Multi-Person Pose Forecasting with Individual Interaction Perceptron and Prior Learning – Supplementary Materials –

Peng Xiao[1], Yi Xie[1], Xuemiao Xu[1,2,3(✉)], Weihong Chen[1], and Huaidong Zhang[1,2(✉)]

[1] South China University of Technology, Guangzhou, China
[2] Guangdong Engineering Center for Large Model and GenAI Technology
[3] Guangdong Provincial Key Lab of Computational Intelligence and Cyberspace Information
{xuemx,huaidongz}@scut.edu.cn

1 Supplemental Experiments

IAFormer is designed to evaluate the people's role in interaction and learn the interaction prior for multi-person forecasting. In this material, we conducted more comprehensive supplemental ablation experiments and analysis of the proposed model. In Sections 1.1, 1.2and 1.3, we validate the IAFormer's design on the number of IAFormer/Encoder/Decoder Blocks and size of Interaction Knowledge Space; In Section 1.4, we analyze the efficiency of our model compared with the existing methods via the Flops/Params measures.

	Number of Blocks	0.2s↓	0.4s↓	0.6s↓	0.8s↓	1s↓	Avg↓
JPE	2 Blocks	36	71	102	133	166	101.6
	5 Blocks	32	65	96	127	159	95.8
	8 Blocks	35	69	100	131	162	99.4
	10 Blocks	34	67	99	129	160	97.8
APE	2 Blocks	25	54	76	94	108	71.4
	5 Blocks	23	50	71	89	103	67.2
	8 Blocks	25	54	74	92	106	70.2
	10 Blocks	24	51	73	90	104	68.4

Table 1: Ablation results (in mm) on number of IAFormer Blocks base on CMU-Mocap (UMPM). [1, 2]

1.1 Ablation on IAFormer Block

In supplemental ablation experiments, the other blocks remain the same as the model in the main text. The IAFormer Block is introduced in section 3.3 in the main text, which combines features from the Interaction Perceptron Module

(IPM) and Interaction Prior Learning Module (IPLM) to predict human pose better. From Table 1, it is evident that the "5 Blocks" case exceeds the other case across all forecasting time. Specifically, the "5 Blocks" case outperforms "2 Blocks" and "10 Blocks" by 5.8 mm and 2 mm JPE in average time. This suggests that the 5 IAFormer Blocks are more suitable for our experimental task.

1.2 Ablation on Multi-Pose Encoder/Decoder Block

The Multi-Pose Encoder/Decoder is based on GCN Block in [3]. The role of them is to perform feature mapping. From Table 2, the model with "5 Blocks" performs the best across 4 different settings. In detail, the "5 Blocks" case outperforms "2 Blocks" and "8 Blocks" by 3.8 mm and 3.2 mm APE in average time, respectively. This suggests that the 5 Multi-Pose Encoder/Decoder Blocks are more suitable for our experimental task.

	Number of Blocks	0.2s↓	0.4s↓	0.6s↓	0.8s↓	1s↓	Avg↓
JPE	2 Blocks	36	71	102	133	165	101.4
	5 Blocks	32	65	96	127	159	95.8
	8 Blocks	36	70	101	132	164	100.6
	10 Blocks	35	69	100	131	162	99.4
APE	2 Blocks	26	54	75	93	107	71.0
	5 Blocks	23	50	71	89	103	67.2
	8 Blocks	26	53	75	92	106	70.4
	10 Blocks	25	53	74	92	106	70.0

Table 2: Ablation results (in mm) on number of Multi-Pose Decoder/Encoder Blocks base on CMU-Mocap (UMPM). [1, 2]

1.3 Ablation on size of Interaction Knowledge Space (IKS)

Table 3 conducts the ablation studies toward the size of IKS in IAFormer. The results show that our method achieves the best performance when the size is set to 256. Note that all the results in this Table outperform SOTA (99.0), indicating that IAFormer is not very sensitive to this parameter. We will add this discussion in the revision.

1.4 Params and Flops analysis.

In order to verify the complexity of our proposed method, we have conducted experiments concerning the model's Params and Flops. Table 4 shows our lightweight IAFormer reaches a state-of-the-art (SOTA) performance with 2.23% FLOPs and 46.19% Params of the JRFormer [5].

Size of IKS	200ms↓	600ms↓	1000ms↓	Avg↓
64	33.9	99.2	160.7	97.9
128	33.4	99.1	160.5	97.7
256	**32.1**	**96.5**	**159.2**	**95.9**
512	32.6	97.7	160.2	96.8
1024	33.3	98.3	160.8	97.5

Table 3: Study of IKS's size on CMU-Mocap(UMPM).

Method	IAFormer (ours)	JRFormer [5]	TBIFormer [4]
FLOPs (G)	**0.54**	24.2	3.11
Params (M)	**2.61**	3.74	5.65
JPE (mm)	**95.9**	99.0	107.0

Table 4: FLOPs (per-sample) and Parameters study.

2 Limitation and future work.

This work only considers pose forecasting within a limited period. We may conduct future research on continuous forecasting based on this work. We will also consider more potential information present in human-human or human-object interactions to achieve more accurate pose forecasting.

References

1. Van der Aa, N., Luo, X., Giezeman, G.J., Tan, R.T., Veltkamp, R.C.: Umpm benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 1264–1269 (2011)
2. CMU-Graphics-Lab: Cmu graphics lab motion capture database (2003), http://mocap.cs.cmu.edu/
3. Ma, T., Nie, Y., Long, C., Zhang, Q., Li, G.: Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6437–6446 (2022)
4. Peng, X., Mao, S., Wu, Z.: Trajectory-aware body interaction transformer for multi-person pose forecasting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17121–17130 (2023)
5. Xu, Q., Mao, W., Gong, J., Xu, C., Chen, S., Xie, W., Zhang, Y., Wang, Y.: Joint-relation transformer for multi-person motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9816–9826 (2023)