Learning Semantic-Aware Disentangled Representation for Flexible 3D Human Body Editing

Xiaokun Sun[1] Qiao Feng[1] Xiongzheng Li[1] Jinsong Zhang[1] Yu-Kun Lai[2] Jingyu Yang[1] Kun Li[1]*
[1]Tianjin University, China [2]Cardiff University, United Kingdom
{sxk_26,fengqiao,lxz,jinszhang,yjy,lik}@tju.edu.cn LaiY4@cardiff.ac.uk

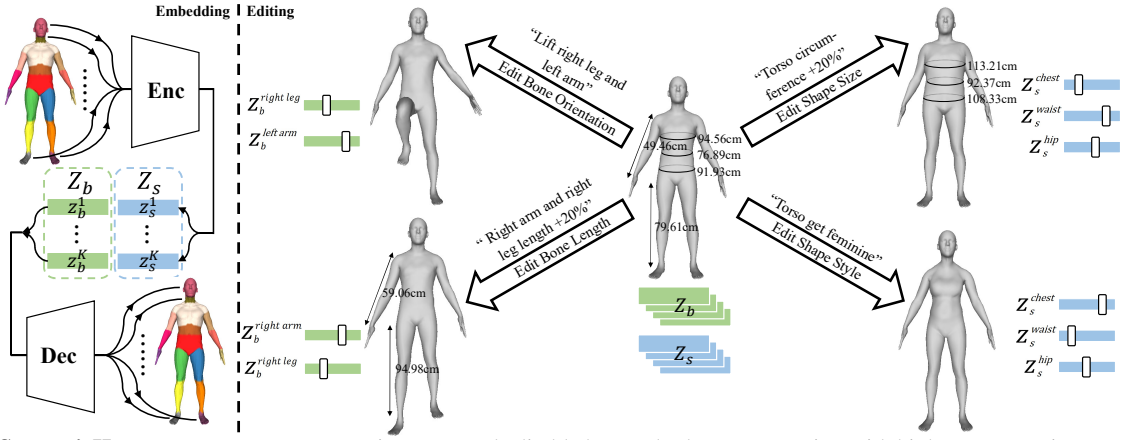

Figure 1. **SemanticHuman**. We present a semantic-aware and editable human body representation with high reconstruction precision in an unsupervised setting. The key idea is to employ a part-aware skeleton-separated decoupling strategy to learn a geometrically meaningful latent space with fine-grained semantics, which benefits part-level personalized human body editing. After embedding a human body into the bone code Z_b and shape code Z_s (left), we can flexibly edit human body attributes by modifying the corresponding latent codes (right).

Abstract

3D human body representation learning has received increasing attention in recent years. However, existing works cannot flexibly, controllably and accurately represent human bodies, limited by coarse semantics and unsatisfactory representation capability, particularly in the absence of supervised data. In this paper, we propose a human body representation with fine-grained semantics and high reconstruction-accuracy in an unsupervised setting. Specifically, we establish a correspondence between latent vectors and geometric measures of body parts by designing a part-aware skeleton-separated decoupling strategy, which facilitates controllable editing of human bodies by modifying the corresponding latent codes. With the help of a bone-guided auto-encoder and an orientation-adaptive weighting strategy, our representation can be trained in an unsupervised manner. With the geometrically meaningful latent space, it can be applied to a wide range of applications, from human body editing to latent code interpolation and shape style transfer. Experimental results on public datasets demonstrate the accurate reconstruction and flexible editing abilities of the proposed method. The code will be available at http://cic.tju.edu.cn/faculty/likun/projects/SemanticHuman.

*Corresponding author

1. Introduction

Learning low-dimensional representations of human bodies plays an important role in various applications including human body reconstruction [4, 19, 32, 37], generation [7, 30, 31] and editing [35, 36, 39]. Existing methods [2, 18, 22, 25, 29] are either too semantically coarse to enable personalized human body editing, or suffer from poor reconstruction performance due to limited representation capability. This paper aims to develop a fine-grained semantic-aware human body representation with flexible representation ability.

Since human bodies are rich in variations of poses and shapes, traditional linear models [1, 25, 29, 35, 36] cannot handle complex nonlinear structures of human body meshes accurately. Therefore, parametric models have been proposed for better representation. The landmark works SCAPE [2] and SMPL [22] represent human bodies by the shape and pose parameters. However, the semantics of their shape parameters are not sufficiently precise, making it impossible to flexibly edit the body shape. Furthermore, the

representation ability of these methods is limited by the linear shape space of human body shapes, and hence their reconstruction accuracy is often unsatisfactory.

With the success of deep learning, the encoder-decoder architecture has demonstrated excellent representation capability [7, 10, 13, 14, 26]. Such methods improve the reconstruction precision by constructing different convolution-like operators for feature extraction on irregular meshes. However, these works lack disentangled representation and fail to obtain promising results for geometrically complex human body parts. Several works [3, 9, 11, 18, 38] pursue the disentanglement of latent representations, *i.e*., each latent code has clear semantics. But these methods either require paired supervised data or have poor performance on the reconstruction, which significantly affects their generalization and robustness. In addition, the semantics of the above representations are coarse, which only enables person-level attribute transfer and cannot be applied to part-level flexible editing.

In this paper, we aim to build a human body representation with fine-grained semantics and high reconstruction-accuracy in an unsupervised setting, which needs to overcome two main challenges. First, how to disentangle the human body to reconstruct precise semantics is a key but difficult problem. Although it is straightforward to decompose a human body into articulated parts for part-level editing, the hidden space of each part is still coupled. Secondly, providing paired supervised data requires a lot of manual effort, and it is very challenging to make the representation disentangled without sacrificing reconstruction accuracy in an unsupervised manner.

To address these challenges, we propose *SemanticHuman*, an editable human body representation with fine-grained semantics and high reconstruction-precision, which facilitates controllable human body editing without paired supervised data. To reconstruct fine-grained semantics, we design a part-aware skeleton-separated decoupling strategy with anatomical priors of the human body. Specifically, we disentangle body part variations into bone-related variations (*e.g*., length and orientation variations) and bone-independent variations (*e.g*., circumference variations). In contrast to the previous pose and shape disentanglement on the entire person [2, 18, 22], this part-aware skeleton-separated decoupling strategy establishes a correspondence between latent vectors and geometric properties of body parts, which benefits part-level controllable editing.

To ensure high reconstruction accuracy and fine-grained semantics of the representation by unsupervised learning, we propose a bone-guided autoencoder architecture and an orientation-adaptive geometry-preserving loss. The bone-guided auto-encoder fuses the geometric features of body parts with their joint information to achieve accurate and efficient modeling of human bodies. Besides, an orientation-adaptive weighting strategy is introduced to compute the geometry-preserving loss, which can provide effective geometric regularization for unsupervised disentanglement and part-level editing. Experimental results on two public datasets with different mesh connectivities demonstrate the high reconstruction-precision and controllable editing capability of the proposed method. An example is given in Fig. 1. The code will be available at http://cic.tju.edu.cn/faculty/likun/projects/SemanticHuman.

Our main contributions are summarized as follows:

- We propose a semantic-aware and editable human body representation with fine-grained representation ability. The latent space of our approach facilitates personalized editing of human bodies by modifying their latent vectors.
- We propose a part-aware skeleton-separated decoupling strategy exploiting structural priors of the human body to learn geometrically meaningful latent codes with fine-grained semantics.
- We propose a bone-guided auto-encoder architecture and an orientation-adaptive geometry-preserving loss to ensure the robust and effective disentanglement of the representation learned without supervision.

2. Related Work

Classical Human Parametric Models. Since human bodies are geometric structures with strong priors, it is straightforward to use a statistical parametric model to represent human bodies. SCAPE [2] is one of pioneering works, which models variations between different human bodies as shape-related and pose-related deformations. SMPL [22] represents the human body more accurately and robustly based on vertices instead of triangle deformations. The SMPL has been extended to represent animals [40], hands [28], and a combination of hands and faces [24].

Deep Learning for Mesh Analysis. Traditional convolutional neural networks cannot be directly applied to meshes with irregular structures. A series of investigations [6, 8, 16, 17, 21, 33] has been devoted to constructing convolution-like operators on irregular meshes. Ranjan *et al*. [26] learn nonlinear representations of human faces by applying spectral convolutions to meshes. Additionally, Bouritsas *et al*. [7] and Gong *et al*. [14] propose to analyze per-vertex spatial features using spiral convolution. Chen *et al*. [10] and Gao *et al*. [13] improve the robustness and efficiency by adopting attention mechanisms in spatial aggregation. Recent works [30, 31] analyze meshes based on as-consistent-as-possible (ACAP) features [12] rather than Euclidean coordinates to learn large-scale deformations on meshes. However, these methods cannot provide explicit

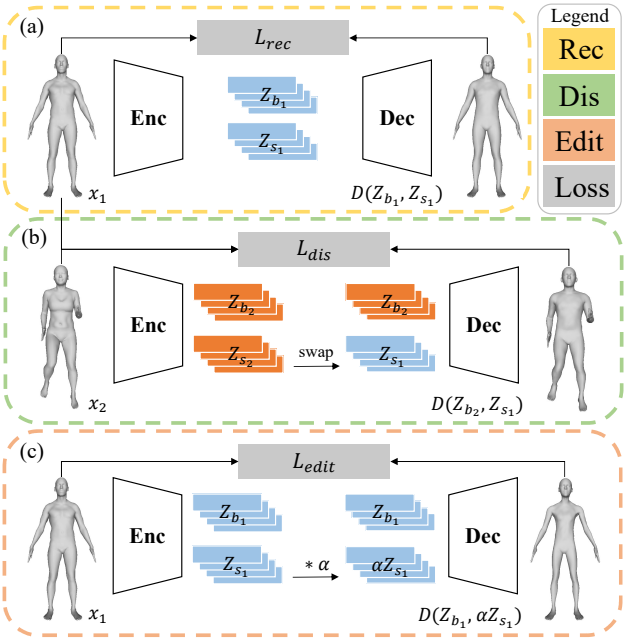

Figure 2. **SemanticHuman** consists of a reconstruction branch, a disentanglement branch, and an editing branch. **(a)** The encoder E maps a mesh x_1 into the bone code $Z_{b_1} = \{z^1_{b_1}, ..., z^K_{b_1}\}$ and shape code $Z_{s_1} = \{z^1_{s_1}, ..., z^K_{s_1}\}$, where K is the number of parts, and the decoder D aims to recover the original mesh $D(Z_{b_1}, Z_{s_1})$. **(b)** With the help of L_{dis}, the swapped codes (Z_{b_2}, Z_{s_1}) will be fed into D to generate the target mesh $D(Z_{b_2}, Z_{s_1})$ retaining the skeleton of x_2 and shape features of x_1. **(c)** By introducing L_{edit}, we force the generated mesh from the scaled codes $(Z_{b_1}, \alpha Z_{s_1})$ to deform as desired, where α is a scale factor.

semantics and fail to handle complex geometries. We alleviate these problems by introducing an autoencoder framework that carries geometric priors of human bodies.

Disentangled Representation for Human Bodies. Jiang *et al*. [18] present a shape and pose disentangled human body representation based on a deep hierarchical neural network, achieving superior reconstruction accuracy. But this method depends on a strong data constraint: each posed mesh must have a paired mesh in a neutral pose. Inspired by the unsupervised disentangled generative model [3], various novel loss functions [9, 11, 38] have been proposed to preserve shape or pose in unsupervised disentanglement. Nevertheless, the supervision provided by these losses is not robust enough, and hence the reconstruction performance of such methods is unsatisfactory. By decoupling pose and shape on the whole human body, these disentangled representations can provide only coarse semantics for latent codes, and thus cannot support part-level human body editing.

Different from prior works, we propose a part-aware skeleton-separated disentanglement strategy, which not only provides precise semantics but also benefits the design of effective information-preserving losses for unsupervised learning.

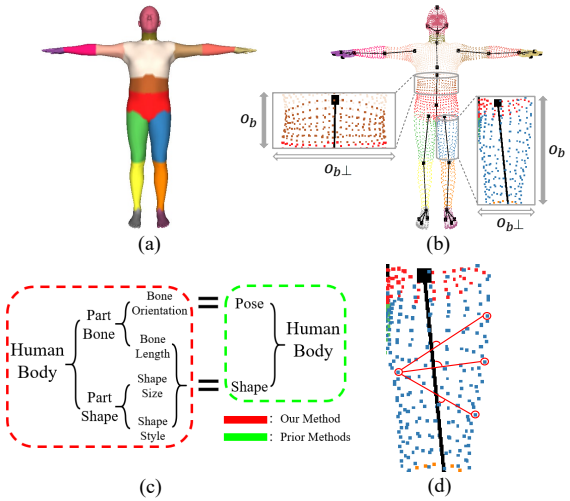

Figure 3. The part-aware skeleton-separated decoupling strategy: **(a)** anatomical human body parts, **(b)** human body bones and joints, **(c)** overview of decoupling strategy, and **(d)** angles between lines and o^k_b.

3. Method

3.1. Overview

Given a set of human meshes with consistent connectivity, our goal is to learn a latent representation that has both fine-grained semantics and high reconstruction-precision in an unsupervised setting. In previous unsupervised decoupling works [3, 9, 11, 38], it is a common way to design novel loss functions capturing the shape or pose information to make representations disentangled. Nonetheless, limited by this traditional disentanglement idea focusing on the whole body, these methods are semantically coarse and not robust enough. In contrast, we design a part-aware skeleton-separated decoupling strategy (Sec. 3.2), which not only provides fine-grained semantics for flexible and precise editing of human attributes (*e.g*., circumference, bone orientation and length) but also facilitates the construction of robust information-preserving losses to achieve unsupervised learning.

Based on this decoupling strategy, we introduce a bone-guided encoder-decoder framework (Sec. 3.3) and exploit three losses (Sec. 3.4) to achieve three core tasks by unsupervised learning: accurate geometric reconstruction (Sec. 3.4.1), unsupervised disentanglement (Sec. 3.4.2) and part-level shape editing (Sec. 3.4.3). An overview of our method is illustrated in Fig. 2, three flows of our pipeline complete the corresponding tasks with the help of specific losses.

3.2. Part-Level Bone and Shape Disentanglement

We leverage an observation that a human body is composed of $K = 17$ anatomical components each containing a bone defined by joints, as shown in Figs. 3 (a) and (b). In

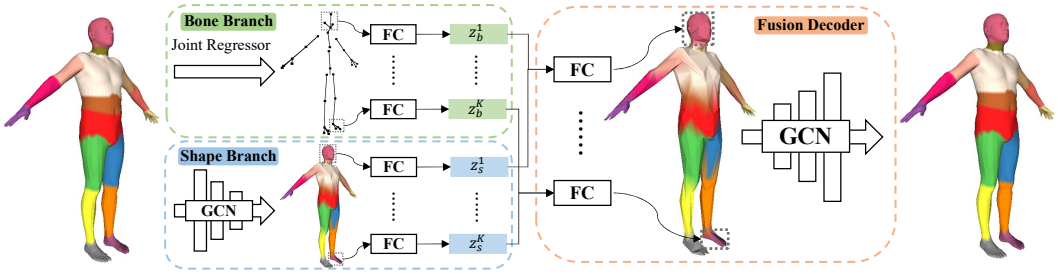

Figure 4. Details of our bone-guided autoencoder.

particular, for geometrically simple human body parts such as the waist, arms, and legs, their shapes can be approximated as cylinders with bones as axes, *i.e.*, the geometry of these parts can be modeled as variations along their bone orientations o_b, and along their orthogonal directions $o_{b\perp}$, which are indicated by solid arrows and hollow arrows in Fig. 3 (b).

Inspired by this observation, we propose a part-aware skeleton-separated decoupling strategy. An overview of our decoupling strategy is shown in Fig. 3 (c). Specifically, we separate body part variations into bone-related variations (*e.g.*, length and orientation variations) and bone-independent shape variations (*e.g.*, size and style variations), which are represented by the bone embedding z_b^k and shape embedding z_s^k for the k-th part, respectively. This part-aware skeleton-separated decoupling strategy establishes a correspondence between latent codes and geometric properties of body parts, which provides fine-grained semantics and enables part-level personalized human body editing.

3.3. Bone-Guided Autoencoder

The basic architecture of our framework is shown in Fig. 4. All three tasks are based on this architecture. Given a human body mesh x, the bone branch and shape branch of encoder E embed the mesh into the bone code $Z_b = \{z_b^1, ..., z_b^K\}$ and shape code $Z_s = \{z_s^1, ..., z_s^K\}$, respectively, where z_b^k and z_s^k are localized latent codes for the k-th body part.

In particular, the bone branch infers localized bone codes $\{z_b^1, ..., z_b^K\}$ with global information (*e.g.*, length and orientation) about each body part from joints B predicted by a linear regressor $J(\cdot)$ [22]. Besides, the shape branch takes mesh x as input with a hierarchical spiral convolution encoder for learning geometric features in multiple scales. Geometric features belonging to each part are subsequently fed into the corresponding fully connected layer according to the part labels of vertices to obtain localized bone codes $\{z_s^1, ..., z_s^K\}$ containing regional geometric details. Finally, the decoder with a similar structure to the shape branch accurately and efficiently reconstructs the original mesh $D(Z_b, Z_s)$ by integrating local and global information of each body part.

3.4. Losses for Unsupervised Learning

With the above framework, we utilize three losses to achieve the corresponding tasks in an unsupervised setting, and the full objective function is defined as:

$$\mathcal{L}_{full} = \mathcal{L}_{rec} + \mathcal{L}_{dis} + \mathcal{L}_{edit}, \tag{1}$$

where $\mathcal{L}_{rec}$ is a geometric reconstruction loss for accurate human body reconstruction, $\mathcal{L}_{dis}$ is a disentanglement loss to ensure the bone and shape disentanglement of body parts, and $\mathcal{L}_{edit}$ is introduced to enable part-level shape editing. The losses and training flows will be described in detail in the following sections.

3.4.1 Accurate Geometric Reconstruction

As shown in Fig. 2 (a), in order to reconstruct the original mesh as accurately as possible, we adopt the geometric reconstruction loss as follows:

$$\mathcal{L}_{rec} = \mathcal{L}_{vert} + \lambda_{edge} \cdot \mathcal{L}_{edge}, \tag{2}$$

where λ_{edge} is the weight of edge regularization. The vertex loss forces the reconstructed mesh $D(E(x))$ to be as close as possible to the original mesh x with the supervision of vertex-wise L1 distance, which is calculated as:

$$\mathcal{L}_{vert} = \|x - D(E(x))\|_1. \tag{3}$$

However, only using this loss to supervise vertex positions cannot avoid producing over-length edges, which seriously affects the smoothness and reasonableness of results. Inspired by [15,34], we address this problem by introducing an edge length regularization L_{edge}, which is formulated as:

$$\mathcal{L}_{edge} = \sum_{p} \sum_{v \in N(p)} \|p - v\|_2^2, \tag{4}$$

where $N(p)$ is the set of 1-ring neighbors of vertex p. This loss ensures the smoothness of output meshes by enforcing their surface to be tight.

3.4.2 Unsupervised Disentanglement

After the reconstruction flow in Sec. 3.4.1, our representation is already capable of accurate reconstruction, but its latent space is still entangled. Based on the part-aware skeleton-separated decoupling strategy (Sec. 3.2), we use $\mathcal{L}_{dis}$ to ensure the bone and shape disentanglement of body parts, which can be defined as:

$$\mathcal{L}_{dis} = \mathcal{L}_{dis\_b} + \lambda_{dis\_s} \cdot \mathcal{L}_{dis\_s}. \quad (5)$$

To achieve unsupervised disentanglement, given two human bodies x_1 and x_2, x_{swp} denotes the generated mesh $D(Z_{b_2}, Z_{s_1})$, which is constructed with the bone latent code from x_2 and shape latent code from x_1, as shown in Fig. 2 (b). Overall, $\mathcal{L}_{dis\_b}$ and $\mathcal{L}_{dis\_s}$ are used to preserve bone information belonging to parts of x_2 and geometric features along $o_{b\perp}$ belonging to parts of x_1, respectively. In this way, our method achieves decoupling via unsupervised learning. x_{swp} should have close joint positions as x_2 as these are largely determined by the bones, so L_{dis_b} is defined as

$$\mathcal{L}_{dis\_b} = \|J(x_2) - J(x_{swp})\|_1, \quad (6)$$

where $J(x)$ is the vector containing joint positions for x obtained through a joint regressor.

However, preserving geometric features is a challenging issue. An intuitive idea to solve this issue is to use the matrix of pairwise Euclidean distances between all vertices in each part to capture regional shape information, which can be defined as:

$$\mathcal{L}_{dis\_s} = \sum_{k=1}^{K} \|D_e(x_1^k) - D_e(x_{swp}^k)\|_1, \quad (7)$$

where x^k is the k-th part-mesh of x, and $D_e(x^k)$ denotes the matrix of pairwise Euclidean distances between all vertices in x^k. $D_e(x^k)$ is of size $n(x^k) \times n(x^k)$, where $n(x^k)$ is the number of vertices of the k-th part. Nevertheless, the bone length is coupled with the Euclidean distance matrix, which leads to incomplete decoupling and affects editing precision (Sec. 4.3).

To alleviate this problem, we propose an orientation-adaptive weighting (OAW) strategy to enforce the representation to focus on shape variations along $o_{b\perp}$. For each body part x^k, we first construct a pairwise angle matrix of size $n(x^k) \times n(x^k)$ by joining every pair of vertices to form a line, and working out the angle $A(x^k)$ (in degrees) the line makes with the bone orientation o_b^k, as illustrated in Fig. 3 (d). As can be seen, the larger the angle, the more significant the line contributes to the shape variation along $o_{b\perp}$. We apply the following thresholding and normalization f to obtain the weight matrix $W(x^k)$ also of the same size by suppressing small angles, and setting the maximum weight to one:

$$W(x^k) = f(A(x^k)), \quad (8)$$

where $f(\cdot)$ is an element-wise mapping function defined as

$$f(a) = \begin{cases} a/90, & \text{if } a > \sigma \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

The weighted pairwise distance matrix is then defined as

$$D_e^w(x^k) = W(x^k) \otimes D_e(x^k), \quad (10)$$

where $\otimes$ is element-wise matrix multiplication. This orientation-adaptive weighting strategy enforces the bone length information to be separated from the Euclidean distance matrix as much as possible. Nonetheless, this strategy also ignores some useful geometric information along o_b, which may result in unreasonable mesh deformations when editing bone orientation in large scale (Sec. 4.3). To address this problem, we further impose part-level volume regularization penalizing the unreasonable shape variations. Ideally, when the shape along $o_{b\perp}$ is retained, the volume of a body part should change in proportion to the bone length, so we introduce a volume constraint, which provides strong geometric supervision to achieve natural and reasonable editing results, which can be calculated as:

$$\mathcal{L}_{vol} = \sum_{k=1}^{K} \|v(x_1^k)/l(x_1^k) - v(x_{swp}^k)/l(x_{swp}^k)\|_1, \quad (11)$$

where $v(\cdot)$ is a function that calculates the volume of a mesh part according to the tetrahedral volume formula, and $l(\cdot)$ is a function that measures the length of a mesh part between the two joints. We can then rewrite $\mathcal{L}_{dis\_s}$:

$$\mathcal{L}_{dis\_s} = \sum_{k=1}^{K} \|D_e^w(x_1^k) - D_e^w(x_{swp}^k)\|_1 + \mathcal{L}_{vol}. \quad (12)$$

3.4.3 Part-Level Shape Editing

With the constraints in Sec. 3.4.2, our autoencoder learns a bone and shape disentangled representation for body parts, and hence we can directly control bone-related attributes (*e.g.*, bone orientation and length) of parts by modifying their joints. Nevertheless, we cannot flexibly edit other bone-independent shape features (*e.g.*, circumference) as desired. To address this issue, we propose an editing flow shown in Fig. 2 (c), which achieves part-level shape editing by forcing the generated mesh from the scaled codes $(Z_{b_1}, \alpha Z_{s_1})$ to deform as desired with L_{edit}:

$$\mathcal{L}_{edit} = \mathcal{L}_{edit\_b} + \lambda_{edit\_s} \cdot \mathcal{L}_{edit\_s} + \lambda_{norm} \cdot \mathcal{L}_{norm}, \quad (13)$$

$$\mathcal{L}_{edit\_s} = \sum_{k=1}^{K} \|\alpha D_e^w(x_1^k) - D_e^w(x_{sca}^k)\|_1, \quad (14)$$

$$\mathcal{L}_{edit\_b} = \|J(x_1) - J(x_{sca})\|_1, \quad (15)$$

where α is a scalar uniformly sampled in $(\alpha_{\min}, \alpha_{\max})$ during training, and x_{sca} denotes the generated mesh $D(Z_{b_1}, \alpha Z_{s_1})$ from the scaled codes, i.e., the edited mesh x_{sca} has bone code from x_1, and shape code from x_1 with flexible scaling to change body shape.

In particular, $\mathcal{L}_{edit\_s}$ constrains the k-th part of x_{sca} reconstructed from the scaled shape codes αZ_{s_1} to have the shape described by $\alpha D_e^W(x_1^k)$. As $D_e^W(\cdot)$ is designed to largely capture the pairwise distances along $o_{b\perp}^k$, i.e. reflecting shape size and style for the part. Elements in the matrix therefore should be scaled accordingly when the part is scaled along $o_{b\perp}^k$. Besides, $\mathcal{L}_{edit\_b}$ is introduced to preserve bone information during the editing process.

However, the editing flow easily causes training collapses. We solve this problem by imposing a vector norm regularization $\mathcal{L}_{norm}$ to establish a mapping between the norm of localized shape codes and the circumference of body parts, which not only benefits converge, but also allows users to edit part shape size on a unified scale. This regularization can be defined as:

$$\mathcal{L}_{norm} = \frac{1}{K}\sum_{k=1}^{K}\left|\|z_{s_1}^k\|_2 - circ(x_1^k)\right|, \qquad (16)$$

where $circ(\cdot)$ is a function that measures the circumference of a part using the identified landmarks. $\mathcal{L}_{edit}$ successfully allows shape codes' norm and direction to represent shape size and style respectively, which not only enables flexible part-level editing but also ensures the continuity of the learned latent space to a certain extent.

3.5. Implementation Details

For the spiral convolution encoder, we use a framework similar to [7]. Specifically, it consists of four spiral convolution layers and downsampling layers, and the structure of the decoder is a mirror of the encoder, except that the downsampling layers are replaced by the upsampling layers. Our algorithm is implemented in PyTorch [23]. All the training and test experiments are carried out on a PC with an RTX 3090 GPU. We train our network for 300 epochs with a learning rate of 1×10^{-3}, a learning rate decay of 0.99 after each epoch and the Adam optimizer [20]. The entire training time takes around 24 hours. We use 16-dimensional latent codes (8 for bones and 8 for shape) to embed each part. More implementation details can be found in the supplementary material.

4. Experiments

4.1. Datasets

DFAUST. The dynamic human body dataset with the same mesh connectivity as SMPL [22] from Bogo *et al*. [5], captures 14 different body motion sequences (*e.g*., hips, running, and jumping) for each of the 10 human subjects. We evenly extract one-twentieth of the original dataset for the convenience of training. Then we randomly split the extracted data into a test set of 182 meshes and a training set of 1936 meshes.

SPRING. The large human body dataset with the same mesh connectivity as SCAPE [2] from Yang *et al*. [35], consists of 3000+ subject meshes with a rough A-pose registered from the CAESAR dataset [27] using a non-rigid deformation algorithm. For the subsequent experiments, the SPRING dataset is randomly split into 2743 training and 305 test meshes.

Data Preprocessing. For the training data, our method requires a joint regressor, body part semantics of each vertex, and landmarks for calculating circumferences. Since the mesh connectivity of datasets is consistent (which is required by nearly all 3D human representation learning works), the SMPL model only needs to be registered once to meet the requirements. More details about data preprocessing are given in the supplementary material.

Method	DFAUST		SPRING	
	E_{avd}	Param(M)	E_{avd}	Param(M)
COMA [26]	6.06	7.54	6.04	6.84
Neural3DMM [7]	5.49	30.35	6.11	27.56
Spiralplus [14]	5.35	15.15	4.99	13.75
Pai3DMM [13]	5.76	15.18	4.45	13.78
Deep3DMM [10]	9.91	8.35	10.88	7.84
Unsup [38]	10.18	12.89	-	-
Ours	**4.70**	**1.59**	**4.33**	**1.47**

Table 1. Quantitative reconstruction results on DFAUST [5] and SPRING [35] datasets. - : not supported for this dataset. Param(M) shows the number of learnable parameters in millions.

Method	DHNN	
	E_{avd}	Param(M)
DHNN [18]	**3.16**	91.63
Ours	3.96	**1.47**

Table 2. Quantitative reconstruction results on DHNN [18]. Param(M) shows the number of learnable parameters in millions.

4.2. Comparison

In the following, we evaluate the representation ability (reconstruction precision) and editing capacity (semantics) of our approach on the DFAUST and SPRING datasets.

Reconstruction Experiments. We first compare four kinds of methods to validate the reconstruction precision of our representation: the spectral-based approach (COMA [26]), the spiral-based methods (Neural3DMM [7] and Spiralplus [14]), the attention-based approaches (Pai3DMM [13] and Deep3DMM [10]), and the disentanglement representation works (DHNN [18] and Unsup [38]). For a fair comparison, we use the official implementation of the compared methods with the same latent space dimension. Note that DHNN [18] did not release the training code, so we train and test our model on its dataset for a fair comparison. we utilize the average point-wise Euclidean distance E_{avd} (in millimeters) between corresponding vertices in the input and its reconstruction as metrics.

As shown in Tab. 1, our proposed approach shows excellent representation capability with a small model size. Fig. 5 visualizes some reconstruction results and their error maps. It can be observed that our approach outperforms other methods in reconstruction accuracy, especially for complex geometric details (*e.g*., faces and hands), which demonstrates the effectiveness of our bone-guided autoencoder.

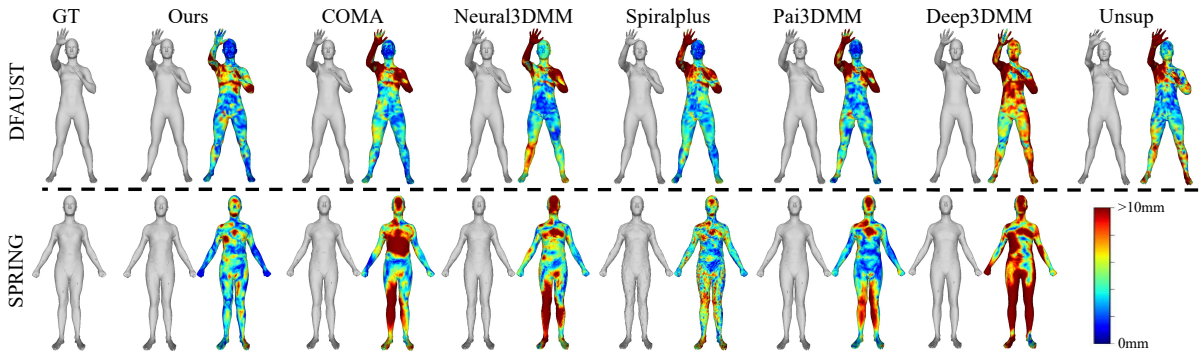

Figure 5. Qualitative reconstruction results on DFAUST [5] and SPRING [35]. The per-vertex Euclidean distance error is color-coded on the reconstructed meshes for visual inspection. Since Unsup [38] has a data constraint that the same subject in different poses should be given, it cannot be trained on the SPRING dataset. Note that our representation is trained without the need of data constraint, which is used in the compared methods.

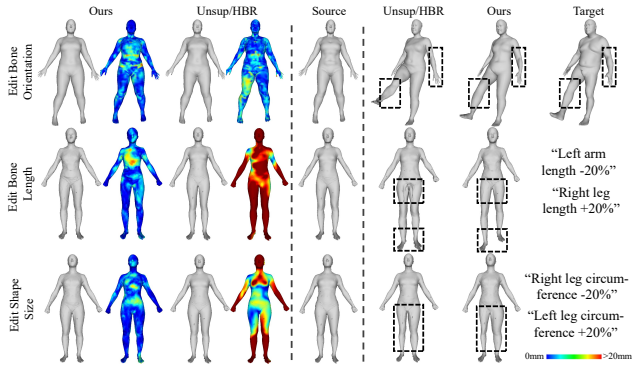

Figure 6. Qualitative editing results by Unsup [38] (first row), HBR [36] (second and third rows) and our method. We show the reconstructed bodies and edited bodies on the left and right of the source mesh.

Method	Bone Orientation		Bone Length		Shape Size	
	E_{joint}	E_{cric}	E_{joint}	E_{cric}	E_{joint}	E_{cric}
Unsup [38]	36.79	14.16	-	-	-	-
HBR [36]	-	-	38.37	15.27	12.87	18.78
Ours	**3.26**	**9.75**	**1.57**	**5.92**	**0.45**	**17.64**

Table 3. Quantitative editing results. - : not supported for this task.

In addition, Tab. 2 gives the quantitative results on the DHNN dataset [18]. It is important to note that the compared method [18] uses a training data assumption (*i.e.*, each posed mesh has a paired mesh in a neutral pose) for shape and pose disentanglement, but our method does not make use of this constraint. Under this unfair condition, the reconstruction accuracy of our method is only slightly lower than DHNN [18], and our model is more lightweight and semantically finer. Please refer to our supplementary material for more experimental results and details.

Editing Experiments. We also demonstrate the flexible editing ability of our model on three editing tasks: editing bone orientation and bone length, and editing part shape size. Since bone orientation is approximately equivalent to pose, we compare with the unsupervised pose-and-shape disentanglement work Unsup [38] on the task of editing bone orientation on the DFAUST dataset. In contrast, bone length and shape size are shape information, so we compare with the Human Body Reshaping work HBR [36] on the task of editing bone length and shape size on the SPRING dataset. As there is no ground truth for these editing tasks, how to measure the editing performance is a problem. We utilize the joint and circumference errors E_{joint}, E_{circ} (in millimeters) to evaluate the accuracy of editing bone and shape.

Specifically, for each test mesh, we randomly select editing targets and then calculate E_{joint}, E_{circ} between the target attribute value and the actual attribute value of the edited human body, extracted using $J(\cdot)$ and $circ(\cdot)$, which can be defined as:

$$E_{joint} = \|T_{joint} - J(x_{edited})\|_2, \tag{17}$$

$$E_{circ} = \frac{1}{K}\sum_{k=1}^{K}\left|T_{circ}^{k} - circ(x_{edited}^{k})\right|, \tag{18}$$

where T_{joint} and T_{circ} are the editing targets of joint positions and circumference.

Tab. 3 gives the quantitative editing results. Our method not only achieves the best performance on all the editing tasks but also preserves other unedited attributes well. Some visual results are shown in Fig. 6. Compared with Unsup [38] and HBR [36], our method can edit the human body more accurately, reasonably and flexibly. Since the attribute error does not always reflect the editing performance of models, we also design a user study for better evaluation (see the supplementary material).

4.3. Ablation Study

Effect of Orientation-Adaptive Weighting Strategy. We validate the effectiveness of the orientation-adaptive weighting strategy by ablating it during training. As shown in Tab. 4, introducing this strategy helps our representation to focus on geometric features along $o_{b\perp}$, leading to higher editing accuracy and more complete decoupling.

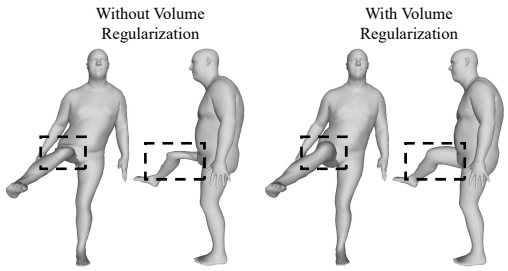

Figure 7. Qualitative results of volume regularization ablation study.

Method	Reconstruction			Editing						
	DFA.	SPR.	Mean	Bone Ori. E_{joint}	Bone Ori. E_{circ}	Bone Length E_{joint}	Bone Length E_{circ}	Shape Size E_{joint}	Shape Size E_{circ}	Mean
Full	4.70	**4.33**	**4.52**	3.26	**9.75**	1.57	**5.92**	**0.45**	17.64	**6.43**
w/o OAW	**4.66**	4.39	4.53	3.02	13.66	1.58	10.95	0.61	20.40	8.37
w/o $\mathcal{L}_{edge}$	5.23	5.01	5.12	3.20	9.76	**1.52**	8.37	0.53	29.48	8.81
w/o $\mathcal{L}_{dis}$	4.81	4.88	4.85	-	-	-	-	1.56	**17.50**	-
w/o $\mathcal{L}_{edit}$	4.87	4.43	4.65	**3.00**	10.19	1.88	8.01	-	-	-

Table 4. Quantitative ablation study for reconstruction and editing (in mm). - : not supported for this task.

Effect of Other Losses. To evaluate the impact of $\mathcal{L}_{edge}$, $\mathcal{L}_{dis}$ and $\mathcal{L}_{edit}$, we remove them from the supervision one by one during training. As compared in Tab. 4, $\mathcal{L}_{edge}$ effectively reduces the reconstruction error, and the use of $\mathcal{L}_{dis}$ and $\mathcal{L}_{edit}$ is required to enable flexible bone and part shape editing. The ablation study proves that all losses are necessary.

Effect of Part-Level Volume Regularization. To analyze the impact of volume regularization, we remove it during the training process. Fig. 7 shows the comparison results of edited bodies. It can be seen that our volume constraint provides strong geometric supervision, resulting in more natural and reasonable editing results.

4.4. Applications

Latent Space Interpolation. Our representation decouples the bone and shape of each part, which allows us to interpolate meshes by linearly interpolating the bone and shape latent codes. Fig. 8 shows reasonable and meaningful results of such interpolation. It is worth noting that, thanks to the latent space with fine-grained semantics, we can perform mesh interpolation at the component level (see Figs. 8 (b) and (c)).

Shape Style Transfer. As mentioned in Sec. 3.4.3, we successfully make the norm and direction of shape vectors represent shape size and style, respectively. Thus, we can transfer shape style by swapping the direction of the shape vectors. Some results are shown in Fig. 9. It can be easily observed that our approach changes the shape style of the source mesh naturally and preserves its other attributes (*e.g.*, bone and circumference).

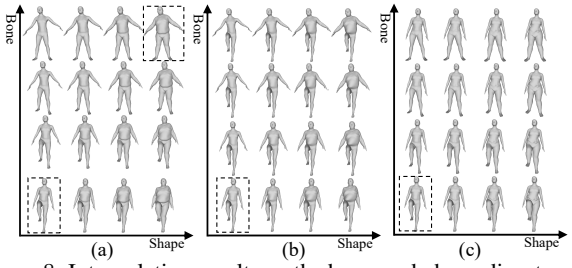

Figure 8. Interpolation results on the bone-and-shape disentangled space of **(a)** the whole body, **(b)** upper body and **(c)** lower body, where the highlighted bodies are the input reconstructed bodies.

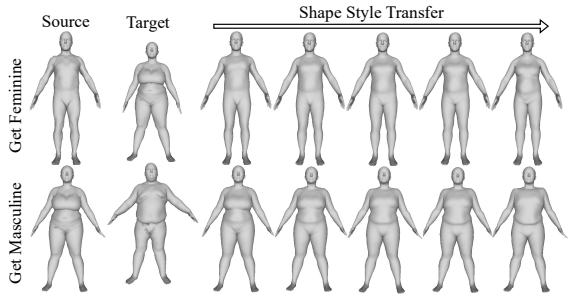

Figure 9. Results of shape style transfer by linearly interpolating the direction of shape codes.

5. Conclusion and Discussion

Conclusion. In this paper, we propose a human body representation with fine-grained semantics and high reconstruction-accuracy in an unsupervised setting. The key idea is to exploit a part-aware skeleton-separated decoupling strategy to establish a correspondence between latent vectors and geometric properties of body parts, which benefits personalized editing of human bodies by modifying the corresponding latent codes. Based on this disentanglement strategy, we propose a bone-guided autoencoder and well-designed losses to learn representation in an unsupervised manner. At last, with the geometrically meaningful latent space, the application can be extended from human body editing to latent code interpolation and shape style transfer.

Limitations. Since our decoupling strategy is based on the cylinder assumption (*i.e.*, the shape of body parts can be approximated as cylinders with bones as axes), we cannot controllably edit the bone length and circumference of parts (*e.g.*, faces, hands, and feet) whose geometry is too complex to meet this assumption. Additionally, editing bone orientation with our method may fail when the target orientation is uncommon in the training data. In further work, we will dig deeper into the prior knowledge about the human body to improve the generalization capability of our representation.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China (62122058 and 62171317).

References

[1] Brett Allen, Brian Curless, and Zoran Popović. The space of human body shapes: reconstruction and parameterization from range scans. *ACM Trans. Graph.*, 2003. 1

[2] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: shape completion and animation of people. *ACM Trans. Graph.*, 2005. 1, 2, 6

[3] Tristan Aumentado-Armstrong, Stavros Tsogkas, Allan Jepson, and Sven Dickinson. Geometric disentanglement for generative latent shape models. In *Int. Conf. Comput. Vis.*, 2019. 2, 3

[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. 1

[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. 6, 7

[6] Davide Boscaini, Jonathan Masci, Simone Melzi, Michael M Bronstein, Umberto Castellani, and Pierre Vandergheynst. Learning class-specific descriptors for deformable shapes using localized spectral convolutional networks. In *Comput. Graph. Forum*, 2015. 2

[7] 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 *Int. Conf. Comput. Vis.*, 2019. 1, 2, 6

[8] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013. 2

[9] Haoyu Chen, Hao Tang, Henglin Shi, Wei Peng, Nicu Sebe, and Guoying Zhao. Intrinsic-extrinsic preserved GANs for unsupervised 3D pose transfer. In *Int. Conf. Comput. Vis.*, 2021. 2, 3

[10] Zhixiang Chen and Tae-Kyun Kim. Learning feature aggregation for deep 3D morphable models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2, 6

[11] Luca Cosmo, Antonio Norelli, Oshri Halimi, Ron Kimmel, and Emanuele Rodola. LIMP: Learning latent shape representations with metric preservation priors. In *Eur. Conf. Comput. Vis.*, 2020. 2, 3

[12] Lin Gao, Yu-Kun Lai, Jie Yang, Ling-Xiao Zhang, Shihong Xia, and Leif Kobbelt. Sparse data driven mesh deformation. *IEEE Trans. Vis. Comput. Graph.*, 2019. 2

[13] Zhongpai Gao, Junchi Yan, Guangtao Zhai, Juyong Zhang, Yiyan Yang, and Xiaokang Yang. Learning local neighboring structure for robust 3D shape representation. In *AAAI*, 2021. 2, 6

[14] Shunwang Gong, Lei Chen, Michael Bronstein, and Stefanos Zafeiriou. SpiralNet++: A fast and highly efficient mesh convolution operator. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019. 2, 6

[15] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. 3D-CODED: 3D correspondences by deep deformation. In *Eur. Conf. Comput. Vis.*, 2018. 4

[16] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Adv. Neural Inform. Process. Syst.*, 2017. 2

[17] Mikael Henaff, Joan Bruna, and Yann LeCun. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163*, 2015. 2

[18] Boyi Jiang, Juyong Zhang, Jianfei Cai, and Jianmin Zheng. Disentangled human body embedding based on deep hierarchical neural network. *IEEE Trans. Vis. Comput. Graph.*, 2020. 1, 2, 3, 6, 7

[19] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 1

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2015. 6

[21] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. In *Int. Conf. Comput. Vis.*, 2019. 2

[22] 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, 4, 6

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. 2019. 6

[24] 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. 2

[25] Leonid Pishchulin, Stefanie Wuhrer, Thomas Helten, Christian Theobalt, and Bernt Schiele. Building statistical shape spaces for 3D human modeling. *Pattern Recognition*, 2017. 1

[26] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3D faces using convolutional mesh autoencoders. In *Eur. Conf. Comput. Vis.*, 2018. 2, 6

[27] Kathleen M Robinette, Hans Daanen, and Eric Paquet. The caesar project: a 3-d surface anthropometry survey. In *International Conference on 3-D Digital Imaging and Modeling*, 1999. 6

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

[29] Hyewon Seo and Nadia Magnenat-Thalmann. An automatic modeling of human bodies from sizing parameters. In *Proceedings of the Symposium on Interactive 3D Graphics*, 2003. 1

[30] Qingyang Tan, Lin Gao, Yu-Kun Lai, and Shihong Xia. Variational autoencoders for deforming 3D mesh models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 1, 2

[31] Qingyang Tan, Lin Gao, Yu-Kun Lai, Jie Yang, and Shihong Xia. Mesh-based autoencoders for localized deformation component analysis. In *AAAI*, 2018. 1, 2

[32] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. Recovering 3D human mesh from monocular images: A survey. *arXiv preprint arXiv:2203.01923*, 2022. 1

[33] Nitika Verma, Edmond Boyer, and Jakob Verbeek. Dynamic filters in graph convolutional networks. *arXiv preprint arXiv:1706.05206*, 2017. 2

[34] Jiashun Wang, Chao Wen, Yanwei Fu, Haitao Lin, Tianyun Zou, Xiangyang Xue, and Yinda Zhang. Neural pose transfer by spatially adaptive instance normalization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 4

[35] Yipin Yang, Yao Yu, Yu Zhou, Sidan Du, James Davis, and Ruigang Yang. Semantic parametric reshaping of human body models. In *International Conference on 3D Vision*, 2014. 1, 6, 7

[36] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. 3D human body reshaping with anthropometric modeling. In *International Conference on Internet Multimedia Computing and Service*, 2017. 1, 7

[37] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In *Int. Conf. Comput. Vis.*, 2021. 1

[38] Keyang Zhou, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Unsupervised shape and pose disentanglement for 3D meshes. In *Eur. Conf. Comput. Vis.*, 2020. 2, 3, 6, 7

[39] Shizhe Zhou, Hongbo Fu, Ligang Liu, Daniel Cohen-Or, and Xiaoguang Han. Parametric reshaping of human bodies in images. *ACM Trans. Graph.*, 2010. 1

[40] Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 2