Deformer: Dynamic Fusion Transformer for Robust Hand Pose Estimation Supplementary Material

Qichen Fu[1] Xingyu Liu[1] Ran Xu[2] Juan Carlos Niebles[2] Kris M. Kitani[1]
[1] Carnegie Mellon University [2] Salesforce Research

A. Overview

This document provides additional implementation and experimental details, as well as qualitative results and analysis. We illustrate the motion discriminator architecture in Appendix B. Then we show additional qualitative results in Appendix C. Finally, we discuss the limitations of our approach in Appendix D.

B. Motion Discriminator

As described in *Sec. 3.2* of the main paper, the motion discriminator $\mathcal{D}$ is learned to supervise the output sequence of the SpatioTemporal Transformer. The architecture of $\mathcal{D}$ is depicted in Fig. 1. Given a hand pose sequence, the motion discriminator first uses a shared linear layer to map the hand pose (represented by MANO parameters) of each timestamp to a high-dimensional feature vector. These frame-wise hand pose features are then input to a two-layer bidirectional Grated Recurrent Unit (GRU), which outputs a new sequence of features that incorporate the information of previous and future hand poses. Instead of using an average or max pooling, we use the self-attention [4] mechanism to adaptively choose the most important features in the sequence and summarize them into a single hidden feature. Finally, a linear layer is learned to predict a value in $[0, 1]$ indicating if the input hand pose sequence is realistic or fake.

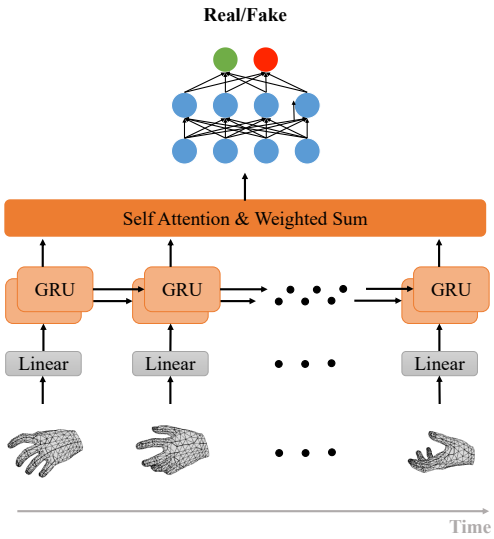

Figure 1: The architecture of motion discriminator $\mathcal{D}$. Given a hand pose sequence represented by MANO parameters, $\mathcal{D}$ uses two GRU layers with self-attention to classify if the input is a realistic or fake sequence.

C. Qualitative Results

In this PDF, we present an exemplary set of our hand pose estimation sequence on the DexYCB [1] dataset in Fig. 3 and on the HO3D[2] dataset in Fig. 4. For the DexYCB dataset, each example sequence is visualized in two rows, where the first row shows the predicted hand mesh aligned with the image and the second row shows the error map. For the HO3D dataset, as its test set ground truth is not public, the error map could not be computed and we only show the predicted mesh. The qualitative results and Fig. 2 demonstrated that our method is more robust and can estimate accurate hand pose under scenes where the hand is visually occluded or blurred. Most errors are located near the hand parts which are invisible (along all frames in the sequence). In certain cases, another factor that constrains the network is that the mesh generated from MANO parameters (with 778 vertices) is not fine enough to represent the detail of hands.

D. Limitations

First, our method is supervised, which relies on an annotated video dataset to train. Second, the MANO hand model we used has a limited number (778) of vertices. We found in certain cases it fails to capture the details of various hands. Finally, as we used the self-attention mechanism to capture the temporal information, the memory of the proposed method is quadratically proportional to the length of the sequence. This fact limits our approach to efficiently capture the long-term hand motion over the whole video.

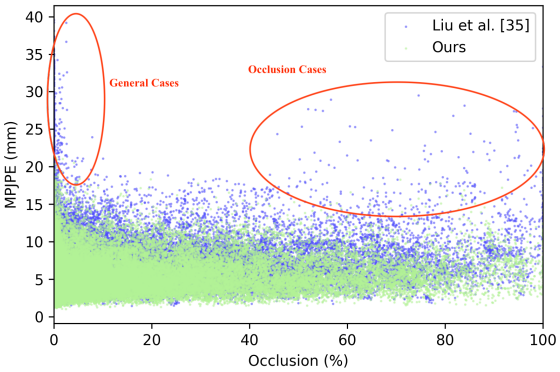

Figure 2: The scatter plot of MPJPE in mm vs. different hand-object occlusion levels on DexYCB test data samples. Compared to [3], our method significantly reduces the hand pose estimation error, especially in occlusion cases.

References

[1] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 1, 3

[2] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3196–3206, 2020. 1, 4

[3] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14687–14697, 2021. 2

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

Figure 3: Qualitative results on the DexYCB[1] dataset. For every sequence sample, the first row shows the predicted hand mesh and the second row shows the error map.

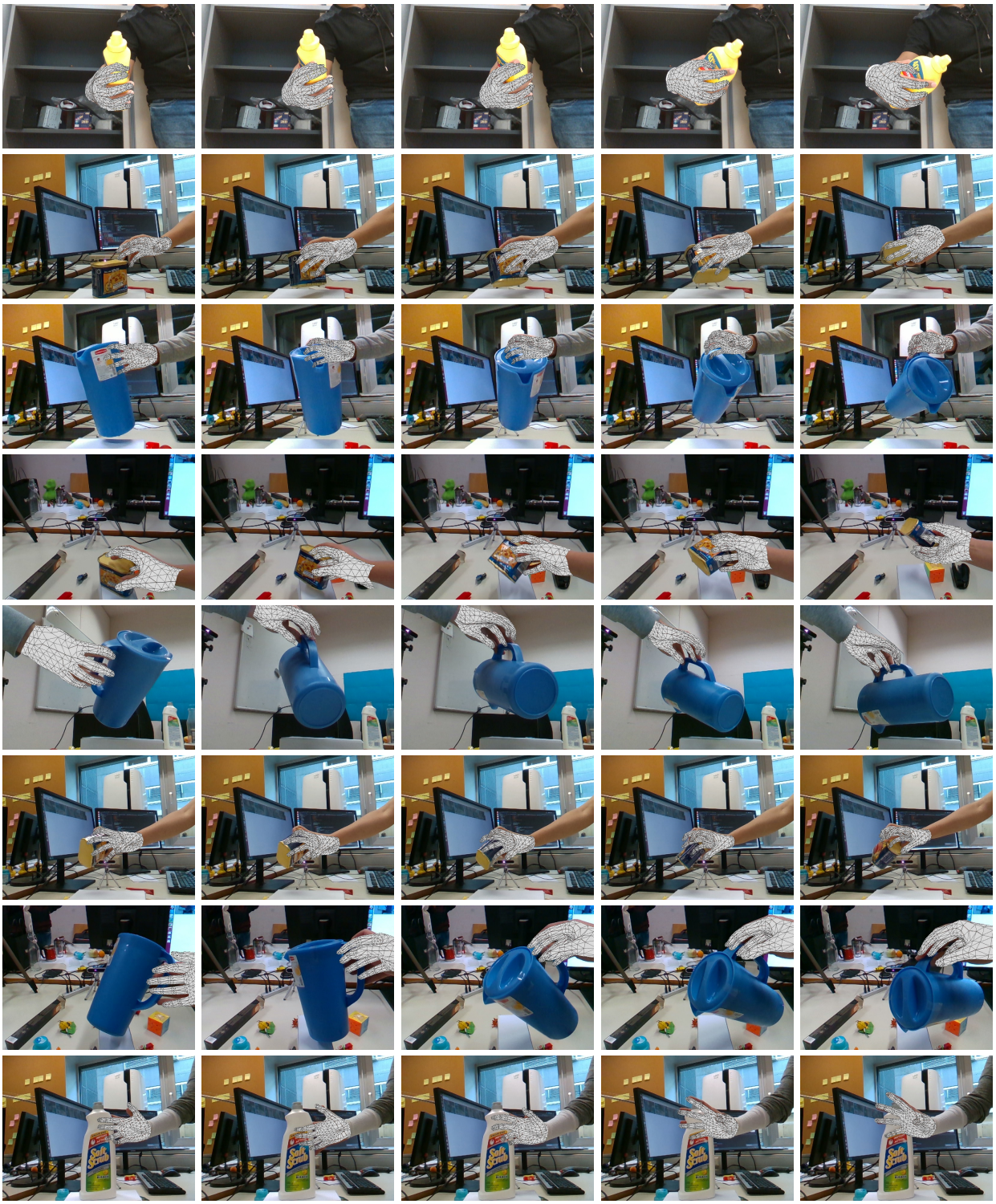

Figure 4: Qualitative results on the HO3D[2] dataset. As the HO3D test set annotation is not public, we only show the predicted hand meshes for every exemplary sequence in each row.