InterCap: Joint Markerless 3D Tracking of Humans and Objects in Interaction
*Supplementary Material**

Yinghao Huang[1], Omid Taheri[1], Michael J. Black[1], and Dimitrios Tzionas[2]

[1] Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2] University of Amsterdam, Amsterdam, The Netherlands
{yhuang2, otaheri, black}@tue.mpg.de, d.tzionas@uva.nl

S.1 Video on our Website

The narrated video on our website (https://intercap.is.tue.mpg.de) presents:

- An explanation of our motivation.
- An overview of our InterCap dataset and method.
- Some videos (input) and reconstructed 4D meshes (output) of our InterCap dataset.
- A qualitative comparison between our InterCap mesh reconstructions to the ones from SMPLify-X [3], ExPose [1], and VIBE [2].

S.2 Optimization Objective Function & Terms

We use the objective function of Eq. 6 of the main paper to jointly refine (via optimization) the body and object motion over the whole sequence. Here we give a detailed explanation of the terms not elaborated in the main paper due to space limitations.

The motion smoothness term $E_{\mathcal{S}}$ penalizes sudden changes in the position of body vertices. It employs the learned motion prior of LEMO [4] and is defined as:

$$E_{\mathcal{S}}(\Theta, \Psi, \Gamma, A; T, \beta^*) = \frac{1}{Q(T-2)} \sum_{t=1}^{T-1} \left\| z_{t+1}^{opt} - z_t^{opt} \right\|^2, \tag{S.1}$$

where T is the sequence length, Q is a constant representing the number of virtual body-markers of LEMO (see [4] for an explanation; they use a different symbol), z_t^{opt} is the latent vector for the t-th frame from LEMO's pre-trained motion auto-encoder (F_S):

$$Z^{opt} = F_S(X_{\Delta}^{opt}) = [z_1^{opt}, z_2^{opt}, ..., z_{T-1}^{opt}], \tag{S.2}$$

where X_{Δ}^{opt} is a (concatenated) vector containing the temporal position change of LEMO's virtual body-markers. For more details, please refer to the paper of LEMO [4].

The vertex acceleration term $E_{\mathcal{A}}$ is a simple hand-crafted motion prior that encourages smooth motion trajectories for the object:

$$E_{\mathcal{A}}(\Xi; T, M) = \frac{1}{T-2} \sum_{t=2}^{T-1} \Big\| W'(\Xi_{t-1}, M) + W'(\Xi_{t+1}, M) - 2 * W'(\Xi_t, M) \Big\|^2 \tag{S.3}$$

where M is the object mesh, and W' denotes the operation of first rigidly deforming the object according to Ξ_t and then concatenating the vertices into a single vector.

The contact term $E_{\mathcal{C}}(\beta^*, \Theta_t, \Psi_t, \Gamma_t, \Xi_t, M)$ encourages the annotated likely contact areas of the body (see Fig. 3 of the main paper) to be in contact with the object:

$$E_{\mathcal{C}}(\beta^*, \Theta_t, \Psi_t, \Gamma_t, \Xi_t, M) = CD\Big(H\big(W(\Theta_t, \Psi_t, \Gamma_t, A, \beta^*)\big), H'\big(W'(\Xi_t, M)\big)\Big) \tag{S.4}$$

where CD refers to the Chamfer Distance function, H is a function that returns only the annotated body-contact vertices of Fig. 3, H' returns the closest points on the object for these body-contact vertices, W' deforms rigidly the object and is explained in the previous paragraph and W similarly (non-rigidly) deforms the SMPL-X mesh and concatenates the vertices into a single vector.

Finally, the ground-support terms $E_{\mathcal{G}}$ and $E_{\mathcal{G}'}$ build on the fact that no human or object vertex, respectively, should be below the ground plane, and penalize any vertex penetrating the ground. Let $p_{\mathcal{G}}$ be a point on the ground plane and $n_{\mathcal{G}}$ be the corresponding normal; both defined are once and offline. Then the term $E_{\mathcal{G}}$ for body-ground penetration is defined as:

$$E_{\mathcal{G}}(\beta^*, \Theta_t, \Psi_t, \Gamma_t) = \Big\| RL\Big(n_{\mathcal{G}} * \big(p_{\mathcal{G}} - W(\beta^*, \Theta_t, \Psi_t, \Gamma_t)\big)\Big)\Big\|^2, \tag{S.5}$$

where RL is the ReLU function, and $*$ here is the inner product of vectors. The term $E_{\mathcal{G}'}$ for object-ground penetration follows a similar formulation:

$$E_{\mathcal{G}'}(\Xi_t, M) = \Big\| RL(n_{\mathcal{G}} * (p_{\mathcal{G}} - W'(\Xi_t, M)))\Big\|^2. \tag{S.6}$$

Bibliography

[1] Choutas, V., Pavlakos, G., Bolkart, T., Tzionas, D., Black, M.J.: Monocular expressive body regression through body-driven attention. In: European Conference on Computer Vision (ECCV). vol. 12355, pp. 20–40 (2020) 1

[2] Kocabas, M., Athanasiou, N., Black, M.J.: VIBE: Video inference for human body pose and shape estimation. In: Computer Vision and Pattern Recognition (CVPR). pp. 5252–5262 (2020) 1

[3] Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: Computer Vision and Pattern Recognition (CVPR). pp. 10975–10985 (2019) 1

[4] Zhang, S., Zhang, Y., Bogo, F., Pollefeys, M., Tang, S.: Learning motion priors for 4D human body capture in 3D scenes. In: Computer Vision and Pattern Recognition (CVPR). pp. 11323–11333 (2021) 1