CHORE: *Supplementary material*

Xianghui Xie[1], Bharat Lal Bhatnagar[1,2], and Gerard Pons-Moll[1,2]

[1] Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
[2] University of Tübingen, Germany
{xxie, bbhatnag}@mpi-inf.mpg.de, gerard.pons-moll@uni-tuebingen.de

In this supplementary, we discuss in more detail about our implementations and show more qualitative examples to highlight the importance of explicit contact modelling and object pose prediction. We show more qualitative results on BEHAVE [**?**], NTU-RGBD [4] as well as images in the wild [3] and discuss limitations of our method in the end.

1 Implementation details

1.1 Networks and hyper-parameters

Network architecture. The input to our network is an RGB image stacked with human and object masks: $512 \times 512 \times 5$. The encoder f^{enc} consists of five stacked hourglass modules, similar to PIFu [6]. All decoders have the same structure: three FC layers with ReLU activation and one output FC layer. Specifically, the input to each decoder is a feature vector of size $256 + 64 + 3 = 323$ and output shape is $2, 14, 9, 6$ for f^u, f^p, f^R, f^c respectively.

Training details. Our query points for implicit reconstruction are generated using multi-distribution sampling strategy used in IFNets [2], we also add random grid samples ($\frac{1}{16}$ of total samplings) inside a fixed volume, as suggested by [6]. We use depth-aware scaling for all meshes such that the SMPL mesh center is always at $z_0 = 2.2m$ and then generate training labels from scaled meshes. The network is trained with joint objective $L = \lambda_u(L_{u_h} + L_{u_o}) + \lambda_p L_p + \lambda_R L_R + \lambda_c L_c$ where we set $\lambda_u, \lambda_p, \lambda_R, \lambda_c$ as 1.0, 0.006, 500, 1000 in our experiments. The model is trained with Adam optimizer of learning rate 0.001. Our model is trained on a cluster server with 3 RTX8000 GPUs, each GPU has 48GB memory capacity. It takes around 30 hours to converge.

Details for joint optimization. The data term of our joint optimization objective defined in Eq. 6 is highly non-convex, hence we solve the problem in 3 stages, namely human reconstruction, object reconstruction and joint optimization. In human reconstruction stage, we optimize only the human parameters until convergence and in object reconstruction stage we optimize the object parameters until convergence. And finally in joint optimization stage, we fix the human parameters and optimize object parameters using the final objective until convergence. The loss weights $\lambda_h, \lambda_{p'}, \lambda_o, \lambda_{\text{occ}}, \lambda_{\text{reg}}, \lambda_c, \lambda_J, \lambda_r$ are set to 900, 0.0025, 8100, 9×10^{-6}, 900, 10^4, 0.25, and 1.

1.2 Adaptive cropping and resizing at test time

The original image from BEHAVE dataset [**?**] is $1536 \times 2048px$, which is too large for our network. We hence make a square crop of size 1200px around the human object bounding box center and resize it to 512px for the network. The original camera intrinsic is used for accurate pixel-aligned training. However, at test time the images have various resolution and *unknown* intrinsic with person at diverse range of depth. Since our network is trained with reduced depth-scale ambiguity, it is able to reconstruct surface centered at a fixed depth and reason scale (object size) from input pixels. We hence crop and resize the human-object patch such that the person in the resized patch appears as if they are at z_0 under the camera intrinsic we used during training. As people usually present diverse poses during human-object interaction, a fixed pixel size for human [7] is not accurate. Therefore, we use SMPL mesh estimated from FrankMocap [5] to compute the patch resizing factor.

More specifically, a depth offset of z_0 is added to the SMPL mesh from FrankMocap (centered at origin). We then compute the body keypoints in 3D and project them to image plane using the camera intrinsic at training time. For an image with different resolution, we first resize and pad it to $1536 \times 2048px$: images with larger width than height are resized to width at 2048px and pad along height direction. Similarly, images with larger height than width are resized to height at 1536px and pad along width direction. We detect the person body keypoints using Openpose [1] in the resized image and compute the height of the bounding box that encloses valid body keypoint detections, denoted as h_a. Let h_p denotes the height of the same set of body keypoints obtained from projecting 3D keypoints of FrankMocap SMPL mesh. The resizing factor is then computed as $s = \frac{h_p}{h_a}$. Intuitively this means if a person's keypoints are smaller/larger than the keypoints projected from FrankMocap mesh, the person is farther/closer than z_0, which requires a scale up/down of the patch so that the person *appears* as if it is at z_0, as desired. We apply this scaling factor to resize the square patch of person and object to 1200px and finally resize it to 512px for the image encoder.

2 Additional qualitative results for ablation studies

2.1 Importance of contacts

We present a key insight that explicitly modelling the contact is important to obtain accurate alignment between human and objects. We show more qualitative examples in Fig. 1. It can be clearly seen that our contact prediction can improve the accuracy and physical plausibility of the joint reconstruction.

2.2 Object pose prediction

In addition to the contacts, we also predict the object orientation and center to initialize and regularize the object fitting, which is important to obtain accurate

object pose and joint reconstruction, see Fig. 2. Without our orientation prediction, the fitting easily gets stuck in local minima, and without object center to initialize and regularize fitting, the object maybe fitted to an incorrect depth but aligned with input, which leads to inaccurate contact and pose in the end.

3 More qualitative examples

We show more comparisons with Weng et al. [8] and PHOSA [9] on the BEHAVE dataset in Fig. 3. It can be seen that baseline methods may give reasonable results in front view, but their error in 3D becomes obvious in side view. In contrast, our method produces coherent 3D reconstruction.

More qualitative comparisons on NTU-RGBD [4] can be found in Fig. 4 and more results from our method are shown in Fig. 5. Fig. 4 shows that our method produces more accurate 3D reconstruction than baselines and in Fig. 5 we show that our method generalizes well across diverse subjects, locations and camera viewpoints.

We also show significantly more qualitative comparisons with PHOSA on in the wild images from COCO [3] and internet images in Fig. 6. Note that images with boxes and yoga balls are from internet and all other images are from COCO. More results from our method on in the wild images are shown in Fig. 7 and Fig. 8. Overall, our method trained on BEHAVE generalizes well to NTU-RGBD as well as in the wild images, without any fine tuning.

One example question from our user studies is shown in Fig. 10.

4 Limitations and future work

Two typical failure cases of our method is shown in Fig. 9. Our method fails when the object is heavily occluded. In these cases, the object orientation prediction is incorrect and due to occlusion, our contact prediction can be noisy. These errors accumulated in the joint fitting and lead to failures.

One direction to improve the robustness of object pose prediction is to first train our network on other pose prediction or synthetic datasets where objects are occluded. Another big challenge of single view reconstruction is the lack of depth information. To add more 3D features to the decoder, one can additionally train a network to predict an intermediate depth map and stack this to the input images. One can also use the depth prediction to lift points to 3D and address the task following single view point cloud reconstruction methods. We leave these directions for future work.

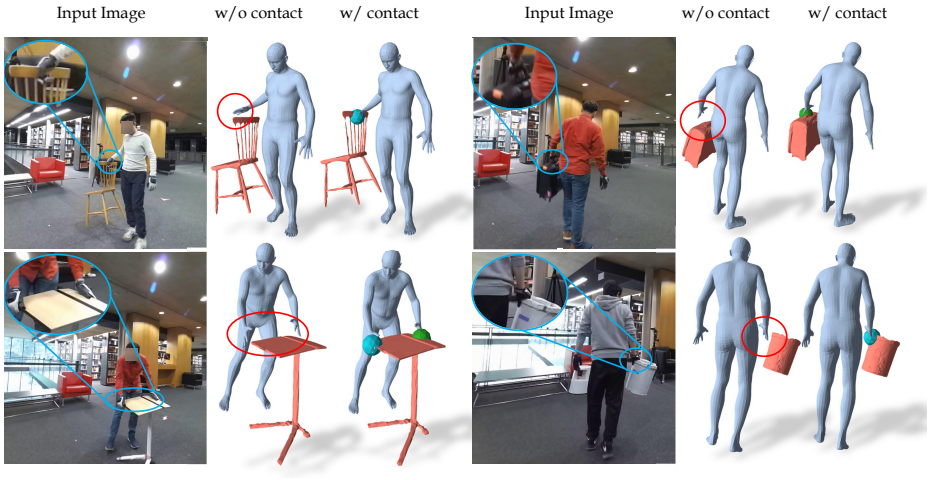

Fig. 1: Importance of contacts. We show more examples where our contact prediction helps improve the accuracy and physical plausibility of the joint reconstruction. Without contacts to snap objects to the correct interacting location with the person, the object maybe optimized towards inaccurate locations, leading to artefacts like floating objects in the air. For instance, the chair and the suitcase shown in the first row, which is not physically plausible. Our contacts prediction corrects these errors and leads more accurate reconstructions.

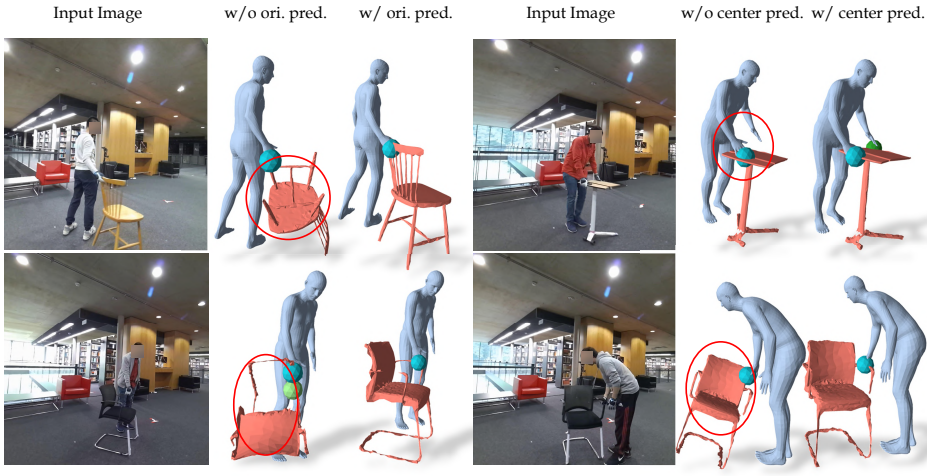

Fig. 2: Importance of object pose prediction. We compare method without orientation prediction (left) or object center prediction (right) with our full model. Without our orientation prediction to initialize object pose, the fitting gets stuck in local minima. Without object center to initialize and regularize object fitting, the object maybe optimized towards inaccurate depth and leads to incorrect contact and pose reconstruction in the end.

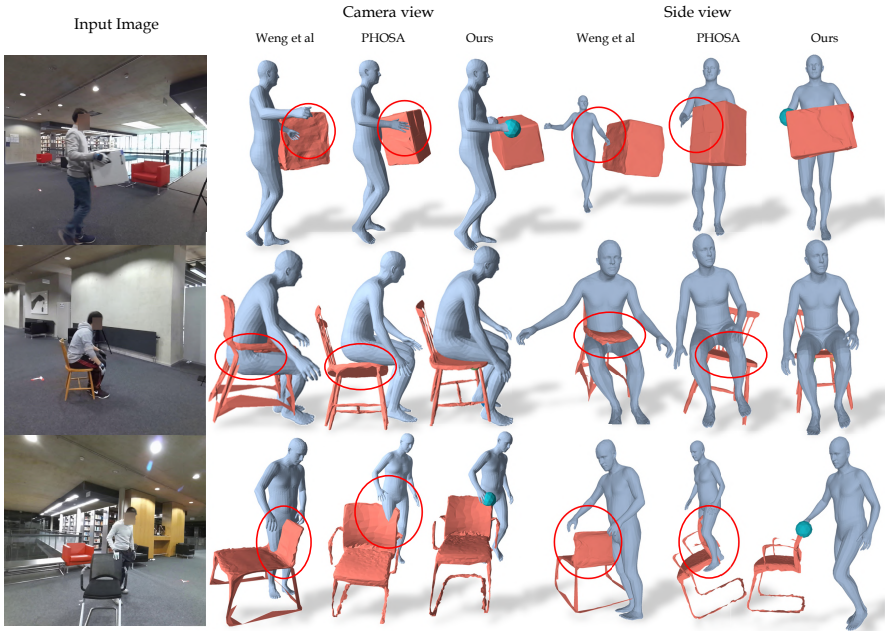

Fig. 3: More results on BEHAVE dataset [**?**]. It can be seen that our reconstruction is more accurate and consistent with input images.

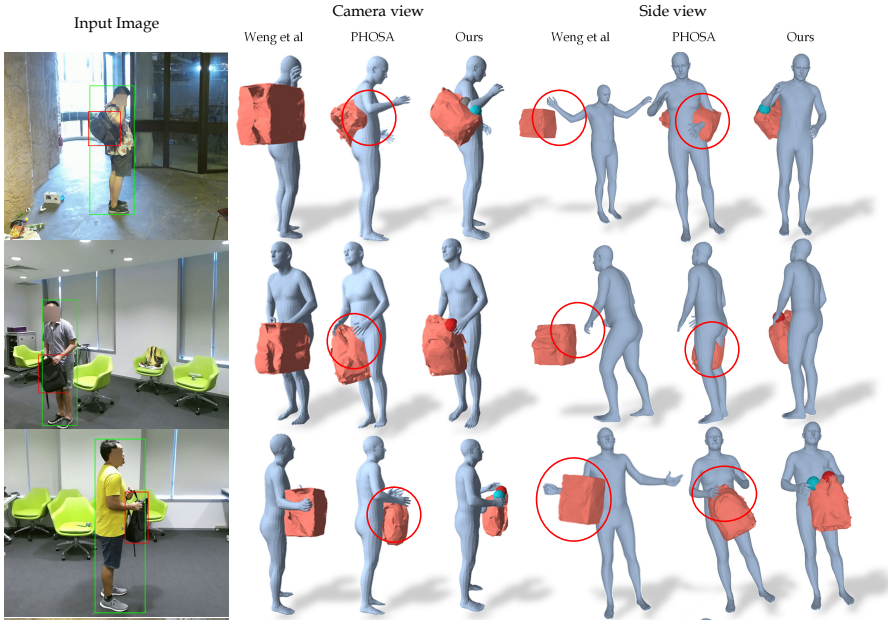

Fig. 4: Comparison with Weng et al. [8] and PHOSA [9] on the NTU-RGBD dataset [4]. Our joint reasoning method produces more accurate 3D reconstruction than baselines.

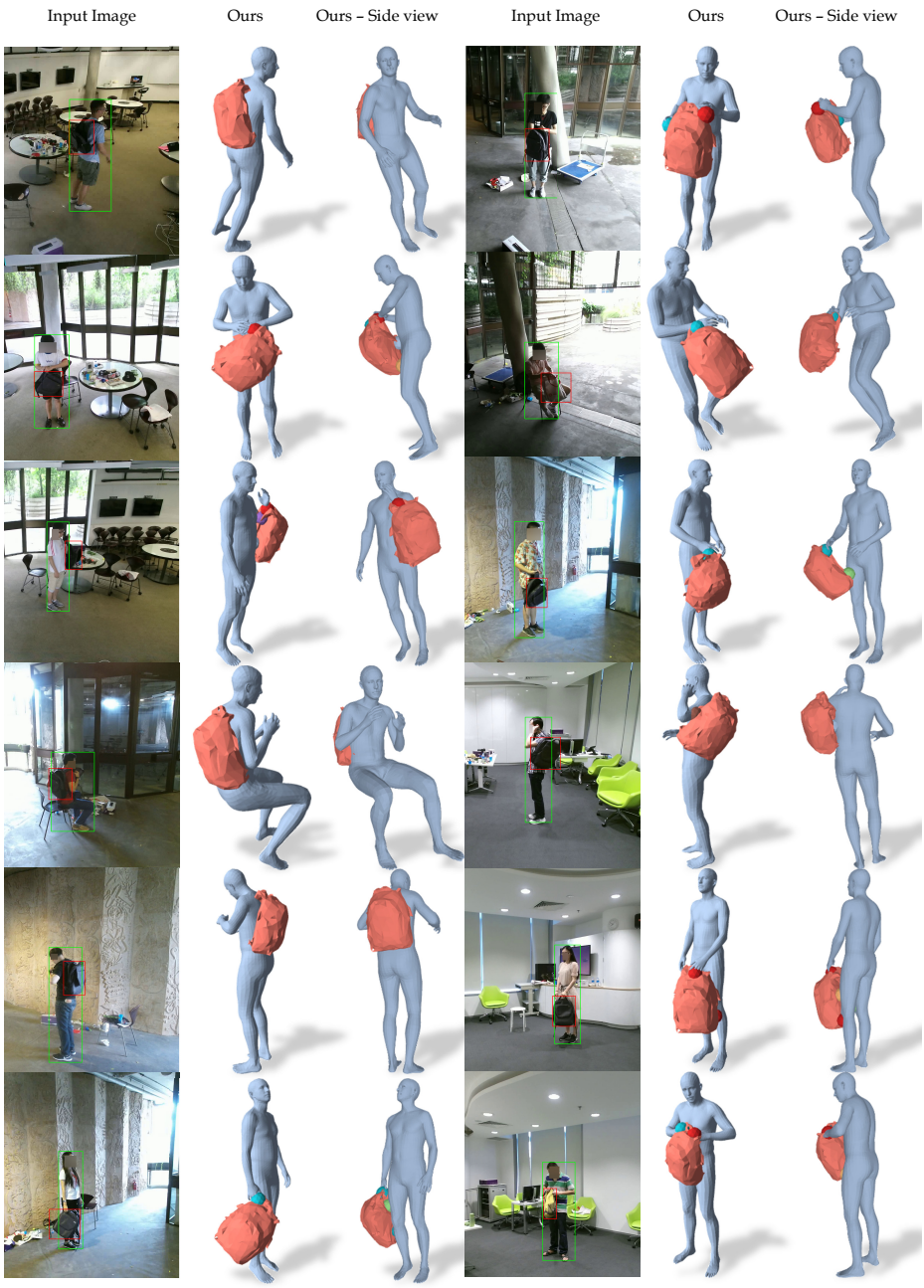

Fig. 5: More results from ours method on the NTU-RGBD [4] dataset. The human and object to be reconstructed are highlighted with green and red boxes respectively. It can be seen that our method generalizes well across diverse subjects, locations and camera viewpoints, without any fine tuning on this dataset.

Fig. 6: Comparison with PHOSA on in the wild images. The human and object to be reconstructed are highlighted with green and red boxes respectively. Our method generalizes better than PHOSA on in the wild images.

Fig. 7: More results from our method on in the wild images. It can be seen that our method generalizes well to images in the wild.

Fig. 8: More results from our method on in the wild images. Our method generalizes well to images in the wild.

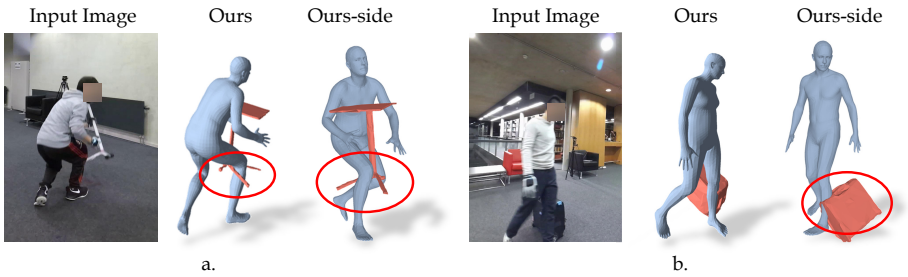

Fig. 9: We show two typical failure cases mainly caused by severe occlusion: inaccurate object orientation (a) and contact (b) prediction. It is difficult to predict accurate pose when over half of the table (a) is occluded and object fitting gets stuck in local minima due to this incorrect orientation initialization. The network also failed to correctly reason whether the suitcase (b) is in contact with the person or not and falsely push it to the person's foot.

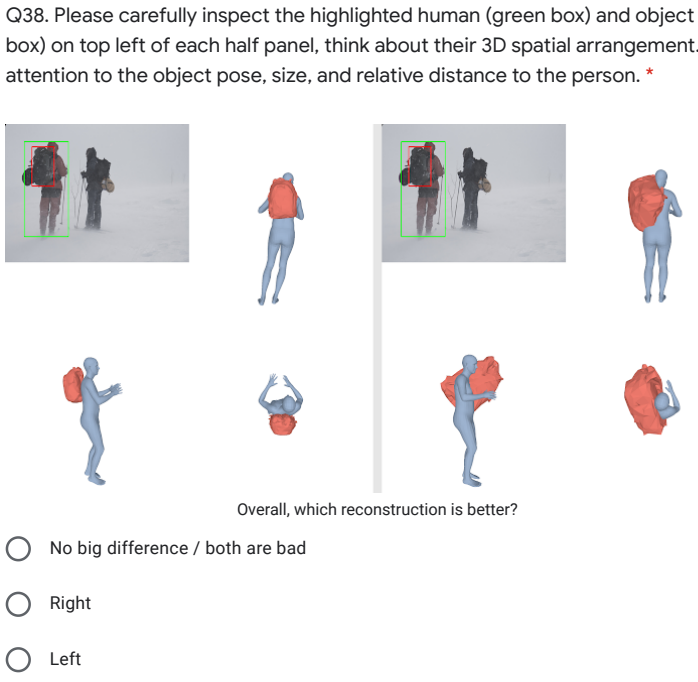

Fig. 10: Example question from our COCO user study survey. Annotators are asked to select which reconstruction is better or there is no big difference. Clockwise from top-left: original image, reconstruction rendered in camera view, side view and top-down view.

References

1. https://github.com/cmu-perceptual-computing-lab/openpose
2. Chibane, J., Mir, A., Pons-Moll, G.: Neural unsigned distance fields for implicit function learning. In: Neural Information Processing Systems (NeurIPS). (December 2020)
3. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 740–755. Springer International Publishing, Cham (2014)
4. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019). https://doi.org/10.1109/TPAMI.2019.2916873
5. Rong, Y., Shiratori, T., Joo, H.: Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In: IEEE International Conference on Computer Vision Workshops (2021)
6. Saito, S., , Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: IEEE International Conference on Computer Vision (ICCV). IEEE (oct 2019)
7. Saito, S., Simon, T., Saragih, J., Joo, H.: Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (June 2020)
8. Weng, Z., Yeung, S.: Holistic 3d human and scene mesh estimation from single view images. arXiv preprint arXiv:2012.01591 (2020)
9. Zhang, J.Y., Pepose, S., Joo, H., Ramanan, D., Malik, J., Kanazawa, A.: Perceiving 3d human-object spatial arrangements from a single image in the wild. In: European Conference on Computer Vision (ECCV) (2020)