Learning Locally Editable Virtual Humans

Hsuan-I Ho Lixin Xue Jie Song Otmar Hilliges
Department of Computer Science, ETH Zürich

Abstract

In this paper, we propose a novel hybrid representation and end-to-end trainable network architecture to model fully editable and customizable neural avatars. At the core of our work lies a representation that combines the modeling power of neural fields with the ease of use and inherent 3D consistency of skinned meshes. To this end, we construct a trainable feature codebook to store local geometry and texture features on the vertices of a deformable body model, thus exploiting its consistent topology under articulation. This representation is then employed in a generative auto-decoder architecture that admits fitting to unseen scans and sampling of realistic avatars with varied appearances and geometries. Furthermore, our representation allows local editing by swapping local features between 3D assets. To verify our method for avatar creation and editing, we contribute a new high-quality dataset, dubbed CustomHumans, for training and evaluation. Our experiments quantitatively and qualitatively show that our method generates diverse detailed avatars and achieves better model fitting performance compared to state-of-the-art methods. Our code and dataset are available at https://ait.ethz.ch/custom-humans.

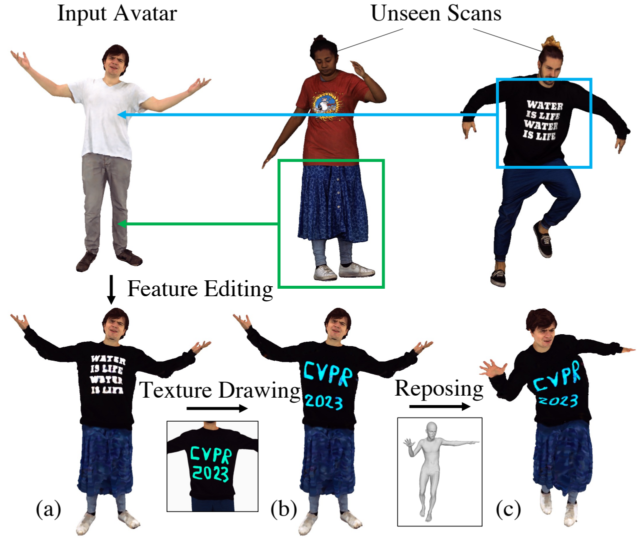

Figure 1. **Creating locally editable avatars:** Given an input avatar, (a) the avatar can be edited by transferring clothing geometry and color details from existing, yet unseen 3D assets. (b) Users can customize clothing details such as logos and letters via drawing on 2D images. (c) The avatars retain local detail consistently under pose changes.

1. Introduction

3D Avatars are an important aspect of many emerging applications such as 3D games or the Metaverse. Allowing for easy personalization of such avatars, holds the promise of increased user engagement. Traditionally, editing 3D assets requires knowledge of computer graphics tools and relies on standardized data formats to represent shapes and appearances. While methods for reconstruction or generative modeling of *learned* avatars achieve impressive results, it is unknown how such neural avatars can be edited and customized. Thus, the goal of our work is to contribute a simple, yet powerful data-driven method for avatar creation and customization (Fig. 1): our method enables (a) the ability to transfer partial geometric and appearance details between 3D assets, and (b) the ability to author details via 2D-3D transfer. The resulting avatars (c) retain consistent local details when posed.

Existing methods do not allow for such capabilities. While 3D generative models of articulated human bodies [5, 21, 25, 42, 75] leverage differentiable neural rendering to learn from images, they cannot control local details due to highly entangled color and geometry in the 2D supervision signal. Generative models trained on 3D data [9, 13, 36, 44, 45] can produce geometric details for surfaces and clothing. However, the diversity of generated samples is low due to the lack of high-quality 3D human scans and not all methods model appearance.

At the core of the issue lies the question of representation: graphics tools use meshes, UV, and texture maps which provide consistent topologies under deformation. However, human avatar methods that are built on mesh-based representations and linear blend skinning (LBS) are limited in their representational power with respect to challenging geometry (e.g., puffy garments) and flexible topologies (e.g., jackets), even with adaptations of additional displacement parameters [36] and mesh subdivision [66].

Inspired by the recent neural 3D representations [40, 62, 70, 72], we propose a novel *hybrid* representation for digital humans. Our representation combines the advantages of consistent topologies of LBS models with the representational power of neural fields. The key idea is to decompose the tasks of deformation consistency on one hand and local surface and appearance description on the other. For the former, we leverage existing parametric body models (e.g., SMPL [33] and SMPL-X [48]). For the latter, we leverage the fixed topology of the poseable mesh to store local feature codebooks. A decoder, shared across subjects, is then conditioned on the local features to predict the final signed distance and color values. Since only local information [15] is exposed to the decoder, overfitting and memorization can be mitigated. We experimentally show that this is crucial for 3D avatar fitting and reposing.

Complementing this hybrid representation, we propose a training pipeline in the auto-decoding generative framework [9, 46, 52]. To this end, we jointly optimize multi-subject feature codebooks and the shared decoder weights via 3D reconstruction and 2D adversarial losses. The 3D losses help in disentangling appearance and geometric information from the input scans, while the latter improves the perceptual quality of randomly generated samples. To showcase the hybrid representation and the generative model we implement a prototypical avatar editing workflow shown in Fig. 1.

Furthermore, to enable research on high-quality 3D avatars we contribute training data for generative 3D human models. We record a large-scale dataset (more than 600 scans of 80 subjects in 120 garments) using a volumetric capture stage [11]. Our dataset consists of high-quality 3D meshes alongside accurately registered SMPL-X [48] models and will be made available for research purposes. Finally, we assess our design decisions in detailed evaluations, both on existing and the proposed datasets.

In summary, our contributions are threefold: (a) a novel hybrid representation for 3D virtual humans that allows for local editing across subjects, (b) a generative pipeline of 3D avatars creation that allows for fitting to unseen 3D scans and random sampling, and (c) a new large-scale high-quality dataset of 3D human scans containing diverse subjects, body poses and garments.

2. Related Work

Controllable human representations. Topics of virtual humans have received much attention in the graphics literature, such as skinning and rigging of articulated meshes [33, 43, 48, 55], physical simulation of clothing [19, 22, 41], and deferred rendering [24, 49, 65]. With the advances in neural rendering [63, 64] and the availability of large-scale human datasets [18, 28, 29, 32, 47, 67, 73, 74], numerous approaches have been proposed to reconstruct [4, 56, 57] and explicitly control [10, 30, 50] human avatars in a data-driven manner.

One branch of work focuses on 2D image synthesis via generative adversarial networks (GANs) [20] and techniques of feature manipulation [28, 54, 60]. Typically, a 2D neural renderer creates human images corresponding to pose and appearance latent codes learned from the training data. Related applications such as virtual try-on [16, 17, 71] and video retargeting [7, 31, 68] have shown promising results in light of photo-realistic image synthesis by GANs [18, 26]. However, these methods do not explicitly reason about complex 3D human geometry and can therefore not produce 3D-consistent results.

A newly emerging line of work aims to create controllable avatars with 3D consistency. Some methods extend existing body models with neural networks to predict displacement layers [3, 6, 36] or textures [51]. Other methods learn to model challenging pose-dependent deformations on avatars either by predicting LBS weights [8, 10, 14, 58, 59, 76] or improving the capabilities of body models [23, 30, 34, 35, 37, 50, 53] with the power of implicit neural fields. However, these approaches mainly focus on modeling a *single* subject in specific clothing and do not scale to create diverse avatars. Our method overcomes this issue by learning a *multi-subject* generative model which produces realistic virtual humans with disentangled controllability over body poses, clothing geometry, and texture.

Generative 3D human models. Existing generative models of human avatars can be loosely split into two main streams: learning 3D-aware neural rendering from 2D images [5, 21, 25, 42, 75] and learning body shapes from 3D supervision [9, 13, 36, 44, 45]. Powered by recent advances in differentiable neural rendering [64] and neural fields [70], much progress has been made in 3D-aware generative models [61]. However, learning to generate detailed clothed avatars from pure 2D supervision [5, 21, 25, 42, 75] is still challenging due to the complex appearance and articulation of bodies, self-occlusions, and highly entangled colors and geometries in images.

More closely related to our setting are methods that learn to generate detailed body shapes from 3D scans or RGB-D data. For instance, CAPE [36] and SMPLicit [13] are generative models for clothed humans. The former exploits VAE-GAN to predict additive displacements based on the SMPL vertices while the latter drape an implicitly modeled garment layer onto SMPL. NPMs [44], and SPAM [45] learn pose and shape latent spaces from 3D supervision, which enables latent code inversion using point clouds or depth sequences. gDNA [9] learns to synthesize body shapes in the canonical space and further improves clothing details with adversarial losses. However, none of the above-mentioned works is able to generate human bodies with appearance and neither allows fine-grained editing of the generated avatars. Our method addresses both issues by learning disentangled

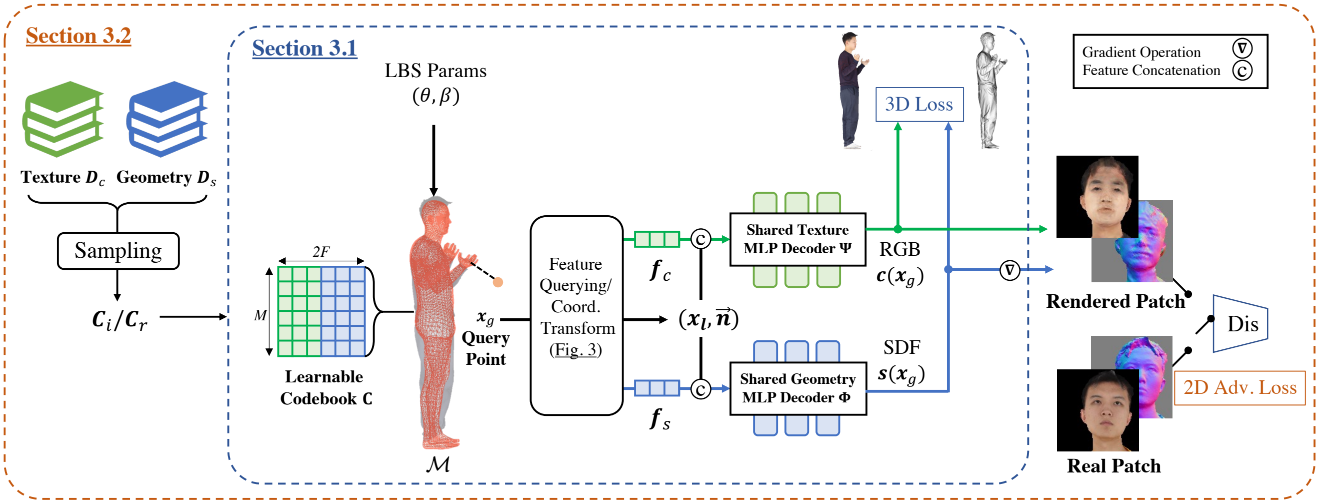

Figure 2. **Proposed framework**. Given a posed scan registered with body pose and shape LBS parameters (θ, β), our proposed human representation stores its local geometry and texture features in a codebook $\mathbf{C}$ which is indexed by the vertex indices of M-vertex LBS body mesh $\mathcal{M}$ (Sec. 3.1). Given a query point coordinates $\mathbf{x}_g$, two separated MLP decoders $\mathbf{\Psi}, \mathbf{\Phi}$ predict signed distances and colors conditioned on the positional features $(\mathbf{x}_l, \vec{\mathbf{n}})$ and the local geometry/texture features $(\mathbf{f_s}$ / $\mathbf{f_c})$ respectively. We train a generative auto-decoder using N posed scans, whose feature codebooks are stored in the dictionary $\mathbf{D_s}, \mathbf{D_c}$. We introduce two sampling strategies to sample codebooks (denoted as $\mathbf{C}_i/\mathbf{C}_r$ respectively) and jointly train our dictionaries and shared MLP decoders with a 3D reconstruction loss and a 2D adversarial loss (Sec. 3.2 & Sec. 3.3).

local representations for multiple subjects. In addition, we experimentally show that our representation significantly improves the performance of model fitting against state-of-the-art human generative models.

3. Method

Our proposed method is summarized in Fig. 2. We first contribute a novel hybrid human representation that stores local geometric and textural information into two aligned feature spaces (Sec. 3.1 and Fig. 3). To allow fitting to new 3D scans and drawing random samples from the underlying data distribution, we design a training strategy to learn a meaningful latent space under the generative adversarial framework to bring in additional 2D adversarial supervision (Sec. 3.2 & Sec. 3.3). Finally, we demonstrate the utility of our method for creating avatars by enabling local feature editing through existing 3D assets or images (Sec. 3.4).

3.1. Hybrid Representation of Humans

To enable 3D avatars with high-fidelity representational power and local editing capabilities, a suitable representation is needed. To this end, we propose a novel *hybrid* representation that combines the advantages of neural fields (flexibility and modeling power) with LBS-articulated mesh models (ease of deformation and full explicit control).

An overview of how we leverage the proposed representation is provided in Fig. 2 in the dotted blue box. Given a human scan, we first create a posed, coarse body mesh $\mathcal{M}$ (shown in red) using the registered body parameters (θ, β) of an LBS body model. The mesh consists of M vertices ($\mathcal{V} \in \mathbb{R}^{M\times 3}$) in the posed space and M_f faces where $\mathcal{F} \in \{1, ..., M\}^{M_f \times 3}$ defines the vertex indices on each face. We then construct a trainable feature codebook $\mathbf{C} \in \mathbb{R}^{M\times 2F}$, which stores F-dimensional local geometry and texture features respectively for each vertex.

Similar to coordinate-based implicit fields, a 3D coordinate $\mathbf{x}_g \in \mathbb{R}^3$ is used to predict its corresponding signed distance $s(\mathbf{x}_g)$, and RGB color $c(\mathbf{x}_g)$. Instead of using global coordinates directly as inputs, we condition neural field decoders on the local triangle coordinates $\mathbf{x}_l \in \mathbb{R}^3$ and the local geometry and texture features $\mathbf{f_s}, \mathbf{f_c} \in \mathbb{R}^F$. We illustrate this conversion from global coordinates to local triangle coordinates in Fig. 3. The global coordinates $\mathbf{x}_g$ are first projected onto the mesh by finding the closest point $\mathbf{x}_c^*$ (Fig. 3, blue dot):

$$\begin{aligned} \mathbf{x}_c^* &= \arg\min_{\mathbf{x}_c} \|\mathbf{x}_g - \mathbf{x}_c\|_2, \\ \mathbf{x}_c &= \mathcal{B}_{u,v}(\mathcal{V}[m_0, m_1, m_2]), \end{aligned} \tag{1}$$

where (m_0, m_1, m_2) are vertex indices of faces $\mathcal{F}$ and $\mathcal{B}_{u,v}(.)$ is the barycentric interpolation function with barycentric coordinates $(u, v, 1-u-v)$. The closest point $\mathbf{x}_c^*$ within the face is used to transform $\mathbf{x}_g$ into a local triangle coordinate system. Hence, $\mathbf{x}_l$ consists of the barycentric coordinates (u, v), the signed distance d between $\mathbf{x}_g$ and $\mathbf{x}_c^*$, i.e., $\mathbf{x}_l \coloneqq (u, v, d)$. We also compute a direction vector $\vec{\mathbf{n}}$ between $\mathbf{x}_g$ and $\mathbf{x}_c^*$ as an additional feature to distinguish points near triangle edges. To query local features $(\mathbf{f_s}, \mathbf{f_c})$, we use the vertex indices on the triangle (m_0^*, m_1^*, m_2^*) to

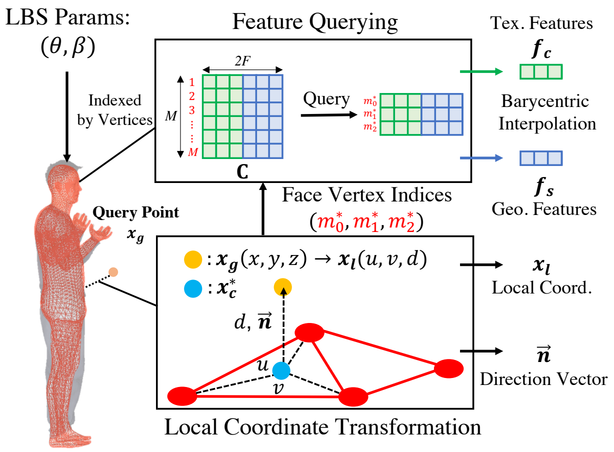

Figure 3. **Local feature querying.** Given an LBS body mesh posed by the parameters (θ, β), we represent a detailed human body as a codebook that stores local texture and geometry features indexed by the vertices on the mesh. An input query point $\mathbf{x_g}$ finds the nearest triangle on the LBS body mesh and returns its vertex indices for local feature lookup. To prevent decoders from memorizing any global information, we transform the position of $\mathbf{x_g}$ into local triangle coordinates (u, v), distance d, and direction $\vec{\mathbf{n}}$ of the closest point (i.e., the blue dot). The final geometry and texture features $\mathbf{f_s}, \mathbf{f_c}$ are fused via barycentric interpolation.

look up three elements in the feature codebook $\mathbf{C}$. We then fuse the three local features via barycentric interpolation.

Finally, we take all the local features $(\mathbf{f_s}, \mathbf{f_c}, \mathbf{x}_l, \vec{\mathbf{n}})$ as input to two separate decoders $\mathbf{\Phi}$ and $\mathbf{\Psi}$ to predict SDF and RGB values respectively:

$$\begin{aligned} \mathbf{\Phi} : \mathbb{R}^F \times \mathbb{R}^3 \times \mathbb{R}^3 &\to \mathbb{R} \\ (\mathbf{f_s}, \mathbf{x}_l, \vec{\mathbf{n}}) &\mapsto s(\mathbf{x}_g), \end{aligned} \tag{2}$$

$$\begin{aligned} \mathbf{\Psi} : \mathbb{R}^F \times \mathbb{R}^3 \times \mathbb{R}^3 &\to \mathbb{R}^3 \\ (\mathbf{f_c}, \mathbf{x}_l, \vec{\mathbf{n}}) &\mapsto c(\mathbf{x}_g). \end{aligned} \tag{3}$$

Note that only local information is exposed to the decoders, which allows us to use the same MLPs across different vertices and subjects. We show that preventing networks from memorizing global information in this way is necessary for local editing and reposing in our experiments (Fig. 9).

3.2. Generative Codebook Sampling

Our goal is to provide means to create and personalize avatars with diverse body shapes, appearances, and local details. To this end, we leverage the above representation to train a *single* multi-subject model which enables the transfer of local features *across* subjects. We note that since the mesh topology of the LBS model is identical, this enables us to learn a shared feature space from multiple posed scans.

To learn the feature representation over a dataset of N scans, it is sufficient to store the codebooks $\mathbf{C}_i$ in two dictionaries $\mathbf{D_s}, \mathbf{D_c} \in \mathbb{R}^{N \times (MF)}$ to represent shape and color information of the i-th subject respectively. The entries $\mathbf{C}_i$ can then be learned jointly with the decoder weights via direct 3D supervision using the i-th scan (Fig. 2). However, we experimentally show that this is insufficient to learn a well-behaved latent space from which we can draw novel samples (see Fig. 10).

Therefore we introduce a codebook sampling strategy that allows us to draw random samples and update the entries of the dictionaries $\mathbf{D_s}, \mathbf{D_c}$ via an additional 2D adversarial loss. More specifically, we follow the auto-decoder architecture [52] and perform PCA on the reshaped dictionary to compute eigenvectors $\mathbf{V} \in \mathbb{R}^{D \times (MF)}$ and fit a normal distribution to the D-dimensional PCA coefficients of N samples. A new *random* codebook $\mathbf{C}_r$ can then be generated by sampling D-dimensional PCA parameters and multiplying them with the eigenvectors $\mathbf{V}$ (See Supp-B.1 for details). Note that our representation disentangles shapes from appearances with separated geometry and texture branches, which enables independent sampling of geometry and texture features.

3.3. Model Training

3D reconstruction loss. To train a codebook $\mathbf{C}_i$ with a single scan, we sample points in a thin shell around the scan. For each coordinate we compute its signed distance s to the input scan, closest texture color $\mathbf{c}$, and surface normal $\mathbf{n}$ on the input scan to attain ground truth values. The codebooks and the decoder weights are then optimized via the following losses:

$$\mathcal{L}_{sdf} = \|s - s(\mathbf{x_g})\|_1 + \lambda_n \|1 - \mathbf{n} \cdot \nabla_{\mathbf{x_g}} s(\mathbf{x_g})\|_1, \tag{4}$$

$$\mathcal{L}_{rgb} = \|\mathbf{c} - c(\mathbf{x_g})\|_1, \tag{5}$$

$$\mathcal{L}_{3D} = \lambda_{sdf} \mathcal{L}_{sdf} + \lambda_{rgb} \mathcal{L}_{rgb}. \tag{6}$$

2D adversarial loss. Adversarial learning does not require exact ground-truth annotations but is trained via a collection of real and fake (rendered) images. Thus, real images are obtained by rasterizing the ground-truth scan to which the coarse body mesh $\mathcal{M}$ was fitted. Color and normal images (denoted as "Real Patch" in Fig. 2) are used for learning texture and geometry respectively. Using the same virtual camera parameters and the coarse mesh $\mathcal{M}$, we attain rendered patches (denoted as "Rendered Patch" Fig. 2) via implicit surface rendering of a sampled codebook $\mathbf{C}_r$. Please refer to Supp-B.2 for more details.

Using these 2D patches, we train dictionaries, decoders, and discriminators jointly with a non-saturating logistic loss $\mathcal{L}_{adv}$ [20], R1 regularization $\mathcal{L}_{R1}$ [38], and path length regularization $\mathcal{L}_{path}$ [27]. Note that these losses do not require exact ground-truth replication. Furthermore, we regularize the feature dictionaries to follow a Gaussian distribution [46] with $\mathcal{L}_{reg} = \|\mathbf{D}\|_F$. In summary, we optimize the discriminator:

$$\mathcal{L}_{dis} = \mathcal{L}_{adv} + \lambda_{R1}\mathcal{L}_{R1}, \tag{7}$$

while updating the remaining components $(\mathbf{D_c}, \mathbf{D_s}, \Phi, \Psi)$:

$$\mathcal{L} = -\mathcal{L}_{adv} + \mathcal{L}_{3D} + \lambda_{path}\mathcal{L}_{path} + \lambda_{reg}\mathcal{L}_{reg}, \tag{8}$$

where $\lambda_{(\cdot)}$ denotes weights to balance the losses.

Since we sample on the fly during training (see Sec. 3.2), the 2D adversarial loss does affect the shared decoders *and* the whole feature dictionaries (See Supp-Fig.14 for details).

3.4. Feature Editing and Avatar Customization

We now describe how we integrate the above-mentioned human representation and the generative architecture into the avatar creation workflow shown in Fig. 1.

Avatar initialization. To simplify the avatar creation process, our method allows users to start with a default example $\mathbf{C}_d$, which can be queried from the trained codebook dictionaries directly with index i ($\mathbf{C}_i$) or randomly sampled from the learned D-dimensional PCA parameters distribution ($\mathbf{C}_r$ in Sec. 3.2).

Model fitting. Being able to extract elements of interest or copying from existing 3D assets is necessary for avatar creation and editing. To this end, we leverage a similar technique to GAN inversion [69], where decoder parameters $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are frozen and we only optimize a new feature codebook $\mathbf{C}_{fit}$ to fit a 3D scan.

Given a 3D target scan and its corresponding body model parameters, we calculate the 3D reconstruction loss in Eq. (6) between the prediction conditioned on $\mathbf{C}_{fit}$ and the ground-truth scan, i.e.,

$$\mathbf{C}_{fit} = \arg\min_{\mathbf{C}}(\lambda_{sdf}\mathcal{L}_{sdf} + \lambda_{rgb}\mathcal{L}_{rgb}). \tag{9}$$

Note that our generative neural fields (MLP decoders) are conditioned on descriptive *local* features. We show that it allows us to accurately fits complex clothing geometry and unseen textural patterns in Sec. 4.3.

Cross-subjects feature editing. With multiple codebooks each representing different 3D assets, we can easily transfer local geometry and texture from one avatar to another, for instance, changing the top wear from the fitted scan $\mathbf{C}_{fit}$ to the initial $\mathbf{C}_d$. Recalling that all codebooks are indexed by an identical mesh topology, users can easily retrieve the index numbers of $\mathcal{V}_{body} \subset \mathcal{V}$ via standard mesh visualization tools such as Blender [12]. Finally, swapping of the corresponding rows in $\mathbf{C}_{fit}$ and $\mathbf{C}_d$ given the vertex indices also swaps the local appearance.

Figure 4. **Randomly sampled texture and body geometry from the model trained on THuman2.0**. *Top*: Given any poses, our randomly sampled geometries contain realistic details such as wrinkles in garments and facial expressions. *Bottom*: Given arbitrary poses and body geometries, our model produces reasonable colors on skin, hair, clothes, and pants in each sample (We turn off the shader for visualizing pure texture colors).

Personalized texture drawing. One can further customize clothing by drawing directly onto 2D images. Due to the learned disentangled feature spaces, we are able to update texture features in $\mathbf{C}_{fit}$ while keeping the geometry features unchanged. When fitting 2D images, users draw on the rasterized images from arbitrary target scans. We then finetune only texture features in $\mathbf{C}_{fit}$ via the RGB loss in Eq. (5) given the corresponding 3D coordinates and colors.

Avatar reposing. Our representation combines a base mesh $\mathcal{M}$ and underlying feature codebooks $\mathbf{C}_{fit}$ that learn *local* geometry and texture on the 3D scans. Hence, one can repose the mesh $\mathcal{M}$ using (θ, β) parameters, which also reposes the avatar correspondingly. Since the geometry codebook does not contain global pose (θ, β), local information can be consistently applied to $\mathcal{M}$ under unseen poses.

4. Experiments

Our goal is locally editable 3D avatar creation. Since we are the first to discuss this problem, we visualize our editing results in Sec. 4.2. Next, we highlight the capability of model fitting by comparing our method with SOTA human generative models in Sec. 4.3. Finally, controlled experiments are presented in Sec. 4.4 and Sec. 4.5 to verify the effectiveness of our design.

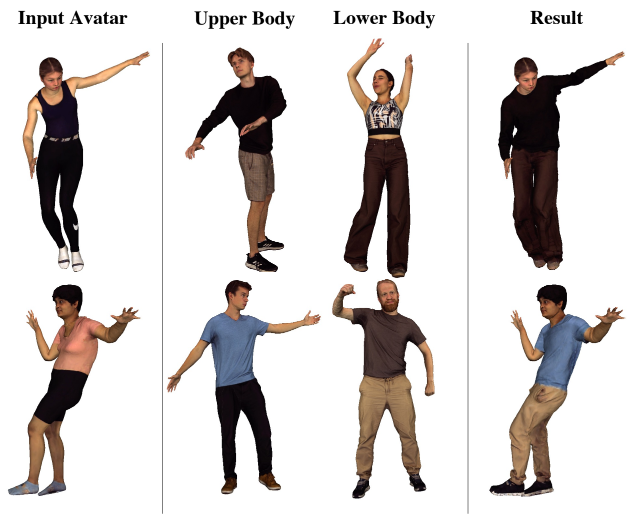

Figure 5. **Cross-subject feature editing results**. We partially transfer local clothing details from the unseen scans (upper and lower body) to the input avatars. The results of the edited avatars are shown in the right column.

4.1. Experiment Settings

Dataset. Most generative human works [9, 44, 45] exploit commercial data [1, 2] for training, which is not easily accessible and limits reproducibility. Furthermore, the quality of publicly available 3D human datasets [66, 73] is not satisfactory. Issues such as non-watertight topologies and noise are very common (See Supp-A.2 for examples and comparison). To bridge this gap, we collect a new dataset named **CustomHumans** for training and evaluation. Here we summarize the datasets used in our experiments.

- **CustomHumans** (Ours) contains more than 600 high-quality scans of 80 participants in 120 garments in varied poses from a volumetric capture stage [11], which is equipped with 106 synchronized cameras (53 RGB and 53 IR cameras). We use our dataset to train models of all quantitative experiments. (Sec. 4.3 $\sim$ Sec. 4.5)
- **THuman2.0** [73] is a dataset containing about 500 scans of humans wearing 150 garments in various poses. Since this dataset has more textural diversity, we train our method on it for qualitative random sampling experiments (Sec. 4.2 and Sec. 4.4).
- **SIZER** [66] is a widely used 3D scan dataset containing A-pose human meshes of 97 subjects in 22 garments. These meshes are used as *unseen* test scans in our fitting experiment (Sec. 4.3).

Evaluation protocol. Following the evaluation protocol in OccNet [39], we quantitatively evaluate the model fitting accuracy using three metrics: **Chamfer distance (CD)**, **normal consistency (NC)**, and **f-Score**.

Figure 6. **Personalized texture editing**. We draw personalized logos on 2D images and fit avatars' texture features to the images. These local textures remain consistent under pose changes.

4.2. Customized Avatars

We visualize the results of our proposed avatar customization workflow described in Sec. 3.4.

Avatar initialization. In Fig. 4, we show random textures and geometries sampled from the model trained on THuman2.0. Our method is able to generate reasonable colors and wrinkles in arbitrary poses. Note that the sampled geometries are shown as the *real* meshes but not as rendered normals as in [9] (See Supp-C.2 for comparisons).

Cross-subjects feature editing. After fitting feature codebooks to 3D scans, we can change the clothes on our avatars by swapping the local features stored on the body vertices. We select the features within the upper body and lower body areas. We then copy these local features to the initial avatars' feature codebooks. As shown in Fig. 5, our method is able to handle multiple garments on different human subjects and preserves consistent details under different body poses or shapes.

Personalized texture drawing. Our method allows users to draw complex letters and logos on images for personalized texture editing. We perform the model fitting and feature editing process but only optimize the texture features in the codebooks using user-edited images and the RGB loss (Eq. (5)). Fig. 6 shows that new texture can be seamlessly applied to the 3D avatars. It is worth noting that resulting avatars enable detailed pose control via the SMPL-X parameters without affecting the fitted texture and geometry.

4.3. Model Fitting Comparison

Since model fitting is an important step in our avatar creation workflow, we compare the capability of feature inversion using unseen 3D scans. The goal of this task is to invert a 3D scan into latent codes while keeping the remaining model parameters fixed. We compare our method with the 3D human generative model gDNA [9], which has achieved state-of-the-art performance in fitting the geometry of 3D human bodies. We also directly compare with SMPL [33] and SMPL+D [3]. Note that SMPL+D is a stronger vertex-based extension that uses a subdivided version of SMPL to directly register surfaces to scans while our method and gDNA optimize latent codes.

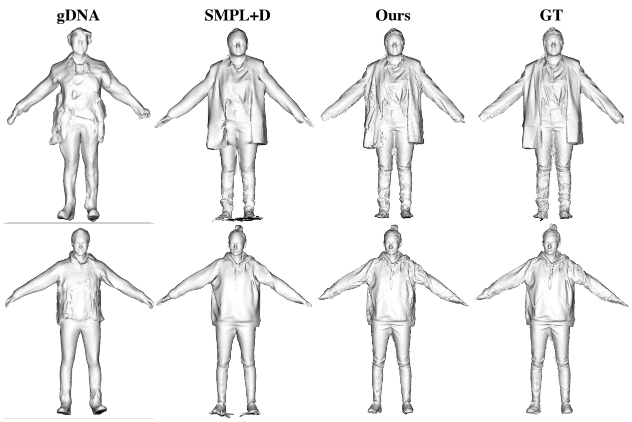

Figure 7. **Qualitative comparison of model fitting on SIZER**. We visualize the fitting results from gDNA [9], SMPL+D [3], and our method. Our results are perceptually close to the ground truth even on the challenging test cases of jackets and loose t-shirts.

Figure 8. **Qualitative comparison of texture fitting**. We compare our method with SMPL+D [3] by fitting to unseen textured meshes. The performance of SMPL+D is limited by its geometry and texture resolution.

From Fig. 7 we can see that SMPL+D handles loose clothing, such as a business suite, better than gDNA. However, the surfaces of SMPL+D results are over-smoothed and do not contain high-frequency details while ours can preserve them. Quantitatively, our method consistently outperforms these methods on all metrics as shown in Tab. 1.

Fig. 8 depicts the result of texture fitting against the SMPL+D baseline. While both methods inherit a fixed mesh topology the quality of SMPL+D is limited by its model resolution. Our method addresses this issue via local neural fields that enable cross-subject feature editing of texture and geometry with enhanced representational power.

4.4. Ablation Study

Effectiveness of local features and shared decoders. To verify the design choice of using local features, we replace the local features $\mathbf{x}_l$ by the global coordinates $\mathbf{x}_g$ for conditioning the decoders. Fig. 9 shows that even though the shared decoders are able to achieve similar reconstruction results when training with global information, they do not maintain consistent performance for model fitting and avatar reposing. This is because the shared decoders tend to memorize global coordinates information in a "per-subject" manner, rather than learning shareable information that can be used across vertices and subjects. On the other hand, our representation ensures only local features defined on the triangle coordinates are exposed to the shared decoders. In such cases, the decoders can better handle unseen body poses or out-of-distribution samples for model fitting and avatar editing.

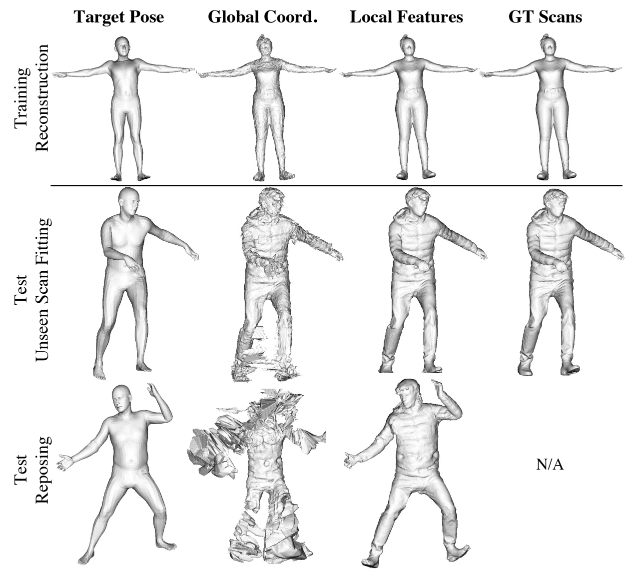

Figure 9. **Comparison of using global/local features for conditioning decoders**. The use of global coordinates causes overfitting and memorization to the shared decoders, which makes it struggle to handle unseen scans or novel body poses.

Method	Pred-to-Scan / Scan-to-Pred (mm)↓	NC↑	f-Score↑
SMPL [33]	13.60 / 18.03	0.849	0.458
gDNA [9]	8.374 / 8.006	0.842	0.718
SMPL+D [3]	5.192 / 2.854	0.911	0.962
Ours	**1.364 / 1.423**	**0.949**	**0.997**

Table 1. **Model fitting comparison on SIZER**. We report Chamfer distance, normal consistency (NC), and f-score between ground truth and the meshes fitted by different methods.

Importance of 2D adversarial loss and 3D disentanglement. As discussed in Sec. 3.3, we introduce feature disentanglement and generative adversarial learning in our training framework. As shown in Fig. 10, sampling within the feature spaces learned without adversarial loss does not yield reasonable body textures. Similarly, training only a single decoder for both geometry and texture does not allow us to maintain desired body geometries when sampling random textures. Our full model can produce disentangled textures, given arbitrary body geometries and poses.

Training Data Percentage	10%	25%	50 %	75%	100%
Chamfer Distance (mm) S-to-P / P-to-S ↓	1.933 / 1.798	1.754 / 1.590	1.543 / 1.456	1.463 / 1.385	1.423 / 1.364
Normal Consistency↑	0.918	0.931	0.935	0.947	0.949
f-Score (%) ↑	99.25	99.38	99.65	99.74	99.75

Table 2. **Generalization analysis on CustomHumans**. We analyze the model fitting performances with regard to different amounts of training data (100% = 100 training scans). We observe consistent performance gain on all evaluation metrics when using more training subjects to train the shared decoders.

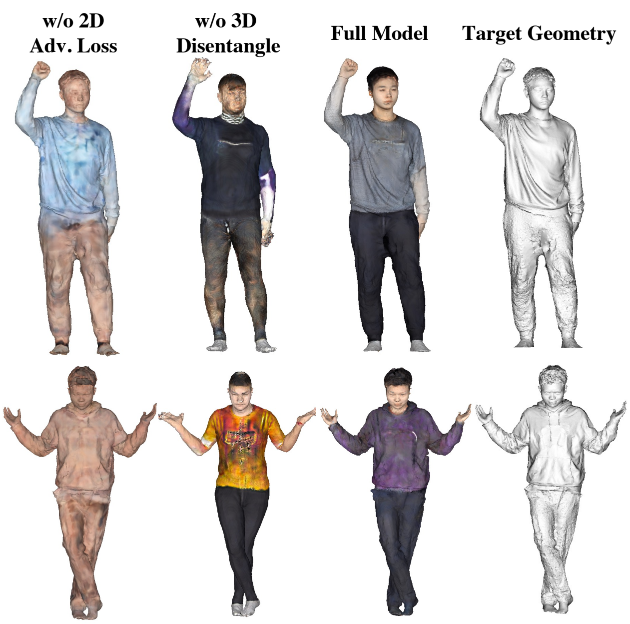

Figure 10. **Ablative comparison of our framework designs**. We visualize the results of transferring random texture to given body geometry. Our full model produces reasonable body texture and is able to maintain fixed geometry for texture editing.

4.5. Generalization Ability Analysis

We are interested in how the amount of training data affects the capacities of the MLP decoders. To analyze this, we design three evaluation protocols: 3D model fitting, avatar reposing, and 2D texture fitting. Tab. 2 summarizes the model fitting performance using different percentages of training data. We observe a 25% accuracy improvement when using the full training set. In Fig. 11 (*Top*) we show that the reposing artifacts caused by self-contact (e.g., fist and elbow) can be reduced when training the MLP decoders with more poses and subjects. In addition, Fig. 11 (*Bottom*) depicts a qualitative comparison of 2D texture editing under different training data percentages. We evaluate texture editing quality by fitting a 2D image with unseen geometric shapes and colors. It can be seen that the model trained on more samples is able to handle a wider range of color distribution. These results confirm the necessity for learning multi-subject shared decoders in our task.

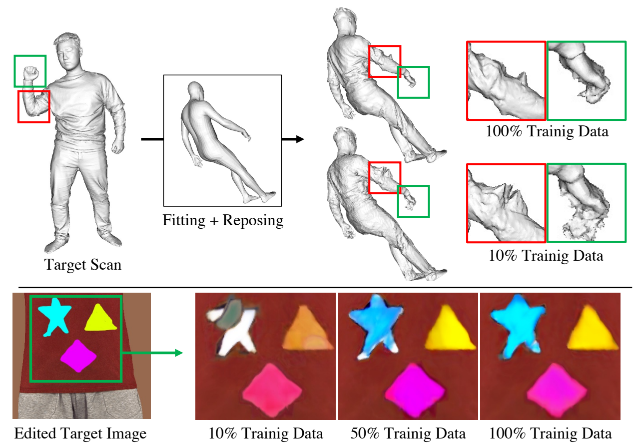

Figure 11. **Qualitative comparison of generalization**. *Top*: We visualize the results of avatar reposing using different percentages of training data (i.e., 10 meshes vs 100 meshes). Artifacts caused by self-contact (e.g., fist and elbow) can be reduced when more training subjects are introduced. *Bottom*: We visualize the results using different percentages of training data. Using more data results in more robust shared decoders (with a wider color range), which is necessary for the avatar creation task.

5. Conclusion

We propose an end-to-end trainable framework for learning 3D human avatars with high fidelity and full editability. By combining neural fields with explicit skinned meshes, our representation addresses the controllability issue of many previous implicit representations. Moreover, we uniquely integrate the proposed human representation into a generative auto-decoding pipeline that enables local editing across *multiple* animation-ready avatars. Through our evaluation on the newly contributed CustomHumans dataset, we demonstrate that our approach achieves higher model fitting accuracy and generates diverse detailed avatars. We believe that this work opens up exciting possibilities for accelerating content creation in the Metaverse.

Acknowledgements We express our gratitude to Stefan Walter and Dean Bakker for infrastructure support, Juan Zarate for managing the capture stage, and Deniz Yildiz and Laura Wülfroth for data capture assistance. We thank Andrew Searle for supporting the capturing system and all the dataset participants.

References

[1] https://3dpeople.com/. 6

[2] https://renderpeople.com/. 6

[3] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1175–1186, 2019. 2, 6, 7

[4] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[5] Alexander W Bergman, Petr Kellnhofer, Yifan Wang, Eric R Chan, David B Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 2

[6] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 311–329. Springer, 2020. 2

[7] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5933–5942, 2019. 2

[8] Xu Chen, Tianjian Jiang, Jie Song, Max Rietmann, Andreas Geiger, Michael J. Black, and Otmar Hilliges. Fast-SNARF: A fast deformer for articulated neural fields. *arXiv*, abs/2211.15601, 2022. 2

[9] Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J Black, Andreas Geiger, and Otmar Hilliges. gdna: Towards generative detailed neural avatars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20427–20437, 2022. 1, 2, 6, 7

[10] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[11] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (TOG)*, 34(4):1–13, 2015. 2, 6

[12] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 5

[13] Enric Corona, Albert Pumarola, Guillem Alenyà, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2

[14] Boyang Deng, JP Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Neural articulated shape approximation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, August 2020. 2

[15] Terrance DeVries, Miguel Angel Bautista, Nitish Srivastava, Graham W Taylor, and Joshua M Susskind. Unconstrained scene generation with locally conditioned radiance fields. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 14304–14313, 2021. 2

[16] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bowen Wu, Bing-Cheng Chen, and Jian Yin. Fw-gan: Flow-navigated warping gan for video virtual try-on. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1161–1170, 2019. 2

[17] Xin Dong, Fuwei Zhao, Zhenyu Xie, Xijin Zhang, Daniel K. Du, Min Zheng, Xiang Long, Xiaodan Liang, and Jianchao Yang. Dressing in the wild by watching dance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3480–3489, June 2022. 2

[18] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–19. Springer, 2022. 2

[19] Rony Goldenthal, David Harmon, Raanan Fattal, Michel Bercovier, and Eitan Grinspun. Efficient simulation of inextensible cloth. *ACM Transactions on Graphics (TOG)*, pages 49–es, 2007. 2

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2672–2680, 2014. 2, 4

[21] Artur Grigorev, Karim Iskakov, Anastasia Ianina, Renat Bashirov, Ilya Zakharkin, Alexander Vakhitov, and Victor Lempitsky. Stylepeople: A generative model of fullbody human avatars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5151–5160, 2021. 1, 2

[22] Artur Grigorev, Bernhard Thomaszewski, Michael J Black, and Otmar Hilliges. HOOD: Hierarchical graphs for generalized modelling of clothing dynamics. 2023. 2

[23] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[24] Pat Hanrahan and Paul Haeberli. Direct wysiwyg painting and texturing on 3d shapes: (an error occurred during the printing of this article that reversed the print order of pages 118 and 119. while we have corrected the sort order of the 2 pages in the dl, the pdf did not allow us to repaginate the 2 pages.). *ACM Transactions on Graphics (TOG)*, 24(4):215–223, 1990. 2

[25] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. Eva3d: Compositional 3d human generation from 2d image collections. *arXiv preprint arXiv:2210.04888*, 2022. 1, 2

[26] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free

generative adversarial networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:852–863, 2021. 2

[27] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 4

[28] Christoph Lassner, Gerard Pons-Moll, and Peter V Gehler. A generative model of people in clothing. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 853–862, 2017. 2

[29] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[30] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Transactions on Graphics (TOG)*, 40(6):1–16, 2021. 2

[31] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5904–5913, 2019. 2

[32] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1096–1104, 2016. 2

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248:1–248:16, October 2015. 2, 6, 7

[34] Qianli Ma, Shunsuke Saito, Jinlong Yang, Siyu Tang, and Michael J. Black. SCALE: Modeling clothed humans with a surface codec of articulated local elements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16082–16093, June 2021. 2

[35] Qianli Ma, Jinlong Yang, Michael J. Black, and Siyu Tang. Neural point-based shape modeling of humans in challenging clothing. In *2022 International Conference on 3D Vision (3DV)*, September 2022. 2

[36] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to Dress 3D People in Generative Clothing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2

[37] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J. Black. The power of points for modeling humans in clothing. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, October 2021. 2

[38] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International Conference on Machine Learning (ICML)*, pages 3481–3490. PMLR, 2018. 4

[39] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4460–4470, 2019. 6

[40] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. 2

[41] Rahul Narain, Armin Samii, and James F O'brien. Adaptive anisotropic remeshing for cloth simulation. *ACM Transactions on Graphics (TOG)*, 31(6):1–10, 2012. 2

[42] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Unsupervised learning of efficient geometry-aware neural articulated representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1, 2

[43] Ahmed AA Osman, Timo Bolkart, and Michael J Black. Star: Sparse trained articulated human body regressor. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 598–613. Springer, 2020. 2

[44] Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. Npms: Neural parametric models for 3d deformable shapes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 6

[45] Pablo Palafox, Nikolaos Sarafianos, Tony Tung, and Angela Dai. Spams: Structured implicit parametric models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 6

[46] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 165–174, 2019. 2, 5

[47] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. Agora: Avatars in geography optimized for regression analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13468–13478, 2021. 2

[48] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2

[49] Hans Køhling Pedersen. Decorating implicit surfaces. In *ACM Transactions on Graphics (TOG)*, pages 291–300, 1995. 2

[50] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9054–9063, 2021. 2

[51] Sergey Prokudin, Michael J Black, and Javier Romero. Smplpix: Neural avatars from 3d human models. In *Proceed-*

ings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1810–1819, 2021. 2

[52] Daniel Rebain, Mark Matthews, Kwang Moo Yi, Dmitry Lagun, and Andrea Tagliasacchi. Lolnerf: Learn from one look. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1558–1567, 2022. 2, 4

[53] Edoardo Remelli, Timur Bagautdinov, Shunsuke Saito, Chenglei Wu, Tomas Simon, Shih-En Wei, Kaiwen Guo, Zhe Cao, Fabian Prada, Jason Saragih, et al. Drivable volumetric avatars using texel-aligned features. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 2

[54] Daniel Roich, Ron Mokady, Amit H. Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 42(1), aug 2022. 2

[55] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 36(6), Nov. 2017. 2

[56] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2

[57] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 84–93, 2020. 2

[58] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 2

[59] Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Zarate, Julien Valentin, Jie Song, and Otmar Hilliges. X-avatar: Expressive human avatars. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[60] Yichun Shi, Divyansh Aggarwal, and Anil K Jain. Lifting 2d stylegan for 3d-aware face generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6258–6266, 2021. 2

[61] Zifan Shi, Sida Peng, Yinghao Xu, Yiyi Liao, and Yujun Shen. Deep generative models on 3d representations: A survey. *arXiv preprint arXiv:2210.15663*, 2022. 2

[62] Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3D shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[63] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. State of the art on neural rendering. In *Annual Conference of the European Association for Computer Graphics (EUROGRAPHICS)*, volume 39, pages 701–727. Wiley Online Library, 2020. 2

[64] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, W Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. Advances in neural rendering. In *Annual Conference of the European Association for Computer Graphics (EUROGRAPHICS)*, volume 41, pages 703–735. Wiley Online Library, 2022. 2

[65] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 2

[66] Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. Sizer: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, August 2020. 1, 6

[67] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 109–117, 2017. 2

[68] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10039–10049, 2021. 2

[69] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *TPAMI*, 2022. 5

[70] 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. In *Annual Conference of the European Association for Computer Graphics (EUROGRAPHICS)*, volume 41, pages 641–676. Wiley Online Library, 2022. 2

[71] Zhenyu Xie, Zaiyu Huang, Fuwei Zhao, Haoye Dong, Michael Kampffmeyer, and Xiaodan Liang. Towards scalable unpaired virtual try-on via patch-routed spatially-adaptive gan. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:2598–2610, 2021. 2

[72] Bangbang Yang, Chong Bao, Junyi Zeng, Hujun Bao, Yinda Zhang, Zhaopeng Cui, and Guofeng Zhang. Neumesh: Learning disentangled neural mesh-based implicit field for geometry and texture editing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 597–614. Springer, 2022. 2

[73] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 2, 6

[74] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. Humbi: A large multiview dataset of human body expressions. In *Proceedings of the IEEE Conference on Computer*

Vision and Pattern Recognition (CVPR), pages 2990–3000, 2020. 2

[75] Jianfeng Zhang, Zihang Jiang, Dingdong Yang, Hongyi Xu, Yichun Shi, Guoxian Song, Zhongcong Xu, Xinchao Wang, and Jiashi Feng. Avatargen: A 3d generative model for animatable human avatars. In *Arxiv*, 2022. 1, 2

[76] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. I M Avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2