# Benchmarks and Challenges in Pose Estimation for Egocentric Hand Interactions with Objects

— Supplementary Material —

## 1 Additional results of AssemblyHands

**Qualitative results:** Figure 1 shows the qualitative results of submitted methods and failure patterns indicated by the red circles. The left hand in the first row grabs the object where the left thumb finger is only visible. While Base fails to infer the plausible pose, JHands enables estimation in such heavy hand-object occlusions compared to the GT. However, the methods PICO-AI and FRDC incorrectly predict the location of the left thumb finger and Phi-AI's prediction of the left index and middle fingers is also erroneous. The second row is the case where two hands and an object are closely interacting, particularly the left thumb finger presents near the right hand. The methods Base, FRDC, and Phi-AI fail to localize the left thumb finger. The third and fourth rows indicate hand images presented near the image edges. The methods Base, PICO-AI, and Phi-AI are prone to produce implausible predictions, including noise and stretched poses due to the distortion effect discussed in Section 5.2 "**Bias of hand position in an image**." The method JHands with distortion correction successfully addresses these edge images.

**Per-view analysis:** Figure 2 shows the detailed statistics and performance of per-view predictions, related to the analysis in Section 5.2 "**Effect of multi-view fusion**." Considering per-sequence results, we find the sample availability (blue bars) and performance (green bars) from cam1 and cam2 vary among different users. In contrast, the number of samples and performance of cam3 and cam4 are mostly stable. This study further necessitates the sample selection and multi-view fusion adaptively for each sequence (user).

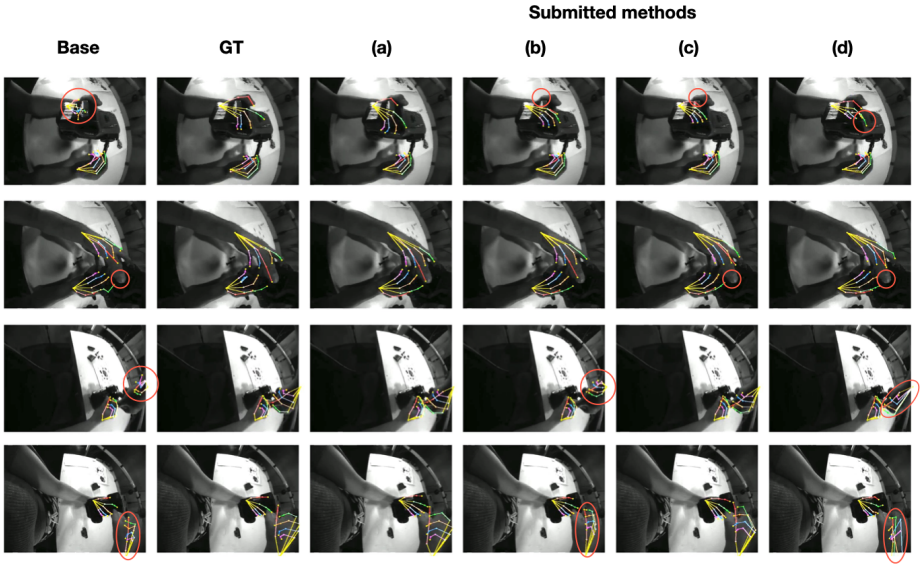


**Fig. 1: Qualitative results of submitted methods in AssemblyHands.** The columns correspond to the results of Base, ground-truth (GT), submitted methods, namely (a) JHands, (b) PICO-AI, (c) FRDC, and (d) Phi-AI. The red circles indicate where failures occur.

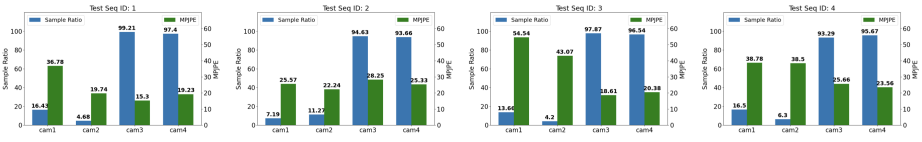


**Fig. 2: Additional results of multi-view fusion in AssemblyHands**. We analyze the availability of samples and performance per camera view. The two lowest cameras (cam3, cam4) out of the four cameras allow us to capture hands most of the time (>93 % of samples). In contrast, the images from cam1 and cam2 are fewer and the error varies in different sequences.