COAP: Compositional Articulated Occupancy of People

Marko Mihajlovic[1], Shunsuke Saito[2], Aayush Bansal[2], Michael Zollhoefer[2], Siyu Tang[1]
[1]ETH Zürich [2]Reality Labs Research

neuralbodies.github.io/COAP

Figure 1. We present **COAP**, a **CO**mpositional **A**rticulated occupancy representation of **P**eople that is robust to highly articulated out-of-distribution pose sequences and that generalizes well to novel identities. The key idea is to decompose a human body into articulated body parts and employ a part-aware encoder-decoder architecture to learn neural articulated occupancy that models complex deformations locally. The displayed example visualizes unseen subjects performing challenging poses from the PosePrior dataset [1]. The color scheme indicates labels of the activated articulated local implicit representations.

Abstract

We present a novel neural implicit representation for articulated human bodies. Compared to explicit template meshes, neural implicit body representations provide an efficient mechanism for modeling interactions with the environment, which is essential for human motion reconstruction and synthesis in 3D scenes. However, existing neural implicit bodies suffer from either poor generalization on highly articulated poses or slow inference time. In this work, we observe that prior knowledge about the human body's shape and kinematic structure can be leveraged to improve generalization and efficiency. We decompose the full-body geometry into local body parts and employ a part-aware encoder-decoder architecture to learn neural articulated occupancy that models complex deformations locally. Our local shape encoder represents the body deformation of not only the corresponding body part but also the neighboring body parts. The decoder incorporates the geometric constraints of local body shape which significantly improves pose generalization. We demonstrate that our model is suitable for resolving self-intersections and collisions with 3D environments. Quantitative and qualitative experiments show that our method largely outperforms existing solutions in terms of both efficiency and accuracy.

1. Introduction

Computers can perceive rich representations of 3D human pose, shape, and motion by regressing the latent parameters of parametric human body models [25, 35, 54]. Conventionally, such generative human body models are represented as polygonal meshes and are easy to deform and animate by leveraging skinning algorithms such as linear blend skinning (LBS) [14]. However, they are not well suited for efficient interactions with 3D graphics environment and resolving self-intersections.

Unlike meshes, neural implicit representations [6, 28, 48] are flexible, continuous, and support efficient intersection tests with the environment. The state-of-the-art neural implicit body models [28, 42, 52] learn an inverse LBS network to convert an arbitrary point in 3D space to the canonical space where identity- and pose-dependent surface deformations are modeled. While being effective in capturing surface deformations in the canonical space, the learned inverse LBS networks often suffer from poor generalization capability to highly articulated unseen poses (Fig. 1). SNARF [6] circumvents the need of learning the inverse LBS network by formulating the inverse mapping as a root-finding problem. However, the model is learned per subject, and the computationally expensive root-finding prevents the practical application of their method for human body reconstruction in 3D scenes. In this work, we present a novel part-aware encoder-decoder architecture that mod-

els compositional neural occupancy representations which are robust, efficient, and can generalize to a large variety of body shapes and highly articulated body poses. We name it COAP (**CO**mpositional **A**rticulated occupancy of **P**eople).

COAP is inspired by two key insights: First, the learned inverse LBS function in LEAP [28] captures spurious long-range correlations, making it hard to generalize to highly articulated unseen poses. To address this, we get rid of the learned LBS and propose a novel local shape encoding that models the neural occupancy of articulated body parts by using a localized context of direct neighbors in the kinematic chain. This localized way of representing the body and its deformations reduces overfitting to the spurious correlations in the training set. Furthermore, given the local part encoding, the final whole human body is represented as a composition of these predicted local neural fields. Instead of a simple per-part combination as in NASA [8], each local encoding in COAP contributes not only to the corresponding body part but also to the deformations of the neighboring body parts. Overall, the compositional neural fields modeled by the part-aware encoder-decoder architecture are effective and greatly benefit generalization (Sec. 5).

Second, prior knowledge about human shapes that is carried by the parametric body models can significantly ease the task of learning robust neural representations. Similar to LEAP [28], we use SMPL [25] as the starting point. Given the input bone transformations, we can effectively extract the relevant local body vertex positions. We leverage the per-part body vertices to create simple geometric primitives (such as 3D boxes) and incorporate them in the neural network architecture. This can be considered as a geometric prior of a local body shape which simplifies the learning problem and helps the neural network to properly allocate its modeling capacity around the surface. As demonstrated in our experiment, the effective fusion of the geometric prior and the learning power of neural networks is vital for the generalization capability of the learned representations.

We systematically evaluate the robustness and the representation power of COAP. We compare with SNARF [6] that is trained per subject and shows impressive results on unseen poses [1]. COAP achieves even better performance while at the same time being more efficient in terms of inference time. We also compare with LEAP [28] and NeuralGIF [48] that produce generalizable neural implicit bodies. Once again, COAP significantly outperforms their results on the PosePrior [1] and the DFaust [5] datasets.

Resolving self-interpenetration of deformable 3D shapes is challenging and has been a long-standing question in computer graphics and vision [4, 11, 18, 35, 37, 40, 50]. We propose a simple, yet effective optimization algorithm based on COAP that can efficiently resolve self-interpenetration among different body parts. Our method can reliably solve the challenging cases that are not addressed by existing solutions [35] (as shown in Sec. 5.3). Furthermore, we demonstrate the utility of COAP for resolving collisions with 3D environments. Prior work [13, 57] requires pre-computed signed distance fields (SDFs) of 3D scenes to perform collision detection between 3D human bodies and the scene geometry, which is cumbersome and does not scale to scenes with moving objects or humans. Our robust and generalizable neural body model can be used to directly detect collisions with raw scans to improve 3D pose and shape estimation (Sec. 5.3).

Contributions. In summary, our main contributions are: (1) a novel neural implicit body model that is robust and efficient, and can generalize to a large variety of human shapes and highly articulated body poses; (2) an effective localized encoder-decoder architecture that leverages local shape encoding and geometric shape priors to learn compositional neural body representations; and, (3) simple and efficient optimization algorithms that reliably resolve challenging self-interpenetration and human-scene interpenetration. Code and models are public[1].

2. Related Work

2.1. Parametric Body Representations

Parametric body models [25, 32, 41, 54] consist of a template mesh with an underlying kinematic skeleton. To animate a body, the canonical skeleton is reposed via forward kinematics and the mesh vertices are deformed by a skinning algorithm [17, 19, 23]. Popular data-driven models such as SMPL [25] and GHUM [54] use the Linear Blend Skinning (LBS) algorithm to deform mesh vertices as a weighted sum of several rigid body part transformations. While human meshes are ubiquitous in computer graphics due to their good animation and rendering properties, they often self-intersect [50] when a human body is reposed, and they are further not suitable for testing interactions with the environment. These two properties are essential for many human-scene interaction applications [13, 39, 55] and registration pipelines [3, 51] which often generate ill-defined models that self-intersect or collide with other objects. We address these two critical problems with our compositional neural implicit representation.

Resolving Self-intersections. Mesh self-intersection is a common problem in computer graphics that occurs when a human body mesh is reposed. To address this problem, most prior techniques [24, 29, 30, 44] build an intermediate volumetric representation (*e.g.* tetrahedral mesh) at every animation step and require an expensive optimization procedure to untangle self-intersecting bodies, which makes them unsuitable for image-based human reconstruction tasks [16, 35, 45]. More efficient methods tailored

[1] neuralbodies.github.io/COAP

for human bodies optimize human pose to resolve self-intersections. Guan *et al*. [11, 12] model each body part by their convex hull, which is in turn employed to create a differentiable penalty function for interpenetrated body parts. Since such an approach imposes a computationally expensive optimization problem, other works have proposed to alleviate the computation bottleneck by over-approximating body parts with simple geometric proxies (*e.g*. spheres [37] or capsules [4]) to compute a differentiable interpenetration term efficiently. A more precise approach has been proposed in [35, 49], which detects and penalizes self-intersected mesh triangles using a BVH tree [46]. However, such a loss term imposes a discretized surface-based error that is prone to local minima, whereas our method is volume-aware and imposes a more robust continuous penalty.

Resolving Collisions with the Environment. Modeling interactions of articulated parametric human bodies with raw scans or other geometries is a hard task. A common approach is to convert raw scans into meshes and penalize collided triangles [49]. However, such methods impose a computationally expensive surface-based loss and are computationally expensive for more complex scenes. Hence most prior works [13, 55–57] circumvent this problem by calculating SDF grids of raw scans, which is an error-prone task and not always possible [15]. Similarly, [16] propose to detect collisions between two human body meshes by dynamically calculating 3D SDF grids, which is memory and computationally expensive (≈25s for 256^3 grids) and erroneous when the meshes self-intersect. Our method circumvents these problems by representing a parametric human body as a volumetric representation that enables efficient differentiable collision checks with other geometries represented by meshes or point clouds.

2.2. Neural Implicit Representations.

Neural implicit representations [7, 27, 34, 36, 53] enable efficient inside/outside tests by representing shapes as signed-distance or occupancy functions parameterized by neural network weights. However, most of these representations are designed for rigid objects and cannot represent highly-articulated humans.

Neural Implicit Bodies. Analogously to mesh-based body models, several recent works [6, 28, 33, 42, 43, 52] have proposed to learn neural implicit bodies. They simplify the learning problem by modeling neural representations in canonical space. NASA [8] learns a subject-specific part-based occupancy representation that is composed via rigid bone transformations in a posed space. However, the composition introduces artifacts around joints, and their low-dimensional pose encoding does not fully remove long-range spurious correlations. LEAP [28] and NeuralGIF [48] propose to learn a generalizable neural implicit human body model in a canonical space and a separate inverse LBS neural network that projects any given query point to the canonical space where reliable occupancy checks are performed. Similarly, SCANimate [42] and MetaAvatar [52] learn subject-specific avatars in a canonical space and an inverse LBS neural network to deform the surface points. These methods alleviate the problem of the artifacts around the joints presented in NASA [8]. However, the learned inverse LBS is less robust to novel motions. imGHUM [2] employs a multi-part model and learns an implicit human representation directly in the posed space. SNARF [6] learns a subject-specific model in a canonical pose, but it circumvents the need for an inverse LBS network by formulating the inverse mapping as a root-finding problem. However, it suffers from computationally expensive inference and requires per subject training, which makes it less suitable for many practical applications. Compared to existing representations, our model better generalizes to novel motions and identities. This is achieved by learning the implicit fields for articulated body parts and leveraging geometric priors and localized encoders that reduce the overfitting caused by spurious correlations.

3. Fundamentals

Modeling Human Bodies. A parametric body model such as SMPL [25] is a data-driven model that is controlled via shape parameters β and pose parameters $\theta \in \mathbb{R}^{K\times 3}$, where K is the number of articulated joints. It builds a human mesh in canonical pose $\bar{V}$ by deforming a pre-defined template mesh $\bar{T}$ via identity-dependent $\mathsf{B}_S(\beta)$ and pose-dependent $\mathsf{B}_P(\theta)$ vertex correctives:

$$\bar{V} = \bar{T} + \mathsf{B}_S(\beta) + \mathsf{B}_P(\theta)\,. \tag{1}$$

After this step, a skeleton composed of joint locations $\mathbf{J} \in \mathbb{R}^{K\times 3}$ in the canonical space is regressed by a learned matrix $\mathcal{J}$:

$$\mathbf{J} = \mathcal{J}(\bar{T} + \mathsf{B}_S(\beta))\,. \tag{2}$$

Reposing. To animate a human body, the skeleton $\mathbf{J}$ in the canonical pose is reposed via the forward kinematics and can be compactly represented by a set of rigid bone transformation matrices $\mathcal{G} = [G_k]_{k=1}^K$ as

$$G_k(\theta, \mathbf{J}) = \prod_{j\in A(k)} \left[\begin{array}{c|c} R(\theta_j) & \mathbf{J}_j \\ \hline \vec{0} & 1 \end{array}\right], \tag{3}$$

where the rotation and the translation parts correspond to the bone orientation and the joint location, respectively. R transforms pose parameters of the part j into rotation matrix, and $A(k)$ defines a kinematic tree as an ordered set of ancestors of the joint k.

Analogously to reposing of the canonical skeleton, the canonical mesh vertices $\bar{V}$ are deformed via the linear blend

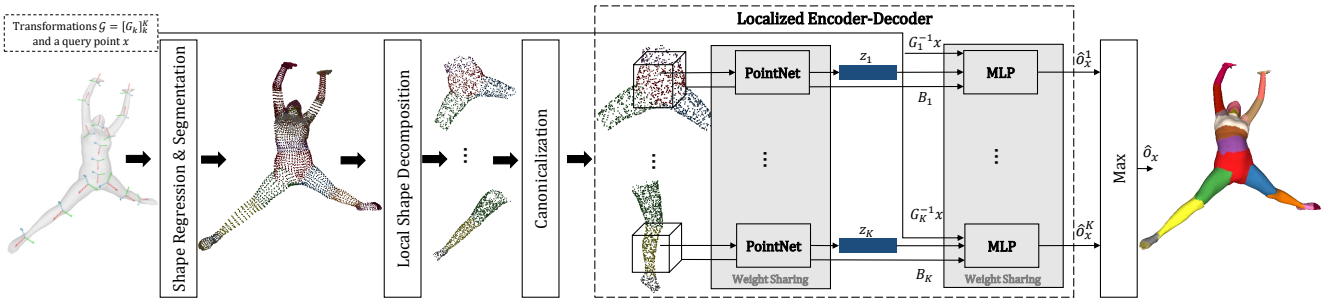

Figure 2. **Overview.** We propose a part-aware neural network that consists of a localized shape encoder and decoder. The model takes the bone transformation matrices $\mathcal{G}$ as input and regresses (via SMPL [25]) a segmented human point cloud. The point cloud is then decomposed to articulated body parts which are canonicalized and encoded by the PointNet [38] encoders. Finally, the decoder MLPs model each articulated body part independently as occupancy fields which are composed to represent the entire human body. The extracted mesh on the right represents the reconstructed human body, segmented according to the predictions of the decoder MLPs, for an unseen subject performing a novel pose from the PosePrior dataset [1].

skinning weights $\mathcal{W} \in \mathbb{R}^{K \times N}$ as a linear combination of rigid transformation matrices that define the mapping from the canonical to a posed space:

$$V_i = \sum_{k=1}^{K} \mathcal{W}_{k,i} G_k(\theta, \mathbf{J}) G_k(\vec{0}, \mathbf{J})^{-1} \bar{V}_i \,, \tag{4}$$

where $G_k(\vec{0}, \mathbf{J})^{-1}$ removes the transformation due to the canonical pose (see [25] for more details).

Shape Regression. The bone transformation matrices G_k define and fully constrain a human skeleton in a posed space. They encapsulate information about the canonical joints $\mathbf{J}$, which enables direct conversion of the transformation matrices into the shape vector β for a small number of shape coefficients via the linear system:

$$\mathsf{B}_S(\beta) = \mathbf{J} - \mathcal{J}\bar{T} \,. \tag{5}$$

Such conversion enables us to interchangeably use the bone transformation matrices G_k to represent shape coefficients β and pose matrices θ and could be directly used to regress a human shape (Eq. (4)). In this work, we use the G_k-notation to be consistent with the previous neural body models [8, 28].

4. COAP

COAP (**CO**mpositional **A**rticulated occupancy of **P**eople) represents articulated human bodies as a differentiable implicit function. It defines the shape volume as the zero-level set $f_\Theta(x|\mathcal{G}) = 0$, in which $x \in \mathbb{R}^3$ is the input query point[2], $\mathcal{G} = [G_k]_{k=1}^K \in \mathbb{R}^{K \times 4 \times 4}$ is the input bone transformations with K being the number of articulated body parts; we use the same number of articulated joints and body parts.

[2]Represented as homogeneous coordinates where appropriate.

On a high level, our method first regresses the surface points of a human body using SMPL [25] and then implements a localized encoder-decoder neural network to represent human bodies as an implicit function. Figure 2 shows an overview of our method.

4.1. Localized Shape Encoder

Body Shape Regression and Segmentation. The input bone transformations are first used to regress the deformed SMPL body vertices V in the posed space (Eq. 4). These vertices V are segmented to different body parts based on the SMPL skinning weights.

Local Shape Decomposition. To encode the pose-dependent shape deformations, it is essential to consider not only the segmented body parts but also their neighborhoods in the kinematic chain. Therefore, for the local shape encoding of a segmented body part, besides its surface points, we also include the surface points that belong to its parent and child body parts in the kinematic chain. Specifically, to compute the surface points for a segmented body part k, we use the skinning weights $W \in \mathbb{R}^{K \times N}$ and select all vertices in V whose weights are larger than a threshold (empirically set to 0.01) for all the body parts that are connected with the body part k. We further extend this decomposition to mesh faces in the template mesh $\bar{T}$, and sample points on the mesh surface. Each local part is represented compactly with a point cloud as an intermediate representation. More details about point sampling are in the supplementary materials.

Canonicalization. Directly encoding the local point clouds as feature vectors makes learning hard since the neural networks need to reason about all possible human poses. Therefore, we simplify the learning problem by canonical-

izing the point cloud of the local part k based on its bone transformation G_k. Let the kth point cloud be denoted as $\mathcal{P}^k$, then each point is projected to a canonical space via the corresponding bone transformation:

$$\hat{\mathcal{P}}_i^k = G_k^{-1}\mathcal{P}_i^k \,, \tag{6}$$

where $\hat{\mathcal{P}}^k$ denotes the canonicalized point cloud of the body part k.

Geometric Prior. To further simplify the learning problem and help the neural networks properly allocate capacity, we build a simple geometric prior by constructing 3D bounding boxes $[B_k]_{k=1}^K \in \mathbb{R}^{K\times 6}$ for local body parts. These geometric primitives B_k over-approximate the central component of the corresponding articulated body part and are estimated deterministically by finding extreme points in the local point clouds and adding an additional 15% padding.

Local Shape Codes. Canonicalized point clouds $\hat{\mathcal{P}}^k$ are then encoded via a PointNet [38] as compact feature vectors $z_k \in \mathbb{R}^{128}$ that carry information about canonical shape and complex local deformations. These feature vectors are further augmented with one-hot encoding vectors for body parts to help the neural network learn a part-specific representation. This localized PointNet encodes each articulated part independently and is implemented as a shared neural network for all articulated parts to reduce overfitting and improve the generalization to novel poses.

4.2. Neural Occupancy Decoder

The second part of our approach is a decoder module that represents articulated body parts as occupancy fields which are composed to form a full human shape. The occupancy decoder takes as input the local shape codes $[z_k]_{k=1}^K$, the geometric prior $[B_k]_{k=1}^K$, the bone transformation matrices $[G_k]_{k=1}^K$, and a query point x for which it predicts whether it is inside of a 3D human body.

Local Occupancy Decoder. First, the input query point $x \in \mathbb{R}^3$ is projected to the canonical space of the respective articulated body part $\hat{x}_k = G_k^{-1}x$. These local queries are augmented with a binary mask $b_k \in \{0, 1\}$ to facilitate the training by reducing the learning space, where b_k indicates whether a local point $\hat{x}_k$ is inside of the created bounding box $B_k \in \mathbb{R}^6$. Next, the local query point $\hat{x}_k \in \mathbb{R}^3$, the binary mask $b_k \in \mathbb{R}$, and the local body code z_k are concatenated as a feature vector and propagated through a 10-layer MLP that predicts occupancy value for the kth articulated part $\hat{o}_k$. The occupancy predictions are further multiplied by the weights b_k to reduce potential spurious correlations. Similar to the local PointNet encoder, all local occupancy decoder MLPs share the same weights and perform occupancy checks independently to reduce overfitting. Please see the supplemental material for details about the neural network architecture.

Occupancy Prediction. The final occupancy prediction for the input query point is then determined as the union of localized occupancy predictions via the max operation:

$$\hat{o}_x = \max[\hat{o}_k]_{k=1}^K \,. \tag{7}$$

Note, there are two key differences between our approach and NASA [8] which also composes per-part occupancy representation to obtain occupancy prediction for full bodies. First, our local shape encoding models a combination of local body parts and their direct neighboring parts along the kinematic chain, whereas NASA only captures single body parts. Second, we leverage shared occupancy decoders and geometric priors, while in NASA, each body part has an independent MLP, leading to poor generalization capability to out-of-distribution poses.

4.3. Training

We use the SMPL [25] body meshes from the AMASS dataset [26] to train our model and the baselines. For each body mesh in the training set, we sample a set of query points P. Half of these points are sampled uniformly inside the local bounding boxes $[B_k]_{k=1}^K$, while the other half is sampled around the mesh surface by using a Gaussian noise $x \sim \mathcal{N}(0, 0.1)$. For each query point, we compute the ground truth occupancy value $o_x \in \{0, 1\}$ for supervision similar to the previous works [8, 28] and activate the network output via the sigmoid function σ. Then, the final supervision loss is a simple mean squared error between the ground truth and the predicted occupancy values:

$$\mathcal{L} = \frac{1}{|P|}\sum\nolimits_{x\sim P}(\sigma(\hat{o}_x) - o_x)^2 \,. \tag{8}$$

We use the batch size of ten and optimize the model parameters via the Adam optimizer [20] with the learning rate of 10^{-4} and its default parameters. The representation fully converges after roughly 300k iterations for most experiments.

5. Experiments

We start by comparing our method with the state-of-the-art subject-specific neural implicit body model SNARF [6] and generalizable implicit body models, LEAP [28] and Neural-GIF [48], in Sec. 5.1. Then, we conduct an ablation study to validate our design choices. We further demonstrate the effectiveness of our representation to untangle self-intersected human bodies in Sec. 5.2 and study the benefit of COAP for estimating human-scene interactions [13] in Sec. 5.3. We conclude the section with a brief overview of the current limitations in Sec. 5.4.

Method	G	t [ms] ↓	Female Subjects					Male Subjects				
			50004	50020	50021	50022	50025	50002	50007	50009	50026	50027
SNARF [6]	✗	809	95.75/84.32	95.42/86.32	95.43/86.07	96.08/85.47	95.57/85.01	96.05/82.50	**95.69/82.11**	94.44/83.41	95.35/83.41	95.22/84.91
COAP	✗	**75**	**95.97/85.35**	95.84/87.62	95.57/86.82	95.98/85.65	95.84/86.28	**96.61/82.96**	95.27/81.90	94.91/84.90	96.07/85.89	95.78/86.90
COAP	✓	**75**	95.83/84.09	**96.95/90.57**	**96.93/90.36**	**96.59/87.16**	**97.24/90.36**	86.75/58.75	93.89/76.72	**96.16/88.15**	**96.79/88.22**	**96.89/89.97**

Table 1. **Single-subject neural implicit models.** Comparison of our model and SNARF [6] on subjects from the DFaust dataset [5] performing novel challenging poses from the PosePrior dataset [1]. While both methods are quite robust, ours is over 10 times faster and can additionally generalize to novel identities as shown in the third row; G denotes whether the model has not seen test subjects during training; values in cells are pairs of the mean IoU on uniformly sampled points and on points sampled around the ground truth meshes.

	PosePrior Dataset [1]		DFaust Dataset [5]	
	IoU Unif.	IoU Surf.	IoU Unif.	IoU Surf.
Neural-GIF [48]	65.83	58.21	64.85	43.22
LEAP [28]	89.36	73.33	87.02	66.35
COAP	**96.97**	**89.92**	**95.41**	**84.44**

Table 2. **Generalization to unseen humans.** Comparison of our model with LEAP [28] and Neural-GIF [48] on the identities from the DFaust [5] and the PosePrior [1] datasets performing challenging novel poses from the PosePrior dataset; values in the table correspond to the mean IoU of uniformly sampled points and of points sampled around the ground truth meshes respectively.

Geometric Prior	One Hot Encoding	IoU Local Boxes [%] ↑	IoU Surface [%] ↑
		91.99	82.81
	✓	92.14	84.46
✓		92.99	85.44
✓	✓	**93.61**	**86.86**

Table 3. **Ablation study** quantifies the impact of the geometric prior b_k and the one hot encoding vectors for the local features (Sec. 4); IoU Local Boxes and IoU Surface are respectively the mean IoU of points uniformly sampled around bounding boxes B_k and points sampled around the ground truth surface. The metrics are computed on the PosePrior sequences [1].

5.1. Generalizable Representation Power

Experimental Setup. For a fair comparison with the baselines, we assume a human skeleton topology with 24 body parts ($K = 24$ in Sec. 4) and use the DFaust [5], MoVi [10] and PosePrior [1] datasets to train and evaluate our representation. We report the **mean inference time** in ms for 10k points, the mean Intersection Over Union (**IoU**) of uniformly samples query points in a bounding box around the ground truth mesh, and the IoU of points sampled around the ground truth surface ($\mathcal{N}(0, 0.01)$) [6, 28].

Single-subject Neural Implicit Models. We start by comparing our method with SNARF [6], a state-of-the-art subject-specific neural implicit body representation. Both methods are trained for each subject in the DFaust dataset [5] and evaluated on the challenging poses from the PosePrior dataset [1]. We observe in Table 1 that both methods are robust for challenging poses, whereas ours is more than **10 times faster** while being more accurate in most scenarios. Our method additionally generalizes to novel identities and motions. As demonstrated in Table 1 (3rd row), our model that is trained on MoVi [10] sequences can be directly used for DFaust subjects with challenging poses and produces even higher accuracy than the per-subject trained models from SNARF [6] in Table 1 (1st row).

Generalization to Unseen Subjects. We now compare our model with two recently proposed neural body representations, LEAP [28] and Neural-GIF [48], which generalize to unseen identities. We train our model on the MoVi [10] dataset and use the pretrained baselines provided by the authors; LEAP trained on the same MoVi dataset, and Neural-GIF on augmented multi-shape SMPL models. As validation datasets, we use novel identities from the PosePrior [1] and the DFaust [5] datasets and sample novel poses from the challenging PosePrior dataset.

Quantitative results are displayed in Table 2 (see Sup. Mat. for qualitative results) and demonstrate that our method significantly outperforms the baselines in terms of accuracy. This robustness comes from the compositional design of our representation and not requiring an inverse LBS network that poorly generalizes to novel motions. This further enables faster and end-to-end training, whereas the baselines employ multi-stage training for the LBS networks that is less stable and more sensitive to hyperparameter tuning.

In summary, our implicit representation is efficient, fast, and robust for articulated human bodies.

Ablations Study. Lastly, we study the impact of the geometric prior and the one-hot encoding vectors (Sec. 4) in Tab. 3. All methods are trained for 200k iterations on the MoVi dataset and evaluated on the PosePrior sequences. We observe that using both geometric prior and one-hot encoding improves the accuracy of our model.

5.2. Resolving Self-intersections

Prior work on neural implicit bodies [6, 28, 48] models humans as a holistic implicit field. Such modeling restricts

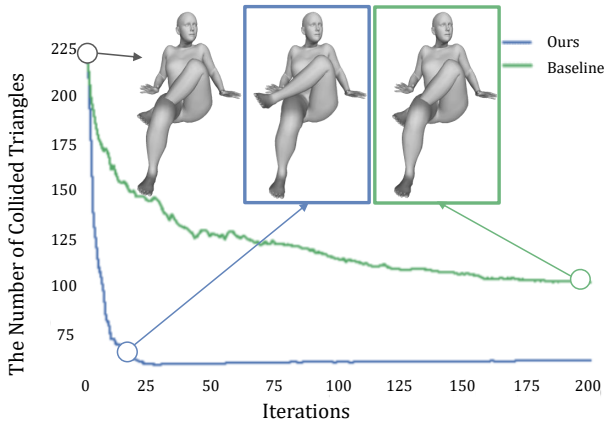

Figure 3. **Resolving self-intersected humans** from the PROX [13] dataset. Our method successfully resolves the challenging self-intersections and converges faster compared to the baseline by a large margin [35, 49].

them from straightforwardly resolving self-intersections. On the contrary, our compositional body model naturally offers this ability and is robust for challenging cases.

Method. Given the self-intersected human body parameters as input (*e.g.* SMPL shape β and pose θ vectors, see Sec. 3), we seek the optimal human pose θ^* such that the human body does not self-intersect. We take inspiration from the traditional computer graphics methods [9, 47] that use geometry proxies to efficiently approximate collisions. We propose to use 3D boxes to approximate body parts in order to efficiently detect potential collided body parts. Based on these collided boxes, we compute their intersected volumes in which we uniformly sample an initial set of points. From this initial set, we select only a subset of points that are inside of at least two body parts by checking our part-wise occupancy predictions. Let this final set be denoted as $\mathcal{S}$, then our self-intersection loss term is defined as:

$$\arg\min_{\theta} \sum\nolimits_{x \in \mathcal{S}} \sigma\left(f_{\Theta}(x|\mathcal{G})\right) . \tag{9}$$

To further prevent unnecessary pose distortions (common in prior approaches [13, 55]), we explicitly disable detecting collisions between kinematically connected body parts that almost always intersect. Please see the supplementary material for additional implementation details.

Evaluation. We use the PROX dataset [13] to study the effectiveness of our method. This dataset contains invalid 3D human bodies whose body parts intersect with each other. From PROX, we sample 100 SMPL bodies by checking the number of self-intersected mesh triangles and compare our method (trained on the MoVi dataset [10]) with the commonly used mesh-based method [35, 49][3] that penalizes intersected mesh triangles via local distance fields.

Both methods optimize the input pose parameters with a simple gradient descent until convergence or the maximum of 200 optimization steps. We quantify the model performance by computing the mean number of self-intersected triangles in the SMPL meshes. Figure 3 illustrates the convergence curve of both methods over 200 optimization steps. Note that our method converges significantly faster and achieves better results compared to the baseline due to the key advantage that our loss term is volume-aware, whereas the baseline imposes the penalty only on the mesh surface.

Qualitative results in Figure 4 illustrate that our method can resolve highly ill-posed self-intersections such as a hand penetrating deeply into the torso.

5.3. Resolving Collisions with 3D Environments

Method. Our method is also compatible with scene-aware human reconstruction methods [13, 39, 55]. These methods convert raw scans of 3D scenes into SDF grids in order to handle collisions. However, such a process is costly and not always feasible. With our representation, one can easily resolve such collisions directly with the raw scans $\mathcal{R} = \{r \in \mathbb{R}^3\}$ by using the following loss term:

$$E_{\text{collision}}(\mathcal{R}) = \frac{1}{|\mathcal{R}|} \sum\nolimits_{r \in \mathcal{R}} \sigma\left(f_{\Theta}(r|\mathcal{G})\right) \mathbb{I}_{f_{\Theta}(r|\mathcal{G})>0} . \tag{10}$$

Evaluation. We demonstrate this application on the lab-controlled portion of the PROX dataset [13], which has accurate scene SDF grids and SMPL registrations (the PROX Quantitative dataset). To impose the collision loss $E_{\text{collision}}$ (10), we directly samples points from a given 3D scan and shift them along the opposite direction of the scan's surface orientation by a displacement sampled from a normal distribution $\mathcal{N}(0.05, 0.05)$. This collision term is then added to the reconstruction terms from the PROX reconstruction pipeline, including 2D joint reprojection E_J, human pose priors E_P, and contact E_C loss terms (see [13] for more details). The final reconstruction loss term is defined as:

$$E = E_J + E_P + E_C + E_{\text{collision}} , \tag{11}$$

which is then optimized with the L-BFGS optimizer [31] until convergence. We see in Table 4 that our method improves the reconstruction accuracy and produces more physically plausible human bodies by reducing collisions with the environment. We also provide the analysis assuming the collision term is derived from a ground truth scene SDF $E_{\text{collision}}^{\text{GT SDF}}$ (third row) for reference.

[3] Code of github.com/vchoutas/torch-mesh-isect

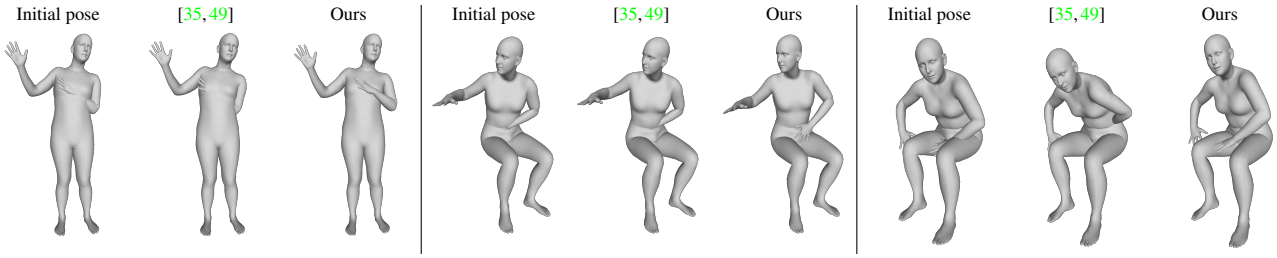

Figure 4. **Resolving self-interactions.** Comparison of our method and the baseline [35, 49] on the PROX [13] dataset. Our optimization method with COAP can resolve highly ill-posed self-intersections such as a hand penetrating deeply into the torso.

	V2V [mm] ↓	PJE [mm] ↓	Penetration ↓
$E_J + E_P + E_C$	154.26	154.39	143.52
$E_J + E_P + E_C + E_{\text{collision}}$	**154.15**	**154.34**	**100.17**
$E_J + E_P + E_C + E_{\text{collision}}^{\text{GT SDF}}$	154.01	154.13	46.84

Table 4. **Collisions with environment.** Experiment performed on the PROX quantitative dataset [13]. We report the mean vertex-to-vertex error (V2V), the mean per-joint error (JPE), and the mean number of penetrated SMPL mesh vertices into the 3D scene geometry (Penetration). We also provide the analysis assuming the collision term is derived from a ground truth scene SDF $E_{\text{collision}}^{\text{GT SDF}}$ (third row) for reference.

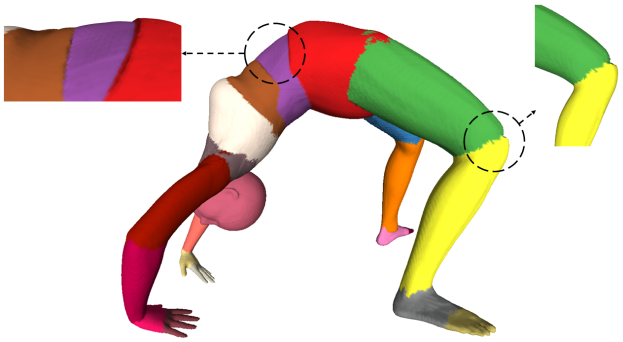

Figure 5. **Limitation.** COAP has difficulties modeling smooth transitions between body parts for out-of-distribution poses. The displayed example is a sample from the PosePrior dataset [1].

We refer the reader to the supplementary video and material for qualitative results and cases where the proposed optimization fails.

5.4. Limitations

Although COAP performs significantly better than previous state-of-the-art models for neural implicit bodies [6, 28, 48] in terms of reconstruction accuracy, sometimes we observe non-smooth connections between body parts (Figure 5) and weak generalization to out-of-distribution extreme body shapes (*e.g.* subject 50002 in Tab. 1 3rd row) if the model is trained on the small number of diverse identities. Additionally, the proposed optimization algorithm for resolving self-intersections sometimes can produce less realistic human pose due to the lack of additional terms that incentivize pose naturalness. We believe that the inference time of COAP (75ms for 10k points) could be improved as it is currently slower compared to the generalizable human bodies LEAP (35ms) and Neural-GIF (22ms). Therefore, exploring even more powerful neural representations and optimization pipelines is an interesting direction for future work.

6. Conclusion and Future Work

Neural implicit representations for human body modeling are a rising research topic. Existing state-of-the-art models have difficulties generalizing to unseen poses and shapes. In this work, we propose COAP, a novel compositional neural occupancy representation, that drastically improves the robustness and the generalization to challenging motions. We decompose the geometry of a full body into local body parts and learn per-part occupancy representations by leveraging the geometric constraints facilitated by the prior knowledge of human body shape. Such part-aware representation enables efficient untangling of challenging self-intersected human bodies and collision detection with other objects.

Future Work. As future work we consider modeling clothing for our neural implicit body model, deploying COAP into 3D human body estimators (*e.g.* [21, 22, 45]) to enforce collision-free predictions during the neural network training, as well as addressing current weaknesses such as generalization to extreme out-of-distribution body shapes and small visible artifacts between body parts for some poses.

Acknowledgments. We thank Shaofei Wang and Yan Zhang for proofreading and Garvita Tiwari for the help with one of the baselines. S. T. and M. M. acknowledge the SNF grant 200021_204840.

Disclaimer. The project was fully completed at ETH Zürich. It was not funded by Meta, nor has it been conducted at Meta.

References

[1] Ijaz Akhter and Michael J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 1, 2, 4, 6, 8

[2] Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. imGHUM: Implicit generative models of 3d human shape and articulated pose. In *Int. Conf. Comput. Vis.*, 2021. 3

[3] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *Eur. Conf. Comput. Vis.*, 2020. 2

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Eur. Conf. Comput. Vis.*, 2016. 2, 3

[5] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: Registering human bodies in motion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 2, 6

[6] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Int. Conf. Comput. Vis.*, 2021. 1, 2, 3, 5, 6, 8

[7] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 3

[8] Boyang Deng, JP Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. NASA: Neural Articulated Shape Approximation. In *Eur. Conf. Comput. Vis.*, 2020. 2, 3, 4, 5

[9] Christer Ericson. *Real-time collision detection*. Crc Press, 2004. 7

[10] Saeed Ghorbani, Kimia Mahdaviani, Anne Thaler, Konrad Kording, Douglas James Cook, Gunnar Blohm, and Nikolaus F Troje. MoVi: A large multipurpose motion and video dataset. *arXiv preprint arXiv:2003.01888*, 2020. 6, 7

[11] Peng Guan. *Virtual human bodies with clothing and hair: From images to animation*. PhD thesis, Brown University Providence, RI, USA, 2012. 2, 3

[12] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *Int. Conf. Comput. Vis.*, 2009. 3

[13] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *Int. Conf. Comput. Vis.*, 2019. 2, 3, 5, 7, 8

[14] Alec Jacobson, Zhigang Deng, Ladislav Kavan, and J. P. Lewis. Skinning: Real-time shape deformation (full text not available). In *ACM SIGGRAPH 2014 Courses*. ACM, 2014. 1

[15] Alec Jacobson, Ladislav Kavan, and Olga Sorkine-Hornung. Robust inside-outside segmentation using generalized winding numbers. *ACM Trans. Graph.*, 2013. 3

[16] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2, 3

[17] Ladislav Kavan, Steven Collins, Jiří Žára, and Carol O'Sullivan. Geometric skinning with approximate dual quaternion blending. *ACM Trans. Graph.*, 2008. 2

[18] Ladislav Kavan and Olga Sorkine. Elasticity-inspired deformers for character articulation. *ACM Trans. Graph.*, 2012. 2

[19] Ladislav Kavan and Jiří Žára. Spherical blend skinning: a real-time deformation of articulated models. In *Interactive 3D graphics and games*, 2005. 2

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Int. Conf. Learn. Represent.*, 2015. 5

[21] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 8

[22] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Int. Conf. Comput. Vis.*, 2019. 8

[23] John P Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Computer graphics and interactive techniques*, 2000. 2

[24] Yijing Li and Jernej Barbič. Immersion of self-intersecting solids and surfaces. *ACM Trans. Graph.*, 2018. 2

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graph.*, 2015. 1, 2, 3, 4, 5

[26] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *Int. Conf. Comput. Vis.*, 2019. 5

[27] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 3

[28] Marko Mihajlovic, Yan Zhang, Michael J Black, and Siyu Tang. LEAP: Learning articulated occupancy of people. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 1, 2, 3, 4, 5, 6, 8

[29] Neil Molino, Robert Bridson, and Ronald Fedkiw. Tetrahedral mesh generation for deformable bodies. In *Symposium on Computer Animation*, 2003. 2

[30] Matthieu Nesme, Paul G Kry, Lenka Jeřábková, and François Faure. Preserving topology and elasticity for embedded deformable models. In *ACM SIGGRAPH*. ACM, 2009. 2

[31] Jorge Nocedal and Stephen J Wright. Nonlinear equations. *Numerical Optimization*, 2006. 7

[32] Ahmed AA Osman, Timo Bolkart, and Michael J Black. Star: Sparse trained articulated human body regressor. In *Eur. Conf. Comput. Vis.*, 2020. 2

[33] Pablo Palafox, Aljaz Bozic, Justus Thies, Matthias Nießner, and Angela Dai. Neural parametric models for 3d deformable shapes. In *Int. Conf. Comput. Vis.*, 2021. 3

[34] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 3

[35] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1, 2, 3, 7, 8

[36] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Eur. Conf. Comput. Vis.*, 2020. 3

[37] Gerard Pons-Moll, Jonathan Taylor, Jamie Shotton, Aaron Hertzmann, and Andrew Fitzgibbon. Metric regression forests for correspondence estimation. *Int. J. Comput. Vis.*, 2015. 2, 3

[38] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 4, 5

[39] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *Int. Conf. Comput. Vis.*, 2021. 2, 7

[40] Damien Rohmer, Stefanie Hahmann, and Marie-Paule Cani. Exact volume preserving skinning with shape control. In *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 2009. 2

[41] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Trans. Graph.*, 2017. 2

[42] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 1, 3

[43] Igor Santesteban, Nils Thuerey, Miguel A Otaduy, and Dan Casas. Self-supervised collision handling via generative 3d garment models for virtual try-on. In *CVPR*, 2021. 3

[44] Eftychios Sifakis, Kevin G Der, and Ronald Fedkiw. Arbitrary cutting of deformable tetrahedralized objects. In *ACM SIGGRAPH*, 2007. 2

[45] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *Int. Conf. Comput. Vis.*, 2021. 2, 8

[46] Matthias Teschner, Stefan Kimmerle, Bruno Heidelberger, Gabriel Zachmann, Laks Raghupathi, Arnulph Fuhrmann, M-P Cani, François Faure, Nadia Magnenat-Thalmann, Wolfgang Strasser, et al. Collision detection for deformable objects. In *Computer graphics forum*, 2005. 3

[47] Jean-Marc Thiery, Émilie Guy, and Tamy Boubekeur. Sphere-meshes: Shape approximation using spherical quadric error metrics. *ACM Trans. Graph.*, 2013. 7

[48] Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Neural-gif: Neural generalized implicit functions for animating people in clothing. In *Int. Conf. Comput. Vis.*, 2021. 1, 2, 3, 5, 6, 8

[49] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *Int. J. Comput. Vis.*, 2016. 3, 7, 8

[50] Rodolphe Vaillant, Loïc Barthe, Gaël Guennebaud, Marie-Paule Cani, Damien Rohmer, Brian Wyvill, Olivier Gourmel, and Mathias Paulin. Implicit skinning: Real-time skin deformation with contact modeling. *ACM Trans. Graph.*, 2013. 2

[51] Shaofei Wang, Andreas Geiger, and Siyu Tang. Locally aware piecewise transformation fields for 3d human mesh registration. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2

[52] Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. Metaavatar: Learning animatable clothed human models from few depth images. In *Adv. Neural Inform. Process. Syst.*, 2021. 1, 3

[53] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. *arXiv preprint arXiv:2111.11426*. 3

[54] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 1, 2

[55] Siwei Zhang, Yan Zhang, Federica Bogo, Pollefeys Marc, and Siyu Tang. Learning motion priors for 4d human body capture in 3d scenes. In *Int. Conf. Comput. Vis.*, 2021. 2, 3, 7

[56] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J. Black, and Siyu Tang. PLACE: Proximity learning of articulation and contact in 3D environments. In *3DV*, 2020. 3

[57] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2, 3