Auxiliary Tasks Benefit 3D Skeleton-based Human Motion Prediction
– Supplementary Material –

A. Baseline Description

Here we introduce the baselines we compare in the paper.

- Res-sup. [8]: The model uses deep recurrent neural networks (RNNs) to model human motion, with the goal of learning time-dependent representations.
- CSM [2]: The model uses a convolutional long-term encoder is used to encode the whole given motion sequence into a long-term hidden variable, and uses a decoder to predict the remainder of the sequence.
- Traj-GCN [7]: The model encodes temporal information by working in trajectory space and designs a graph convolutional network to learn graph connectivity automatically.
- DMGNN [4]: The model uses a multiscale graph to extract features at individual scales and fuse features across scales for a more comprehensive motion feature learning.
- MSRGCN [1]: The model uses a series of GCNs that are used to extract features from fine to coarse scale and then from coarse to fine scale, enforcing the network to learn more representative features.
- HisRep [6]: The model extracts motion attention to capture the similarity between the current motion context and the historical motion sub-sequences.
- STSGCN [9]: The method models the human pose dynamics only with a graph convolutional network, consisting of temporal evolutions and spatial joint interactions.
- PGBIG [5] : The model designs a multi-stage prediction framework where each stage predicts initial guess for the next stage. Each stage's model consists of spatial dense graph convolutional Networks (S-DGCN) and temporal dense graph convolutional networks.
- SPGSN [3]: The model proposes adaptive graph scattering leveraging multiple trainable band- pass graph filters to decompose pose features into richer graph spectrum bands and models body-parts separately.

B. More Implementation Details

Here we further introduce more implementation details that are specific to different tasks. For Human3.6M short-term tasks, we set the number of attention layers L as 3, the dimension of each head as 32, and weight decay as 1e-12. For Human3.6M long-term tasks, we set the number of attention layers L as 4, the dimension of each head as 32, and weight decay as 1e-4. For CMU Mocap short-term tasks, we set the number of attention layers L as 4, the dimension of each head as 32, and weight decay as 1e-12. We reduce the learning rate using a ratio 0.9 for every 5 epochs. For CMU Mocap long-term tasks, we set the number of attention layers L as 5, the dimension of each head as 16, and weight decay as 5e-6. We additionally concatenate the velocity and acceleration with the position. We reduce the learning rate using a ratio 0.9 for every 5 epochs. For 3DPW short-term tasks, we set the number of attention layers L as 3, the dimension of each head as 32, and weight decay as 5e-8. We reduce the learning rate using a ratio 0.9 for every 5 epochs. For 3DPW long-term tasks, we set the number of attention layers L as 3, the dimension of each head as 16, and weight decay as 5e-4. We additionally concatenate the velocity with the position. We reduce the learning rate using a ratio 0.9 for every 5 epochs.

Table 1: Effect of different numbers of attention layers L on H3.6M.

L	80ms	160ms	320ms	400ms	AVG
1	9.6	21.2	45.6	57.0	33.4
2	9.5	20.9	44.4	55.2	32.5
3	9.5	**20.6**	**43.4**	**54.1**	**31.9**
4	**9.4**	20.6	43.7	54.5	32.1

C. Additional Experimental Results

C.1. Effect of Number of Attention Layers

Table 1 shows the effect of different numbers of two iterative attention layers L on the H3.6M dataset. We find that i) initially increasing L leads to better performance since the model network learning capability increases; and ii) when the number of layers is sufficient, the performance tends to be stable.

C.2. Comparison of Model Size and Performance

To verify the applicability of AuxFormer, we compare AuxFormer to existing methods in terms of the parameter

Table 2: Comparison of model size and performance in short-term prediction on H3.6M dataset.

	DMGNN	Traj-GCN	MSR-GCN	HisRep	SPGSN	AuxFormer (Ours)
ParaSize (M)	4.82	2.56	6.30	3.18	5.66	1.00
AVG MPJPE	49.0	38.6	38.1	36.4	34.5	**31.9**

Table 3: Comparisons of long-term prediction on Human3.6M. Results at 560ms and 1000ms in the future are shown. **Bold**/underline font represent the best/second best result.

Motion	Walking		Eating		Smoking		Discussion		Directions		Greeting		Phoning		Posing	
millisecond	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms
Res-Sup.	81.7	100.7	79.9	100.2	94.8	137.4	121.3	161.7	110.1	152.5	156.3	184.3	143.9	186.8	165.7	236.8
Traj-GCN	54.1	59.8	53.4	77.8	50.7	72.6	91.6	121.5	71.0	101.8	115.4	148.8	69.2	103.1	114.5	173.0
DMGNN	71.4	85.8	58.1	86.7	50.9	72.2	81.9	138.3	102.1	135.8	144.5	170.5	71.3	108.4	125.5	188.2
MSRGCN	52.7	63.0	52.5	77.1	49.5	71.6	88.6	117.6	71.2	100.6	116.3	147.2	68.3	104.4	116.3	174.3
PGBIG	48.1	56.4	51.1	76.0	46.5	69.5	87.1	118.2	**69.3**	**100.4**	**110.2**	143.5	**65.9**	102.7	106.1	164.8
SPGSN	46.9	53.6	**49.8**	**73.4**	46.7	68.6	89.7	118.6	70.1	100.5	111.0	143.2	66.7	**102.5**	110.3	165.4
Ours	**43.8**	**52.0**	50.3	74.7	**42.0**	**63.0**	**77.6**	**102.3**	71.6	103.0	110.5	**141.6**	66.6	**102.5**	**91.6**	**137.1**

Motion	Purchases		Sitting		Sitting Down		Taking Photo		Waiting		Walking Dog		Walking Together		Average	
millisecond	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms
Res-Sup.	119.4	176.9	166.2	185.2	197.1	223.6	107.0	162.4	126.7	153.2	173.6	202.3	94.5	110.5	129.2	165.0
Traj-GCN	102.0	143.5	78.3	119.7	100.0	150.2	77.4	119.8	79.4	108.1	111.9	148.9	55.0	65.6	81.6	114.3
DMGNN	104.9	146.1	75.5	115.4	118.0	174.1	78.4	123.7	85.5	113.7	183.2	210.2	70.5	86.9	93.6	127.6
MSRGCN	101.6	139.2	78.2	120.0	102.8	155.5	77.9	121.9	76.3	106.3	111.9	148.2	52.9	65.9	81.1	114.2
PGBIG	**95.3**	**133.3**	**74.4**	**116.1**	**96.7**	**147.8**	**74.3**	118.6	**72.2**	**103.4**	104.7	139.8	51.9	64.3	76.9	110.3
SPGSN	96.5	133.9	75.0	116.2	98.9	149.9	75.6	**118.2**	73.5	103.6	**102.4**	138.0	49.8	60.9	77.4	109.6
Ours	96.9	134.8	76.1	119.7	98.5	151.1	78.9	123.9	74.6	106.4	103.3	**133.3**	**47.3**	**58.8**	**75.3**	**107.0**

Table 4: Prediction MPJPEs of methods on CMU Mocap for both short-term and long-term prediction, as well as the average prediction results across all the actions. **Bold**/underline font represent the best/second best result.

Motion	Basketball					Basketball Signal					Directing Traffic					Jumping				
millisecond	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
Res-sup. [8]	15.45	26.88	43.51	49.23	88.73	20.17	32.98	42.75	44.65	60.57	20.52	40.58	75.38	90.36	153.12	26.85	48.07	93.50	108.90	162.84
DMGNN [?]	15.57	28.72	59.01	73.05	138.62	5.03	9.28	20.21	26.23	52.04	10.21	20.90	41.55	52.28	111.23	17.42	26.82	38.27	40.08	46.40
Traj-GCN [7]	11.68	21.26	40.99	50.78	97.99	3.33	6.25	13.58	17.98	54.00	6.92	13.69	30.30	39.97	114.16	14.53	24.20	37.44	41.10	51.73
MSR-GCN [1]	10.28	18.94	37.68	47.03	86.96	3.03	5.68	12.35	16.26	47.91	5.92	12.09	28.36	38.04	111.04	12.84	20.42	30.58	34.42	48.03
STSGCN [9]	12.56	23.04	41.92	50.33	94.17	4.72	6.69	14.53	17.88	49.52	6.41	12.38	29.05	38.86	109.42	16.70	27.58	36.15	36.42	55.34
PGBIG [5]	9.53	17.53	**35.32**	**44.23**	84.14	2.71	4.88	10.77	14.63	50.19	**4.83**	**9.77**	**23.62**	**32.23**	**102.32**	12.69	23.18	38.31	42.24	51.71
SPGSN [3]	10.24	18.54	38.22	48.68	89.58	2.91	5.25	11.31	15.01	**47.31**	5.52	11.16	25.48	37.06	108.14	10.75	16.67	26.07	30.08	52.92
Ours	**9.35**	**17.06**	35.46	45.50	**80.77**	**2.62**	**4.79**	**10.57**	**14.20**	48.19	6.27	12.37	28.93	38.43	111.97	**9.98**	**15.78**	**25.31**	**28.81**	**41.64**
Motion	Running					Soccer					Walking					Washing Window				
millisecond	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
Res-sup. [8]	25.76	48.91	88.19	100.80	158.19	17.75	31.30	52.55	61.40	107.37	44.35	76.66	126.83	151.43	194.33	22.84	44.71	86.78	104.68	202.73
DMGNN [?]	17.42	26.82	38.27	40.08	46.40	14.86	25.29	52.21	65.42	111.90	9.57	15.53	26.03	30.37	67.01	7.93	14.68	33.34	44.24	82.84
Traj-GCN [7]	14.53	24.20	37.44	41.10	51.73	13.33	24.00	43.77	53.20	108.26	6.62	10.74	17.40	20.35	34.41	5.96	11.62	24.77	31.63	66.85
MSR-GCN [1]	12.84	20.42	30.58	34.42	48.03	10.92	19.50	37.05	46.38	99.32	6.31	10.30	17.64	21.12	39.70	5.49	11.07	25.05	32.51	71.30
STSGCN [9]	16.70	27.58	36.15	36.42	55.34	13.49	25.24	39.87	51.58	109.63	7.18	10.99	17.84	22.61	44.12	6.79	12.10	24.92	36.66	69.48
PGBIG [5]	12.69	23.18	38.31	42.24	51.71	11.09	20.62	39.48	48.72	99.98	6.23	10.34	16.84	19.76	**33.92**	**4.63**	**9.16**	**20.87**	**27.34**	65.69
SPGSN [3]	10.75	16.67	26.07	30.08	52.92	10.86	18.99	**35.05**	**45.16**	99.51	6.32	10.21	16.34	20.19	34.83	4.86	9.44	21.50	28.37	**65.08**
Ours	**9.98**	**15.78**	**25.31**	**28.81**	**41.64**	**10.01**	**18.21**	36.31	45.79	**95.98**	**5.76**	**9.16**	**15.69**	**18.80**	34.81	4.69	9.39	21.87	28.83	72.90
Motion	Average																			
millisecond	80	160	320	400	1000															
Res-sup. [8]	24.21	43.75	76.19	88.93	139.00															
DMGNN [?]	14.07	24.44	45.90	55.45	104.33															
Traj-GCN [7]	9.94	18.02	33.55	40.95	81.85															
MSR-GCN [1]	8.72	15.83	30.57	38.10	79.01															
STSGCN [9]	10.80	18.19	31.18	41.05	81.76															
PGBIG [5]	8.20	15.41	30.13	37.27	76.69															
SPGSN [3]	8.30	14.80	28.64	36.96	77.82															
Ours	**7.54**	**13.78**	**27.95**	**35.39**	**76.32**															

numbers and prediction results in short-term prediction on H3.6M. The results is shown in Table 2. We can see that our AuxFormer has the lowest MPJPE and the smallest model size, showing better applicability.

D. Limitation and Future Work

This work considers the masked prediction and denoising as auxiliary tasks for more comprehensive spatial-temporal dependency modeling. A possible future work is exploring more auxiliary tasks and more corruption strategies. Also, this work considers deterministic human motion prediction. We will explore the effect of auxiliary tasks in stochastic human motion prediction where the model is required to predict diverse future motions.

References

[1] 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. 1, 2

[2] 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. 1

[3] Maosen Li, Siheng Chen, Zijing Zhang, Lingxi Xie, Qi Tian, and Ya Zhang. Skeleton-parted graph scattering networks for 3d human motion prediction. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, pages 18–36. Springer, 2022. 1, 2

[4] 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 Computer Vision and Pattern Recognition*, pages 214–223, 2020. 1

[5] Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6437–6446, 2022. 1, 2

[6] 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. 1

[7] 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

[8] 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. 1, 2

[9] 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. 1, 2