Capturing and Inferring Dense Full-Body Human-Scene Contact -Supplementary Material-

Chun-Hao P. Huang[1] Hongwei Yi[1] Markus Höschle[1] Matvey Safroshkin[1] Tsvetelina Alexiadis[1]
Senya Polikovsky[1] Daniel Scharstein[2] Michael J. Black[1]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]Middlebury College
{paul.huang, firstname.lastname, black}@tuebingen.mpg.de, schar@middlebury.edu

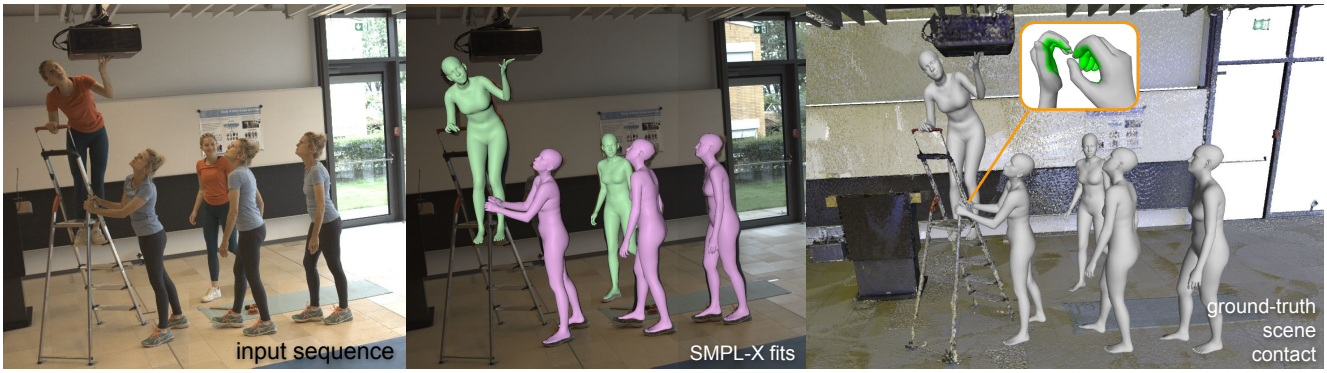

Figure R.1. The **RICH dataset** contains multiple people interacting with a real scene. It provides complex natural images, precise 3D scene scans, pseudo ground-truth SMPL-X bodies, and dense body contact labels.

The Supplementary Material consists of this document and a video. They include additional information and visualizations of our dataset, method, and results.

1. SMPL-X vs. SMPL HSC labels

We build RICH by fitting a SMPL-X template to multi-view data and compute the human-scene contact (HSC) as explained in the Sec. 3 and Sec. 5 of the main paper (Fig. R.1). The contact labels are defined in SMPL-X format and we map them to SMPL format for training BSTRO. This is feasible since there is an 1-to-1 correspondence between SMPL-X and SMPL vertices below the neck, as shown in Fig. R.2.

With this mapping, we convert the ground-truth HSC labels from SMPL-X to SMPL without losing information. As a result, we benefit from realistic hand articulation in SMPL-X and still keep the dimension of the output space small (SMPL). Such a mapping also makes RICH a suitable HSC benchmark for both body models. Since the two models share the set of vertices of interest, choosing either of them does not influence the detection scores or errors.

On the other hand, the human pose and shape (HPS) parameters of the two models differ. Converting HPS parameters between SMPL-X and SMPL requires extra processing [1] and one always loses the hand articulation when converting from SMPL-X to SMPL. Therefore, RICH provides only SMPL-X as pseudo ground truth. To evaluate methods that regress SMPL parameters using RICH, users should convert SMPL to SMPL-X, which does not result in a loss of information.

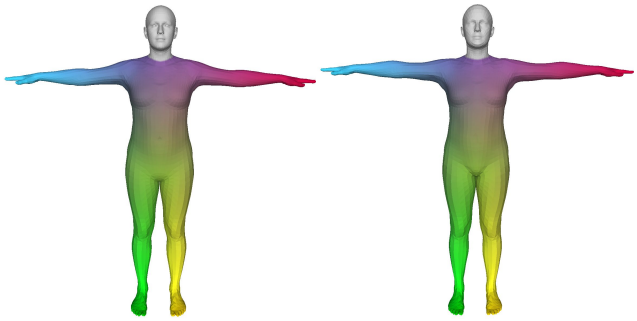

Figure R.2. SMPL-X and SMPL bodies share the same set of vertices for regions below the neck. The same vertices are visualized in the same colors.

2. RICH Dataset

The 134 multi-view videos in RICH are recorded at a rate of 30 frames per second. We separate them into subsets of 57, 27, 50 for training, validation, and testing purposes, respectively. This amounts to 277K, 142K, 121K images of 4K resolutions (in total 540K), and 36K, 18K, 31K 3D SMPL-X bodies along with dense scene-contact labels (in total 540K) in each subset. By "body" here, we mean any SMPL-X mesh. Note that the number of unique "people" in the dataset is much smaller than the number of bodies because every posed mesh constitutes a separate body.

Compared to the recent HPS dataset AGORA [7], RICH has more 3D bodies (85K vs. 4K), more images (540K vs. 19K) and more accurate body shapes (registrations to minimally-closed scans [3] vs. clothed scans [9]). It has more subjects in varied body shapes than 3DPW [8] (22 vs. 18) and subjects are in natural clothing as opposed to those in Human3.6M [4]. Last but not least, RICH provides high-quality scene scans and scene contact labels that none of the above datasets provides.

3. Bone-orientation Term E_O

Following the illustration in Fig. 2(a) of the main paper, the bone-orientation term E_O factors out the residual of the parent joint ϵ_1 from the residual of the child joint ϵ_2:

$$\begin{aligned} r_2 &= \epsilon_2 - \epsilon_1, \\ &= (j'_2 - j_2) - (j'_1 - j_1), \\ &= (j'_2 - j'_1) - (j_2 - j_1), \\ &= b'_2 - b_2, \end{aligned}$$

where $b'_2 = j'_2 - j'_1$ and $b_2 = j_2 - j_1$ denote the "bone vector" of target points (detected landmarks) and estimated SMPL-X joints respectively. It follows that

$$\|r_2\|_2^2 = \|b'_2\|_2^2 + \|b_2\|_2^2 - b_2^\top b'_2. \tag{1}$$

Since b'_2 involves only the detected landmarks and $\|b_2\|$ is fixed given a constant body shape β, the first two terms are constant when optimizing the multi-view objective E_{mv}. $\|r_2\|_2^2$ is therefore minimized when $b_2^\top b'_2$ is maximized, i.e., when b_2 has the same orientation as b'_2.

4. BSTRO Implementation Details

We sample RICH-train to build the image-HSC pairs $(I, \mathbf{c})$ for training BSTRO. For each sequence, we consider only every other frame, and for each frame, we use the dynamic view and one randomly selected static view, or two static views if no moving camera is available. This sampling strategy ensures sufficient variations in viewpoints and background, while keeping the total number of the training pairs tractable.

We train with in total 21K $(I, \mathbf{c})$ pairs from RICH-train and use Adam [5] optimizer with an initial learning rate of 1e-4 for 100 epochs. The HR-Net backbone is initialized with the weights pre-trained on ImageNet [2], Human3.6M [4] or 3DPW [8]. The best checkpoint is selected by the best performance on RICH-validation with RICH-test completely withheld. We refer interested readers to [6] for the architecture of the multi-layer transformer.

References

[1] https://github.com/vchoutas/smplx/tree/master/transfer_model. 1

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 2

[3] David Hirshberg, Matthew Loper, Eric Rachlin, and Michael J. Black. Coregistration: Simultaneous alignment and modeling of articulated 3D shape. In *European Conference on Computer Vision*, 2012. 2

[4] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 2

[5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 2

[6] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[7] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[8] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European Conference on Computer Vision*, 2018. 2

[9] Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4191–4200, 2017. 2