Multi-Person 3D Motion Prediction with Multi-Range Transformers

Jiashun Wang[1] Huazhe Xu[2] Medhini Narasimhan[2] Xiaolong Wang[1]
[1]UC San Diego [2]UC Berkeley

Abstract

We propose a novel framework for multi-person 3D motion trajectory prediction. Our key observation is that a human's action and behaviors may highly depend on the other persons around. Thus, instead of predicting each human pose trajectory in isolation, we introduce a Multi-Range Transformer *model which contains of a local-range encoder for individual motion and a global-range encoder for social interactions. The Transformer decoder then performs prediction for each person by taking a corresponding pose as a query which attends to both local and global-range encoder features. Our model not only outperforms state-of-the-art methods on long-term 3D motion prediction, but also generates diverse social interactions. More interestingly, our model can even predict 15-person motion simultaneously by automatically dividing the persons into different interaction groups. Videos are available at:* https://anonymousaut.github.io/Anonymous-Result/

1. Introduction

Given a few time steps of human motion, we are able to forecast how the person will continue to move and imagine the complex dynamics of their motion in the future. The ability to perform such predictions allows us to react and plan our own behaviors. Similarly, a predictive model for human motion is an essential component for many real world computer vision applications such as surveillance systems, and collision avoidance for robotics and autonomous vehicles. The research on 3D human motion prediction has caught a lot of attention in recent years [35, 34], where deep models are designed to take a few steps of 3D motion as inputs and predict a long-term future 3D motion as the outputs.

While encouraging results have been shown in previous work, most of the research focus on single human 3D motion prediction. Our key observation is that, how a human acts and behaves may highly depend on the people around. Especially during interactions with multiple agents, an agent will need to predict the other agents' intentions, and then response accordingly [42]. Thus instead of predicting each human motion in isolation, we propose to build a model to predict multi-person 3D motion and interactions. Such a model will need have the following properties: (i) understand each agent's own motion in previous time steps to obtain smooth and natural future motion; (ii) within a crowd of agents, understand which agents are interacting with each other and learn to predict based on the social interactions; (iii) the time scale for prediction needs to be long-term.

In this paper, we introduce a Multi-Range Transformer for multi-person 3D motion trajectory prediction. The Transformer [48] has shown to be very effective in modeling long-term relations in language modeling [12] and recently in visual recognition [13]. Inspired by these encouraging results, we propose to explore Transformer models for predicting long-term human motion (3 seconds into the future). Our Multi-Range Transformer contains a local-range Transformer encoder for each individual person trajectory, a global-range Transformer encoder for modeling social interactions, and a Transformer decoder for predicting each person's future motion trajectory in 3D.

Specifically, given the human pose joints (with 3D locations in the world coordinate) in 1-second time steps as inputs, the local-range Transformer encoder processes each person's trajectory separately and focuses on the local motion for smooth and natural prediction. The global-range Transformer encoder performs self-attention on 3D pose joints across different persons and different time steps, and it automatically learns which persons that one person should be attending to model their social interactions. Our Transformer decoder will then take a single human 3D pose in *one time step* as the query input and encoder features as the key and value inputs to compute attention for prediction. We perform prediction for different persons by using different query pose inputs. By using only one time step person pose as the query for the decoder instead of a sequence of motion steps, we create a bottleneck to force the Transformer to exploit the relations between different time steps and persons in the encoders, instead of just repeating the existing motion alone [34].

We perform our experiments on multiple datasets including CMU-Mocap [1], MuPoTS-3D [38], 3DPW [49], Panop-

tic [20] for multi-person motion prediction in 3D (with $2 \sim 15$ persons). Our method achieves a significant improvement over state-of-the-art approaches for long-term predictions and the gain enlarges as we increase the future prediction time steps from 1 second to 3 seconds. Qualitatively, we visualize that our method can predict interesting behaviors and interactions between different persons while previous approaches will repeat the same poses as it goes to further steps in the future.

2. Related Work

3D Motion Prediction. Predicting future human pose in 3D has been widely studied with Recurrent Neural Networks (RNNs) [11, 15, 26, 32, 36, 40, 16, 18, 22, 55, 17]. For example, Fragkiadaki *et al.* [15] propose a Encoder-Recurrent-Decoder (ERD) model which incorporates nonlinear encoder and decoder networks before and after recurrent layers. Besides using RNNs, temporal convolution networks have also show promising results on modeling long-term motion [30, 24, 9, 9, 35, 34, 5]. Most of these studies fix the pose center and ignore the global body trajectory. Instead of solving two problems separately, recent works start looking into jointly predict human pose and the trajectory in the world coordinate [53, 56, 54, 50, 10].

Social interaction with multiple persons. Multi-person trajectory prediction has been a long standing problem in decades [21, 37, 51, 41, 57, 7, 6, 19, 14, 39, 29, 33, 43, 8, 28, 46, 52, 47]. For example, Alahi *et al.* [7] present a LSTM [23] model which jointly reasons across multiple individuals in a scene. However, most of these approaches focus on the global movement of the humans. To model more fine-grained human-human interactions, recent research have proposed to predict multi-person poses and trajectories at the same time [45, 44, 3, 2]. Inspired by these works, we propose a novel Multi-Range Transformer which scales up the long-term prediction with even more than 10 persons.

3. Method

Given a scene with N persons and their corresponding history motion, our goal is to predict their future 3D motion. Specifically, given $X^n_{1:k} = [x^n_1, ..., x^n_k]$ representing the history motion of person n where $n = 1, ..., N$, and k is the time step. We aim to predict the future motion $X^n_{k+1:T}$ where T represents the end of the sequence. We use a vector $x^n_k \in \mathbb{R}^{3J}$ containing the Cartesian coordinates of the J skeleton joints to represent the pose of the person n at time step k. In contrast to most previous motion prediction works which center the pose (joint positions) at the origin, we instead use the absolute joint positions in the world coordinate. In our method, x^n_k contains both the trajectory and the pose information. For simplicity, we omit subscript n when n only represents an arbitrary person, e.g., taking $x^n_{1:k}$ as $x_{1:k}$.

3.1. Network Architecture

The proposed architecture is composed of a motion predictor $\mathcal{P}$ and a motion discriminator $\mathcal{D}$. In the predictor $\mathcal{P}$, two Transformer-based encoders encode the individual (local) and global motion separately and one Transformer-based decoder decodes a smooth and natural motion sequence. The motion discriminator $\mathcal{D}$ is a Transformer-based classifier to determine whether the generated motion is natural. The network architecture is shown in Fig. 1.

3.1.1 Local-range Transformer Encoder

We first use our Local-range Transformer encoder to process this person's history motion. We use offset $\Delta x_i = x_{i+1} - x_i$ between two time steps to represent the motion. We apply Discrete Cosine Transform (DCT) and a linear layer to $\Delta x_{1:k}$ and then add the sinusoidal positional embedding [48] to get the local motion embedding $l_{1:k} = [l_1, ..., l_k]$. We concatenate them as a set of tokens $E_{loc} = [l_1, ..., l_k]^T$ and feed them to the Transformer encoder. We have L alternating layers in the local-range Transformer encoder and we introduce the technique we use in each layer. Firstly, a Multi-Head Attention is used for extracting the motion information,

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= [\text{head}_1; ...; \text{head}_h]W^O \\ \text{where head}_i &= \text{softmax}(\frac{Q^i(K^i)^T}{\sqrt{d_K}})V^i \end{aligned} \quad (1)$$

W^O is a projection parameter matrix, d_K is the dimension of the key and h is the number of the heads we use. We use self-attention and get the query Q_{loc}, key K_{loc}, and value V_{loc} from E_{loc} for each head_i as $Q^i_{loc} = E_{loc}W^{(Q,i)}_{loc}, K^i_{loc} = E_{loc}W^{(K,i)}_{loc}, V^i_{loc} = E_{loc}W^{(V,i)}_{loc}$ where $W^{(Q,i)}_{loc}$, $W^{(K,i)}_{loc}$, $W^{(V,i)}_{loc}$ are projection parameter matrices. loc represents the local-range. We then employ a residual connection and the layer normalization techniques to our architecture. We further apply a feed forward layer, again followed by a residual connection and a layer normalization following [48]. The whole process forms one layer of local-range Transformer encoder. We stack L such Transformer encoders to update the local motion embedding and obtain the local motion feature $e_{1:k} = [e_1, ..., e_k]$ as the output, with e_i represents the feature for time step i.

3.1.2 Global-range Transformer Encoder

We aim to encode all the N people's motion in the scene. In our method, this only needs to be calculated one time and then can contact with any person's local motion feature to predict the correspond person's future motion. We first apply a linear layer to each person's motion $x^n_{1:k}$ and plus the sinusoidal positional embedding to get the global motion embedding $g^{1:N}_{1:k} = [g^1_1, ..., g^1_k, ..., g^N_1, ..., g^N_k]$ for N persons in k time steps. We use L layers of Transformers to encode the global motion embedding. We apply the Multi-head Attention mechanism similar

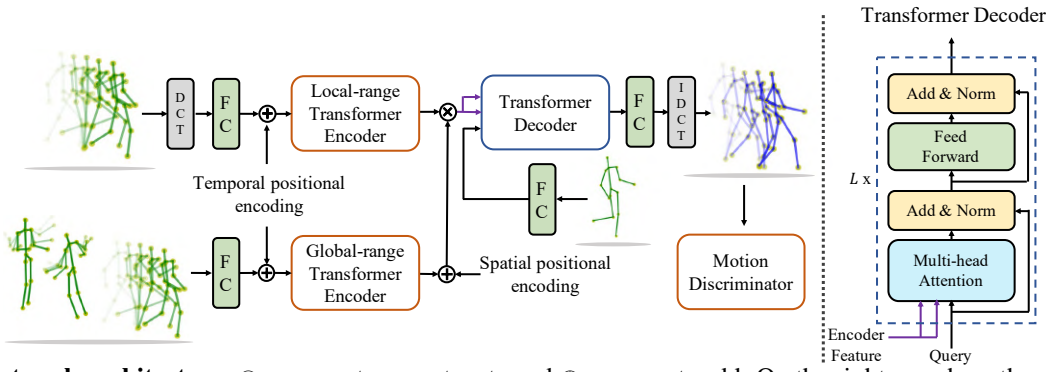

Figure 1: **Network architecture.** $\otimes$ represents concatenate and $\oplus$ represents add. On the right, we show the architecture of the Transformer decoder. The encoder architecture is similar with the decoder except we perform self-attention.

to Eq. 1 to the global embedding which calculated as $Q^i_{glob} = E_{glob}W^{(Q,i)}_{glob}, K^i_{glob} = E_{glob}W^{(K,i)}_{glob}, V^i_{glob} = E_{glob}W^{(V,i)}_{glob}$ where $E_{glob} = [g^1_1, ..., g^1_k, ..., g^N_1, ..., g^N_k]^T$. $W^{(Q,i)}_{glob}$, $W^{(K,i)}_{glob}$ and $W^{(V,i)}_{glob}$ are projection parameter matrices. *glob* represents the global-range. Then we feed them to the normalization and feed-forward layers same as local-range Transformer encoder. After applying such L Transformer encoders to the global embedding, we can get the output $o^{1:N}_{1:k}$. We add our spatial positional encoding to it and get the global motion feature $f^{1:N}_{1:k}$. We concatenate the local $e_{1:k}$ and global $f^{1:N}_{1:k}$ features together as $H = [e_1, ..., e_k, f^1_1, ..., f^1_k, ..., f^N_1, ..., f^N_k]^T$ and feed them to the decoder.

Spatial positional encoding. We propose a spatial positional encoding (SPE) technique on the outputs of the global-range Transformer encoder. Before forwarding to the Transformer decoder, we want to provide the spatial distance between the query token x_k and the tokens of every time step of each person $x^{1:N}_{1:k}$. Intuitively, the location information helps clustering different persons in different social interactions groups, especially in a scene with a crowd of persons. We calculate SPE as,

$$\text{SPE}(x^n_t, x_k) = \exp(-\frac{1}{3J}||x^n_t - x_k||^2) \tag{2}$$

3.1.3 Transformer Decoder

We send the local-global motion feature H together with a static human pose x_k at time step k into the decoder. We use the similar Multi-head attention mechanism as the Transformer encoders. But differently, we take the single pose as the query and use the feature from the encoders to get keys and values. Specifically, We apply a linear layer to x_k and then get q in order to get the query Q_{dec}, we get the key K_{dec} and value V_{dec} both from the local-global motion feature H as $Q^i_{dec} = q^T W^{(Q,i)}_{dec}, K^i_{dec} = HW^{(K,i)}_{dec}, V^i_{dec} = HW^{(V,i)}_{dec}$ where $W^{(Q,i)}_{dec}$,$W^{(K,i)}_{dec}$ and $W^{(V,i)}_{dec}$ are projection parameter matrices. At the end of the decoder, we apply two fully connected layers followed by Inverse Discrete Cosine Transform (IDCT) [4, 35] and output an offset motion sequence $[\Delta\hat{x}_k, ..., \Delta\hat{x}_{T-1}]$ which can easily lead to the future 3D motion trajectory $\hat{x}_{k+1:T}$. The architecture outputs sequence directly prevents generating freezing motion [31]. Note we also add residual connections and layer normalization between layers.

3.1.4 Motion Discriminator

The design of such encoder-decoder architecture helps to predict the future motion. To ensure a natural and continuous long-term motion, we use a discriminator $\mathcal{D}$ to adversially train the Predictor $\mathcal{P}$. The output motion $\hat{x}_{k+1:T}$ is given as input to the Transformer encoder with the same architecture of the local-range encoder and we further use another two fully connected layers to predict values $\in \{1, 0\}$ representing that $\hat{x}_{k+1:T}$ are real or fake poses. We use the ground-truth future poses to provide as the positive examples. We train the predictor $\mathcal{P}$ and discriminator $\mathcal{D}$ jointly.

3.2. Training

We train our predictor $\mathcal{P}$ with both the reconstruction loss and the adversarial loss as $L_{\mathcal{P}} = \lambda_{rec}L_{rec} + \lambda_{adv}L_{adv}$ where $\lambda_{adv} = 1$ and $\lambda_{rec} = 3 \times 10^{-4}$ are constant coefficients to balance the training loss. We calculate the L_{rec} and L_{adv} as follows,

$$\begin{aligned} L_{rec} &= \frac{1}{T-k}\Sigma^{T-1}_{t=k}||\Delta\hat{x}_t - \Delta x_t||^2 \\ L_{adv} &= \frac{1}{T-k}||\mathcal{D}(\hat{x}_{k+1:T}) - \mathbf{1}||^2 \end{aligned} \tag{3}$$

We train our discriminator $\mathcal{D}$ following [27] with loss $L_{\mathcal{D}}$,

$$L_{\mathcal{D}} = \frac{1}{T-k}||\mathcal{D}(\hat{x}_{k+1:T})||^2 + \frac{1}{T-k}||\mathcal{D}(y_{k+1:T}) - \mathbf{1}||^2 \tag{4}$$

where $\hat{x}_{k+1:T}$ is from the predicted motion and $y_{k+1:T}$ is from the real motion. We train the discriminator that classifies the real ones as $\mathbf{1}$, where $\mathbf{1} \in \mathbb{R}^{T-k}$ represents all the poses are natural.

4. Experiments

We perform our experiments on multiple datasets with two settings. The first setting consists of a small number of people ($2 \sim 3$). We use CMU-Mocap as the training data. We mix and make all the CMU-Mocap data consists

method	CMU-Mocap (3 persons)			MuPoTS-3D (2 $\sim$ 3 persons)			3DPW (2 persons)			Mix1 (9 $\sim$ 15 persons)			Mix2 (11 persons)		
	1 s	2s	3s	1 s	2s	3s	1 s	2s	3s	1 s	2s	3s	1 s	2s	3s
LTD [35]	1.37	2.19	3.26	1.19	1.81	2.34	4.67	7.10	8.71	2.10	3.19	4.15	1.72	2.58	3.45
HRI [34]	1.49	2.60	3.07	0.94	1.68	2.29	4.07	6.32	8.01	1.80	3.14	4.21	1.60	2.71	3.67
SocialPool [2]	1.15	2.71	3.90	0.92	1.67	2.51	4.17	7.17	9.27	1.85	3.39	4.84	1.72	3.06	4.26
Ours w/o Global	0.99	1.71	2.50	0.92	1.67	2.50	4.17	6.85	8.91	1.77	3.10	4.19	1.42	2.29	3.06
Ours w/o $\mathcal{D}$	1.13	1.84	2.57	0.92	1.62	2.26	4.17	6.41	8.09	1.75	3.00	4.00	1.34	2.19	2.95
Ours w/o SPE	1.05	1.68	2.37	0.92	**1.51**	2.23	3.92	6.18	**7.79**	1.75	3.09	4.13	1.31	2.15	2.92
Ours	**0.96**	**1.57**	**2.18**	**0.89**	1.59	**2.22**	**3.87**	**6.12**	7.83	**1.73**	**2.99**	**3.97**	**1.29**	**2.09**	**2.82**

Table 1: MPJPE on different datasets. We compare the MPJPE with the previous SOTA methods and ablative baselines of predicting 1, 2 and 3 seconds motion. Best results are shown in boldface.

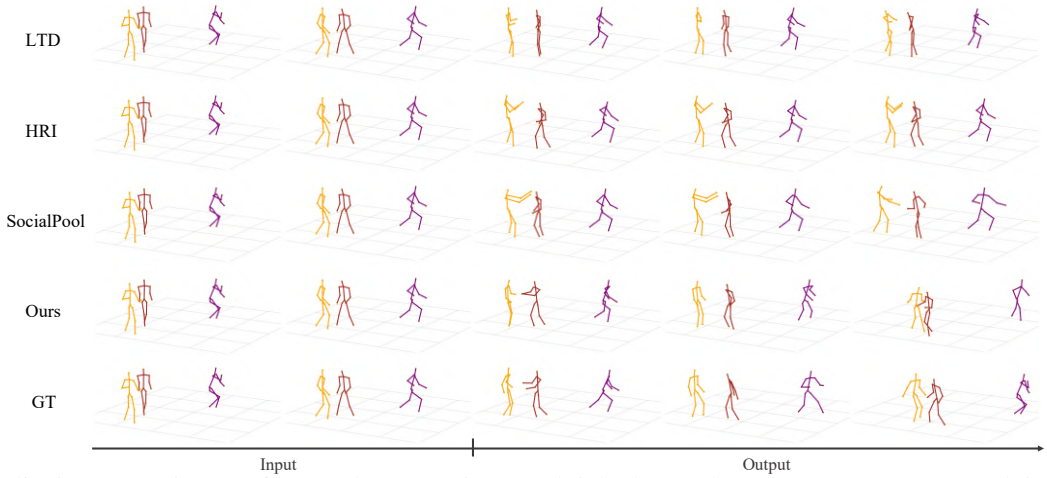

Figure 2: Qualitative comparison. Left two columns are input and right three columns are outputs. Our result is the closest to the real record and the others fail to predict a walking motion and predict a less accurate interaction motion.

of 3 persons in each scene. We sample test set from CMU-Mocap in a similar way. We also test on MuPoTS-3D and the 3DPW dataset with the model trained on the CMU-Mocap dataset. The second setting consists of scenes with more people ($9 \sim 15$). For the training data, We sample motions from CMU-Mocap and Panoptic and then mix them. For the test data, we sample one version from both CMU-Mocap and Panoptic, namely Mix1. And we sample one version from CMU-Mocap, MuPoTS-3D and 3DPW, namely Mix2. In our experiment, we take 1-second motion as input and predict the future 3 seconds(15 fps).

We use Mean Per Joint Position Error (MPJPE) [25] without aligning as the metric to compare the prediction results in 1, 2 and 3 seconds. We select two competitive state-of-the-art single person motion prediction methods: LTD [35] is a graph-based method and HRI [34] is an attention-based method. Most relevant to our work is SocialPool [2], a method uses social pool to model the interaction.

We report MPJPE in 0.1 meters of 1, 2 and 3 seconds predicted motion on different datasets in Tab. 1. In both cases with a small number and a large number of people, our method achieves state-of-the-art performance for different prediction time lengths. We achieve up to 20% improvement when compared to the previous single-person-based methods [35, 34] and achieve up to 30% improvement compared to the multi-person-based method [2]. In SocialPool [2], the same global feature is added to all the persons which interferes with the model's prediction for each individual, especially when there are a large number of people. However, in our design the model can use the features corresponding to one person to query the global motion feature which automatically allows it to use the motion information belonging to other persons. We also perform ablation study on different modules of our network by removing the global-range encoder, discriminator and the spatial positional encoding respectively to prove the effectiveness of each module.

We provide qualitative comparisons on CMU-Mocap in Fig. 2. Our predictions are more natural and smooth while being close to the real record. SocialPool [2] will quickly produce freezing motion, which is consistent with the claims in [5, 31]. Decoding based on an input seed sequence (HRI [34]) or adding the input sequential residual (LTD [35]) to the output, will make the predicted motion have hysteresis and repeat the history. For example, in a forward motion, the prediction may jump back into temporally unreasonable position and then continue to move forward. However, our method, using a static pose as query and predicting a Δx sequence, could solve this problem effectively.

Conclusion. In this paper, we propose a novel framework to predict multi-person 3D motion. We design a Multi-Range Transformer architecture which outperforms state-of-the-art on long-term 3D motion prediction.

References

[1] Cmu graphics lab motion capture database. http://mocap.cs.cmu.edu/. 1

[2] 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, 4

[3] 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. *arXiv preprint arXiv:2104.04029*, 2021. 2

[4] Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1):90–93, 1974. 3

[5] Emre Aksan, Peng Cao, Manuel Kaufmann, and Otmar Hilliges. Attention, please: A spatio-temporal transformer for 3d human motion prediction. *arXiv preprint arXiv:2004.08692*, 2020. 2, 4

[6] 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

[7] Alexandre Alahi, Vignesh Ramanathan, and Li Fei-Fei. Socially-aware large-scale crowd forecasting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2203–2210, 2014. 2

[8] Javad Amirian, Jean-Bernard Hayet, and Julien Pettré. Social ways: Learning multi-modal distributions of pedestrian trajectories with gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[9] 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

[10] 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

[11] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014. 2

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[14] Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection. *Neural networks*, 108:466–478, 2018. 2

[15] 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. 2

[16] 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. 2

[17] Anand Gopalakrishnan, Ankur Mali, Dan Kifer, Lee Giles, and Alexander G Ororbia. A neural temporal model for human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12116–12125, 2019. 2

[18] 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. 2

[19] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018. 2

[20] Lei Tan Lin Gui Bart Nabbe Iain Matthews Takeo Kanade Shohei Nobuhara Hanbyul Joo, Hao Liu and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, 2015. 2

[21] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995. 2

[22] 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. 2

[23] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[24] 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

[25] 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. 4

[26] 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. 2

[27] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[28] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, S Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *arXiv preprint arXiv:1907.03395*, 2019. 2

[29] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 336–345, 2017. 2

[30] 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

[31] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Learn to dance with aist++: Music conditioned 3d dance generation, 2021. 3, 4

[32] Zimo Li, Yi Zhou, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. *arXiv preprint arXiv:1707.05363*, 2017. 2

[33] Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. It is not the journey but the destination: Endpoint conditioned trajectory prediction. In *European Conference on Computer Vision*, pages 759–776. Springer, 2020. 2

[34] 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, 2, 4

[35] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9489–9497, 2019. 1, 2, 3, 4

[36] 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. 2

[37] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 935–942. IEEE, 2009. 2

[38] 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. 1

[39] Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14424–14432, 2020. 2

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

[41] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th International Conference on Computer Vision*, pages 261–268. IEEE, 2009. 2

[42] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978. 1

[43] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1349–1358, 2019. 2

[44] Tianmin Shu, Xiaofeng Gao, Michael S Ryoo, and Song-Chun Zhu. Learning social affordance grammar from videos: Transferring human interactions to human-robot interactions. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 1669–1676. IEEE, 2017. 2

[45] Tianmin Shu, Michael S Ryoo, and Song-Chun Zhu. Learning social affordance for human-robot interaction. *arXiv preprint arXiv:1604.03692*, 2016. 2

[46] Jianhua Sun, Qinhong Jiang, and Cewu Lu. Recursive social behavior graph for trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 660–669, 2020. 2

[47] Jianhua Sun, Yuxuan Li, Hao-Shu Fang, and Cewu Lu. Three steps to multimodal trajectory prediction: Modality clustering, classification and synthesis. *arXiv preprint arXiv:2103.07854*, 2021. 2

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 1, 2

[49] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018. 1

[50] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. *arXiv preprint arXiv:2012.05522*, 2020. 2

[51] Kota Yamaguchi, Alexander C Berg, Luis E Ortiz, and Tamara L Berg. Who are you with and where are you going? In *CVPR 2011*, pages 1345–1352. IEEE, 2011. 2

[52] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *European Conference on Computer Vision*, pages 507–523. Springer, 2020. 2

[53] Ye Yuan and Kris Kitani. Diverse trajectory forecasting with determinantal point processes. *arXiv preprint arXiv:1907.04967*, 2019. 2

[54] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. Simpoe: Simulated character control for 3d human pose estimation. *arXiv preprint arXiv:2104.00683*, 2021. 2

[55] Jason Y Zhang, Panna Felsen, Angjoo Kanazawa, and Jitendra Malik. Predicting 3d human dynamics from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7114–7123, 2019. 2

[56] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. *arXiv preprint arXiv:2012.00619*, 2020. 2

[57] Bolei Zhou, Xiaogang Wang, and Xiaoou Tang. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2871–2878. IEEE, 2012. 2