Decompose More and Aggregate Better: Two Closer Looks at Frequency Representation Learning for Human Motion Prediction – Supplementary Material –

Xuehao Gao[1], Shaoyi Du[1], Yang Wu[2], Yang Yang[1,*]

[1]Xi'an Jiaotong University, [2]Tencent AI Lab

{gaoxuehao.xjtu, dushaoyi}@gmail.com, dylanywu@tencent.com

Abstract

In this supplementary material, we analyze more experiment results of the proposed human motion prediction system. First, we analyze its computational efficiency in section 1. We then supplement its more detailed prediction performances in section 2. In section 3, we further investigate the effect of its multiple frequency representations with t-SNE visualizations. In section 4, we evaluate its few-sample prediction performances on different human activities.

1. Analysis for Computational Efficiency

To verify the applicability of our proposed human motion prediction system, we compare it with other methods in terms of their running time, FLOPs and MPJEP performances in long-term prediction (1000ms) on H3.6M dataset. As shown in Table 1, our method outperforms the state-of-the-art MPJPE performance by a large margin with a comparable computational cost. It indicates that we develop a powerful baseline model for the real-time and precise human motion prediction, enjoying a wide prospect in future real-world applications.

Table 1. Efficiency Comparisons between different human motion prediction methods on their running time costs and MPJPEs.

Model	Time Cost (ms)	FLOPs (M)	MPJPE (mm)
DMGNN [3]	91.1	8.8	137.2
MSR-GCN [1]	83.8	7.9	114.2
PGBIG [4]	49.3	4.7	110.3
SPGSN [2]	61.9	5.8	109.6
Ours	72.6	6.6	100.3

2. Analysis for More Prediction Comparisons

In the paper (Table 3), we report MPJPE results averaged over all kinds of human motions at each future time step on CMU Mocap dataset. In this appendix, we provide more detailed prediction performance comparisons to verify the effectiveness of our method. First, in Table 2, we supplement detailed MPJPE performances of each kind of human motion on CMU Mocap dataset. Then, in Table 3, we use mean angle error (MAE) to evaluate the prediction performance on the angle representation of H3.6M dataset. As verified in these comparisons, our proposed human motion prediction system outperforms baseline methods at all actions and all evaluation metrics, achieving state-of-the-art short-term and long-term prediction performances. These performance improvements on these common-used large-scale datasets and evaluation metrics verify the effectiveness of the proposed method, indicating that it develops a strong baseline model for robust human motion prediction.

3. Analysis for Multiple Frequency Features

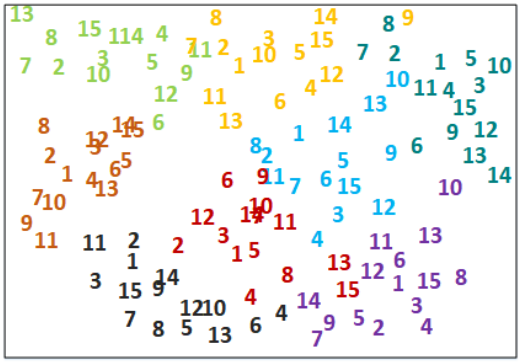
In this section, we further investigate the effectiveness of the proposed decomposition-aggregation scheme in the frequency representation learning task with qualitative analysis. As shown in Figure 1, we first select a motion sample from each motion category in the CMU Mocap dataset and then plot the t-SNE visualization of their multiple frequency representations augmented in the decomposition stage. As presented in the paper (Section 4), FDU component develops 15 filters to embed an input human motion into 15 feature spaces, and the No.1 filter focuses on extracting its initial frequency representation. Figure 1 verifies that FDU significantly enriches the spectral encoding of an input body motion. Besides, multiple frequency representations of a body motion encapsulate shared semantic similarities to reflect their consistent motion categories. By factorizing

Table 2. Supplement to comparisons of short-term and long-term prediction on CMU Mocap dataset.

scenarios	basketball					basketball signal					directing traffic					jumping				
millisecond	80ms	160ms	320ms	400ms	1000	80ms	160ms	320ms	400ms	1000	80ms	160ms	320ms	400ms	1000	80ms	160ms	320ms	400ms	1000
DMGNN [3]	15.6	28.7	59.0	73.1	138.6	5.0	9.3	20.2	26.2	52.0	10.2	20.9	41.6	52.3	111.2	32.0	54.3	96.7	119.9	224.6
MSR-GCN [1]	10.3	18.9	37.7	47.0	87.0	3.0	5.7	12.4	16.3	47.9	5.9	12.1	28.4	38.0	111.0	15.0	28.7	55.9	69.1	124.8
PGBIG [4]	9.5	17.5	35.3	44.2	84.1	2.7	4.9	10.8	14.6	50.2	4.8	9.8	23.6	32.3	102.3	13.9	27.8	55.8	69.0	125.6
SPGSN [2]	10.2	18.5	38.2	48.7	89.6	2.9	5.3	11.3	15.0	47.3	5.5	11.2	25.5	37.1	108.1	14.9	28.2	56.7	71.2	125.2
Ours	**9.1**	**16.1**	**34.2**	**44.0**	**83.1**	**2.5**	**4.7**	**10.1**	**14.0**	**46.3**	**4.6**	**9.1**	**22.4**	**31.1**	**100.9**	**13.0**	**26.3**	**52.1**	**67.3**	**122.9**
scenarios	soccer					walking					wash window					average				
millisecond	80ms	160ms	320ms	400ms	1000ms	80ms	160ms	320ms	400ms	1000ms	80ms	160ms	320ms	400ms	1000ms	80ms	160ms	320ms	400ms	1000ms
DMGNN [3]	14.9	25.3	52.2	65.4	111.9	9.6	15.5	26.0	30.4	67.0	7.9	14.7	33.3	44.2	82.8	13.6	24.1	47.0	58.8	112.6
MSR-GCN [1]	10.9	19.5	37.1	46.4	99.3	6.3	10.3	17.6	21.1	39.7	5.5	11.1	25.1	32.5	71.3	8.1	15.2	30.6	38.6	83.0
PGBIG [4]	11.1	20.6	39.5	48.7	99.9	**6.2**	10.3	16.8	19.8	33.9	4.6	**9.2**	20.9	27.3	65.7	7.6	14.3	29.0	36.6	80.1
SPGSN [2]	10.9	**19.0**	35.1	45.2	99.5	6.3	10.2	16.3	20.2	34.8	4.9	9.4	21.5	28.4	65.1	8.3	14.8	28.6	37.0	77.8
Ours	**10.6**	**19.0**	**34.6**	**43.1**	**97.4**	**6.2**	**10.0**	**16.0**	**18.8**	**31.2**	**4.5**	9.3	**20.1**	**25.3**	**64.1**	**6.4**	**13.9**	**27.9**	**36.0**	**75.4**

Table 3. Comparisons of mean angle errors on H3.6M dataset at 80ms, 160ms, 320ms, 400ms, 560ms, and 1000ms.

millisecond	80ms	160ms	320ms	400ms	560ms	1000ms
DMGNN [3]	0.38	0.65	0.94	1.04	1.24	1.64
MSR-GCN [1]	0.35	0.61	0.98	1.11	1.31	1.67
PGBIG [4]	0.30	0.54	0.89	1.02	1.23	1.61
SPGSN [2]	0.31	0.52	0.86	1.01	1.18	1.60
Ours	**0.27**	**0.48**	**0.81**	**0.96**	**1.01**	**1.52**



Color: Category of human motions
Number: Index of multi-view frequency representations

Figure 1. The t-SNE visualization of augmented multi-view frequency features on CMU Mocap dataset.

the frequency representation learning into a decomposition-aggregation scheme, we collect richer frequency representations from the proposed FDU and FAU components for robust human motion prediction.

4. Analysis for Few-sample Prediction

In the paper (Figure 4), we report few-sample prediction performances averaged over all human motions on the H3.6M dataset. Here, we supplement its detailed action-specific results to investigate the few-sample prediction performances on different kinds of human motions. As shown in Figure 2, we first choose four activities (walking, smoking, discussion, and posing) as examples and then provide five configurations (10%, 30%, 50%, 70%, and 90%) for few-sample training on each activity. We can see that our proposed decomposition-aggregation scheme has a clear advantages on extracting richer frequency representations from few training samples for robust human motion prediction. For example, when training with 10%-only samples, our system outperforms state-of-the-art methods by large margins: 20% on walking, 25% on smoking, 21% on discussion, and 19% on posing. These significant performance gains on few-sample prediction suggest that by enriching the spectral encoding of an input body motion, the decomposition-aggregation scheme extracts richer frequency features for robust motion prediction, making it less prone to overfitting on limited motion samples.

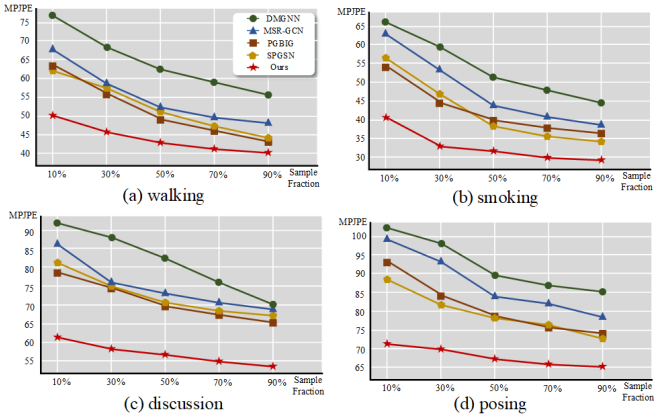


Figure 2. Few-sample prediction performances on different kinds of human motions.

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 *ICCV*, pages 11447–11456, 2021. 1, 2

[2] Maosen Li, Siheng Chen, Zijing Zhang, Lingxi Xie, Qi Tian, and Ya Zhang. Skeleton-parted graph scattering networks for 3d human motion prediction. In *ECCV*, 2022. 1, 2

[3] 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 *CVPR*, pages 211–220, 2020. 1, 2

[4] 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 *CVPR*, pages 6437–6446, 2022. 1, 2