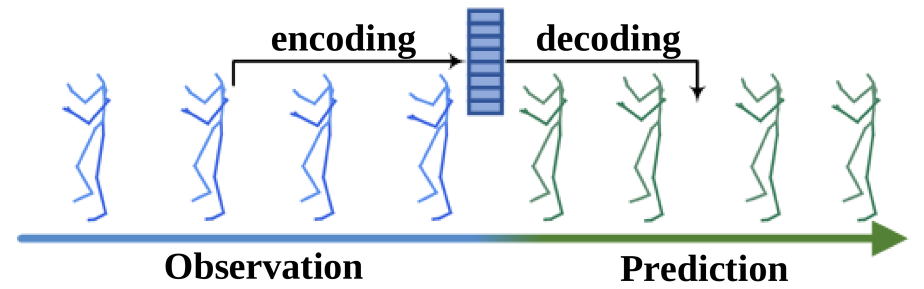
Overlooked Poses Actually Make Sense: Distilling Privileged Knowledge for Human Motion Prediction

Xiaoning Sun[1], Qiongjie Cui[1], Huaijiang Sun[1], Bin Li[2], Weiqing Li[1], Jianfeng Lu[1]
[1]Nanjing University of Science and Technology, [2]Tianjin AiForward Science and Technology Co., Ltd., China

TASK

Human Motion Prediction: Predicting a future motion sequence based on the historical observed one.

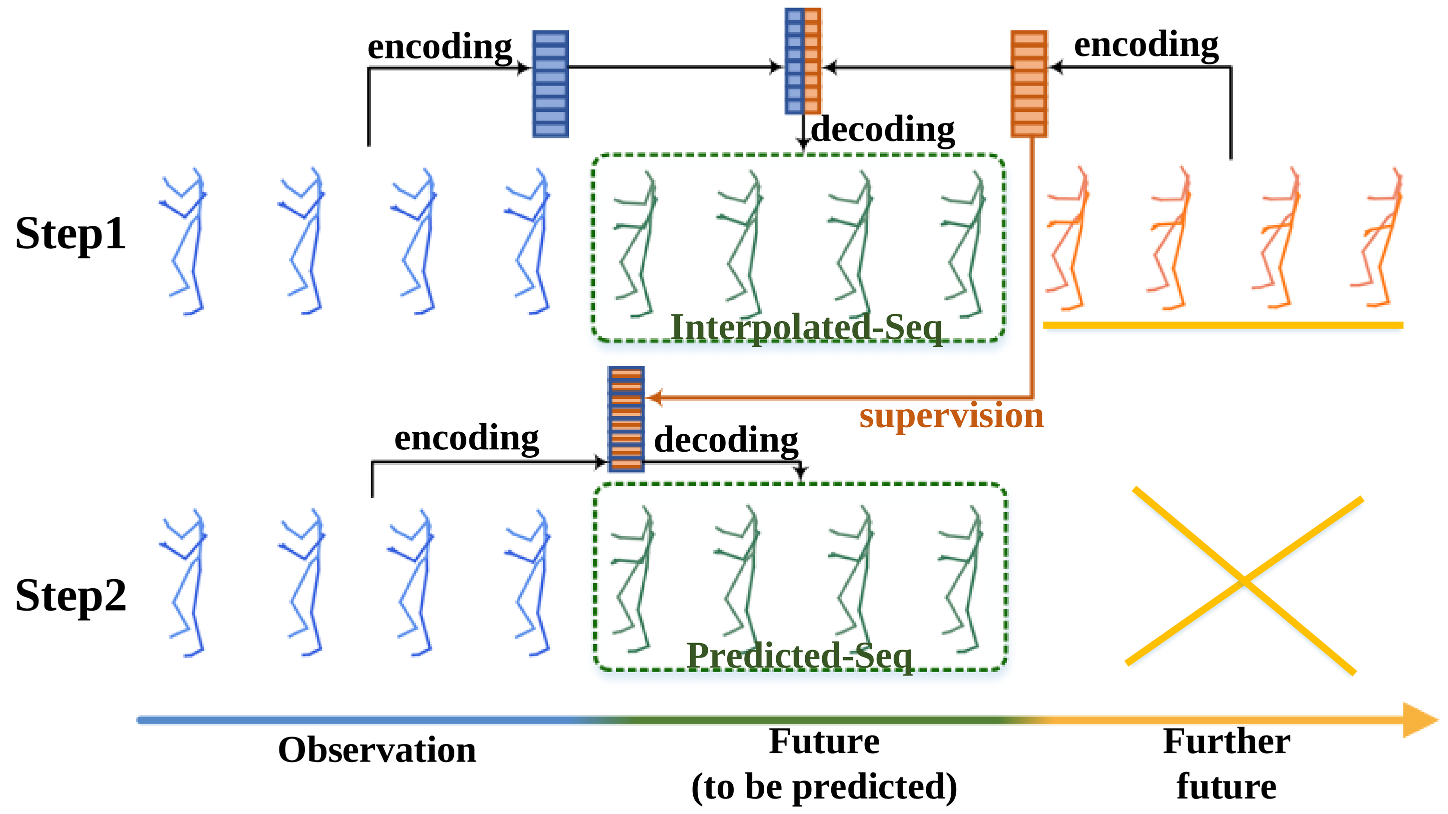


Current works focus on directly finding the extrapolation mapping relation in-between, which may lead to failing results due to the inherent challenge of multivariate time series extrapolation problem.

INSIGHT

Generally, sequence interpolation is easier to operate and often yields an overall better result than extrapolation. We introduce the overlooked poses which exist *after* the predicted sequence, in the hope of building a prediction pattern that shares similar spirit with interpolation.

A new challenge: These poses would not exist in the real prediction phase, so how could the model acquire the information from them and therefore achieve better performance?

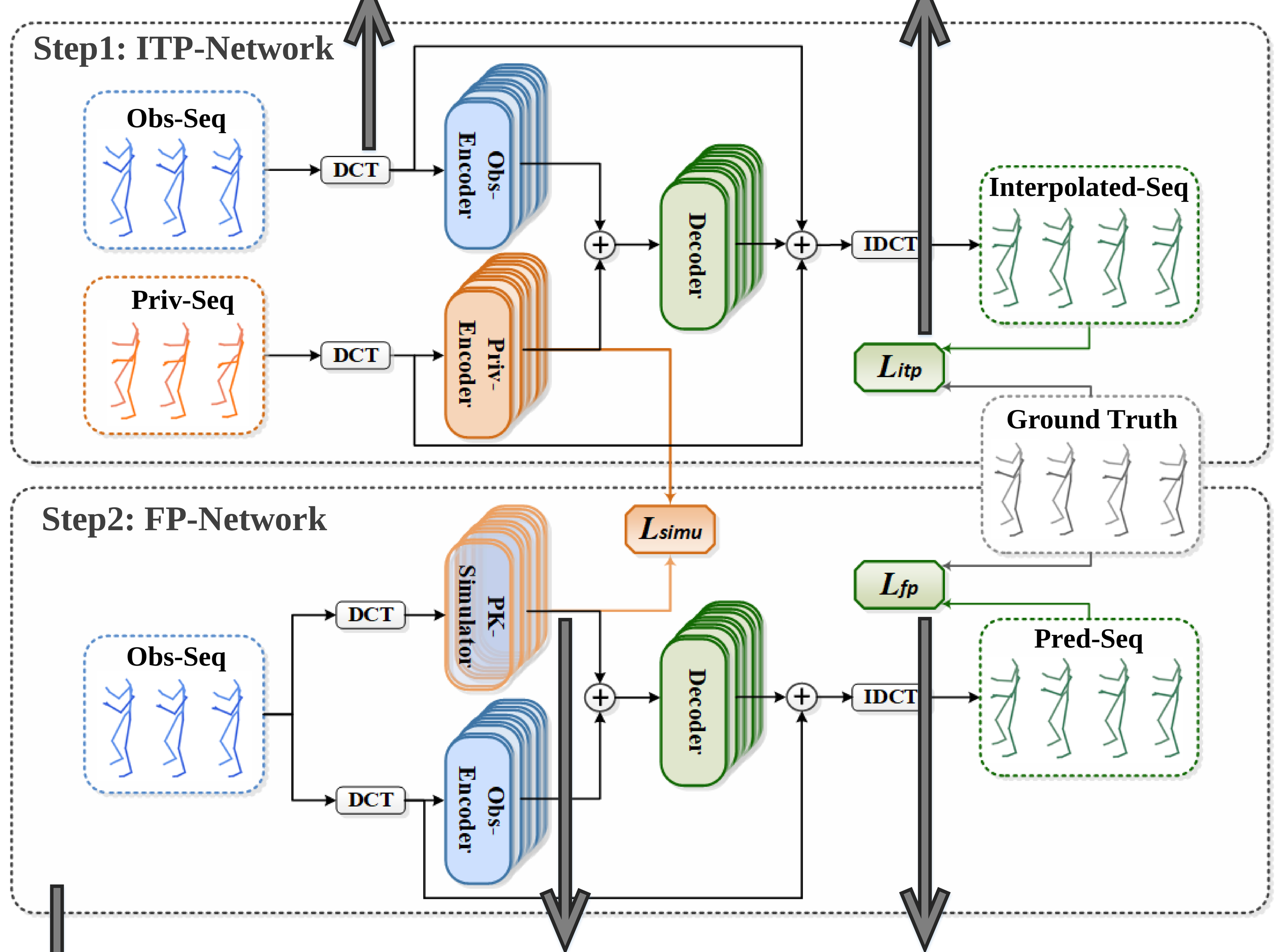


 # Poses in orange: named as privileged sequence that can be used during interpolation but discarded during prediction (i.e. extrapolation)
 # Step1 (interpolation step): learn a representation of privileged information
 # Step2 (extrapolation step): use the representation as supervision to transfer the privileged knowledge into prediction phase.

PK-GCN

DCT transformation on the observed/privileged sequence before being fed into the network.

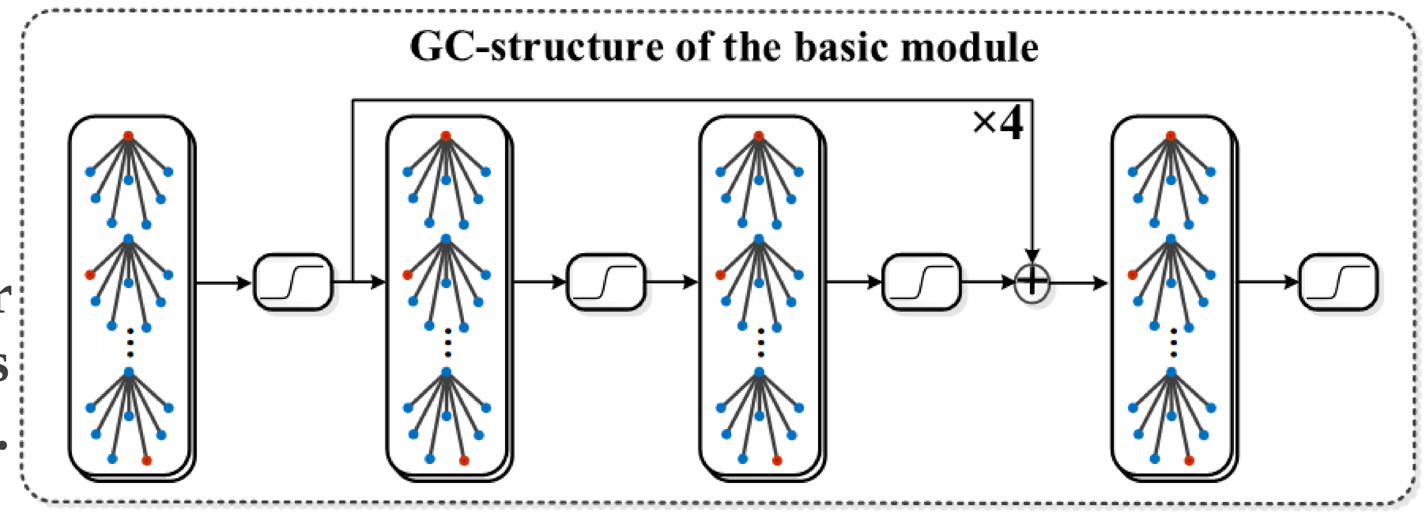
Interpolation loss that measures the discrepancy between the interpolated sequence and GT to ensure the privileged information is learned.



Knowledge distillation realized by a simulator that takes as input the observed sequence but approximates the privileged representation learned in Step1 (the parameters in Step1 is fixed now). The approximation is measured by our simulation loss.

Final prediction loss that measures the discrepancy between the predicted sequence and GT.

Module structures of all encoders, the simulator and decoders are the same.



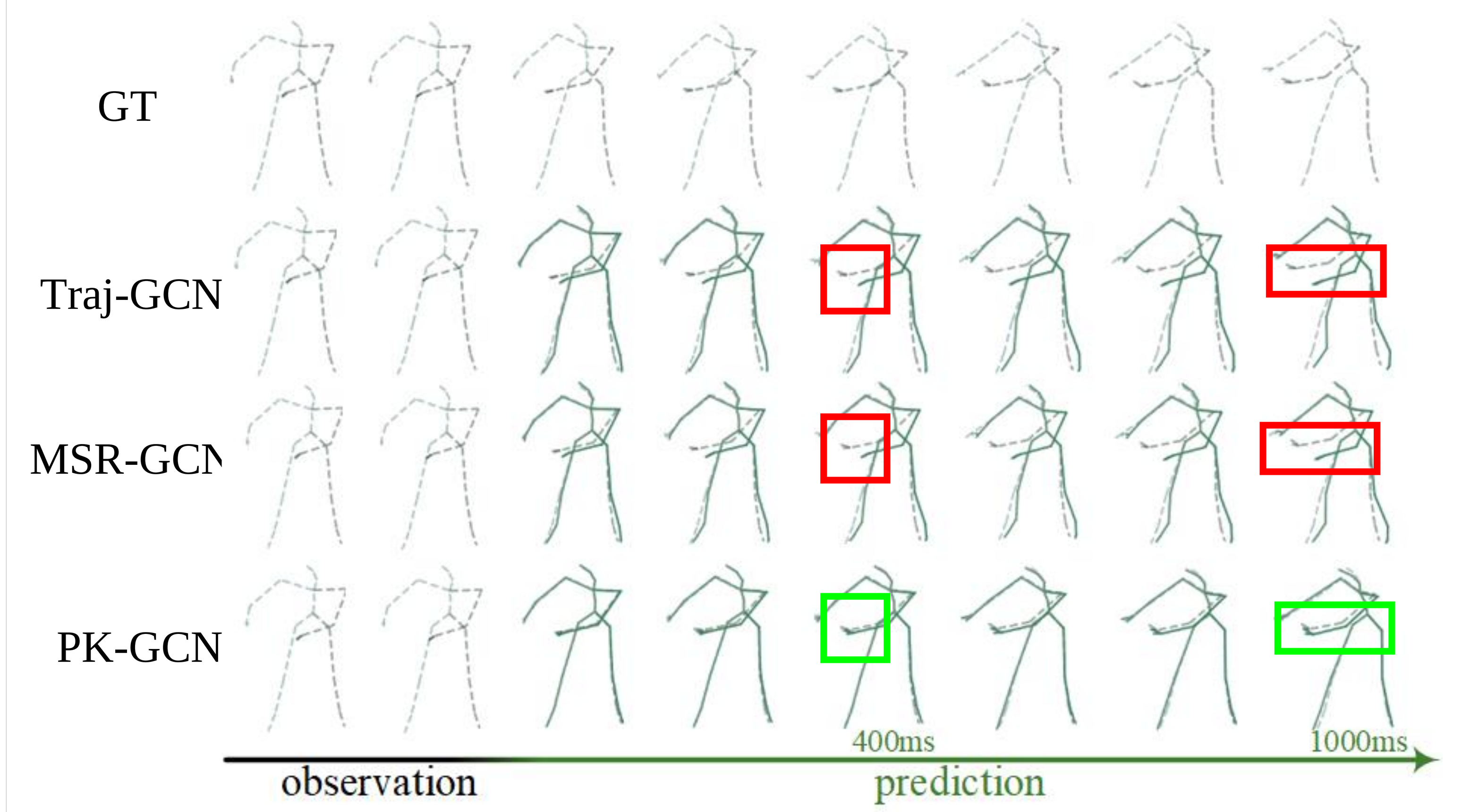
Experiments

Metrics.
MPJPE: Mean Per Joint Position Error.
MAE: Mean Angle Error.

Human3.6M

	MPJPE						MAE					
millisecond (ms)	80	160	320	400	560	1000	80	160	320	400	560	1000
Res. sup [1]	34.7	62.0	101.1	115.5	135.8	167.3	0.36	0.67	1.02	1.15	-	-
Traj-GCN [2]	12.7	26.1	52.3	63.5	81.6	114.3	0.32	0.55	0.91	1.04	1.27	1.66
MSR-GCN [3]	12.1	25.6	51.6	62.9	81.1	114.2	-	-	-	-	-	-
STS-GCN [4]	16.3	28.1	54.4	65.8	85.1	117.0	0.31	0.57	0.91	1.03	1.22	1.61
PK-GCN	**10.8**	**23.3**	**48.2**	**57.4**	**76.1**	**106.4**	**0.29**	**0.54**	**0.85**	**0.96**	**1.15**	**1.57**

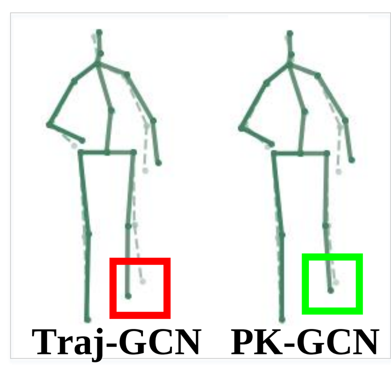
	CMU-MPJPE						3DPW-MPJPE				
millisecond (ms)	80	160	320	400	560	1000	200	400	600	800	1000
Res. sup [1]	24.0	43.0	74.5	87.2	105.5	136.3	113.9	173.1	191.9	201.1	210.7
Traj-GCN [2]	11.5	20.4	37.8	46.8	55.8	86.2	35.6	67.8	90.6	106.9	117.8
MSR-GCN [3]	**8.1**	18.7	34.2	42.9	53.7	83.0	-	-	-	-	-
PK-GCN	9.4	**17.1**	**32.8**	**40.3**	**52.2**	**79.3**	**34.8**	**66.2**	**88.1**	**104.3**	**114.2**



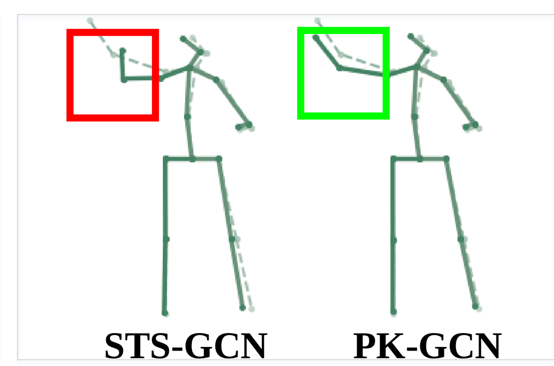
Visualized comparisons of predictions on a motion sequence *Washwindow* in CMU-Mocap. Red boxes indicate unexpected deviations. Ours in green boxes are closer to GT.

References:
[1] Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. In: CVPR. pp. 2891–2900 (2017)
[2] Mao, W., Liu, M., Salzmann, M., Li, H.: Learning trajectory dependencies for human motion prediction. In: ICCV. pp. 9489–9497 (2019)
[3] Dang, L., Nie, Y., Long, C., Zhang, Q., Li, G.: MSR-GCN: Multi-scale residual graph convolution networks for human motion prediction. In: ICCV. pp. 11467–11476 (2021)
[4] Sofianos, T., Sampieri, A., Franco, L., Galasso, F.: Space-time-separable graph convolutional network for pose forecasting. In: ICCV. pp. 11209–11218 (2021)

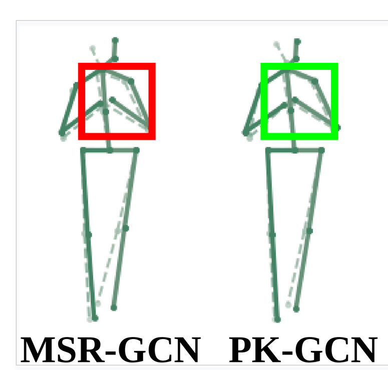
1000ms *walking*



1000ms *discussion*



1000ms *greeting*



1000ms *takingphoto*