AttriHuman-3D: Editable 3D Human Avatar Generation with Attribute Decomposition and Indexing

Fan Yang[1] Tianyi Chen[2] Xiaosheng He[1] Zhongang Cai[1,3]
Lei Yang[3] Si Wu[2] Guosheng Lin[1*]
[1]S-Lab, Nanyang Technological University
[2]School of Computer Science and Engineering, South China University of Technology
[3]SenseTime Research
fan007@e.ntu.edu.sg,csttychen@mail.scut.edu.cn, gslin@ntu.edu.sg

Abstract

Editable 3D-aware generation, which supports user-interacted editing, has witnessed rapid development recently. However, existing editable 3D GANs either fail to achieve high-accuracy local editing or suffer from huge computational costs. We propose AttriHuman-3D, an editable 3D human generation model, which address the aforementioned problems with attribute decomposition and indexing. The core idea of the proposed model is to generate all attributes (e.g. human body, hair, clothes and so on) in an overall attribute space with six feature planes, which are then decomposed and manipulated with different attribute indexes. To precisely extract features of different attributes from the generated feature planes, we propose a novel attribute indexing method as well as an orthogonal projection regularization to enhance the disentanglement. We also introduce a hyper-latent training strategy and an attribute-specific sampling strategy to avoid style entanglement and misleading punishment from the discriminator. Our method allows users to interactively edit selected attributes in the generated 3D human avatars while keeping others fixed. Both qualitative and quantitative experiments demonstrate that our model provides a strong disentanglement between different attributes, allows fine-grained image editing and generates high-quality 3D human avatars.

1. Introduction

As the increasing demand for digital content creation in various domains such as gaming, virtual reality and e-commerce, the need for realistic and customizable 3D human avatars has grown significantly. Traditional 2D editable generation models control the generation style with the manipulation of high-level style latent codes [18–20], which is ambiguous and hard to achieve precise local editing, limiting its application in the real-world scenario. Recently, several approaches have been proposed to enable user manipulation on the free-view generation results of 3D-aware GANs for rigid objects [22, 34, 35] (e.g. human faces and cars). However, editable 3D-aware generation for human avatars could be more challenging due to its high variance in human geometry and appearance.

Some existing editable 3D-aware GANs for human portraits represent the overall regions of the generated objects with a single NeRF and perform local editing by asking users to manually edit the generated semantic masks [34, 35]. Although being very efficient, these methods suffer from the high variance of human shapes, poses and outfits. Directly manipulating the semantic masks requires strong painting skills for the users and involve a lot of tedious manual works, making it hard to be applied in the real-world scenarios. One possible solution is to model and manipulate different semantic regions independently. CN-eRF [22] proposes to model each semantic region with different generators. Although generating different semantic parts independently, training several deep generators in parallel will extremely increase the memory footprint, hindering their applications in real-world scenarios. Moreover, currently available human datasets are limited and almost all of them suffer from the style entanglement between different attributes, which means the style of one attribute may be highly influenced by the style of another attribute. For example, some pairs always show in the existing datasets: dresses always with high-heel shoes, t-shirts always with sports shoes. Men rarely with dresses. Such correlation may also be learned by the discriminator, leading to misleading punishment in the training process and artifacts in the editing stage.

To address these problems, we propose AttriHuman-

*Guosheng Lin is the corresponding author.

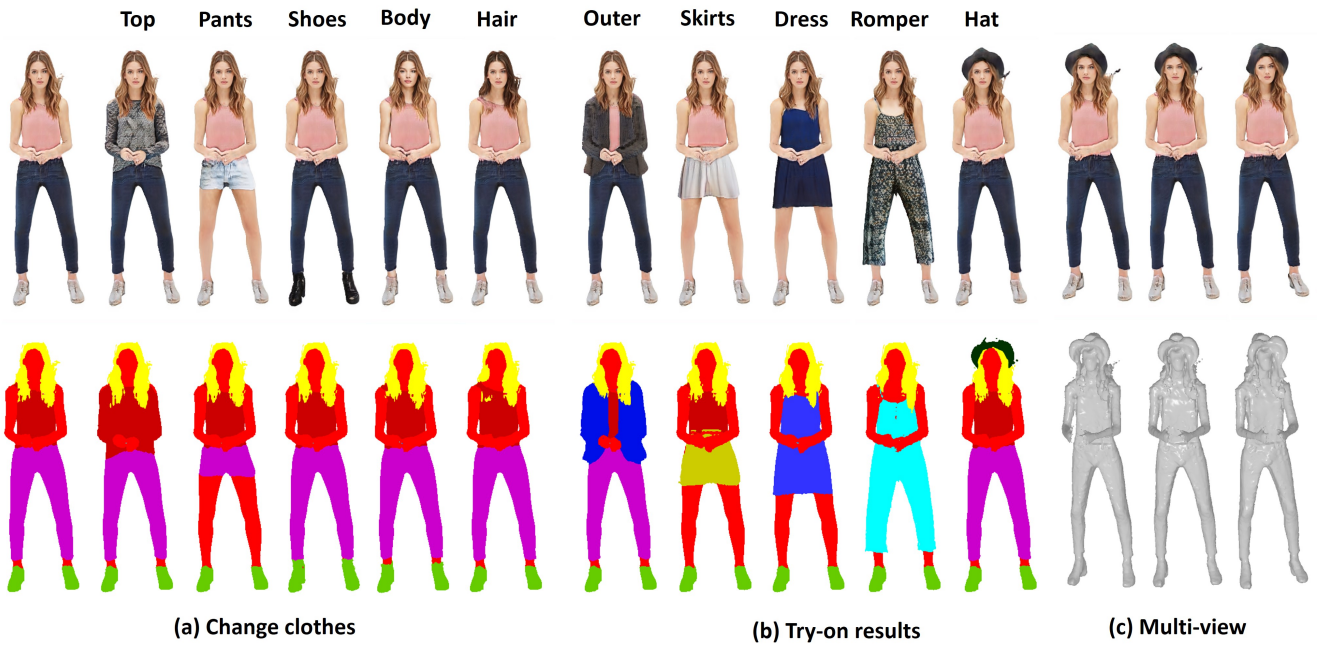

Figure 1. Our AttriHuman-3D achieves strong disentanglement between different attributes, generates high-quality view-consistent 3D human avatars which allows fine-grained editing. From left to right, we show the generation results by editing different attributes, try-on results by modifying the selected attribute sets and generation results of view-consistent images.

3D, an editable 3D human generation model with attribute decomposition and indexing. Specifically, we propose to generate all attributes (e.g. hair, clothes, main body and so on) in an overall feature space with six feature planes and then extract independent attribute-specific features with different attribute indexes. Different from CNeRF [22] that generates each attribute with an independent generator, our method benefits from the "all-to-one" design, facilitating the information sharing across different attributes and achieving much higher computational efficiency. Our method is inspired by the recently proposed tensor decomposition technologies in 4D dynamic NeRFs [3, 9], we formulate the generation space as a 4D space-attribute field which could be decomposed into six feature planes and efficiently generate with a single 2D CNN-based feature generator. However, trivially extending the tensor decomposition technologies in space-time field into space-attribute domain leads to degraded generation results. Different from time dimension which is continuous and has clear definition, the attribute dimension could be disjoint and ambiguous, making it hard to define the index of different attributes. To address this problem, we further propose an implicit indexing module to learn the index of each human attribute in the attribute dimension. To enhance the disentanglement of different attributes, we further propose an orthogonal projection regularization to enforce the orthogonality of different indexes, which enables us to extract independent attribute features from the generated feature planes.

To address the implicit style entanglement between different attributes in the existing datasets, we introduce a hyper-latent training strategy which conditions the overall training progress with hyper-latent label to avoid the misleading punishment from the discriminator and an attribute-specific sampling strategy to split different attributes with pre-defined sampling bounding boxes to remove the influence of other attributes defined in different bounding boxes.

Qualitative and quantitative experiments conducted on the fashion datasets demonstrate that our model provides a strong disentanglement between different attributes, allows fine-grained image editing and generates high-quality view-consistent 3D human avatars. In summary, our main contributions can be listed as follows: 1) We propose a novel editable 3D human avatar generation model, which achieves fully disentangled control over the generated human avatars with attribute decomposition and indexing. 2) We propose a novel implicit indexing method with an orthogonal projection regularization to enhance the disentanglement of different attributes. 3) We introduce a hyper-latent training strategy and an attribute-specific sampling strategy to address the style entanglement between different attributes in the existing human datasets, leading to better editing performance.

2. Related Work

2.1. 3D-Aware GANs

Generative Adversarial Networks (GANs) [11] have achieved remarkable success in the field of 2D image synthesis [2, 17–19, 37]. Recently, the implicit neural representation have been widely explored to be combined with GANs to learn 3D-aware image synthesis with only 2D images as supervision [5, 6, 25, 27, 32, 33]. Particularly, GRAF [31] and Pi-GAN [5] propose to learn 3D-aware image and geometry generation with implicit neural radiance fields as generators. However, these methods are limited to synthesis only low-resolution images due to the large computational costs of the MLP querying in the volumne rendering process. To solve this problem, recent methods tend to adopt a two-stage generation process by firstly rendering low-resolution images and then passing it to 2D CNN-based decoders to generate images with higher resolution [6, 25, 27]. Although providing high-quality generation of 3D-aware images, these methods fail to support user-interacted local editings. In the task of 3D human avatar generation, most of the recent works generate the 3D human avatars with predefined parametic human templates [1, 8, 14, 16, 26, 39–41]. Among them, EVA3D [14] proposes a compositional human NeRF representation which divides the human body into different local parts and achieves high-resolution human image generation. However, all of them do not support user-interacted editing of generated human avatars. Some recent works [4, 12, 15, 28] also explore using text-image embedding to edit the generation results. Although achieving good performance, these methods can not achieve fully disentangled control over different semantic parts and suffer from long optimization time.

2.2. Spatial decomposition

NeRF proposes to use fully implicit networks to represent 3D density and color fields, which is slow in querying and computational infeasible to generate high-resolution images directly. A recent trend is to reduce the rendering time of NeRF with explicit geometric representations, including sparse 3D grids [38], point clouds [36] and so on. However, there is a tradeoff between the rendering speed and the memory costs of different geometric representations. Although explicit methods reduce the optimization time, it requires more memory footprint.

Hybrid representation [3, 6, 7, 9, 24] has been proposed to improve the computational efficiency of NeRF by combining the explicit representations with implicit layers. In particular, TensoRF [7] proposes to decompose the voxel tensor into feature planes and vectors, which greatly improves the computational efficiency. While all these methods focus on the decomposition of 3D feature volumes. Hexplane [3] further extends it into 4D dynamic scenes. They decompose the 4D space-time volume into six feature planes spanning each pair of coordinate axes (e.g. XY, ZT).

3. Method

3.1. Efficient 4D Space-Attribute Decomposition

we formulate the generated human avatars in a 4D space-attribute field (e.g. hair, clothes, body shapes and so on), which could be denoted as $(x, y, z, a) \in \mathbb{R}^4$ where $a \in \{clothes, hair, shoes...\}$ denotes the attribute dimension. Given a selected attribute $attr_i$ and its attribute index $a_i = Index(attr_i)$, we are capable to extract the corresponding attribute features independently from the generated feature planes. Recently, CNeRF [22] propose to generate the 4D volumes $(x, y, z, a_i), i \in \{1, 2, ..., N\}$ with N independent generators. However, training several independent generators in parallel will exponentially increase the computational cost. Inspired by the tensor decomposition technologies in 4D dynamic NeRFs [3], we propose to decompose the 4D space-attribute field into six feature planes. Specifically, we decompose the 4D space-attribute volume $V \in \mathbb{R}^{XYZAF}$ into six planes : P_r^{XY}, P_r^{YZ}, P_r^{XZ}, P_r^{XA}, P_r^{YA}, P_r^{ZA}, and represent the space-attribute fields D as:

$$\begin{aligned} D(x, y, z, a) = &(P_{xy}^{XYR_1} \odot P_{za}^{ZAR_1})V^{R_1F} \\ &+(P_{xz}^{XZR_2} \odot P_{ya}^{YAR_2})V^{R_2F} \\ &+(P_{yz}^{YZR_3} \odot P_{xa}^{XAR_3})V^{R_3F}, \end{aligned} \tag{1}$$

where $\odot$ denotes the element-wise product, $P_{za}^{ZAR_1}$ denotes a 3D tensor with shape $Z \times A \times R_1$ and V^{R_1F} denotes a 2D tensor with shape $R_1 \times F$. In our experiments, we find that relaxing Eq 1 by setting V^{R_1F}, V^{R_2F} and V^{R_3F} to be constant leading to faster convergence without losing performance. Therefore, we set V^{R_1F}, V^{R_2F} and V^{R_3F} to be one in our all experiments.

Leveraging the proposed tensor decomposition technology, we are capable to efficiently generate the decomposed feature planes with a single 2D CNN-based feature generator (like StyleGan2 [19]), which greatly reduces the computational cost as well as facilitates information sharing across different attributes and encourages consistency among the generated features of different attributes.

3.2. Implicit Attribute Indexing

We propose implicit attribute indexing module to predict the indexes of different attributes. As the attribute dimension could be disjoint and ambiguous, trivially adopting the identical mapping indexing method as dynamic NeRFs [3] leads to degraded generation results (Figure 5). To address this problem, we propose an implicit indexing module to learn the index of each predefined attribute in the attribute

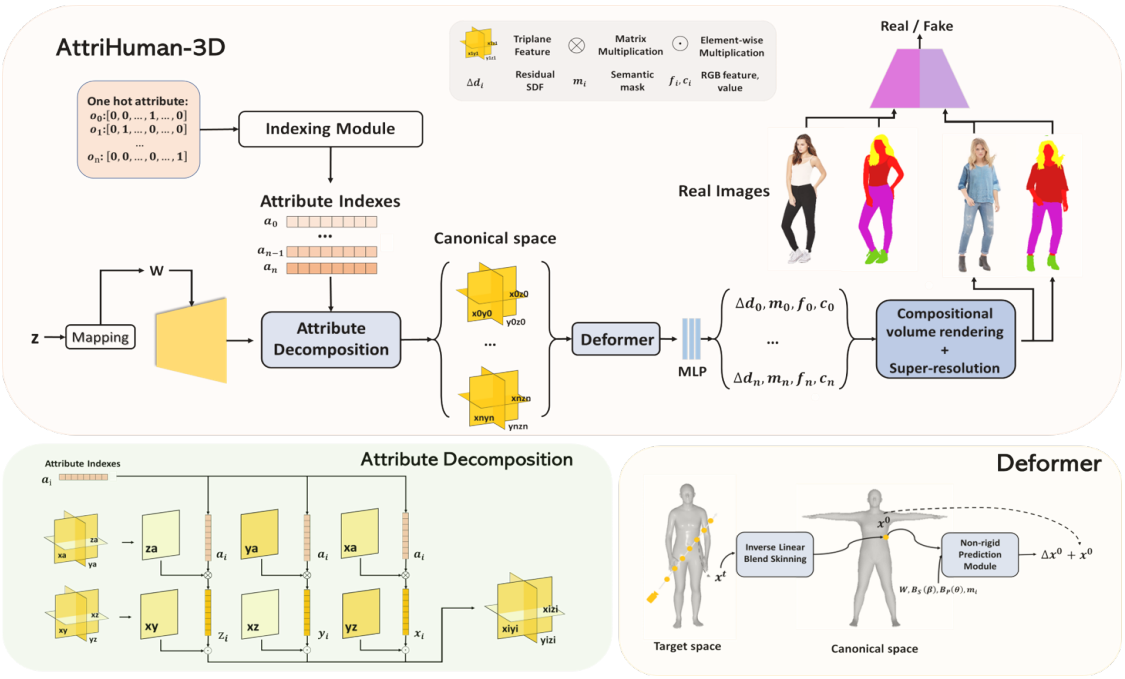

Figure 2. The overall framework of our model. We generate the decomposed feature plane with StyleGANv2-based generator and predict the indexes of each attribute with implicit indexing module. We model the deformation between canonical space and target space with deformer module and synthesis final image with compositional volume rendering and super-resolution module. Detailed structure of the attribute decompose module and deformer module is shown at the bottom, where $B_s(\beta)$, $B_P(\theta)$ represents the SMPL parameters randomly sampled from the dataset, n denotes the total number of selected attributes.

dimension. In detail, the implicit indexing module I is formulated as an MLP network with eight layers. The input for the indexing module is a one-hot label $o_i \in \mathbb{R}^N$, where N is the number of total attributes. To stabilize the training, we normalize the predicted index before using it for attribute indexing. The total function of the implicit indexing module can be simply formulated as $a_i = norm(I(o_i))$, where $norm$ denotes the $L2$ normalization. The independent feature D_i for attribute i can then be extracted from the generated feature fields D as:

$$\begin{aligned} D_i = &(P_{xy}^{XYR_1} \odot (P_{za}^{ZAR_1} \times a_i^{AR_1})) \\ &+(P_{xz}^{XZR_2} \odot (P_{ya}^{YAR_2} \times a_i^{AR_2})) \\ &+(P_{yz}^{YZR_3} \odot (P_{xa}^{XAR_3} \times a_i^{AR_3})). \end{aligned} \tag{2}$$

Note we omit V^* here for simplicity. During our experiments, we found that learning the indexing module in an unsupervised style fails to achieve strong disentanglement between different attributes, leading to collapsed geometry and chaotic textures in the editing stage. To address this problem, we further propose an orthogonal projection regularization to enforce the orthogonality between indexes for different attributes.

Orthogonal Projection Regularization. Given a predicted index a_i, the orthogonal projection regularization could be defined as:

$$L_{orth} = \sum_{\substack{i,j \in B \\ i \neq j}} |\langle a_i, a_j \rangle|, \qquad \langle x_i, x_j \rangle = \frac{x_i \cdot x_j}{\|x_i\|_2 \cdot \|x_j\|_2}, \tag{3}$$

which involves no additional learnable parameters and can be applied directly on the mini-batch level for each iteration. In our experiments, we found that the orthogonal projection regularization is crucial to enforce the implicit indexing module to learn orthogonal attribute indices, facilitates the disentanglement between different attributes and leads to better editing results.

3.3. Overall Framework

Our framework is built on EG3D [6], involving a StyleGAN2-based [19, 34] feature generator, an implicit attribute indexing module, a compositional volume renderer and a 2D CNN-based up-sampler. The overall framework of our model is shown in Figure 2.

In detail, we adopt StyleGAN2-based generator to generate the overall feature planes. To achieve better generation quality and efficiency, we adopt an improved version of generator structure following IDE3d [34]. We then sample a set of attributes O_i from the real-world dataset, where $O_i = (o_0, ..., o_k), k < N$, N represents the total number

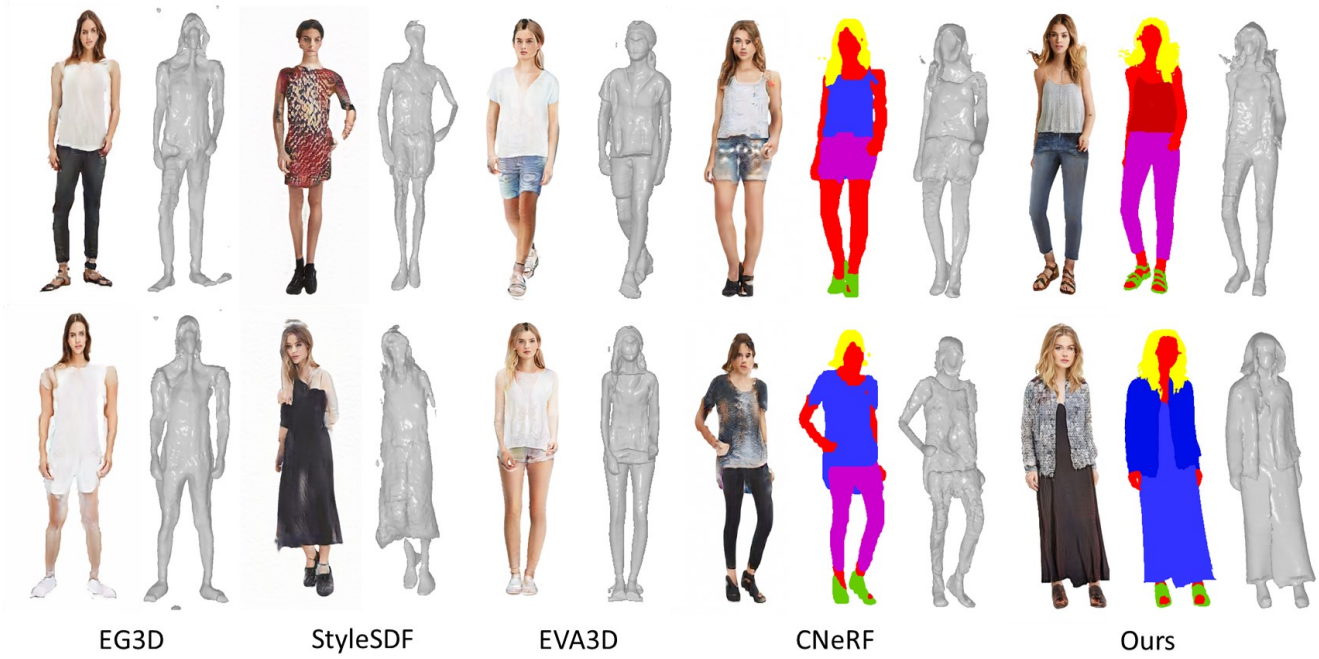

Figure 3. Qualitative Comparisons of our methods with EG3D, StyleSDF, EVA3D, CNeRF. The RGB images, generated segmentation masks and 3D meshes demonstrated that our method achieves high-quality human avatar generation. Moreover, the main contribution of our model is to support interactive user editing, which is not supported by EG3D, StyleSDF and EVA3D.

of attribute categories predefined on the dataset. In our experiments, we set 11 attribute categories including the human body, hair, shirts, pants, skirts, dress and so on. During the generation process, each attribute is embedded as a one-hot label o_i and is passed into the implicit indexing module to get the learned index in the attribute dimension. The features of each attribute are then extracted from the generated feature planes. Given a sampled point on a certain camera ray, we query its feature vectors by projecting the points onto the axis-aligned extrated feature planes and decode the residual signed distance values Δd_i, semantic masks m_i, RGB features and values f_i, c_i with a multi-head MLP, which shares the weights across different attributes.

We disentangle the pose and shape of human by shaping the generated space-attribute fields in a common canonical space and find the deformation between the canonical space with each observation posed space leveraging the predefined skinning weights on the SMPL human templates. In detail, we define the transformation of a point x^t from observation space t to canonical space as:

$$\begin{bmatrix} x^0 \\ 1 \end{bmatrix} = \sum_{v_i \in \mathcal{N}(x)} \frac{\omega_i}{\sum \omega} M(\beta_0, \theta_0)(M(\beta_t, \theta_t))^{-1} \\ , M(\beta, \theta) = \begin{bmatrix} I & B_S(\beta) + B_P(\theta) \\ 0^T & 1 \end{bmatrix}, \tag{4}$$

where $\mathcal{N}$ represents the nearest k points that are found among the vertices of SMPL posed mesh $M(\beta_t, \theta_t)$, $w_i = 1/\|x^t - v_j\|$ is the transformation weight, $B_S(\beta)$ and $B_P(\theta)$ denote the shape/pose blend shape for vertexes on the SMPL mesh, respectively. Similar to AvatarGen [40], we further involve a per-attribute non-rigid deformation network, which predicts the non-rigid offsets Δx^t, to accommodate the non-rigid deformation between the observation and the canonical spaces. Denote the transformation defined in E.q. 4 as T, the final deformation can be formulated as $x^0 = T(x^t) + \Delta x^t$, $\Delta x^t = MLP(embed(x^0), B_S(\beta), B_P(\theta), m_i)$, where m_i represents the semantic mask for attribute i.

As we formulate all the generation in the canonical space, it's easy for us to conduct editing by directly exchanging the feature planes. To edit certain attribute o_i, we could change the generated feature planes of o_i with another ones generated from new sampled latent codes. In this way, we support precise editing on certain selected attributes while keeping others fixed. We show more details in the supplemental materials.

3.4. Compositional Volume rendering

We fuse and render the output of all attributes with compositional volume rendering. Specifically, we first fuse the outputs of all selected attributes with the predicted semantic masks. Given a sampled points x in attribute i with predicted residual signed distance values Δd_i, semantic mask m_i, RGB features and values f_i, and c_i, the fusion step can

be formulated as:

$$d(x) = d_t(x) + \sum_i^N m_i'(x) \cdot \Delta d_i(x), \tag{5}$$

$$f(x) = \sum_i^N m_i'(x) \cdot f_i(x), \qquad c(x) = \sum_i^N m_i'(x) \cdot c_i(x), \tag{6}$$

where $m_i'(x) = \frac{exp(m_i(x))}{\sum_j^N exp(m_j(x))}$ denotes the softmax operation, d_t represents the template SDF queried from the SMPL template, N denotes the total number of semantic categories. We then adopt volume rendering [23] to render the low-resolution RGB image and semantic mask. The rendered RGB feature along with the segmentation masks are then fed into the super-resolution module to generate the final high-resolution images.

3.5. Training

Hyper-latent Training Strategy. During our experiments, we found that the style entanglement existing in the current human datasets [21]. One of the most significant problems is that almost all datasets have few or no images containing men wearing dresses. However, as attribute is defined according to semantic areas without definition of human genders. Problems come when we generate a body area of men with attributes of dress. In this condition, the generated image would become out-of-distribution to the discriminator and it would give misleading punishment. As a result, the feature plane of attribute dress would try to influence the feature of attribute human body, making it to look like women (Figure 7), leading to slower convergency and artifacts in the editing stage. To address this problem, we condition the linear mapping module of StyleGAN [10] with hyper-label to control the gender of the generated human bodies and sample reasonable attribute sets according to different genders. In this way, we avoid the misleading punishment from the discriminator and achieves better generation results.

Attribute-specific Sampling Strategy. While hyper-latent training strategy is adopted to address the entanglement between attributes with large spatial overlap, like body and dress. We propose attribute-specific sampling strategy to address the entanglement between attributes with little overlap areas. For example, the style entanglement of dresses always with high-heel shoes, T-shirts always with snipers. In detail, we define attribute-specific bounding boxes for each attribute and sample corresponding attribute feature only inside the bounding box. In this way, we limit the influence of each attribute strictly inside the bounding box, enhance the disentanglement between attributes with little or no overlaps and further improve the computational efficiency by sampling much fewer points for each attribute. We present more details of the pre-defined bounding boxes in the supplementary material.

Loss. We use the non-saturating GAN loss L_{gan} with R1 regularization L_{Greg}. We regularize the learned SDFs by pushing the derivation of delta SDF values to be zero: $L_{eik} = \sum_x (\|\nabla d_i(x)\|) - 1)^2$, and further prevent false or non-visible surfaces with $L_{surf} = \sum_x exp(-100|d_i(x)|)$. Moveover, to constrain the generated residual signed distance fields to be consistent with the SMPL template mesh, we guide the predicted residual signed distance fields with a minimum regularization: $L_{rsdf} = \sum_x (\|\Delta d(x)\|)$. To prevent the learned non-rigid deformation from collapsing, we add a non-rigid regularization to regularize the learned non-rigid deformation to be small: $L_{nonrig} = \sum_x (\|\Delta x^t\|)$. The overall loss function is formulated as:

$$\begin{aligned} L_{overall} = L_{gan} + \lambda_{Greg} L_{Greg} + \lambda_{eik} L_{eik} + \lambda_{surf} L_{surf} \\ + \lambda_{rsdf} L_{rsdf} + \lambda_{nonrig} L_{nonrig} + \lambda_{orth} L_{orth}, \end{aligned} \tag{7}$$

where λ_* denotes the weights of each loss item.

4. Experiments

Datasets. We train and evaluate our model mainly on the Deepfashion dataset [21], which contains single-view human images with different clothes and the corresponding semantic masks. We select 11903 images that contain the full human bodies, crop and resize all the images to coarsely align the human bodies. All images are resized to 512×512 for training. We use an off-the-shelf pose estimator [29] to estimate SMPL parameters and camera parameters of each image in the dataset. For each human image, we convert its corresponding semantic masks into attribute sets by checking whether the semantic region is above zero. we adopt 11 attributes for training, including Outer, Top, Skirts, Dress, Pants, Rompers, Hats, Glasses, Body, Shoes and Haircut.

4.1. Comparisons.

Baselines. We compare our AttriHuman-3D model with several state-of-the-art 3D-aware GANs, including EG3D [6], StyleSDF [27], CNeRF [22] and EVA3D [14]. EG3D, STyleSDF and CNeRF are designed for 3D generation of rigid / half-rigid objects such as human faces. Among them, CNeRF is also capable to support user-interacted editing for the generation results, which is similar to our task. EVA3D is also designed for human avatar generation, but it does not support interactive editing, which is the main contribution of our model.

Quantitative Evaluations. The quantitative comparisons between our methods with baselines are shown in Table 1. In detail, we adopt three commonly used image quality evaluation metrics including Frechet Inception Distance (FID) [13] to evaluate the visual quality and diversity of the rendered images, Percentage of Correct Key-

Methods	FID ↓	FID_{edit} ↓	PCK↑	Depth↓	E
EG3D [6]	25.90	-	79.63	0.0331	×
StyleSDF [27]	26.32	-	75.13	0.0376	×
CNeRF [22]	27.69	30.31	71.56	0.0421	✓
EVA3D [14]	15.91	-	87.50	0.0272	×
fixed indexing	21.20	23.58	88.41	0.0357	✓
w/o OPR	19.81	22.31	88.72	0.0322	✓
w/o HL	17.24	18.11	89.63	0.0311	✓
w/o ASP	17.02	17.32	89.97	0.0305	✓
Ours(full)	16.85	17.43	89.91	0.0302	✓

Table 1. Quantitative comparisons of our method with other methods, where OPR denotes orthogonal projection regularization, GS denotes Hyper-Latent training strategy and ASP denotes Attribute-specific Sampling strategy. E represents editable ability.

Methods	Memory	Parameters	Time(s)
EG3D	6G	31M	0.06
StyleSDF	6G	-	0.13
CNeRF	16G	-	0.23
EVA3D	7G	-	0.16
11G-base	Out of Mem	311M	-
Ours(full)	9G	59M	0.09

Table 2. Efficiency comparisons of our method with other methods. Our model is much more efficient compared to CNeRF and 11G-base model without using the proposed tensor decomposition technique.

points (PCK) [1] to evaluate the effectiveness of the pose controllability. Pseudo Depth [30] to evaluate the consistency between the generated geometry and RGB images. As shown in Table 1, our model achieves comparable results to the SOTA method EVA3D, while capable to support user-interacted editing. We also compare the FID of the edited images (FID_{edit}) with CNeRF and the result demonstrates our model achieves better editing results with a little drop in the FID for the edited images. Moreover, we also compare the computational efficiency of our model with others in Table 2, which shows our model only involves limited extra parameters to achieve editing capability. Benefiting from the proposed tensor decomposition technology our model is much higher efficient compared to CNeRF and 11G-base model which directly uses 11 generators for the generation.

Qualitative Evaluations. Figure 3 shows the qualitative comparisons of our methods with other baselines. Specifically, we show the rendered images, segmentation masks and the corresponding meshes. 3D GANs designed for rigid/half-rigid objects including EG3D, StyleSDF and CNeRF fail to learn high-quality generation of human avatars due to the high variance of human poses and appearance. While EVA3D generates realistic 3D human avatars, it fails to support user-interacted editing. Our model achieves comparable generation results to EVA3D while capable to support semantic part-level editing that other methods do not support. We also compare the editing results of our model with CNeRF. As shown in Figure 4, benefit from our hyper-latent and attribute-specific training strategy, our model achieves better disentanglement and more precise control over the target semantic region compared to CNeRF.

Figure 4. Qualitative comparisons of the editing results between our methods and CNeRF. From left to right we show the editing RGB and residual of changing the Top, Pants and Haircut. Benefit from our hyper-latent and attribute-specific training strategy, our model achieves better disentanglement and more precise control over the target semantic region compared to CNeRF.

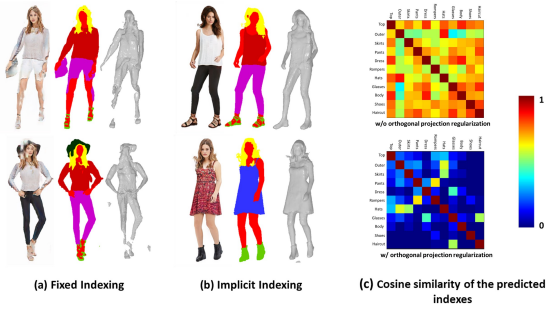

Figure 5. Ablation of the implicit indexing module. Figure (a) and (b) compare the ablation results of our implicit mapping module with fixed identical mapping as dynamic NeRFs [3]. Figure (c) shows cosine similarity of the predicted indexes with or without the proposed orthogonal projection regularization.

4.2. Ablations Studies.

Implicit Indexing Module. Trivially adopting the fixed identical indexing method in dynamic NeRFs [3] leads to degraded generation results as shown in Figure 5. In our implicit indexing module, the proposed orthogonal projection regularization serves an important role in encouraging the predicted implicit attribute indexes to be orthogonal to each other, facilitating the disentanglement between different attributes. As shown in Figure 5(c) , Figure 6, with-

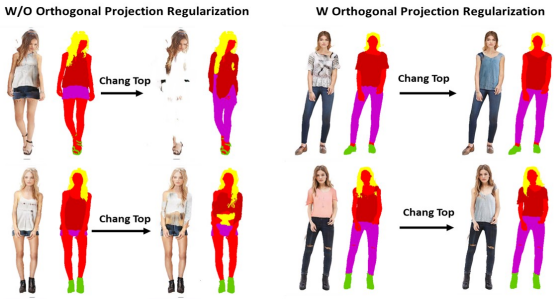

Figure 6. Ablation of the orthogonal projection regularization. we achieve better editing results with the proposed OPR loss.

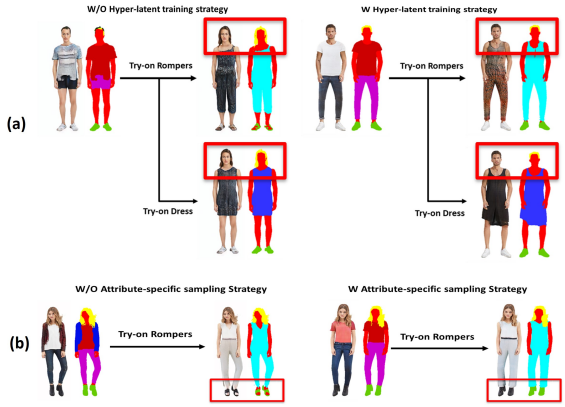

Figure 7. Ablation of the Hyper-latent training strategy and Attribute-specific Sampling training strategy.

out the orthogonal projection regularization, the learned attribute indexes tend be entangled with each other, leading to collapsed results in the editing stage. When we remove the original feature of the top clothes and change it into another one, the network fails to generate satisfactory edited images. We also conduct quantitative comparisons in Table 1. Removing the implicit indexing module or orthogonal projection regularization leads to a significant performance drop in all metrics, which indicates the effectiveness of the proposed implicit module and orthogonal projection regularization.

Training Strategies. During training, we adopt the hyper-latent training strategy and attribute-specific sampling strategy to address the implicit style entanglement between different attributes in the existing dataset. As shown in Figure 7 (a), without the hyper-latent training strategy, the discriminator tend to give misleading penalty to push the attribute features of dresses or rompers to influence the body attribute and try to change the original appearance of the generated bodies to make them visually more suitable for a female subject. With the proposed hyper-latent training strategy, we avoid the misleading punishment from the discriminator and achieve more accurate editing results

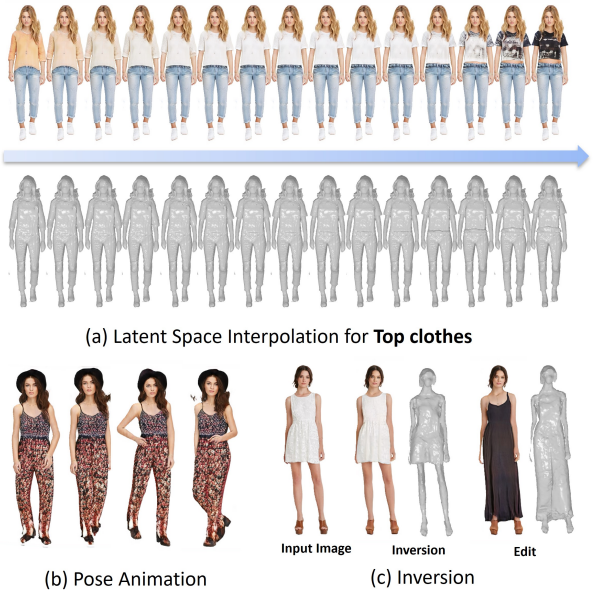

Figure 8. Application of AttriHuman-3D. (a) Interpolation on the latent code of Top Clothes gives smooth transformations between two samples. (b) Pose animation results in a generated avatar. (c) Inversion of real image

without changing the appearance of the generated human. The Attribute-specific sampling strategy enforces the style disentanglement between attributes with little or no overlay such as dress and shoes. As shown in Figure 7, with attribute-specific sampling strategy, we avoid the style influence from rompers on the shoes, and achieve more precise editing results. Quantitative results shown in Table 1 also indicate that the proposed training strategies lead to better generation results.

5. Conclusion

In this paper, we propose AttriHuman-3D, a novel editable 3D human avatar generation model.Our method allows users to interactively edit selected attributes in the generated 3D human avatars while keeping others fixed. Both qualitative and quantitative experiments demonstrate that our model provides a strong disentanglement between different attributes, allows fine-grained image editing and generates high-quality 3D human avatars

6. Acknowledgement

This study is supported under the RIE2020 Industry Alignment Fund - Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). This research is also supported by the MOE AcRF Tier 2 grant (MOE-T2EP20220-0007).

References

[1] Alexander Bergman, Petr Kellnhofer, Wang Yifan, Eric Chan, David Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields. *Advances in Neural Information Processing Systems*, 35:19900–19916, 2022. 3, 7

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natual image synthesis. In *Proc. International Conference on Learning Representation*, 2019. 3

[3] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 130–141, 2023. 2, 3, 7

[4] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. *arXiv preprint arXiv:2304.00916*, 2023. 3

[5] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021. 3

[6] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 3, 4, 6, 7

[7] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 333–350. Springer, 2022. 3

[8] Zijian Dong, Xu Chen, Jinlong Yang, Michael J Black, Otmar Hilliges, and Andreas Geiger. Ag3d: Learning to generate 3d avatars from 2d image collections. *arXiv preprint arXiv:2305.02312*, 2023. 3

[9] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023. 2, 3

[10] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 1–19. Springer, 2022. 6

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. Neural Information Processing Systems*, 2014. 3

[12] Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. *arXiv preprint arXiv:2303.12789*, 2023. 3

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[14] 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. 3, 6, 7

[15] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*, 2022. 3

[16] Suyi Jiang, Haoran Jiang, Ziyu Wang, Haimin Luo, Wenzheng Chen, and Lan Xu. Humangen: Generating human radiance fields with explicit priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12543–12554, 2023. 3

[17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *Proc. International Conference on Learning Representation*, 2018. 3

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1

[19] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3, 4

[20] 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*, 34:852–863, 2021. 1

[21] 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*, pages 1096–1104, 2016. 6

[22] Tianxiang Ma, Bingchuan Li, Qian He, Jing Dong, and Tieniu Tan. Semantic 3d-aware portrait synthesis and manipulation based on compositional neural radiance field. *arXiv preprint arXiv:2302.01579*, 2023. 1, 2, 3, 6, 7

[23] B Mildenhall, PP Srinivasan, M Tancik, JT Barron, R Ramamoorthi, and R Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, 2020. 6

[24] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 3

[25] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021. 3

[26] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Unsupervised learning of efficient geometry-aware neural articulated representations. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 597–614. Springer, 2022. 3

[27] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13503–13513, 2022. 3, 6, 7

[28] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 3

[29] 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/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 6

[30] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 7

[31] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020. 3

[32] Chaoyue Song, Jiacheng Wei, Ruibo Li, Fayao Liu, and Guosheng Lin. 3d pose transfer with correspondence learning and mesh refinement. *Advances in Neural Information Processing Systems*, 34:3108–3120, 2021. 3

[33] Chaoyue Song, Jiacheng Wei, Ruibo Li, Fayao Liu, and Guosheng Lin. Unsupervised 3d pose transfer with cross consistency and dual reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[34] Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *ACM Transactions on Graphics (TOG)*, 41(6):1–10, 2022. 1, 4

[35] Jingxiang Sun, Xuan Wang, Yong Zhang, Xiaoyu Li, Qi Zhang, Yebin Liu, and Jue Wang. Fenerf: Face editing in neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7672–7682, 2022. 1

[36] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5438–5448, 2022. 3

[37] Fan Yang and Guosheng Lin. Ct-net: Complementary transfering network for garment transfer with arbitrary geometric changes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9899–9908, 2021. 3

[38] Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. *arXiv preprint arXiv:2112.05131*, 2021. 3

[39] Jichao Zhang, Enver Sangineto, Hao Tang, Aliaksandr Siarohin, Zhun Zhong, Nicu Sebe, and Wei Wang. 3d-aware semantic-guided generative model for human synthesis. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 339–356. Springer, 2022. 3

[40] 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 *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pages 668–685. Springer, 2023. 5

[41] Xuanmeng Zhang, Jianfeng Zhang, Rohan Chacko, Hongyi Xu, Guoxian Song, Yi Yang, and Jiashi Feng. Getavatar: Generative textured meshes for animatable human avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2273–2282, 2023. 3