Multi-Person Extreme Motion Prediction

Wen Guo[1,2*], Xiaoyu Bie[1*], Xavier Alameda-Pineda[1], Francesc Moreno-Noguer[2]
[1]Inria, Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
[2]Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain
[1]{wen.guo,xiaoyu.bie,xavier.alameda-pineda}@inria.fr, [2]fmoreno@iri.upc.edu

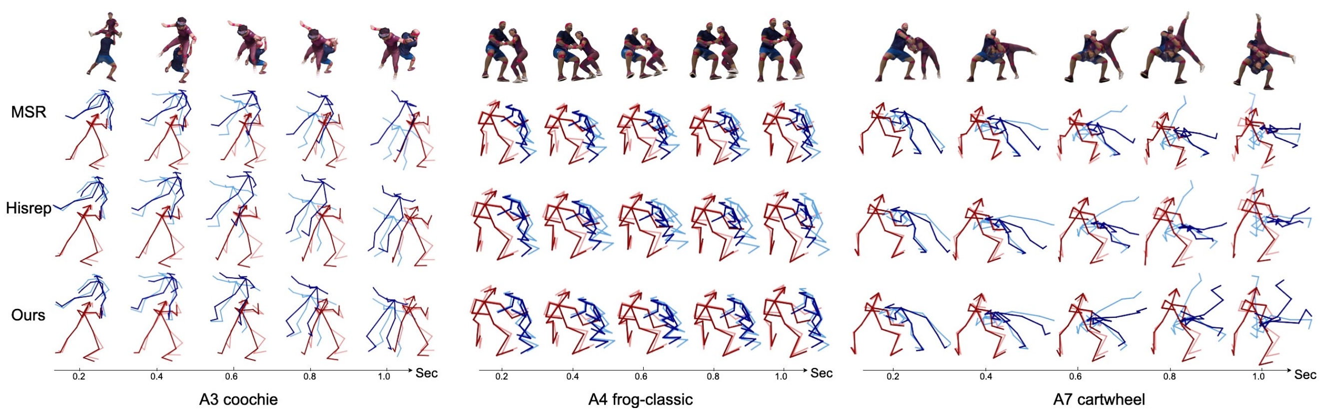

Figure 1. **Collaborative human motion prediction. 1st row:** 3D sample meshes from our ExPI Dataset (just for visualization purposes). **2nd-4th rows:** Motion prediction results by MSR [19], Hisrep [47], and our method. Dark red/blue indicate prediction results, and light red/blue are the ground truth. By exploiting the interaction information, our approach of collaborative motion prediction achieves significantly better results than methods that independently predict the motion of each person.

Abstract

Human motion prediction aims to forecast future poses given a sequence of past 3D skeletons. While this problem has recently received increasing attention, it has mostly been tackled for single humans in isolation. In this paper, we explore this problem when dealing with humans performing collaborative tasks, we seek to predict the future motion of two interacted persons given two sequences of their past skeletons. We propose a novel cross interaction attention mechanism that exploits historical information of both persons, and learns to predict cross dependencies between the two pose sequences. Since no dataset to train such interactive situations is available, we collected ExPI (Extreme Pose Interaction) dataset, a new lab-based person interaction dataset of professional dancers performing Lindy-hop dancing actions, which contains 115 sequences with 30K frames annotated with 3D body poses and shapes. We thoroughly evaluate our cross interaction network on ExPI and show that both in short- and long-term predictions, it consistently outperforms state-of-the-art methods for single-person motion prediction. Our code and dataset are available at: https://team.inria.fr/robotlearn/multi-person-extreme-motion-prediction/

1. Introduction

The goal of human motion prediction is to predict future motions from previous observations. With the successful development of deep human pose estimation from single image [9, 18, 27, 37, 51, 52, 55, 56, 58, 59, 67], motion prediction begins to draw an increasing attention [3, 8, 16, 22, 23, 26, 29, 33, 38, 43, 47, 49, 50, 60]. Most existing works formulate motion prediction as a sequence-to-sequence task, where past observations of 3D skeleton data are used to forecast future skeleton movements. A common denominator of all these approaches is that they treat each pose

*Equal contribution.

†This research was supported by ANR-3IA MIAI (ANR-19-P3IA-0003), ANR-JCJC ML3RI (ANR-19-CE33-0008-01), H2020 SPRING (funded by EC under GA #871245), by the Spanish government with the project MoHuCo PID2020-120049RB-I00 and by an Amazon Research Award. We also thank the Kinovis platform at Inria Grenoble, Laurence Boissieux and Julien Pansiot at Inria Grenoble for their help, and Nvidia for hardware donation under the Academic Hardware Grant Program. And we thank Yuming Du, Yang Xiao, Meng Guo, Anand Ballou, Louis Airale for their kind suggestions and discussions.

sequence as an independent and isolated entity: the motion predicted for one person relies solely on her/his past motion. However, in real world scenarios people interact with each other, and the motion of one person is typically dependent on or correlated with the motion of other people. Thus, we could potentially improve the performance of motion prediction by exploiting such human interaction.

Based on this intuition, in this paper we present a novel task: *collaborative motion prediction*, which aims to jointly predict the motion of two persons strongly involved in an interaction. To the best of our knowledge, previous publicly available datasets for 3D human motion prediction like 3DPW [67] and CMU-Mocap [25] that involve multiple persons only include weak human interactions, e.g., talking, shaking hands etc. Here we move a step further and analyse situations where the motion of one person is highly correlated to the other person, which is often seen in team sports or collaborative assembly tasks in factories.

With the goal to foster research on this new task, we collected the ExPI (Extreme Pose Interaction) dataset, a large dataset of professional dancers performing Lindy Hop aerial steps.[1] To perform these actions, the two dancers perform different movements that require a high level of synchronisation. These actions are composed of extreme poses and require strict and close cooperation between the two persons, which is highly suitable for the study of human interactions. Some examples of this highly interacted dataset are shown in Figure 2. Our dataset contains 115 sequences of 2 professional couples performing 16 different actions. It is recorded in a multiview motion capture studio, and the 3D poses and 3D shapes of the two persons are annotated for all the 30K frames. We have carefully created train/test splits, and proposed two different extensions of the pose evaluation metrics for collaborative motion prediction task.

To model such strong human-to-human interactions, we introduce a novel Cross-Interaction Attention (XIA) module, which is based upon a standard multi-head attention [64] and exploits historical motion data of the two persons simultaneously. For a pair of persons engaging in the same activity, XIA module extracts the spatial-temporal motion information from both persons and uses them to guide the prediction of each other.

We exhaustively evaluate our approach and compare it with state-of-the-art methods designed for single human motion prediction. Note that in our dataset of dancing actions, movements are performed at high speed. The long term predictions are very challenging in this case. Nevertheless, the results demonstrate that our approach consistently outperforms these methods by a large margin, with $10 \sim 40\%$ accuracy improvement for short (≤ 500 ms) and $5 \sim 30\%$ accuracy improvement for long term prediction (500 ms $\sim$ 1000 ms).

[1]The Lindy Hop is an African-American couple dance born in the 1930's in Harlem, New York, see [54].

Our key contributions can be summarized as follows:

- We introduce the task of collaborative motion prediction, to focus on the estimation of future poses of people in highly interactive setups.
- We collect and make publicly available ExPI, a large dataset of highly interacted extreme dancing poses, annotated with 3D joint locations and body shapes. We also define the benchmark with carefully selected train/test splits and evaluation protocols.
- We propose a method with a novel cross-interaction attention (XIA) module that exploits historical motion of two interacted persons to predict their future movements. Our model can be used as a baseline method for collaborative motion prediction.

2. Related Work

2.1. 3D Human Motion Prediction

Due to the inherent sequential structure of human motion, 3D human motion prediction has been mostly addressed with recurrent models. For instance, Fragkiadaki *et al*. [22] propose an encoder-decoder framework to embed human poses and an LSTM to update the latent space and predict future motion. Jain *et al*. [33] split human body into sub-parts and forward them via structural RNNs. Martinez *et al*. [50] introduce a residual connection to model the velocities instead of the poses themselves. Interestingly, they also show that a model trained with diverse action data performs better than those trained with single actions. However, although RNNs achieve great success in motion prediction, they suffer from containing the entire history with a fixed-size hidden state and tend to converge to a static pose. Some works alleviate this problem by using RNN variants [15, 45], sliding windows [10, 11], convolutional models [29, 30, 39] or adversarial training [26].

Since human body is a non-rigid and structured data, directly encoding the whole body into a compact latent embedding will neglect the spatial connectivity of human joints. To this end, Mao *et al*. [49] introduces a feed forward graph convolutional network (GCN) [35, 65] with learnable adjacent matrix. This approach was later boosted with self-attention on an entire piece of historical information [47] or a selection of them [41]. Recently, GCN based methods are further developed by leveraging multi-scale supervision [19], space-time-separable graph [63], and contextual information [1, 2]. In terms of GCN design, Cui *et al*. [17] argue that training the adjacent matrix from scratch ignores the natural connections of human joints, and propose to use a semi-constrained adjacent matrix. Li *et al*. [42] combine a graph scattering network with a hand-crafted adjacent matrix. Other works also exploit the use of transformers [64] to replace GCN in human motion prediction [3, 12].

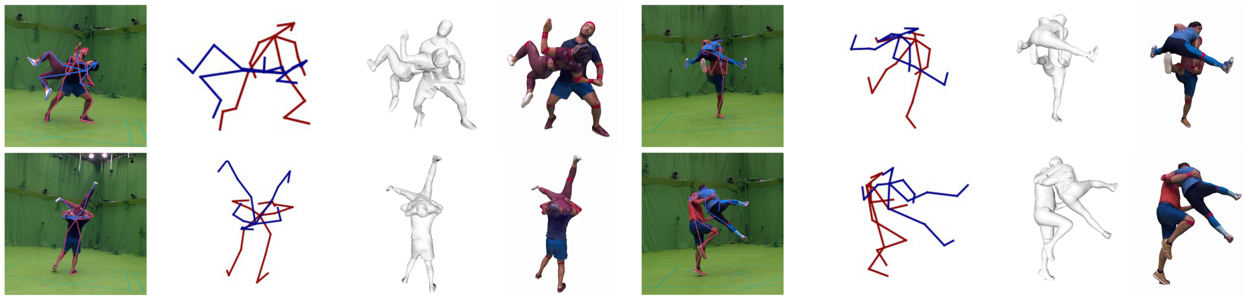

Figure 2. Some samples of the ExPI dataset: RGB image with projected 2D skeletons, 3D pose, mesh and textured mesh.

Considering that human actions are essentially stochastic in the future, some works leverage on generative models (*e.g*. VAEs and GANs) [5, 6, 13, 48, 57, 70, 71, 73]. Nevertheless, although these models can generate diverse future motions, their prediction accuracy still needs to be further improved when compared to deterministic models.

2.2. Contextual Information in Human Interaction

Humans never live in isolation, but perform continuous interactions with other people and objects. Modeling such interactions and the contextual information has been proven to be effective in the topic of 3D human pose estimation [27, 28, 34, 68, 69, 72]. Contextual information has also been shown to be beneficial in predicting human path trajectories. For this purpose, recent works explore the use of multi-agent context with social pooling mechanisms [4], tree-based role alignment [20], soft attention mechanisms [66] and graph attention networks [31, 36, 40].

Unlike the trajectory forecasting problem that focuses on a single center point, motion prediction aims at predicting the dynamics of the whole human skeleton. Incorporating contextual information in such a situation is still much unexplored. Corona *et al*. [16] expand the use of contextual information into motion prediction with a semantic-graph model, but only weak human-to-human or human-to-object correlations are modeled. Cao *et al*. [14] involve scene context information into the motion prediction framework, but without human-to-human interaction. More recently, Adeli *et al*. [1, 2] develop a social context aware motion prediction framework, where interactions between humans and objects are modeled either with a social pooling [1] or with a graph attention network [2]. However, they only study in 2D space [7] or with weak human interactions [67]. Since in this dataset [67], most of the actions involve weak interactions like shaking hands or walking together. In any event, none of these papers explores the situation we contemplate in this paper, in which humans do perform highly interactive actions.

2.3. Datasets

Using deep learning methods to study 3D human pose tasks relies on high-quality datasets. Most previous 3D human datasets are single person [32, 46, 62] or made of pseudo 3D poses [53, 67]. Other datasets which contain lab-based 3D data usually do not have close interactions [25, 44, 53, 61]. Recently, some works start to focus on the importance of context information and propose datasets to model interaction of synthetic persons with scenes [14]. Furthermore, Fieraru *et al*. [21] created a dataset of human interaction with a contact-detection-aware framework, but this dataset just contains several daily scenarios with mild human interactions and it is not released yet at the time of our submission. Thus, we believe the ExPI dataset we present here, where the actions of people are highly correlated, fills an empty space in the current datasets of human 3D pose/motion.

3. Problem Formulation

As discussed in the introduction, the task of single person human motion prediction is well established. It is defined as learning a mapping $\mathcal{M} : P_{t_{\mathrm{I}}:t-1} \rightarrow P_{t:t_{\mathrm{E}}}$ to estimate the future movements $P_{t:t_{\mathrm{E}}}$ from the previous observation $P_{t_{\mathrm{I}}:t-1}$, where t_{I} (t_{E}) denotes the initial (ending) frame of a sequence, and P_t denotes the pose at time t.

In this work, we extend the problem formulation to collaborative motion prediction of two interacted persons. While our formulation is general and could work for any kind of interactions, for the sake of consistency throughout the paper, we will denote by ℓ and f variables corresponding to the leader and the follower respectively (see Section 4 on the dataset description). Therefore, the collaborative motion prediction task is defined as learning a mapping:

$$\mathcal{M}_{\mathrm{C}} : P^{\ell}_{t_{\mathrm{I}}:t-1}, P^{f}_{t_{\mathrm{I}}:t-1} \rightarrow P^{\ell}_{t:t_{\mathrm{E}}}, P^{f}_{t:t_{\mathrm{E}}}. \tag{1}$$

Since the two persons are involved in the same interaction, we believe it is possible to better predict the motion of a person by exploit the pose information of her/his interacted partner. From now on, we will use $P^c_t = [P^l_t, P^f_t]$ to denote the joint pose of the couple (two actors) at time t, and P_t to denote either of them.

In the following parts of the paper, we will provide an experimental framework for the collaborative motion prediction task, consisting of a dataset and evaluation metrics,

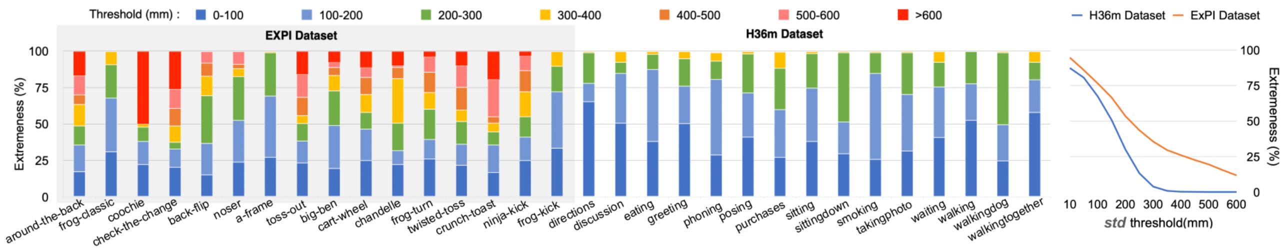

Figure 3. Extremeness. **Left**: percentage of joints whose *std* is among a certain threshold (in different colors), for different actions. Actions with more red colors are more extreme. **Right**: percentage of joints whose *std* is beyond a certain threshold.

to foster research in this direction. And we will also introduce our proposed method for this task.

4. The Extreme Pose Interaction Dataset

We present the Extreme Pose Interaction (ExPI) Dataset, a new person interaction dataset of Lindy Hop dancing actions. In Lindy Hop, the two dancers are called *leader* and *follower*.[2] We recorded 2 couples of dancers in a multi-camera setup equipped also with a motion-capture system. In this section we firstly describe the recording procedure, then give a comprehensive analysis of our dataset.

4.1. Dataset Overview

Dataset Structure. 16 different actions are performed in ExPI dataset, some by the 2 couples of dancers, some by only one of the couples. Each action was repeated five times to account for variability. More precisely, for each recorded sequence, ExPI provides: (i) Multi-view videos at 25FPS from all the cameras in the recording setup; (ii) Mocap data (3D position of 18 joints for each person) at 25FPS synchronised with the videos.; (iii) camera calibration information; and (iv) 3D shapes as textured meshes for each frame. Overall, the dataset contains 115 sequences with 30k visual frames for each view point and 60k 3D instances annotated.

Dataset Collection and Post-processing. The data were collected in a multi-camera platform equipped with 68 synchronised and calibrated color cameras and a motion capture system with 20 mocap cameras.[3] When collecting the motion capture data, some points are missed by the system due to occlusions or tracking losses, which is a common phenomena in lab-based interacted Mocap datasets [21]. To overcome this issue and ensure the quality of the data, we spent months to manually label the missing points.
More details about the data structure and data post-processing are provided in the supplementary material.

4.2. Data Analysis

Diversity. Similar to Ionescu *et al*. [32], we analyse the diversity of our dataset by checking how many *distinct* poses have been obtained. We consider two poses to be *distinct*, if at least one of the J joints for one pose P^c_m is different from the corresponding joint of the other pose P^c_n, beyond a certain tolerance τ (mm):

$$\max_{j\in[1,J]} \|P^c_{m,j} - P^c_{n,j}\| > \tau, \tag{2}$$

where $m, n \in \mathcal{D}$ denote any two poses in the dataset $\mathcal{D}$. Then we define *diversity* of the dataset as the percentage of *distinct* poses among all the poses. According to Ionescu *et al*. [32], the diversity of H3.6M[4] is 24% and 12% when setting the tolerance τ to 50 mm and 100 mm, respectively. While the diversities of ExPI for the same threshold values are 52% and 23%, which are much more diverse.

Extremeness. To measure the extremeness of a pose sequence, we first compute the standard deviation (*std*) over time for each dimension of the $xyz-$coordinate for every joint. Then, the extremeness of the joint j is defined as its maximum per-coordinate standard deviation: $\varepsilon_j = \max\{\sigma^x_j, \sigma^y_j, \sigma^z_j\}$. Finally, the extremeness of an action is evaluated by computing the percentage of joint extremeness values ε_n within various intervals $[\varepsilon_{\min}, \varepsilon_{\max}]$. Figure 3 reports the extremeness of ExPI dataset compared to H3.6M in two different ways: (i) a per-action plot reporting extremeness on various color-coded intervals (left); (ii) computing the percentage of joints more extreme than a certain *std* value(right). From both plots it is clear than the ExPI dataset is significantly more extreme than the H3.6M dataset.

5. Method

We introduce our approach for collaborative motion prediction, aiming to set the first performance baseline to help future developments.

5.1. Pipeline

The idea of our method is to learn two person-specific motion prediction mappings, and to propose a strategy to share information between these two mappings. The possibility to include information from the other person involved

[2] This is the standard gender-neutral terminology for Lindy-Hop.
[3] Kinovis https://kinovis.inria.fr/
[4] Licence for H3.6M dataset http://vision.imar.ro/human3.6m/eula.php

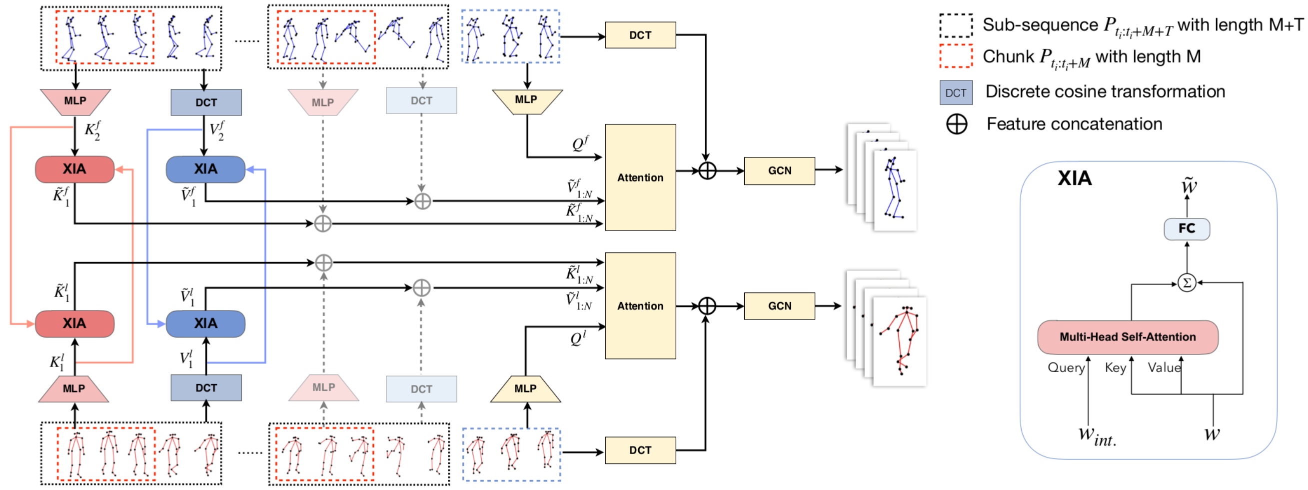

Figure 4. **Left:** Computing flow of the proposed method. Two parallel pipelines – for the leader and the follower – are implemented. The key-value pairs are refined by XIA modules (we just visualize XIA modules for the first sub-sequences, while it is the same for the following sub-sequences). **Right:** Cross-interaction attention (XIA) module. In order to refine w with the help of the corresponding interaction information $w_{int.}$, the multi-head attention is queried by $w_{int.}$ and take w as key and value.

in the interaction, should push the network to learn a better representation for motion prediction. The overall pipeline is described in Figure 4-left.

For the two single person motion prediction mappings, we draw inspiration from [47], using an attention model for leaning temporal attention w.r.t. the past motions, and a predictor based on Graph Convolutional Network (GCN) [35] to model the spatial attention among joints using an adjacency matrix. The temporal attention model aims to find the most relative sub-sequence in the past by measuring the similarity between the last observed sub-sequence and a set of past sub-sequences. In this attention model, the query Q is learnt by MLP from the last observation $P_{t-1-M:t-1}$ (blue dashed rectangle in Figure 4-left, length M). The keys K_i are learnt by MLP from the starting chunk of sub-sequences $P_{t_i:t_i+M}$(red dashed rectangles in Figure 4-left, length M). And the values V_i consist of DCT representations built from the sub-sequences $P_{t_i:t_i+M+T}$ (black dashed rectangles in Figure 4-left, length $M+T$), where t_i with $i \in \{1,\dots,N\}$ indicates the start frame of each past sub-sequence.

Training such strategy separately for each actor does not account for any interaction between the two dancing partners. To deal with this, we design a cross-interaction attention (XIA) module based on multi-head attention, to introduce guidance form the interacted person. In the next section we introduce this XIA module.

5.2. Cross-Interaction Attention (XIA)

XIA aims to share motion information between the two predictors. In particular, we denote the query and the key-value pairs for one person by Q and $\{K_i, V_i\}_{i=1}^N$ respectively, and use the superscript f and ℓ to indicate the two person, follower and leader. We naturally cast the collaborative human motion prediction task into learning how to jointly exploit the information in (K_i, V_i) when querying with Q to predict motion of each person.

Our intuition is that the pose information (key-value pairs) of one person can be used to transform the pose information of the other person for better motion prediction. We implement this intuition with the help of the proposed *cross-interaction attention module*. Such a module takes as input w and the corresponding vector from the interacted pose $w_{int.}$, and uses multi-head self attention to get the refined vector $\tilde{w}$ (see Figure 4-right):

$$\tilde{w} = \text{XIA}(w_{int.}, w) = \text{FC}(\text{MHA}(w_{int.}, w, w) + w), \quad (3)$$

where $\text{MHA}(q,k,v)$ stands for multi-head attention with query q, key k and value v, and FC indicates fully connected layers. We use different XIA modules to update keys and values mentioned in Section 5.1: in our implementation, XIA modules for keys have 8 attention heads, and XIA for values have a single attention head. Moreover, we add a skip-connection for the MHA module followed by 2 FC layers. XIA modules for leader/follower do not share weights.

The proposed XIA module is integrated at several stages of the computing flow as shown in Figure 4. More precisely, we refine all keys:

$$\tilde{K}_i^\ell = \text{XIA}(K_i^\ell, K_i^f), \qquad \tilde{K}_i^f = \text{XIA}(K_i^f, K_i^\ell), \quad (4)$$

and analogously for the values.

XIA could be potentially generalised to any number of participants by considering either several XIA modules and fusing their outcome, or performing the fusion at the input of XIA module.

Table 1. Results on common action split with the two evaluation metrics (in mm). Lower value means better performance. Obviously, our proposal outperforms all the other methods both on JME and AME.

	Action	A1 A-frame				A2 Around the back				A3 Coochie				A4 Frog classic				A5 Noser				A6 Toss Out				A7 Cartwheel				AVG			
	Time (sec)	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0
JME	Res-RNN [50]	83	141	182	236	127	224	305	433	99	177	239	350	74	135	182	250	87	152	201	271	93	166	225	321	104	189	269	414	95	169	229	325
	LTD [49]	70	125	157	189	131	242	321	426	102	194	260	357	62	117	155	197	72	131	173	231	81	151	200	280	112	223	315	442	90	169	226	303
	Hisrep [47]	52	103	139	188	96	186	256	349	57	118	167	240	45	93	131	180	51	105	149	214	61	125	176	252	71	150	222	333	62	126	177	251
	MSR [19]	56	100	**132**	**175**	102	187	256	365	65	120	166	244	50	95	127	172	54	100	138	202	70	132	182	258	82	154	218	321	69	127	174	248
	Ours	**49**	**98**	140	192	**84**	**166**	**234**	**346**	**51**	**105**	**154**	**234**	**41**	**84**	**120**	**161**	**43**	**90**	**132**	**197**	**55**	**113**	**163**	**242**	**62**	**130**	**192**	**291**	**55**	**112**	**162**	**238**
AME	Res-RNN [50]	59	102	132	167	62	112	152	229	57	102	139	215	48	85	113	157	51	90	120	167	53	94	126	183	74	131	178	265	58	102	137	197
	LTD [49]	51	92	116	132	51	91	116	**148**	43	80	103	130	38	70	89	111	39	70	90	116	42	75	94	123	52	101	139	198	45	83	107	137
	Hisrep [47]	34	69	97	130	44	84	**115**	150	32	65	91	121	27	56	82	112	28	58	85	121	34	66	88	115	42	83	120	171	34	69	97	131
	MSR [19]	41	75	99	**126**	54	96	129	180	41	74	98	135	34	61	82	106	33	59	79	**109**	42	71	93	124	57	103	146	210	43	77	104	141
	Ours	**32**	**68**	**99**	128	**41**	**82**	116	163	**29**	**58**	**84**	**116**	**24**	**50**	**73**	**96**	**24**	**51**	**75**	**109**	**31**	**62**	**86**	**114**	**41**	**81**	**115**	**160**	**32**	**65**	**93**	**127**

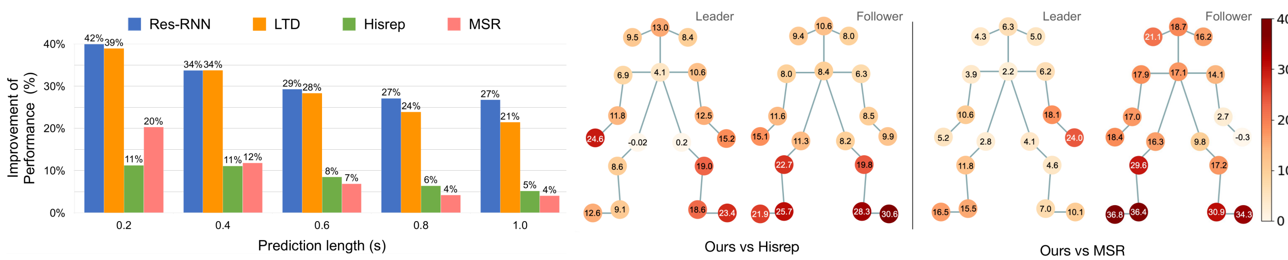

Figure 5. **Left:** Percentages of improvement of our method comparing with different state-of-the-art methods, measured by average JME error on the common action split, at different forecast time. Lower value means closer performance with our model. Our method surpasses these methods up to $10 \sim 40\%$ on short term, and $5 \sim 30\%$ on long term. **Right:** Joint-wise JME improvement(mm) of our method over Hisrep [47] and MSR [42]. Darker color means larger improvement.

5.3. Pose Normalization

Raw poses of ExPI are represented in world coordinate. Similar with single person motion prediction, we normalize the data by removing the global displacement of the poses based on a selected root joint. While our task aims at predicting not only the the distinct poses but also the relative position of the two person, so we have to normalize by the same person to keep the information of their related position. We could either normalize by leader/follower and we choose to normalize by the leader for better visualisation. Specifically, for each frame, we take the the root joint (middle of the two hips) of the leader as coordinate origin, use the root point and left hip of the leader to define x-axis, and use the neck of leader to determine XOZ plane. We normalize all the joints of both persons to this coordinate, then the pose errors can be computed directly in this coordinate. More precisely, we represent the raw poses in world coordinate as $P_w \in \{P_w^\ell, P_w^f\}$, and $T_{P^\ell_{w,t}}$ is the rigid transformation aligning the two actors to the leader's coordinate system. The normalized coordinates are thus $P_t^\ell = T_{P^\ell_{w,t}} P^\ell_{w,t}$, and $P_t^f = T_{P^\ell_{w,t}} P^f_{w,t}$. In the following P shall always represent the normalized pose unless specified otherwise.

6. Experimental Evaluation

This section describes the experimental protocol on ExPI, and discuss the results of our proposed method.

6.1. Splitting the ExPI Dataset

As described in Sect. 4.1, we record 16 actions in ExPI dataset. Seven of them are common actions (A_1 to A_7) which are performed by both of the 2 couples: we denote them as $\mathcal{A}_c^1$ performed by couple 1 and $\mathcal{A}_c^2$ by couple 2. The other actions are couple-specific, which are performed only by one couple: we denote the actions performed by couple 1 (A_8 to A_{13}) as $\mathcal{A}_u^1$, and actions by couple 2 (A_{14} to A_{16}) as $\mathcal{A}_u^2$. With these notations, we propose three data splits.

Common action split. Similar to [32], we consider the common actions performed by different couples of actors as train and test data. More precisely, $\mathcal{A}_c^2$ is the train dataset and $\mathcal{A}_c^1$ is the test dataset. Thus, train and test data contain the same actions but performed by different people.

Single action split. Similar to [22, 33], we train 7 action specific models separately for each common action, by taking one action from couple 2 as train set and the related one from couple 1 as test set.

Unseen action split. The train set is the entire set of common actions $\{\mathcal{A}_c^1, \mathcal{A}_c^2\}$. We regard the extra couple-specific actions $\{\mathcal{A}_u^1, \mathcal{A}_u^2\}$ as unseen actions and use them as our test set. Thus the train and test data contain both couples of actors, but the test actions are not used in training.

To sum up, common action split is designed for a single model on different actions, single action split is designed for action-wise models, and unseen action split focuses on testing unseen actions to measure methods generalization.

Table 2. Results on single action split with the two evaluation metrics (in mm). Lower value means better performance. Seven action-wise models are trained independently. Our method performs the best in 5 actions, and close to the best for the other 2 actions.

	Action	A1 A-frame				A2 Around the back				A3 Coochie				A4 Frog classic				A5 Noser				A6 Toss Out				A7 Cartwheel			
	Time (sec)	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0
JME	Res-RNN [50]	75	131	171	226	122	215	287	403	97	174	235	329	73	131	177	246	76	136	184	255	100	184	252	357	88	162	219	293
	LTD [49]	70	126	155	183	131	243	312	415	102	194	252	338	62	117	**153**	203	71	131	171	231	81	151	199	299	112	223	306	411
	Hisrep [47]	66	118	153	190	128	231	308	417	74	143	205	295	64	120	159	191	63	121	166	227	90	168	232	312	88	166	232	332
	MSR [19]	**64**	**108**	**136**	**171**	119	210	282	385	79	144	189	265	**59**	**103**	**134**	**173**	65	118	162	225	86	151	201	283	96	178	255	362
	Ours	**64**	120	160	199	**109**	**200**	**275**	**381**	**59**	**117**	**174**	**277**	60	116	162	209	**53**	**106**	**152**	**221**	**65**	**122**	**166**	**223**	**74**	**144**	**203**	**301**
AME	Res-RNN [50]	56	99	129	163	61	110	150	229	53	96	131	188	46	81	106	142	44	79	106	147	53	100	162	176	70	133	163	198
	LTD [49]	51	93	114	127	**51**	**91**	**116**	**162**	43	80	100	**126**	38	70	**88**	118	39	70	90	125	42	75	**93**	123	52	101	137	188
	Hisrep [47]	45	83	106	118	57	102	135	178	39	72	100	132	41	77	103	119	35	70	97	125	46	82	107	137	48	90	121	169
	MSR [19]	46	79	**98**	**118**	60	107	141	192	48	86	111	150	39	**68**	**88**	**111**	39	69	91	**121**	55	93	117	156	66	118	163	222
	Ours	**43**	**84**	115	131	53	99	136	185	**35**	**68**	**98**	140	**37**	74	106	128	**29**	**59**	**86**	125	**39**	**72**	94	**119**	**43**	**82**	**112**	**152**

6.2. Evaluation Metrics

The most common metric for evaluating 3D joint position in pose estimation and motion prediction tasks is the mean per joint position error MPJPE$(P, G) = \frac{1}{J}\sum_{j=1}^{J}\|P_j - G_j\|_2$, where J is the number of joints, P_j and G_j are the estimated and ground truth position of joint j. Based on MPJPE, we propose two different metrics to evaluate the multi-person motion task.

Joint mean error (JME): We Propose *Joint Mean per joint position Error* to measure poses of different persons in a same coordinate, and denote it as JME for simplicity:

$$\text{JME}(P, G) = \text{MPJPE}(P, G), \tag{5}$$

where P and G are normalized (see Section 5.3) prediction and ground truth. JME provides an overall idea for the performance of collaborative motion prediction by considering the two interacted persons jointly as a whole, measuring both the error of poses and the error of their relative position.

Aligned mean error (AME): We propose *Aligned Mean per joint position Error* to measure pure pose error without the position bias. We first erase the errors on the relative position between the two persons by normalizing the poses independently to obtain $\hat{P}, \hat{G}$. However the precision of $\hat{P}$ is importantly influenced by the joints that are used to determine the coordinate (hips and back). To mitigate this effect, we compute the best rigid alignment T_A between the estimated pose and the ground-truth using Procrustes analysis [24]:

$$\text{AME}(P, G) = \text{MPJPE}(T_A(\hat{P}, \hat{G}), \hat{G}), \tag{6}$$

where $\hat{P} \in [\hat{P}^{\ell}, \hat{P}^{f}]$ are independently normalized predictions $\hat{P}^{\ell}_t = T_{P^{\ell}_t} P^{\ell}_t$ and $\hat{P}^{f}_t = T_{P^{f}_t} P^{f}_t$, and T_P is the normalisation transformation computed from the pose P as defined in Section 5.3. The same calculation is done for the ground truth $\hat{G}$. This normalization is only used for evaluation purpose.

6.3. Implementation Details

Since this is the first time the collaborative motion prediction task is presented in the literature, there are no available methods to compare with. Thus we choose 4 code-released state-of-the-art methods of single person motion prediction [19, 47, 49, 50], and implement their released codes[5] on ExPI dataset. For fair comparison, all these models are trained with 50 frames of input, train/test for the leader and the follower separately.

We train our model for 25 epochs and calculate the average MPJPE loss of 10 predicted frames. As the data is normalized by the leader, the corresponding branch converges faster, so we compensate by exponentially down-weighting the loss of the leader with the number of epochs ϵ, using the loss function: $\mathcal{L} = \mathcal{L}_f + 10^{-\epsilon}\mathcal{L}_l$,.

When predicting longer horizons, we use the predicted motion as input to predict future motion. Inspired by [47], we take 64 sub-sequences for each sequence to reduce the variance of the test results. Overall, we have $7k$ and $2.3k$ sub-sequences for training and testing respectively in the common action split and the single action split, and $12k$ / $2.9k$ training/testing samples in the unseen action split.

6.4. Results and Discussion

Common action split. Table 1 reports the results on the common action split. We observe that our proposed method outperforms other methods systematically almost for all actions, in all metrics and for different testing time. In Figure 5-left we calculate the percentage of improvement of our method compared with the state-of-the-art methods, and find that we significantly surpass these methods up to $10 \sim 40\%$ on short term, and $5 \sim 30\%$ on long term. We further compare our per-joint results with Hisrep [47] and MSR [19] in Figure 5-right, and observe that our proposed method gets better results on almost on all the joints. More importantly, the keypoints of the limbs (joints of arms and legs) are improved largely. This is reasonable as interaction between persons comes mostly through the limbs, while joints on the torso have little influence on it. So our

[5] All the codes we use are under MIT license.

Table 3. Action-wise results on unseen action split with the two evaluation metrics (in mm). Lower value means better performance. Our method still performs the best on most of the unseen actions and on the average result.

	Action	A8			A9			A10			A11			A12			A13			A14			A15			A16			AVG		
	Time (sec)	0.2	0.6	1.0	0.2	0.6	1.0	0.2	0.6	1.0	0.2	0.6	1.0	0.2	0.6	1.0	0.2	0.6	1.0	0.2	0.6	1.0	0.2	0.6	1.0	0.2	0.6	1.0	0.2	0.6	1.0
JME	Res-RNN [50]	239	312	371	193	256	303	189	257	310	305	425	520	215	289	348	165	214	252	214	293	357	149	187	210	167	226	277	204	273	327
	LTD [49]	239	324	394	175	226	259	148	191	220	176	240	286	143	178	192	146	193	226	252	333	387	174	228	264	139	184	217	177	233	272
	Hisrep [47]	195	**283**	**358**	121	169	206	92	129	**160**	129	193	245	**80**	**104**	121	112	154	187	157	219	257	134	190	233	**96**	**146**	**187**	124	176	**218**
	MSR [19]	297	368	451	250	317	395	173	231	303	241	335	416	280	345	449	158	195	246	173	231	289	**95**	**117**	**134**	153	216	268	202	261	327
	Ours	**191**	287	377	**118**	**165**	**203**	**91**	**129**	162	**122**	**183**	**232**	81	107	**128**	**106**	**150**	**185**	**156**	**216**	**256**	126	175	213	**96**	152	205	**121**	**174**	**218**
AME	Res-RNN. [50]	124	165	195	125	157	181	131	166	189	148	198	240	149	169	192	102	128	147	181	237	279	100	129	144	93	124	147	128	164	190
	LTD [49]	**95**	**123**	**146**	85	106	116	74	91	101	86	115	137	98	125	134	85	110	124	106	136	155	91	119	135	72	96	116	88	113	129
	Hisrep [47]	101	144	176	61	82	94	**49**	**67**	**80**	73	105	129	**53**	**73**	**86**	64	89	**104**	86	120	**142**	73	104	128	54	82	**104**	68	96	**116**
	MSR [19]	377	463	315	360	467	308	260	276	212	158	191	211	524	699	344	212	245	167	262	232	230	**67**	**86**	**98**	116	133	142	258	308	225
	Ours	**95**	137	171	**58**	**80**	**93**	51	70	84	**70**	**105**	**134**	**53**	**73**	88	**63**	**88**	**104**	**82**	**116**	**142**	69	97	120	**52**	**79**	**104**	**66**	**94**	**116**

Table 4. Ablations. *'mix /cat /sep'* use the single person motion prediction model (Hisrep [47]) for multi-person by: mixing two poses together / concatenate two poses as a single vector / train two person-specific models. *'w.o. XIA'* indicates training leader and follower in parallel using our defined loss without XIA module; *'XIA kqv / kq / kv / v'* use XIA module to update key, value and query of the temporal attention, or just some of them.

	JME					AME				
Time (sec)	0.2	0.4	0.6	0.8	1.0	0.2	0.4	0.6	0.8	1.0
mix	69	132	185	233	271	41	77	104	126	142
cat	61	123	176	223	262	37	71	99	121	138
sep	62	126	177	218	251	34	69	97	116	131
w.o. XIA	58	120	174	217	249	33	68	98	118	131
XIA kq	58	118	169	211	245	33	67	95	114	128
XIA kqv	57	117	170	215	251	32	65	95	116	131
XIA v	56	116	168	210	244	32	66	94	113	127
XIA kv	**55**	**112**	**162**	**204**	**238**	**32**	**65**	**93**	**112**	**127**

cross-interaction attention is able to improve the accuracy on the limbs more than on the torso. We could also notice the large improvement on the feet of the follower which usually fly in the air, indicating that our method works even better for these extreme high dynamic joints.

Single action split and unseen action split. We also reported our proposed method by reporting the results on single action split and unseen action split. For single action split, XIA outperforms the state-of-the-art methods also on action-specific models, as shown in Table 2. Interestingly, we observe that the performance on single action split is worse than the corresponding results on common action split, meaning that training on different actions helps regularising the network for this very challenging collaborative extreme motion prediction task. Regarding unseen action split shown in Table 3, we can see that XIA still outperforms the state-of-the-art methods on most of the actions, demonstrating the robustness of our method.

Qualitative results. Figure 1 shows some example of our visualisation results compared to Hisrep *et al.* [47], MSR [19] and the ground truth, on the common action split. We can see that the poses estimated by our method are much closer to the ground truth than the other methods, and it works well even on some extreme actions where other methods totally fail (Figure 1-right). More qualitative examples could be found in the supplementary material.

Ablation study. Taking Hisrep [47] as example, we first tried 3 different ways of training the single-person motion prediction models on our multi-person dataset: (i) 'mix': train a single model using data of the two poses $\{P^l, P^f\}$; (ii) 'cat': concatenate the two poses as a single input vector $[P^l, P^f]$; (iii) 'sep': train two person-specific models for P^l and P^f. Since 'sep' gives best performance, all the state-of-the-art methods reported above in this paper is using this setting. As for our collaborative motion prediction model, we report performances of several different design choices of our model. We found that updating the key and values of the temporal attention using our XIA module provide the best results. We demonstrate the interest of the design of our method as the proposed one is the best in performance and our method significantly improves all the single-person motion prediction results.

Limitations. Collecting clean and reusable 3D pose data requires specific equipment and recording extreme poses requires actors with specific skills, thus ExPI is rare and difficult to reproduce/extend. This is clearly a limitation in the era of data-hungry deep learning architectures. Besides, predicting very long future (beyond 2s for example) is still an open problem, specially for the fast movements of ExPI.

7. Conclusion

Current motion prediction methods are restricted to single person. We move beyond existing approaches for 3D human motion prediction by considering a scenario with two persons performing highly interactive activities. We collected a new dataset called ExPI of professional actors performing dancing actions. ExPI is annotated with sequences of 3D body poses and shapes, opening the door to not only being applied for interactive motion prediction but also for single-frame pose estimation or multi-view 3D reconstruction. In order to learn the interacted motion dynamics, we introduce a baseline method trained with ExPI that exploits historical information of both people in an attention-like fashion. Results of our method show consistent improvement compared to methods that independently predict the motion of each person.

References

[1] Vida Adeli, Ehsan Adeli, Ian Reid, Juan Carlos Niebles, and Hamid Rezatofighi. Socially and contextually aware human motion and pose forecasting. *IEEE Robotics and Automation Letters*, 5(4):6033–6040, 2020. 2.1, 2.2

[2] Vida Adeli, Mahsa Ehsanpour, Ian Reid, Juan Carlos Niebles, Silvio Savarese, Ehsan Adeli, and Hamid Rezatofighi. Tripod: Human trajectory and pose dynamics forecasting in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13390–13400, October 2021. 2.1, 2.2

[3] Emre Aksan, Peng Cao, Manuel Kaufmann, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. *arXiv e-prints*, pages arXiv–2004, 2020. 1, 2.1

[4] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016. 2.2

[5] Sadegh Aliakbarian, Fatemeh Saleh, Lars Petersson, Stephen Gould, and Mathieu Salzmann. Contextually plausible and diverse 3d human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11333–11342, 2021. 2.1

[6] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5223–5232, 2020. 2.1

[7] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5167–5176, 2018. 2.2

[8] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1418–1427, 2018. 1

[9] Abdallah Benzine, Bertrand Luvison, Quoc Cuong Pham, and Catherine Achard. Deep, robust and single shot 3d multi-person human pose estimation from monocular images. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 584–588. IEEE, 2019. 1

[10] Judith Butepage, Michael J Black, Danica Kragic, and Hedvig Kjellstrom. Deep representation learning for human motion prediction and classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6158–6166, 2017. 2.1

[11] Judith Bütepage, Hedvig Kjellström, and Danica Kragic. Anticipating many futures: Online human motion prediction and generation for human-robot interaction. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 4563–4570. IEEE, 2018. 2.1

[12] Yujun Cai, Lin Huang, Yiwei Wang, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Xu Yang, Yiheng Zhu, Xiaohui Shen, et al. Learning progressive joint propagation for human motion prediction. In *European Conference on Computer Vision*, pages 226–242. Springer, 2020. 2.1

[13] Yujun Cai, Yiwei Wang, Yiheng Zhu, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Chuanxia Zheng, Sijie Yan, Henghui Ding, et al. A unified 3d human motion synthesis model via conditional variational auto-encoder. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11645–11655, 2021. 2.1

[14] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *European Conference on Computer Vision*, pages 387–404. Springer, 2020. 2.2, 2.3

[15] Hsu-kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1423–1432. IEEE, 2019. 2.1

[16] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6992–7001, 2020. 1, 2.2

[17] Qiongjie Cui, Huaijiang Sun, and Fei Yang. Learning dynamic relationships for 3d human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6519–6527, 2020. 2.1

[18] Rishabh Dabral, Nitesh B Gundavarapu, Rahul Mitra, Abhishek Sharma, Ganesh Ramakrishnan, and Arjun Jain. Multi-person 3d human pose estimation from monocular images. In *2019 International Conference on 3D Vision (3DV)*, pages 405–414. IEEE, 2019. 1

[19] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11467–11476, October 2021. 1, 2.1, 1, 2, 6.3, 6.4, 3, 6.4

[20] Panna Felsen, Patrick Lucey, and Sujoy Ganguly. Where will they go? predicting fine-grained adversarial multi-agent motion using conditional variational autoencoders. In *Proceedings of the European conference on computer vision (ECCV)*, pages 732–747, 2018. 2.2

[21] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7214–7223, 2020. 2.3, 4.1

[22] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4346–4354, 2015. 1, 2.1, 6.1

[23] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. Learning human motion models for long-term predictions. In *2017 International Conference on 3D Vision (3DV)*, pages 458–466. IEEE, 2017. 1

[24] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. 6.2

[25] CMU graphics lab. Cmu graphics lab motion capture database., 2009. http://mocap.cs.cmu.edu//. 1, 2.3

[26] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José MF Moura. Adversarial geometry-aware human motion prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 786–803, 2018. 1, 2.1

[27] Wen Guo, Enric Corona, Francesc Moreno-Noguer, and Xavier Alameda-Pineda. Pi-net: Pose interacting network for multi-person monocular 3d pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2796–2806, 2021. 1, 2.2

[28] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2282–2292, 2019. 2.2

[29] Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7134–7143, 2019. 1, 2.1

[30] Daniel Holden, Jun Saito, Taku Komura, and Thomas Joyce. Learning motion manifolds with convolutional autoencoders. In *SIGGRAPH Asia 2015 Technical Briefs*, pages 1–4. 2015. 2.1

[31] Yingfan Huang, HuiKun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6272–6281, 2019. 2.2

[32] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 2.3, 4.2, 4.2, 6.1

[33] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 5308–5317, 2016. 1, 2.1, 6.1

[34] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2020. 2.2

[35] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. 2.1, 5.1

[36] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *Advances in Neural Information Processing Systems*, 32:137–146, 2019. 2.2

[37] Laxman Kumarapu and Prerana Mukherjee. Animepose: Multi-person 3d pose estimation and animation. *Pattern Recognit. Lett.*, 147:16–24, 2021. 1

[38] Anliang Li, Shuang Wang, Wenzhu Li, Shengnan Liu, and Siyuan Zhang. Predicting human mobility with federated learning. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, pages 441–444, 2020. 1

[39] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5226–5234, 2018. 2.1

[40] Jiachen Li, Hengbo Ma, Zhihao Zhang, Jinning Li, and Masayoshi Tomizuka. Spatio-temporal graph dual-attention network for multi-agent prediction and tracking. *arXiv preprint arXiv:2102.09117*, 2021. 2.2

[41] Jiachen Li, Fan Yang, Hengbo Ma, Srikanth Malla, Masayoshi Tomizuka, and Chiho Choi. Rain: Reinforced hybrid attention inference network for motion forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16096–16106, October 2021. 2.1

[42] Maosen Li, Siheng Chen, Zihui Liu, Zijing Zhang, Lingxi Xie, Qi Tian, and Ya Zhang. Skeleton graph scattering networks for 3d skeleton-based human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 854–864, October 2021. 2.1, 5

[43] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yan-Feng Wang, and Cewu Lu. Transferable interactiveness prior for human-object interaction detection. *arXiv preprint arXiv:1811.08264*, 2018. 1

[44] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019. 2.3

[45] Zhenguang Liu, Shuang Wu, Shuyuan Jin, Qi Liu, Shijian Lu, Roger Zimmermann, and Li Cheng. Towards natural and accurate future motion prediction of humans and animals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10004–10012, 2019. 2.1

[46] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5442–5451, 2019. 2.3

[47] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision*, pages 474–489. Springer, 2020. 1, 1, 2.1, 5.1, 1, 5, 2, 6.3, 6.4, 3, 4, 6.4

[48] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Generating smooth pose sequences for diverse human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13309–13318, 2021. 2.1

[49] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Con-*

ference on Computer Vision, pages 9489–9497, 2019. 1, 2.1, 1, 2, 6.3, 3

[50] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2891–2900, 2017. 1, 2.1, 1, 2, 6.3, 3

[51] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. Xnect: Real-time multi-person 3d human pose estimation with a single rgb camera. *arXiv preprint arXiv:1907.00837*, 2019. 1

[52] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018. 1

[53] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *3D Vision (3DV), 2018 Sixth International Conference on*. IEEE, sep 2018. 2.3

[54] Terry Monaghan. Why study the lindy hop? *Dance Research Journal*, 33(2):124–127, 2001. 1

[55] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10133–10142, 2019. 1

[56] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2823–2832, 2017. 1

[57] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10985–10995, October 2021. 2.1

[58] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3433–3441, 2017. 1

[59] Grégory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 1

[60] Caner Sahin, Guillermo Garcia-Hernando, Juil Sock, and Tae-Kyun Kim. Instance-and category-level 6d object pose estimation. In *RGB-D Image Analysis and Processing*, pages 243–265. Springer, 2019. 1

[61] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 2.3

[62] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1-2):4, 2010. 2.3

[63] Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. Space-time-separable graph convolutional network for pose forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11209–11218, October 2021. 2.1

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017. 1, 2.1

[65] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. 2.1

[66] Anirudh Vemula, Katharina Muelling, and Jean Oh. Social attention: Modeling attention in human crowds. In *2018 IEEE international Conference on Robotics and Automation (ICRA)*, pages 4601–4607. IEEE, 2018. 2.2

[67] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision*, pages 601–617, 2018. 1, 2.2, 2.3

[68] Can Wang, Jiefeng Li, Wentao Liu, Chen Qian, and Cewu Lu. Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation. In *European Conference on Computer Vision*, pages 242–259. Springer, 2020. 2.2

[69] Zhe Wang, Liyan Chen, Shaurya Rathore, Daeyun Shin, and Charless Fowlkes. Geometric pose affordance: 3d human pose with scene constraints. *arXiv preprint arXiv:1905.07718*, 2019. 2.2

[70] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 265–281, 2018. 2.1

[71] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision*, pages 346–364. Springer, 2020. 2.1

[72] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2148–2157, 2018. 2.2

[73] Yi Zhou, Zimo Li, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. Auto-conditioned recurrent networks for ex-

tended complex human motion synthesis. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. 2.1