Multi-Granularity Interaction for Multi-Person 3D Motion Prediction

Chenchen Liu and Yadong Mu

***Abstract*—Multi-person 3D motion prediction is an emerging task that involves predicting the future 3D motion of multiple individuals based on current observations. In contrast to motion prediction for a single person, this task requires a strong emphasis on learning the interacting dynamics among multiple individuals. Broadly speaking, current methods can be categorized into two groups: The first group involves the straightforward adaptation of models originally developed for single-person scenarios to multi-person scenarios, which is evidently suboptimal. The second group focuses on utilizing off-the-shelf tools like graph convolutional networks to model interactions. While this approach has shown improved results, the interactions primarily consider entire human identities rather than finer details. This motivates the introduction of our novel solution to address this limitation and enhance the task's performance. In this work, we strive to craft a novel framework that can effectively address two key issues ignored in previous works, namely the multi-granularity interaction and time-varying inter-person dynamics. In implementation in accord with above aims, the proposed model has mainly comprised two modules: a person-level interaction module and a part-level interaction module. The former is designed to learn the holistic and dynamic interaction among multiple persons in a coarse-grained sense. Critically, we would emphasize that a unique trait of the former module is learning temporal dynamics. For example, it recognizes that two individuals exhibit a strong correlation during handshaking but less correlation after parting ways. The latter part-level interaction module learns the interaction between the body joints of different persons. This module operates at a more fine-grained level, distinguishing it from existing approaches. By aggregating information from both granularities, our model enables accurate motion prediction. To validate the effectiveness of the proposed model, we conducted comprehensive experiments on three benchmark datasets: 3DPW, CMU-Mocap, and MuPoTS-3D. The results of these evaluations unequivocally demonstrate the empirical superiority of our model compared to previous state-of-the-art methods.**

***Index Terms*—Human 3D motion prediction, neural networks, multi-granularity interaction**

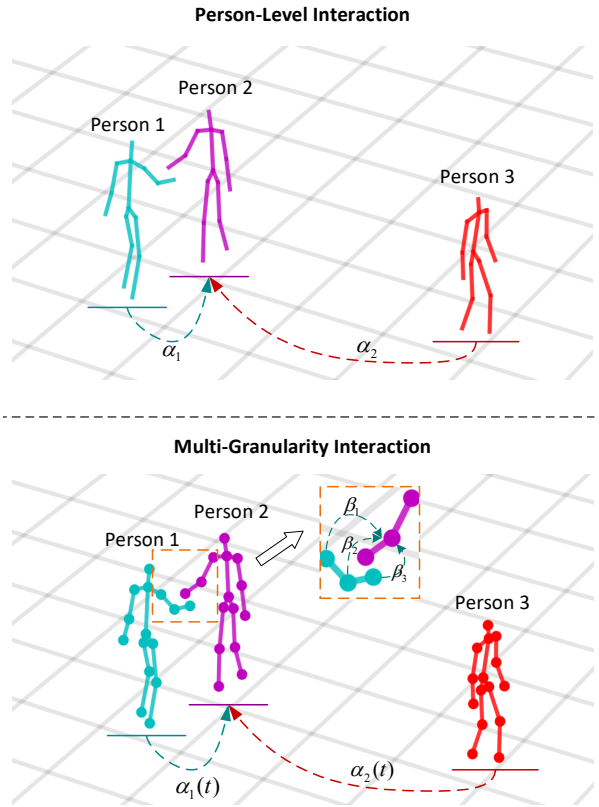

Fig. 1. A comparison of our proposed multi-granularity interaction with person-level interaction. Person-level interaction only concerns the interaction scores between people (*e.g.*, the quantities α_1 and α_2 as in the top panel). In this work, we advocate a multi-granularity method for learning the interactions. Specifically, the proposed model considers both fine-grained skeleton joint level interaction (*e.g.*, β_1, β_2 and β_3 in the bottom panel) and time-varying person-level interaction.

I. Introduction

THE technique of human 3D motion prediction has a wide range of real-world applications, ranging from 3D character animations [1]–[3], decision-making systems for autonomous driving [4]–[6], human-robot interaction [7], [8], and human-centric video generation [9], [10]. Booted by the impressive development of deep learning, recent years have witnessed unprecedented progress toward human motion predictions. A lot of research efforts have been made on single-person motion prediction [11]–[22] and human trajectory prediction [23]–[32]. However, the single-person motion prediction task only pays attention to the pose dynamics, ignoring the global motion of the person. The human trajectory prediction task regards each person in the scene as a whole, yet does not predict the pose dynamics of a specific person. [33] first investigates the problem of multi-person motion prediction in the literature. This task simultaneously predicts the global motion of multiple people as well as the pose dynamics of each individual. In this paper, we mainly focus on the multi-person 3D motion prediction task.

The common ideas in current multi-person 3D motion prediction methods are mainly derived from single-person motion

Chenchen Liu and Yadong Mu are with the Wangxuan Institute of Computer Technology, Peking University, Beijing 100871, China (e-mail: liuchenchen@pku.edu.cn, myd@pku.edu.cn). The research is supported by National Key R&D Program of China (2022ZD0160305) and Beijing Natural Science Foundation (Z190001).

Corresponding author: Yadong Mu.

prediction and human trajectory prediction. The methods can be roughly divided into two categories: methods based on single-person motion modeling [34], [35] and methods based on multi-person interaction information modeling [33], [36]. With the application of deep neural models to time series tasks, [11]–[16], [37], [38] have made great progress in single-person motion sequence prediction tasks. Along these works, [35] boils it down into multiple sub-problems, each of which corresponds to motion prediction for a unique individual in the scene and can be solved via existing single-person motion prediction. Regarding multi-person interaction modeling, the work in [33] proposes a TRiPOD model that uses attention graphs to model human-to-human and human-to-objects interactions in multi-person sequences. [36] introduces a more powerful sequence model Transformer [39] into this task, and proposes a multi-range Transformers model. A global-range Transformer branch is used to capture the interactions between multiple persons, and a local-range Transformer branch is devised for a single person's motion. Notably, this method achieves state-of-the-art results on most benchmarks.

Although the above lines of methods have achieved excellent results for many challenging cases, there is still a large space for further improvement. In a multi-person scene, each individual's movement mutually affects one another. For methods directly borrowing single-person motion prediction [34], [35], the interaction among persons is completely ignored, leading to the non-optimal prediction of future motion. For other works [33], [36] dedicated to multi-person interaction modeling, they all regard the human body as a whole when modeling the interaction, ignoring the mutual influence of the body parts between the two people. An instance appears in Figure 1, where an action of "handshaking" is happening. We argue that multi-granularity interactions should be taken into account for this case. The whole body is seemingly undergoing a relationship of "slowly approaching". Meanwhile, on the level of human skeletons, there are also local interactions such as "shaking" of the hands. Unfortunately, all of the existing works [33], [36] only model the global relationship, skipping exploring the power of learning local part interactions.

Aiming at attacking the above problems, this paper proposes a multi-granularity interaction-based multi-person motion prediction model. The model can simultaneously model the global interaction between multiple people and the local interaction between different people's body parts, enhancing the quality of feature representation for each person. Specifically, the proposed model includes two branches: a branch for *global trajectory prediction* and a second branch for *local pose dynamics prediction*. The former reads the global coordinates of the human body joints as its input. Through a global interaction module (**GIM**), we predict the interaction scores between the persons in the sequence, and meanwhile also predict the motion trajectory of the global joint of the human body. The latter branch is fed with the local coordinates of each joint of the human body relative to the global joint point as the input. It goes through a part attention module (**PAM**) to get the attention scores between different joints of different persons. Afterward, the global interaction score and part attention score are modulated to get multi-granularity interactions between different persons. The human-oriented features after multi-granularity interactive encoding are sent to an encoder-decoder model based on LSTM or transformer to obtain the final prediction results.

Our key technical contributions can be summarized as:

1) We introduce a new multi-granular interaction-based multi-person motion prediction framework, in which a global trajectory prediction branch and a local pose dynamics prediction branch are designated to model the interactions between trajectories and joints, respectively. The human joint features are fused according to the interaction information at two granularities to obtain a better representation of human features, thereby obtaining high-quality human motion prediction results.

2) The method proposed in this paper has strong generalization performance. We introduce the multi-granularity interaction modeling framework into the current two state-of-the-art models DViTA [35] and MRT [36]. Comprehensive evaluations are conducted on three datasets (3DPW, CMU-Mocap and MuPoTs-3D). The experimental results faithfully show that both models achieve consistent improvement after incorporating our proposed method.

II. Related Work

A. Human Trajectory Prediction

The task of human trajectory prediction regards the human body as a whole. The goal is to predict a set of 3D coordinates for each human characterizing its global motion. Datasets in this task are often collected from dense crowds in traffic scenes and have a wide range of applications in autonomous driving tasks. Related work can be mainly divided into regression-based or generative models. Most of the early work in this task is based on regression models. [40] propose Gaussian process dynamical models (GPDMs) for time-series analysis in human trajectories. With the development of deep learning, time series models based on Recurrent Neural Networks (RNNs) have been successfully used in sequence prediction [41], [42], machine translation [43], [44], image captioning [45]–[49], video description [50]–[56] and others. RNNs-based methods [57]–[59] have achieved better results than Gaussian process regression. [57] proposes a model called Social LSTM, which can simultaneously consider common sense rules and social conventions when walking, and predict the movement path of all persons. Recently, generative models have become state-of-the-art for trajectory prediction due to recent advances in deep generative models. The previous regression models can only generate a single trajectory prediction, while the generative model can generate the distribution of potential future trajectories, which plays a more important role in the decision-making system of autonomous driving. Conditional Variational Autoencoder (CVAE) [23]–[25], [60]–[65] and Generative Adversarial Network (GAN) [27], [28], [30], [66]–[70] are two commonly used models in generative methods. [61] is a representative method based on CVAE, which combines tools from recurrent sequences and variational deep generative modeling to produce a distribution of future trajectories for each agent in a scene. [27] proposes a Multi-Agent Tensor

Fusion (MATF) model, which can encode the historical trajectories of multiple agents and scene context into a Multi-Agent Tensor, then capture the interactions between agents and use an adversarial loss to learn stochastic directions.

B. Single-Person Motion Prediction

This task predicts the motion for a period of time in the future given the historical motion of a single person. In this task, the local information of the human joint points relative to a fixed point is often used, rather than the global information like in the human trajectory prediction task. Similar to most sequence-to-sequence tasks, the methods [11]–[13], [71]–[73] commonly used in this task are also based on RNN/LSTM models. [11] proposes an Encoder-Recurrent-Decoder (ERD) framework to model human kinematics and learn human dynamics representations. [13] proposes to use an RNN model with residuals architecture to model first-order motion derivatives, which can generate more smooth and more accurate short-term prediction results. [19] proposes a diffusion convolutional recurrent predictor for spatial and temporal movement forecasting. In view of the problem of error accumulation in long-term prediction based on RNN/LSTM models, some works [14], [15] have begun to try to replace RNN/LSTM with fully connected or convolutional networks. [16] proposes to use Discrete Cosine Transform (DCT) to encode temporal information and use GCNs [74] to encode spatial structure, and the model has proven highly effective for the human motion prediction task. Based on [15], [18] proposes a Multi-Scale Residual Graph Convolution Network (MSR-GCN), where GCNs are used to extract features from fine to coarse scale and then from coarse to fine scale. This method shows stronger feature expression ability.

C. Multi-Person Motion Prediction

As the prediction of human motion becomes increasingly important in real-world applications, recent works are no longer limited to the study of human trajectory prediction or single-person motion. [33] first proposed a multi-person motion prediction task. Currently, common methods in this task can be divided into two categories: traditional methods [34], [35] based on single-person motion prediction and more recent methods [33], [36] based on multi-person information interaction modeling. The former simplifies the multi-person motion prediction task into a single-person motion prediction task, divides a multi-person sequence into multiple single-person sequences, and then uses the single-person motion prediction methods introduced in Section II-B to model each single-person sequence and predict the results, and finally merge the results of each single-person sequences. For example, a VAE-based method is proposed in [34], in which the encoder and decoder composed of LSTM are used to complete the encoding of the observed sequence and the prediction of the unknown sequence in the single-person sequence, respectively. The latter no longer regards each person in the sequence as an independent individual and considers the interaction between multiple persons in the sequence and the interaction between person and scene in the modeling process to obtain a better feature representation of each person, and then complete the prediction of the unknown sequence. [33] proposes a TRiPOD framework that uses two attention graphs to model the human-to-human and human-to-object information interactions. With the application of Transformer in computer vision, [36] proposed a multi-range Transformer method, which uses a local-range Transformer to encode the motion of a single person in the sequence, and uses the global-range Transformer to encode the motion of multiple persons, and then send both to a Transformer-based decoder to predict the future motion. The above two methods take the human body as a whole when modeling the interaction between persons, ignoring the mutual influence between the body parts of different persons. The multi-granularity interaction method proposed in this paper considers two scales of the whole body and body parts at the same time, and can better model the interaction information.

D. Skeleton-based Action Recognition

Skeleton-based action recognition aims to identify human actions based on the 2D or 3D positions of body joints, which are often represented as skeletal structures. With the development of graph convolutional network, [75] introduced the Spatial Temporal Graph Convolutional Network (ST-GCN), which employs graph convolutional network to model the spatial and temporal information in skeletal data. [76] proposes the Actional-Structural Graph Convolutional Network (AS-GCN), which combines actional and structural links into a generalized skeleton graph, utilizing actional-structural graph convolution and temporal convolution as building blocks for learning spatial and temporal features, enhancing the model's ability to capture detailed action patterns. [77] proposes a novel Shift Graph Convolutional Network (Shift-GCN), which addresses the limitations of the traditional ST-GCN by incorporating a shift operation for better temporal modeling. The aforementioned GCN-based methods have proposed promising approaches for spatiotemporal modeling of human motion and have been applied in human motion sequence prediction tasks, such as the work [38]. However, these methods focus primarily on single-person temporal modeling and do not consider the multi-granularity information interaction between persons in multi-person sequences. Recently, the work most closely related to multi-person motion prediction is the group activity recognition task proposed by [78]. This work estimates multi-person skeletons from existing real-world video datasets (i.e., Kinetics and Volleyball-Activity) and releases two new group activity recognition benchmarks. The proposed Zoom Transformer model in this work is mainly to extract high-level group activity patterns and pay more attention to the overall context in the group, which is different from the goal of focusing on the modeling and prediction of each person's motion details in multi-person motion prediction.

III. The Proposed Method

In general human activities have social attributes. When humans are in a multi-person scene, both the global trajectories of individuals, as well as local movements of the limbs, will be affected by the surrounding people. Based on this observation,

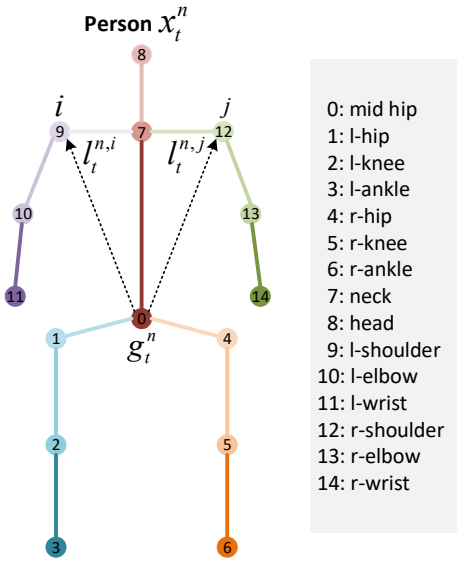

Fig. 2. Representation of person n at frame t in the sequence. Let g_t^n be the "mid hip" joint in the world coordinate system, and $l_t^{n,i}$ is the local coordinate of joint i relative to g_t^n. The method in this paper consists of two branches to predict g_t^n and l_t^n, respectively. The rightmost shows the correspondence between the joint id and the joint name.

we design a multi-granular interaction-based multi-person motion prediction method, as shown in Figure 3. For the two branches in the model, the global trajectory prediction branch inspects the global motion of multiple people and renders the global trajectory prediction of the center point of each human body. The core of this branch is a global interaction module. A second local pose dynamics predicting branch takes the local motion relative to the center of the human body as the input and predicts human pose dynamics. This branch mainly includes a part attention module for predicting the attention scores between human joints and a multi-granularity interaction fusion Module. The details of each branch are described below.

A. Problem Definition

Given a sequence with the historical motion of N persons, our goal is to predict their future 3D motion. Formally, given the part motion of a person n as $\mathbf{X}_{1:T_o}^n = \{x_1^n, x_2^n, ..., x_t^n, ..., x_{T_o}^n\}$, where $n \in \{1, 2, ..., N\}$, t is the time step, and T_o is the length of the observation frames. The model is desired to predict future motions $\mathbf{X}_{T_o+1:T}^n$, where T represents the end of the sequence. $x_t^n \in \mathbb{R}^{3J}$ (J is the total count of human skeleton joints) represents the pose of person n at the time step t. Let $x_t^{n,i}$ be the state of the joint i. We split the joint $x_t^{n,i}$ into two parts, including global g_t^n and local $l_t^{n,i}$:

$$x_t^{n,i} = g_t^n + l_t^{n,i}, \tag{1}$$

where g_t^n is the spatial location of the center of person n in the world coordinate. $l_t^{n,i}$ is the local coordinate of the joint i relative to g_t^n, as shown in Figure 2. The trajectory $g_{1:t}^n$ represents the global movements of person n in the world coordinate and $l_{1:t}^n$ indicates the local movements of all person's joints. The two branches in the proposed method are primarily used to predict $g_{T_o+1:T}^n$ and $l_{T_o+1:T}^n$, respectively. Following previous common practice, we adopt velocities instead of raw coordinates as the input to the model. The velocity of the pose at time t is $v_t^{n,i} = x_t^{n,i} - x_{t-1}^{n,i}$, and the corresponding global and local velocities are $v_{g_t^n}$ and $v_{l_t^{n,i}}$, respectively.

B. Global Trajectory Prediction

As stated before, this branch is mainly used for global interactions prediction and human center trajectory forecasting. As shown in Figure 3, this branch consists of two parts: person-level feature projection and global interaction module. Each part is described in detail below.

Person-level feature projection. Intuitively, to judge whether there is an interaction between two people in the scene, we can observe their distance and their respective moving speeds. Therefore, the input of each person in this branch can be represented as

$$p_t^n = x_t^n \oplus v_t^n, \tag{2}$$

where x_t^n is the original world coordinate of person n at time t, v_t^n is the velocity relative to the previous frame, and $\oplus$ represents the concatenation operation. $p_t^n \in \mathbb{R}^{6J}$. All persons in the sequence can be represented as $\mathbf{P} \in \mathbb{R}^{T_o \times N \times 6J}$. In this branch, we treat each person as a whole, where $6J$ is the dimension of all parameters that describe each person. We use a fully connected sub-network to project each person's feature to d_g-dimensional to obtain person-level feature representations, namely

$$\mathbf{P}' = \text{PROJ}(\mathbf{P}; \mathbf{W}_{PRG}^G), \tag{3}$$

where $\mathbf{P}' \in \mathbb{R}^{T_o \times N \times d_g}$, "PROJ" represents the feature projection network, here is a fully connected sub-network, $\mathbf{W}_{PRG}^G$ encapsulates all the parameters of the projection network. $\mathbf{P}'$ is sent to the global interaction module to calculate the interaction scores.

Global interaction module. We propose a straightforward scheme for predicting the interaction between any two persons. For person n and person m, let their projected features be $p_t'^n$ and $p_t'^m$ at time step t. Since the relative distance and relative motion between two persons are more concerned, we subtract the two features and send it to a multi-layered fully connected sub-network for rendering an interaction score:

$$\alpha_t^{m,n} = \text{GIM}(p_t'^m - p_t'^n; \mathbf{W}_{INTER}^G), \tag{4}$$

$$\alpha_t^m = \text{softmax}(\{\alpha_t^{m,1}, \alpha_t^{m,2}, ..., \alpha_t^{m,N}\}), \tag{5}$$

where $\alpha_t^{m,n}$ is the interaction score between person m and n on frame t. α_t^m is the global interaction score of person m with everyone except himself, and $\mathbf{W}_{INTER}^G$ represents the parameters of global interaction module. After going through the softmax function, the sum of each person's interaction score with other people in the sequence is 1. The interaction scores $\mathbf{A}$ will be used by the multi-granularity interaction fusion module as in Section III-C.

Global feature fusion. Next, we use the calculated interaction score to perform a weighted summation of the features

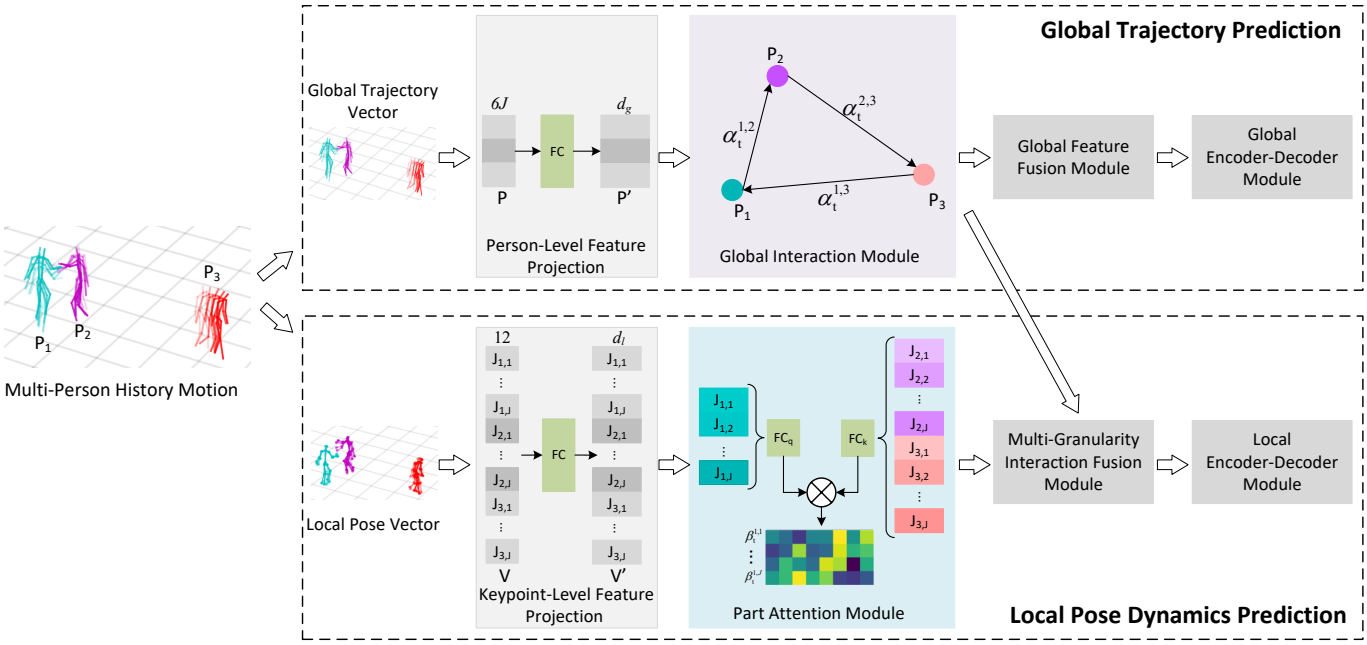

Fig. 3. The computational pipeline of our proposed method. See the main text for more explanations.

of other people in the scene and fuse it as the context information with each person's original feature. The calculation is performed as follows:

$$p_t''^m = \eta p_t'^m + (1-\eta)\sum_{n=1}^{N} \alpha_t^{m,n} \times p_t'^n, \tag{6}$$

where η is the harmonic parameter and we set it as 0.7 without further empirical fine-tuning, $p_t''^m$ is obtained after fusing the features of other people in the sequence according to the global interaction information. The fused features are fed into an encoder-decoder network based on LSTM or Transformer. Finally, the prediction results of the global trajectory $g_{T_o+1:T}$ of the human body are received as the output.

C. Local Pose Dynamics Prediction

In the process of human-to-human communication, in addition to the influence of the overall trajectory, there is also the interaction between the joints of different bodies. For example, when two people shake hands, they will stretch out their right hands at the same time. Obviously, these two right-hand joints exhibit instantaneous strong interaction. To model such local joint interactions, we propose the local pose dynamics prediction branch, which consists of three parts (keypoint-level feature projection, part attention module, and multi-granularity interaction fusion module) as shown in Figure 3.

Keypoint-level feature projection. In this branch, we take the velocity $\{v_{l_t^n}, v_{l_{t-1}^n}, v_{l_{t-2}^n}, v_{l_{t-3}^n}\}$ of the human joints at the current frame and the previous three frames as the input. As stated above, $v_{l_t^n}$ denotes the velocity of all joints of person n at time step t. All persons in the sequences can be represented as $\mathbf{V} \in \mathbb{R}^{T_o \times NJ \times 12}$. Unlike the case of person-level modeling, the central features here are defined for human joints. A fully connected sub-network is adapted to project the feature dimension of each joint to d_l, namely

$$\mathbf{V}' = \mathrm{PROJ}(\mathbf{V}; \mathbf{W}_{PRG}^L), \tag{7}$$

where $\mathbf{V}' \in \mathbb{R}^{T_o \times NJ \times d_l}$, and $\mathbf{W}_{PRG}^L$ are the parameters of the joint projection network.

Part attention module. It is curated to obtain the interaction between each joint and all the joints of others. A method similar to self attention [39] is adopted. Let $v_t'^{n,i}$ denote i-th joint of person n at time step t, and $v_t'^{N\setminus n}$ denote all other joints after removing person n. The notations $v_t'^{n,i} \in \mathbb{R}^{1\times d_l}$ and $v_t'^{N\setminus n} \in \mathbb{R}^{(N-1)J\times d_l}$. The attention between $v_t'^{n,i}$ and $v_t'^{N\setminus n}$ is defined as follows:

$$f_q = \mathrm{FC}_k\left(v_t'^{n,i}, \mathbf{W}_{PAM}^Q\right), \tag{8}$$

$$f_k = \mathrm{FC}_q\left(v_t'^{N\setminus n}, \mathbf{W}_{PAM}^K\right), \tag{9}$$

$$\beta_t^{n,i} = \mathrm{softmax}\left(\frac{f_q f_k^{\mathrm{T}}}{\sqrt{d}}\right), \tag{10}$$

where $\mathbf{W}_{PAM}^Q, \mathbf{W}_{PAM}^K \in \mathbb{R}^{d_l \times d}$ represent the parameters of the corresponding fully connected networks. $\beta_t^{n,i} \in \mathbb{R}^{1\times(N-1)J}$ are the attention scores. For all joints in the sequence, we can get attention $\mathbf{B} \in \mathbb{R}^{T_o \times NJ \times (N-1)J}$.

Multi-granularity interaction fusion module. In a multi-person scene, only two people interacting on the global trajectory will have this mutual influence between their joints, so we need a multi-granularity interaction fusion module to fuse the interactions at these two granularities. From the global interaction module and part attention module, we can get both person-level and part-level interaction information, $\mathbf{A}$ and $\mathbf{B}$, respectively. The calculation of the multi-granularity score

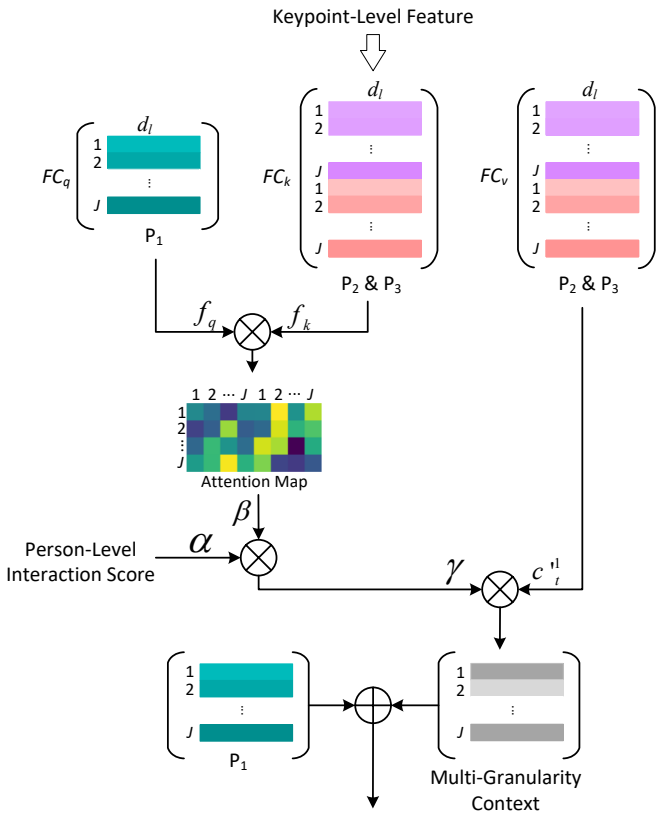

Fig. 4. The computational pipeline of part attention module and multi-granularity interaction fusion module. J is the number of joints for each person. P_1, P_2, and P_3 represent person 1, person 2, and person 3 in the motion sequence, respectively. FC_q, FC_k, and FC_v denote the fully connected layers for encoding the query, key, and value features in the self-attention mechanism, respectively.

between person n and the others in the sequences is defined as follows:

$$\begin{aligned} \alpha_t'^n &= \text{repeat}(\alpha_t^n, J), \\ \varphi_t^{n,i} &= \beta_t^{n,i} \times \alpha_t'^n, \\ \gamma_t^{n,i} &= \text{softmax}(\varphi_t^{n,i}), \end{aligned} \tag{11}$$

where the repeat operation is to duplicate the person-level interaction score J times to get $\alpha_t'^n \in \mathbb{R}^{1\times(N-1)J}$. The purpose is to align with $\beta_t^{n,i}$, $\varphi_t^{n,i}$ is an intermediate variable in the calculation of $\gamma_t^{n,i}$. $\gamma_t^{n,i} \in \mathbb{R}^{1\times(N-1)J}$ is the new interaction score between the ith joint and all other joints in the sequence that fuses the two granular interaction information. Then we fuse the joint features according to $\gamma_t^{n,i}$,

$$\begin{aligned} c_t^{n,i} &= \text{Concat}(v_t^{1,1}, v_t^{1,2}, ..., v_t^{1,J}, ..., v_t^{N,J}), \\ c_t'^{n,i} &= \text{FC}_v(c_t^{n,i}, \mathbf{W}_{PAM}^V), \\ v_t''^{n,i} &= v_t'^{n,i} + \text{MatMul}(\gamma_t^{n,j}, c_t'^{n,i}), \end{aligned} \tag{12}$$

where $c_t^{n,i} \in \mathbb{R}^{(N-1)J\times d_l}$ represents the intermediate variable obtained by concatenating the joint features of all persons except n, "Concat" represents concatenation operator, "MatMul" represents the matrix multiplication operator, and $v_t''^{n,i} \in \mathbb{R}^{1\times d_l}$ is obtained after fusing the features of other joint points by using multi-granularity interaction information. The specific calculation flow of the part attention module and multi-granularity interaction fusion module is shown in Figure 4. The fused features are sent to an encoder-decoder network as detailed in Section III-D, which predicts local pose dynamics $l_{T_o:T}$.

D. Encoder-Decoder Framework

After obtaining the global and the local pose feature, we send them into the global and local encoder-decoder frameworks, respectively. Regarding the encoder-decoder framework, there are two commonly used models recently, either based on RNN/LSTM or based on Transformer. The representative works are DViTA [35] and MRT [36] respectively. Since our method is essentially precoding human features, it can be easily combined with both of these two methods. In the experimental section, we will verify in detail the effect of the proposed GIM and PAM modules when bridged with these two base methods. Below, let us take LSTM as an example to introduce the structure of the encoder-decoder part.

The encoder LSTM conveys the part history of each person's global or local features. Taking the global feature $p_t''^n$ of person n at time t for instance,

$$h_t = \text{LSTM}(h_{t-1}, p_t''^n; \mathbf{W}_{enc}^{global}), \tag{13}$$

where $\mathbf{W}_{enc}^{global}$ are parameters of the global feature encoder. After encoding the input, the decoder LSTM takes the last observed frame $g_{T_o}^n$ and the last hidden state h_{T_o} of the encoder, and outputs the predicted hidden state:

$$\begin{aligned} h_{T_o+t+1} &= \text{LSTM}\left(h_{T_o+t}, g_{T_o+t}^n; \mathbf{W}_{dec}^{global}\right), \\ g_{T_o+t+1}^n &= \phi\left(h_{T_o+t+1}; \mathbf{W}_\phi\right), \end{aligned} \tag{14}$$

where ϕ is a fully connected layer. Following this step, we can iteratively generate global trajectory prediction results $g_{T_o:T}^n$. In the local pose dynamics prediction branch, we use a similar decoder-decoder structure for generating human pose dynamics $l_{T_o:T}^n$.

E. Loss Function

Following the majority of previous works [33], [35], [36], we adopt Mean Square Error (MSE) loss as the optimization objective to supervise the training process of the model. On the predicted frame k, x_t^n and $x_t'^n$ are the groundtruth and predicted motion of person n, respectively. The MSE loss can be calculated as

$$\mathcal{L}_t = \frac{1}{N\times J}\sum_{n=1}^{N}(\|x_t^n - x_t'^n\|^2). \tag{15}$$

For the sequence prediction problem, the human poses on the long-term frame are more difficult to be precisely predicted compared with the short-term ones. Therefore, we adopt the cumulative learning strategy to gradually increase the learning difficulty of the model. An additional frame is incrementally added to the loss computation after every E training epoch.

IV. Experiments

A. Dataset Description

We adopt three multi-person motion prediction benchmarks in the experiments: 3DPW [79], CMU-Mocap [80], and MuPoTs-3D [81]. Details of the three datasets are presented below.

3DPW [79]. The samples are videos taken from a moving phone camera. It contains 60 videos and about 68k frames covering multiple scenarios and actions, with the annotated 3D positions of 24 keypoints. It is first used as a benchmark for human 3D pose estimation tasks in wild scenes. The work in [33] first re-purposes it for the multi-person 3D motion prediction task. Here we follow the setting of [33]. The 3DPW dataset consists of three parts: training, validation, and testing, including 221, 36, and 85 sequences, respectively. The length of each sequence is 30 frames, of which 16 are assumed to be observed and 14 are to be predicted by the model.

CMU-Mocap [80]. It is a human 3D pose video dataset collected by multi-view cameras and markers in a laboratory scene. A total of 112 subjects are included, most of which have only one person's movements, and a small number of scenes contain two persons' movements and interactions. The work of [36] samples and mixes two parts including one person and two persons, constructing a dataset that includes 3 persons in the sequence. The dataset consists of training and testing splits, with 6,000 and 800 sequences respectively. Each sequence has a length of 120 frames, 30 of which are known, and the rest are left for prediction. We fully follow the setting of [36] and use the benchmark constructed by them to validate our proposed method.

MuPoTS-3D [81]. It is a human 3D pose dataset collected by a multi-view mark-less motion-capture system, including 5 indoor and 15 outdoor scenes, a total of 20 sequences, and 8 subjects. There are typical $2 \sim 3$ persons in each sequence. Following [36], we use this dataset to verify the generalization performance of the model trained on the CMU-Mocap dataset.

Mix1 & Mix2. In order to evaluate the performance of our proposed model in scenarios involving a larger number of individuals, we adopt the methodology presented in the MRT paper. We sample and combine data from the CMU-Mocap and Panoptic [82] datasets to generate a mixed training set. This training set contains approximately 3,000 samples, each featuring 9 to 15 people in the scene. Regarding the test set, it is divided into two parts: Mix1, which combines CMU-Mocap and Panoptic data, and Mix2, which blends CMU-Mocap, MuPoTS-3D, and 3DPW data. Mix1 is composed of 800 samples and includes scenes with 9 to 15 people, while Mix2 consistently features 11 persons in each of its 400 samples.

B. Evaluation Protocol

Following common practice, we use the metric of mean per joint position error (MPJPE) in the evaluation. The MPJPE calculates the mean error between the predicted joint positions and the groundtruth. For 3DPW dataset, following [33], we calculate the MPJPE on the predicted frame at 5 moments in the future 100, 240, 500, 640, and 900 milliseconds. For CMU-Mocap and MuPoTS-3D datasets, we simultaneously report the results on short-term 1 second and long-term 3 seconds. All units in the experiments are in millimeters.

C. Base Models

In order to verify the effectiveness of the multi-granularity interaction-based model proposed in this paper, we select two state-of-the-art models DViTA [35] and MRT [36] for the experiments.

DViTA[1]: This method simplifies multi-person motion prediction into multiple single-person motion prediction subtasks, and adopts a simple LSTM-based encoder-decoder structure. It reported excellent results on 3DPW dataset. We incorporate the method proposed in this paper into the original DViTA model (*i.e.*, first get the features of each person that integrates multi-granularity interactive information, and then send it to the original DViTA model).

MRT[2]: MRT is a method based on multi-person interaction modeling. The core of this method is a multi-range transformer. Specifically, it uses a local-range transformer to encode single-person motion sequences, and another global-range transformer to encode the motion of multiple people. This method treats the person as a whole and does not model the interactions between joints. Similar to the processing of the DViTA model, we first use the method proposed in this paper to pre-code multi-person motion sequences and then feed them into the MRT model.

D. Implementation Details

In this paper, the proposed method is mainly experimentally validated on two different models, DViTA and MRT, with varying implementation details for each model. For the DViTA model, we use a given 15-frame historical motion sequence as input to predict the next 14 frames. This model is based on LSTM with a hidden state dimension of 64. We add the GIM and PAM modules proposed in this paper before the LSTM to pre-encode the input motion sequence with multi-granularity interaction, and then feed this feature into the LSTM of the DViTA model. The entire GIM, PAM modules, and the original DViTA model are jointly trained from scratch. The model is implemented using PyTorch and trained using the Adam optimizer. The initial learning rate is 0.004, and the ReduceLROnPlateau strategy is used for adaptive updating. During the training process, the mean squared error loss function introduced in Section III-E adopted, with $E = 4$ in the cumulative learning strategy, meaning that a new frame is added to the loss function calculation every four epochs. The model is trained for a total of 100 epochs.

For the MRT model, it is capable of handling both long-term predictions of 3 seconds and short-term predictions of 1 second. For the 3-second long-term prediction, we mainly adhere to the settings in the original paper, using a 1-second historical motion as input to recursively predict the motion for

[1] https://github.com/vita-epfl/decoupled-pose-prediction
[2] https://github.com/jiashunwang/MRT

TABLE I
MPJPE ON PREDICTING 1, 2, AND 3 SECONDS MOTION ON CMU-MOCAP AND MUPOTS-3D TEST SET. WE VERIFY THE IMPACT OF OUR PROPOSED MODULES GIM AND PAM ON THE BASELINE MODEL MRT.

Method	CMU-Mocap			MuPoTS-3D		
	1s	2s	3s	1s	2s	3s
MRT [36]	165.8	275.3	373.2	109.6	200.4	302.6
MRT + GIM	164.1	269.4	361.3	108.2	196.3	291.8
MRT + PAM	159.5	262.7	351.4	104.3	192.7	286.3
MRT + GIM + PAM	157.2	257.5	344.6	102.1	182.9	265.8

TABLE II
MPJPE ON PREDICTING 1, 2, AND 3 SECONDS MOTION ON MIX1 AND MIX2 DATASET.

Method	Mix1 (9∼15 persons)			Mix2 (11 persons)		
	1s	2s	3s	1s	2s	3s
MRT [36]	189.1	279.9	342.5	210.6	357.7	477.3
MRT + GIM	179.5	272.8	329.8	202.8	349.2	472.3
MRT + PAM	180.9	274.0	332.0	203.9	352.9	468.4
MRT + GIM + PAM	178.2	269.1	324.9	201.8	346.7	460.7

the upcoming 3 seconds. During training, the learning rates for the predictor and discriminator are set at 3e-4 and 5e-4, respectively, and the model undergoes 200 epochs of training. For experiments with 2∼3 people, the batch size is set to 32, and for experiments with 9∼15 people, the batch size is set to 8. For the 1-second short-term prediction, we employ a cumulative learning strategy, training the model for 200 epochs with a value of E set to 10. The learning rates for the predictor and discriminator are 1e-4 and 3e-4, respectively.

E. Experimental Results and Analysis

As introduced in Section III, the proposed method mainly includes two modules, the global interaction module (GIM) and part attention module (PAM), which model the interaction between human bodies from the global trajectory and local joint granularities, respectively. First, we verify the impact of these two modules on the base models DViTA and MRT. Critically, DViTA and MRT models differ in their ability to predict unknown sequence lengths. The original model of DViTA can only predict human motion in the next 14 frames, and MRT can predict motion in the next 3 seconds. The MRT-related verification in the following experiments is performed on all three datasets and 3DPW and CMU-Mocap for DViTA-related. The experimental setting of DViTA is to predict the future 14 frames.

The experimental results of MRT on CMU-Mocap and MuPoTS-3D are shown in Table I. The models are trained on the training set of CMU-Mocap and tested on the test set of CMU-Mocap and MuPoTS-3D. The purpose is to simultaneously verify the accuracy of the model on CMU-Mocap and the generalization performance on MuPoTS-3D. Experimental results show that both our proposed GIM and PAM modules bring effective improvements to the original MRT model. Among them, the improvement brought by the PAM module is more significant than that of the GIM module (373.2 → 361.3 *v.s.* 373.2 → 351.4 in 3s motion prediction), because the MRT model itself has modeled the global interaction of the human body, but it lacks the modeling of the interaction between the body joints. The experimental results are consistent with our intuition. After integrating the interaction of GIM and PAM at two granularities, the model can achieve the best results (373.2 → 344.6).

TABLE III
MPJPE ON CMU-MOCAP VALIDATION DATASET. WE VERIFY THE IMPACT OF GIM AND PAM MODULES ON THE BASELINE MODEL DVITA.

Method	CMU-Mocap prediction time in milliseconds					
	Avg.	100	240	500	640	900
DViTA [35]	77.58	22.82	45.93	85.15	101.97	132.06
DViTA + GIM	76.54	24.80	47.25	83.79	99.08	127.78
DViTA + PAM	76.69	25.17	47.41	83.53	99.12	128.20
DViTA+GIM+PAM	75.97	23.84	46.08	83.39	98.84	127.72

TABLE IV
MPJPE ON MOPUTS-3D DATASET. WE VERIFY THE IMPACT OF GIM AND PAM MODULES ON THE BASELINE MODEL DVITA.

Method	MoPuTS-3D prediction time in milliseconds					
	Avg.	100	240	500	640	900
DViTA [35]	51.48	16.01	30.42	56.70	67.88	86.41
DViTA + GIM	49.24	16.59	30.03	53.66	63.86	82.04
DViTA + PAM	49.06	16.35	29.85	53.65	63.82	81.62
DViTA+GIM+PAM	47.74	15.05	27.65	51.78	62.64	81.58

TABLE V
MPJPE ON 3DPW VALIDATION DATASET. WE VERIFY THE IMPACT OF THE PAM MODULE ON THE TWO BASELINES DVITA AND MRT.

Method	3DPW-Validation prediction time in milliseconds					
	Avg.	100	240	500	640	900
DViTA [35]	56.74	13.87	30.84	62.80	75.56	100.64
DViTA + PAM	54.21	15.90	32.10	59.71	70.78	92.58
MRT [36]	53.06	17.16	31.68	57.82	69.04	89.56
MRT + PAM	52.61	16.76	30.87	57.31	68.31	89.78

As shown in Table II, the performance of the proposed GIM and PAM modules on the Mix1 and Mix2 test sets is presented. As mentioned earlier, the Mix1 and Mix2 datasets contain scenes with a larger number of people, ranging from 9 to 15 individuals. The experimental results demonstrate that both GIM and PAM exhibit positive effects, with the GIM module showing a more significant improvement compared to the PAM module. This suggests that in scenes with a higher number of people, the scale of keypoints considered by the PAM module expands, and its performance may be affected by noise. Therefore, it requires cooperation with the GIM module to achieve better results.

Similarly, we also verified the effects of GIM and PAM on the DViTA method on the CMU-Mocap and MuPoTS-3D datasets, and the experimental results are shown in Table III and Table IV. We can see that GIM and PAM modules can consistently improve the DViTA model. Since the original DViTA model is based on a single-person sequence prediction method and does not model interactions between multiple persons, the improvement brought by the GIM module is more significant.

The experimental results of the two models, DViTA and MRT, on the 3DPW validation set are shown in Table V. Since there are only two people in the sequence of the 3DPW dataset, the global interaction score between the two people

TABLE VI
MPJPE ON 3DPW TEST DATASET.

Method	3DPW-Test					
	prediction time in milliseconds					
	Avg.	100	240	500	640	900
PF-RNN [13]						
+ ST-GAT [83]	157.79	67.12	116.53	164.61	189.82	250.88
Mo-Att [84]						
+ ST-GAT [83]	149.63	62.41	94.59	153.24	188.02	249.91
SC-MPF [85]	123.23	45.44	73.73	129.23	159.47	208.31
TRiPOD [33]	84.21	31.04	50.80	84.74	104.05	150.41
LSTMV_LAST [34]	82.96	25.89	47.57	86.39	106.65	148.28
DViTA [35]	65.67	19.53	36.89	68.29	85.45	118.21
DViTA + PAM	64.71	20.18	37.25	66.79	84.10	115.24
MRT [36]	65.57	21.80	37.09	68.33	84.58	116.07
MRT + PAM	64.24	21.16	35.90	67.00	83.37	113.77

TABLE VII
MPJPE ON 3DPW VALIDATION DATASET. WE VERIFY THE EFFECT OF CUMULATIVE LEARNING ON THE METHOD PROPOSED IN THIS PAPER. "CL" CL IS AN ACRONYM FOR CUMULATIVE LEARNING.

Method	3DPW-Validation					
	prediction time in milliseconds					
	Avg.	100	240	500	640	900
DViTA [35]	56.74	13.87	30.84	62.80	75.56	100.64
Ours (no CL)	55.01	14.20	31.50	61.32	73.78	94.25
Ours (CL, E=2)	54.63	15.14	31.97	60.59	72.16	93.29
Ours (CL, E=4)	54.21	15.90	32.10	59.71	70.78	92.58
Ours (CL, E=6)	54.36	16.52	34.04	59.30	69.93	92.05

obtained by our GIM module is always 1, and the effect of the GIM module cannot be reflected. Therefore, we only verify the impact of the PAM module. It can be seen from the experimental results that the PAM module brings positive improvements to both models. The improvement brought by PAM is larger for the relatively weak DViTA model than MRT (56.74 → 54.21 *v.s.* 53.06 → 52.61). In order to compare with other methods more thoroughly, we select the MRT model to conduct experiments on the 3DPW test set, as shown in Table VI. Experimental results show that our method can still bring significant improvements to DViTA and MRT.

F. Ablation Study and Efficiency Analysis

In order to verify the impact of cumulative learning mentioned in Section III-E on the model performance, we designed an ablation study on the 3DPW validation set using DViTA as the baseline model. The experimental results are shown in Table VII. In the table, the "Ours" method represents the DViTA method with our added PAM module. From the results, we can see that even without cumulative learning, our method can still bring improvements to the baseline model, reducing the average error from 56.74 to 55.01. At the same time, we also conducted an ablation study on the selection of the hyperparameter E in cumulative learning, and it can be observed that when E is set to 4, the model can achieve a relatively balanced performance in short-term (100 ms) and long-term (900 ms) frame predictions, with the minimum prediction error.

To investigate the impact of the proposed method in this paper on the running efficiency of the original baseline model, especially when the number of persons in the scene increases and whether the running time of the model becomes unacceptable, we conducted a comparison experiment on the model's computational complexity using the test set Mix1. First, we calculated the parameter sizes of the MRT model and the model proposed in this paper, which are 6.6 million and 6.9 million, respectively. The parameter size of the model does not change with the increase in the number of characters in the scene. Next, we calculated the computational complexity GFLOPS (Giga Floating-point Operations Per Second) of the models when there are 9 to 15 people in the scene, as shown in Figure 5. From the figure, it can be seen that the computational complexity of both the proposed model and the MRT model basically maintains linear growth as the number of persons in the scene increases. Due to the additional introduction of GIM and PAM modules in our model, the computational complexity is relatively higher, but such a relative increase is acceptable in scenes with up to 15 persons.

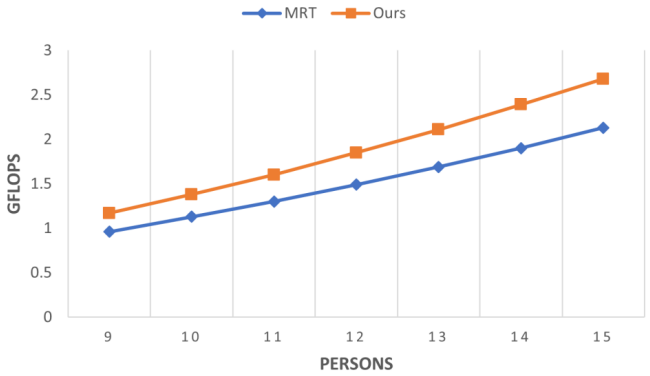

Fig. 5. The change in the model's computational complexity (GFLOPS) as the number of persons in the test set Mix1 varies.

G. Qualitative Analysis

To investigate whether the effects of GIM and PAM modules are consistent with our intuition, we visualize the person-level interaction scores and part-level attention on several examples, as shown in Figure 6 and Figure 7.

Person-level interaction. There are three examples in Figure 6. For each example, we visualize each person's interaction scores with others over time. In the first example, Person 1 is gradually approaching person 3. It can be seen that the interaction scores "P3→P1" and "P1→P3" tend to increase significantly. In the second example, person 2 and person 3 are far away from person 1, and their influence on person 1 is relatively close, thus the interaction scores "P2 → P1" and "P3 → P1" are both around 0.5. For person 2 and person 3, the distance between them is closer and there is a stronger mutual influence, so the values of "P3 → P2" and "P2 → P3" are both significantly higher. The situation for the third example is similar to the second. The results of the three examples above collectively show that the person-level interaction scores modeled by our GIM module are consistent with our intuition.

Part-level attention. Figure 7 show the effect of the part attention module. In the scene in Figure 7 (a), Person 2 and Person 3 are approaching each other, Person 2 is moving his

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3298755

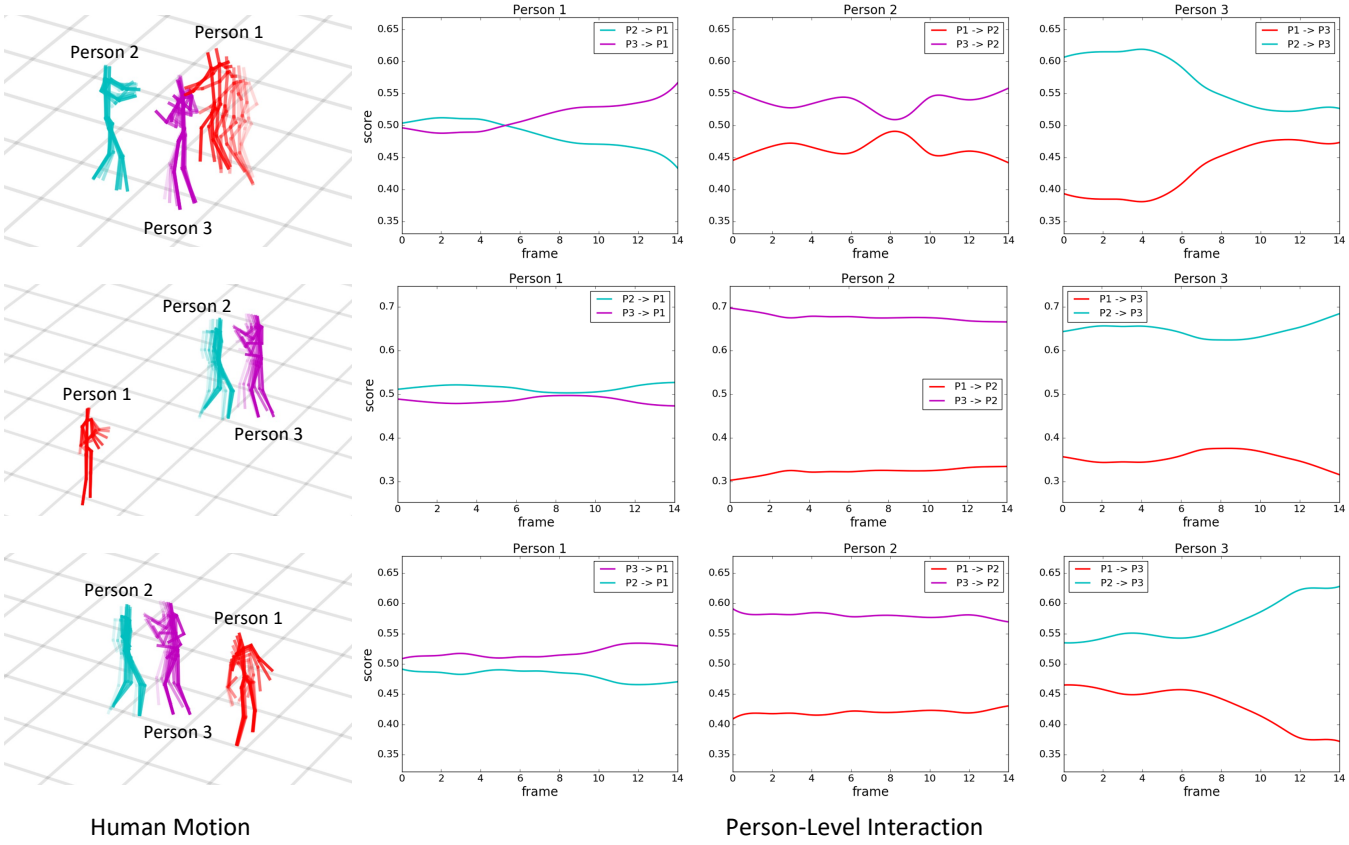

Fig. 6. Visualization of person-level interactions. The figure includes three examples. For each instance, the leftmost represents the dynamics of the persons in the sequences, and the three images on the right represent the changes in the person-level interaction scores between different persons over time, for example, P2 $\rightarrow$ P1, P3 $\rightarrow$ P1 represent the influence of person 2 and person 3 on person 1, respectively. Better viewing in color mode.

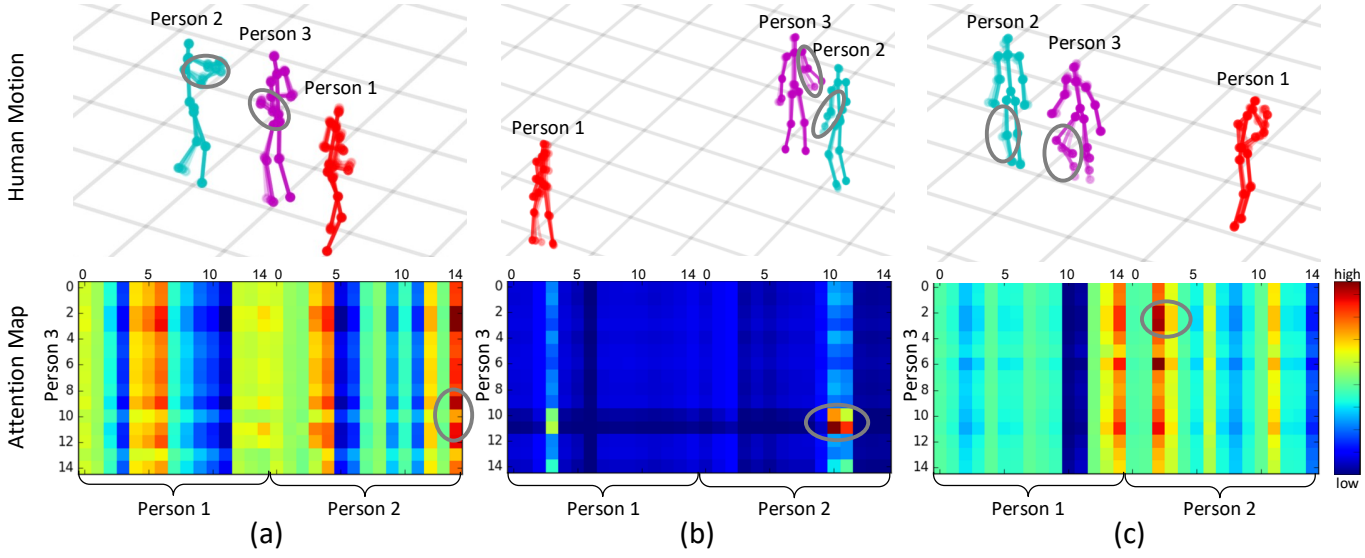

Fig. 7. Visualization of part-level attention. The figure includes three examples. For each instance, the upper image shows the dynamics of the three persons in the sequence, and the lower image shows the attention map between the 15 joints of person 3 and the other two persons' joints. Better viewing in color mode. The correspondence between joint id and joint name in the attention map is shown in Figure 2.

right hand, and Person 3 is moving his left arm. From the attention map, we can see that the response values between the left arm (Joint ID 9$\sim$11) of Person 3 and the right hand (Joint ID 14) of Person 2 are higher (in gray circle). In Figure 7 (b), the left arms of Person 2 and Person 3 are waving, and the other parts of the body are basically still, thus the parts with higher response values in the attention map are concentrated in the joints 10 and 11 of Person 2 and Person 3. In Figure 7

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3298755

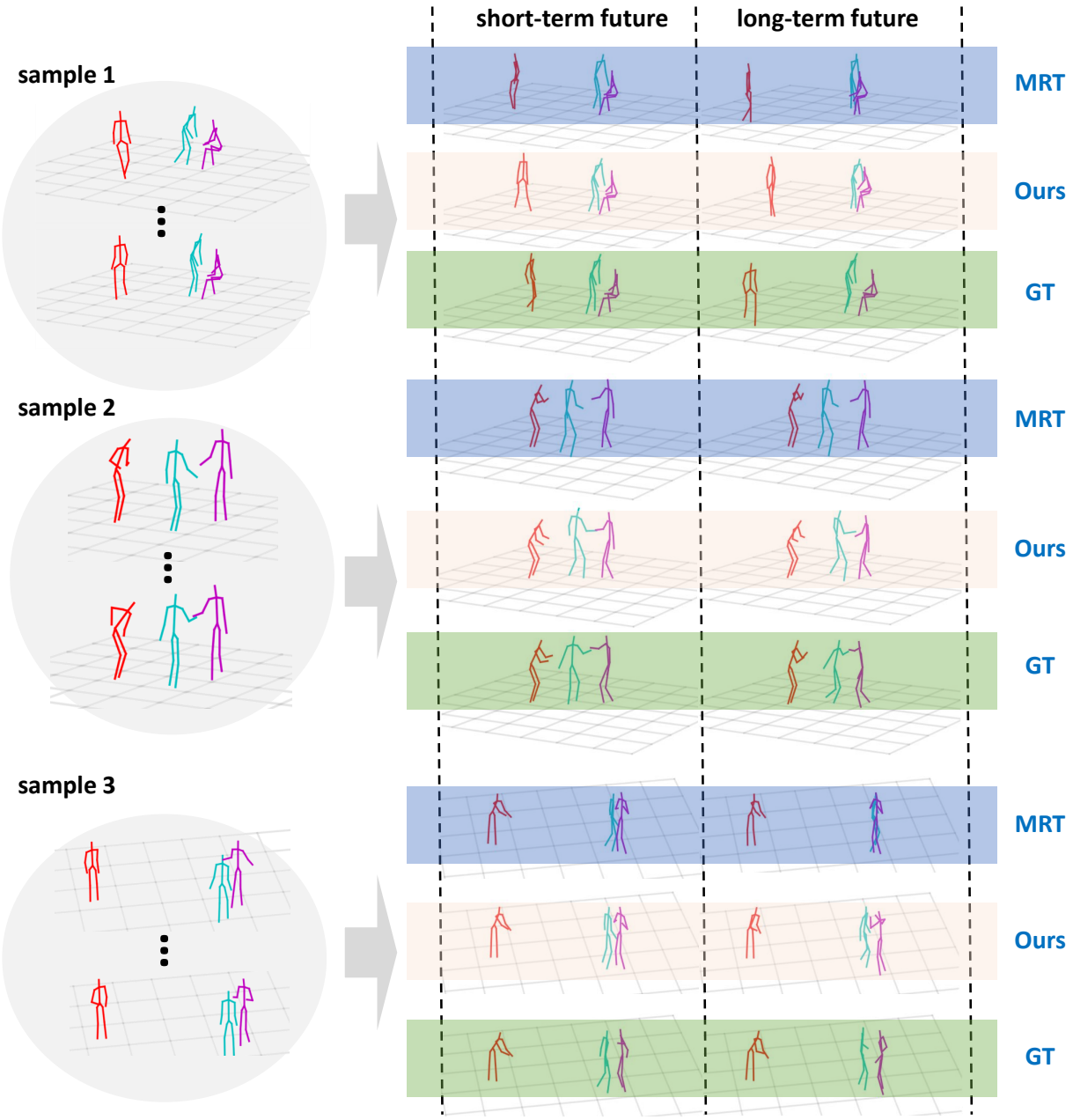

Fig. 8. Visualization of the prediction results of Multi-Range Transformer (MRT), our proposed method, and ground-truth human poses. Left poses are fed as the known sequence and poses on the right panel are inferred by the models.

(c), Person 2 and Person 3 are doing squats, so the response values of their legs is relatively high.

Visualization. Figure 8 shows the visualization results after incorporating our proposed method into MRT. It can be seen from the figure that the original MRT model tends to have abnormal deformation of human scale and unreasonable distance between multiple people in the later stage of sequence prediction. With our method, the above situation is notably alleviated. This is also consistent with the previous conclusion in Table I that MRT+PAM+GIM performs better in long-term sequence prediction.

V. Conclusions

In this paper, we propose a multi-granularity-based multi-person 3D motion prediction framework. The framework mainly includes two modules, Global Interaction Module (GIM) and the Part Attention Module (PAM). The GIM module uses the global position and motion of multiple people to model the person-level interactions, and the PAM module uses the motion of the human joints to model the joint-level interactions. Finally, the person-level and part-level interactions are fused to obtain a multi-granularity interaction score. This framework makes up for the problem of only modeling global person-level interactions but ignoring the interaction between joints in previous methods. On the three datasets of

CMU-Mocap, MuPoTS-3D, and 3DPW, our proposed method brings significant improvement to the baseline models DViTA and MRT.

REFERENCES

[1] K. Chen, Z. Tan, J. Lei, S. Zhang, Y. Guo, W. Zhang, and S. Hu, "Choreomaster: choreography-oriented music-driven dance synthesis," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 145:1–145:13, 2021.

[2] R. Li, S. Yang, D. A. Ross, and A. Kanazawa, "AI choreographer: Music conditioned 3d dance generation with AIST++," in *ICCV*, pp. 13381–13392, 2021.

[3] Z. Ye, H. Wu, J. Jia, Y. Bu, W. Chen, F. Meng, and Y. Wang, "Choreonet: Towards music to dance synthesis with choreographic action unit," in *ACM Multimedia*, pp. 744–752, 2020.

[4] X. Zhang, H. Chen, W. Yang, W. Jin, and W. Zhu, "Pedestrian path prediction for autonomous driving at un-signalized crosswalk using w/cdm and msfm," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 5, pp. 3025–3037, 2020.

[5] K. Li, M. Shan, K. Narula, S. Worrall, and E. M. Nebot, "Socially aware crowd navigation with multimodal pedestrian trajectory prediction for autonomous vehicles," in *ITSC*, pp. 1–8, 2020.

[6] W. Chen, F. Wang, and H. Sun, "S2tnet: Spatio-temporal transformer networks for trajectory prediction in autonomous driving," in *ACML*, pp. 454–469, 2021.

[7] H. S. Koppula and A. Saxena, "Anticipating human activities for reactive robotic response," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013.

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

[9] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now," in *ICCV*, pp. 5932–5941, 2019.

[10] C. Yang, Z. Wang, X. Zhu, C. Huang, J. Shi, and D. Lin, "Pose guided human video generation," in *ECCV*, pp. 204–219, 2018.

[11] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *ICCV*, pp. 4346–4354, 2015.

[12] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-rnn: Deep learning on spatio-temporal graphs," in *CVPR*, pp. 5308–5317, 2016.

[13] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *CVPR*, pp. 4674–4683, 2017.

[14] J. Bütepage, M. J. Black, D. Kragic, and H. Kjellström, "Deep representation learning for human motion prediction and classification," in *CVPR*, pp. 1591–1599, 2017.

[15] C. Li, Z. Zhang, W. S. Lee, and G. H. Lee, "Convolutional sequence to sequence model for human dynamics," in *CVPR*, pp. 5226–5234, 2018.

[16] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning trajectory dependencies for human motion prediction," in *ICCV*, pp. 9488–9496, 2019.

[17] 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.

[18] L. Dang, Y. Nie, C. Long, Q. Zhang, and G. Li, "MSR-GCN: multi-scale residual graph convolution networks for human motion prediction," in *ICCV*, pp. 11447–11456, 2021.

[19] Q. Men, E. S. Ho, H. P. Shum, and H. Leung, "A quadruple diffusion convolutional recurrent network for human motion prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 9, pp. 3417–3432, 2020.

[20] C. Liu and Y. Mu, "Searching motion graphs for human motion synthesis," in *ACM Multimedia*, pp. 871–879, 2021.

[21] R. Zhao, H. Su, and Q. Ji, "Bayesian adversarial human motion synthesis," in *CVPR*, 2020.

[22] 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.

[23] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. S. Torr, and M. Chandraker, "DESIRE: distant future prediction in dynamic scenes with interacting agents," in *CVPR*, pp. 2165–2174, 2017.

[24] N. Deo and M. M. Trivedi, "Multi-modal trajectory prediction of surrounding vehicles with maneuver based lstms," in *Proceedings of the IEEE Intelligent Vehicles Symposium*, pp. 1179–1184, 2018.

[25] A. Sadeghian, F. Legros, M. Voisin, R. Vesel, A. Alahi, and S. Savarese, "Car-net: Clairvoyant attentive recurrent network," in *ECCV*, pp. 162–180, 2018.

[26] K. Chen, X. Song, and X. Ren, "Pedestrian trajectory prediction in heterogeneous traffic using pose keypoints-based convolutional encoder-decoder network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1764–1775, 2020.

[27] T. Zhao, Y. Xu, M. Monfort, W. Choi, C. L. Baker, Y. Zhao, Y. Wang, and Y. N. Wu, "Multi-agent tensor fusion for contextual trajectory prediction," in *CVPR*, pp. 12126–12134, 2019.

[28] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, "Sophie: An attentive GAN for predicting paths compliant to social and physical constraints," in *CVPR*, pp. 1349–1358, 2019.

[29] A. D. Berenguer, M. Alioscha-Perez, M. C. Oveneke, and H. Sahli, "Context-aware human trajectories prediction via latent variational model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1876–1889, 2020.

[30] V. Kosaraju, A. Sadeghian, R. Martín-Martín, I. D. Reid, H. Rezatofighi, and S. Savarese, "Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks," in *NeurIPS*, pp. 137–146, 2019.

[31] N. Shafiee, T. Padir, and E. Elhamifar, "Introvert: Human trajectory prediction via conditional 3d attention," in *CVPR*, pp. 16815–16825, 2021.

[32] H. Sun, Z. Zhao, Z. Yin, and Z. He, "Reciprocal twin networks for pedestrian motion learning and future path prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1483–1497, 2021.

[33] V. Adeli, M. Ehsanpour, I. D. Reid, J. C. Niebles, S. Savarese, E. Adeli, and H. Rezatofighi, "Tripod: Human trajectory and pose dynamics forecasting in the wild," in *ICCV*, pp. 13370–13380, 2021.

[34] A. Saadat, N. Fathi, and S. Saadatanejad, "Towards human pose prediction using the encoder-decoder lstm," in *ICCV Workshops*, 2021.

[35] B. Parsaeifard, S. Saadatnejad, Y. Liu, T. Mordan, and A. Alahi, "Learning decoupled representations for human pose forecasting," in *ICCV Workshops*, pp. 2294–2303, 2021.

[36] J. Wang, H. Xu, M. Narasimhan, and X. Wang, "Multi-person 3d motion prediction with multi-range transformers," *NeurIPS*, pp. 6036–6049, 2021.

[37] L. Gui, Y. Wang, X. Liang, and J. M. F. Moura, "Adversarial geometry-aware human motion prediction," in *ECCV*, pp. 823–842, 2018.

[38] S. Yan, Z. Li, Y. Xiong, H. Yan, and D. Lin, "Convolutional sequence generation for skeleton-based action synthesis," in *ICCV*, pp. 4393–4401, 2019.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, pp. 5998–6008, 2017.

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

[41] A. Graves, "Generating sequences with recurrent neural networks," *CoRR*, vol. abs/1308.0850, 2013.

[42] N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised learning of video representations using lstms," in *ICML*, pp. 843–852, 2015.

[43] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *NeurIPS*, pp. 3104–3112, 2014.

[44] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *EMNLP*, pp. 1412–1421, 2015.

[45] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko, "Long-term recurrent convolutional networks for visual recognition and description," in *CVPR*, pp. 2625–2634, 2015.

[46] Y. Wang, N. Xu, A.-A. Liu, W. Li, and Y. Zhang, "High-order interaction learning for image captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4417–4430, 2022.

[47] Q. Feng, Y. Wu, H. Fan, C. Yan, M. Xu, and Y. Yang, "Cascaded revision network for novel object captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3413–3421, 2020.

[48] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," in *ICCV*, pp. 2407–2415, 2015.

[49] A.-A. Liu, Y. Zhai, N. Xu, W. Nie, W. Li, and Y. Zhang, "Region-aware image captioning via interaction learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3685–3696, 2022.

[50] A. Wu, Y. Han, Y. Yang, Q. Hu, and F. Wu, "Convolutional reconstruction-to-sequence for video captioning," *IEEE Transactions on*

Circuits and Systems for Video Technology, vol. 30, no. 11, pp. 4299–4308, 2020.

[51] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," in *CVPR*, pp. 4594–4602, 2016.

[52] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *CoRR*, vol. abs/1411.2539, 2014.

[53] L. Yao, A. Torabi, K. Cho, N. Ballas, C. J. Pal, H. Larochelle, and A. C. Courville, "Video description generation incorporating spatio-temporal features and a soft-attention mechanism," *CoRR*, vol. abs/1502.08029, 2015.

[54] T. Wang, H. Zheng, M. Yu, Q. Tian, and H. Hu, "Event-centric hierarchical representation for dense video captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1890–1900, 2021.

[55] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *CVPR*, pp. 2625–2634, 2015.

[56] Z. Zhang, D. Xu, W. Ouyang, and C. Tan, "Show, tell and summarize: Dense video captioning using visual cue aided sentence summarization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 9, pp. 3130–3139, 2020.

[57] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: human trajectory prediction in crowded spaces," in *CVPR*, pp. 961–971, 2016.

[58] J. Morton, T. A. Wheeler, and M. J. Kochenderfer, "Analysis of recurrent neural networks for probabilistic modeling of driver behavior," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 5, pp. 1289–1298, 2017.

[59] A. Vemula, K. Muelling, and J. Oh, "Social attention: Modeling attention in human crowds," in *ICRA*, pp. 1–7, 2018.

[60] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *NeurIPS*, pp. 3483–3491, 2015.

[61] B. Ivanovic and M. Pavone, "The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs," in *ICCV*, pp. 2375–2384, 2019.

[62] N. Rhinehart, R. McAllister, K. M. Kitani, and S. Levine, "PRECOG: prediction conditioned on goals in visual multi-agent settings," *CoRR*, vol. abs/1905.01296, 2019.

[63] Y. Liu, Q. Yan, and A. Alahi, "Social NCE: contrastive learning of socially-aware motion representations," in *ICCV*, pp. 15098–15109, 2021.

[64] K. Mangalam, H. Girase, S. Agarwal, K. Lee, E. Adeli, J. Malik, and A. Gaidon, "It is not the journey but the destination: Endpoint conditioned trajectory prediction," in *ECCV*, pp. 759–776, 2020.

[65] C. Yu, X. Ma, J. Ren, H. Zhao, and S. Yi, "Spatio-temporal graph transformer networks for pedestrian trajectory prediction," in *ECCV*, pp. 507–523, 2020.

[66] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, pp. 2672–2680, 2014.

[67] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: socially acceptable trajectories with generative adversarial networks," in *CVPR*, pp. 2255–2264, 2018.

[68] G. Chen, J. Li, J. Lu, and J. Zhou, "Human trajectory prediction via counterfactual analysis," in *ICCV*, pp. 9804–9813, 2021.

[69] H. Sun, Z. Zhao, and Z. He, "Reciprocal learning networks for human trajectory prediction," in *CVPR*, pp. 7414–7423, 2020.

[70] P. Dendorfer, S. Elflein, and L. Leal-Taixé, "MG-GAN: A multi-generator model preventing out-of-distribution samples in pedestrian trajectory prediction," in *ICCV*, pp. 13138–13147, 2021.

[71] E. Barsoum, J. Kender, and Z. Liu, "HP-GAN: probabilistic 3d human motion prediction via GAN," in *CVPR Workshops*, pp. 1418–1427, 2018.

[72] E. Corona, A. Pumarola, G. Alenya, and F. Moreno-Noguer, "Context-aware human motion prediction," in *CVPR*, 2020.

[73] 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.

[74] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *ICLR*, 2017.

[75] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI*, 2018.

[76] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *CVPR*, pp. 180–189, 2020.

[77] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *CVPR*, pp. 3595–3603, 2019.

[78] J. Zhang, Y. Jia, W. Xie, and Z. Tu, "Zoom transformer for skeleton-based group activity recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8646–8659, 2022.

[79] 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 *ECCV*, pp. 614–631, 2018.

[80] MoCap, "Carnegie-mellon mocap database. (available from: http://mocap.cs.cmu.edu/," 2012.

[81] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll, and C. Theobalt, "Single-shot multi-person 3d pose estimation from monocular rgb," in *3D Vision (3DV)*, 2018.

[82] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. S. Godisart, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic studio: A massively multiview system for social interaction capture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[83] Y. Huang, H. Bi, Z. Li, T. Mao, and Z. Wang, "STGAT: modeling spatial-temporal interactions for human trajectory prediction," in *ICCV*, pp. 6271–6280, 2019.

[84] W. Mao, M. Liu, and M. Salzmann, "History repeats itself: Human motion prediction via motion attention," in *ECCV*, pp. 474–489, 2020.

[85] V. Adeli, E. Adeli, I. Reid, J. C. Niebles, and H. Rezatofighi, "Socially and contextually aware human motion and pose forecasting," *IEEE Robotics Autom. Lett.*, vol. 5, no. 4, pp. 6033–6040, 2020.

Chenchen Liu received his B.E. degree from Jilin University, Changchun, China, in 2017. He is currently a Ph.D. candidate at the Wangxuan Institute of Computer Technology, Peking University. His research interests include single or multi-person motion prediction and synthesis, and human 3D pose estimation.

Yadong Mu is a tenured Associate Professor at Wangxuan Institute of Computer Technology, Peking University. He obtained both the B.S. and Ph.D. degrees from Peking University. Before joining Peking University, he had ever worked as research fellow at National University of Singapore, research scientist at Columbia University, researcher at Huawei Noah's Ark Lab in Hong Kong, and senior scientist at AT&T Labs. His research interest is in broad research topics in computer vision and machine learning.