Supplementary: Hierarchical Temporal Transformer for 3D Hand Pose Estimation and Action Recognition from Egocentric RGB Videos

We attach a supplementary video `supp.mp4` to support our discussion for hand pose estimation and action recognition in the main text, which could be played by most players. The video contains two parts with qualitative analysis for videos from the FPHA [1] and H2O [2] dataset.

The first part focuses on hand pose, where we show our qualitative results for hand pose estimation on both datasets. In this part we also verify our design choice of leveraging the short-term temporal cue for pose module, by comparing ours with alternative setups without the temporal cue or with a long-term temporal cue. For these three setups, we visualize the estimated hand pose; for setups using the temporal cue, we also visualize the attention weights among frames for the final layer of the pose block.

The second part focuses on action, using videos of different actions from the H2O [2] dataset. For our setup, we visualize the attention weights of the final layer of the action block, from the action token to the other per-frame tokens, to inspect the pattern of attention distribution for different actions.

References

[1] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 409–419, 2018. 1

[2] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10138–10148, 2021. 1