Interacting Attention Graph for Single Image Two-Hand Reconstruction (Supplementary Document)

Mengcheng Li[1], Liang An[1], Hongwen Zhang[1], Lianpeng Wu[2], Feng Chen[1], Tao Yu[1], Yebin Liu[1]
[1]Tsinghua University [2]Hisense Inc.

1. Video Results

Our supplementary video (refer to our project page) is composed of two parts: live demos and comparisons.

1.1. Live Demos

To fully demonstrate the supreme performance of our IntagHand model, we build an online real-time two-hand reconstruction system (see Supplementary Video). It is well known that the illumination and the background of the training images from InterHand2.6M [2] are ambient and dark. Therefore, neural networks trained on the InterHand2.6M [2] dataset work unstably in real world scenarios.

To handle this issue, we simply build a synthetic dataset to assist training. Specifically, we render synthetic two-hand images using the hand poses and viewpoints provided by InterHand2.6M together with manually designed hand textures and random background images from ImageNet [1], see Fig. 1. Moreover, we add additional per-pixel gaussian noise to simulate the inherent noise of commercial cameras. Finally, we fine-tune our IntagHand model for 50 epochs with 10^{-5} learning rate, where each mini-bacth contains half InterHand2.6M samples and half synthetic samples. Note that, we *only* fine-tune our model for live demo, and all the qualitative and quantitative results in the main paper *do not* incorporate fine-tuning.

We simply utilize a common USB camera to obtain live video stream. Our online system runs at 20fps on single NVIDIA RTX 3090 GPU where the image capture process takes 2ms, the network inference takes 33ms and the rendering process takes 15ms. Note that, without rendering, our model runs at 30fps.

Compared with previous tracking based live demos [3], our results demonstrate closer interaction and more flexible movement benefiting from the strong representation power of our IntagHand model.

1.2. Additional Comparisons

The existing state-of-the-art two-hand reconstruction method is Zhang *et al*. [4]. To compare with it, we demonstrate free viewpoint renderings in the supplementary video on the InterHand2.6M [2] dataset and the RGB2Hands [3] dataset. Furthermore, we compare with [4] on the offline real-life video to show the superior stability of our method. For fair comparison, we only train our model on the InterHand2.6M [2] dataset during all the comparisons.

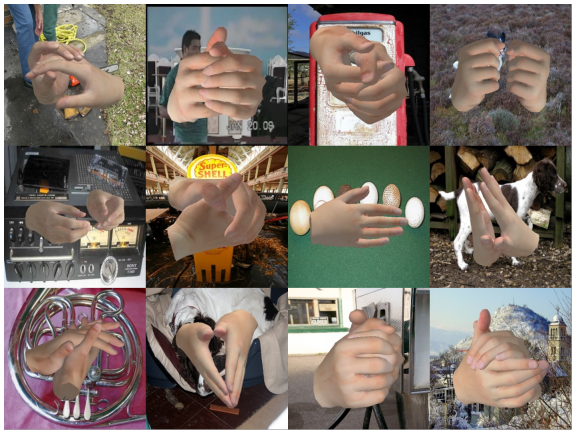

Figure 1. Samples from the synthetic dataset.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[2] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline for 3D interacting hand pose estimation from a single rgb image. In *ECCV*, 2020. 1

[3] Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A. Otaduy, Dan Casas, and Christian Theobalt. RGB2Hands: real-time tracking of 3D hand interactions from monocular rgb video. In *SIGGRAPH Asia*, 2020. 1

[4] Baowen Zhang, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, and Hongan Wang. Interacting two-hand 3D pose and shape reconstruction from single color image. In *ICCV*, 2021. 1