LAL: Enhancing 3D Human Motion Prediction with Latency-aware Auxiliary Learning

Supplementary Material

1. More Experimental Results

In main paper Table 3, we omit some prediction errors at certain timestamps and running time of CMU-Mocap [2] and 3DPW [6] due to limited space. Here, we provide the complete results of LTD [4], STS [5] and SPGSN [3] with LAL-T_L in Table 1 and 2.

millisecond	CMU-Mocap average 80	160	320	400	1000	running time (ms)
LTD [4]	9.9	18.0	33.6	41.0	81.9	39.29
LTD-LAL-1	**9.3**	**17.4**	**32.1**	**39.8**	**81.8**	39.20
LTD-LAL-2	-	**16.7**	**30.9**	**39.2**	**80.7**	39.04
LTD-LAL-3	-	**15.6**	**29.4**	**38.5**	**79.7**	38.93
LTD-LAL-4	-	-	**28.7**	**38.3**	**78.4**	38.87
LTD-LAL-5	-	-	**28.1**	**37.8**	**76.5**	38.74
STS [5]	10.8	18.2	31.2	41.1	81.8	30.44
STS-LAL-1	11.4	**17.9**	**30.5**	**41.4**	**81.3**	30.39
STS-LAL-2	-	**16.8**	**28.4**	**39.5**	**80.7**	30.30
STS-LAL-3	-	**15.3**	**28.0**	**37.2**	**79.1**	30.18
STS-LAL-4	-	-	**27.2**	**37.6**	**76.2**	30.12
STS-LAL-5	-	-	**26.7**	**36.9**	**75.8**	30.06
SPGSN [3]	8.3	14.8	28.6	37.0	77.8	49.37
SPGSN-LAL-2	-	**13.4**	**27.1**	**35.7**	**76.8**	49.29
SPGSN-LAL-3	-	**12.5**	**25.8**	**34.9**	**75.1**	49.26
SPGSN-LAL-4	-	-	**25.1**	**34.8**	**74.9**	49.12
SPGSN-LAL-5	-	-	**24.5**	**33.9**	**74.2**	49.08

Table 1. Comparisons of MPJPE average errors in CMU-Mocap, where $T_L = 1, 2, 3, 4, 5$. Our LAL helps achieve progress without extra time cost during testing.

Similar to main paper Table 2 and 3, there is no LTD-LAL-1 for 3DPW, also no SPGSN-LAL-1 for CMU-Mocap and 3DPW, as their running time has already exceeded the *maximum* latency duration that it supports. Note that the sampling rate of 3DPW differs from the other two datasets, so the timestamps appear different (but we keep aligned with baselines to ensure fair comparison). Overall, LAL achieves improved results over baselines constantly, without the need for extra time cost during testing, which shows both the effectiveness and efficiency of our latency-aware prediction manner.

millisecond	3DPW Average 200	400	500	800	1000	running time (ms)
LTD [4]	35.6	67.5	80.2	106.8	117.8	41.66
LTD-LAL-2	**34.3**	**65.1**	**78.8**	**106.6**	**116.4**	41.50
LTD-LAL-3	**33.1**	**65.0**	**77.0**	**104.8**	**115.3**	41.44
LTD-LAL-4	**31.7**	**63.8**	**76.8**	**103.1**	**113.7**	41.29
LTD-LAL-5	**31.2**	**62.3**	**75.9**	**101.0**	**110.8**	41.20
STS [5]	37.8	67.5	77.3	106.7	112.2	31.84
STS-LAL-1	**37.2**	**66.8**	**76.5**	108.4	**111.4**	31.76
STS-LAL-2	**36.0**	**65.4**	**76.4**	**105.1**	**110.8**	31.75
STS-LAL-3	**33.8**	**64.8**	**75.3**	**103.6**	**109.0**	31.64
STS-LAL-4	**33.4**	**63.0**	**74.7**	**102.8**	**107.6**	31.55
STS-LAL-5	**34.6**	**61.9**	**73.1**	**101.6**	**106.6**	31.43
SPGSN [3]	32.9	64.5	76.2	104.0	111.1	50.61
SPGSN-LAL-2	**30.9**	**63.1**	**75.6**	**103.4**	**109.3**	50.46
SPGSN-LAL-3	**30.6**	**61.2**	**74.3**	**101.7**	**107.8**	50.39
SPGSN-LAL-4	**28.4**	**59.4**	**72.8**	**100.1**	**106.9**	50.35
SPGSN-LAL-5	**27.1**	**58.9**	**71.3**	**99.8**	**106.3**	50.28

Table 2. Comparisons of MPJPE average errors in 3DPW. The results and time consumption appear similar trend as in Table 1.

2. Ablation Experiments

2.1. Training loss hyper-parameters

In main paper Figure 8, we show different $(\lambda_{\mathcal{P}}, \lambda_{\mathcal{A}})$ settings of Eq. (4), with STS [5] as the compared baseline and $T_L = 3$. Here, we provide the corresponding ablation with baseline LTD [4] and SPGSN [3] in Figure 1, and the improvement is most significant when $(\lambda_{\mathcal{P}} = 1.0, \lambda_{\mathcal{A}} = 1.0)$ for both of them. Note that these experiments are all conducted on Human3.6M [1], and the hyper-parameter setting remain the same for each dataset and each T_L value.

2.2. Ablations on refinement step

Refinement loss hyper-parameters. We provide the results of different (λ_Z, λ_U) weight settings in our main paper refinement loss Eq. (6). As shown in Figure 2 (left and middle), all of them can bring benefits more or less. We select $\lambda_Z = 1.0, \lambda_U = 1.0$, and maintain this setting for all baselines in our experiments.

Why not merge the refinement loss into main training.

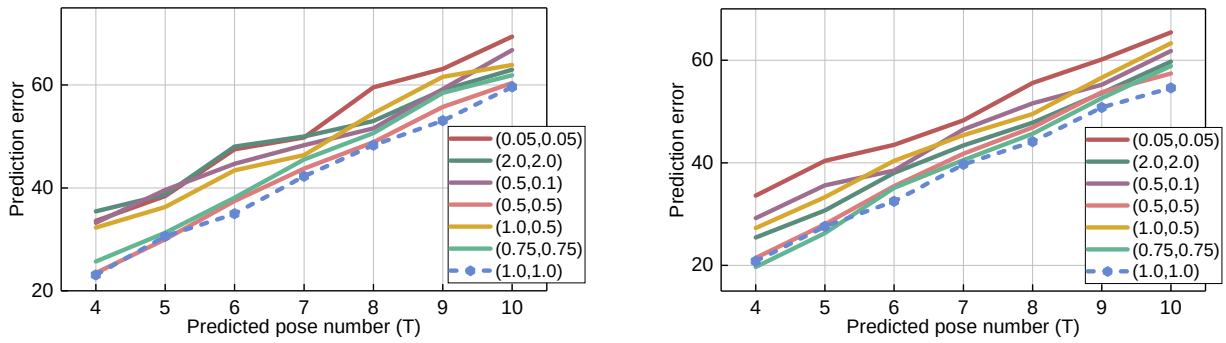

Figure 1. Results of different $(\lambda_{\mathcal{P}}, \lambda_{\mathcal{A}})$ settings of Eq. (4) in LTD-LAL-3 (left) and SPGSN-LAL-3 (right). The best is achieved when $\lambda_{\mathcal{P}} = 1.0, \lambda_{\mathcal{A}} = 1.0$, drawn in dashed blue line.

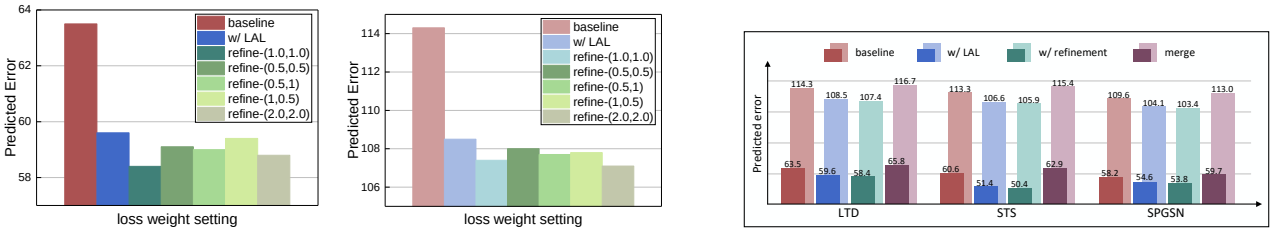

Figure 2. Left and middle: results of different (λ_Z, λ_U) settings in Eq. (6) for LTD-LAL-3. Left subfigure shows errors at 400 ms, while middle shows 1000 ms. We set $\lambda_Z = 1.0, \lambda_U = 1.0$ that yields constant improvement. Right: merging refinement loss into LAL as an end-to-end training manner, showing no improvement compared to baselines. Dark colors indicate 400 ms errors while faded 1000 ms.

As we treat the refinement as an optional post-processing step, we should validate the effectiveness of using such separate training rather than end-to-end manner. We merge the feature statistics-based alignment loss into our main training, which let the primary prediction branch approximate the distribution of the auxiliary latent and final outputs from scratch, i.e., directly involving $\mathcal{L}_{\text{fea},Z}$ and $\mathcal{L}_{\text{fea},U}$ into our LAL. However, the auxiliary branch at the beginning is not well-trained that tends to generate wrong patterns of outputs (especially the final output U), which may confuse the primary branch whether to fit towards the ground truth or the unstable auxiliary outputs. The results are shown in Figure 2 (right) purple bars, with no improvement and even higher errors than baselines, which verifies the above reasoning.

Limitations. In real-world scenarios, the intelligent system may confront incomplete observations due to the occlusion of human body or noise perturbation, which greatly impedes the prediction accuracy. Improving algorithm robustness against these factors is a worthy study that will further contribute to this task.

References

[1] 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 transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 1

[2] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *CVPR*, 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 *ECCV*, pages 18–36. Springer, 2022. 1

[4] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *ICCV*, pages 9489–9497, 2019. 1

[5] Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. Space-time-separable graph convolutional network for pose forecasting. In *ICCV*, pages 11209–11218, 2021. 1

[6] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, pages 601–617, 2018. 1