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

Jiashun Wang
UC San Diego

Huazhe Xu
UC Berkeley

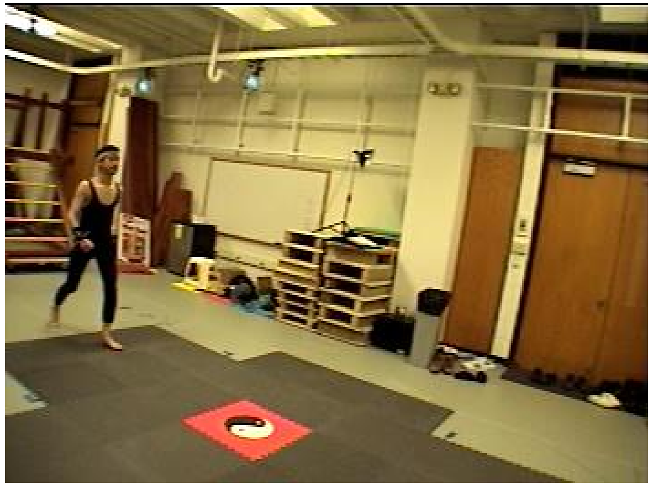
Medhini Narasimhan
UC Berkeley

Xiaolong Wang
UC San Diego

Webpage:

Background

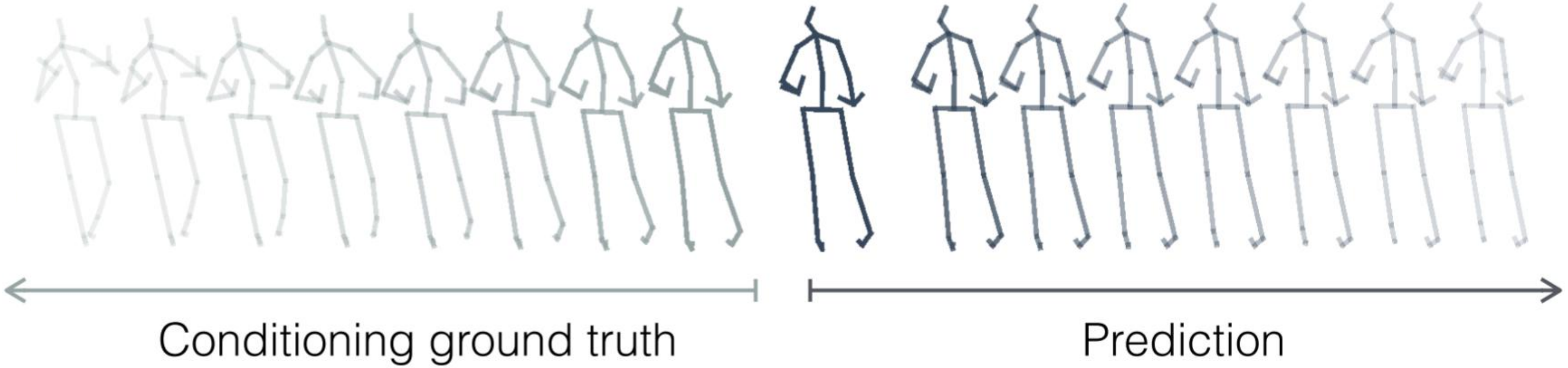
- Modeling 3D human motion has been a long-standing problem in computer vision and computer graphics commuity



http://mocap.cs.cmu.edu/

Background

- Predicting future motion sequences given a sequence of history

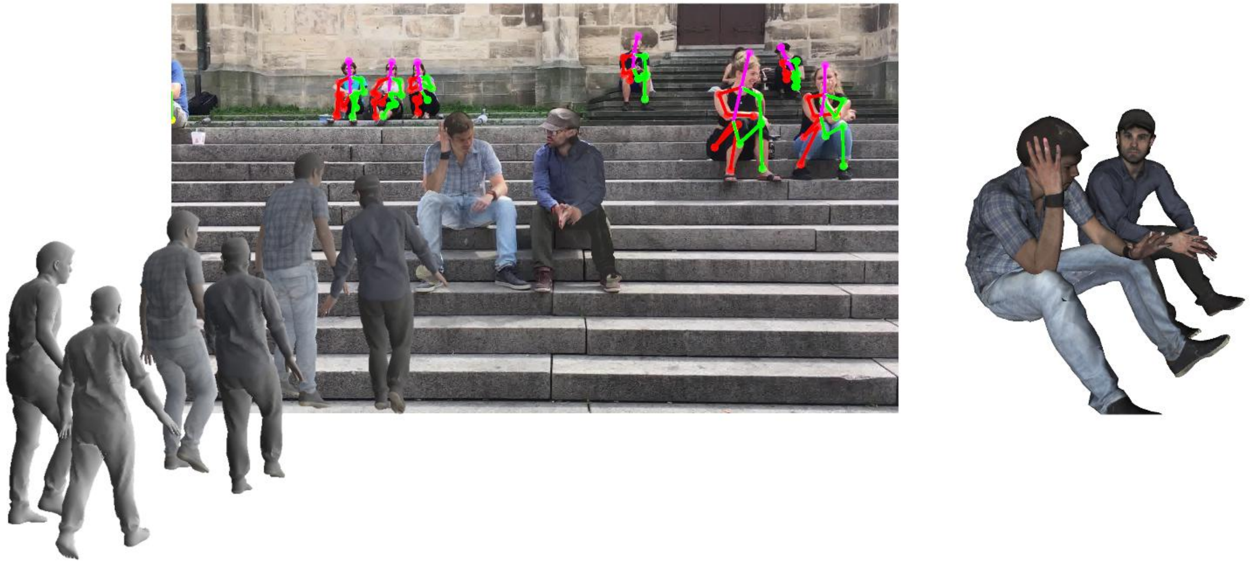


  - Focusing on single person motion
  - Usually neglecting the movement of the root joint

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.

Background

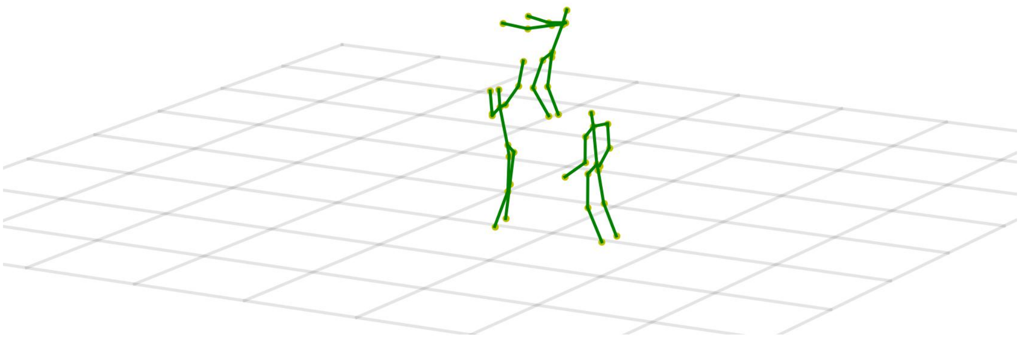
- Multi-person motion prediction is relatively under-explored and more challenging
  - Considering multi-person interaction
  - Modeling pose and trajectory jointly is needed, e.g. catching

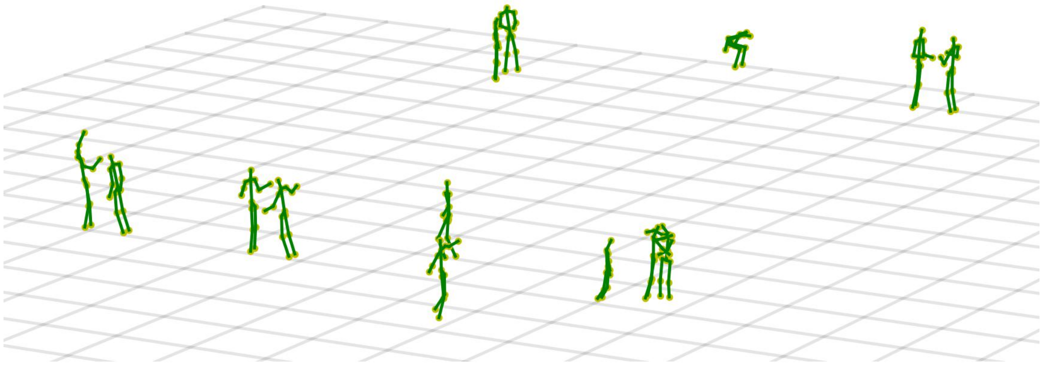


Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In European Conference on Computer Vision (ECCV), sep 2018. 2

Our task

- Given a scene with N persons and their corresponding history motion, we aim to predict their future 3D motion.





Green represents the input and Blue represents the output

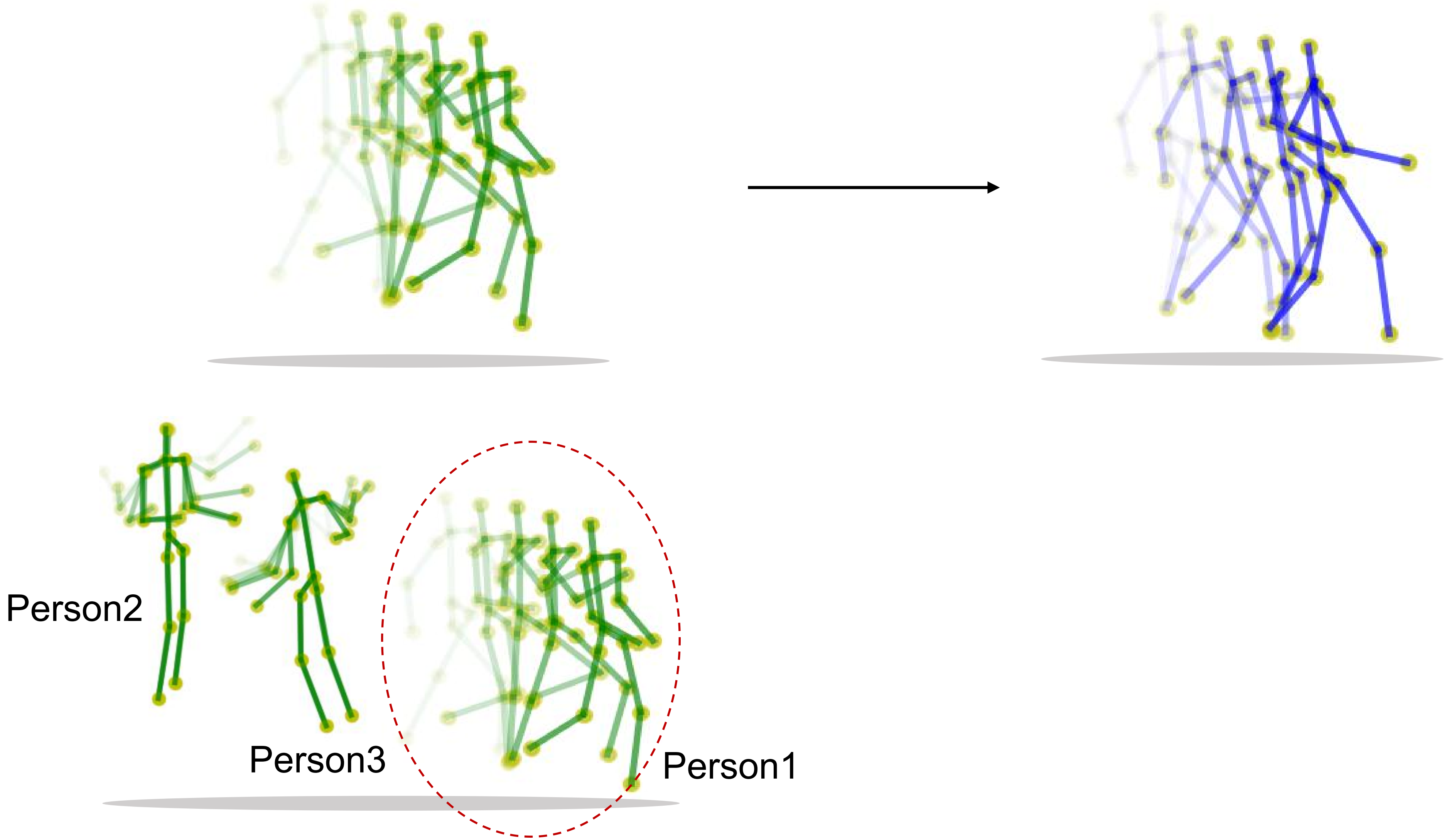
Representation

Input: $X_{1:k}^{n} = [x_1^n,\ ..., x_k^n]$

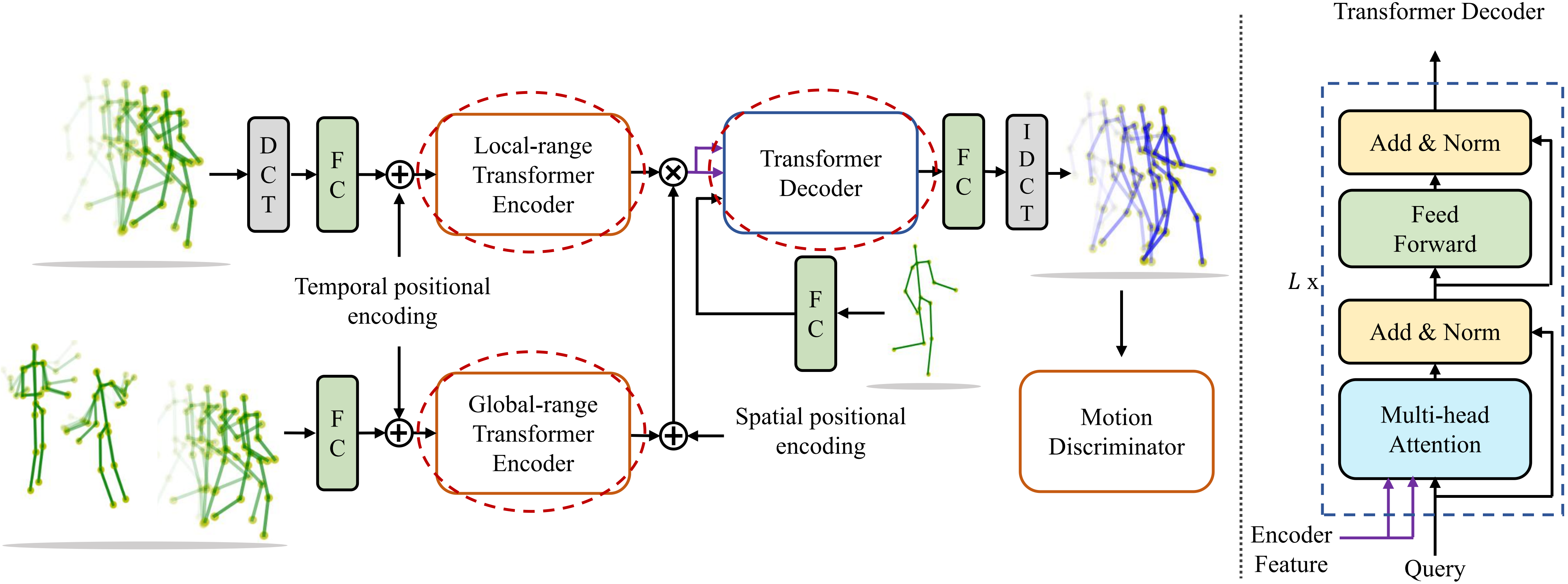
Output: $X_{k+1:T}^{n}$

$x_k^n \in R^{3J}$ represent the pose of the person n at time k

Method

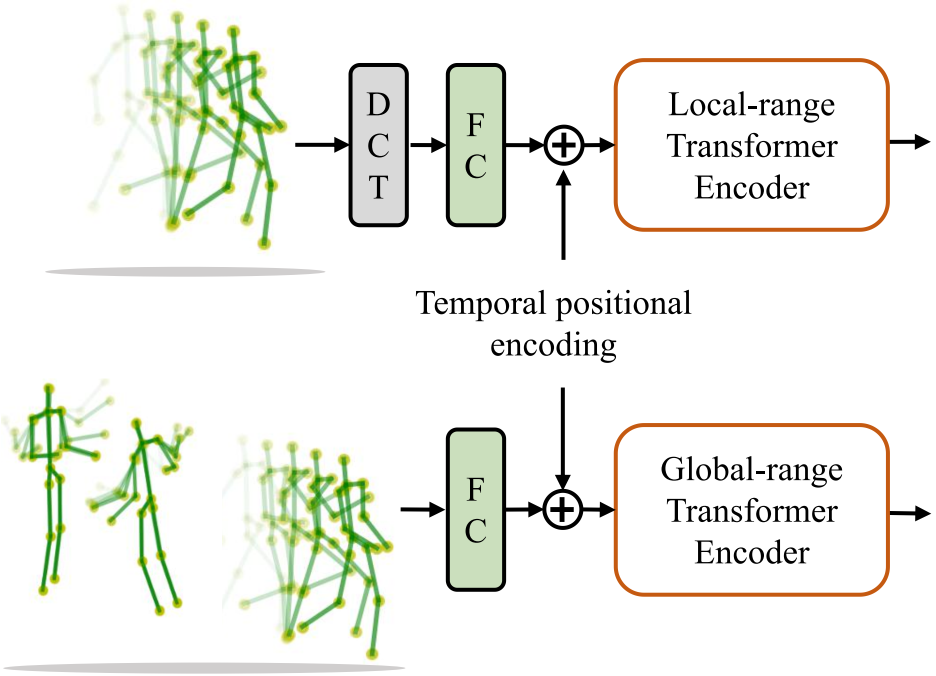


Method



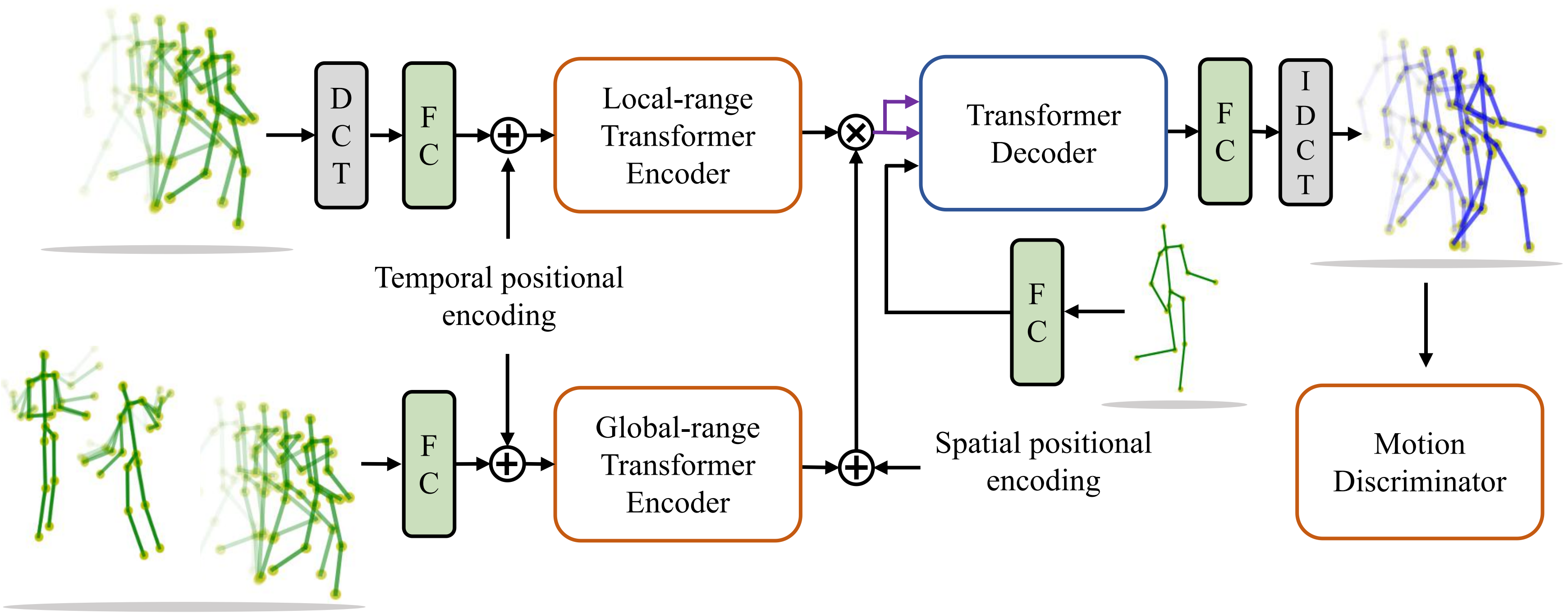
- We propose our Multi-Range Transformers to solve the problem
  - Local-range transformer encoder
  - Global-range transformer encoder
  - Transformer decoder

Method



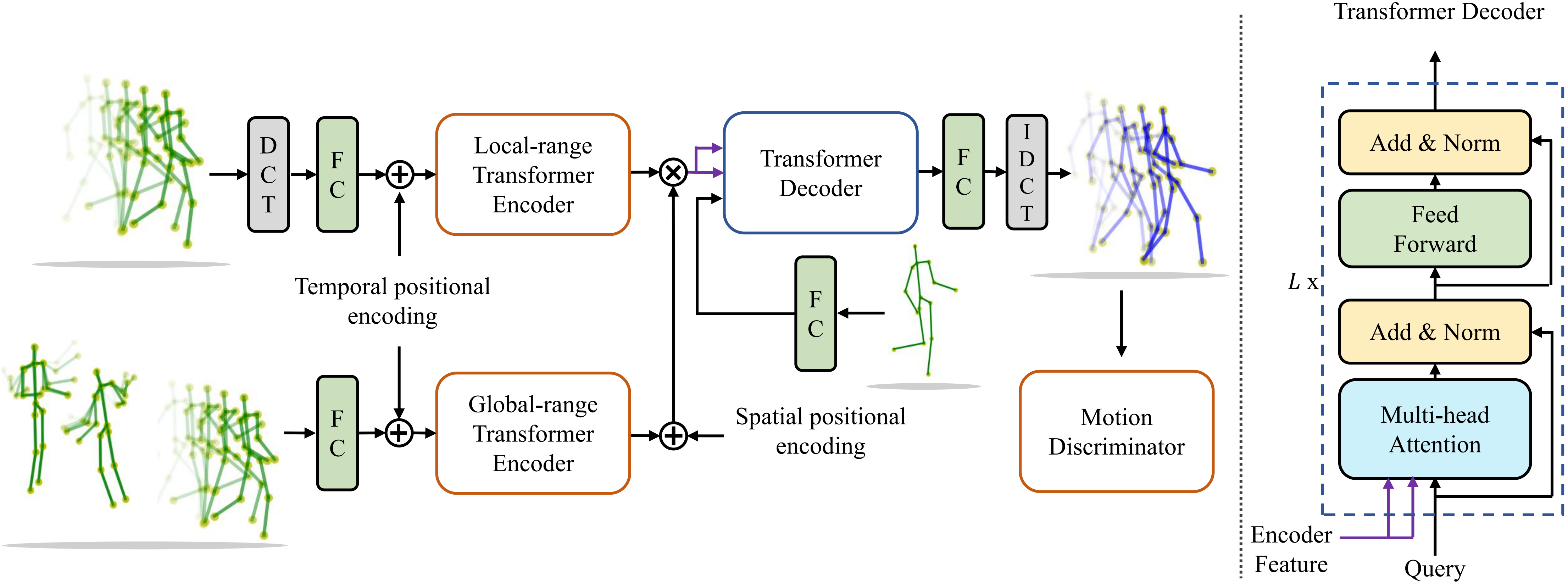
- Local-range Transformer encoder: individual motion
- Global-range Transformer encoder: global motion

Method



- A single pose as the query
- Decoder outputs a sequence directly

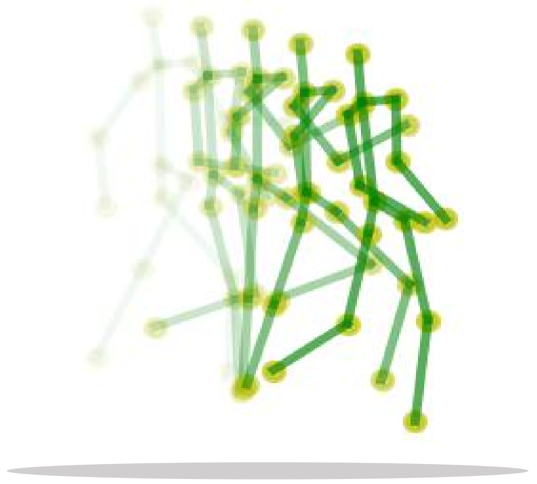
Method

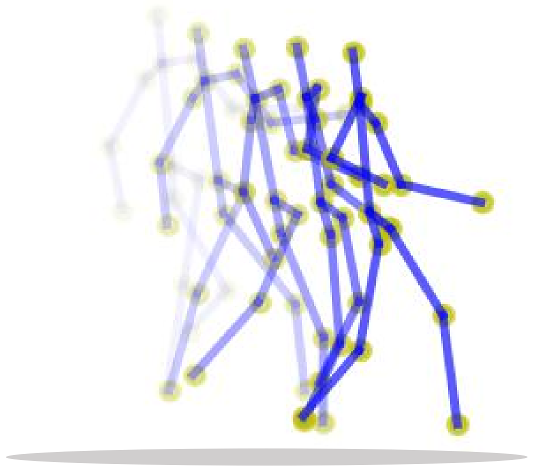


- On the right, we show the architecture of the Transformer decoder.
- The transformer encoder performs self-attention

Why local-range and global-range?

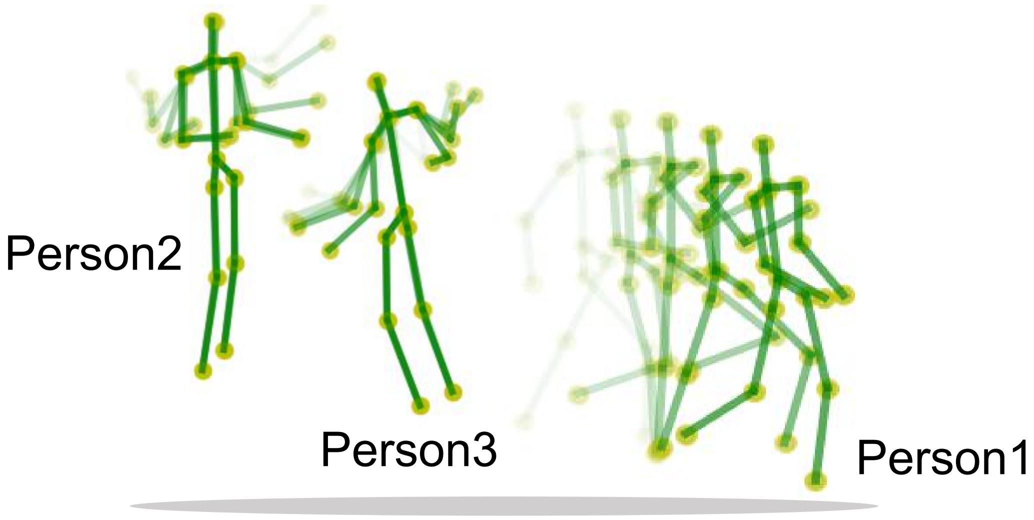
- Local
  - The task of synthesizing a natural motion based on previous states itself is actually a challenging task
  - To ensure the smoothness of the motion, the model requires dense sampling of the input sequence





Why local-range and global-range?

- Global
  - The interaction of all the persons in the whole scene, sparse sampling of the sequences are used
  - Compute the global feature once



Method

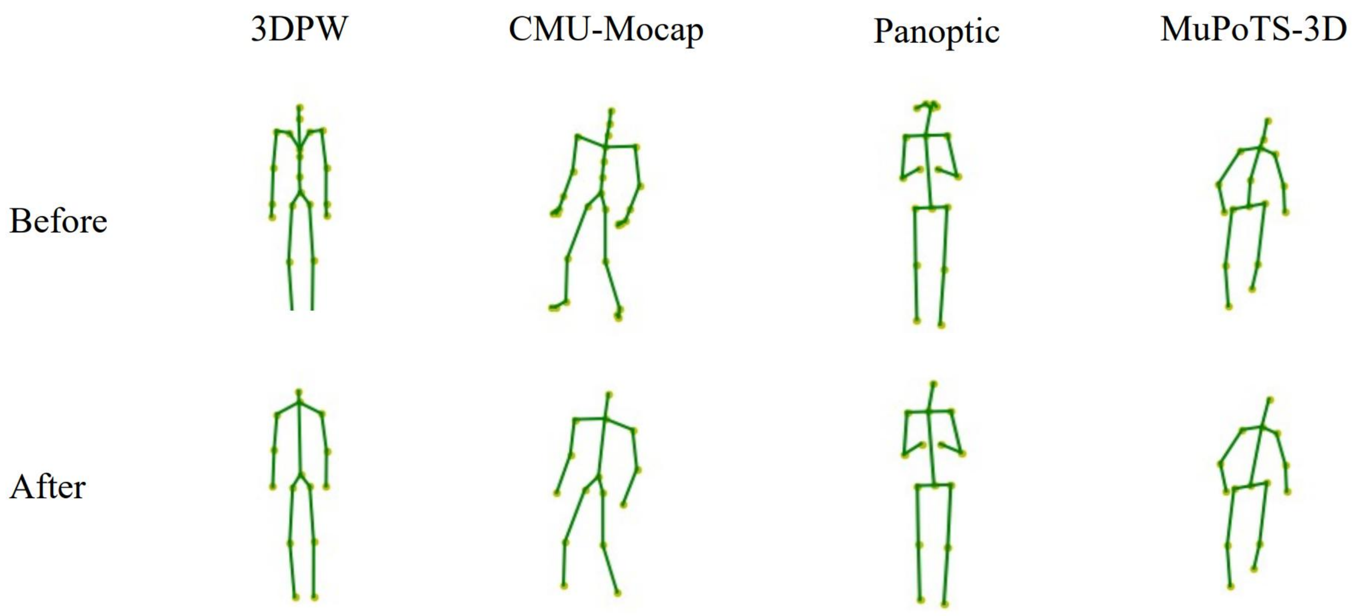
- Spatial Positional Encoding (SPE)
  - SPE encodes the spatial distance between the query token x_k and the tokens of every time step of each person $x_{1:k}^{1:N}$

$$\mathrm{SPE}(x_t^n, x_k) = \exp(-\frac{1}{3J}||x_t^n - x_k||_2^2)$$

  - Helpful in a scene with a crowd of persons

Experiment

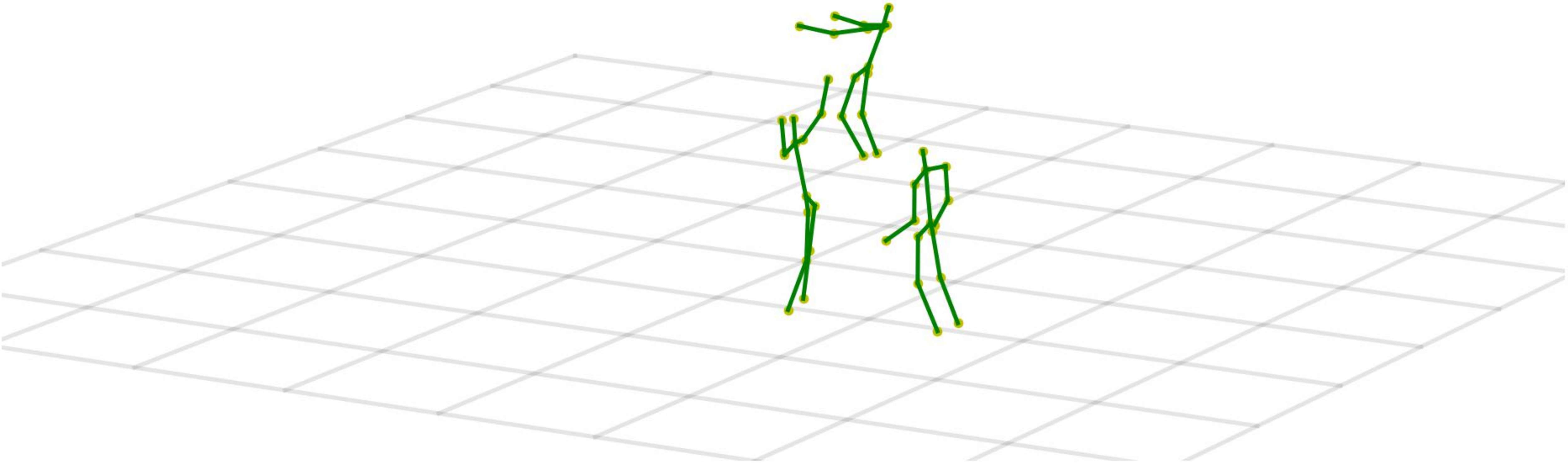
- Data
  - 2-3 persons (3DPW, CMU-Mocap and MuPoTS-3D)
  - 9-15 persons (Mix1 and Mix2)



Qualitative results

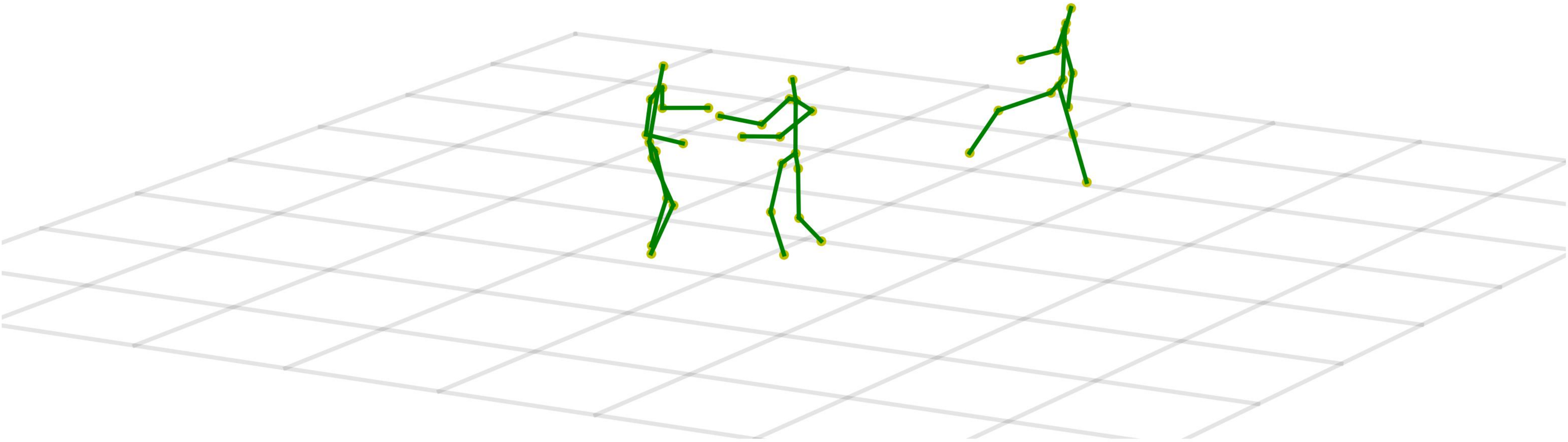
We show some examples of the multi-person motion prediction results

Example 1 (3 persons)



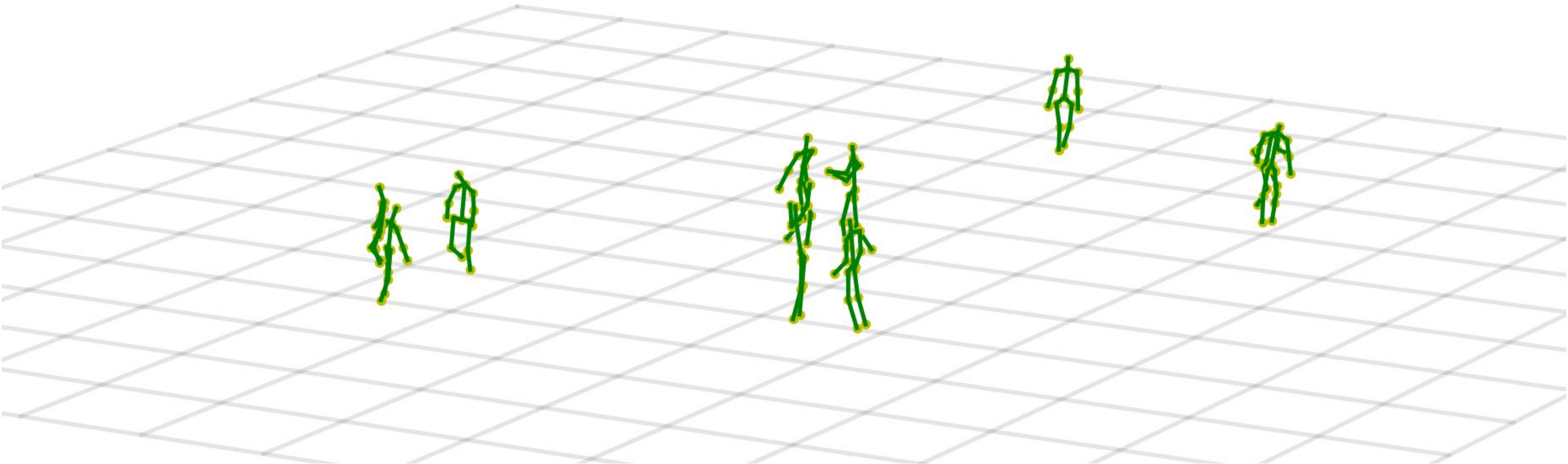
Green represents the input and Blue represents the output

Example 2 (3 persons)



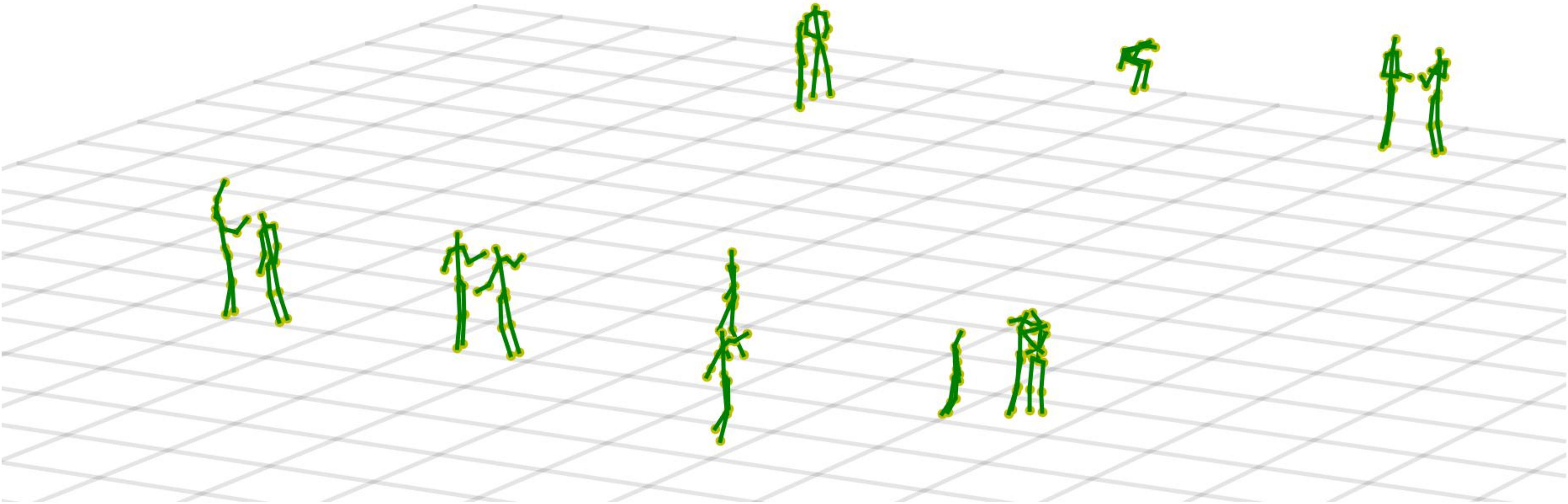
Green represents the input and Blue represents the output

Example 3 (10 persons)



Green represents the input and Blue represents the output

Example 4 (14 persons)



Green represents the input and Blue represents the output

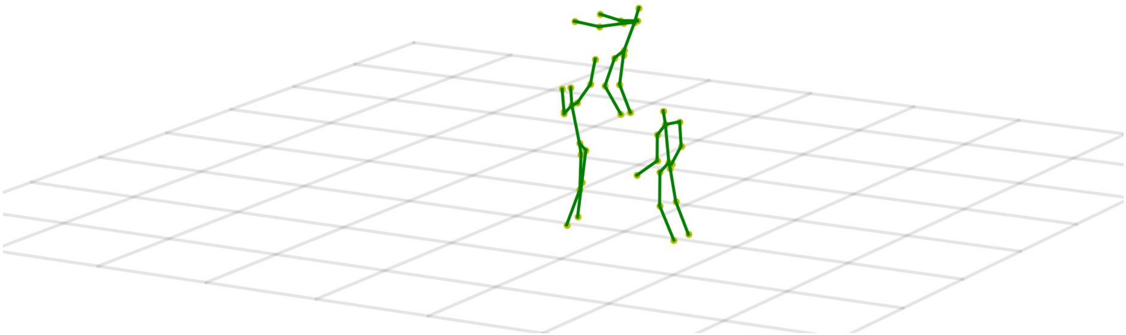
Qualitative results

We compare our method with the other methods on different datasets

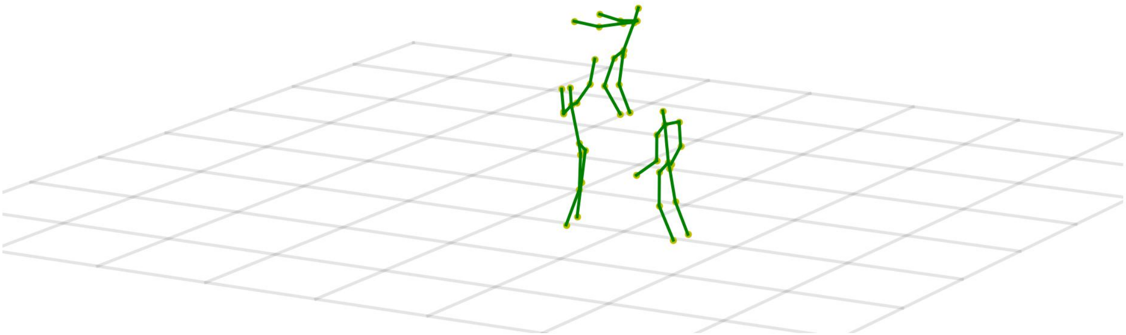
Example 1

LTD is affected by the past positions.

Ours

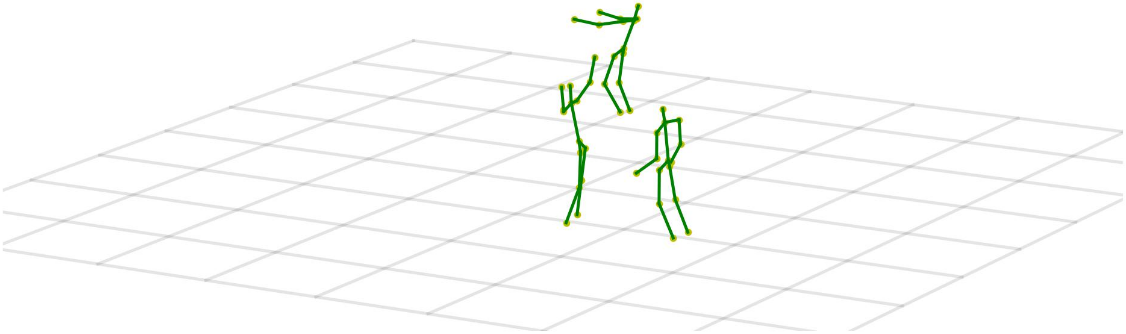


LTD



Green represents the input and Blue represents the output.
Input data is from CMU-Mocap.

GT

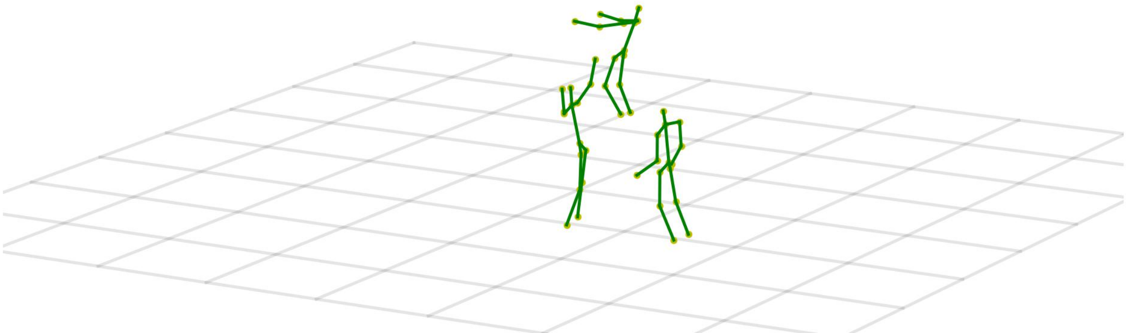


Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision pages 9489–9497, 2019

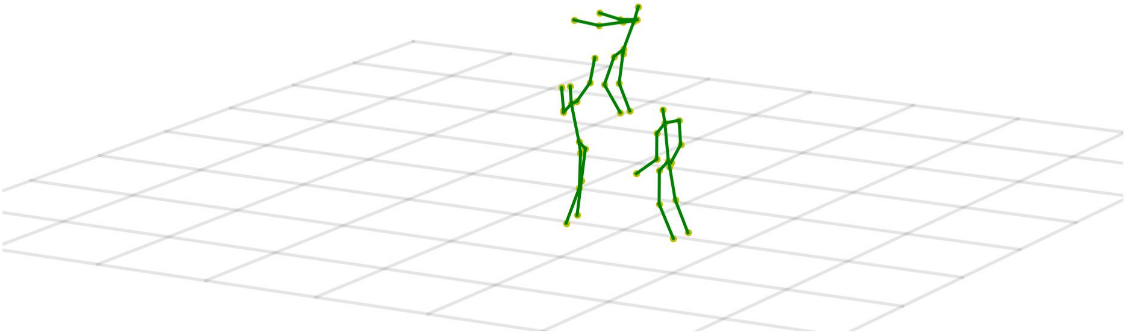
Example 1

HRI is affected by the past positions.

Ours

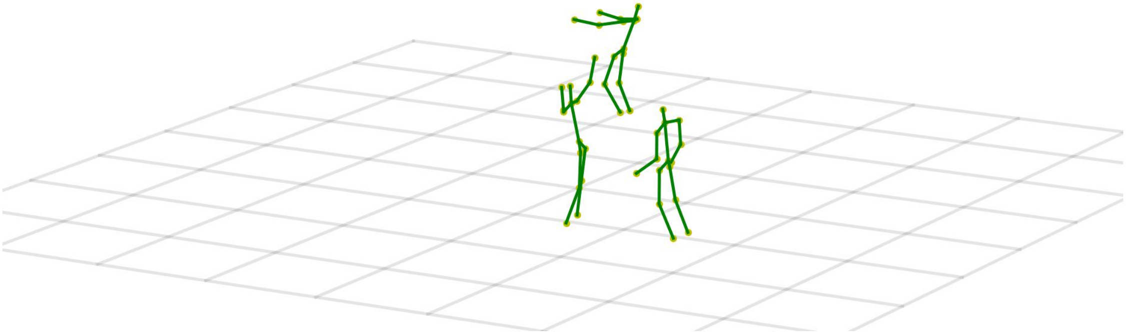


HRI



Green represents the input and Blue represents the output.
Input data is from CMU-Mocap.

GT

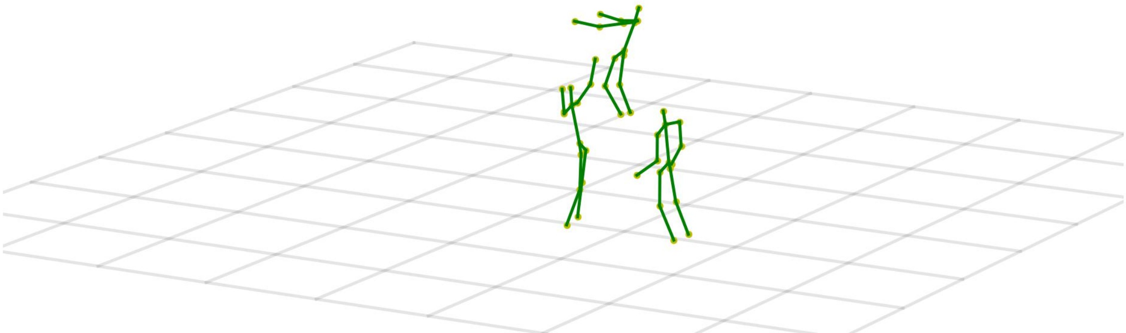


Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In European Conference on Computer Vision, pages 474–489. Springer, 2020.

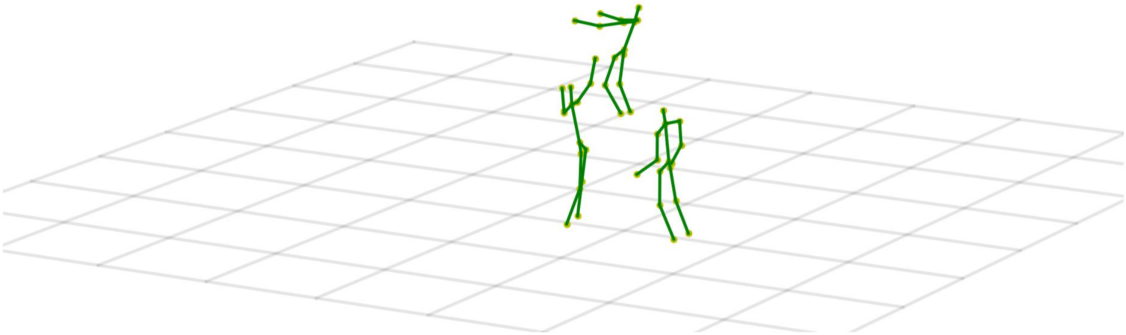
Example 1

SocialPool predicts freezing motions quickly.

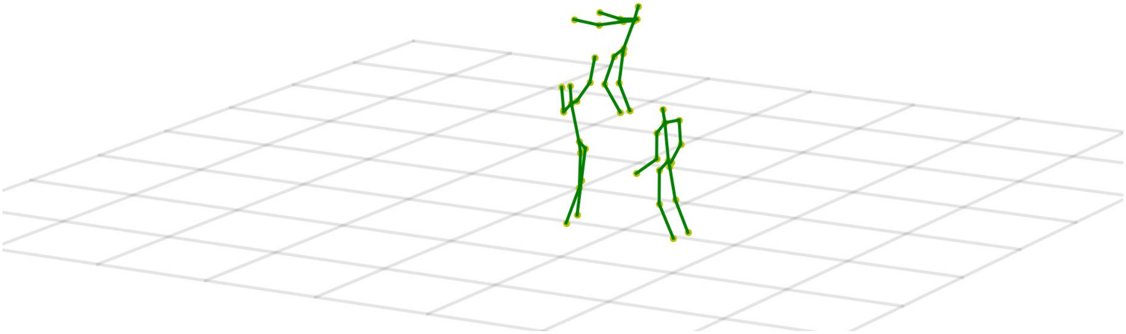
Ours



SocialPool



GT



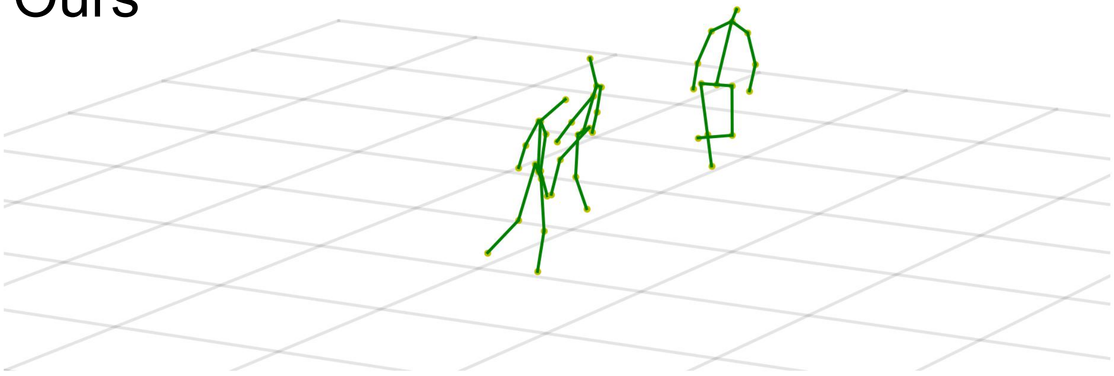
Green represents the input and Blue represents the output.
Input data is from CMU-Mocap.

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.

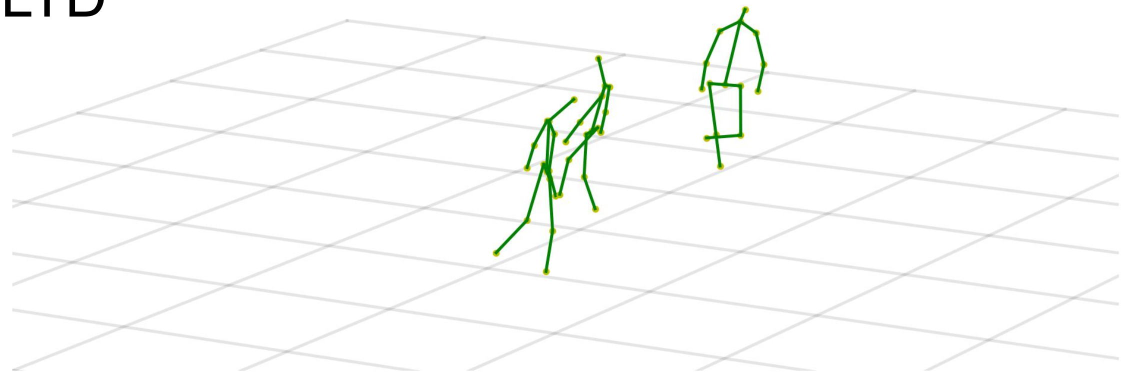
Example 2

LTD fails to predict the correct walking motions.

Ours



LTD



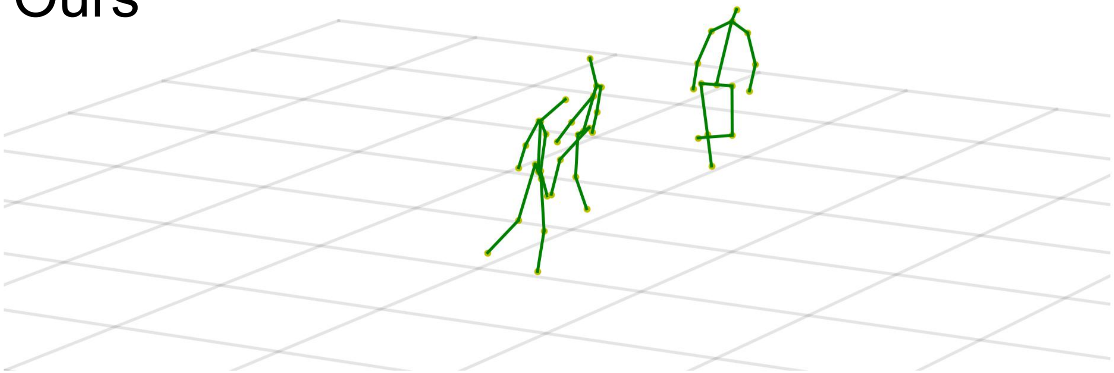
GT

Green represents the input and Blue represents the output.
Input data is from MuPoTS-3D.

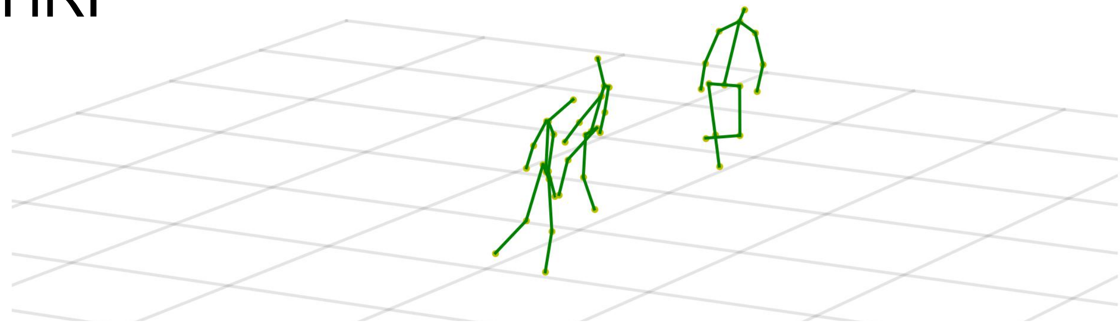
Example 2

HRI predicts freezing motions quickly.

Ours



HRI



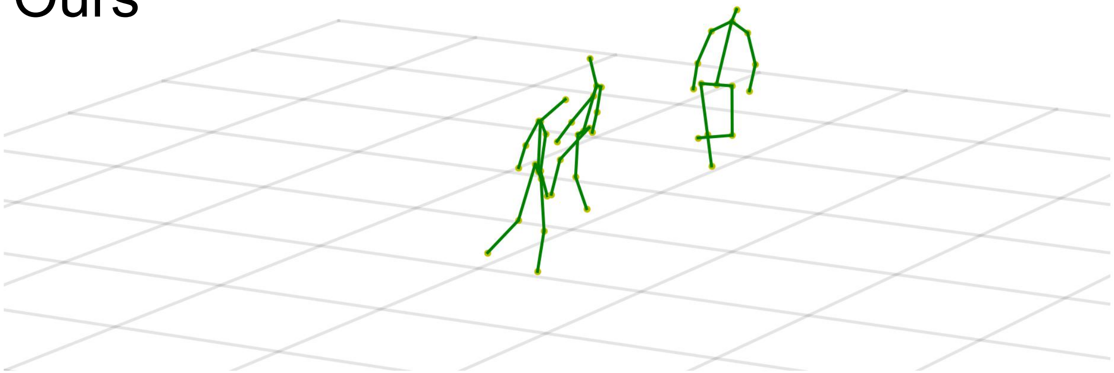
Green represents the input and Blue represents the output.
Input data is from MuPoTS-3D.

GT

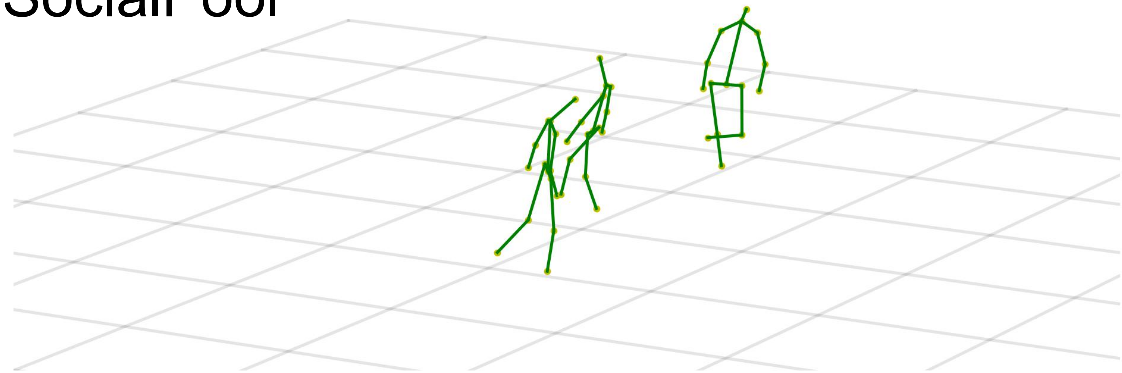
Example 2

SocialPool predicts freezing motions quickly.

Ours

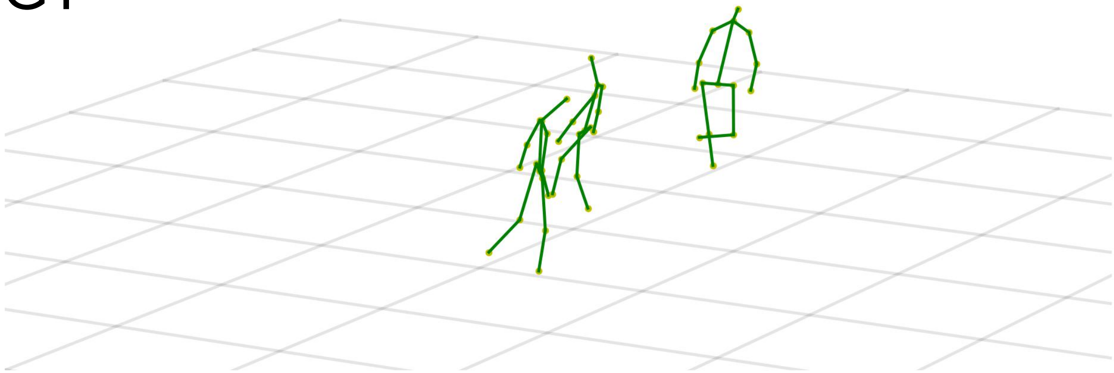


SocialPool



Green represents the input and Blue represents the output.
Input data is from MuPoTS-3D.

GT



Quantitative results

We compare the mean per joint position error(MPJPE), user study and moving distance with the other methods.

MPJPE

	CMU-Mocap 3 persons			Mix1 9~15 persons		
	1s	2s	3s	1s	2s	3s
LTD	1.37	2.19	3.26	2.10	3.19	4.15
HRI	1.49	2.60	3.07	1.80	3.14	4.21
SocialPool	1.15	2.71	3.90	1.85	3.39	4.84
Ours	**0.96**	**1.57**	**2.18**	**1.73**	**2.99**	**3.97**

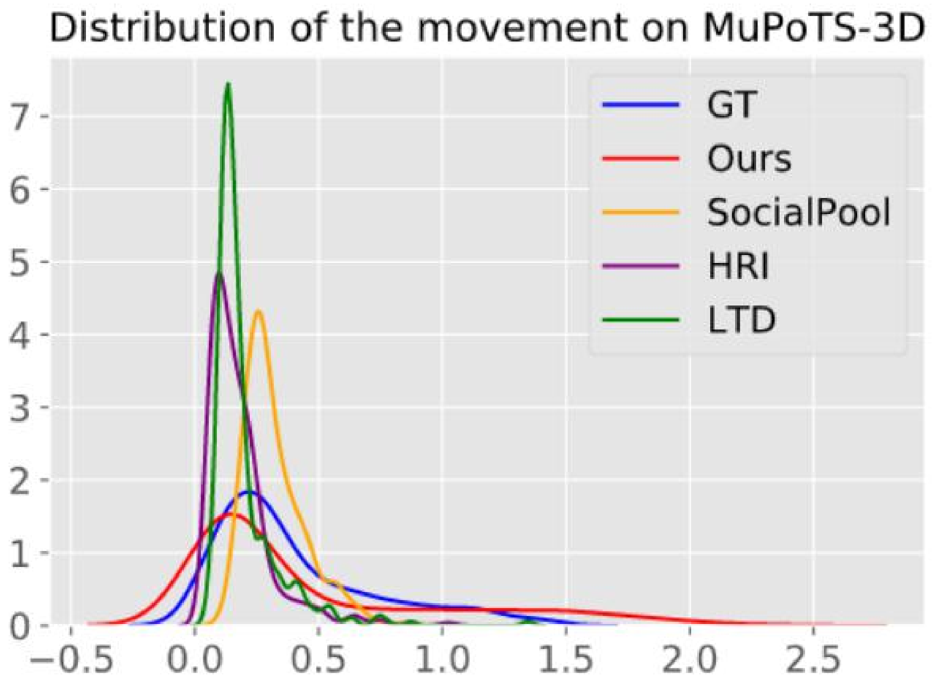
- We report the MPJPE in 0.1 meters of 1 second, 2 seconds and 3 seconds motion.
- In both cases with a small number and a large number of people, our method achieves state-of-the-art performance for different prediction time lengths.

User Study

	Mix1 9~15 persons	**Mix2** 11 persons
LTD	3.71±0.93	3.75±0.90
HRI	3.67±0.89	3.71±0.90
SocialPool	3.62±0.92	3.49±1.02
Ours	**3.74**±**0.83**	**3.77**±**0.82**
GT	3.77±0.81	3.88±0.79

- We report the average and the standard error of the score.
- Our results get better reviews consistently cross all datasets

Distribution of the movement



- We compare the distribution of the movement between the start and end of the outputs.
- Other methods intend to predict a motion with less movement while ours is the most closest to the ground truth.

Thank you!