CrossHOI: Learning Cross-View Representations for Monocular 3D Human-Object Interaction Reconstruction

Supplementary Material

1. Efficiency

Our method improves substantially over the CONTHO baseline with minor overhead, and is significantly more efficient than HOI-TG (27% fewer parameters and higher FPS) while achieving better overall performance, indicating a favorable efficiency-accuracy trade-off.

Method	Params(M)↓	Infer. Speed(FPS)↑	Acc.(Contact_p) ↑
CONTHO [28]	83	1.68	0.628
HOI-TG [36]	123	1.33	0.662
Ours	90	1.60	0.687

Table S1. Efficiency-accuracy comparison with SOTA methods.

2. Cross-Dataset Generalization of the Generator.

To evaluate cross-dataset generalization, we restrict the experiments to overlapping object categories between BEHAVE and InterCap (e.g., Chair, Table, Box). We conduct cross-dataset evaluation by training the cross-view generator on InterCap and testing on BEHAVE. Two variants are compared: Ours-B2B (BEHAVE→BEHAVE) and Ours-I2B (InterCap→BEHAVE). Although a slight performance drop is observed due to domain gap, the cross-view generator remains effective and consistently outperforms the baseline (Tab. R2), demonstrating transferable representations rather than dataset-specific priors.

Method	CD_human↓	CD_object↓	Contact_p ↑	Contact_r ↑
CONTHO [28]	5.18	8.87	0.615	0.481
Ours-B2B	4.36(15.8%)	7.78	0.684(6.9pp)	0.572
Ours-I2B	4.42(14.7%)	7.83	0.679(6.4pp)	0.568

Table S2. Cross-dataset evaluation of the generator.

3. Additional Qualitative Results

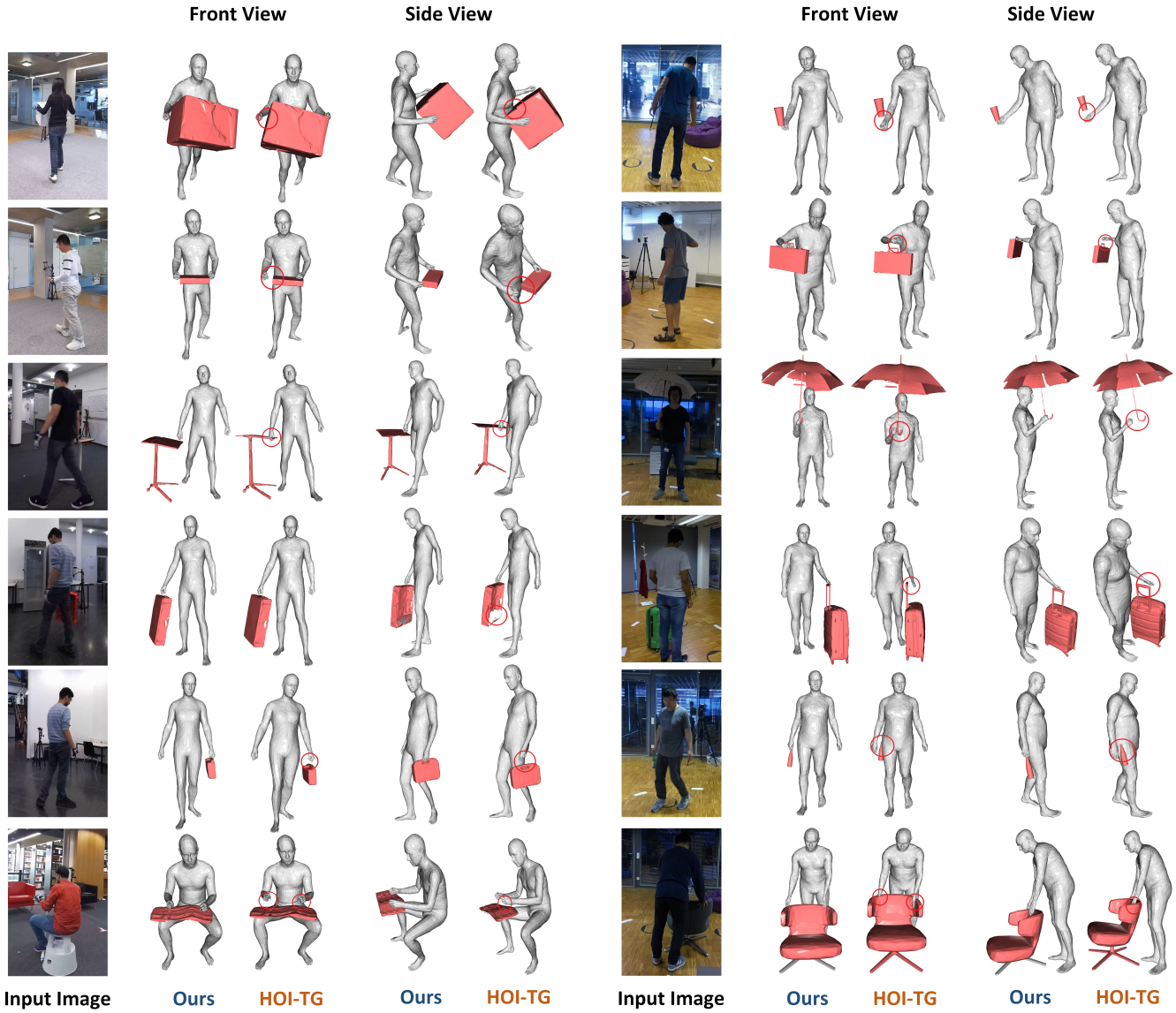

Figure 1. Qualitative comparison of 3D human-object reconstruction with HOI-TG [36] on the BEHAVE dataset (left) and the InterCap dataset (right). The incorrectly predicted contact regions are highlighted with red circles.