Supplementary Materials: Geometry-Aware Deep Learning for 3D Skeleton-Based Motion Prediction

Mayssa Zaier, Hazem Wannous
IMT Nord Europe
Lille, France
mayssa.zaier@imt-nord-europe.fr, hazem.wannous@imt-nord-europe.fr

Hassen Drira
University of Strasbourg
Strasbourg, France
hdrira@unistra.fr

1. Description of the logarithmic and exponential maps

It is helpful to define two Riemannian geometric tools: one for mapping points from the pre-shape space to a tangent space, and another for mapping points from a tangent space back to the pre-shape space. A pictorial description is given in Fig. 1.

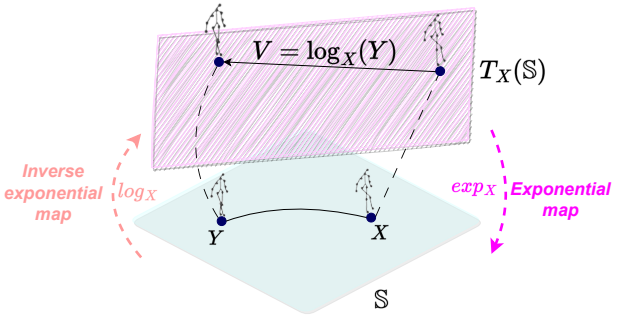

Figure 1. Illustration of the logarithmic mapping to $T_X(\mathbb{S})$

Let $X, Y \in R^{n\times k}$ be two skeletons residing on the n-Sphere $\mathbb{S}$. The first task can be achieved via the *logarithmic map* also known as the *inverse exponential mapping* which allows the unique mapping of any given point residing on the manifold $\mathbb{S}$ to the tangent space, $\log_X : S \to T_X(\mathbb{S})$, defined as Eq. (1) (for $X, Y \in \mathbb{S}$) where $\theta = \cos^{-1}(\langle X, Y\rangle)$ is the arc-length distance between X and Y on the sphere $\mathbb{S}$. The second task is carried out via the *exponential map*, $\exp_X : T_X(\mathbb{S}) \to \mathbb{S}$, defined as Eq. (2) (for $X \in \mathbb{S}$ and $V \in T_X(\mathbb{S})$), where $||V|| = \sqrt{V^T V}$.

$$\log_X(Y) = \frac{\theta}{\sin(\theta)}(Y - \cos(\theta)X) \tag{1}$$

$$Y = \cos(||V||)X + \sin(V)\frac{V}{||V||} \tag{2}$$

2. Evaluation Metrics

To rigorously assess the performance of our pose estimation models, we have used two evaluation metrics: the *Mean Per Joint Position Error (MPJPE)* and the *Mean Angle Error (MAE)*. These metrics have been instrumental in quantifying the accuracy of joint position predictions and orientation estimations, respectively. Employing these metrics has enabled us to conduct a detailed and nuanced evaluation of our models, ensuring that our findings are both robust and reliable. The MPJPE is calculated as Eq. (3).

$$\text{MPJPE}(x;\hat{x}) = \frac{1}{N}\sum_{i=1}^{N} \|\hat{x}(i) - x(i)\| \tag{3}$$

where N is the number of processed joints, $\hat{x}(i)$ the estimated coordinates, and $x(i)$ the ground truth position of the ith joint. For each joint, we compute the Euclidean distance between the estimated coordinate and the ground truth coordinate. Then, we take the mean of these distances across all joints. In contrast, MAE is defined as Eq. (4).

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - x_i| \tag{4}$$

where n is the total number of observations, y_i the observed value, and x_i the predicted value for the ith observation.

For each observation, we compute the absolute difference between the predicted value (x_i) and the actual value (y_i). Then, we take the mean of these absolute differences across all observations.

Why Prefer MPJPE? By covering larger ranges of errors, MPJPE provides a clearer comparison between different models' performance. It's more sensitive to the nuances of human movement, making it a superior choice for tasks that require precise joint localization, such as motion capture and animation. MPJPE is a robust and descriptive metric for evaluating pose estimation models, providing a detailed assessment of positional accuracy. It is generally preferred when a holistic evaluation of a model's accuracy is required, making it an invaluable tool in the advancement of human pose estimation technology.
Unlike the traditional *MAE*, which measures angular deviation, MPJPE calculates the mean Euclidean distance between the estimated and actual joint positions in three-dimensional space. This provides a more nuanced view of a model's performance, capturing the full complexity of human movement.

3. More Quantitative Results

We present the prediction results on all the contained actions in the H3.6m dataset for both short-term and long-term prediction. These results provide sufficient information for a detailed comparison of the algorithm development in future works.

First, we present the MPJPE of various models on H3.6m for short-term motion prediction, where the detailed results of all actions are shown in Table 1. We see that *MAN-TF* obtains superior performance at most timestamps. Compared to the baselines, *MAN-TF* consistently achieves significantly lower MPJPEs on average.

Long-term motion prediction aims to predict the poses over 400 milliseconds, which is challenging due to the pose variation and elusive human intention. Table 2 presents the prediction MPJPEs of various methods at the 560 ms and 1000 ms on all actions. We see that *MAN-TF* achieves more effective prediction on most actions and has lower MPJPEs in average.

MPJPE for 5-Second Prediction Window. The reason for choosing to represent 3D human pose in a non-euclidean space is to achieve more accurate long-term predictions. While we reported the MPJPE errors at 1 second which is the conventional long-term timestamp, it is useful to demonstrate predictions for 5 seconds and beyond.

For the purpose of comparison, we have selected several baseline models. Among the state-of-the-art methods designed for long-term prediction, we chose *HisRep* [9] and *TIM-GCN* [5]. HisRep is evaluated in two variants: *HisRep10* which is recognized as the best model in [9] is trained to output 10 frames and then iteratively uses the predicted frames for extended predictions; and *HisRep125* which directly predicts 125 future frames based on 150 past frames. *TIM-GCN* is a model trained on subsequences of lengths 10, 50, and 100 frames, to predict 125 future frames to accommodate longer-term predictions over 5 seconds. Additionally, we compare against *LTMPUK* [4], *Mix&Match* [1] and *DLow* [14], state-of-the-art methods for multiple long-term motion prediction, both trained to forecast 125 future frames using 100 past frames.

To evaluate the accuracy of the predicted poses over a 5-second prediction window, we calculate the Mean Per Joint Position Error (MPJPE). Given that our model predicts 125 future frames and assuming a frame rate of 25 frames per second (fps), the 5-second duration corresponds to 125 frames. For each frame t in the prediction window, we compute the euclidean distance between the predicted and ground truth joint positions for each joint j. We refer this error by $\text{Error}_{t,j}$. Next, we sum these errors for all joints and average them to obtain the MPJPE for frame t (Eq. (5)) where N is the total number of joints.

$$\text{MPJPE}_t = \frac{1}{N}\sum_{j=1}^{N}\text{Error}_{t,j} \tag{5}$$

Finally, we average the MPJPE values across all 125 frames to derive the overall MPJPE for the 5-second prediction window as follows:

$$\text{MPJPE}_{5\text{sec}} = \frac{1}{125}\sum_{t=1}^{125}\text{MPJPE}_t \tag{6}$$

This metric allows us to quantitatively assess the performance of our models in long-term motion prediction tasks. In table 3, we evaluate the MPJPE losses of the predicted sequences. We observe that our method outperforms the others in having low MPJPE and our performance is noticeably better. Additionally, these experiments confirm the efficacy of combining these two representations.

Methods	5save↓
TIM-GCN [5]	196
HisRep10 [9]	197
HisRep125 [9]	191
Mix&Match [1]	244
DLow [14]	189
LTMPUK [4]	196
Ours (MAN-TF)	**185**
Ours (MAN-TF Lie alone)	194
Ours (MAN-TF Kendall alone)	193

Table 3. Results on the MPJPE errors on the H3.6m dataset. Lower MPJPE values indicate closer predictions to ground truth future motion. We have highlighted close best results in bold.

Table 1. Prediction MPJPEs of various models for short-term motion prediction on H3.6m and the average MPJPEs across all the actions.

Methods	Walking				Eating				Smoking				Discussion			
	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
Res-sup [11]	29.36	50.82	76.03	81.52	16.84	30.60	56.92	68.65	22.96	42.64	70.24	83.68	32.94	61.18	90.92	96.19
CSM [6]	21.70	43.56	66.29	75.48	14.50	26.13	47.47	55.63	19.42	37.70	62.49	68.55	26.35	53.41	79.12	83.01
SkelNet [3]	20.49	34.36	59.64	68.76	11.80	22.38	39.88	48.11	11.33	23.71	45.30	52.85	21.79	40.24	65.93	77.91
DMGNN [8]	17.32	30.67	54.56	65.20	10.96	21.39	36.18	43.88	8.97	17.62	32.05	40.30	17.33	34.78	61.03	69.80
Traj-GCN [10]	12.29	23.03	39.77	46.12	8.36	16.90	33.19	40.70	7.94	16.24	31.90	38.90	12.50	27.40	58.51	71.68
HisRep [9]	10.53	19.96	34.88	42.05	7.39	15.53	31.26	38.58	7.17	14.54	28.83	35.67	10.89	25.19	56.15	69.30
MSR-GCN [2]	12.16	22.65	38.64	45.24	8.39	17.05	33.03	40.43	8.02	16.27	31.32	38.15	11.98	26.76	57.08	69.74
STSGCN [12]	16.26	24.63	40.06	45.94	14.32	22.14	37.91	45.03	13.10	20.20	37.71	44.65	14.33	24.28	52.62	68.53
SPGSN [7]	10.14	19.39	34.80	41.47	7.07	14.85	30.48	37.91	6.72	13.79	27.97	34.61	10.37	23.79	53.61	67.12
EqMotion [13]	9.0	17.5	32.6	39.2	**6.3**	13.6	28.9	36.5	**5.5**	**11.3**	23.0	29.3	8.2	18.8	42.1	53.9
MAN-TF	**8.27**	**16.47**	**25.81**	**38.5**	6.82	**12.32**	**26.79**	**35.11**	12.92	15.47	**22.42**	**28.19**	**7.68**	**12.66**	**34.69**	**50.32**
MAN-TF(Lie alone)	14.76	25.53	28.41	40.92	10.33	21.26	27.99	42.22	15.63	24.63	31.51	36.25	11.84	18.85	42.55	54.69
MAN-TF(Kendall alone)	13.64	24.42	27.29	39.05	10.25	21.17	27.95	42.22	16.52	23.53	30.54	35.23	10.72	17.86	41.47	53.54
MAN-TF (Euclidien space)	14.91	25.58	30.34	41.61	10.42	21.07	28.87	43.02	16.73	24.94	31.60	37.14	11.56	19.80	43.59	55.79
Methods	Directions				Greeting				Phoning				Posing			
	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
Res-sup [11]	35.36	57.27	76.30	87.67	34.46	63.36	124.60	142.50	37.96	69.32	115.00	126.73	36.10	69.12	130.46	157.08
CSM [6]	27.07	44.72	63.94	75.37	28.63	60.69	119.25	139.92	25.66	40.13	63.06	78.01	22.02	40.34	93.65	119.32
SkelNet [3]	16.06	27.12	62.97	72.75	24.71	56.90	111.74	134.25	18.91	34.69	59.34	72.09	18.51	34.67	80.83	106.39
DMGNN [8]	13.14	24.62	64.68	81.86	23.30	50.32	107.30	132.10	12.47	25.77	48.08	58.29	15.27	29.27	71.54	96.65
Traj-GCN [10]	8.97	19.87	43.35	53.74	18.65	38.68	77.74	93.39	10.24	21.02	42.54	52.30	13.66	29.89	66.62	84.05
HisRep [9]	7.77	18.23	41.34	51.61	15.47	34.04	73.77	88.90	9.78	20.98	39.81	50.87	13.23	27.70	63.68	81.82
MSR-GCN [2]	8.61	19.65	43.28	53.82	16.48	36.95	77.32	93.38	10.10	20.74	41.51	51.26	12.79	29.38	66.95	85.01
STSGCN [12]	14.24	24.27	44.24	53.21	15.02	30.70	67.11	87.63	14.88	21.40	46.55	52.03	15.01	25.69	58.38	73.08
SPGSN [7]	7.35	17.15	39.80	50.25	14.64	32.59	70.64	86.44	8.67	18.32	38.73	48.46	10.73	25.31	59.91	76.46
EqMotion [13]	6.3	15.8	38.9	50.1	12.7	30.1	68.3	85.2	7.4	16.7	36.9	47.0	**8.2**	**18.9**	43.4	57.5
MAN-TF	**5.75**	**13.03**	**33.4**	**47.5**	**9.36**	**26.76**	**53.22**	**68.15**	**6.07**	**14.22**	**35.71**	**46.6**	14.48	25.19	**40.56**	**56.18**
MAN-TF(Lie alone)	9.20	17.75	39.08	49.43	17.31	29.84	62.55	71.89	10.71	15.98	40.31	50.08	16.35	27.05	42.55	58.23
MAN-TF(Kendall alone)	8.09	16.82	38.06	48.95	17.23	28.76	61.31	72.84	10.66	15.99	41.5	51.09	15.33	26.02	41.57	57.39
MAN-TF (Euclidien space)	12.06	18.71	39.58	50.17	18.38	30.84	64.05	75.45	11.70	16.24	41.19	51.76	15.57	27.35	43.81	59.08
Methods	Purchases				Sitting				Sitting Down				Taking Photo			
	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
Res-sup [11]	36.33	60.30	86.53	95.92	42.55	81.40	134.70	151.78	47.28	85.95	145.75	168.86	26.10	47.61	81.40	94.73
CSM [6]	25.69	47.85	82.49	93.90	22.25	34.67	58.72	75.80	23.67	51.76	102.93	119.47	20.29	38.92	61.14	77.40
SkelNet [3]	21.04	40.59	79.97	88.66	15.55	28.70	49.35	62.87	17.64	38.88	85.30	101.71	15.74	32.83	48.62	63.90
DMGNN [8]	21.35	38.71	75.67	82.74	11.92	25.11	44.59	50.20	14.95	32.88	77.06	93.00	13.61	28.95	45.99	58.76
Traj-GCN [10]	15.60	32.78	65.72	79.25	10.62	21.90	46.33	57.91	16.14	31.12	61.47	75.46	9.88	20.89	44.95	56.58
HisRep [9]	14.63	32.81	65.18	78.27	10.21	20.36	43.68	53.62	15.54	29.97	59.31	72.25	9.09	20.10	44.60	55.72
MSR-GCN [2]	14.75	32.39	66.13	79.64	10.53	21.99	46.26	57.80	16.10	31.63	62.45	76.84	9.89	21.01	44.56	56.30
STSGCN [12]	15.26	26.27	63.45	74.25	15.19	22.95	46.82	58.34	16.70	28.05	56.15	72.03	16.61	24.84	45.98	61.79
SPGSN [7]	12.75	28.58	61.01	74.38	9.28	19.40	42.25	53.56	14.18	27.72	56.75	70.74	8.79	18.90	41.49	52.66
EqMotion [13]	11.2	26.8	60.5	75.2	**8.1**	**18.0**	41.2	52.9	**13.0**	26.5	56.2	70.7	**7.9**	17.7	40.9	52.8
MAN-TF	**11.14**	**24.61**	**56.9**	**73.61**	9.08	18.56	**38.6**	**49.55**	14.34	**25.85**	**55.89**	**69.24**	8.19	**16.77**	**39.93**	**50.82**
MAN-TF(Lie alone)	12.06	25.68	57.21	76.69	10.18	19.37	41.94	52.99	15.49	26.57	58.77	72.98	9.13	17.63	40.62	52.45
MAN-TF(Kendall alone)	13.09	26.71	58.38	77.99	10.22	19.42	42.12	53.17	15.60	26.58	58.68	71.91	9.27	18.71	41.71	52.54
MAN-TF (Euclidien space)	14.12	28.55	60.42	78.59	11.21	20.28	42.78	53.92	15.71	27.65	58.91	73.09	9.14	19.68	42.37	52.78
Methods	Waiting				Walking Dog				Walking Together				Average			
	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
Res-sup [11]	30.62	57.82	106.22	121.45	64.18	102.10	141.07	164.35	26.79	50.07	80.16	92.23	34.66	61.97	101.08	115.49
CSM [6]	19.14	33.11	69.72	95.21	58.67	97.36	129.74	158.57	22.60	38.51	71.13	84.37	25.17	45.92	78.08	93.33
SkelNet [3]	16.31	29.90	63.86	84.59	54.61	93.23	124.12	155.79	19.01	32.40	63.35	73.18	20.23	38.04	69.35	84.25
DMGNN [8]	12.20	24.17	59.62	77.54	47.09	93.33	160.13	171.20	14.34	26.67	50.08	63.22	16.95	33.62	65.90	79.65
Traj-GCN [10]	11.43	23.99	50.06	61.48	23.39	46.17	83.47	95.96	10.47	21.04	38.47	45.19	12.68	26.06	52.27	63.51
HisRep [9]	10.58	23.75	49.30	60.26	21.77	43.38	78.53	90.21	9.88	19.51	35.91	42.60	11.60	24.40	49.75	60.78
MSR-GCN [2]	10.68	23.06	48.25	59.23	20.65	42.88	80.35	93.31	10.56	20.92	37.40	43.85	12.11	25.56	51.64	62.93
STSGCN [12]	16.30	27.33	48.12	59.79	16.48	37.63	70.60	86.33	11.38	22.39	39.90	47.48	15.34	25.52	50.64	60.61
SPGSN [7]	9.21	19.79	43.10	54.14	17.83	37.15	71.74	84.91	8.94	18.19	33.84	40.88	10.44	22.33	47.07	58.26
EqMotion [13]	**7.6**	**17.4**	39.9	51.1	16.6	36.4	72.5	86.2	**7.8**	**16.1**	30.6	37.1	**9.1**	20.1	43.7	55.0
MAN-TF	9.82	18.24	**35.67**	**49.39**	**15.19**	**34.77**	**69.82**	**84.75**	8.23	16.52	**29.67**	**36.14**	9.82	**19.43**	**39.94**	**52.27**
MAN-TF(Lie alone)	10.83	23.06	37.33	52.75	18.35	37.39	72.73	89.29	9.18	20.59	36.18	43.39	12.76	23.41	43.98	56.28
MAN-TF(Kendall alone)	10.72	22.89	36.30	51.65	17.39	36.26	71.49	88.24	8.93	19.48	35.07	42.36	12.51	22.97	43.56	55.88
MAN-TF (Euclidien space)	11.01	23.95	39.38	53.87	19.38	38.54	72.91	91.03	10.17	21.70	36.47	43.61	13.47	24.32	45.08	57.39

Table 2. Prediction MPJPEs of methods for long-term prediction on H3.6m and the average MPJPEs across all the actions.

Methods	Walking		Eating		Smoking		Discussion		Directions		Greeting		Phoning		Posing	
	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms
Res-sup [11]	81.73	100.68	79.87	100.20	94.83	137.44	121.30	161.70	110.05	152.48	156.32	184.29	143.92	186.79	165.41	236.79
CSM [6]	78.04	94.58	72.14	96.87	66.61	89.80	108.20	142.13	97.80	132.82	151.50	175.37	83.46	127.55	137.72	210.90
SkelNet [3]	73.58	91.84	63.58	90.88	58.96	80.53	98.28	135.68	93.77	124.89	148.38	168.06	75.42	113.34	131.90	196.21
Traj-GCN [10]	54.05	59.75	53.39	77.75	50.74	72.62	91.61	121.53	71.01	101.79	113.87	145.19	69.55	104.19	114.52	171.10
DMGNN [8]	71.36	85.82	58.11	86.66	50.85	72.15	81.90	106.32	102.06	135.75	144.51	170.54	71.33	108.37	125.45	188.18
MSR-GCN [2]	52.72	63.05	52.54	77.11	49.45	71.64	88.59	117.59	71.18	100.59	116.24	147.23	68.28	104.36	116.26	174.33
STSGCN [12]	57.64	66.74	58.46	75.08	55.55	74.13	84.20	107.74	75.61	109.89	**79.32**	**103.75**	79.19	109.88	80.82	**107.58**
SPGSN [7]	46.89	53.59	49.76	73.39	46.68	68.62	89.68	118.55	70.05	100.52	110.98	143.21	66.70	102.52	110.34	165.39
EqMotion [13]	43.4	52.8	48.4	73.0	41.0	63.4	75.3	105.6	70.4	101.3	108.7	142.0	64.7	101.0	84.9	139.4
MAN-TF	**40.4**	**50.6**	**47.3**	**72.9**	**39.8**	**62.7**	**73.1**	**104.3**	**68.4**	**99.8**	89.2	107.8	**61.8**	**99.9**	**79.6**	109.7
MAN-TF(Lie alone)	42.80	54.33	53.46	75.43	45.41	64.84	78.14	114.39	72.30	103.57	93.16	118.05	65.25	104.32	86.39	118.15
MAN-TF(Kendall alone)	41.70	53.17	52.42	76.49	44.46	63.83	77.04	112.51	71.29	102.20	93.05	117.95	64.22	104.43	87.47	119.21
MAN-TF (Euclidien space)	43.53	55.19	53.20	77.37	46.19	65.71	79.43	114.63	72.17	103.62	95.08	119.22	66.97	105.27	89.20	120.10
Methods	Purchases		Sitting		SittingDown		TakingPhoto		Waiting		WalkingDog		WalkingTog		Average	
	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms
Res-sup [11]	119.36	176.92	166.20	185.16	197.09	223.58	107.03	162.38	126.70	153.14	173.61	202.31	94.51	110.48	129.19	164.96
CSM [6]	113.44	167.61	98.04	134.70	148.87	196.75	94.75	144.52	106.03	125.60	168.71	183.42	93.90	106.60	107.94	141.95
SkelNet [3]	109.51	155.72	84.76	127.11	125.89	184.24	86.43	130.41	90.49	112.02	166.65	176.79	79.07	99.25	99.11	132.46
Traj-GCN [10]	99.24	137.28	77.63	118.36	100.91	157.32	78.73	120.06	79.08	103.83	138.24	150.63	51.67	61.10	81.07	113.01
DMGNN [8]	104.86	146.09	75.51	115.44	118.04	174.05	78.38	123.65	85.54	113.68	183.20	210.17	70.46	86.93	93.57	127.62
MSR-GCN [2]	101.63	139.15	78.19	120.02	102.83	155.45	77.94	121.87	76.33	106.25	111.87	148.21	52.93	65.91	81.13	114.18
STSGCN [12]	**87.10**	**119.26**	82.32	119.83	**92.60**	**129.67**	87.70	119.79	86.41	118.04	86.79	118.33	75.33	95.83	80.66	113.33
SPGSN [7]	96.53	133.88	75.00	116.24	98.94	149.88	75.58	118.22	73.50	103.62	102.37	137.96	49.84	60.86	77.40	109.64
EqMotion [13]	93.5	134.5	74.7	116.6	98.1	149.9	76.7	122.0	71.4	104.6	104.8	141.2	44.5	56.0	73.4	106.9
MAN-TF	89.6	122.4	**72.2**	**114.8**	95.8	142.9	**74.8**	**116.5**	**70.4**	**102.28**	**85.8**	**117.3**	**43.2**	**55.8**	**68.76**	**98.65**
MAN-TF(Lie alone)	96.01	127.78	76.52	116.74	104.75	148.36	81.91	120.03	79.38	109.33	89.86	121.59	51.21	63.06	74.44	104.01
MAN-TF(Kendall alone)	97.21	128.87	77.65	117.86	103.70	147.26	82.06	121.08	78.33	105.14	88.75	119.64	50.23	62.75	73.97	103.49
MAN-TF (Euclidien space)	98.09	129.64	77.83	117.95	104.72	149.33	83.42	121.47	80.63	110.58	91.15	124.37	51.94	63.52	75.57	105.19

References

[1] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5223–5232, 2020. 2

[2] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11467–11476, 2021. 3

[3] Xiao Guo and Jongmoo Choi. Human motion prediction via learning local structure representations and temporal dependencies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2580–2587, 2019. 3

[4] Sena Kiciroglu, Wei Wang, Mathieu Salzmann, and Pascal Fua. Long term motion prediction using keyposes. In *2022 International Conference on 3D Vision (3DV)*, pages 12–21. IEEE, 2022. 2

[5] Tim Lebailly, Sena Kiciroglu, Mathieu Salzmann, Pascal Fua, and Wei Wang. Motion prediction using temporal inception module. In *Proceedings of the Asian conference on computer vision*, 2020. 2

[6] 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. 3

[7] Maosen Li, Siheng Chen, Zijing Zhang, Lingxi Xie, Qi Tian, and Ya Zhang. Skeleton-parted graph scattering networks for 3d human motion prediction. In *European conference on computer vision*, pages 18–36. Springer, 2022. 3

[8] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *Proceedings of the IEEE/CVF conference on CVPR*, pages 214–223, 2020. 3

[9] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 474–489. Springer, 2020. 2, 3

[10] 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. 3

[11] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on CVPR*, pages 2891–2900, 2017. 3

[12] Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. Space-time-separable graph convolutional network for pose forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11209–11218, 2021. 3

[13] Chenxin Xu, Robby T Tan, Yuhong Tan, Siheng Chen, Yu Guang Wang, Xinchao Wang, and Yanfeng Wang. Eqmotion: Equivariant multi-agent motion prediction with invariant interaction reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1410–1420, 2023. 3

[14] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 346–364. Springer, 2020. 2