Prior-aware Dynamic Temporal Modeling Framework for Sequential 3D Hand Pose Estimation (Supplemental Material)

In the supplemental material, we provide:

- more details of network structure in Sec. 1,
- the details of loss function in Sec. 2,
- more qualitative results in Sec. 3,
- discussion on limitation and future work in Sec. 4.

Note that all the notation and abbreviations here are consistent with the main manuscript.

1. Details of Network Structure

In this section, we introduce the details of the network structure. Given the input 3D skeleton sequence $\mathbf{X} \in \mathbb{R}^{T\times J\times C}$, we first project it to a high-dimensional feature $\mathbf{F}^{init} \in \mathbb{R}^{T\times J\times C}$ using a linear layer, then add learnable spatial positional encoding $\mathbf{P}^{\mathrm{s}} \in \mathbb{R}^{1\times J\times C}$ and learnable temporal positional encoding $\mathbf{P}^{\mathrm{t}} \in \mathbb{R}^{T\times 1\times C}$ to it. Similar to MotionBERT [6], residual connection and layer normalization (LayerNorm) are used to all self-attention module and the dynamic temporal module result.

2. Definition of Loss Function

Similar to previous methods [2, 4], we supervise the joints and mesh vertices as follows:

$$L_{joint} = \sum_{i=0}^{T-1}\sum_{j=0}^{J-1} SL1\left(\mathbf{P}_{i,j}, \mathbf{P}_{i,j}^{gt}\right), \tag{1}$$

$$L_{mesh} = \sum_{i=0}^{T-1}\sum_{j=0}^{V-1} SL1\left(\mathbf{V}_{i,j}, \mathbf{V}_{i,j}^{gt}\right), \tag{2}$$

$$L_{mano} = \sum_{i=0}^{T-1} SL1\left(\theta_i, \theta_i^{gt}\right) + SL1\left(\beta, \beta^{gt}\right), \tag{3}$$

where $SL1$ represents the smooth L1 loss [1, 3]; T, J, V represents the number of frames, the number of joints and the number of mesh vertices, respectively.

In order to make the network estimation result smoother, we use acceleration loss [5] to supervise the predicted joints as follows:

$$L_{acc} = \sum_{i=0}^{T-3}\sum_{j=0}^{J-1} SL1\left(\mathbf{A}_{i,j}^{3D}, \mathbf{A}_{i,j}^{3D,gt}\right), \tag{4}$$

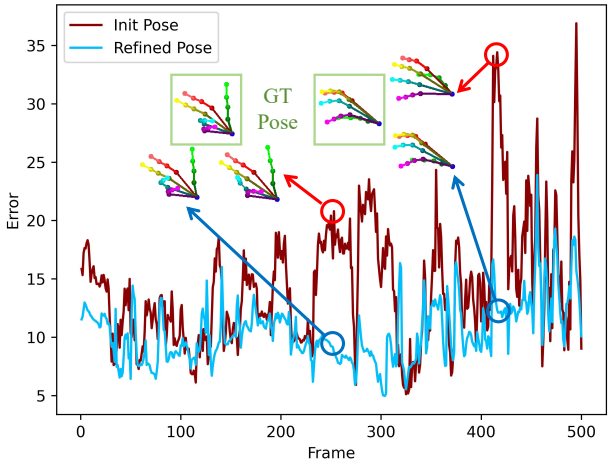

Figure 1. The initial error curve and refined error curve. We show the initial poses and the refined poses at some specific frames.

where $\mathbf{A}^{3D}$ is the computed acceleration from predicted pose and $\mathbf{A}^{3D,gt}$ is the ground-truth acceleration.

3. More Qualitative Results

As shown in Fig. 1, our method can generate smoother estimation results and correct the erroneous estimation of specific patterns. In particular, our method effectively reduces some extremely large errors, which is very important for some practical hand interaction applications. On the other hand, our method can correct some local errors, such as the ring finger error. At the same time, benefiting from the global temporal modeling, our method can correct the global hand pose, such as the direction of the hand.

Meanwhile, we provide qualitative results for four representative scenarios in the supplementary video, including self-occlusion, subtle high-frequency hand motion, large hand motion, and global hand rotation. For prediction errors caused by self-occlusion, our method can use timing information to correct the hand pose, thereby obtaining an smoother and more accurate results. For scenes with subtle and high-frequency changes in motion trends, our method can effectively avoid latency problems. In scenes with large hand movements and global hand rotations, our method can

achieve global motion smoothness and refine the inaccurate local joints, which shows the importance of combining global long-term temporal modeling with local short-term temporal modeling.

4. Limitation and Future Work

Although our method has achieved significant improvements in hand-object and two-hand interaction scenarios, it does not explicitly model the interactions, which may limit its performance in these complex scenarios. Additionally, similar to MotionBERT, our method could leverage large-scale pre-training on multiple datasets to further enhance 3D hand pose estimation performance. It could also be fine-tuned to adapt to various downstream tasks, such as gesture recognition.

References

[1] Ross Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015. 1

[2] Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. Interacting attention graph for single image two-hand reconstruction. In *CVPR*, pages 2761–2770, 2022. 1

[3] Pengfei Ren, Haifeng Sun, Qi Qi, Jingyu Wang, and Weiting Huang. Srn: Stacked regression network for real-time 3d hand pose estimation. In *BMVC*, page 112, 2019. 1

[4] Pengfei Ren, Chao Wen, Xiaozheng Zheng, Zhou Xue, Haifeng Sun, Qi Qi, Jingyu Wang, and Jianxin Liao. Decoupled iterative refinement framework for interacting hands reconstruction from a single rgb image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1

[5] Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. Smoothnet: A plug-and-play network for refining human poses in videos. In *European Conference on Computer Vision*. Springer, 2022. 1

[6] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1