Context-aware Human Motion Prediction

Enric Corona Albert Pumarola Guillem Alenyà Francesc Moreno-Noguer

Institut de Robòtica i Informàtica Industrial, CSIC-UPC, 08028, Barcelona, Spain
{ecorona, apumarola, galenya, fmoreno}@iri.upc.edu

Abstract

The problem of predicting human motion given a sequence of past observations is at the core of many applications in robotics and computer vision. Current state-of-the-art formulate this problem as a sequence-to-sequence task, in which a historical of 3D skeletons feeds a Recurrent Neural Network (RNN) that predicts future movements, typically in the order of 1 to 2 seconds. However, one aspect that has been obviated so far, is the fact that human motion is inherently driven by interactions with objects and/or other humans in the environment.

In this paper, we explore this scenario using a novel context-aware motion prediction architecture. We use a semantic-graph model where the nodes parameterize the human and objects in the scene and the edges their mutual interactions. These interactions are iteratively learned through a graph attention layer, fed with the past observations, which now include both object and human body motions. Once this semantic graph is learned, we inject it to a standard RNN to predict future movements of the human/s and object/s. We consider two variants of our architecture, either freezing the contextual interactions in the future of updating them. A thorough evaluation in the "Whole-Body Human Motion Database" [29] shows that in both cases, our context-aware networks clearly outperform baselines in which the context information is not considered.

1. Introduction

The ability to predict and anticipate future human motion based on past observations is essential for interacting with other people and the world around us. While this seems a trivial task for a person, it involves multiple sensory modalities and complex semantic understanding of the environment and the relations between all objects in it. Modeling and transferring this kind of knowledge to autonomous agents would have a major impact in many different fields, mainly in human-robot interaction [30] and autonomous driving [47], but also in motion generation for computer graphics animation [31] or image understanding [10].

The explosion of deep learning, combined with large-scale datasets of human motion such as Human3.6M [24] or the CMU motion capture dataset [34], has led to a significant amount of recent literature that tackles the problem of forecasting 3D human motion from past observations [14, 25, 43, 20, 3, 37, 15, 42, 26, 66]. These algorithms typically formulate the problem as sequence-to-sequence task, in which past observations represented 3D skeleton data are injected to a Recurrent Neural Network (RNN) which then predicts movements in the near future (less than 2 seconds).

Nevertheless, while promising results have been achieved, we argue that the standard definition of the problem used so far lacks an important factor, which is the influence of the rest of the environment on the movement of the person. For instance, if a person is carrying a box, the configuration of the body arms and legs will be highly constrained by the 3D position of that box. Discovering such interrelations between the person and the object/s of the context (or another person he/she is interacting with), and how these interrelations constrain the body motion, is the principal motivation of this paper.

In order to explore this new paradigm, we devise a context-aware motion prediction architecture, that models the interactions between all objects of the scene and the human using a directed semantic graph. The nodes of this graph represent the state of the person and objects (*e.g.* positional features) and the edges their mutual interactions. These interactions are iteratively learned with the past observations of the human and objects motion and fed into a standard RNN which is then responsible for predicting the future movement of all elements in the scene (for both rigid objects and non-rigid human skeletons). Additionally, we propose a variant of this model that also predicts the evolution of the adjacency matrix representing the interaction between the elements of the scene.

Presumably, one of the reasons why current state-of-the-art has not considered an scenario like ours is because all methods are trained and evaluated on benchmarks (mostly

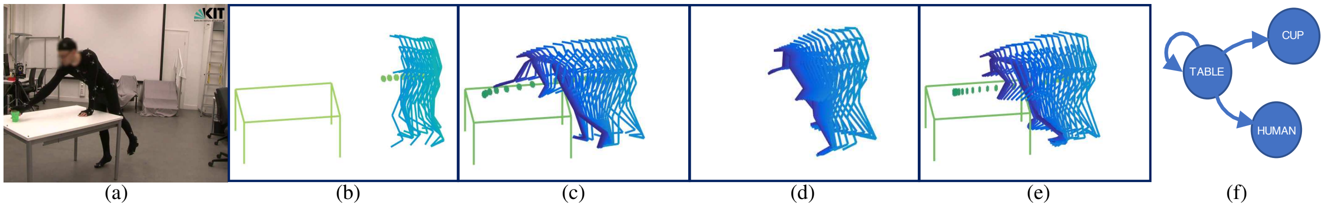


Figure 1: **Context-aware human motion prediction.** (a) Sample image of a sequence with a person placing a cup on a table. This image is shown solely for illustrative purposes, our approach only relies on positional data. (b) Past observations of all elements of the scene, the person, the cup and the table. (c) Ground truth future movements. (d) Human motion predicted using [43], consisting of an RNN that is agnostic of the context information. Note that there is a large gap with the ground truth, especially for the final frames of the sequence. (e) Cup and human motion prediction obtained with our context-aware model. While the arm of the person is not fully extended, the forecasted motion highly resembles the ground truth. Interestingly, the interaction with the table also helps to set the motion boundaries. (f) Main interactions that are learned with our approach in which dominates the influence of the table over both the cup and the person.

the aforementioned Human3.6M dataset [24]) annotated only with human motion. In this paper, we thoroughly evaluate our approach in the "Whole-Body Human Motion Database" [29], that contains about 200 videos of people performing several tasks and interacting with objects. This dataset is annotated with MoCap data for the humans and rigid displacement for the rest of objects, being thus, a perfect benchmark to validate our ideas. We also evaluate our method in the CMU MoCap database [34] with only two people being tracked. The results obtained in both datasets show that our methodology is able to accurately predict the future motion of people and objects while simultaneously learning very coherent interaction relations. Additionally, all context-aware versions, clearly outperform the baselines which uniquely rely on human past observations of the human (see Fig. 1). Since all previous works evaluate their methods using past observations of ground truth skeletons, we finally discuss the applicability of state-of-art motion prediction methods, with an ablation study of our models and baselines when considering noisy observations.

2. Related work

Human motion prediction. Since the release of large-scale MoCap datasets [52, 24, 29], there has been a growing interest in the problem of estimating 3D human pose from single images [5, 52, 49, 60, 56, 55, 44, 57]. More recently, the community is focusing in predicting 3D human motion from past observations. Most approaches build upon RNNs [14, 43, 20, 3, 50, 1] that encode historical motion of the human and predict the future configuration that minimizes different sort of losses. Martinez *et al.* [43], for instance, minimize the L2 distance and provide one of the baselines in our work. This work also compares against a zero-velocity baseline, which despite steadily predicting the last observed frame, yields very reasonable results under the L2 metric. This phenomenon has been recently discussed by Ruiz *et al.* [54], that argue that L2 distance is not an appropriate metric to capture the actual distribution of human motion, and that a network trained using only this metric is prone to converge to a mean body pose. To better capture real distributions of human movement, recent approaches use adversarial networks [17, 2] in combination with geometric losses [3, 20, 54, 33].

There exist alternative approaches other than RNNs. For instance, Jain *et al.* [25] consider a hand-crafted spatial-temporal graph adapted to the skeleton shape. Li *et al.* [37] use Convolutional Neural Networks to encode and decode skeleton sequences instead of RNNs.

All methods described in this section formulate the human prediction problem without considering the context information. In this paper, we aim to fill this gap.

Rigid 3D object motion prediction. While there is a vast amount of works on 3D object reconstruction [51, 19, 41], detection [9, 18, 11] and tracking [8, 4], only very few approaches address the problem of predicting future rigid motion [6, 63, 32, 59]. Among these, it is worth to mention Byravan *et al.* [6], that predict the future 3D pose given an image of an object and the action being applied to it. In our case, the action applied to each object is implicitly encoded in the previous observations.

Human-Object Interaction (HOI). Even though our work does not aim to identify Human-Object relationships, we have been inspired by a few papers on this topic. The standard formulation of the problem consists in representing an image with several detected objects and people as a graph encoding the context [22, 45, 53, 39, 16], or some other structured representation [36, 65, 12]. The most recent approaches [37, 53, 22] extract features of the detected entities using some image-based classification CNNs. Then, they compare pairs of features to predict their mutual interaction. Qi *et al.* [53] refine the representations and predicted interactions in a recursive manner. In this work, we use a similar idea to progressively refine the estimation of the interactions between objects.

Graph-based context reasoning. A few works leverage context information to boost the performance of different tasks[48, 38, 23, 35, 46]. Graph Convolutional Networks

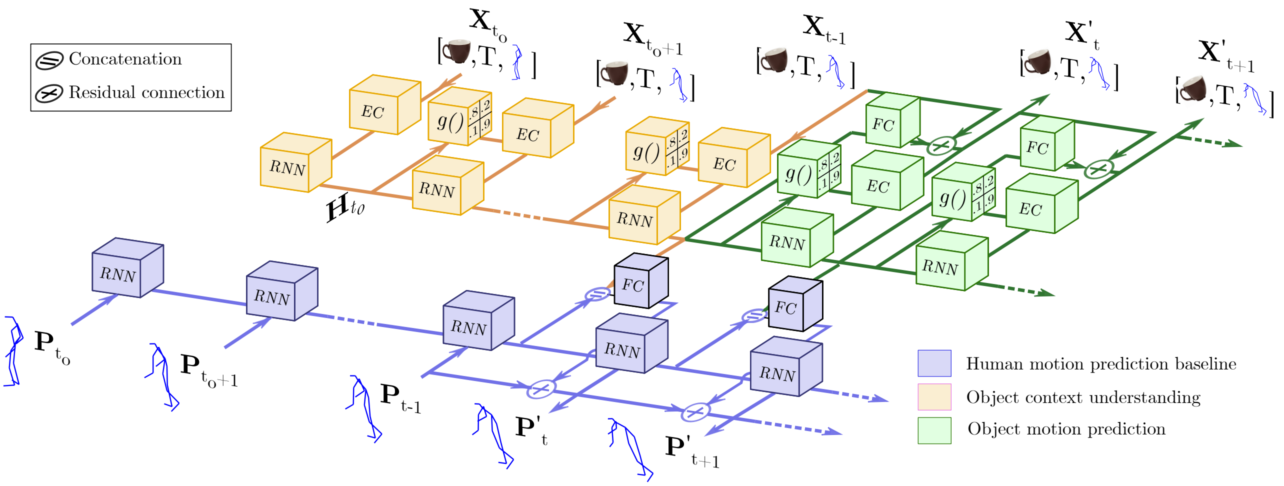


Figure 2: **Overview of our context-aware motion prediction model.** The blue branch represents a basic RNN that encodes past poses and decodes future human motion using a residual layer [43]. The upper branch corresponds to an RNN that encodes the contextual representation for each object in the scene. This branch contains two modules (depicted in brown and green). In brown, the past object position, class, and human joints are used to predict interactions and context feature vectors. The node corresponding to the human context representation is then used in conjunction with the human motion hidden state, to predict human motion. In green, the model is extended to predict motion of all observed objects. Best viewed in color.

(GCNs) [28] were recently proposed for improved semi-supervised classification. Jain *et al.* [25] used Structural RNNs to model spatio-temporal graphs. Wang *et al.* [61] propose to use GCNs, in which the interactions between objects depend on the intersection over union of their detected bounding boxes. Chen *et al.* [10] introduce an approach for image segmentation in which features from a 2D image coordinate space are represented in a graph reasoning space.

3. Problem formulation

Recent methods for human motion prediction consist of a model $\mathcal{M}$, typically a deep neural network, that encodes motion from time t_o until $t-1$. The goal is then to predict future human motion until t_f, namely $P_{t:t_f}$, where P stands for the human pose represented by 3D joint coordinates. Previous approaches have formulated the problem as $\mathcal{M} : P_{t_o:t-1} \rightarrow P_{t:t_f}$, *i.e.* future motion is estimated only from past observations. In this paper, we conjecture that future motion is also driven by the context and the action the human is performing. We therefore consider other objects O of type T in the scene with which the human may interact. The objects can be other people or any object in the scene. We will design our approach to be able to predict the motion of such objects of the context.

Additionally, the influence that objects will have in the future motion of other objects is unclear. Thus, we also aim to build a model that learns these interactions in an unsupervised manner. Considering all this, we reformulate our problem as the estimation of the following mapping:

$$\mathcal{M} : \{P_{t_o:t-1}, O_{t_o:t-1}, T\} \rightarrow \{P_{t:t_f}, O_{t:t_f}, I_{t_o:t_f}\} \;, \quad (1)$$

where I corresponds to the predicted interactions.

4. Approach

Figure 2 shows the main architecture used in this work. It consists of two branches that separately process human motion and object relationships. We use the latter to obtain a representation for all the observed entities, including the human, which we then use to predict both human and object motion prediction. We next describe these two branches.

4.1. Human motion branch

This branch builds upon the RNN network proposed by Martinez *et al.* [43]. This model, depicted in blue in Figure 2, is based on a residual architecture [21] that, at each step, uses a fully connected layer to predict the velocity of the body joints. As in a typical sequence-to-sequence network, the predictions are fed to the next step.

4.2. Context branch

The context information is represented using a directed graph structure where each node denotes an object or person. We then store a state for each entity and frame, encoding context information relevant to each node. These states are iteratively refined as new observations are processed.

Object representation. At each frame t, we define a matrix $X_t \in \mathbb{R}^{N \times F_0} = [O_t, T_t, P_t]$ that gathers the representation of all N nodes. F_0 is the length of the state vector of each node. This state vector contains the object 3D bounding box O_t, their object type T as a one-hot vector, and the joints of the person P_t. If the node does not correspond to a person, the joints in the representation are set to a zero vector of same size. The object type helps to identify the task the human is performing and the motion defined for that task.

By doing this, we aim to capture the semantic difference between the motion of a person when handling a knife or when using a whisk.

Modelling contextual object representations. Recent works on Graph Convolutional Networks (GCNs) [28] have shown very promising results in a variety of problems requiring the manipulation of graph-structured data. In GCNs, a feature vector of a certain node R_i is expressed as a function of other nodes x, as $R_i = \sigma(\sum_j^N \tilde{A}_{ij} W x_j)$, where W are trainable weights, σ is an activation and N the number of nodes of the graph connected to the i-th node. $\tilde{A} \in \mathbb{R}^{N\times N}$ is a normalized weighted adjacency matrix that defines interactions between nodes.

Graph Attention Networks (GATs) [58] have been proposed as an extension of GCNs, and introduce an attention model on every graph node. In this paper we also investigate the use of Edge Convolutions [62], which are indeed very similar to GATs. In ECs the update rule for a feature vector of each entity considers the representations of other relevant objects as follows:

$$R_i = \sigma(\sum_j^N \tilde{A}_{ij} W[x_i; x_i - x_j]). \tag{2}$$

The intuition behind this equation is that x_i encodes a global representation of the node, while $x_i - x_j$ provides local information. EC proposes combining both types of information in an asymmetric graph function.

We keep track of the context representations during all observations through a second RNN. Each node on the scene has a hidden state H that is updated every frame t:

$$H_i^{t+1} = \text{RNN}(R_i^t, H_i^t). \tag{3}$$

Learning interactions. As we shall see in the experimental section, we initially evaluate a simplified version of our Context-RNN (C-RNN) that uses a heuristic to define the adjacency matrices, setting $A_{ij} = 1$ if the center of gravity of objects i and j is closer than 1 meter.

In practice, interactions between entities are not known a priory, and furthermore, they change over time. Our goal is to automatically learn these changing interactions with no supervision. For this purpose we devise an iterative process in which, for the first frame, we set A to a diagonal matrix, *i.e.* $\tilde{A}_{t_0} = I_N$, meaning that the initial hidden representation of every object depends only on itself. We then predict the value of the interaction between two objects given the hidden state of both. We consider asymmetric weighted adjacency matrices, that for a frame t are estimated as:

$$A_{ij}^t = g(H_i^t, H_i^t - H_j^t), \tag{4}$$

with similar structure as in Eq. 2. The function g represents the output of a neural network layer, in our case a fully connected. We normalize the interactions for each node using a Softmax function, which we shall denote $\tilde{A}$.

Intuitively, we can consider this as a complete graph, where a graph attention mechanism [58] decides on the strength of interactions based on past observations. Note that while existing works typically use binary adjacency matrices from ground truth relationships [28], spatial assumptions [61] or K-NN on node representations [62], in this work we consider a differentiable continuous space of interactions, learned using back-propagation. In the rest of the paper we will denote the models that learn interactions with the suffix "-LI" (*e.g.* C-RNN+LI).

Object motion prediction. We propose two methods that exploit context at different levels. First, in the blue+brown modules of Fig. 2, we consider a model that reasons about the past context observations and iteratively improves hidden representations. The refined context representation of the human node is concatenated to the baseline branch (in blue) representation at every time step, and used by a fully connected layer to predict human velocity in that step. This is followed by a residual layer that yields skeleton poses.

Our second approach consists of the complete model depicted in Fig. 2 which, apart from past context, predicts object motion for all objects using a residual fully connected layer on each object hidden state. Analogous to the human motion branch, the predicted positions are forwarded to the next step, allowing to extend the context analysis into the future. The joints in the feature representations for those nodes describing people are also updated with the joint predictions of the human branch.

Additionally, when tracking several people, the human motion branch is repeated for each of them, and the model provides complete future motion for all available entities. In the rest of the document, we will denote the models that predict object motion with the suffix "-OPM".

5. Implementation details

Our model builds on the residual architecture of Martinez *et al*. [43] to allow an unbiased comparison with their work. The size of the human and object RNN hidden representations are 1024 and 256, respectively.

After the motion seed, we sample an observation every 100 ms. In all experiments, we encode and decode 10 (1 sec.) and 20 frames (2 sec.) respectively. Larger encoding times did not help in improving the results and significantly increased training time. We augment the train set through random rotation over the height Z in the range $(-180, 180]^\circ$ and random translation $X, Y \in (-1500, 1500)$mm.

We use a similar approach as in [53] to obtain the adjacency matrix. We build a 4D matrix A such that A_{ij} contains the hidden representations $[H_i^t; H_i^t - H_j]$ of nodes i and j, extending over the channel dimension. The function $g(\cdot)$ is formed by two Convolutional Layers of output kernel size 1 to make computation faster. We do not use bias term in these Convolutional layers nor in the Edge Convolutions.

Human motion	Passing objects				Grasping objects				Cutting food				Mixing objects				Cooking			
Time (s)	0.5	1	1.5	2	0.5	1	1.5	2	0.5	1	1.5	2	0.5	1	1.5	2	0.5	1	1.5	2
ZV[43]	**34**	**81**	120	153	89	222	333	421	54	132	198	258	102	262	396	495	**24**	**53**	70	80
RNN[43]	50	99	132	162	82	158	211	254	48	103	140	180	68	135	190	226	27	54	65	71
QuaterNet	62	145	211	267	208	209	248	292	87	211	308	389	192	237	296	345	39	87	121	144
C-RNN	47	102	141	177	76	149	203	247	49	100	124	158	70	158	214	247	26	**53**	**63**	69
C-RNN+OMP	53	99	127	155	128	154	197	239	49	96	121	149	**61**	**127**	**168**	**199**	29	55	65	70
C-RNN+LI	43	89	117	**142**	**72**	**141**	**188**	**230**	**47**	**92**	**117**	**147**	72	145	194	219	27	**53**	**63**	69
C-RNN+OMP+LI	44	89	**116**	**142**	115	156	204	251	48	95	121	**147**	77	152	195	219	26	**53**	**63**	**68**
Object motion	Passing objects				Grasping objects				Cutting food				Mixing objects				Cooking			
Time (s)	0.5	1	1.5	2	0.5	1	1.5	2	0.5	1	1.5	2	0.5	1	1.5	2	0.5	1	1.5	2
ZV	48	118	181	237	65	152	226	289	**29**	70	104	132	50	126	188	229	16	**33**	**44**	**53**
RNN	49	107	154	198	64	139	201	257	**29**	70	105	134	47	113	166	199	17	36	48	58
C-RNN+OMP	**44**	92	122	150	**55**	**103**	**136**	**167**	31	64	83	97	**29**	**65**	**90**	**110**	**15**	**33**	46	56
C-RNN+OMP+LI	**44**	**91**	**119**	**142**	58	112	152	186	**29**	**62**	**81**	**92**	51	106	145	171	**16**	34	46	55

Table 1: **Class-specific models results.** In this table, every action is independently trained. The results report the mean Euclidean error (in mm), for the 2s prediction of the human motion (top) and object motion (bottom). In all cases, 1s of past observations is provided. The context-based models we propose in this paper are those with the suffixes "OMP" and "LI". They provide the best results in most sequences.

All	Human motion prediction																			
Time (s)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
ZV [43]	24	46	67	87	106	125	143	160	176	190	205	219	231	244	256	267	279	290	300	310
RNN [43]	29	44	57	68	78	87	96	104	113	121	128	136	143	150	157	164	171	177	184	191
C-RNN	27	46	58	69	79	87	96	104	113	121	129	137	144	152	160	166	174	181	188	196
C-RNN+OMP	46	83	76	82	87	95	101	108	116	123	131	138	146	153	160	167	174	182	189	197
C-RNN+LI	**21**	**39**	**52**	**63**	**72**	**80**	**89**	**97**	**104**	**111**	**118**	**125**	**131**	**137**	**144**	**150**	**157**	**163**	**170**	**177**
C-RNN+OMP+LI	39	77	77	76	80	87	94	101	108	114	120	126	133	139	145	151	158	165	171	178
All	Object motion prediction																			
ZV	**13**	**25**	**35**	**44**	52	60	68	77	84	90	96	102	109	115	120	125	131	135	140	144
RNN	15	28	38	46	53	60	68	74	80	85	91	97	102	107	112	117	121	125	130	135
C-RNN+OMP	15	26	36	**44**	**50**	**55**	**61**	**67**	**73**	**79**	**84**	89	94	99	104	108	113	117	121	125
C-RNN+OMP+LI	16	29	39	46	52	57	63	69	75	**79**	**84**	**88**	**93**	**97**	**101**	**105**	**110**	**114**	**117**	**121**

Table 2: **Training with all actions simultaneously.** For each method we train a single model using all actions simultaneously. See also caption in Table 1.

Object representations are formed first by the bounding box position, defined by the minimum and maximum 3D Cartesian points.

We train the model to minimize $L2$ distance between the predicted and the actual future motion $\mathcal{L} = ||M(P_{t_o:t-1}) - P_{t:t_f}||_2$. The model is trained until convergence, using Adam [27] with learning rate of 0.0005, beta1 0.5, beta2 0.99 and batch size 16.

6. Experiments

6.1. Preliminaries

Datasets. Large-scale MoCap datasets [29, 24, 34] provide annotations on the human poses but do not give any annotation about objects of the scene or any relevant context information. Therefore, most recent works on human motion prediction are evaluated without considering context information. Martinez *et al*. [43] show that for certain cases, even a simple zero-velocity baseline may yield better results than context-less learning models.

To demonstrate the merits of our approach, we leverage on the Whole-Body Human Motion (WBHM) Database [40], a large-scale publicly available dataset containing 3D raw data of multiple individuals and objects. In particular, we use all the activities where human joints are provided and include at least a table. This results in 190 videos and 198K frames, and a total of 15 tracked object classes. We use the raw recordings Vicon files at 100 Hz to obtain the bounding box of each object in each frame, and select 18 joints to represent the human skeleton.

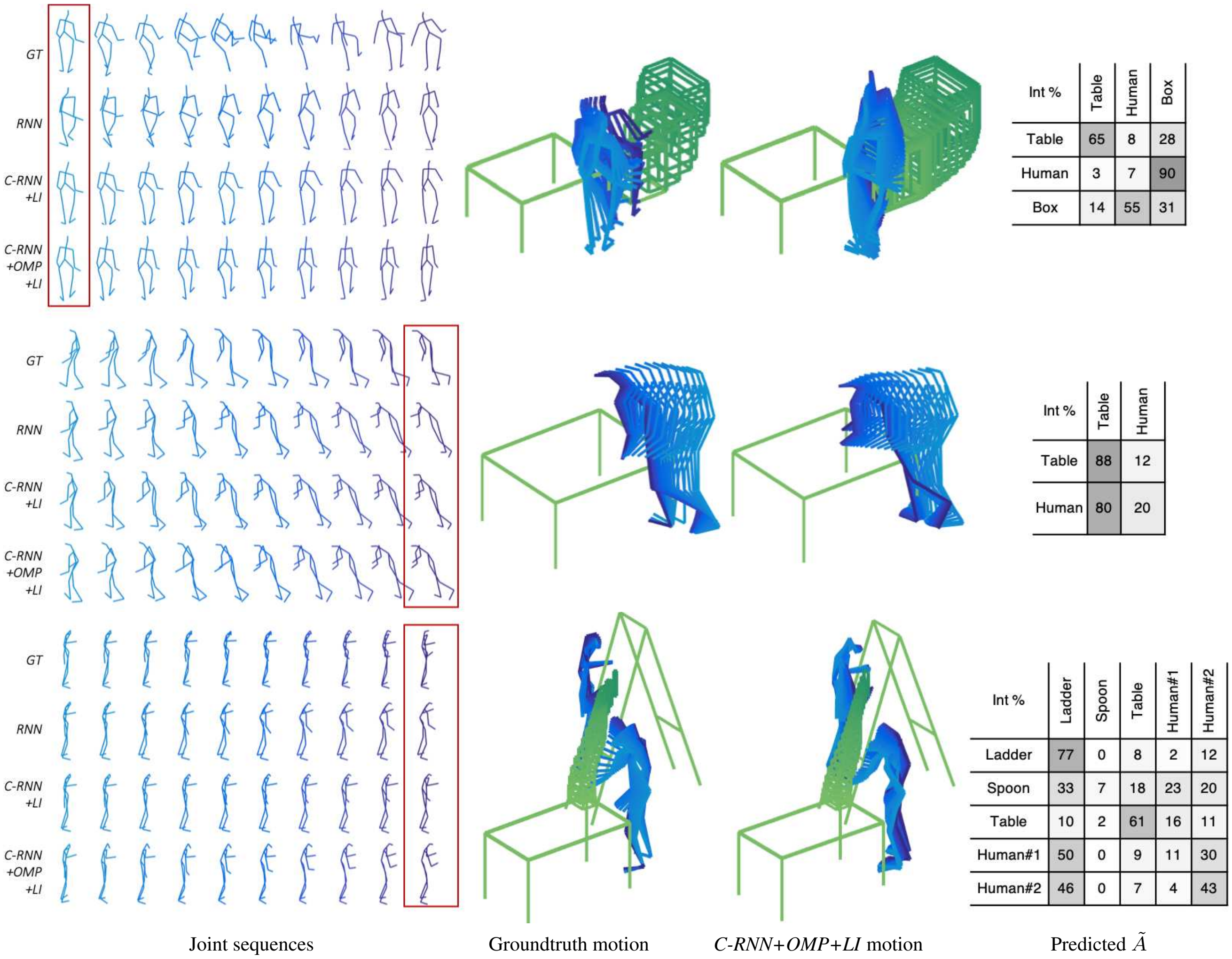
We extract different actions representing different levels of complexity on the contextual information. The statistics of this dataset are the following:

	Passing objects	Grasping/ leaving object	Cutting food	Mixing objects	Cooking	All
# objects	4	5	6	9	12	15
# people	2	1	1	1	1	1/2
# videos	18	36	10	17	35	190
# frames	30k	31k	11k	14k	54k	198k

We will report results on both action-specific models and also on models trained with the entire dataset.

We also run experiments on the CMU Mocap Database [34]. We select the actions that include two people interacting, which include 34 videos with different activities like dancing, talking with hand gestures or boxing. In this case, the objects are not annotated, but we will show that context information from the two users is useful to improve over context-less models.

Baselines. We compare our models to the context-less models proposed in [43]. First, we consider the basic residual



Int %	Table	Human	Box
Table	65	8	28
Human	3	7	90
Box	14	55	31

Int %	Table	Human
Table	88	12
Human	80	20

Int %	Ladder	Spoon	Table	Human#1	Human#2
Ladder	77	0	8	2	12
Spoon	33	7	18	23	20
Table	10	2	61	16	11
Human#1	50	0	9	11	30
Human#2	46	0	7	4	43

Figure 3: **Qualitative motion generation up to two seconds. Left:** Predicted sample frames of our approaches and the baselines. **Center:** Detail of the predictions obtained with our approaches, compared with the ground truth. Human and object motion are represented from light blue to dark blue and light green to dark green, respectively. Actions, from top to bottom are: A human supports on a table to kick a box, human leaning on a table, and two people (one of them standing on a ladder) passing an object. **Right:** Predicted adjacency matrices representing the interactions learned by our model. Note that these relations are directional (*e.g.* in the last example the ladder highly influences the motion of the Human#1 (50%) but the human has little influence over the ladder (11%). Best viewed in color with zoom.

RNN. We also consider a Zero-Velocity (ZV) baseline that constantly predicts the last observed frame. We also compare to QuaterNet [50] using their available code, to predict absolute motion prediction. For object motion prediction, we also use a ZV and RNN models [43], where the position of an object is defined by its 3D bounding box.

Our models. We run our context-aware models (*C-RNN*), incrementally adding the main ideas described in the paper.

The basic *C-RNN* in our experiments uses the spatial heuristic described in Section 4.2 where interactions depend only on the distance between objects. This model processes context during the past frames, and then uses the last hidden state of the human node for human motion prediction at each step. This is extended by additionally predicting object motion (*OMP*) and recomputing object interaction from the previous assumption on the predicted positions. We then evaluate the efficiency of our model for learning interactions (*LI*). Like in the previously defined experiments, we evaluate a model that considers past contextual information and a model that prolongs object analysis into the future.

Evaluation metric. Previous works on human motion prediction focus mainly on predicting relative motion [43, 20, 50], using joint angles. However, our model reasons about the full scene and is able to predict absolute motion in Cartesian coordinates. Therefore, we use the mean Euclidean Distance (in mm) between predictions and real future motion, obtained from the unnormalized predictions in the 3D space. For human motion prediction, we take into account the 18 joints defined in the human skeleton. For objects, we consider the eight 3D vertices of their bounding boxes.

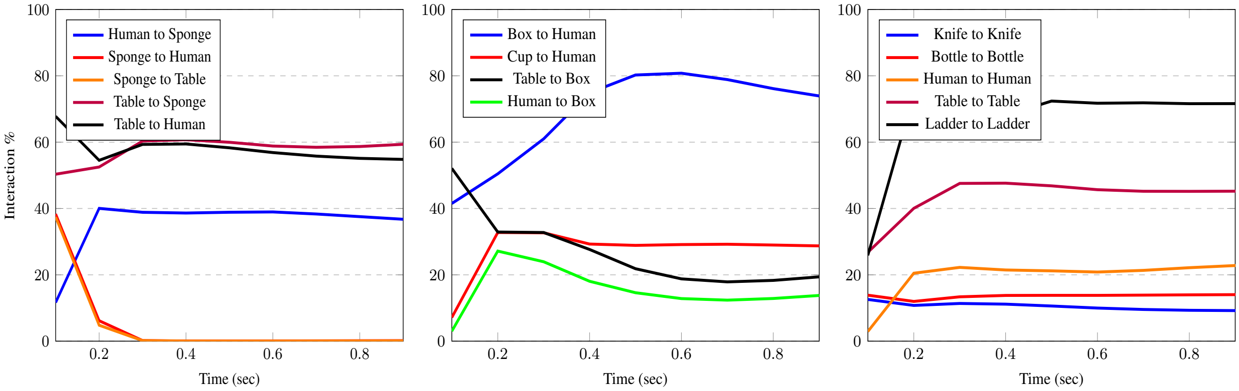


Figure 4: **Average interactions refined by the model during the past observations of the context**. In the left and center plots, we depict relevant interactions for *table cleaning* and *moving box* activities respectively. In the first case, notice the table affects significantly the sponge and human, which initially moves towards the table to clean it. Similarly, in the second case, the human moves towards a box on the ground, picks it up and puts it on the table. The right plot shows average self-interaction percentages among all the test samples, for relevant object types. We found that non-moving objects like tables or ladders consistently have very little influence from other objects. Likewise, passive objects that are often moved by a human, such as knives or bottles, are more influenced by them and leave self-influence relatively low.

6.2. Results on the WBHM Dataset

Quantitative results. Table 1 summarizes the performance of class-specific models trained on different activities. Table 2 provides results at much higher temporal resolution for models trained using all the dataset, reporting the mean Euclidean distance between predictions and ground truth every 100 ms. In all cases, 1 second of past observations is provided and 2 seconds are predicted.

The performance of models that consider a threshold-based binary interaction vary significantly between classes, suggesting they are effectively unable to understand the context as done by models that learn the actual interactions (*LI*). Notice that even the basic *C-RNN* does not yet provide a consistent improvement compared to state-of-art models. The same model that additionally learns interactions (C-RNN+LI) obtains a significant boost in most cases. Nonetheless, activities such as passing objects or grasping require attending to items that are at variable distances.

Regarding the complexity of the scene, most improvement comes from scenes with a small number of objects where interactions are well defined and actions are more predictable. For cooking activities, there are several objects in a table next to the human. Different motion options are possible and, as uncertainty grows, the model seems unable to confidently understand interactions. Because of this, context-aware models do not provide such a significant improvement as in previous activities. Considering all actions simultaneously seems to favor even more the context-aware approaches and, specially, those that learn interactions (*C-RNN+LI* and *C-RNN+OPM+LI*).

Qualitative results. Figure 3-left shows the motion generation results of our two main models, compared to the baseline [43] on different classes. We did not include the Zero-Velocity baseline as it does not provide interesting motion even though it has remained a difficult baseline on uncertain activities. We have marked some specific frames in which context-aware approaches improve the RNN baseline.

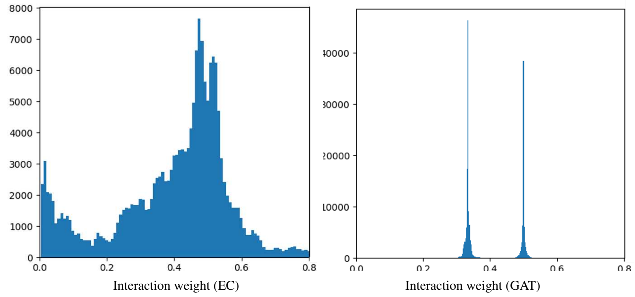


Figure 5: **Interaction strength histogram predicted by EC and GAT models.** These include interactions predicted among all humans and objects after two observations are given to the networks. For simplicity, we depict the histogram from tasks whose context only contain two or three nodes. On the left, interactions learnt by EC-based model, spanning a wide range of values, up to interaction strengths of more than 80%. On the right, GAT-based model, which predicts all interaction weights similarly and therefore we can only see peaks at $1/2$ and $1/3$.

For human motion prediction, poses generated are frequently more semantically-related to their closest objects than context-less models. For instance, as shown in the last action of Figure 3, people holding objects tend to move the relevant hand. For object motion prediction, context-less model predictions hardly move from their original position.

Regarding the interactions predicted by the model, we notice coherent patterns in many activities. For example, drinking videos generate strong Cup-Human relationships. In Figure 4, we represent the average predicted interactions for different actions. These are gathered from the *C-*

	Noise-free input				Noisy input (HMP)											
Model	HMP Models				25 mm				50 mm				100 mm			
Time (s)	0.5	1	1.5	2	0.5	1	1.5	2	0.5	1	1.5	2	0.5	1	1.5	2
ZV [43]	61	150	223	281	70	156	227	284	87	168	236	292	122	195	260	313
RNN [43]	55	110	148	179	72	122	156	185	83	127	162	192	112	184	214	226
C-RNN+LI	52	104	**136**	**161**	**58**	**107**	**139**	**166**	**69**	**113**	147	**175**	**99**	136	175	208
C-RNN+OMP+LI	56	109	140	165	62	111	142	167	71	116	**146**	171	103	132	**162**	**187**
Model	OMP Models				25 mm				50 mm				100 mm			
ZV	42	100	149	188	53	106	154	191	**66**	118	164	199	106	151	187	223
RNN	41	93	135	169	52	99	141	174	69	113	151	183	**105**	142	181	208
C-RNN+OMP+LI	**40**	**81**	**109**	**129**	**51**	**88**	**115**	**134**	67	**100**	**126**	**144**	106	**132**	**156**	**172**

Table 3: **Robustness to noise in Human and Object Motion Prediction.** Average performance of the principal models when using the original test set (Noise-free input), compared to their performance when seeing noisy observations.

CMU MoCap Dataset								
Time (s)	0.5		1.0		1.5		2.0	
	μ	σ	μ	σ	μ	σ	μ	σ
ZV	127	32	271	66	374	86	460	97
RNN	125	28	267	58	378	77	477	92
QuaterNet	138	26	279	58	378	82	466	95
C-RNN+LI	**124**	27	**257**	53	**352**	65	**435**	78

Table 4: **Mean and Std of prediction errors (mm) on the CMU dataset.** Our Context-aware model C-RNN+LI outperforms baselines even though context only consists of two people.

RNN+OMP+LI. This model provides more intense Object-Object interactions than *C-RNN+LI*, which does not need to obtain such meaningful representations for objects as only human contextual representations are used. Note that the models learn to predict interactions that provide information relevant to future pose, and thus improve motion predictions. Interactions here do not necessarily respond to actual action relationships.

We finally study the effect of the Graph architecture in the learned interactions. Graph Attention Networks (GATs) and Edge Convolutions (EC) provide an attention mechanism to measure the interaction strength. Nevertheless, we found that GAT-based networks consider all interactions of similar importance, while EC-based architectures are able to predict a continuous and wide range of attention values. We show this in Figure 5.

6.3. Results on the CMU MoCap Dataset

We train the models again on the CMU MoCap Database, obtaining the results depicted in Table 4. In this setup, the users perform very energetic activities like dancing or boxing, which implies that absolute motion is larger, and error on the CMU MoCap database being in average more than twice that in the former database. In this case, only two nodes are observed in each video for the two people being tracked. Since no information about actions or objects is given, we do not provide results on *OMP*. However, we find our proposed model *C-RNN+LI* outperforms all other baselines significantly, specially in the long-term.

6.4. Robustness to noise

All previous works on human motion prediction use ground truth MoCap data as past observations. Nevertheless, real applications will receive joint observations from *e.g.* human pose estimation models, such as OpenPose [7] or AlphaPose [13, 64], which are prone to suffer from noise and mis-detections, specially under strong occlusions. In these subsection, we therefore evaluate the resilience of our proposed models and previous baselines to noise in the input observations. Predictions are evaluated on the original ground truth data. The 3D coordinates of past observations (both in human and objects positions) are corrupted by additive Gaussian noise $\mathcal{N}(0, \sigma^2)$. In Table 3 we show the results of this experiment, with different values of σ. Interestingly, the error in the predictions gracefully increases with the noise, but still, our approach performs consistently better than those approaches that do not consider the context information. Indeed, the best context-aware models (*C-RNN+LI* and *C-RNN+OMP+LI*) with noise up to $\sigma = 50mm$, perform better than context-less baselines with no noise in the input.

7. Conclusion

In this work, we explore a context-aware motion prediction architecture, using a semantic-graph representation where objects and humans are represented by nodes independently of the number of objects or complexity of the environment. We extensively analyze their contribution for human motion prediction. The results observed in different actions suggest that the models proposed are able to understand human activities significantly better than state-of-art models which do not use context, improving both human and object motion prediction.

Acknowledgements: This work has been partially funded by the Spanish government under projects HuMoUR TIN2017-90086-R, ERA-Net Chistera project IPALM PCI2019-103386 and María de Maeztu Seal of Excellence MDM-2016-0656. We also thank Nvidia for hardware donation.

References

[1] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *ICCV*, 2019. 2

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. 2

[3] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *CVPR-Workshop*, 2018. 1, 2

[4] Erkan Baser, Venkateshwaran Balasubramanian, Prarthana Bhattacharyya, and Krzysztof Czarnecki. Fantrack: 3d multi-object tracking with feature association network. In *IV*, 2019. 2

[5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016. 2

[6] Arunkumar Byravan and Dieter Fox. Se3-nets: Learning rigid body motion using deep neural networks. In *ICRA*, 2017. 2

[7] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018. 8

[8] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[9] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Andrew G Berneshawi, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3d object proposals for accurate object class detection. In *Advances in Neural Information Processing Systems*, pages 424–432, 2015. 2

[10] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Shuicheng Yan, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. *arXiv preprint arXiv:1811.12814*, 2018. 1, 3

[11] Enric Corona, Kaustav Kundu, and Sanja Fidler. Pose estimation for objects with rotational symmetry. In *IROS*, 2018. 2

[12] Enric Corona, Albert Pumarola, Guillem Alenyà, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *CVPR*, 2020. 2

[13] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *CVPR*, 2017. 8

[14] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *ICCV*, 2015. 1, 2

[15] 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)*, 2017. 1

[16] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Actions and attributes from wholes and parts. In *ICCV*, 2015. 2

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 2

[18] Alexander Grabner, Peter M Roth, and Vincent Lepetit. 3d pose estimation and 3d model retrieval for objects in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3022–3031, 2018. 2

[19] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan Russell, and Mathieu Aubry. AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In *CVPR*, 2018. 2

[20] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José MF Moura. Adversarial geometry-aware human motion prediction. In *ECCV*, 2018. 1, 2, 6

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[22] Roei Herzig, Moshiko Raboh, Gal Chechik, Jonathan Berant, and Amir Globerson. Mapping images to scene graphs with permutation-invariant structured prediction. In *NIPS*, 2018. 2

[23] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *CVPR*, 2018. 2

[24] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 2014. 1, 2, 5

[25] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *CVPR*, 2016. 1, 2, 3

[26] Kyung-Rae Kim, Whan Choi, Yeong Jun Koh, Seong-Gyun Jeong, and Chang-Su Kim. Instance-level future motion estimation in a single image based on ordinal regression. In *ICCV*, 2019. 1

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[28] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 3, 4

[29] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *IJRR*, 2013. 1, 2, 5

[30] Hema S Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *TPAMI*, 2016. 1

[31] Lucas Kovar, Michael Gleicher, and Frédéric Pighin. Motion graphs. In *SIGGRAPH*, 2008. 1

[32] Philipp Kratzer, Marc Toussaint, and Jim Mainprice. Motion prediction with recurrent neural network dynamical models and trajectory optimization. *arXiv preprint arXiv:1906.12279*, 2019. 2

[33] Jogendra Nath Kundu, Maharshi Gor, and R Venkatesh Babu. Bihmp-gan: Bidirectional 3d human motion prediction gan. *arXiv preprint arXiv:1812.02591*, 2018. 2

[34] CMU Graphics Lab. Cmu motion capture database. `http://mocap.cs.cmu.edu/`. 1, 2, 5

[35] Colin Lea, Austin Reiter, René Vidal, and Gregory D Hager. Segmental spatiotemporal cnns for fine-grained action segmentation. In *ECCV*, 2016. 2

[36] Ruiyu Li, Makarand Tapaswi, Renjie Liao, Jiaya Jia, Raquel Urtasun, and Sanja Fidler. Situation recognition with graph neural networks. In *ICCV*, 2017. 2

[37] 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, 2

[38] Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Structure inference net: object detection using scene-level context and instance-level relationships. In *CVPR*, 2018. 2

[39] Chih-Yao Ma, Asim Kadav, Iain Melvin, Zsolt Kira, Ghassan AlRegib, and Hans Peter Graf. Attend and interact: Higher-order object interactions for video understanding. In *CVPR*, 2018. 2

[40] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. The kit whole-body human motion database. In *ICAR*, 2015. 5

[41] Priyanka Mandikal and Venkatesh Babu Radhakrishnan. Dense 3d point cloud reconstruction using a deep pyramid network. In *WACV*, 2019. 2

[42] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *ICCV*, 2019. 1

[43] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *CVPR*, 2017. 1, 2, 3, 4, 5, 6, 7, 8

[44] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *CVPR*, 2017. 2

[45] Alejandro Newell and Jia Deng. Pixels to graphs by associative embedding. In *NIPS*, 2017. 2

[46] Bingbing Ni, Xiaokang Yang, and Shenghua Gao. Progressively parsing interactional objects for fine grained action detection. In *CVPR*, 2016. 2

[47] Brian Paden, Michal Čáp, Sze Zheng Yong, Dmitry Yershov, and Emilio Frazzoli. A survey of motion planning and control techniques for self-driving urban vehicles. *IV*, 2016. 1

[48] Sarah Parisot, Sofia Ira Ktena, Enzo Ferrante, Matthew Lee, Ricardo Guerrero, Ben Glocker, and Daniel Rueckert. Disease prediction using graph convolutional networks: Application to autism spectrum disorder and alzheimer's disease. *Medical image analysis*, 48, 2018. 2

[49] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *CVPR*, 2017. 2

[50] Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion. *arXiv preprint arXiv:1805.06485*, 2018. 2, 6

[51] Albert Pumarola, Stefan Popov, Francesc Moreno-Noguer, and Vittorio Ferrari. C-Flow: Conditional Generative Flow Models for Images and 3D Point Clouds. In *CVPR*, 2020. 2

[52] Albert Pumarola, Jordi Sanchez-Riera, Gary Choi, Alberto Sanfeliu, and Francesc Moreno-Noguer. 3dpeople: Modeling the geometry of dressed humans. In *ICCV*, 2019. 2

[53] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, 2018. 2, 4

[54] Alejandro Hernandez Ruiz, Juergen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. In *ICCV*, 2019. 2

[55] Edgar Simo-Serra, Ariadna Quattoni, Carme Torras, and Francesc Moreno-Noguer. A joint model for 2d and 3d pose estimation from a single image. In *CVPR*, 2013. 2

[56] Edgar Simo-Serra, Arnau Ramisa, Guillem Alenyà, Carme Torras, and Francesc Moreno-Noguer. Single image 3d human pose estimation from noisy observations. In *CVPR*, 2012. 2

[57] Edgar Simo-Serra, Carme Torras, and Francesc Moreno-Noguer. 3d human pose tracking priors using geodesic mixture models. 2017. 2

[58] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017. 4

[59] Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. Sfm-net: Learning of structure and motion from video. *arXiv preprint arXiv:1704.07804*, 2017. 2

[60] Chunyu Wang, Yizhou Wang, Zhouchen Lin, Alan L Yuille, and Wen Gao. Robust estimation of 3d human poses from a single image. In *CVPR*, 2014. 2

[61] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV*, 2018. 3, 4

[62] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *arXiv preprint arXiv:1801.07829*, 2018. 4

[63] Jiayin Xie and Nilanjan Chakraborty. Rigid body motion prediction with planar non-convex contact patch. In *ICRA*, 2019. 2

[64] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose flow: Efficient online pose tracking. *arXiv preprint arXiv:1802.00977*, 2018. 8

[65] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *CVPR*, 2016. 2

[66] Jason Y Zhang, Panna Felsen, Angjoo Kanazawa, and Jitendra Malik. Predicting 3d human dynamics from video. In *ICCV*, 2019. 1