Populating 3D Scenes by Learning Human-Scene Interaction

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.

Abstract

Humans live within a 3D space and constantly interact with it to perform tasks. Such interactions involve physical contact between surfaces that is semantically meaningful. Our goal is to learn how humans interact with scenes and leverage this to enable virtual characters to do the same. To that end, we introduce a novel Human-Scene Interaction (HSI) model that encodes proximal relationships, called POSA for "Pose with prOximitieS and contActs". The representation of interaction is body-centric, which enables it to generalize to new scenes. Specifically, POSA augments the SMPL-X parametric human body model such that, for every mesh vertex, it encodes (a) the contact probability with the scene surface and (b) the corresponding semantic scene label. We learn POSA with a VAE conditioned on the SMPL-X vertices, and train on the PROX dataset, which contains SMPL-X meshes of people interacting with 3D scenes, and the corresponding scene semantics from the PROX-E dataset. We demonstrate the value of POSA with two applications. First, we automatically place 3D scans of people in scenes. We use a SMPL-X model fit to the scan as a proxy and then find its most likely placement in 3D. POSA provides an effective representation to search for "affordances" in the scene that match the likely contact relationships for that pose. We perform a perceptual study that shows significant improvement over the state of the art on this task. Second, we show that POSA's learned representation of body-scene interaction supports monocular human pose estimation that is consistent with a 3D scene, improving on the state of the art. Our model and code are available for research purposes at https://posa.is.tue.mpg.de.

1. Introduction

Humans constantly interact with the world around them. We move by walking on the ground; we sleep lying on a bed; we rest sitting on a chair; we work using touchscreens and keyboards. Our bodies have evolved to exploit the affordances of the natural environment and we design objects

to better "afford" our bodies. While obvious, it is worth stating that these physical interactions involve contact. Despite the importance of such interactions, existing representations of the human body do not explicitly represent, support, or capture them.

In computer vision, human pose is typically estimated in isolation from the 3D scene, while in computer graphics 3D scenes are often scanned and reconstructed without people. Both the recovery of humans in scenes and the automated synthesis of realistic people in scenes remain challenging problems. Automation of this latter case would reduce animation costs and open up new applications in augmented reality. Here we take a step towards automating the realistic placement of 3D people in 3D scenes with realistic contact and semantic interactions (Fig. 1). We develop a novel *body-centric* approach that relates 3D body shape and pose to possible world interactions. Learned parametric 3D human models [2, 28, 39, 46] represent the shape and pose of people accurately. We employ the SMPL-X [46] model, which includes the hands and face, as it supports reasoning about contact between the body and the world.

While such body models are powerful, we make three *key observations*. First, human models like SMPL-X [46] do not explicitly model contact. Second, not all parts of the body surface are equally likely to be in contact with the scene. Third, the poses of our body and scene semantics are highly intertwined. Imagine a person sitting on a chair; body contact likely includes the buttocks, probably also the back, and maybe the arms. Think of someone opening a door; their feet are likely in contact with the floor, and their hand is in contact with the doorknob.

Based on these observations, we formulate a novel model, that makes human-scene interaction (HSI) an explicit and integral part of the body model. The key idea is to encode HSI in an ego-centric representation built in SMPL-X. This effectively extends the SMPL-X model to capture contact and the semantics of HSI in a body-centric representation. We call this POSA for *"Pose with prOximitieS and contActs"*. Specifically, for every vertex on the body and every pose, POSA defines a probabilistic feature map that encodes the probability that the vertex is in contact with the world and the distribution of semantic labels associated with that contact.

POSA is a conditional Variational Auto-Encoder (cVAE), conditioned on SMPL-X vertex positions. We train on the PROX dataset [22], which contains 20 subjects, fit with SMPL-X meshes, interacting with 12 real 3D scenes. We also train POSA using use the scene semantic annotations provided by the PROX-E dataset [65]. Once trained, given a posed body, we can sample likely contacts and semantic labels for all vertices. We show the value of this representation with two challenging applications.

First, we focus on automatic scene population as illustrated in Fig. 1. That is, given a 3D scene and a body in a particular pose, where in the scene is this pose most likely? As demonstrated in Fig. 1 we use SMPL-X bodies fit to commercial 3D scans of people [45], and then, conditioned on the body, our cVAE generates a target POSA feature map. We then search over possible human placements while minimizing the discrepancy between the observed and target feature maps. We quantitatively compare our approach to PLACE [64], which is SOTA on a similar task, and find that POSA has higher perceptual realism.

Second, we use POSA for monocular 3D human pose estimation in a 3D scene. We build on the PROX method [22] that hand-codes contact points, and replace these with our learned feature map, which functions as an HSI prior. This automates a heuristic process, while producing lower pose estimation errors than the original PROX method.

To summarize, POSA is a novel model that intertwines SMPL-X pose and scene semantics with contact. To the best of our knowledge, this is the first learned human body model that incorporates HSI in the model. We think this is important because such a model can be used in all the same ways that models like SMPL-X are used but now with the addition of body-scene interaction. The key novelty is posing HSI as part of the body representation itself. Like the original learned body models, POSA provides a platform that people can build on. To facilitate this, our model and code are available for research purposes at `https://posa.is.tue.mpg.de`.

2. Related Work

Humans & Scenes in Isolation: For *scenes* [68], most work focuses on their 3D shape in isolation, e.g. on rooms devoid of humans [3, 10, 16, 56, 60] or on objects that are not grasped [8, 11, 13, 61]. For *humans*, there is extensive work on capturing [5, 26, 40, 54] or estimating [41, 51] their 3D shape and pose, but outside the context of scenes.

Human Models: Most work represents 3D humans as body skeletons [26, 54]. However, the 3D body surface is important for physical interactions. This is addressed by learned parametric 3D body models [2, 28, 39, 44, 46, 62]. For interacting humans we employ SMPL-X [46], which models the body with full face and hand articulation.

HSI Geometric Models: We focus on the spatial relationship between a human body and objects it interacts with. Gleicher [18] uses contact constraints for early work on motion re-targeting. Lee et al. [34] generate novel scenes and motion by deformably stitching "motion patches", comprised of scene patches and the skeletal motion in them. Lin et al. [37] generate 3D skeletons sitting on 3D chairs, by manually drawing 2D skeletons and fitting 3D skeletons that satisfy collision and balance constraints. Kim et al. [30] automate this, by detecting sparse contacts on a 3D object mesh and fitting a 3D skeleton to contacts while avoid-

ing penetrations. Kang et al. [29] reason about the physical comfort and environmental support of a 3D humanoid, through force equilibrium. Leimer et al. [35] reason about pressure, frictional forces and body torques, to generate a 3D object mesh that comfortably supports a given posed 3D body. Zheng et al. [66] map high-level ergonomic rules to low-level contact constraints and deform an object to fit a 3D human skeleton for force equilibrium. Bar-Aviv et al. [4] and Liu et al. [38] use an interacting agent to describe object shapes through detected contacts [4], or relative distance and orientation metrics [38]. Gupta et al. [21] estimate human poses "afforded" in a depicted room, by predicting a 3D scene occupancy grid, and computing support and penetration of a 3D skeleton in it. Grabner et al. [20] detect the places on a 3D scene mesh where a 3D human mesh can sit, modeling interaction likelihood with GMMs and proximity and intersection metrics. Zhu et al. [67] use FEM simulations for a 3D humanoid, to learn to estimate forces, and reasons about sitting comfort.

Several methods focus on dynamic interactions [1, 25, 47]. Ho et al. [25] compute an "interaction mesh" per frame, through Delaunay tetrahedralization on human joints and unoccluded object vertices; minimizing their Laplacian deformation maintains spatial relationships. Others follow an object-centric approach [1, 47]. Al-Asqhar et al. [1] sample fixed points on a 3D scene, using proximity and "coverage" metrics, and encode 3D human joints as transformations w.r.t. these. Pirk et al. [47] build functional object descriptors, by placing "sensors" around and on 3D objects, that "sense" 3D flow of particles on an agent.

HSI Data-driven Models: Recent work takes a data-driven approach. Jiang et al. [27] learn to estimate human poses and object affordances from an RGB-D scene, for 3D scene label estimation. SceneGrok [52] learns action-specific classifiers to detect the likely scene places that "afford" a given action. Fisher et al. [17] use SceneGrok and interaction annotations on CAD objects, to embed noisy 3D room scans to CAD mesh configurations. PiGraphs [53] maps pairs of {verb-object} labels to "interaction snapshots", i.e. 3D interaction layouts of objects and a human skeleton. Chen et al. [12] map RGB images to "interaction snapshots", using Markov Chain Monte Carlo with simulated annealing to optimize their layout. iMapper [42] maps RGB videos to dynamic "interaction snapshots", by learning "scenelets" on PiGraphs data and fitting them to videos. Phosa [63] infers spatial arrangements of humans and objects from a single image. Cao et al. [9] map an RGB scene and 2D pose history to 3D skeletal motion, by training on video-game data. Li et al. [36] follow [59] to collect 3D human skeletons consistent with 2D/3D scenes of [55, 59], and learn to predict them from a color and/or depth image. Corona et al. [14] use a graph attention model to predict motion for objects and a human skeleton, and their evolv-

ing spatial relationships.

Another HSI variant is Hand-Object Interaction (HOI); we discuss only recent work [7, 15, 57, 58]. Brahmbhatt et al. [7] capture fixed 3D hand-object grasps, and learn to predict contact; features based on object-to-hand mesh distances outperform skeleton-based variants. For grasp generation, 2-stage networks are popular [43]. Taheri et al. [57] capture moving SMPL-X [46] humans grasping objects, and predict MANO [50] hand grasps for object meshes, whose 3D shape is encoded with BPS [48]. Corona et al. [15] generate MANO grasping given an object-only RGB image; they first predict the object shape and rough hand pose (grasp type), and then they refine the latter with contact constraints [23] and an adversarial prior.

Closer to us, PSI [65] and PLACE [64] populate 3D scenes with SMPL-X [46] humans. Zhang et al. [65] train a cVAE to estimate humans from a depth image and scene semantics. Their model provides an implicit encoding of HSI. Zhang et al. [64], on the other hand, explicitly encode the scene shape and human-scene proximal relations with BPS [48], but do not use semantics. Our key difference to [64, 65] is our human-centric formulation; inherently this is more portable to new scenes. Moreover, instead of the sparse BPS distances of [64], we use dense body-to-scene contact, and also exploit scene semantics like [65].

3. Method

3.1. Human Pose and Scene Representation

Our training data corpus is a set of n pairs of 3D meshes

$$\mathcal{M} = \{\{M_{b,1}, M_{s,1}\}, \{M_{b,2}, M_{s,2}\}, \ldots, \{M_{b,n}, M_{s,n}\}\}$$

comprising body meshes $M_{b,i}$ and scene meshes $M_{s,i}$. We drop the index, i, for simplicity when we discuss meshes in general. These meshes approximate human body surfaces $\mathcal{S}_b$ and scene surfaces $\mathcal{S}_s$. Scene meshes $M_s = (V_s, F_s, L_s)$ have a varying number of vertices $N_s = |V_s|$ and triangle connectivity F_s to model arbitrary scenes. They also have per-vertex semantic labels L_s. Human meshes are represented by SMPL-X [46], i.e. a differentiable function $M(\theta, \beta, \psi) : \mathbb{R}^{|\theta| \times |\beta| \times |\psi|} \rightarrow \mathbb{R}^{(N_b \times 3)}$ parameterized by pose θ, shape β and facial expressions ψ. The pose vector $\theta = (\theta_b, \theta_f, \theta_{lh}, \theta_{rh})$ is comprised of body, $\theta_b \in \mathbb{R}^{66}$, and face parameters, $\theta_f \in \mathbb{R}^9$, in axis-angle representation, while $\theta_{lh}, \theta_{rh} \in \mathbb{R}^{12}$ parameterize the poses of the left and right hands respectively in a low-dimensional pose space. The shape parameters, $\beta \in \mathbb{R}^{10}$, represent coefficients in a low-dimensional shape space learned from a large corpus of human body scans. The joints, $J(\beta)$, of the body in the canonical pose are regressed from the body shape. The skeleton has 55 joints, consisting of 22 body joints, 30 hand joints, and 3 head joints (neck and eyes). The mesh is posed using this skeleton and linear blend skinning.
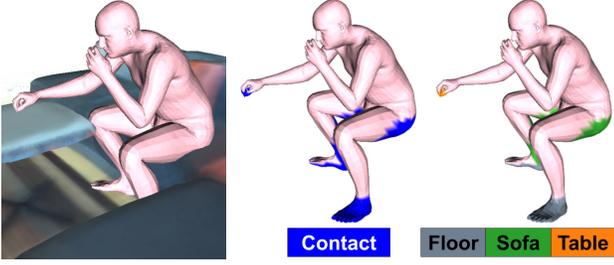
Figure 2: Illustration of our proposed representation. From left to right: An example of a SMPL-X mesh M_b in a scene M_s, contact f_c, and scene semantics f_s on it. For f_c, blue means the body vertex is likely in contact. For f_s, the colors correspond to the scene semantics.

Body meshes $M_b = (V_b, F_b)$ have a fixed topology with $N_b = |V_b| = 10475$ vertices $V_b \in \mathbb{R}^{(N_b \times 3)}$ and triangles F_b, i.e. all human meshes are in correspondence. For more details, we refer the reader to [46].

3.2. POSA Representation for HSI

We encode the relationship between the human mesh $M_b = (V_b, F_b)$ and the scene mesh $M_s = (V_s, F_s, L_s)$ in an egocentric feature map f that encodes per-vertex features on the SMPL-X mesh M_b. We define f as:

$$f : (V_b, M_s) \rightarrow [f_c, f_s], \qquad (1)$$

where f_c is the contact label and f_s is the semantic label of the contact point. N_f is the feature dimension.

For each vertex i on the body, V_b^i, we find its closest scene point $P_s = \operatorname{argmin}_{P_s \in \mathcal{S}_s} \|P_s - V_b^i\|$. Then we compute the distance f_d:

$$f_d = \|P_s - V_b^i\| \in \mathbb{R}. \qquad (2)$$

Given f_d, we can compute whether a V_b^i is in contact with the scene or not, with f_c:

$$f_c = \begin{cases} 1 & f_d \leq \textit{Contact Threshold}, \\ 0 & f_d > \textit{Contact Threshold}. \end{cases} \qquad (3)$$

The contact threshold is chosen empirically to be 5 cm. The semantic label of the contacted surface f_s is a one-hot encoding of the object class:

$$f_s = \{0, 1\}^{N_o}, \qquad (4)$$

where N_o is the number of object classes. The sizes of f_c, f_s, and f are 1, 40 and 41 respectively. All the features are computed once offline to speed training up. A visualization of the proposed representation is in Fig. 2.

3.3. Learning

Our goal is to learn a probabilistic function from body pose and shape to the feature space of contact and semantics. That is, given a body, we want to sample labelings of the vertices corresponding to likely scene contacts and their corresponding semantic label. Note that this function, once learned, only takes the body as input and not a scene – it is a *body-centric* representation.

To train this we use the PROX [22] dataset, which contains bodies in 3D scenes. We also use the scene semantic annotations from the PROX-E dataset [65]. For each body mesh M_b, we factor out the global translation and rotation R_y and R_z around the y and z axes. The rotation R_x around the x axis is essential for the model to differentiate between, e.g., standing up and lying down.

Given pose and shape parameters in a given training frame, we compute a $M_b = M(\theta, \beta, \psi)$. This gives vertices V_b from which we compute the feature map that encodes whether each V_b^i is in contact with the scene or not, and the semantic label of the scene contact point P_s.

We train a conditional Variational Autoencoder (cVAE), where we condition the feature map on the vertex positions, V_b, which are a function of the body pose and shape parameters. Training optimizes the encoder and decoder parameters to minimize $\mathcal{L}_{\text{total}}$ using gradient descent:

$$\mathcal{L}_{\text{total}} = \alpha * \mathcal{L}_{KL} + \mathcal{L}_{rec}, \qquad (5)$$

$$\mathcal{L}_{KL} = KL(Q(z|f, V_b) \| p(z)), \qquad (6)$$

$$\mathcal{L}_{rec}\left(f, \hat{f}\right) = \lambda_c * \sum_i \text{BCE}\left(f_c^{\,i}, \hat{f}_c^{\,i}\right)$$
$$+ \lambda_s * \sum_i \text{CCE}\left(f_s^i, \hat{f}_s^{\,i}\right), \qquad (7)$$

where $\hat{f}_c$ and $\hat{f}_s$ are the reconstructed contact and semantic labels, *KL* denotes the Kullback Leibler divergence, and $\mathcal{L}_{rec}$ denotes the reconstruction loss. BCE and CCE are the binary and categorical cross entropies respectively. The α is a hyperparameter inspired by Gaussian β-VAEs [24], which regularizes the solution; here $\alpha = 0.05$. $\mathcal{L}_{rec}$ encourages the reconstructed samples to resemble the input, while $\mathcal{L}_{KL}$ encourages $Q(z|f, V_b)$ to match a prior distribution over z, which is Gaussian in our case. We set the values of λ_c and λ_s to 1.

Since f is defined on the vertices of the body mesh M_b, this enables us to use graph convolution as our building block for our VAE. Specifically, we use the Spiral Convolution formulation introduced in [6, 19]. The spiral convolution operator for node i in the body mesh is defined as:

$$f_k^i = \gamma_k \left(\|_{j \in S(i,l)} f_{k-1}^j \right), \qquad (8)$$

where γ_k denotes layer k in a multi-layer perceptron (MLP) network, and $\|$ is a concatenation operation of the features
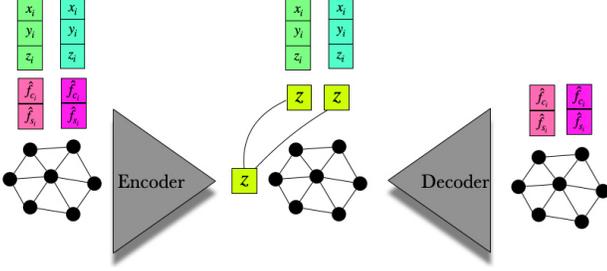
Figure 3: cVAE architecture. For each vertex on the body mesh, we concatenate the vertex positions x_i, y_i, z_i, the contact label f_c, and the corresponding semantic scene label f_s. The latent vector z is concatenated to the vertex positions, and the result passes to the decoder which reconstructs the input features $\hat{f}_c, \hat{f}_s$.

of neighboring nodes, $S(i, l)$. The spiral sequence $S(i, l)$ is an ordered set of l vertices around the central vertex i. Our architecture is shown in Fig. 3. More implementation details are included in the Sup. Mat. For details on selecting and ordering vertices, please see [19].

4. Experiments

We perform several experiments to investigate the effectiveness and usefulness of our proposed representation and model under different use cases, namely generating HSI features, automatically placing 3D people in scenes, and improving monocular 3D human pose estimation.

4.1. Random Sampling

We evaluate the generative power of our model by sampling different feature maps conditioned on novel poses using our trained decoder $P(f_{\text{Gen}}|z, V_b)$, where $z \sim \mathcal{N}(0, I)$ and f_{Gen} is the randomly generated feature map. This is equivalent to answering the question: "In this given pose, which vertices on the body are likely to be in contact with the scene, and what object would they contact?" Randomly generated samples are shown in Fig. 4.

We observe that our model generalizes well to various poses. For example, notice that when a person is standing with one hand pointing forward, our model predicts the feet and the hand to be in contact with the scene. It also predicts the feet are in contact with the floor and hand is in contact with the wall. However this changes for the examples when a person is in a lying pose. In this case, most of the vertices from the back of the body are predicted to be in contact (blue color) with a bed (light purple) or a sofa (dark green).

These features are predicted from the body alone; there is no notion of "action" here. Pose alone is a powerful predictor of interaction. Since the model is probabilistic, we can sample many possible feature maps for a given pose.
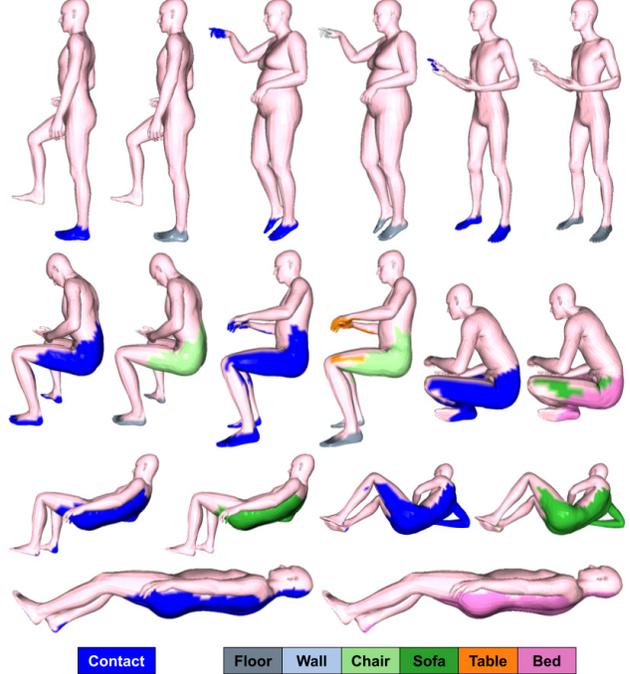

Figure 4: Random samples from our trained cVAE. For each example (image pair) we show from left to right: f_c and f_s. The color code is at the bottom. For f_c, blue means contact, while pink means no contact. For f_s, each scene category has a different color.

4.2. Affordances: Putting People in Scenes

Given a posed 3D body and a 3D scene, can we place the body in the scene so that the pose makes sense in the context of the scene? That is does the pose match the affordances of the scene [20, 30, 32]? Specifically, given a scene, M_s, semantic labels of objects present, and a body mesh, M_b, our method finds where in M_s this given pose is likely to happen. We solve this problem in two steps.

First, given the posed body, we use the decoder of our cVAE to generate a feature map by sampling $P(f_{\text{Gen}}|z, V_b)$ as in Sec. 4.1. Second, we optimize the objective function:

$$E(\tau, \theta_0, \theta) = \mathcal{L}_{afford} + \mathcal{L}_{pen} + \mathcal{L}_{reg}, \qquad (9)$$

where τ is the body translation, θ_0 is the global body orientation and θ is the body pose. The afforance loss $\mathcal{L}_{afford} =$

$$\lambda_1 * ||f_{\text{Gen}_c} \cdot f_d||_2^2 + \lambda_2 * \sum_i \text{CCE}\left(f_{\text{Gen}_s}^i, f_s^i\right), \qquad (10)$$

f_d and f_s are the observed distance and semantic labels, which are computed using Eq. 2 and Eq. 4 respectively. f_{Gen_c} and f_{Gen_s} are the generated contact and semantic labels, and $\cdot$ denotes dot product. λ_1 and λ_2 are 1 and 0.01 respectively. $\mathcal{L}_{pen}$ is a penetration penalty to discourage the
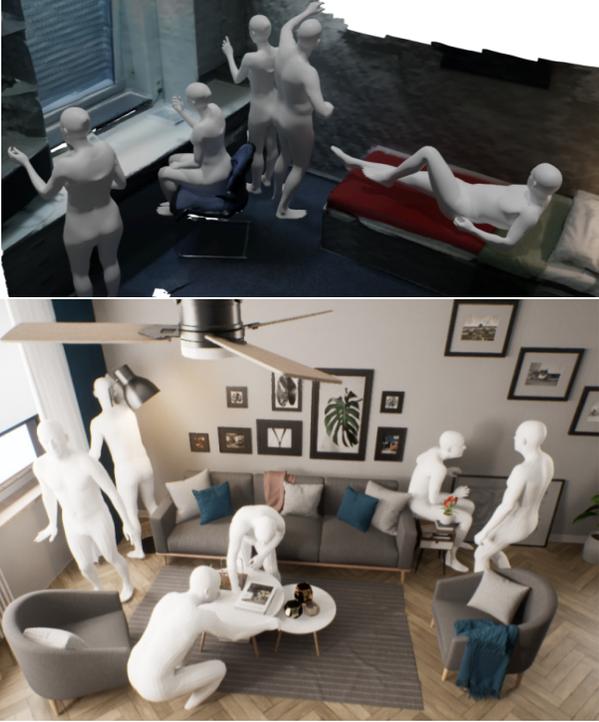
Figure 5: **(Top):** SMPL-X meshes automatically placed in a real scene from the PROX test set. The body shapes and poses here are drawn from the PROX test set and were not used in training. **(Bottom):** SMPL-X meshes automatically placed in a synthetic scene.

body from penetrating the scene:

$$\mathcal{L}_{pen} = \lambda_{pen} * \sum_{f_d^i < 0} \left(f_d^i \right)^2. \tag{11}$$

$\lambda_{pen} = 10$. $\mathcal{L}_{reg}$ is a regularizer that encourages the estimated pose to remain close to the initial pose θ_{init} of M_b:

$$\mathcal{L}_{reg} = \lambda_{reg} * ||\theta - \theta_{\text{init}}||_2^2. \tag{12}$$

Although the body pose is given, we optimize over it, allowing the θ parameters to change slightly since the given pose θ_{init} might not be well supported by the scene. This allows for small pose adjustment that might be necessary to better fit the body into the scene. $\lambda_{reg} = 100$.

The input posed mesh, M_b, can come from any source. For example, we can generate random SMPL-X meshes using VPoser [46] which is a VAE trained on a large dataset of human poses. More interestingly, we use SMPL-X meshes fit to realistic Renderpeople scans [45] (see Fig. 1).

We tested our method with both real (scanned) and synthetic (artist generated) scenes. Example bodies optimized to fit in a real scene from the PROX [22] test set are shown in Fig. 5 (top); this scene was not used during training. Note that people appear to be interacting naturally with the

Figure 6: Unmodified clothed bodies (from Renderpeople) automatically placed in real scenes from the Replica dataset.

scene; that is, their pose matches the scene context. Figure 5 (bottom) shows bodies automatically placed in an artist-designed scene (Archviz Interior Rendering Sample, Epic Games)[1]. POSA goes beyond previous work [20, 30, 65] to produce realistic human-scene interactions for a wide range of poses like lying down and reaching out.

While the poses look natural in the above results, the SMPL-X bodies look out of place in realistic scenes. Consequently, we would like to render realistic people instead, but models like SMPL-X do not support realistic clothing and textures. In contrast, scans from companies like Renderpeople (Renderpeople GmbH, Köln) are realistic, but have a different mesh topology for every scan. The consistent topology of a mesh like SMPL-X is critical to learn the feature model.

Clothed Humans: We address this issue by using SMPL-X fits to clothed meshes from the AGORA dataset [45]. We then take the SMPL-X fits and minimize an energy function similar to Eq. 9 with one important change. We keep the pose, θ, fixed:

$$E(\tau, \theta_0) = \mathcal{L}_{afford} + \mathcal{L}_{pen}. \tag{13}$$

Since the pose does not change, we just replace the SMPL-X mesh with the original clothed mesh after the optimization converges; see Sup. Mat. for details.

Qualitative results for real scenes (Replica dataset [56]) are shown in Fig. 6, and for a synthetic scene in Fig. 1. More results are shown in Sup. Mat. and in our video.

4.2.1 Evaluation

We quantiatively evaluate POSA with two perceptual studies. In both, subjects are shown a pair of two rendered scenes, and must choose the one that best answers the question "Which one of the two examples has more realistic (i.e. natural and physically plausible) human-scene interaction?" We also evaluate physical plausibility and diversity.

---

[1]https://docs.unrealengine.com/en-US/Resources/Showcases/ArchVisInterior/index.html

	Generation ↑	PROX GT ↓
PLACE [64]	48.5%	51.5%
POSA (contact only)	46.9%	53.1%
POSA (contact + semantics)	**49.1%**	50.1%
POSA-*clothing* (contact)	55.0%	45.0%
POSA-*clothing* (semantics)	**60.6%**	39.4%

Table 1: Comparison to PROX [22] ground truth. Subjects are shown pairs of a generated 3D human-scene interaction and PROX ground truth (GT), and must chose the most realistic one. A higher percentage means that subjects deemed this method more realistic.

	POSA-variant ↑	PLACE ↓
POSA (contact only)	60.7%	39.3%
POSA (contact + semantics)	**61.0%**	39.0%

Table 2: POSA compared to PLACE for 3D human-scene interaction generation. See Tab. 1 caption.

Comparison to PROX ground truth: We follow the protocol of Zhang et al. [64] and compare our results to randomly selected examples from PROX ground truth. We take 4 real scenes from the PROX [22] test set, namely MPH16, MPH1Library, N0SittingBooth and N3OpenArea. We take 100 SMPL-X bodies from the AGORA [45] dataset, corresponding to 100 different 3D scans from Renderpeople. We take each of these bodies and sample one feature map for each, using our cVAE. We then automatically optimize the placement of each sample in all the scenes, one body per scene. For unclothed bodies (Tab. 1, rows 1-3), this optimization changes the pose slightly to fit the scene (Eq. 9). For clothed bodies (Tab. 1, rows 4-5), the pose is kept fixed (Eq. 13). For each variant, this optimization results in 400 unique body-scene pairs. We render each 3D human-scene interaction from 2 views so that subjects are able to get a good sense of the 3D relationships from the images. Using Amazon Mechanical Turk (AMT), we show these results to 3 different subjects. This results in 1200 unique ratings. The results are shown in Tab. 1. POSA (contact only) and the state-of-the-art method PLACE [64] are both almost indistinguishable from the PROX ground truth. However, the proposed POSA (contact + semantics) (row 3) outperforms both POSA (contact only) (row 2) and PLACE [64] (row 1), thus modeling scene semantics increases realism. Lastly, the rendering of high quality clothed meshes (bottom two rows) influences the perceived realism significantly.

Comparison between POSA and PLACE: We follow the same protocol as above, but this time we directly compare POSA and PLACE. The results are shown in Tab. 2. Again, we find that adding semantics improves realism. There are likely several reasons that POSA is judged more realistic than PLACE. First, POSA employs denser contact information across the whole SMPL-X body surface, com-

	Non-Collision ↑	Contact ↑
PSI [65]	0.94	0.99
PLACE [64]	**0.98**	0.99
POSA (contact only)	0.97	**1.0**
POSA (contact + semantics)	0.97	0.99

Table 3: Evaluation of the physical plausibility metric. Arrows indicate that higher scores are better.

pared to PLACE's sparse distance information through its BPS representation. Second, POSA uses a human-centric formulation, as opposed to PLACE's scene-centric one, and this can help generalize across scenes better. Third, POSA uses semantic features that help bodies do the right thing in the scene, while PLACE does not. When human generation is imperfect, inappropriate semantics may make the result seem worse. Fourth, the two methods are solving slightly different tasks. PLACE generates a posed body mesh for a given scene, while our method samples one from the AGORA dataset and places it in the scene using a generated POSA feature map. While this gives PLACE an advantage, because it can generate an appropriate pose for the scene, it also means that it could generate an unnatural pose, hurting realism. In our case, the poses are always "valid" by construction but may not be appropriate for the scene. Note that, while more realistic than prior work, the results are not always fully natural; sometimes people sit in strange places or lie where they usually would not.

Physical Plausibility: We take 1200 bodies from the AGORA [45] dataset and place all of them in each of the 4 test scenes of PROX, leading to a total of 4800 samples, following [64, 65]. Given a generated body mesh, M_b, a scene mesh, M_s, and a scene signed distance field (SDF) that stores distances d_j for each voxel j, we compute the following scores, defined by Zhang et al. [65]: (1) the *non-collision score* for each M_b, which is the ratio of body mesh vertices with positive SDF values divided by the total number of SMPL-X vertices, and (2) the *contact score* for each M_b, which is 1 if at least one vertex of M_b has a non-positive value. We report the mean non-collision score and mean contact score over all 4800 samples in Tab. 3; higher values are better for both metrics. POSA and PLACE are comparable under these metrics.

Diversity Metric: Using the same 4800 samples, we compute the diversity metric from [65]. We perform K-means ($k = 20$) clustering of the SMPL-X parameters of all sampled poses, and report: (1) the entropy of the cluster sizes, and (2) the cluster size, i.e. the average distance between the cluster center and the samples belonging in it. See Tab. 4; higher values are better. While PLACE generates poses and POSA samples them from a database, there is little difference in diversity.

	Entropy ↑	Cluster Size ↑
PSI [65]	**2.97**	2.53
PLACE [64]	2.91	**2.72**
POSA (contact only)	2.94	2.28
POSA (contact + semantics)	2.92	2.27

Table 4: Evaluation of the diversity metric. Arrows indicate that higher scores are better.

4.3. Monocular Pose Estimation with HSI

Traditionally, monocular pose estimation methods focus only on the body and ignore the scene. Hence, they tend to generate bodies that are inconsistent with the scene. Here, we compare directly with PROX [22], which adds contact and penetration constraints to the pose estimation formulation. The contact constraint snaps a fixed set of contact points on the body surface to the scene, if they are "close enough". In PROX, however, these contact points are manually selected and are independent of pose.

We replace the hand-crafted contact points of PROX with our learned feature map. We fit SMPL-X to RGB image features such that the contacts are consistent with the 3D scene and its semantics. Similar to PROX, we build on SMPLify-X [46]. Specifically, SMPLify-X optimizes SMPL-X parameters to minimize an objective function of multiple terms: the re-projection error of 2D joints, priors and physical constraints on the body; $E_{\text{SMPLify-X}}(\beta, \theta, \psi, \tau) =$

$$E_J + \lambda_\theta E_\theta + \lambda_\alpha E_\alpha + \lambda_\beta E_\beta + \lambda_{\mathcal{P}} E_{\mathcal{P}} \quad (14)$$

where θ represents the pose parameters of the body, face (neck, jaw) and the two hands, $\theta = \{\theta_b, \theta_f, \theta_h\}$, τ denotes the body translation, and β the body shape. E_J is a re-projection loss that minimizes the difference between 2D joints estimated from the RGB image I and the 2D projection of the corresponding posed 3D joints of SMPL-X. $E_\alpha(\theta_b) = \sum_{i \in (elbows, knees)} \exp(\theta_i)$ is a prior penalizing extreme bending only for elbows and knees. The term $E_{\mathcal{P}}$ penalizes self-penetrations. For details please see [46].

We turn off the PROX contact term and optimize Eq. 14 to get a pose matching the image observations and roughly obeying scene constraints. Given this rough body pose, which is not expected to change significantly, we sample features from $P(f_{\text{Gen}}|z, V_b)$ and keep these fixed. Finally, we refine by minimizing $E(\beta, \theta, \psi, \tau, M_s) =$

$$E_{\text{SMPLify-X}} + ||f_{\text{Gen}_c} \cdot f_d|| + \mathcal{L}_{pen} \quad (15)$$

where $E_{\text{SMPLify-X}}$ represents the SMPLify-X energy term as defined in Eq. 14, f_{Gen_c} are the generated contact labels, f_d is the observed distance, and $\mathcal{L}_{pen}$ represents the body-scene penetration loss as in Eq. 11. We compare our results to standard PROX in Tab. 5. We also show the results

(mm)	PJE ↓	V2V ↓	p.PJE ↓	p.V2V ↓
RGB	220.27	218.06	73.24	60.80
PROX	167.08	166.51	**71.97**	**61.14**
POSA	**154.33**	**154.84**	73.17	63.23

Table 5: Pose estimation results for PROX and POSA. *PJE* is the mean per-joint error and *V2V* is the mean vertex-to-vertex Euclidean distance between meshes (after only pelvis joint alignment). The prefix "p" means that the error is computed after Procrustes alignment to the ground truth; this hides many errors, making the methods comparable.

of RGB-only baseline introduced in PROX for reference. Using our learned feature map improves accuracy over the PROX's heuristically determined contact constraints.

5. Conclusions

Traditional 3D body models, like SMPL-X, model the a priori probability of possible body shapes and poses. We argue that human poses in isolation from the scene, make little sense. We introduce POSA, which effectively upgrades a 3D human body model to explicitly represent possible human-scene interactions. Our novel, body-centric, representation encodes the contact and semantic relationships between the body and the scene. We show that this is useful and supports new tasks. For example, we consider placing a 3D human into a 3D scene. Given a scan of a person with a known pose, POSA allows us to search the scene for locations where the pose is likely. This enables us to populate empty 3D scenes with higher realism than the state of the art. We also show that POSA can be used for estimating human pose from an RGB image, and that the body-centered HSI representation improves accuracy. In summary, POSA is a good step towards a richer model of human bodies that goes beyond pose to support the modeling of HSI.

Limitations: POSA requires an accurate scene SDF; a noisy scene mesh can lead to penetration between the body and scene. POSA focuses on a single body mesh only. Penetration between clothing and the scene is not handled and multiple bodies are not considered. Optimizing the placement of people in scenes is sensitive to initialization and is prone to local minima. A simple user interface would address this, letting naive users roughly place bodies, and then POSA would automatically refine the placement.

References

[1] Rami Ali Al-Asqhar, Taku Komura, and Myung Geol Choi. Relationship descriptors for interactive motion adaptation. In *Symposium on Computer Animation (SCA)*, pages 45–53, 2013.

[2] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: Shape Completion and Animation of PEople. *Transactions on Graphics (TOG), Proceedings SIGGRAPH*, 24(3):408–416, 2005.

[3] Iro Armeni, Sasha Sax, Amir R. Zamir, and Silvio Savarese. Joint 2D-3D-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.

[4] Ezer Bar-Aviv and Ehud Rivlin. Functional 3D object classification using simulation of embodied agent. In *British Machine Vision Conference (BMVC)*, pages 307–316, 2006.

[5] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: Registering human bodies in motion. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5573–5582, 2017.

[6] Giorgos Bouritsas, Sergiy Bokhnyak, Stylianos Ploumpis, Michael Bronstein, and Stefanos Zafeiriou. Neural 3D morphable models: Spiral convolutional networks for 3D shape representation learning and generation. In *International Conference on Computer Vision (ICCV)*, pages 7212–7221, 2019.

[7] Samarth Brahmbhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. ContactPose: A dataset of grasps with object contact and hand pose. In *European Conference on Computer Vision (ECCV)*, volume 12358, pages 361–378, 2020.

[8] Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. Benchmarking in manipulation research: Using the Yale-CMU-Berkeley object and model set. *Robotics & Automation Magazine (RAM)*, 22(3):36–52, 2015.

[9] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qizhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *European Conference on Computer Vision (ECCV)*, volume 12346, pages 387–404, 2020.

[10] 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.

[11] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. ShapeNet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[12] Yixin Chen, Siyuan Huang, Tao Yuan, Yixin Zhu, Siyuan Qi, and Song-Chun Zhu. Holistic++ scene understanding: Single-view 3D holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *International Conference on Computer Vision (ICCV)*, pages 8647–8656, 2019.

[13] Sungjoon Choi, Qian-Yi Zhou, Stephen Miller, and Vladlen Koltun. A large dataset of object scans. *arXiv preprint arXiv:1602.02481*, 2016.

[14] Enric Corona, Albert Pumarola, Guillem Alenyà, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6990–6999, 2020.

[15] Enric Corona, Albert Pumarola, Guillem Alenyà, Francesc Moreno-Noguer, and Grégory Rogez. GanHand: Predicting human grasp affordances in multi-object scenes. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5030–5040, 2020.

[16] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2443, 2017.

[17] Matthew Fisher, Manolis Savva, Yangyan Li, Pat Hanrahan, and Matthias Nießner. Activity-centric scene synthesis for functional 3D scene modeling. *Transactions on Graphics (TOG), Proceedings SIGGRAPH Asia*, 34(6):179:1–179:13, 2015.

[18] Michael Gleicher. Retargeting motion to new characters. In *Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 33–42, 1998.

[19] 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.

[20] Helmut Grabner, Juergen Gall, and Luc Van Gool. What makes a chair a chair? In *Computer Vision and Pattern Recognition (CVPR)*, pages 1529–1536, 2011.

[21] Abhinav Gupta, Scott Satkin, Alexei A Efros, and Martial Hebert. From 3D scene geometry to human workspace. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1961–1968, 2011.

[22] 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.

[23] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11807–11816, 2019.

[24] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017.

[25] Edmond S. L. Ho, Taku Komura, and Chiew-Lan Tai. Spatial relationship preserving character motion adaptation. *Transactions on Graphics (TOG), Proceedings SIGGRAPH*, 29(4):33:1–33:8, 2010.

[26] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments.

Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 36(7):1325–1339, 2014.

[27] Yun Jiang, Hema Koppula, and Ashutosh Saxena. Hallucinated humans as the hidden context for labeling 3D scenes. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2993–3000, 2013.

[28] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8320–8329, 2018.

[29] Changgu Kang and Sung-Hee Lee. Environment-adaptive contact poses for virtual characters. *Computer Graphics Forum (CGF)*, 33(7):1–10, 2014.

[30] Vladimir G Kim, Siddhartha Chaudhuri, Leonidas Guibas, and Thomas Funkhouser. Shape2Pose: Human-centric shape analysis. *Transactions on Graphics (TOG), Proceedings SIGGRAPH*, 33(4):120:1–120:12, 2014.

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

[32] Hedvig Kjellström, Javier Romero, and Danica Kragić. Visual object-action recognition: Inferring object affordances from human demonstration. *Computer Vision and Image Understanding (CVIU)*, 115(1):81–90, 2011.

[33] 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.

[34] Kang Hoon Lee, Myung Geol Choi, and Jehee Lee. Motion patches: Building blocks for virtual environments annotated with motion data. *Transactions on Graphics (TOG), Proceedings SIGGRAPH*, 25(3):898–906, 2006.

[35] Kurt Leimer, Andreas Winkler, Stefan Ohrhallinger, and Przemyslaw Musialski. Pose to seat: Automated design of body-supporting surfaces. *Computer Aided Geometric Design (CAGD)*, 79:101855, 2020.

[36] Xueting Li, Sifei Liu, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. Putting humans in a scene: Learning affordance in 3D indoor environments. In *Computer Vision and Pattern Recognition (CVPR)*, pages 12368–12376, 2019.

[37] Juncong Lin, Takeo Igarashi, Jun Mitani, Minghong Liao, and Ying He. A sketching interface for sitting pose design in the virtual environment. *Transactions on Visualization and Computer Graphics (TVCG)*, 18(11):1979–1991, 2012.

[38] Zhenbao Liu, Caili Xie, Shuhui Bu, Xiao Wang, Junwei Han, Hongwei Lin, and Hao Zhang. Indirect shape analysis for 3D shape retrieval. *Computers & Graphics (CG)*, 46:110–116, 2015.

[39] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *Transactions on Graphics (TOG), Proceedings SIGGRAPH Asia*, 34(6):248:1–248:16, 2015.

[40] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*, pages 5441–5450, 2019.

[41] Thomas B. Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding (CVIU)*, 104(2-3):90–126, 2006.

[42] Aron Monszpart, Paul Guerrero, Duygu Ceylan, Ersin Yumer, and Niloy J Mitra. iMapper: interaction-guided scene mapping from monocular videos. *Transactions on Graphics (TOG), Proceedings SIGGRAPH*, 38(4):92:1–92:15, 2019.

[43] Arsalan Mousavian, Clemens Eppner, and Dieter Fox. 6-DOF GraspNet: Variational grasp generation for object manipulation. In *International Conference on Computer Vision (ICCV)*, pages 2901–2910, 2019.

[44] Ahmed A. A. Osman, Timo Bolkart, and Michael J. Black. STAR: Sparse trained articulated human body regressor. In *European Conference on Computer Vision (ECCV)*, volume 12351, pages 598–613, 2020.

[45] 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.

[46] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[47] Sören Pirk, Vojtech Krs, Kaimo Hu, Suren Deepak Rajasekaran, Hao Kang, Yusuke Yoshiyasu, Bedrich Benes, and Leonidas J Guibas. Understanding and exploiting object interaction landscapes. *Transactions on Graphics (TOG)*, 36(3):31:1–31:14, 2017.

[48] Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *International Conference on Computer Vision (ICCV)*, pages 4331–4340, 2019.

[49] 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.

[50] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *Transactions on Graphics (TOG), Proceedings SIGGRAPH Asia*, 36(6):245:1–245:17, 2017.

[51] Nikolaos Sarafianos, Bogdan Boteanu, Bogdan Ionescu, and Ioannis A. Kakadiaris. 3D human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding (CVIU)*, 152:1–20, 2016.

[52] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. SceneGrok: Inferring action maps in 3D environments. *Transactions on Graphics (TOG), Proceedings SIGGRAPH Asia*, 33(6):212:1–212:10, 2014.

[53] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. PiGraphs: Learning interaction snapshots from observations. *Transactions on Graphics (TOG), Proceedings SIGGRAPH*, 35(4):139:1–139:12, 2016.

[54] Leonid Sigal, Alexandru Balan, and Michael J Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision (IJCV)*, 87(1-2):4–27, 2010.

[55] Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas A. Funkhouser. Semantic scene completion from a single depth image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 190–198, 2017.

[56] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

[57] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, volume 12349, pages 581–600, 2020.

[58] He Wang, Sören Pirk, Ersin Yumer, Vladimir Kim, Ozan Sener, Srinath Sridhar, and Leonidas Guibas. Learning a generative model for multi-step human-object interactions from videos. *Computer Graphics Forum (CGF)*, 38(2):367–378, 2019.

[59] Xiaolong Wang, Rohit Girdhar, and Abhinav Gupta. Binge watching: Scaling affordance learning from sitcoms. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3366–3375, 2017.

[60] Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson Env: Real-world perception for embodied agents. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9068–9079, 2018.

[61] Yu Xiang, Wonhui Kim, Wei Chen, Jingwei Ji, Christopher Choy, Hao Su, Roozbeh Mottaghi, Leonidas Guibas, and Silvio Savarese. ObjectNet3D: A large scale database for 3D object recognition. In *European Conference on Computer Vision (ECCV)*, volume 9912, pages 160–176, 2016.

[62] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D human shape and articulated pose models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6183–6192, 2020.

[63] Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3D human-object spatial arrangements from a single image in the wild. In *European Conference on Computer Vision (ECCV)*, volume 12357, pages 34–51, 2020.

[64] 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.

[65] 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.

[66] Youyi Zheng, Han Liu, Julie Dorsey, and Niloy J Mitra. Ergonomics-inspired reshaping and exploration of collections of models. *Transactions on Visualization and Computer Graphics (TVCG)*, 22(6):1732–1744, 2016.

[67] Yixin Zhu, Chenfanfu Jiang, Yibiao Zhao, Demetri Terzopoulos, and Song-Chun Zhu. Inferring forces and learning human utilities from videos. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3823–3833, 2016.

[68] Michael Zollhöfer, Patrick Stotko, Andreas Görlitz, Christian Theobalt, Matthias Nießner, Reinhard Klein, and Andreas Kolb. State of the art on 3D reconstruction with RGB-D cameras. *Computer Graphics Forum (CGF)*, 37(2):625–652, 2018.