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

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. We visualize part of the prediction results with multi-person interactions in Fig. 1. More Videos are available at: https://anonymousaut.github.io/Anonymous-Result/

Given a scene with N persons and their corresponding history motion, our goal is to predict their future 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.

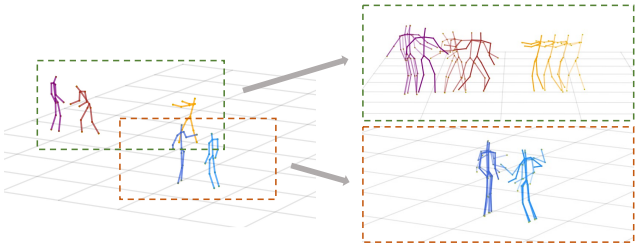

Figure 1: Our motion prediction results. **Left**: The last time step of the input sequence. **Right**: The predicted motion with multi-person social interactions.

We perform our experiments on multiple datasets. We evaluate on CMU-Mocap [1], MuPoTS-3D [6] and the 3DPW [7] dataset and the mix of them, namely Mix1 and Mix2. We use Mean Per Joint Position Error (MPJPE) [3] 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 [5] is a graph-based method and HRI [4] 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 [5, 4] and achieve up to 30% improvement compared to the multi-person-based method [2].

	CMU-Mocap			MuPoTS-3D			3DPW			Mix1			Mix2		
	1 s	2s	3s	1 s	2s	3s	1 s	2s	3s	1 s	2s	3s	1 s	2s	3s
LTD [5]	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 [4]	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.

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. 1

[3] 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. 1

[4] 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

[5] 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

[6] 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

[7] 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