Populating 3D Scenes by Learning Human-Scene Interaction
Supplementary Material

Mohamed Hassan Partha Ghosh Joachim Tesch Dimitrios Tzionas Michael J. Black
Max Planck Institute for Intelligent Systems, Tübingen, Germany
{mhassan, pghosh, jtesch, dtzionas, black}@tuebingen.mpg.de

Figure 1: POSA automatically places 3D people in 3D scenes such that the interactions between the people and the scene are both geometrically and semantically correct. POSA exploits a new learned representation of human bodies that explicitly models how bodies interact with scenes.

Appendices

A. Training Details

The global orientation of the body is typically irrelevant in our body-centric representation, so we rotate the training bodies around the y and z axes to put them in a canonical orientation. The rotation around the x axis, however, is essential to enable the model to differentiate between standing up and lying down. The semantic labels for the PROX scenes are taken from Zhang et al. [9], where scenes were manually labeled following the object categorization of Matterport3D [1], which incorporates 40 object categories.

Our encoder-decoder architecture is similar to the one introduced in Gong et al. [2]. The encoder consists of 3 spiral convolution layers interleaved with pooling layers $3 \times \{\text{Conv}\,(64) \rightarrow \text{Pool}\,(4)\} \rightarrow \text{FC}\,(512)$. Pool stands for a downsampling operation as in COMA [7], which is based on contracting vertices. FC is a fully connected layer and the number in the bracket next to it denotes the number of units in that layer. We add 2 additional fully connected layers to predict the parameters of the latent code, with fully connected layers of 256 units each. The input to the encoder is a body mesh M_b where, for each vertex, i, we concatenate V_b^i vertex positions, and f vertex features. For computational efficiency, we first downsample the input mesh by a factor of 4. So instead of working on the full mesh resolution of 10475 vertices, our input mesh has a resolution of 655 vertices. The decoder architecture consists of spiral convolution layers only $4 \times \{\text{Conv}\,(64)\} \rightarrow \text{Conv}\,(N_f)$. We attach the latent vector z to the 3D coordinates of each vertex similar to Kolotouros et al. [5].

We build our model using the PyTorch framework. We use the Adam optimizer [4], batch size of 64, and learning rate of $1e^{-3}$ without learning rate decay.

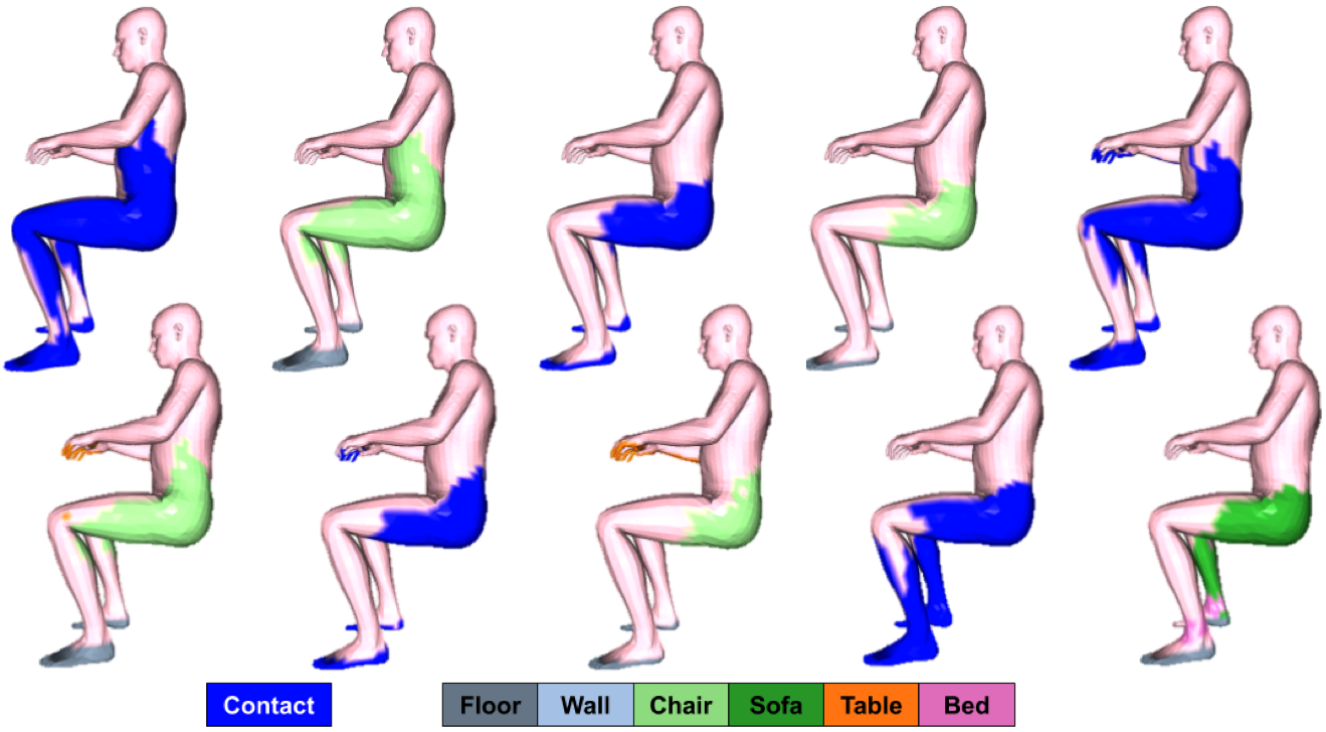

Figure S.1: Random samples from our trained cVAE for the same pose. For each example we show from left to right: f_c and f_s. The color code is at the bottom. For f_c, blue means contact, while pink means no contact. For f_s, each scene category has a different color.

B. SDF Computation

For computational efficiency, we employ a precomputed 3D signed distance field (SDF) for the static scene $\mathcal{S}_s$, following Hassan et al. [3]. The SDF has a resolution of $512 \times 512 \times 512$. Each voxel c_j stores the distance $d_j \in \mathbb{R}$ of its centroid $P_j \in \mathbb{R}^3$ to the nearest surface point $P_s \in \mathcal{S}_s$. The distance d_j has a positive sign if P_j lies in the free space outside physical scene objects, while it has a negative sign if it is inside a scene object.

C. Random Samples

We show multiple randomly sampled feature maps for the same pose in Fig. S.1. Note how POSA generate a variety of valid feature maps for the same pose. Notice for example that the feet are always correctly predicted to be in contact with the floor. Sometimes our model predicts the person is sitting on a chair (far left) or on a sofa (far right).

The predicted semantic map f_s is not always accurate as shown in the far right of Fig. S.1. The model predicts the person to be sitting on a sofa but at the same time predicts the lower parts of the leg to be in contact with a bed which is unlikely.

D. Affordance Detection

The complete pipeline of the affordance detection task is shown in Fig. S.2. Given a clothed 3D mesh that we want to put in a scene, we first need a SMPL-X fit to the mesh; here we take this from the AGORA dataset [6]. Then we generate a feature map using the decoder of our cVAE by sampling $P(f_{\text{Gen}}|z, V_b)$. Next we minimize the energy function in Eq. 1.

$$E(\tau, \theta_0) = \mathcal{L}_{afford} + \mathcal{L}_{pen} \tag{1}$$

Finally, we replace the SMPL-X mesh with the original clothed.

We show additional qualitative examples of SMPL-X meshes automatically placed in real and synthetic scenes in Fig. S.4. Qualitative examples of clothed bodies placed in real and synthetic scenes are shown in Fig. S.5. We show qualitative comparison between our results and PLACE [8] in Fig. S.6.

E. Failure Cases

We show representative failure cases in Fig. S.3. A common failure mode is residual penetrations; even with the penetration penalty the body can still penetrate the scene. This can happen due to thin surfaces that are not captured

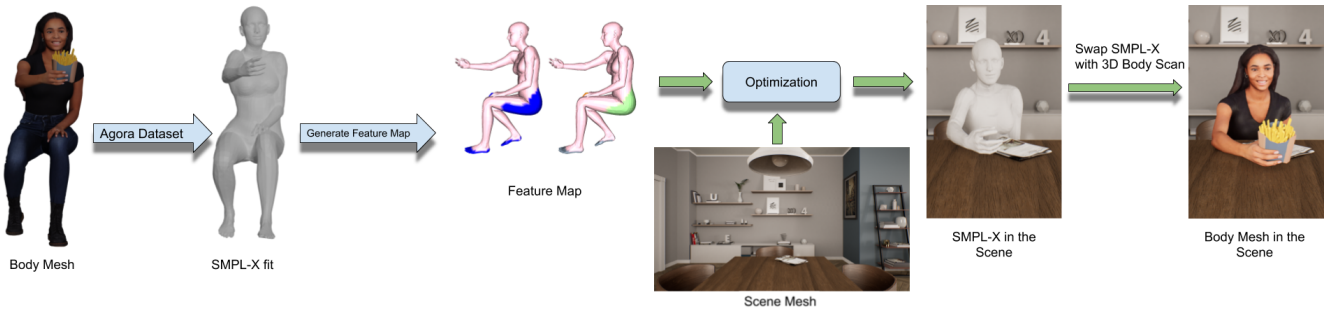

Figure S.2: Putting realistic people in scenes. Pipeline of affordance detection using meshes with clothing. SMPL-X acts as a proxy for the clothed scan. POSA is used to sample features for this pose. These features are then used with the scene mesh to optimize the placement of the body. After convergence, we simply replace SMPL-X with the clothed scan.

Figure S.3: Failure cases.

by our SDF and/or because the optimization becomes stuck in a local minimum. In other cases, the feature map might not be right. This can happen when the model does not generalize well to test poses due to the limited training data.

F. Effect of Shape

Fig. S.7 shows that our model can predict plausible feature maps for a wide range of human body shapes.

G. Scene population.

In Fig. S.8 we show the three main steps to populate a scene: **(1)** Given a scene, we create a regular grid of candidate positions (Fig. S.8.1). We place the body, in a given pose, at each candidate position and evaluate Eq. 10 once. **(2)** We then keep the 10 best candidates with the lowest energy (Fig. S.8.2), and **(3)** iteratively optimize Eq. 10 for these; Fig. S.8.3 shows results at three positions, with the best one highlighted with green.

References

[1] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*, pages 667–676, 2017.

[2] Shunwang Gong, Lei Chen, Michael Bronstein, and Stefanos Zafeiriou. SpiralNet++: A fast and highly efficient mesh convolution operator. In *International Conference on Computer Vision Workshops (ICCVw)*, pages 4141–4148, 2019.

[3] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constrains. In *International Conference on Computer Vision (ICCV)*, pages 2282–2292, 2019.

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

[5] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4501–4510, 2019.

[6] Priyanka Patel, Chun-Hao Paul Huang, Joachim Tesch, David Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Computer Vision and Pattern Recognition (CVPR)*, 2021.

[7] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. Generating 3D faces using convolutional mesh autoencoders. In *European Conference on Computer Vision (ECCV)*, volume 11207, pages 725–741, 2018.

[8] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J. Black, and Siyu Tang. PLACE: Proximity learning of articulation and contact in 3D environments. In *International Conference on 3D Vision (3DV)*, pages 642–651, 2020.

Figure S.4: Qualitative examples of SMPL-X meshes automatically placed in real and synthetic scenes. The body shapes and poses were not used in training.

[9] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J. Black, and Siyu Tang. Generating 3D people in scenes without people. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6193–6203, 2020.

Figure S.5: Clothed bodies (from Renderpeople) automatically placed in real and synthetic scenes.

Figure S.6: Qualitative examples from POSA (pink) and PLACE [8] (silver).

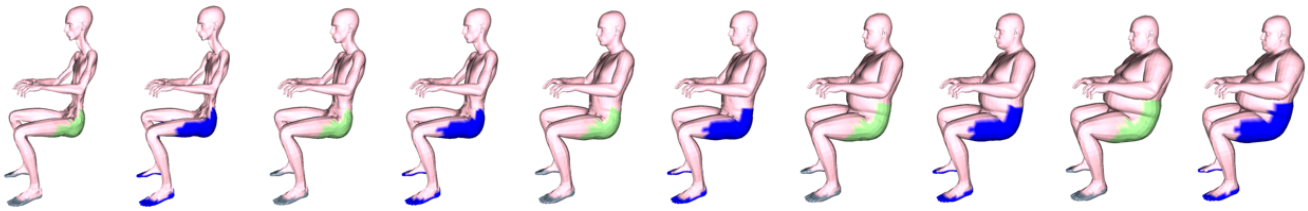

Figure S.7: Generated feature maps for various body shapes.

Figure S.8: Main steps of our method for scene population. **(1)** Grid with all candidate positions. **(2)** The 10 best positions. **(3)** Final result.