MotionAug: Augmentation with Physical Correction for Human Motion Prediction
——Supplementary Material——

Takahiro Maeda Norimichi Ukita
Toyota Technological Institute, Japan
`{sd21601, ukita}@toyota-ti.ac.jp`

A. Dataset

In the following, we give detailed information about the HDM05 Motion Database [8] and post-processing for our experiments. HDM05 contains dynamic motion sequences such as *kick*, *punch*, and *jump* classes for 50 minutes in total. On the other hand, the standard benchmark Human3.6m [3] contains relatively static motions such as *eating*, *talking on the phone*, and *smoking* classes for about 1200 minutes in total. We find HDM05 more challenging due to dynamic motions and fewer data. Therefore, HDM05 is suitable for validating our motion data augmentation. We followed the post-processing procedure of a motion synthesis method [2] that cuts long motion sequences to clips of each motion class and retargets them to the uniform skeleton based on CMU Mocap [1] for VAE to learn motions independently from skeletal differences.

B. Additional Experiments on PD-residual Forces

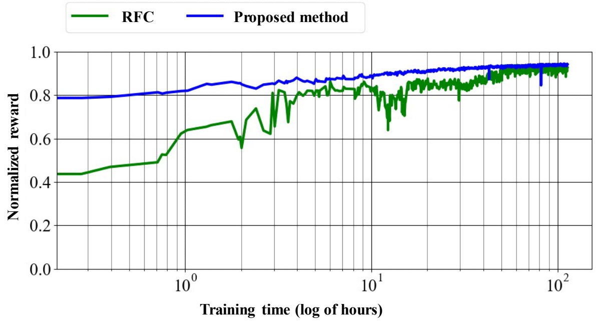
In Sec. 4.2, the convergence time comparison is shown only on the *kick* class motion. We further conducted the convergence time comparison on the *walk* and *punch* class motions used to compare motion data augmentation Sec. 4.3.

Implementation Details: We use the same setting as Sec. 4.2 to train imitation learning.

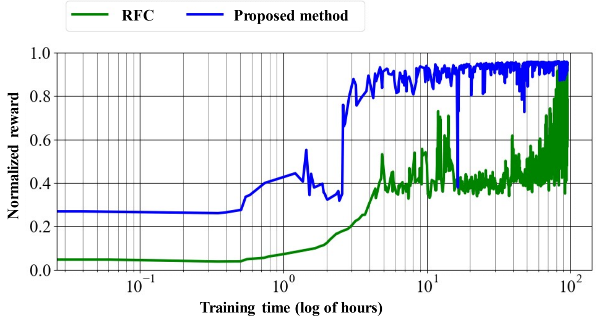
Results: The results are shown in Fig. 11. Again, imitation learning with our PD-residual forces converges faster and stabler than RFC [9].

C. Visualization of Sampling-near-samples

We visualized the sampled latent variables from the prior distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and proposed sampling-near-samples.



(a) comparison on the walk motion.



(b) comparison on the punch motion.

Figure 11. **Normalized rewards vs. training time in logarithmic scale.**

Implementation Details: We applied dimension reduction by PCA and subsequent UMAP [7] to map the latent variables to dim = 13 and dim = 2 respectively.

Results: The plot of the latent variables is shown in Fig. 12. One can observe that our sampling-near-samples method successfully samples representations(▲, ■, ⋆) located near the train set (▲, ■, ⋆). However, the prior distribution samples representations (●) from unlearned regions that the train set does not cover. As the visualization suggests, our sampling-near-samples method is robust to the sparsity

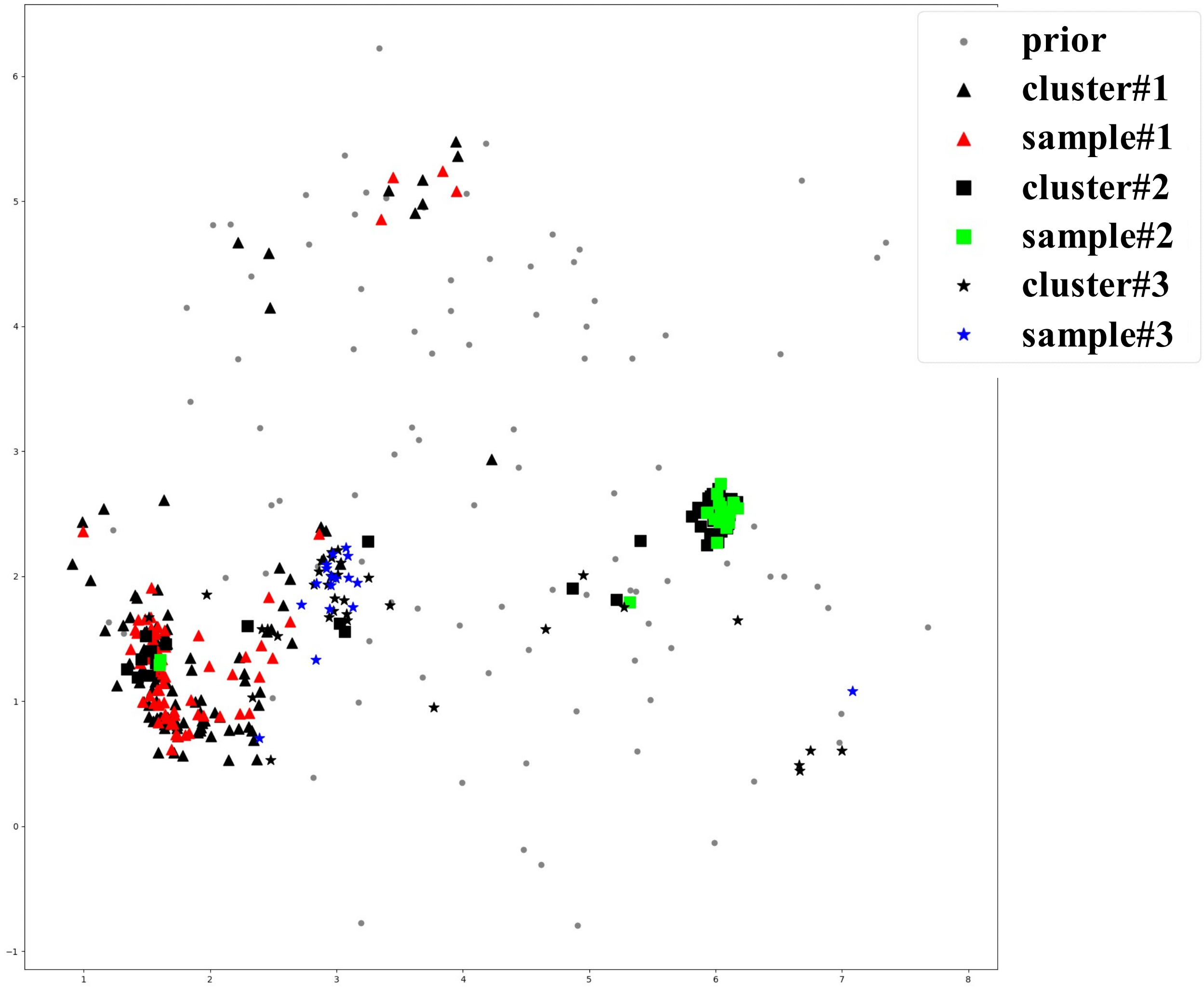


Figure 12. **Visualization of the sampled latent variables from the prior distribution and proposed sampling-near-samples.** We cluster the train set to three clusters denoted by black triangles(▲), squares(■), and stars(⋆).The prior distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ samples the latent variables denoted by gray dots(●) distributed on the unlearned regions that the train set does not cover. The samples from each cluster(▲, ■, ⋆) by our sampling-near-samples locate near the train set and succeed to synthesize dynamic motions even with inefficient data.

by sampling from the reliably learned regions.

D. Experiments on Additional Action Classes

We conducted further experiments on augmentation for human motion prediction to verify the effectiveness of our approach. Five action classes (*grab*, *deposit*, *jog*, *sneak*, *throw*) are added for evaluation. We chose the hand joint for *grab*, *deposit*, and *throw* to modify the motion sequences. For *grab*, *deposit*, *jog*, *sneak*, *throw* classes, the IK target positions are sampled from $([0.5, 2.0]r, [1.0, 1.0]h, [-1.7, 1.7] + \theta)$, $([0.5, 2.0]r, [1.0, 1.0]h, [-1.7, 1.7] + \theta)$, $([0.5, 2.0]r, [1.0, 1.0]h, [-0.3, 0.3] + \theta)$, $([0.5, 2.0]r, [1.0, 1.0]h, [-0.3, 0.3] + \theta)$, and $([0.5, 2.0]r, [1.0, 1.0]h, [-1.7, 1.7] + \theta)$ respectively. Other experimental settings except action classes follow Sec 4.3. The results are shown in Tables 3 and 4. Our proposed method outperforms in most cases on the RNN-based model. However, on the GCN-based model, our motion generation achieves the best performance in most cases. We can still optimize the augmentation parameters for better performances of the GCN-based model.

Table 3. Quantitative results of motion data augmentation on the RNN-based human motion prediction [5] on *grab*, *deposit*, *jog*, *sneak*, *throw*.

Methods			action class & timesteps [ms]									
	physical correction	motion debiasing	grab↓		deposit↓		jog↓		sneak↓		throw↓	
			100	400	100	400	100	400	100	400	100	400
No aug	-	-	0.71	1.92	0.72	**1.88**	0.91	1.67	0.46	1.27	1.24	**2.61**
Noise	-	-	0.71	1.92	0.69	1.92	0.91	1.65	0.45	1.21	1.24	2.56
VAE			0.83	2.06	0.83	2.18	0.97	1.79	0.53	1.50	**1.20**	**2.61**
IK			0.89	2.26	0.85	2.22	0.94	1.82	0.55	1.46	1.46	3.16
VAE & IK			0.78	1.99	0.78	2.08	0.90	1.80	0.46	1.38	1.27	2.81
VAE	✓		0.84	2.25	0.72	2.04	1.15	2.27	0.55	1.57	1.25	2.71
IK	✓		0.88	2.23	0.86	2.23	0.96	2.14	0.54	1.56	1.41	3.16
VAE & IK	✓		0.87	2.30	0.85	2.31	0.99	2.35	0.49	1.50	1.40	3.19
VAE	✓	✓	0.69	1.87	0.71	1.92	0.87	1.71	0.43	1.26	1.21	**2.61**
IK	✓	✓	0.70	1.81	0.74	1.99	**0.73**	**1.40**	0.41	1.16	1.37	2.93
VAE & IK	✓	✓	**0.64**	**1.68**	**0.72**	1.94	0.77	1.50	**0.39**	**1.14**	1.31	2.87

Table 4. Quantitative results of motion data augmentation on the SOTA GCN-based human motion prediction [4] on *grab*, *deposit*, *jog*, *sneak*, *throw*.

Methods			action class & timesteps [ms]									
	physical correction	motion debiasing	grab↓		deposit↓		jog↓		sneak↓		throw↓	
			100	400	100	400	100	400	100	400	100	400
No aug	-	-	0.56	1.79	0.55	**1.71**	0.66	1.33	0.27	0.94	1.03	2.45
Noise	-	-	0.56	1.79	0.55	1.70	0.66	1.31	0.27	0.93	1.03	2.40
VAE			0.52	1.65	0.53	**1.68**	0.67	1.32	0.30	0.95	1.06	**2.28**
IK			0.58	1.79	0.56	1.76	0.62	**1.24**	0.29	0.92	1.08	2.45
VAE & IK			0.53	1.69	**0.52**	**1.68**	**0.61**	1.26	0.27	**0.90**	**1.02**	2.31
VAE	✓		0.59	1.80	0.57	1.74	0.72	1.49	0.33	1.06	1.05	2.43
IK	✓		0.55	1.77	0.53	1.76	0.69	1.51	0.30	1.09	1.07	2.57
VAE & IK	✓		0.55	1.80	0.54	1.74	0.73	1.68	0.30	1.10	1.05	2.47
VAE	✓	✓	0.51	1.64	0.55	1.77	0.72	1.45	0.29	1.03	1.09	2.37
IK	✓	✓	0.52	1.67	0.55	1.73	0.63	1.28	0.28	0.98	1.08	2.39
VAE & IK	✓	✓	**0.48**	**1.60**	0.53	1.70	0.64	1.34	**0.26**	**0.90**	1.06	2.39

E. Experiments on Transformer-based Human Motion Prediction Model

We further validated the effectiveness of our approach on the transformer-based human motion prediction model [6]. Other experimental settings except the prediction model follow Sec 4.3 and Sec D. The results are shown in Tables 5 and 6. Our proposed method outperforms in most cases.

F. Performance Comparison on Augmentation by DeepMimic

We evaluated the augmentation by DeepMimic on human motion prediction. All experimental settings, including the sampling of targets, follow Sec 4.3. The result is shown in Table 7. DeepMimic has limited performance compared to our approach.

References

[1] Fernando De la Torre, Jessica Hodgins, Adam Bargteil, Xavier Martin, Justin Macey, Alex Collado, and Pep Beltran. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. 2009. 1

[2] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Trans. Graph.*, 35(4):138:1–138:11, 2016. 1

[3] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, 2014. 1

Table 5. Quantitative results of motion data augmentation on the Transformer-based human motion prediction [6].

Methods			Prediction errors↓[rad] on each action class & timesteps [ms]								
	physical correction	motion debiasing	punch			kick			walk		
			100	200	400	100	200	400	100	200	400
No aug	-	-	0.61	1.66	2.55	0.5	1.48	2.20	0.27	0.93	1.65
Noise	-	-	0.60	1.65	2.53	0.50	1.47	2.20	0.27	0.88	1.55
VAE			**0.47**	**1.06**	**1.51**	0.40	1.01	1.42	0.22	0.58	0.89
IK			0.49	1.14	1.58	0.43	1.15	1.67	0.30	0.86	1.31
VAE & IK			**0.47**	**1.06**	1.49	0.41	1.00	1.38	0.26	0.79	1.25
VAE	✓		0.51	1.22	1.71	0.44	1.15	1.64	0.22	0.58	0.92
IK	✓		0.51	1.10	1.56	0.45	1.17	1.70	0.28	0.79	1.24
VAE & IK	✓		0.55	1.25	1.83	0.46	1.15	1.68	0.22	0.59	0.96
VAE	✓	✓	0.52	1.18	1.69	0.42	1.04	1.48	0.23	0.60	0.97
IK	✓	✓	**0.47**	1.12	1.63	0.39	0.86	1.18	0.21	0.59	0.95
VAE & IK	✓	✓	0.51	1.18	1.73	**0.38**	**0.84**	**1.15**	**0.19**	**0.48**	**0.77**

Table 6. Quantitative results of motion data augmentation on the Transformer-based human motion prediction [6] on *grab*, *deposit*, *jog*, *sneak*, *throw*.

Methods			action class & timesteps [ms]									
	physical correction	motion debiasing	grab↓		deposit↓		jog↓		sneak↓		throw↓	
			100	400	100	400	100	400	100	400	100	400
No aug	-	-	0.26	1.83	0.25	**1.80**	0.26	1.11	0.14	1.04	0.48	2.36
Noise	-	-	0.26	1.83	0.25	1.80	0.27	1.13	0.14	1.01	0.48	2.36
VAE			0.24	1.20	**0.23**	1.34	0.26	1.13	0.13	0.80	0.46	2.02
IK			0.23	1.25	**0.23**	1.32	0.25	1.02	0.11	0.71	0.47	2.00
VAE & IK			0.32	1.32	0.26	1.41	0.25	1.02	0.11	0.69	0.56	2.49
VAE	✓		0.23	1.29	0.24	1.47	0.26	1.09	0.12	0.72	0.47	2.15
IK	✓		0.23	1.21	0.24	1.34	0.25	1.07	0.11	0.69	**0.41**	**1.56**
VAE & IK	✓		0.23	1.24	0.29	1.75	0.25	1.08	0.12	0.72	0.44	1.73
VAE	✓	✓	0.22	1.20	**0.23**	1.43	0.25	1.00	0.11	0.66	0.45	1.96
IK	✓	✓	**0.20**	**1.04**	**0.23**	**1.29**	**0.23**	0.85	**0.10**	**0.60**	0.43	1.73
VAE & IK	✓	✓	0.26	1.11	0.25	1.53	**0.23**	**0.83**	**0.10**	**0.60**	0.45	2.02

Table 7. Performance comparison of DeepMimic augmentation on *kick*.

Prediction errors on action class↓	kick		
timesteps[ms]	100	200	400
GCN+No Aug	1.08	1.68	2.26
GCN+ours	**0.52**	**1.23**	**1.74**
GCN+ours(w/o residual force)	1.07	1.65	2.08
GCN+DeepMimic augmentation	1.13	1.72	2.24

[4] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *ICCV*, pages 9488–9496, 2019. 3

[5] Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *CVPR*, pages 4674–4683, 2017. 3

[6] Ángel Martínez-González, Michael Villamizar, and Jean-Marc Odobez. Pose transformers (POTR): human motion prediction with non-autoregressive transformers. In *CVPR Workshops*, pages 2276–2284, 2021. 3, 4

[7] Leland McInnes and John Healy. UMAP: uniform manifold approximation and projection for dimension reduction. *CoRR*, abs/1802.03426, 2018. 1

[8] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation mocap database hdm05. Technical Report CG-2007-2, Universität Bonn, June 2007. 1

[9] Ye Yuan and Kris Kitani. Residual force control for agile human behavior imitation and extended motion synthesis. In *NeurIPS*, 2020. 1