Learning Progressive Joint Propagation for Human Motion Prediction Supplementary

In this supplementary document, we provide materials not included in the main paper due to space constraints. Firstly, we provide more details about the DCT/IDCT process in Section 1. Secondly, we show the detailed architecture of our proposed network in Section 2. Lastly, detailed quantitative results are elaborated in Section 3.

1 Details of DCT/IDCT process

Following [3], we employ DCT to encode the temporal trajectories into the frequency domain. A key benefit is that, by discarding the high-frequencies, the DCT can provide a more compact representation to captures the smoothness of human motion. Specifically, given the trajectory $\widetilde{\mathbf{x}}_j = (\mathbf{x}_{j,1}, \mathbf{x}_{j,2}, ..\mathbf{x}_{j,t})$, where j denotes the index of joint, the corresponding l^{th} DCT coefficient can be computed as:

$$C_{j,l} = \sqrt{\frac{2}{T}} \sum_{t=1}^{T} x_{j,t} \sqrt{\frac{1}{1+\delta_{l,1}}} cos(\frac{\pi}{2T})(2t-1)(l-1) \tag{1}$$

where $\sigma_{m,n}$ denotes the *Kronecker* delta function :

$$\delta_{m,n} = \begin{cases} 1 & \text{if } m = n \\ 0 & \text{if } m \neq n \end{cases} \tag{2}$$

Similarly, the final output can be transformed to temporal domain via the Inverse Discrete Cosine Transform (IDCT).

$$x_{j,t} = \sqrt{\frac{2}{T}} \sum_{l=1}^{T} C_{j,l} \sqrt{\frac{1}{1+\delta_{l,1}}} cos(\frac{\pi}{2T})(2t-1)(l-1) \tag{3}$$

2 Network Architecture

Figure 1 illustrates the detailed architectures of our proposed network, including the transformer-encoder (left), dictionary (left), and the progressive decoder (right). Note that all operations (*i.e.* Linear, Feed Forward, Attention) are deployed for each joint with shared parameters. For more details of the progressive decoder and dictionary module, please kindly refer to Section 3.4 and 3.5 in our main paper.

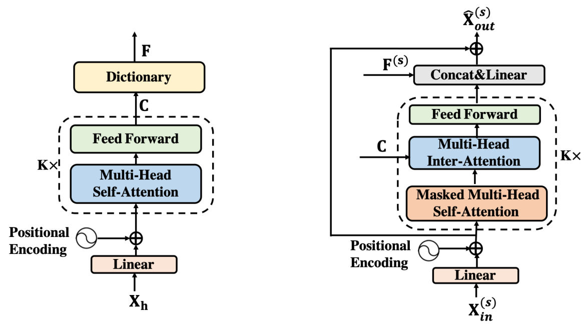

Fig. 1. Detailed architecture of our proposed network. Left: **Encoder and Dictionary.** $\mathbf{X}_h$ denotes the encoded historical trajectories of each joint. $\mathbf{C}$ is the generated context feature and $\mathbf{F}$ is the future dynamics summarized from the learned dictionary. Right: **Progressive decoder at stage** s**.** $\mathbf{X}_{in}^{(s)}$, $\hat{\mathbf{X}}_{out}^{(s)}$ are the input and corresponding output of the progressive decoder at stage s. $\mathbf{F}^{(s)}$ is the future dynamics used in stage s. More details of the progressive decoding process can be found in Section 3.4 in our main paper.

3 More Quantitative Results

3.1 Comparison with the State-of-the-art Methods

As mentioned in Section 4.3, we provide the full table of our proposed method compared with the state-of-the-art approaches [4, 3, 2, 5, 1] on Human3.6M dataset. As can be seen in Table 1 and Table 2, our results outperforms other methods on both MAE and MPJPE protocols.

Table 1. Short-term prediction results in Mean Angle Error (MAE) on Human3.6M for all actions. The best result is marked in bold.

	Walking				Eating				Smoking				Discussion				Directions				Greeting			
milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
zero-velocity [4]	0.39	0.68	0.99	1.15	0.27	0.48	0.73	0.86	0.26	0.48	0.97	0.95	0.31	0.67	0.94	1.04	0.39	0.59	0.79	0.89	0.54	0.89	1.30	1.49
Residual sup. [4]	0.28	0.49	0.72	0.81	0.23	0.39	0.62	0.76	0.33	0.61	1.05	1.15	0.31	0.68	1.01	1.09	0.26	0.47	0.72	0.84	0.75	1.17	1.74	1.83
convSeq2Seq [2]	0.33	0.54	0.68	0.73	0.22	0.36	0.58	0.71	0.26	0.49	0.96	0.92	0.32	0.67	0.94	1.01	0.39	0.60	0.80	0.91	0.51	0.82	1.21	1.38
AGED w/o adv [1]	0.28	0.42	0.66	0.73	0.22	0.35	0.61	0.74	0.30	0.55	0.98	0.99	0.30	0.63	0.97	1.06	0.26	0.46	0.71	0.81	0.61	0.95	1.44	1.61
AGED w/ adv [1]	0.22	0.36	0.55	0.67	0.17	**0.28**	0.51	0.64	0.27	0.43	**0.82**	0.84	0.27	0.56	**0.76**	**0.83**	0.23	0.39	0.63	0.69	0.56	0.81	1.30	1.46
Imitation [5]	0.21	0.34	0.53	0.59	0.17	0.30	0.52	0.65	0.23	0.44	0.87	0.85	0.23	0.56	0.82	0.91	0.27	0.46	0.81	0.89	0.43	0.75	1.17	1.33
GNN [3]	0.18	0.31	**0.49**	0.56	0.16	0.29	0.50	0.62	0.22	0.41	0.86	0.80	0.20	**0.51**	0.77	0.85	0.26	0.45	0.71	0.79	0.36	0.60	0.95	1.13
ours	**0.17**	**0.30**	0.51	**0.55**	**0.16**	0.29	**0.50**	**0.61**	**0.21**	**0.40**	0.85	**0.78**	**0.19**	0.54	0.89	0.94	**0.22**	**0.39**	**0.62**	**0.69**	**0.34**	**0.58**	**0.94**	**1.12**

	Phoning				Posing				Purchase				Sitting				Sitting Down				Taking Photo			
milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
zero-velocity [4]	0.64	1.21	1.65	1.83	0.28	0.57	1.13	1.37	0.62	0.88	1.19	1.27	0.40	1.63	1.02	1.18	0.39	0.74	1.07	1.19	0.25	0.51	0.79	0.92
Residual sup. [4]	0.23	0.43	0.69	0.82	0.36	0.71	1.22	1.48	0.51	0.97	1.07	1.16	0.41	1.05	1.49	1.63	0.39	0.81	1.40	1.62	0.24	0.51	0.90	1.05
convSeq2Seq [2]	0.59	1.13	1.51	1.65	0.29	0.60	1.12	1.37	0.63	0.91	1.19	1.29	0.39	0.61	1.02	1.18	0.41	0.78	1.16	1.31	0.23	0.49	0.88	1.06
AGED w/o adv [1]	0.23	0.42	0.61	0.79	0.34	0.70	1.19	1.40	0.46	0.89	1.06	1.11	0.46	0.87	1.23	1.51	0.38	0.77	1.18	1.41	0.24	0.52	0.92	1.01
AGED w/ adv [1]	**0.19**	**0.34**	**0.50**	**0.68**	0.31	0.58	1.21	1.34	0.46	0.78	**1.01**	**1.07**	0.41	0.76	1.05	1.19	0.33	0.62	0.98	1.10	0.23	0.48	0.81	0.95
Imitation [5]	0.54	1.05	1.40	1.56	0.27	0.55	1.16	1.41	0.54	0.78	1.07	1.14	0.29	0.48	0.87	1.04	0.34	0.68	1.01	1.14	0.17	0.37	0.60	0.72
GNN [3]	0.53	1.02	1.35	1.48	0.19	0.44	**1.01**	**1.24**	0.43	0.65	1.05	1.13	0.29	0.45	0.80	0.97	0.30	0.61	0.90	1.00	0.14	0.34	**0.58**	**0.70**
ours	0.46	0.90	1.2	1.37	**0.19**	**0.44**	1.07	1.3	**0.38**	**0.64**	1.13	1.21	**0.27**	**0.44**	**0.78**	**0.96**	**0.27**	**0.54**	**0.88**	**0.97**	**0.13**	**0.33**	0.60	0.74

	Waiting				Walking Dog				Walking Together				Average			
milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
zero-velocity [4]	0.34	0.67	1.22	1.47	0.60	0.98	1.36	1.50	0.33	0.66	0.94	0.99	0.40	0.78	1.07	1.21
Residual sup. [4]	0.28	0.53	1.02	1.14	0.56	0.91	1.26	1.40	0.31	0.58	0.87	0.91	0.36	0.67	1.02	1.15
convSeq2Seq [2]	0.30	0.62	1.09	1.30	0.59	1.00	1.32	1.44	0.27	0.52	0.71	0.74	0.38	0.68	1.01	1.13
AGED w/o adv [1]	0.31	0.64	1.08	1.12	0.51	0.87	1.21	1.33	0.29	0.51	0.72	0.75	0.32	0.62	0.96	1.07
AGED w/ adv [1]	0.24	0.50	1.02	1.13	0.50	0.81	1.15	1.27	0.23	0.41	0.56	0.62	0.31	0.54	0.85	0.97
Imitation [5]	0.25	0.53	0.96	1.19	0.46	0.80	1.12	1.31	0.19	0.41	0.61	0.64	0.31	0.57	0.90	1.02
GNN [3]	0.23	0.50	0.91	1.14	0.46	0.79	1.12	1.29	0.15	0.34	0.52	0.57	0.27	0.51	0.83	0.95
ours	**0.21**	**0.48**	**0.86**	**1.08**	**0.40**	**0.75**	**1.05**	**1.23**	**0.14**	**0.32**	**0.52**	**0.55**	**0.25**	**0.49**	**0.83**	**0.94**

Table 2. Short-term prediction results in Mean Per Joint Position Error (MPJPE) on Human3.6M for all actions. The best result is marked in bold. A $3D$ suffix to a method indicates that the method was directly trained on 3D joint positions. Otherwise, the results were obtained by converting the joint angle to 3D positions. The best result is marked in bold.

	Walking				Eating				Smoking				Discussion				Directions				Greeting			
milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Residual sup. [4]	21.7	38.1	58.9	68.8	15.1	28.6	54.8	67.4	20.8	39.0	66.1	76.1	26.2	51.2	85.8	94.6	27.9	44.8	63.5	78.2	29.3	56.0	110.2	125.6
Residual sup. 3D [4]	23.8	40.4	62.9	70.9	17.6	34.7	71.9	87.7	19.7	36.6	61.8	73.9	31.7	61.3	96.0	103.5	36.5	56.4	81.5	97.3	37.9	74.1	1390	158.8
convSeq2Seq [2]	21.8	37.5	55.9	63.0	13.3	24.5	48.6	60.0	15.4	25.5	39.3	44.5	23.6	43.6	68.4	74.9	26.7	43.3	59.0	72.4	30.4	58.6	110.0	122.8
convSeq2Seq 3D [2]	17.1	31.2	53.8	61.5	13.7	25.9	52.5	63.3	11.1	21.0	33.4	38.3	18.9	39.3	67.7	75.7	22.0	37.2	59.6	73.4	24.5	46.2	90.0	103.1
GNN [3]	11.1	19.0	32.0	39.1	9.2	19.5	40.3	48.9	9.2	16.6	26.1	29.0	11.3	23.7	41.9	46.6	11.2	23.2	52.7	64.1	14.2	**27.7**	67.1	82.9
GNN 3D [3]	8.9	15.7	29.2	**33.4**	8.8	18.9	39.4	47.2	7.8	14.9	25.3	28.7	9.8	22.1	**39.6**	**44.1**	12.6	24.4	48.2	58.4	14.5	30.5	74.2	89.0
Ours	9.6	18.0	33.1	39.1	9.1	19.5	40.2	48.8	7.2	14.2	24.7	29.7	9.0	22.9	43.4	51.6	**9.3**	**22.0**	51.6	63.2	15.4	30.7	71.8	82.8
Ours 3D	**7.9**	**14.5**	**29.1**	34.5	**8.4**	**18.1**	**37.4**	**45.3**	**6.8**	**13.2**	**24.1**	**27.5**	**8.3**	**21.7**	43.9	48.0	11.1	22.7	**48.0**	**58.4**	**13.2**	28.0	**64.5**	**77.9**

	Phoning				Posing				Purchase				Sitting				Sitting Down				Taking Photo			
milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Residual sup. [4]	28.7	50.9	88.0	99.7	30.5	59.4	118.7	144.7	33.3	58.2	85.4	93.7	32.6	65.2	113.7	126.2	33.0	64.1	121.7	146	21.2	40.3	72.2	86.2
Residual sup. 3D [4]	25.6	44.4	74.0	84.2	27.9	54.7	131.3	160.8	40.8	71.8	104.2	109.8	34.5	69.9	126.3	141.6	28.6	55.3	101.6	118.9	23.6	47.4	94.0	112.7
convSeq2Seq [2]	22.4	38.4	65.0	75.4	22.4	42.1	87.3	106.1	28.4	53.8	82.1	93.1	24.7	50.0	88.6	100.4	23.9	39.9	74.6	89.8	18.4	32.1	60.3	72.5
convSeq2Seq 3D [2]	17.2	29.7	53.4	61.3	16.1	35.6	86.2	105.6	29.4	54.9	82.2	93.0	19.8	42.4	77.0	88.4	17.1	34.9	66.3	77.7	14.0	27.2	53.8	66.2
GNN [3]	13.5	22.5	45.2	52.4	11.1	27.1	69.4	86.2	20.4	42.8	69.1	78.3	11.7	27.0	55.9	66.9	11.5	25.4	53.9	**65.6**	8.3	15.8	38.5	49.1
GNN 3D [3]	11.5	20.2	37.9	**43.2**	9.4	23.9	66.2	82.9	19.6	38.5	64.4	72.2	10.7	24.6	50.6	62.0	11.4	27.6	56.4	67.6	6.8	15.2	**38.2**	49.6
Ours	12.8	24.1	43.6	51.6	8.4	25.3	68.9	88.7	20.2	43.5	66.2	74.9	10.6	24.8	50.4	**59.1**	11.5	**25.3**	**53.9**	69.2	7.1	14.8	39.2	**48.9**
Ours 3D	**10.8**	**19.6**	**37.6**	46.8	**8.3**	**22.8**	**65.6**	**81.8**	**18.5**	**38.1**	**61.8**	**69.6**	**9.5**	**23.9**	**49.8**	61.8	**11.2**	29.9	59.8	68.4	**6.3**	**14.5**	38.8	49.4

	Waiting				Walking Dog				Walking Together				Average			
milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Residual sup. [4]	24.9	50.0	96.5	114.0	53.8	90.9	134.6	156.9	19.7	38.2	62.9	72.3	27.9	51.6	88.9	103.4
Residual sup. 3D [4]	29.5	60.5	119.9	140.6	60.5	101.9	160.8	188.3	23.5	45.0	71.3	82.8	30.8	57.0	99.8	115.5
convSeq2Seq [2]	24.9	50.2	101.6	120.0	56.4	94.9	136.1	156.3	21.1	38.5	61.0	70.4	24.9	44.9	75.9	88.1
convSeq2Seq 3D [2]	17.9	36.5	72.9	90.7	40.6	74.7	116.6	138.7	15.0	29.9	54.3	65.8	19.6	37.8	68.1	80.2
GNN [3]	12.1	27.5	67.3	85.6	35.8	63.6	106.7	126.8	11.7	23.5	46.0	53.5	13.5	27.0	54.2	65.0
GNN 3D [3]	9.5	22.0	57.5	73.9	32.2	58.0	102.2	122.7	8.9	18.4	35.3	44.3	12.1	25.0	51.0	61.3
Ours	9.2	23.6	62.0	79.6	29.9	60.3	104.3	127.2	10.2	22.0	44.5	52.9	11.9	26.1	53.2	64.5
Ours 3D	**8.4**	**21.5**	**53.9**	**69.8**	**22.9**	**50.4**	**100.8**	**119.8**	**8.7**	**18.3**	**34.2**	**44.1**	**10.7**	**23.8**	**50.0**	**60.2**

References

1. Gui, L.Y., Wang, Y.X., Liang, X., Moura, J.M.: Adversarial geometry-aware human motion prediction. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 786–803 (2018)
2. Li, C., Zhang, Z., Sun Lee, W., Hee Lee, G.: Convolutional sequence to sequence model for human dynamics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5226–5234 (2018)
3. Mao, W., Liu, M., Salzmann, M., Li, H.: Learning trajectory dependencies for human motion prediction. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9489–9497 (2019)
4. Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2891–2900 (2017)
5. Wang, B., Adeli, E., Chiu, H.k., Huang, D.A., Niebles, J.C.: Imitation learning for human pose prediction. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7124–7133 (2019)