art model for human motion prediction.

[2]Note the unit of the converted 3D coordinates data on H3.6M is millimeter, and the unit of the skeletal data on G3D and FNTU is meter.

C. Comparison With Baselines

Results on H3.6M: Table I reports the short-term prediction errors for 15 activities and their average performance. Specifically, compared with the results of the RNN model [11] and the CNN model [16], the errors of our method decrease significantly, which demonstrates the effectiveness of our proposed method. These possible reasons are two folds: (1) according to the definition of pose space and trajectory space in Section III-B, the trajectory space contains richer trajectory information, and in the trajectory space, we can conveniently capture the global temporal information and dynamic local-global correlations among joint trajectories of human motion. Moreover, the RNN model [11] and the CNN model [16] capture the motion dynamics in the pose space, while our method captures the motion dynamics in the trajectory space. (2) The RNN model [11] can not capture long-term temporal information and the correlations among joint trajectories well. The CNN model [16] ignored the global temporal co-occurrence modeling and failed to model the dynamic local-global correlations among joint trajectories. But our model captures motion dynamics of the human motion sequence by simultaneously learning the coupled spatio-temporal features, dynamic local-global features and global temporal co-occurrence features, considering the correlations among joint trajectories and the temporal cues among the input poses carefully. Therefore, our model can better capture the motion dynamic law. Compared with [31], our model achieves the lowest errors at all time-steps on average. Although LearnTrajDep [31] also captures motion dynamics in the trajectory space, compared with LearnTrajDep [31], our model has two major advantages. (1) Our model achieves trajectory space transformation, captures the motion dynamics, and predicts future poses end to end, while LearnTrajDep [31] captures the motion dynamics and predicts

TABLE II

LONG-TERM PREDICTION OVER 15 ACTIVITIES ON H3.6M. THE 3D ERRORS FOR 4 ACTIVITIES OF "RESIDUAL SUP" [11] AND "CONVSEQ2SEQ" [16] MODELS ARE PROVIDED IN [31], AND THE RESULTS OF "LEARNTRAJDEP" [31] FOR ALL ACTIVITIES IS REPRODUCED USING THE AVAILABLE PRE-TRAINED MODEL

Milliseconds	Walking		Eating		Smoking		Discussion		Directions		Greeting		Phoning		Posing	
	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000
Residual sup [11]	73.8	86.7	101.3	119.7	85.0	118.5	120.7	147.6	–	–	–	–	–	–	–	–
convSeq2Seq [16]	59.2	71.3	66.5	85.4	42.0	67.9	84.1	116.9	–	–	–	–	–	–	–	–
LearnTrajDep [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
TrajectoryCNN (Ours)	**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**

Milliseconds	Purchases		Sitting		Sitting Down		Taking Photo		Waiting		Walking Dog		Walking Together		Average	
	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000
Residual sup [11]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
convSeq2Seq [16]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
LearnTrajDep [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
TrajectoryCNN (Ours)	**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**

TABLE III

SHORT AND LONG-TERM PREDICTION ON CMU-MOCAP

Milliseconds	Basketball					Basketball Signal					Directing Traffic				
	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
LearnTrajDep [31]	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
TrajectoryCNN (Ours)	**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**

Milliseconds	Jumping					Running					Soccer				
	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
LearnTrajDep [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**
TrajectoryCNN (Ours)	**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

Milliseconds	Walking					Washwindow					Average				
	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
LearnTrajDep [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
TrajectoryCNN (Ours)	**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**

future poses separately and relies on manual features. Therefore, our model is more flexible in capturing motion dynamics and predicting future poses that lead to superior performance. (2) Our model can capture the global temporal co-occurrence features of the input poses by learning free parameters for each pose, while LearnTrajDep [31] can not since there is no parameters that can be learned for each pose using DCT. Based on the two advantages above, our model achieves the best performance, showing the effectiveness of our model again.

For long-term prediction, as is shown in Table II, compared with all the baselines, our model achieves the best performance on average for both 560ms and 1000ms, especially in 1000ms, which further verifies the effectiveness of our proposed method. The most important reason is that our model can simultaneously capture the global temporal co-occurrence features and coupled spatio-temporal features of previous poses, while other methods can not.

Results on CMU-Mocap: Table III reports the 3D errors for short-term and long-term prediction on CMU-Mocap. Taking the action "Directing Traffic" as an example, the errors of our method decrease significantly in all cases, especially in the case of 1000ms. In general, our method outperforms all baselines by a large margin for both short-term and long-term prediction, demonstrating the effectiveness of our proposed method powerfully.

Results on 3DPW: Table IV reports the 3D errors for short-term and long-term prediction on 3DPW. For a more difficult dataset such as the dataset in the wild (i.e. 3DPW), compared with all baselines, the errors of our method decrease at all time-steps by a larger margin for both short-term and long-term prediction, showing the effectiveness of our method powerfully. The main reasons are: (1) Residual sup [11] can not model the spatial correlations and long-term temporal features well using GRUs. (2) convSeq2Seq [16] can not capture the spatial features of the human body well using a large convolutional filter, and they also failed to capture the global temporal co-occurrence features since the temporal dimension is not specified as the channel of the input tensor and their convolutional model can not learn the free parameters for each pose. (3) LearnTrajDep [31] can not capture the global temporal co-occurrence features of the input poses since there are no parameters needed to be learned using DCT. Moreover, their model is not built end to end and thus is not flexible enough in capturing motion dynamics for predicting future poses. Differently, we achieve trajectory space transformation, capture motion dynamics, and predict future poses end to end. In addition, our method captures the motion dynamics with coupled spatio-temporal features, dynamic local-global features, and global temporal co-occurrence features simultaneously, considering the spatio-temporal features of the input sequence carefully. Therefore, we can achieve better results.

Results on G3D: the results on G3D are reported in Table V. Our method achieves state-of-the-art performance, demonstrating the effectiveness of our proposed method again. Compared with PredCNN′ [9], [10] and PISEP^2 [9], the errors of our method decrease significantly. The possible reasons are two

TABLE IV
SHORT AND LONG-TERM PREDICTION ON 3DPW

Milliseconds	200	400	600	800	1000
Residual sup [11]	113.9	173.1	191.9	201.1	210.7
convSeq2Seq [16]	71.6	124.9	155.4	174.7	187.5
LearnTrajDep [31]	35.6	67.8	90.6	106.9	117.8
TrajectoryCNN (Ours)	**30.0**	**59.7**	**85.3**	**99.0**	**107.7**

TABLE V
HUMAN MOTION PREDICTION ON G3D AND FNTU

Model	G3D		FNTU	
	MSE	MAE	MSE	MAE
PredCNN′ [10], [9]	0.1882	1.5713	0.1665	1.6394
PISEP^2 [9]	0.1199	1.1101	0.1210	1.1651
LearnTrajDep [31]	0.1013	0.9068	0.1696	1.2280
TrajectoryCNN (Ours)	**0.0937**	**0.8663**	**0.1055**	**1.0114**

folds: (1) our model considers the spatio-temporal structure carefully by capturing the coupled spatial-temporal features from both the spatial and temporal dimensions. PredCNN′ and PISEP^2 separately modeled the spatial and temporal information of previous poses, ignoring the correlations among the spatial and temporal dimensions. Therefore, these models break the natural state of human motion and thus can not capture the motion dynamics well. (2) We specify the temporal dimension as the channel of the input tensor in our CNN model. Therefore, we can easily learn free parameters for each pose to capture the global temporal co-occurrence relationships of all input poses. PredCNN′ [9], [10] and PISEP^2 [9] modeled multiple temporal scale information hierarchically and the poses in each temporal scale shared the same weights, making it difficult to learn free parameters for each pose and thus can not capture the global temporal co-occurrence features of all input poses. Compared with [31], our method achieves the best results at both MSE and MAE, showing the effectiveness of our method in modeling motion dynamics and predicting human motion end to end.

Results on FNTU: the results on FNTU are reported in Table V. Similarly, our conclusion remains unchanged. Our method outperforms all baselines by a large margin, showing the effectiveness again.

Discussion: as is shown in Table I∼Table V, compared to the results on H3.6M and CMU-Mocap, the performance gaps between the baseline [31] and our method on 3DPW, G3D, and FNTU are enlarged. The reasons can be analyzed for these two perspectives. (1) **From the perspective of the dataset,** the prediction of human motion on 3DPW, G3D, and FNTU is more difficult than that on H3.6M, and CMU-Mocap. The reasons are as follows: the used datasets in this paper can be roughly divided into two categories: body-centered coordinate data (i.e. H3.6M, CMU-Mocap, and 3DPW) and camera-centered coordinate data (i.e. G3D and FNTU). *a*) Body-centered coordinate data (i.e. H3.6M, CMU-Mocap, and 3DPW): for these datasets, the center of the 3D coordinate system is located in the center of the human body, and the 3D coordinate data excludes the global rotations and translations of the humans. Moreover, 3DPW is a newly released dataset collected in challenging outdoor scenes with a moving camera, including walking in the city, going up-stairs, having coffee or taking the bus. Therefore, the prediction of human motion on 3DPW is more challenging than that on H3.6M and CMU-Mocap. *b*) Camera-centered coordinate data (i.e. G3D and FNTU): for these datasets, the center of the 3D coordinate system is located in the center of the camera, and the coordinate data contains the global rotations and translations of the humans. Moreover, due to the various distances or views of the placement of the camera, the problem of predicting human motion is very challenging. Because the coordinate value of distant joints is greater than that of near joints. When the center of the 3D coordinate system is located on the camera, it will lead to various physical structural characteristics of the human body even for the same people. Therefore, the predictive task on these datasets is very challenging. (2) **From the perspective of the method,** our method can better capture motion dynamics to describing the complex human motion. Compared with the baseline [31], our method achieves improved performance especially on the more challenging datasets such as 3DPW (the result is shown in Table IV), G3D and FNTU (the results are shown in Table V), which benefits from the two advantages of our method: *i*) our model is built end to end, and thus is flexible in capturing motion dynamics and predicting future poses, while [31] is not built end to end and relies on manual features (i.e. Using DCT to capture the temporal dependencies of the input poses); *ii*) our model can capture the global temporal co-occurrence relationships of the input poses by learning free parameters for each pose, while the baseline [31] can not since there is no parameters can be learned for each pose using DCT.

D. Ablation Analysis

In this section, we verify our network from coupled spatio-temporal modeling, global temporal co-occurrence modeling, and dynamic local-global modeling.

Evaluation of Coupled Spatio-Temporal Modeling (C-ST): we prevent the network capturing the coupled spatio-temporal features by modifying the filter size of the whole network to verify the importance of this. We consider these two experiments: (1) the filter size is set to 1×1: in this case, the filter covers one joint trajectory along one axis (w/o C-ST(#1)); (2) the filter size is set to 1×3: in this case, the filter covers one joint trajectory (w/o C-ST(#2)). Although there is weight shared mechanism in the convolutional neural network, it is hard to capture the strong correlations among joint trajectories of the same limb. In a word, the network mainly focuses on modeling the temporal information, while ignores the spatial modeling.

As is shown in Table VI, compared with the results of "w/o C-ST(#1)" and "w/o C-ST(#2)", the errors of our proposed "TrajectoryCNN" decrease significantly on five datasets, proving that simultaneously modeling the spatial and temporal features is critical for the network performance. For example, "TrajectoryCNN" on H3.6M achieves lower errors at all timestamps, especially for the later predictions. The possible reason is: at the early timestamps, the spatial features of future

TABLE VI

EVALUATION OF COUPLED SPATIO-TEMPORAL MODELING. HERE, W/O C-ST(#1) DENOTES THE FIRST EXPERIMENT,AND W/O C-ST(#2) DENOTES THE SECOND EXPERIMENT. THE ERRORS ON H3.6M, CMU-MOCAP AND 3DPW DATASETS ARE AVERAGED OVER ALL ACTIVITIES

Model	MPJPE on H3.6M			
	80ms	160ms	320ms	400ms
w/o C-ST(#1)	12.4	31.0	67.6	81.6
w/o C-ST(#2)	11.8	27.5	58.1	69.4
TrajectoryCNN	**10.2**	**23.2**	**49.3**	**59.7**
Model	MPJPE on CMU-Mocap			
	80ms	160ms	320ms	400ms
w/o C-ST(#1)	9.7	20.9	49.0	65.1
w/o C-ST(#2)	9.4	18.4	39.0	50.5
TrajectoryCNN	**8.3**	**15.6**	**33.4**	**43.1**
Model	MPJPE on 3DPW			
	200ms		400ms	
w/o C-ST(#1)	36.8		74.5	
w/o C-ST(#2)	34.2		69.0	
TrajectoryCNN	**30.0**		**59.7**	
Model	G3D		FNTU	
	MSE	MAE	MSE	MAE
w/o C-ST(#1)	0.1542	1.0681	0.1566	1.2057
w/o C-ST(#2)	0.1300	1.0032	0.1488	1.2052
TrajectoryCNN	**0.0937**	**0.8663**	**0.1055**	**1.0114**

poses are similar to the later observed poses, but at the later timestamps, the spatial features of future poses may vary greatly. Therefore, ignoring spatial modeling may have much more effect on the later predictions. "w/o C-ST(#1)" and "w/o C-ST(#2)" can not model spatial information well, which may lead to the worse performance at the later timestamps.

Evaluation of Global Temporal Co-Occurrence Modeling (GTC): to show the effectiveness of global temporal co-occurrence information, we reorganize the input tensor (frames are set as width, joints are set as height, and coordinates (e.g. x, y and z) are set as depth) as done by prior literature [43]–[45]. In this case, the filter covers local spatial and local temporal information. Because of the local weight shared mechanism of convolutional networks, it is difficult to learn free parameters for each pose. In this case, the network can not model the global temporal co-occurrence relationships of all input poses.

Experimental results are reported in Table VII. Compared with "w/o GTC", in general, the errors of "TrajectoryCNN" decrease on three datasets, especially on the more challenging datasets such as H3.6M and 3DPW datasets, showing the effectiveness of global temporal co-occurrence modeling. The possible reason is: for a more challenging dataset, human motion is more complex, modeling the global temporal co-occurrence relationships of the input poses can better capture motion dynamics for predicting human motion. As is shown in Table VII, although the overall improvement on CMU-Mocap dataset is limited, the improvement for the short-term prediction can not be ignored. This further shows the effectiveness of global temporal co-occurrence modeling. To sum up, we can conclude that the global temporal co-occurrence modeling can improve the final performance of the network, especially on a more challenging dataset.

TABLE VII

EVALUATION OF GLOBAL TEMPORAL CO-OCCURRENCE MODELING. HERE, "W/O GTC" DENOTES "WITHOUT GLOBAL TEMPORAL CO-OCCURRENCE MODELING". THE AVERAGES ERRORS ON THREE DATASETS ARE AVERAGED OVER ALL ACTIVITIES

Model	MPJPE on H3.6M				
	80ms	160ms	320ms	400ms	average
w/o GTC	12.0	25.5	51.1	60.7	37.3
TrajectoryCNN	**10.2**	**23.2**	**49.3**	**59.7**	**35.6**
Model	MPJPE on CMU-Mocap				
	80ms	160ms	320ms	400ms	average
w/o GTC	10.6	16.8	**32.8**	**41.8**	25.5
TrajectoryCNN	**8.3**	**15.6**	33.4	43.1	**25.1**
Model	MPJPE on 3DPW				
	200ms		400ms		average
w/o GTC	34.0		63.9		49.0
TrajectoryCNN	**30.0**		**59.7**		**44.9**

Fig. 5. Remove residual connections in the trajectory block.

TABLE VIII

EVALUATION OF DYNAMIC LOCAL-GLOBAL MODELING. HERE, F_i DENOTES THE i-TH PREDICTIVE POSE, "W/O DLG" DENOTES "WITHOUT DYNAMIC LOCAL-GLOBAL MODELING". THE ERRORS ON H3.6M, CMU-MOCAP AND 3DPW DATASETS ARE AVERAGED OVER ALL ACTIVITIES

Model	MPJPE on H3.6M			
	80ms	160ms	320ms	400ms
w/o DLG	35.7	37.2	65.8	79.8
TrajectoryCNN	**10.2**	**23.2**	**49.3**	**59.7**
Model	MPJPE on CMU-Mocap			
	80ms	160ms	320ms	400ms
w/o DLG	44.3	36.1	54.9	69.8
TrajectoryCNN	**8.3**	**15.6**	**33.4**	**43.1**
Model	MPJPE on 3DPW			
	200ms		400ms	
w/o DLG	54.3		78.6	
TrajectoryCNN	**30.0**		**59.7**	

Model	MSE on FNTU				
	F_2	F_4	F_8	F_{10}	average
w/o DLG	0.1671	0.1412	0.3373	0.3636	0.2523
TrajectoryCNN	**0.0365**	**0.0673**	**0.1524**	**0.2101**	**0.1166**
Model	MAE on FNTU				
	F_2	F_4	F_8	F_{10}	average
w/o DLG	1.5053	1.6507	2.2590	2.4883	1.9758
TrajectoryCNN	**0.5350**	**0.8160**	**1.3389**	**1.5943**	**1.0711**

Evaluation of Dynamic Local-Global Modeling (DLG): to verify the effectiveness of dynamic local-global modeling, as is shown in Fig. 5, we remove all residual connections in the trajectory block. Without the residual connections in the trajectory block, fine-grained features (including dynamic local features) from the lower layers can not be fused with the coarse-grained features (including dynamic global features) from the higher layers. In this case, the network is difficult to capture the dynamic local-global features well. Experimental results are shown in Table VIII. Compared with "w/o DLG", the errors of "TrajectoryCNN" significantly decrease on all

TABLE IX

EVALUATION OF GENERALIZATION. IN THE EXPERIMENT, WE TRAIN ON FNTU DATASET AND DIRECTLY TEST ON G3D DATASET, WHICH IS CONSISTENT WITH THE BASELINES [9], [10], [31]

Methods	MSE	MAE
PredCNN′ [10], [9]	0.2432	2.1706
PISEP^2 [9]	0.1446	1.2713
LearnTrajDep [31]	0.1353	1.0409
TrajectoryCNN (Ours)	**0.1179**	**0.9353**

datasets, which demonstrates the effectiveness of dynamic local-global modeling.

E. Evaluation of Generalization

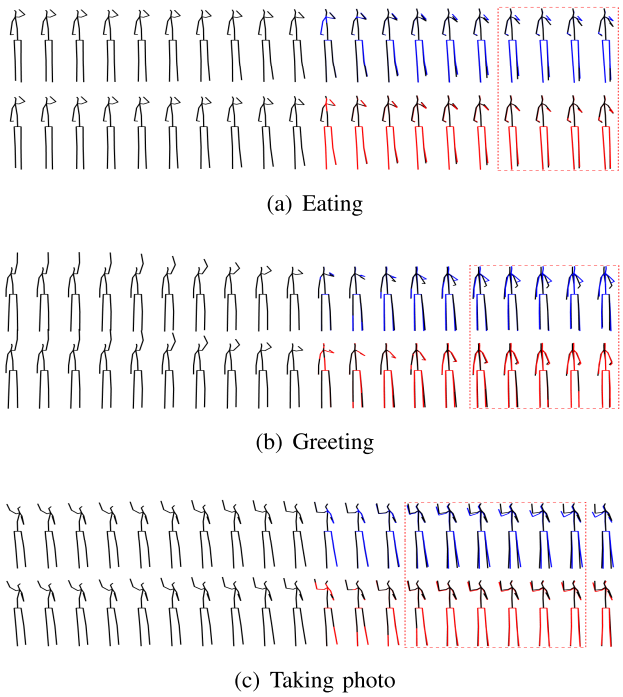
Considering the differences of these datasets, we decide to verify the generalization performance of the proposed network on G3D and FNTU. For this, we pre-train our model on FNTU and directly test on G3D. As is shown in Table IX, our method achieves the best performance on unseen data. Specifically, compared with [31], the MSE and MAE of our network decrease by 0.0174 and 0.1056, respectively. The possible reason is that our model considers modeling motion dynamics and predicting future poses end to end, which may lead to better generalization of our proposed model to novel actions.

F. Qualitative Analysis

To show the qualitative performance of our proposed network, we also visualize the predictive results frame-by-frame on H3.6M, G3D, and FNTU.[3] Fig. 6 and Fig. 7 show the predictive results for short-term and long-term prediction on H3.6M, respectively. Our method achieves the best visualized performance for both short-term and long-term predictions on H3.6M. As is denoted in Fig. 6, for the left hand of the poses in both Fig. 6(a) and Fig. 6(b), and the right hand of the poses in Fig. 6(c), the performance of our method is better than the baseline. As is denoted in Fig. 7, the right hand of predicted poses of our proposed method is closer to the groundtruth. This may benefit from two folds: (1) we model the motion dynamics of the input sequence by simultaneously extracting the dynamic local-global features, the coupled spatio-temporal features, and the global temporal co-occurrence features; (2) our proposed network can achieve trajectory space transformation, capture motion dynamics, and predict human motion end to end, while the baseline can not. Therefore, our method is more flexible in capturing motion dynamics, and thus we can achieve better results.

Fig. 8 and Fig. 9 show the visualization performance on G3D and FNTU, respectively, and our method achieves the best performance on both datasets. As is denoted in Fig. 8, compared with [31], two hands in Fig. 8(a) and the left leg in Fig. 8(b) are the closest to the groundtruth. As is shown in Fig. 9, for the right hand in Fig. 9(a) and the upper body in Fig. 9(b), the results of our method are significantly better than [31], which demonstrates the effectiveness of our proposed method again.

[3]Because there are not available pre-trained models on CMU-Mocap and 3DPW datasets, we only provide the qualitative results on H3.6M, G3D, and FNTU.



(a) Eating

(b) Greeting

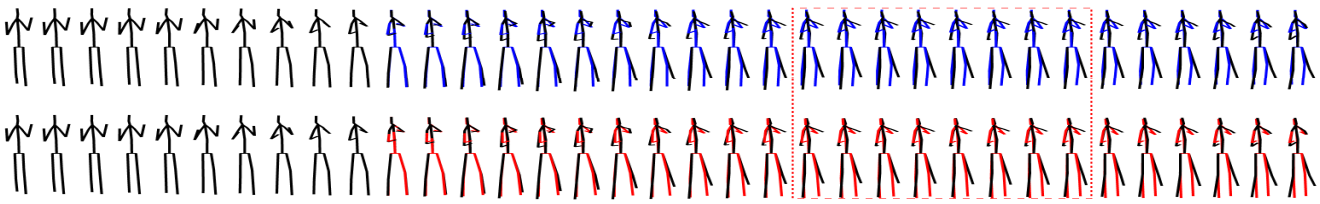
(c) Taking photo

Fig. 6. Visualization of frame-wise performance for short-term prediction on H3.6M. For each group poses, from top to bottom, we show the results of LearnTrajDep [31] and the results of our proposed method, where the black poses denote the groundtruth, the blue poses and the red poses denote the predictive poses.

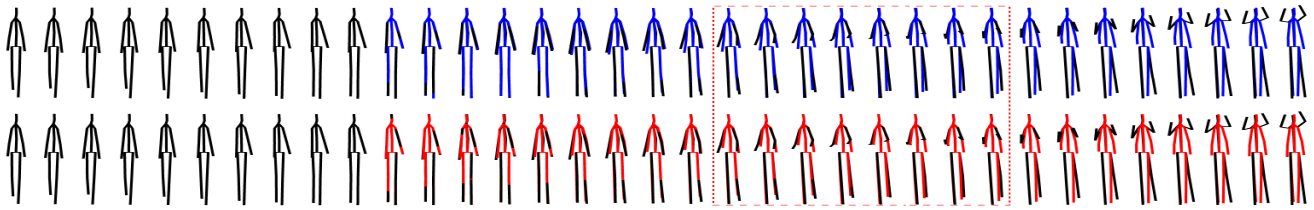
More visualization results are reported in Fig. 10 and Fig. 11 to show the advantages of our proposed model in capturing motion dynamics.

1) Global Temporal Co-Occurrence Modeling: as is discussed in Section III-C, the global temporal co-occurrence relationships are encoded during the trajectory space transformation, and therefore we visualize the convolutional kernels of this convolutional layer to intuitively show the modeling, and the results are shown in Fig. 10. Here, the horizontal axis denotes the frames across the temporal dimension, and the vertical axis denotes the weights of convolutional kernels. Taking "kernel-1" as an example, this kernel covers the whole temporal axis of the input sequence and each frame has its free weights. Therefore, with the free weights for each frame, we can conveniently model the different importance of different frames for capturing the motion dynamics to better predict the future poses. To sum up, the kernel encodes the global temporal co-occurrence relationships of the input frames. Since different kernels pay different attention to different frames, we think that different convolutional kernels encode different global temporal co-occurrence relationships. As is shown in Fig. 2, under the modeling of the following layers, the network can automatically learn different global temporal co-occurrence relationships for different input sequences.

2) Dynamic Local-Global Modeling With Residual Connections: we visualize the intermediate features maps of each trajectory block in our encoder to show how the residual

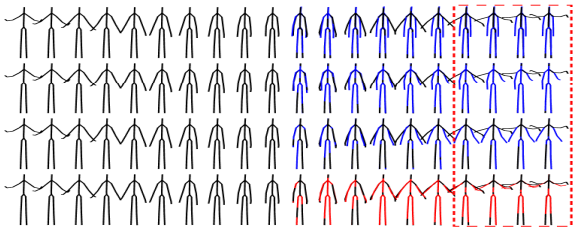


(a) Posing

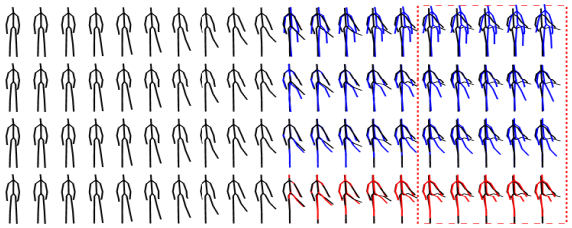


(b) Waiting

Fig. 7. Visualization of frame-wise performance for long-term prediction on H3.6M. For each group poses, from top to bottom, we show the results of LearnTrajDep [31] and the results of our proposed method, where the black poses denote the groundtruth, the blue poses and the red poses denote the predictive poses.

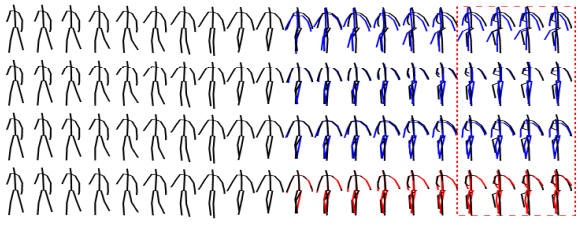


(a)



(b)

Fig. 8. Visualization of frame-wise performance on G3D. For each group poses, from top to bottom, we show the results of PredCNN [9], [10], the results of PISEP^2 [9], the results of LearnTrajDep [31] and the results of our proposed method, where the black poses denote the groundtruth, the blue poses and the red poses denote the predictive poses.



(a)

(b)

Fig. 9. Visualization of frame-wise performance on FNTU. For each group poses, from top to bottom, we show the results of PredCNN [9], [10], the results of PISEP^2 [9], the results of LearnTrajDep [31] and the results of our proposed method, where the black poses denote the groundtruth, the blue poses and the red poses denote the predictive poses.

connections help enhance the coarse-grained features with the point-level features from lower layers, and the results are shown in Fig. 11. Taking the motion sequence in Fig. 11(a) as an example, the movement of the human body mainly occurs in the joints of the right hand, left leg, and right leg. Comparing the visualization results of each trajectory block between Fig. 11(b) and Fig. 11(c), we can see that the response of the moving joints in Fig. 11(c) is larger than that in Fig. 11(b). The possible reason is that: the residual connections in the trajectory block can help enhance the coarse-grained features with the fine-grained features, enabling the network to capture the moving joints of human movement.

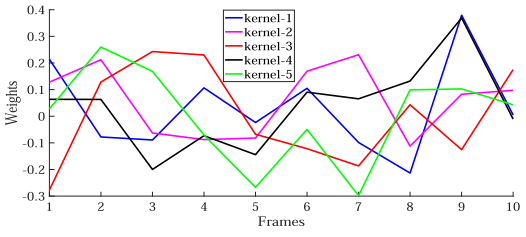
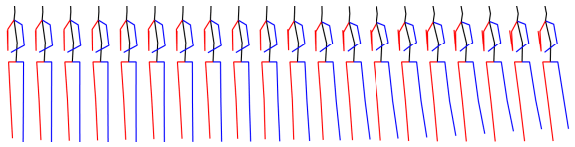
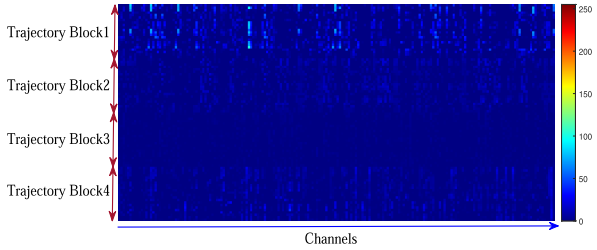


Fig. 10. Visualization of convolutional kernels on H3.6M. Different colors denotes different kernels.

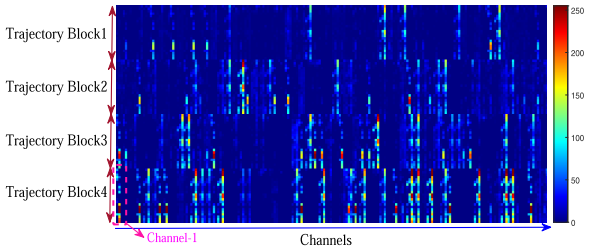
Therefore, we can better capture the dynamic local-global features of human motion using the residual connections in the trajectory block.



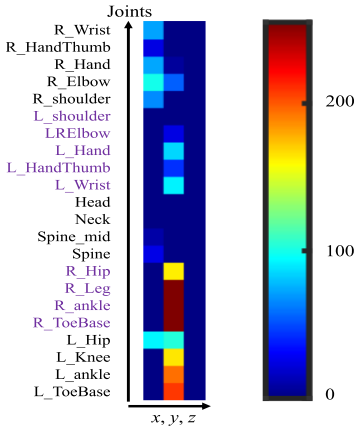
(a) Human motion sequence (smoking)



(b) Without residual connections in the trajectory block



(c) With residual connections in the trajectory block



(d) Joint anotations of feature maps

Fig. 11. Visualization of feature maps on H3.6M. Fig. 11(a) shows the human motion sequence corresponding to Fig. 11(b) and Fig. 11(c). For Fig. 11(b) and Fig. 11(c), the height is concatenated by the output feature maps of trajectory blocks, and the width is concatenated by different channels of output feature maps. From top to bottom of each trajectory block, as is shown in Fig. 3, we show the joints of the right arm, left arm, trunk, right leg, and left leg. Taking the feature map denoted in Fig. 11(c) at channel-1 of the trajectory block4 as an example, we show the joint annotations of feature maps in Fig. 11(d).

V. Conclusion

In this work, we propose an effective end-to-end spatio-temporal feature learning network, TrajectoryCNN, to capture the motion dynamics of the previous human motion sequence in the trajectory space and predict future human motion sequence in a non-recursive manner. A major difference between our method and other existing methods is that our model mainly captures the motion dynamics of input pose sequence in the trajectory space while other methods model their motion dynamics in the pose space or in the frequency domain. More importantly, our new proposed network can simultaneously capture the coupled spatio-temporal information and global temporal co-occurrence dependencies of previous poses, moreover, it can model the different correlations among joint trajectories of different parts. Evaluations are carried out on five benchmark datasets, and our method achieves state-of-the-art performance on all datasets. Experiments also show that simultaneously modeling the spatial and temporal information is critical to the final performance of the network, and the global temporal co-occurrence modeling can further improve the performance of the final network, especially on a more challenging dataset.

Acknowledgment

The research in this paper used the NTU RGB+D Action Recognition Dataset made available by the ROSE Lab at the Nanyang Technological University, Singapore.

References

[1] J. Hamill and K. M. Knutzen, *Biomechanical Basis of Human Movements*. Philadelphia, PA, USA: Lippincott Williams & Wilkins, 2006.
[2] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proc. CVPR*, 2017, pp. 6299–6308.
[3] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. NeurIPS*, 2014, pp. 568–576.
[4] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," 2018, *arXiv:1806.11230*. [Online]. Available: http://arxiv.org/abs/1806.11230
[5] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. CVPR*, 2015, pp. 1110–1118.
[6] K. Soomro, H. Idrees, and M. Shah, "Online localization and prediction of actions and interactions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 459–472, Feb. 2019.
[7] J. Liu, A. Shahroudy, G. Wang, L.-Y. Duan, and A. C. Kot, "Skeleton-based online action prediction using scale selection network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 6, pp. 1453–1467, Jun. 2020.
[8] J. Butepage, M. J. Black, D. Kragic, and H. Kjellstrom, "Deep representation learning for human motion prediction and classification," in *Proc. CVPR*, 2017, pp. 6158–6166.
[9] X. Liu, J. Yin, H. Liu, and Y. Yin, "PISEP^2: Pseudo image sequence evolution based 3D pose prediction," 2019, *arXiv:1909.01818*. [Online]. Available: http://arxiv.org/abs/1909.01818
[10] Z. Xu, Y. Wang, M. Long, J. Wang, and M. KLiss, "PredCNN: Predictive learning with cascade convolutions," in *Proc. IJCAI*, 2018, pp. 2940–2947.
[11] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *Proc. CVPR*, 2017, pp. 2891–2900.
[12] L.-Y. Gui, Y.-X. Wang, D. Ramanan, and J. M. Moura, "Few-shot human motion prediction via meta-learning," in *Proc. ECCV*, 2018, pp. 432–450.
[13] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-RNN: Deep learning on spatio-temporal graphs," in *Proc. CVPR*, 2016, pp. 5308–5317.
[14] A. Gopalakrishnan, A. Mali, D. Kifer, L. Giles, and A. G. Ororbia, "A neural temporal model for human motion prediction," in *Proc. CVPR*, 2019, pp. 12116–12125.
[15] H. Wang and J. Feng, "VRED: A position-velocity recurrent encoder-decoder for human motion prediction," 2019, *arXiv:1906.06514*. [Online]. Available: http://arxiv.org/abs/1906.06514
[16] C. Li, Z. Zhang, W. Sun Lee, and G. Hee Lee, "Convolutional sequence to sequence model for human dynamics," in *Proc. CVPR*, 2018, pp. 5226–5234.
[17] Z. Liu *et al.*, "Towards natural and accurate future motion prediction of humans and animals," in *Proc. CVPR*, 2019, pp. 10004–10012.
[18] 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 *Proc. IJCAI*, 2018, pp. 786–792.
[19] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," in *Proc. AAAI*, 2016, pp. 3691–3697.

[20] D. Guo, W. Zhou, H. Li, and M. Wang, "Online early-late fusion based on adaptive HMM for sign language recognition," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 1, pp. 1–18, Jan. 2018.

[21] D. Guo, S. Tang, and M. Wang, "Connectionist temporal modeling of video and language: A joint model for translation and sign labeling," in *Proc. IJCAI*, 2019, pp. 751–757.

[22] V. Vukotić, S.-L. Pintea, C. Raymond, G. Gravier, and J. C. Van Gemert, "One-step time-dependent future video frame prediction with a convolutional encoder-decoder neural network," in *Proc. ICIAP*. Cham, Switzerland: Springer, 2017, pp. 140–151.

[23] H.-K. Chiu, E. Adeli, B. Wang, D.-A. Huang, and J. C. Niebles, "Action-agnostic human pose forecasting," in *Proc. WACV*, 2019, pp. 1423–1432.

[24] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *Proc. ICCV*, s2015, pp. 4346–4354.

[25] J. N. Kundu, M. Gor, and R. V. Babu, "BIHMP-GAN: Bidirectional 3D human motion prediction GAN," in *Proc. AAAI*, vol. 33, 2019, pp. 8553–8560.

[26] L.-Y. Gui, Y.-X. Wang, X. Liang, and J. M. Moura, "Adversarial geometry-aware human motion prediction," in *Proc. ECCV*, 2018, pp. 786–803.

[27] X. Guo and J. Choi, "Human motion prediction via learning local structure representations and temporal dependencies," in *Proc. AAAI*, vol. 33, 2019, pp. 2580–2587.

[28] S. Cho and H. Foroosh, "A temporal sequence learning for action recognition and prediction," in *Proc. WACV*, 2018, pp. 352–361.

[29] D. Pavllo, C. Feichtenhofer, M. Auli, and D. Grangier, "Modeling human motion with quaternion-based neural networks," *Int. J. Comput. Vis.*, pp. 1–18, Oct. 2019.

[30] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. CVPR*, 2018, pp. 6450–6459.

[31] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning trajectory dependencies for human motion prediction," in *Proc. ICCV*, 2019, pp. 9489–9497.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[33] A. Deshpande. (2016). *A Beginner's Guide to Understanding Convolutional Neural Networks*. [Online]. Available: https://adeshpande3.github.io/A-Beginner%27s-Guide-To-Understanding-Convolutional-Neural-Networks/

[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. CVPR*, 2016, pp. 2818–2826.

[35] Y. Li *et al.*, "Efficient convolutional hierarchical autoencoder for human motion prediction," *Vis. Comput.*, vol. 35, nos. 6–8, pp. 1143–1156, Jun. 2019.

[36] P. K. P. Kingma and W. Max, "Auto-encoding variational bayes," 2013, *arXiv:1312.6114*. [Online]. Available: https://arxiv.org/abs/1312.6114

[37] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. Metaxas, "Learning to forecast and refine residual motion for image-to-video generation," in *Proc. ECCV*, 2018, pp. 387–403.

[38] X. Peng, R. S. Feris, X. Wang, and D. N. Metaxas, "A recurrent encoder-decoder network for sequential face alignment," in *Proc. ECCV*, 2016, pp. 38–56.

[39] 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, Jul. 2014.

[40] T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3D human pose in the wild using IMUs and a moving camera," in *Proc. ECCV*, 2018, pp. 601–617.

[41] V. Bloom, D. Makris, and V. Argyriou, "G3D: A gaming action dataset and real time action recognition evaluation framework," in *Proc. CVPRW*, 2012, pp. 7–12.

[42] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. CVPR*, 2016, pp. 1010–1019.

[43] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *Proc. ACPR*, 2015, pp. 579–583.

[44] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *Proc. CVPR*, 2017, pp. 3288–3297.

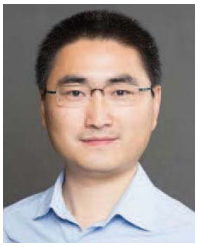
[45] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in *Proc. ICMEW*, 2017, pp. 597–600.

Xiaoli Liu received the bachelor's and M.Eng. degrees from the University of Jinan, Jinan, China, in 2015 and 2018, respectively. She is currently pursuing the Ph.D. degree with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China. Her research interests include computer vision, machine learning and image processing, and deep learning.

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.

Jin Liu received the B.S. degree in information engineering from Ludong University, Yantai, China, in 2019. He is currently pursuing the M.S. degree in mechanical engineering with the School of Modern Post, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include robotics and computer vision.

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.



Jun Liu (Member, IEEE) received the Ph.D. degree from the University of Toronto, in 2016. From 2017 to 2019, he worked as a Postdoctoral Fellow with the Weill Cornell Medical College, Cornell University. He is currently an Assistant Professor with the Department of Mechanical Engineering, City University of Hong Kong. His research interests include micro-nano robotics, medical robotics, and medical image analysis. His research has been recognized in the field of robotics and automation by winning multiple awards including the Best Student Paper Award and the Best Medical Robotics Paper Finalist Award from the IEEE International Conference on Robotics and Automation, in 2014, and the IEEE Transactions on Automation Science and Engineering Best New Application Paper Award, in 2018.

Huaping Liu (Senior Member, IEEE) received the Ph.D. degree from Tsinghua University, Beijing, China, in 2004. He is currently an Associate Professor with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His research interests include robot perception and learning. He serves as an Associate Editor of several journals including the IEEE ROBOTICS AND AUTOMATION LETTERS, *Neurocomputing, Cognitive Computation*, and some conferences including the *International Conference on Robotics and Automation* and the *International Conference on Intelligent Robots and Systems*. He also served as a Program Committee Member of RSS2016 and IJCAI2016.