Joint Reconstruction of 3D Human and Object via Contact-Based Refinement Transformer

Supplementary Material

In this supplementary material, we present additional technical details and more experimental results that could not be included in the main manuscript due to the lack of pages. The contents are summarized below:

S1. Comparison with CHAIRS

Table S1 shows that our CONTHO mostly outperforms CHAIRS [16], a recently published 3D human and object reconstruction method. There are two core differences in the evaluation protocol between CHAIRS [16] and the other state-of-the-art methods [48, 55]. First, CHAIRS [16] only reports reconstruction scores in specific object classes, such as chairs, table, yoga ball, and suitcase. Second, its evaluation does not perform the Procrustes alignment before measuring the Chamfer distance. Consequently, for a fair comparison with CHAIRS, Table S1 reports the performance of our method following the CHAIRS evaluation process. Similarly to the other previous methods [48, 55], the CHAIRS is also an optimization-based method, which first reconstructs an object voxel and optimizes the 3D object mesh template on the reconstructed object voxel. The optimization process totally depends on the 3D object voxel reconstruction without considering image features during the optimization. Biased on the optimization target, CHAIRS often fails when the initial reconstruction provides an imperfect or noisy object voxel. In contrast, our learning-based approach produces outputs based on data-driven knowledge obtained during training rather than optimizing on imperfect targets. Furthermore, CHAIRS does not predict or exploit human-object contact for joint reconstruction. Unlike CHAIRS, our CONTHO utilizes human-object contact information as a key signal for reconstruction by estimating and exploiting human-object contact maps for 3D human and object joint reconstruction.

S2. Comparison of running time

Table S2 shows that our CONTHO takes the shortest computational time compared to previous 3D human and object reconstruction methods. The running time is measured in the same environment with the Intel Xeon Gold 6248R CPU and RTX 2080 Ti GPU. For all methods, we exclude the pre-processing stage of acquiring human and object silhouettes. PHOSA [55] and CHORE [48] demand extremely long times, as their optimization processes iteratively fit 3D meshes with more than 100 iterations. On the other hand, our CONTHO takes much less time, requiring only a single feed-forward for the inference under the learning-based approach. Thus, our CONTHO has a significant advantage in running time compared to previous methods [48, 55].

Methods	Chair		Table		Yogaball		Suitcase	
	$CD_{human}\downarrow$	$CD_{object}\downarrow$	$CD_{human}\downarrow$	$CD_{object}\downarrow$	$CD_{human}\downarrow$	$CD_{object}\downarrow$	$CD_{human}\downarrow$	$CD_{object}\downarrow$
CHAIRS [16]	13.77	12.73	11.53	15.22	10.82	**9.88**	9.53	15.84
CONTHO (Ours)	**5.94**	**9.91**	**6.57**	**9.30**	**6.42**	10.67	**4.81**	**8.36**

Table S1. **Quantitative comparison of 3D human and object reconstruction with CHAIRS [16].**

PHOSA [55]	CHORE [48]	**CONTHO (Ours)**
165.30	312.20	**0.077**

Table S2. **Running time comparison between different methods.** The unit is seconds per frame.

S3. Evaluation on object contact map

Compared to the previous human-object contact estimation methods [14, 43], which only estimate a human contact map, our CONTHO additionally estimates an object contact map. We evaluate CONTHO in object contact maps by using the same evaluation metrics of human contact maps (*i.e.,* $\text{Contact}_{\text{p}}^{\text{est}}$ and $\text{Contact}_{\text{r}}^{\text{est}}$), described in Section 5. For the BEHAVE dataset [2], the precision and recall scores of object contact estimation are 0.530 and 0.217, respectively. For the InterCap dataset [15], the precision and recall scores are 0.627 and 0.323, respectively. Although our qualitative results (Figure 5) show plausible estimations about object contact maps, the scores of object contact maps are relatively lower than those of human contact maps. This is due to the lack of considerations for 3D object symmetry from these evaluation metrics (i.e., $\text{Contact}_{\text{p}}^{\text{est}}$ and $\text{Contact}_{\text{r}}^{\text{est}}$). For example, when touching a ball with a sphere shape, any part of the ball can be the correct contact region. However, the evaluation metrics are vertex-

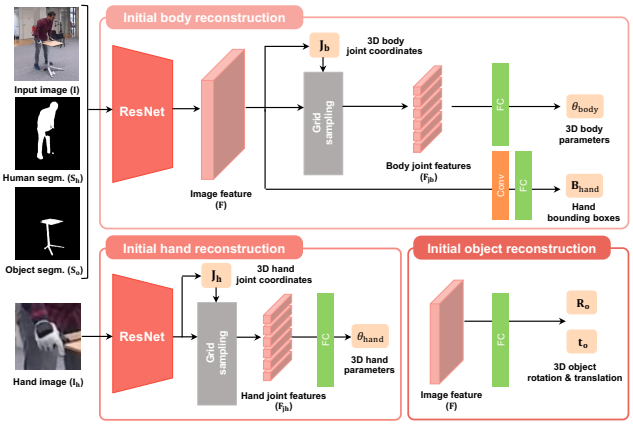


Figure S1. **The detailed architecture of the initial reconstruction stage in CONTHO.**

to-vertex comparisons between predictions and GTs, where the vertex-to-vertex pairs are fixed. Thus, the evaluation metrics cannot consider the symmetry of 3D objects, which results in an underestimation of the performance of object contact maps.

S4. Details of initial reconstruction

S4.1. Architectural design

Our initial reconstruction follows Hand4Whole [27] with a few modifications. Figure S1 specifically describes the backbone network in Figure 3, which consists of two separate ResNets for reconstructing 1) 3D human body along with 3D object and 2) 3D human hands, respectively.

Initial body reconstruction. To predict 3D body parameters θ_{body}, we first extract an image feature $\mathbf{F}$ with a backbone network (*i.e.*, ResNet-50 [13]). Then, we apply a 1-by-1 convolutional layer, followed by soft-argmax operation [41], to extract 3D human body joints $\mathbf{J}_{\text{b}}$ from the image feature $\mathbf{F}$ in a differentiable way. The 3D human body joints $\mathbf{J}_{\text{b}}$ are utilized to extract body joint features $\mathbf{F}_{\text{jb}}$ by conducting grid sampling onto (x, y) positions of 3D body joint coordinates $\mathbf{J}_{\text{b}}$ of the image feature $\mathbf{F}$. Based on the extracted body joint features $\mathbf{F}_{\text{jb}}$, we predict 3D body parameters θ_{body} with a fully-connected layer after flattening the 3D joint features $\mathbf{F}_{\text{jb}}$. The image feature $\mathbf{F}$ is used further for the three subsequent processes. First, the 2D hand bounding boxes $\mathbf{B}_{\text{hand}}$ are obtained by forwarding the image feature to a combination of convolutional layers and fully-connected (FC) layers, where the 2D hand bounding boxes are passed to initial hand reconstruction. Second, the image feature is also used in initial object reconstruction. Third, we obtain the 3D vertex features, which are the core components of our CONTHO (Section 3).

Initial hand reconstruction. Using the estimated 2D bounding boxes $\mathbf{B}_{\text{hand}}$, we crop the input image $\mathbf{I}$ to obtain hand image $\mathbf{I}_{\text{h}}$, which serves as an input for initial hand reconstruction. Similarly to the initial body reconstruction, our initial hand reconstruction acquires 3D hand parameters θ_{hand} with a similar process of extracting 3D hand joint coordinates $\mathbf{J}_{\text{h}}$ and utilizing the 3D joint coordinates $\mathbf{J}_{\text{h}}$ to obtain 3D joint features of hand $\mathbf{F}_{\text{jh}}$, which will be passed to a fully-connected layer. The 3D body parameters θ_{body} from initial body reconstruction and 3D hand parameters θ_{hand} from initial hand reconstruction are passed to SMPL+H model to get initial human mesh $\mathbf{M}_{\text{h}}$.

Initial object reconstruction. To predict 3D object parameters ($\mathbf{R}_{\text{o}}$ and $\mathbf{t}_{\text{o}}$), we process image feature $\mathbf{F}$ from the initial body reconstruction and predict 3D object rotation $\mathbf{R}_{\text{o}}$ and translation $\mathbf{t}_{\text{o}}$ with a fully-connected layer. In the end, the initial reconstruction stage obtains the initial object mesh $\mathbf{M}_{\text{o}}$ along with the initial human mesh $\mathbf{M}_{\text{h}}$, which are passed to the next stage of CONTHO, the 3D-guided contact estimation.

S4.2. Loss function design

In Section 3, the L_{init} is designed to supervise the output of the initial reconstruction stage, with a few modifications of Hand4Whole [27]'s loss function. The L_{init} is defined as

$$L_{\text{init}} = L_{\text{param}} + L_{\text{coord}} + L_{\text{hbox}}. \tag{4}$$

Parameter loss (L_{param}). We minimize L1 loss between the predicted and GT parameters for 3D human body mesh θ_{body}, 3D human hand mesh θ_{hand}, and 3D object mesh ($\mathbf{R}_{\text{o}}$, and $\mathbf{t}_{\text{o}}$).

Coordinate loss (L_{coord}). We utilize L1 loss between the predicted and GT human joint coordinates. Specifically, the human joint coordinates consist of three types: 1) the extracted 3D body joint coordinates $\mathbf{J}_{\text{b}}$ and 3D hand joint coordinates $\mathbf{J}_{\text{h}}$, 2) 3D joint coordinates regressed from 3D human mesh $\mathbf{M}_{\text{h}}$ with pre-defined regression matrix of SMPL+H, and 3) 2D joint coordinates, obtained by projecting the 3D joint coordinates from 3D human mesh $\mathbf{M}_{\text{h}}$ to image space.

Hand bounding boxes loss (L_{hbox}). We implement L1 loss between the predicted and GT bounding boxes of hand. Specifically, the L1 distance for the center and scale of the bounding boxes are computed following Hand4Whole [27].

S5. More examples of undesired correlation

In Figure S3, we provide more examples of the undesired correlation between human and object. In these examples, we further show that the Transformer baseline with naive use of contact, learns undesired correlation between human and object. As shown in Figure S3, the Transformer baseline outputs objects to face toward a certain human body part (*e.g.,* the back of a chair and a monitor display facing

toward human face) or to be positioned in frequent interacting poses (*e.g.,* a keyboard in typing position and chair in regular sitting position). Unlike the baseline method, our CONTHO does not suffer from the undesired correlation with the novel contact-based refinement Transformer.

In Figure S4, we additionally show the undesired correlation between human and object with sensitivity tests. In addition to the main manuscript, which only showed sensitivity tests of object errors, we show the results of our sensitivity test of human errors for the Transformer baseline and our CRFormer with more results for sensitivity tests of object errors on BEHAVE [2] and InterCap [15]. The results from sensitivity tests show that the Transformer baseline contains high sensitivity in the object regions for human errors and in human regions for object errors, indicating a high correlation between human and object. Unlike the Transformer baseline, our CRFormer gives reasonable sensitivity of human regions for human errors and object regions for object errors on both datasets.

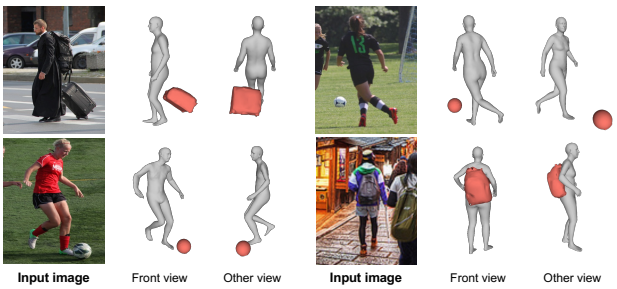
S6. More qualitative results

We provide more qualitative comparisons of human-object contact estimation and 3D human and object reconstruction under two experimental protocols: 1) training & evaluating all methods on BEHAVE [2] and 2) training & evaluating all methods on InterCap [15]. Figure S5 and Figure S6 show that our CONTHO vastly outperforms previous contact estimation methods (BSTRO [14] and DECO [43]) on BEHAVE [2] and InterCap [15]. Figure S7 and Figure S8 show that our CONTHO produces much accurate reconstruction results than previous reconstruction methods (PHOSA [55] and CHORE [48]) on BEHAVE [2] and InterCap [15].

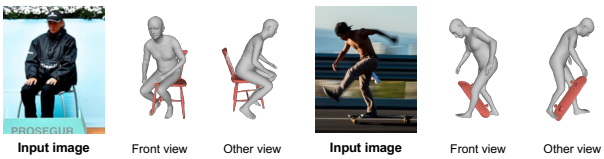
S7. Limitations and future works

Generalization to in-the-wild images. Figure S2 shows qualitative reconstruction results of CONTHO on in-the-wild images of MSCOCO [25] and MPII [1]. To obtain the results, we used the network trained on BEHAVE [2], without any fine-tuning. As shown in Figure S2 (b), our CONTHO fails on some in-the-wild images. This is mainly due to domain gap between training datasets [2] and in-the-wild images. As the training datasets are acquired in restricted environments, the datasets contain much less diverse image appearances compared to in-the-wild datasets. Due to such a domain gap problem, generalization on in-the-wild images is one of the challenges to be solved.

Diversity of object shape. Our CONTHO covers a limited number of 3D object categories, included in the training datasets. However, real-world objects are more diverse than the restrained categories of training objects. Collecting more 3D object data in the real-world and learning interaction with the objects are crucial future research directions.



(a) Reconstruction results on in-the-wild images



(b) Failure cases on in-the-wild images

Figure S2. **Qualitative results of CONTHO on in-the-wild images.**

Video as input. Our CONTHO aims to jointly reconstruct 3D human and object from a single image. The recent emergence of a video-based method for 3D human and object reconstruction [49] suggests that reconstructing human and object from a video could be a promising direction. Despite its strong performance, it utilizes future frames for the results of a certain frame, which we call an offline approach. Instead, we think an online approach, which does not assume the availability of the future frame, could be closer to real-world applications, and we aim to extend our CONTHO for such an online approach.

License of the Used Assets

- BEHAVE dataset [2] is available for the sole purpose of performing non-commercial scientific research.
- InterCap dataset [15] is released for non-commercial scientific research purposes.
- BSTRO codes [14] are released under the MPI license.
- DECO codes [43] are released for non-commercial scientific research purposes.
- PHOSA codes [55] are released under CC BY-NC 4.0.
- CHORE codes [48] is available for the sole purpose of performing non-commercial scientific research.

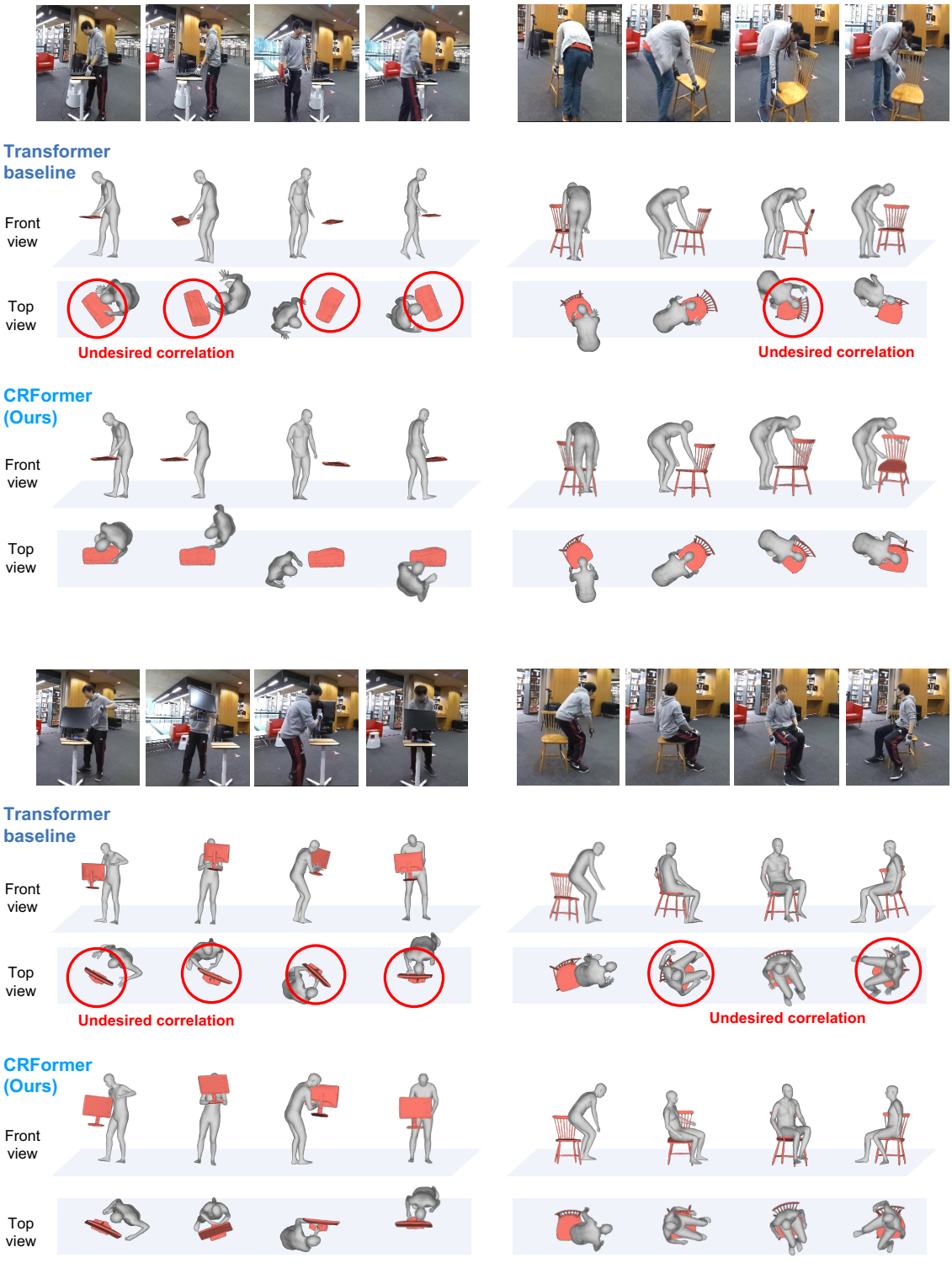


Figure S3. **Undesired correlation between human and object on BEHAVE [2].**

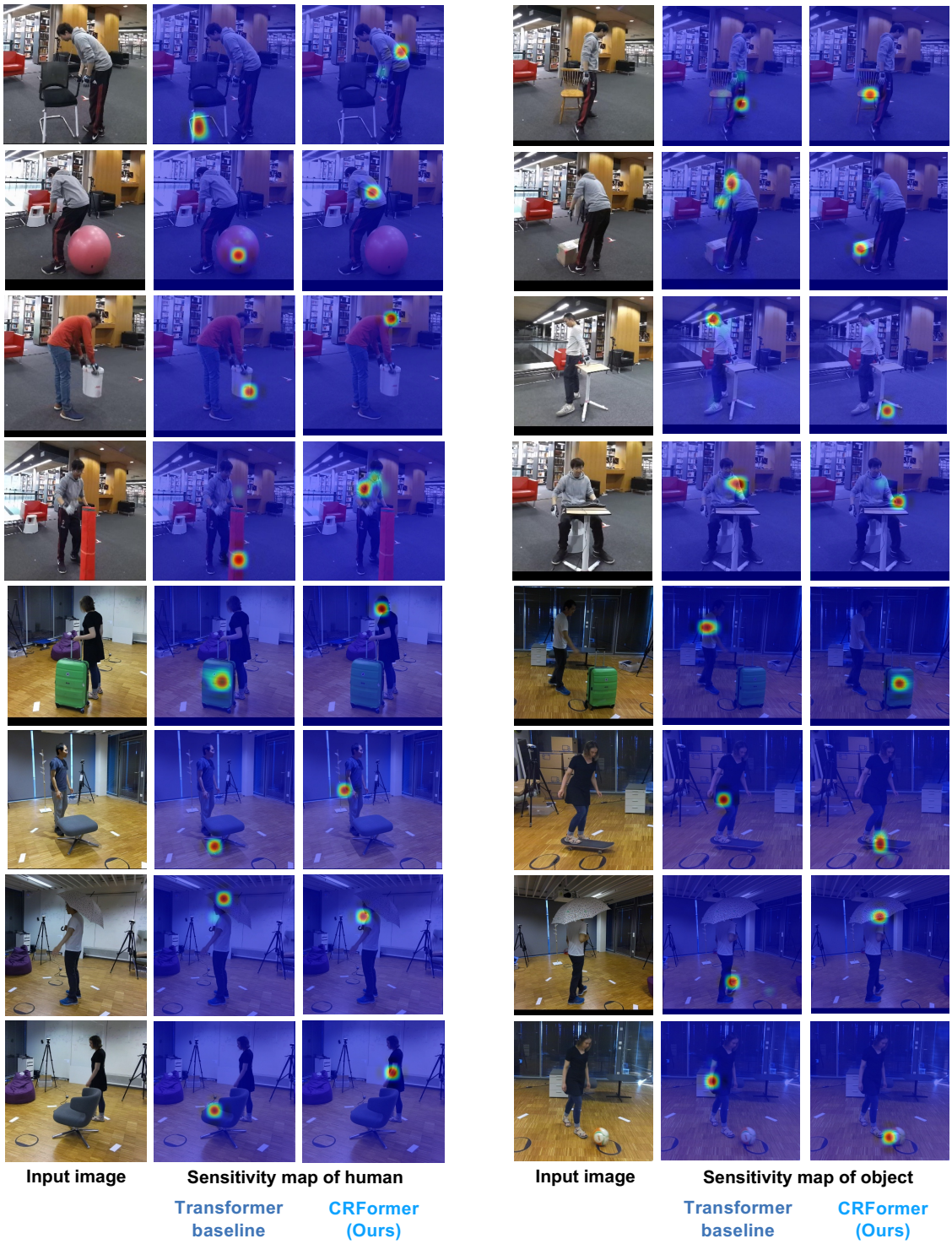


Figure S4. **Sensitivity tests of human errors (left) and object errors (right) on BEHAVE [2] and InterCap [15].**

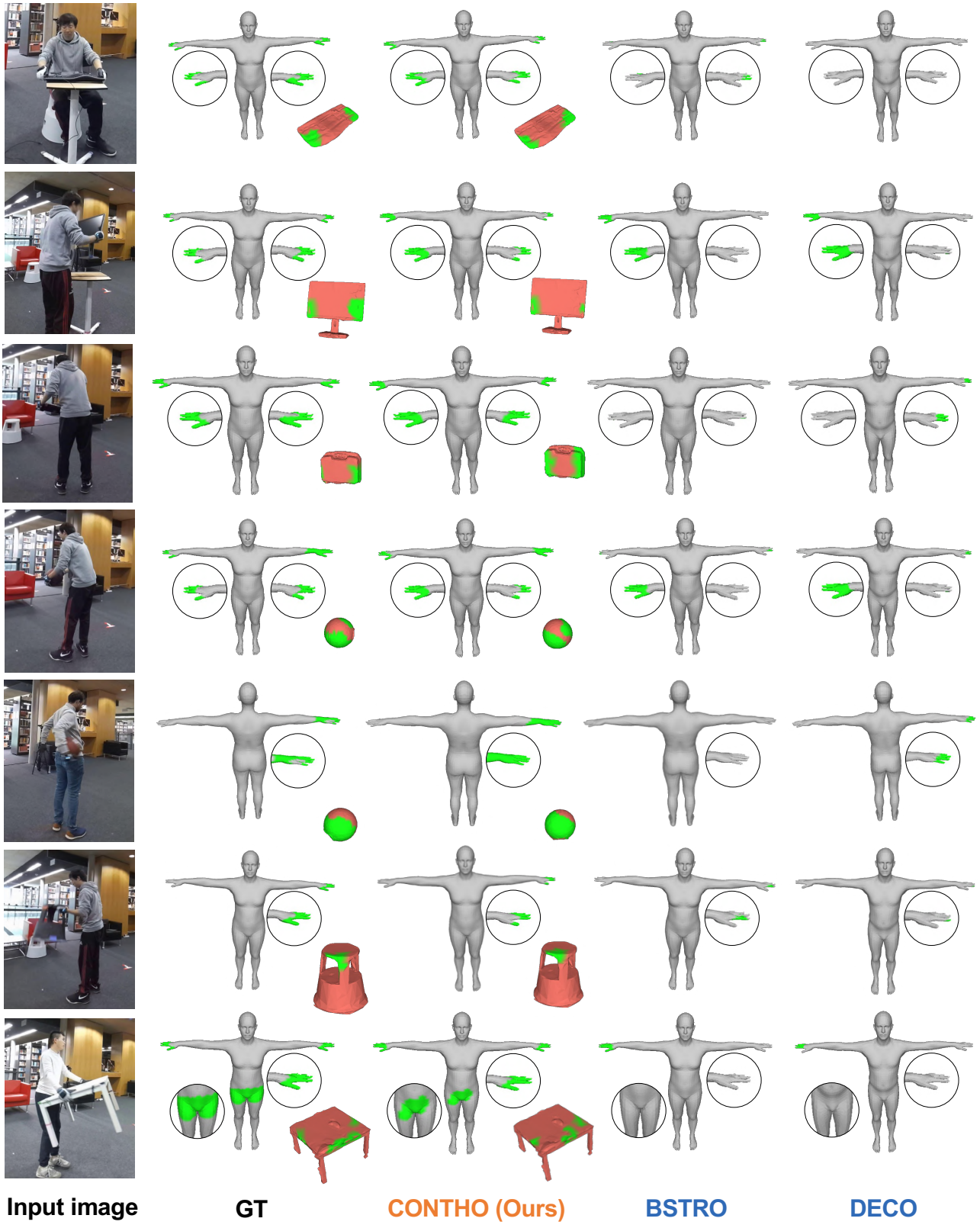


Figure S5. **Qualitative comparison of human-object contact estimation with BSTRO [14] and DECO [43], on BEHAVE [2].** The green color indicates the contacting regions.

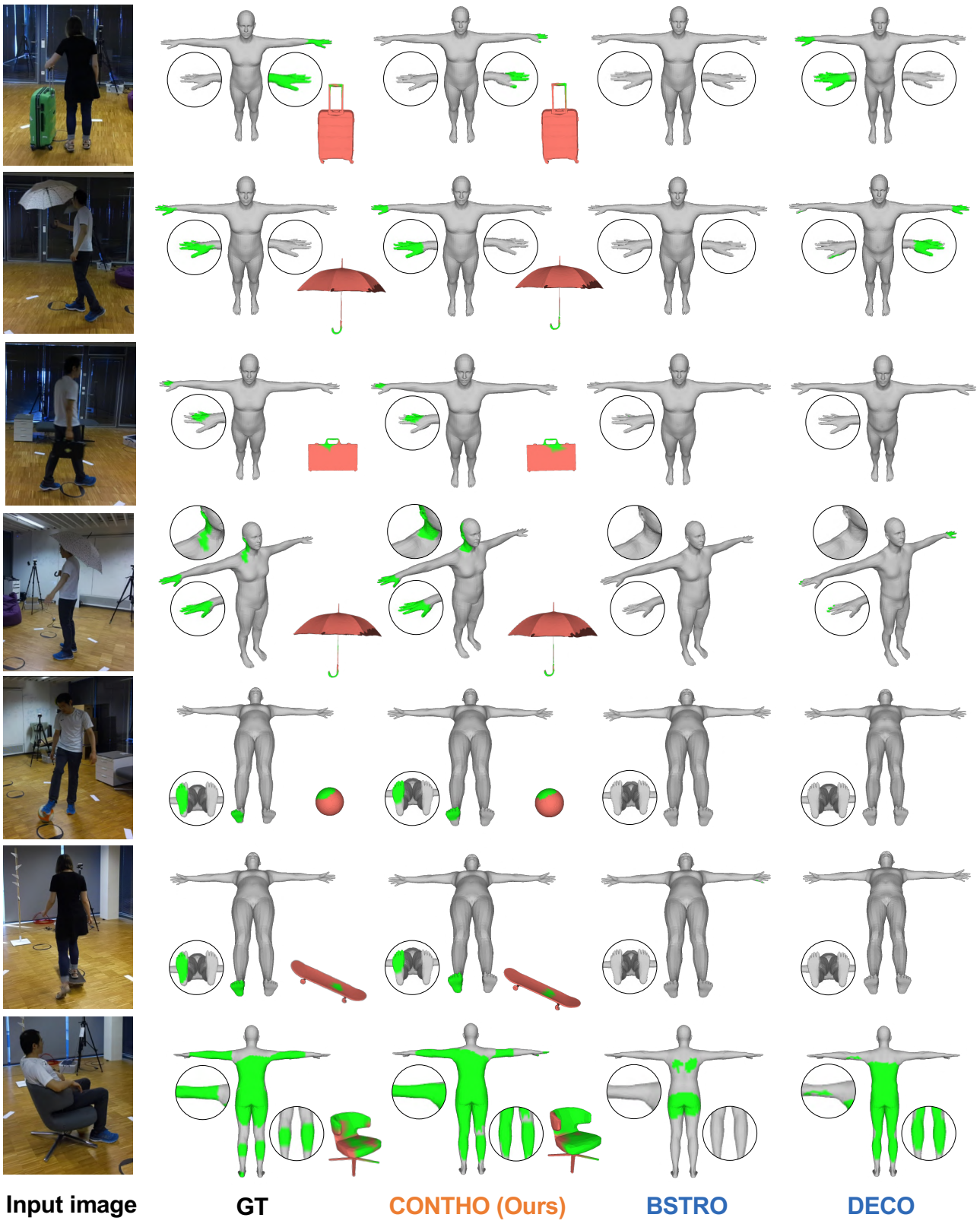


Figure S6. **Qualitative comparison of human-object contact estimation with BSTRO [14] and DECO [43], on InterCap [15].** The green color indicates the contacting regions.

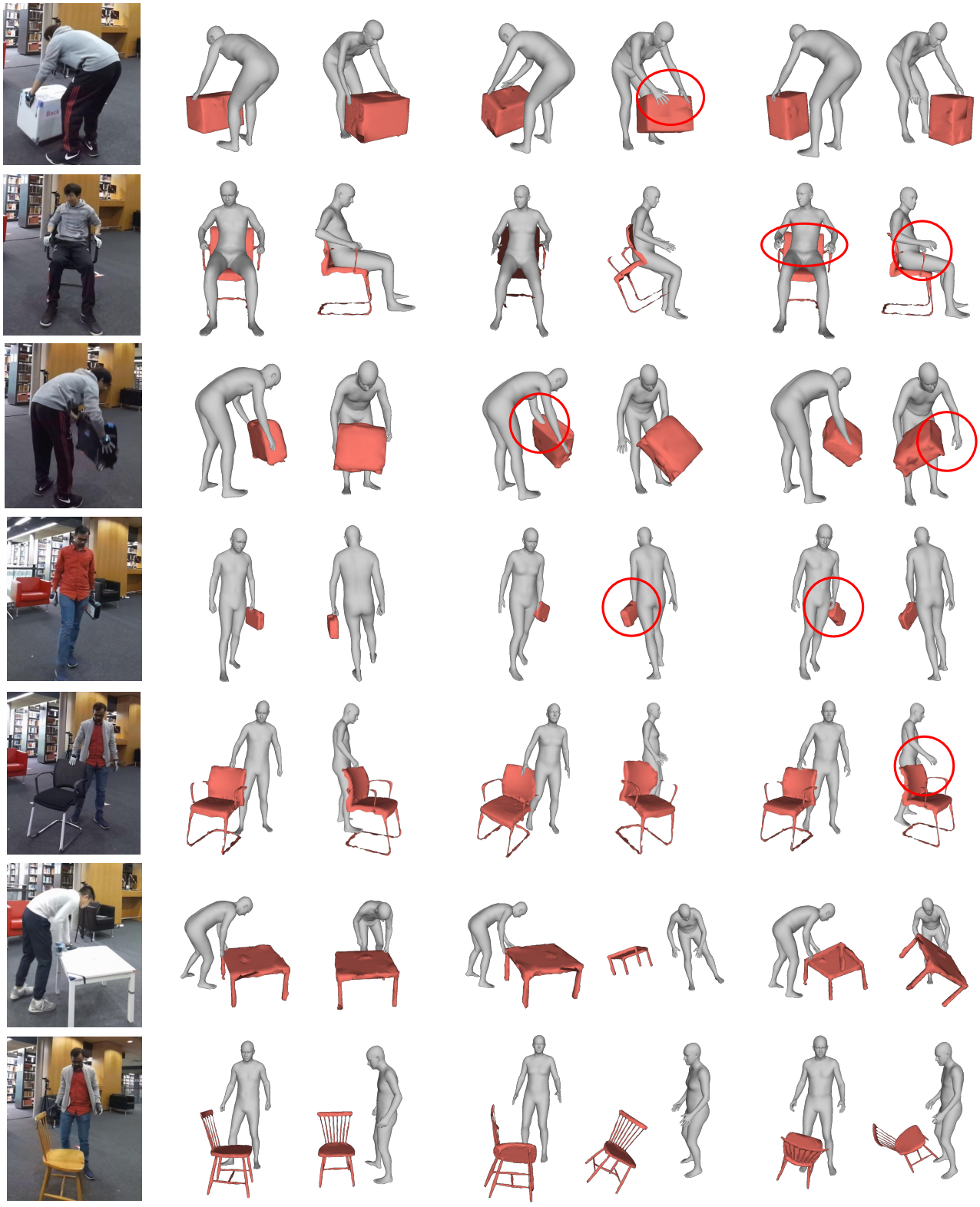


Figure S7. **Qualitative comparison of 3D human and object reconstruction with PHOSA [55] and CHORE [48], on BEHAVE [2].** We highlight their representative failure cases with red circles.

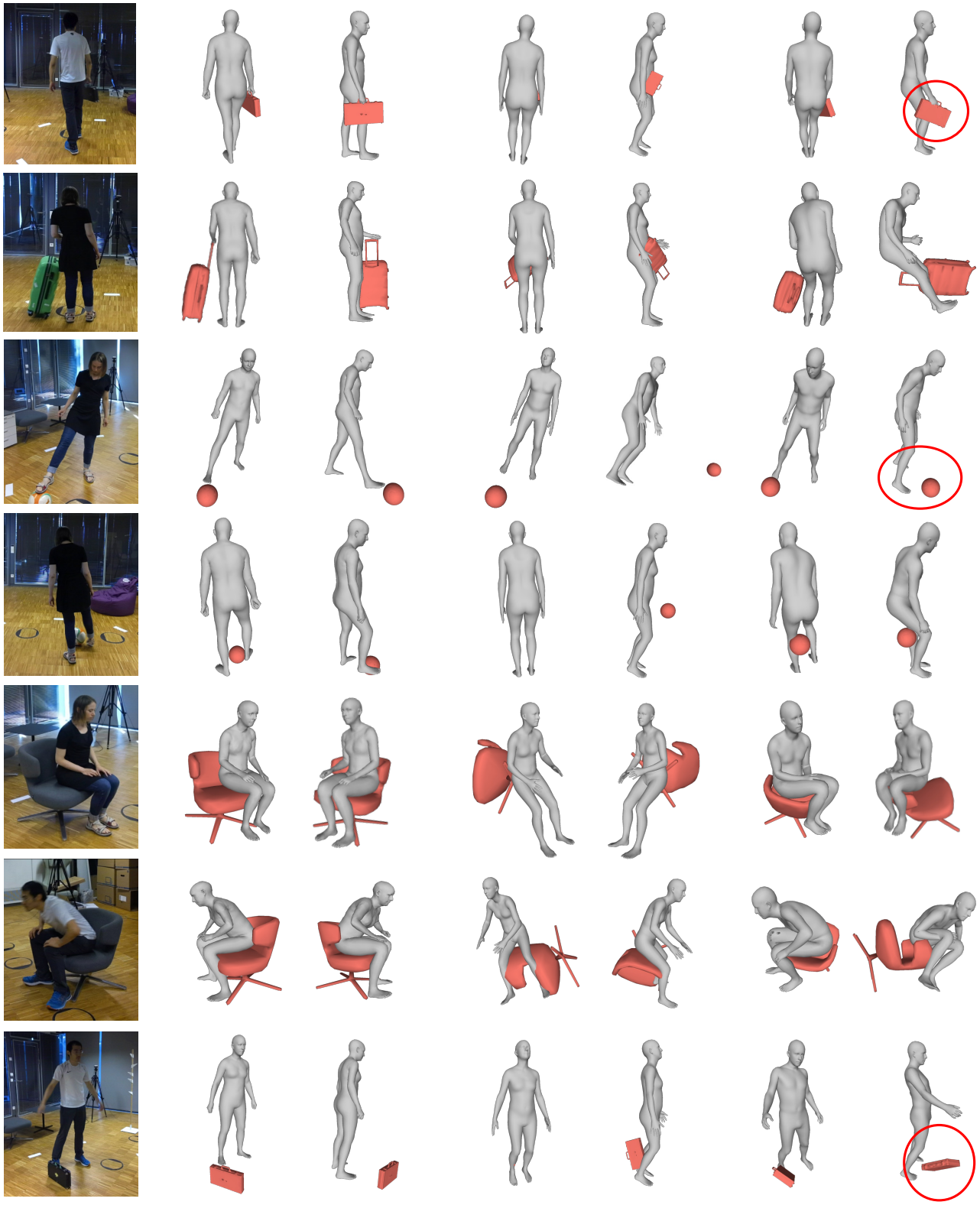


Figure S8. **Qualitative comparison of 3D human and object reconstruction with PHOSA [55] and CHORE [48], on InterCap [15].** We highlight their representative failure cases with red circles.