HuGeDiff: 3D Human Generation via Diffusion with Gaussian Splatting

Maksym Ivashechkin
m.ivashechkin@surrey.ac.uk
Oscar Mendez
o.mendez@surrey.ac.uk
Richard Bowden
r.bowden@surrey.ac.uk

Centre for Vision, Speech, and Signal Processing (CVSSP)
University of Surrey
Guildford, United Kingdom

Abstract

3D human generation is an important problem with a wide range of applications in computer vision and graphics. Despite recent progress in generative AI, such as diffusion models or rendering methods like Neural Radiance Fields or Gaussian Splatting, controlling the generation of accurate 3D humans from text prompts remains an open challenge. Current methods struggle with fine detail, accurate rendering of hands and faces, human realism, and controlability over appearance. The lack of diversity, realism, and annotation in human image data also remains a challenge, hindering the development of a foundational 3D human model. We present a weakly supervised pipeline that attempts to address these challenges. In the first step, we generate a photorealistic human image dataset with controllable attributes such as appearance, race, gender, etc., using a *state-of-the-art* image diffusion model. Next, we propose an efficient mapping approach from image features to 3D point clouds using a transformer-based architecture. Finally, we close the loop by training a point-cloud diffusion model that is conditioned on the same text prompts used to generate the original samples. We demonstrate orders-of-magnitude speed-ups in 3D human generation compared to the *state-of-the-art* approaches, along with significantly improved text-prompt alignment, realism, and rendering quality. The code and data are available at https://github.com/ivashmak/hugediff.git.

1 Introduction

3D human modeling, including reconstruction, rendering, and generation, plays an increasingly vital role in modern computer vision, supporting applications in augmented/virtual reality, human-computer interaction, gaming, and telepresence. Recent advances in the field, alongside increased computational power, have enabled the creation of photorealistic avatars. However, the field still faces many challenges, including model generalization, data diversity and privacy (*e.g.*, person appearance), clothing variability, method performance, etc.

Efficient generative 3D human models can greatly benefit the field by enabling appearance transfer, which helps preserve anonymity and increase diversity. However, such a generative model requires large-scale and diverse datasets, which remain a major challenge. Most real-world human datasets are constrained by privacy concerns, financial limitations, commercial restrictions, small sample sizes, or demographic biases (limited to the population of the country where the dataset was collected). In turn, the lack of data can severely affect the ability of the models to generalize. To mitigate this issue, recent works [73, 86] have begun generating synthetic data or combining existing datasets. However, the realism of such synthetic data is inherently limited by the quality of the generative models. Although early works [57, 78] often produced artifacts and lacked realism, modern diffusion models [33, 51] have made significant progress toward generating photorealistic images. In the literature [53], authors propose collecting samples from public repositories (*e.g.,* YouTube), but the ethical implications of this approach are questionable due to uncertain permissions.

Many existing methods for 3D human reconstruction from a single image rely on SMPL-X UV maps [73, 80, 86]. However, this is suboptimal for several reasons. First, the XYZ-to-UV mapping is typically neither one-to-one nor unique due to discontinuities and distortions that arise when unwrapping a 3D model onto a 2D plane. As a result, this process can cause repetitions in the mapping, where multiple 3D vertices may correspond to the same UV coordinate. Second, it requires prior access to UV maps, which restricts its applicability in real-world scenarios where such mappings are unavailable. In contrast, we propose a universal and straightforward approach to map image features directly to 3D structure without relying on predefined UV coordinates. Compared to related work [53, 81], our method is more general and easier to apply.

Our approach directly addresses key challenges such as data scarcity, generalization, diversity, and reconstruction. We introduce a weakly supervised pipeline that integrates data generation, 3D reconstruction, rendering, and generative modeling. The process starts by producing diverse single-view human images from different camera angles using an *off-the-shelf* diffusion model. We then reconstruct 3D human representations by lifting image features from arbitrary viewpoints and assigning them to a 3D point cloud using an attention-based mechanism. Finally, we complete the human appearance and close the loop by generating 3D human parameters through a guided diffusion model conditioned on textual prompts.

By leveraging diffusion models for initial data generation, we gain explicit control over diversity in human appearance, including variations in clothing, race, gender, etc., which benefits the training of a generative 3D human model. In parallel, we introduce efficient architectures for the 3D reconstruction (uplift) stage and point-cloud noise prediction within the diffusion process. Our contributions are summarized as follows:

1. A diverse, high-resolution, photorealistic, AI-generated dataset with corresponding 3D human reconstructions.
2. An efficient and universal UV-free transformer-based reconstruction method that maps image features to 3D point clouds.
3. A guided point-cloud diffusion sampler conditioned on text prompts for human rendering via Gaussian Splatting.

2 Related Work

Most 3D human models are explicitly represented as meshes (*e.g.*, [19, 37, 43, 53, 60, 73, 81, 86] etc.), which provides fast training and inference. The most widely used mesh representation is SMPL(-X) [40, 47]. Implicit human representations [50, 61, 70], which use continuous functions (*e.g.*, signed distance fields), are less common due to their high computational cost and slower inference. However, they are more flexible, template-free, and capable of capturing fine-grained details.

Defining a suitable human model is crucial for 3D reconstruction from images, as it provides strong priors that simplify the task. Several methods [73, 80, 86] leverage priors from the SMPL-X framework by mapping UV coordinates to 3D mesh vertices. A common approach involves extracting image features using a pretrained model (*e.g.*, Sapiens [27]), projecting them into UV space, and uplifting the UV features to 3D meshes via the predefined mapping. Other approaches, such as LHM [53] and GTA [81], utilize transformer-based architectures. For instance, LHM adopts a multimodal transformer [11] tailored for human heads, while GTA leverages a triplane representation. Point-E [45] of Nichol *et al.* proposes an alternative method which combines CLIP [54] embeddings for integrating text prompts and image features.

Supervision in reconstruction models often involves rendering the 3D output and computing losses against ground-truth images. Earlier works [25, 48] used mesh-based rendering, which is fast but limited by mesh resolution. NeRF-based methods [59, 62, 66] model continuous color and density fields, providing strong generalization and high fidelity. Although they are computationally expensive and slow to train and run, even with faster variants [2, 6, 12, 67, 71]. The recently proposed 3D Gaussian Splatting (3DGS) [26] introduces a discrete set of Gaussian primitives, enabling real-time rendering while achieving quality comparable to or better than NeRF-like approaches.

Training data for 3D human models typically include multi-view images to resolve single-view ambiguities. Some of the largest multi-view human datasets include MVHumanNet [69], THuman [63, 75, 83], Human3.6M [24], and ActorsHQ [23]. However, limited diversity and control in human appearance have led to a growing interest in generating synthetic datasets or augmenting real ones. For example, IDOL [86] of Zhuang *et al.* proposes human image generation, and Champ [85] augments the data to simulate multiple views using generative models conditioned on single-view images and meshes. SIGMAN [73] and LHM [53] combine various public real-world and synthetic datasets to improve generalization.

An obvious solution is to approach 3D human generation from a text prompt as a two-stage process [3, 5, 22, 31, 32, 36, 38, 74], where multi-view images are generated first by Stable Diffusion [57] or ControlNet [78], and then a 3D model is fitted to them. However, since many approaches build on older diffusion models, these methods lack realism and can take too long to generate. Methods such as [1, 18, 73, 80] generate 3D humans in a zero-shot manner, but they lack appearance control. The SMPLitex [4] has both appearance control and instant generation from text, but generates only UV texture maps. Gong *et al.* in [15] and Fu *et al.* in [14] struggle to render detailed hands and faces. Human3Diffusion [72] and PSHuman [35] provide only an image-conditioned 3D human generation framework.

3 Methodology

Our pipeline is illustrated in Fig. 1 and consists of several key stages. First, we generate a dataset of diverse humans and perform an initial reconstruction by fitting the SMPL-X parameters to each image. Next, we train an efficient transformer-based reconstruction model that estimates the 3DGS parameters, anchored to the canonical SMPL-X mesh, using Sapiens features extracted from the image. Finally, we train a text-conditioned guided diffusion

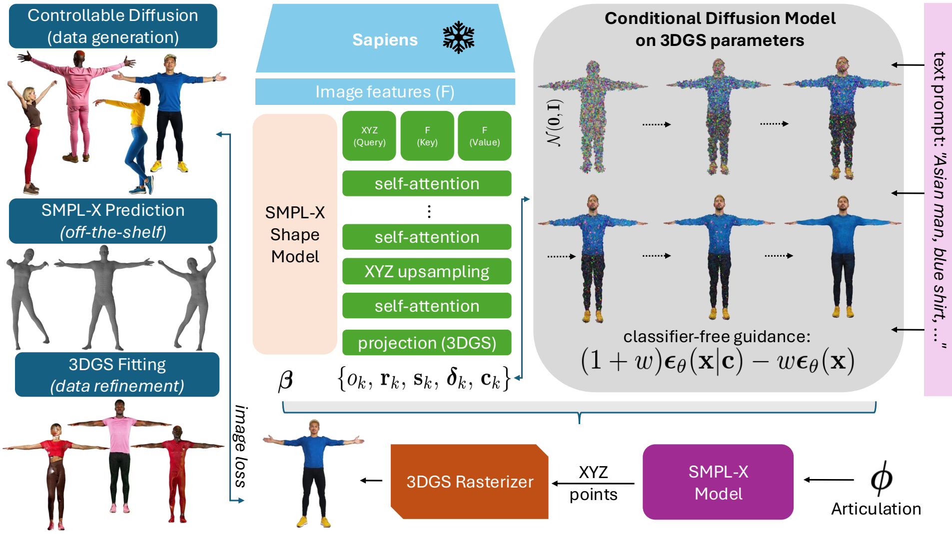

Figure 1: **Overview of the proposed pipeline**. The left side shows the data generation part, the middle part is the 3DGS reconstruction from image features. The right side highlights the text-conditioned diffusion model.

process on 3DGS parameters that learns the human appearance distribution. We describe each component in detail below.

3.1 Dataset generation

We employ the FLUX diffusion model [33] to generate approximately 10,000 diverse human images, conditioned on text prompts describing attributes such as race, gender, hair color, and clothing. Compared to alternatives like SDXL [51] or ControlNet [78], FLUX produced the most photorealistic and prompt-aligned results. Each generated image is segmented using SAM-2 [55] to remove the background, and then processed with Sapiens [27] to extract image features. We estimate initial SMPL-X parameters using ExPose [8], and refine hand poses with HaMeR [49]. Since ExPose reconstructions are often misaligned, we initialize our Gaussian splatting model with the coarse mesh and jointly optimize SMPL-X and 3DGS parameters by minimizing rendering loss with respect to the original image. Fitting 3DGS to a single image is highly under-constrained, so we regularize the process by setting all opacities to one and aligning rotations perpendicular to surface normals. This enables fast and stable optimization.

3.2 Single Image Reconstruction

After obtaining reliable SMPL-X parameters, our next step is to create a generalizable model that reconstructs a 3D human from a single image in terms of 3DGS parameters. For image $\mathbf{I}$, the human reconstruction is represented by SMPL-X pose vector $\boldsymbol{\phi}$, shape vector $\boldsymbol{\beta}$, facial expressions $\boldsymbol{\psi}$, and set of Gaussian primitives $\mathcal{G} = \{o_k, \mathbf{r}_k, \mathbf{s}_k, \boldsymbol{\delta}_k, \mathbf{c}_k \mid k = 1, \ldots, N\}$ of N points anchored to SMPL-X model θ. The 3DGS parameters $(o, \mathbf{r}, \mathbf{s}, \mathbf{c}, \boldsymbol{\delta}, \mathbf{c})$ are attached to canonical point set $\mathbf{X}_0 = \theta(\boldsymbol{\phi}_0, \boldsymbol{\beta}_0, \boldsymbol{\psi}_0) \in \mathbb{R}^{N\times 3}$, and correspond to opacities, rotations, anisotropic scales, colors, and offsets to the canonical pose coordinates ($\mathbf{X}_0$), respectively.

To reduce the complexity of estimating this high-dimensional parameter space from a single image, we make the following assumptions. First, the exact human articulation ($\boldsymbol{\phi}$) and facial expressions ($\boldsymbol{\psi}$) are less critical for our task than accurate appearance modeling. *Off-the-shelf* estimators, such as ExPose or OSX [37] and HaMeR for hands, can already estimate an approximate human articulation from a single image. Second, since all points lie on the mesh surface, we assume, without loss of generality, that they are non-transparent ($o_k = 1$), as transparency would introduce holes in the mesh. Furthermore, placing rotations perpendicular to the mesh normals implies better alignment with surface geometry, spatial consistency, and higher robustness to various articulations. Finally, the formalized problem is the following:

$$\mathcal{H}(\mathbf{I}) \sim (\mathcal{G}, \boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{\psi}) \sim (\{\mathbf{s}_k, \boldsymbol{\delta}_k, \mathbf{c}_k\}, \boldsymbol{\beta}), \tag{1}$$

We leverage a pre-trained foundation Sapiens model for image processing that has a large human prior and good generalization. The Sapiens model returns image features $\mathbf{F}$ as the spatial grid $H' \times W'$ with C channels from the input image $\mathbf{I}$ of $H \times W \times 3$. The next step is to map the $M = (H' \cdot W')$ feature cells to the N mesh points, which is a non-trivial problem that requires assigning 2D image-based information to a 3D space. Previous work has approached this problem using the UV template SMPL-X mapping [73, 86], which determines the correspondence of the 2D UV coordinates to the 3D model using standard graphics techniques. However, taking inspiration from template-free methods [53, 81], we train an efficient transformer-based model to map image features to the point cloud. To preserve high-frequency details and maintain efficiency, we design a lightweight cross-attention module between canonical points and image features. The query is a canonical set of points ($\mathbf{X}_0$) and keys with values are image features ($\mathbf{F}$). We add a linear projection l along with positional encoding to match the same dimensionality (d) of key ($\mathbf{K}$) and query ($\mathbf{Q}$) and reshape the tensors to allow multi-head attention (h – number of heads). The value ($\mathbf{V}$) of cross-attention is image features split among h heads:

$$\mathbf{Q} = l_{\mathrm{PE}}(\mathbf{X}_0) \in \mathbb{R}^{N\times h\times 1\times d}; \;\; \mathbf{K} = l_{\mathrm{PE}}(\mathbf{F}) \in \mathbb{R}^{1\times h\times M\times d}; \;\; \mathbf{V} = \mathit{split}(\mathbf{F}, h) \in \mathbb{R}^{1\times h\times M\times \frac{C}{h}} \tag{2}$$

The attention weights are computed by applying a softmax to the multiplication of the query and the key matrices. Afterwards, the per-point features ($\mathbf{F}_X$) are computed as follows:

$$\mathbf{F}_X = \mathrm{softmax}(\mathbf{Q}\mathbf{K}^\top)\,\mathbf{V} \in [\mathbb{R}^{N\times h\times 1\times d}\,\mathbb{R}^{1\times h\times d\times M}\,\mathbb{R}^{1\times h\times M\times \frac{C}{h}} \to \mathbb{R}^{N\times h\times 1\times \frac{C}{h}} \to \mathbb{R}^{N\times C}], \tag{3}$$

In our derivation, we assume broadcasting of tensors for dimension value "1" (similarly to PyTorch [46]). Additionally, the transpose operator ($^\top$) is applied only to the last two dimensions.

The spatial dimensions of the feature grid (H', W') (last output of the ViT [10] encoder) normally do not exceed 64×64, and the number of feature channels are 2048 in the largest Sapiens model. The main computational bottleneck arises from the number of points N, which significantly increases the processing cost. As a result, the final complexity is $\mathcal{O}\big(NhM \max(d, \frac{C}{h})\big)$. The number of points N is a scalable parameter that influences the quality of the reconstruction, but it must be handled with care to manage the computational demands. To address this, we introduce an upsampling module that enables the mapping of image features to the anchor set of mesh points, and further enhances the output fidelity with a denser point cloud.

Analogously to the assignment of image features to a point cloud, we design a point upsampling mechanism based on cross-attention. Let $\mathbf{X}'_0 \in \mathbb{R}^{n\times 3}$ be the subsampled point set (*i.e.*, $n < N$), and $K \in \{1,\dots,n\}^{N\times k}$ be a matrix of k nearest neighbor indices of $\mathbf{X}_0$ to $\mathbf{X}'_0$. The upsampling transformer takes the full point cloud $\mathbf{X}_0$ as a query, a subsampled point cloud indexed with a neighbors matrix as keys, and associated features as values. The multihead cross-attention is then defined as:

$$\mathbf{F}_{X_0} \leftarrow \mathrm{softmax}\big(l_{\mathrm{PE}}(\tilde{\mathbf{X}}_0) l_{\mathrm{PE}}(\tilde{\mathbf{X}}'_0[K])^\top\big)\,\mathbf{F}_{X'_0}[K]. \tag{4}$$

Connections among neighboring points in graph-structured point clouds are essential for local feature propagation and to add high-frequency detail. The graph-convolutional methods like GCN [29] or GAT [64] improve the local context within neighborhoods. However, these approaches yielded limited improvement in experiments. Consequently, we designed a self-attention module for point clouds, where the keys incorporate both feature embeddings and relative positional distances to neighboring points. Let K' be the neighbors indices of $\mathbf{X}$ to itself ($\mathbf{X}$), then features are updated as follows:

$$\mathbf{F}_X \leftarrow \mathrm{softmax}\left(l_{\mathrm{PE}}(\mathbf{F}_X)\big(l_{\mathrm{PE}}(\mathbf{F}_X[K']) + l_{\mathrm{PE}}(\tilde{\mathbf{X}}_0 - \tilde{\mathbf{X}}_0[K'])\big)^\top\right)\mathbf{F}_X[K']. \tag{5}$$

For $\mathbf{F}_X \in \mathbb{R}^{N\times h\times 1\times d}$, the operator $[\dots]$ means indexing tensor with indices K, *e.g.*, $\mathbf{F}_X[K] \in \mathbb{R}^{N\times h\times k\times d}$. The point set $\tilde{\mathbf{X}}$ indicates expanded embeddings of the point cloud $\mathbf{X}$ to the shape $(N \times h \times 1 \times d)$.

3DGS Regression. The 3DGS parameters are estimated in the final layers of our model. After upsampled point features are refined via the self-attention module, we apply a set of linear layers to regress the colors, scales, and displacements relative to the canonical mesh. Each output is constrained to a pre-defined range to limit and regularize 3DGS, *e.g.,* prevent excessively large scales or displacements. For finer structures, such as the head and hands, we apply stricter constraints due to their smaller spatial extent. To enforce those constraints, we use a sigmoid for RGB colors and sinusoidal normalization for scales and displacements. For displacements and scales, we found that sigmoid can suffer from vanishing gradients, leading

Multi-view Regularization. Estimating a full human appearance from a single-view or partial observation is inherently ambiguous, as many plausible reconstructions exist. Without sufficient regularization and constraints, neural models tend to eventually overfit to the training data. In the context of a 3D human appearance from a single image, we observe that after a few training epochs, the model starts mirroring the front-facing appearance to the back. For example, resulting in artifacts such as “Janus faces”, *i.e.*, a face appearing on both the front and back of the head. To mitigate this, prior works (*e.g.,* [30, 86]) leverage *off-the-shelf* generative models conditioned on the input view to obtain multi-view supervision. In contrast, we eliminate reliance on the external frameworks by implicitly utilizing a human 3D prior. Specifically, we train a separate text-conditioned reconstruction model by only replacing the image features $\mathbf{F}$ with text-prompt encodings. Although this model tends to generate an “average'; human appearance, it is capable of hallucinating high-detail textures. We render its output from four canonical views (front, back, left, right) and use these renderings as multi-view supervision for regularizing the feature-based model.

Shape Model. The SMPL-X shape coefficients $\boldsymbol{\beta}$ are predicted using an MLP that takes both spatial and averaged features. In particular, the feature matrix $\mathbf{F}$ is first averaged along height and width dimensions, and then separately averaged along the channel dimension (C).

Method	**Time** (minutes)	**Method**	**Time** (minutes)
HumanGaussian [38]	68.79	DreamWaltz-G [21, 22]	201.95
TADA [36]	191.74	DreamAvatar [3]	350.50
GaussianDreamer [74]	11.66	SMPLitex [4]	0.29
EVA3d* [18]	0.16	StructLDM* [20]	0.10
GSM* [1]	0.12	Ours	0.31

Table 1: Time comparison for one sample generation. The methods denoted with * do not have control over appearance.

The resulting features are concatenated and flattened into a single vector. This provides the model with both spatially-aware and globally-aggregated information, enabling it to estimate an approximate human body shape. For text-guided models, the shape is predicted directly from the text encoding.

3.3 Gaussian Splatting Diffusion

The final stage of our pipeline generates 3D human parameters from the same text prompts used to guide the image diffusion model. The diffusion model introduces greater stochasticity and variety into appearance generation than its deterministic text-conditioned reconstruction counterpart. We adopt diffusion models for this task, as they have demonstrated *state-of-the-art* performance in generative modeling [9, 17], outperforming earlier approaches such as VAEs [28] and GANs [16].

Most approaches employ latent diffusion models [56] to accelerate training and inference. For instance, methods [7, 34, 77, 84] encode point clouds via point-based models such as PointNet(++) [5, 52], PointTransformer [82], DGCNN [65], PointConv [68], etc., and the noise predictor is typically a U-Net [58]. The trade-off for denoising the latent space is a potential quality reduction in generation, because of error accumulation within the encoder-decoder and diffusion process. To ensure the best quality generation, we apply our diffusion model to the 3DGS parameters, *i.e.*, scales, displacements, and colors, estimated from the previous steps of the pipeline. The U-Net 1D works on sequence data, which requires reordering of the point cloud. With a human point cloud derived from the SMPL-X mesh, there is no reordering to sort points consecutively. Therefore, we make the following changes. Firstly, we replace "Conv1D" with a lightweight MLP that learns weights to combine the $k-$ nearest neighbor features of a point. Secondly, we replace "Conv1D" downsampling and "Upsample" layers with multi-head attention on the nearest neighbors, the same as in the reconstruction model.

4 Experiments

Implementation Details. The original SMPL-X mesh contains around 10,475 mesh vertices. However, such a low resolution leads to poor rendering and a lack of detail when used for a Gaussian splatting framework. Consequently, we densified the mesh to 84,317 points by recursively splitting edges and faces if the length or area exceeds pre-defined thresholds. This allowed the 3DGS fitting to reach almost 30 PSNR on reconstruction in less than a minute on NVIDIA GeForce RTX 3090.

We trained our models with the AdamW [41] optimizer. For the training of the 3D

Method	Le Chat Aln./Aes.	Gemini Aln./Aes.	Grok Aln./Aes.	Claude Aln./Aes.	Copilot Aln./Aes.
TADA [36]	77.7/68.3	62.7/46.7	49.0/38.7	82.7/76.1	81.9/75.9
HumanGaussian [38]	85.3/75.7	76.0/62.3	76.3/67.0	83.7/75.8	86.0/72.7
DreamAvatar [3]	73.3/70.0	61.0/38.7	59.3/38.7	83.0/81.1	85.0/68.7
GaussianDreamer [74]	82.7/68.7	49.7/43.7	35.3/42.3	65.0/79.3	84.3/77.2
DreamWaltz-G [21, 22]	86.3/78.7	58.7/63.0	60.3/70.0	62.0/78.0	72.3/76.3
SMPLitex [4]	58.6/72.0	32.7/11.3	21.7/32.0	47.3/71.7	66.7/59.4
Ours	**89.3/91.0**	**83.3/74.7**	**93.0/85.5**	**95.0/85.0**	**88.3/82.6**

Table 2: Comparison evaluation of rendered images in terms of text-prompt alignment ("Aln.") and image aesthetics ("Aes.") using large language models.

Method	Google Gemini preference ranking (%)		
	1st	2nd	3rd
TADA [36]	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
HumanGaussian [38]	0.00 ± 0.00	2.22 ± 3.14	93.33 ± 5.44
DreamAvatar [3]	0.00 ± 0.00	4.44 ± 3.14	0.00 ± 0.00
GaussianDreamer [74]	0.00 ± 0.00	2.22 ± 3.14	2.22 ± 3.14
DreamWaltz-G [21, 22]	11.11 ± 15.71	80.00 ± 18.86	4.44 ± 3.14
SMPLitex [4]	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Ours	$\mathbf{88.89 \pm 15.71}$	11.11 ± 15.71	0.00 ± 0.00

Table 3: Ranked preferences of rendered human appearances as judged by Google Gemini. Results are reported as mean ± standard deviation over three runs.

reconstruction model, we utilize LPIPS [79] (0.15), L1 (1.0), and SSIM (0.25) losses. For the diffusion model, we used the "*pred_x0*" objective, number of timestamps 1000, and the L2 loss.

The time comparison is shown in Table 1. Our method outperforms the text-conditioned *state-of-the-art* models (sometimes by orders of magnitude) in terms of time to generate and render a human. Text-conditioned methods such as HumanGaussian [38], DreamWaltz-G [22] or DreamAvatar [3] first generate training images, then fit a 3DGS to them, and then render (*e.g.*, Instant-NGP [44] or NeRF [42]). With the exception of SMPLitex [4], no other text-guided model in the competitors list renders a 3D human directly. The methods that unconditionally generate humans have a comparative speed, but they cannot align the generation with the text prompt.

Evaluation. We quantitatively evaluate 3D human generation in Table 2, focusing on prompt alignment and image aesthetics (including realism). CLIP-based metrics [54] are commonly used to measure consistency between text prompts and generated images [38]. However, we found CLIP scores to be unreliable due to factors such as resolution differences and variations in human articulation. Large language models (LLMs), in contrast, do not necessarily rely on CLIP, and many can directly evaluate images and text.

We rendered 15 images per method; however, the high computational cost of some approaches makes large-scale quantitative evaluation impractical. Our method achieved the best results in image aesthetics and text prompt alignment, outperforming *state-of-the-art* methods by a high margin, as evaluated by all five language models. In our evaluation, we excluded some relevant works (SIGMAN [73], Text2Avatar [15], DreamHuman [31], and [32]) as the code is not publicly available.

Figure 2: Comparison of rendered images generated with text prompt conditioning.

From our previous evaluation, we found Google Gemini to be the most reliable LLM for judgment. Consequently, we used Gemini to rank the rendered results based on aesthetic quality. The outcomes are reported in Table 3, where our method was ranked first in nearly 89% of cases and second in the remainder.

Fig. 2 shows a qualitative comparison with text-guided *state-of-the-art* methods. None of the competitors match the prompt we provided. The reason for this could be that most of the methods use Stable Diffusion (SD) [57] or ControlNet [78] internally to generate training images, and they struggle with prompt guidance. Furthermore, the generated appearances look rather cartoonish, human proportions are strange, and the hand quality is poor.

Fig. 3 illustrates a visual comparison to the *state-of-the-art* models that generate without control of human appearance. The reason for this is that these methods were trained primarily on real-world fashion datasets (*e.g.*, SHHQ [13], DeepFashion [39], UBCFashion [76]) that may have too diverse clothing to annotate with text. Furthermore, those datasets could contain a bias towards female participants, because methods such as EVA3D [18] and GSM [1] tend to generate women. In Fig. 3, we highlight the rendering artifacts that occurred on hands, feet, and faces, where the *state-of-the-art* methods struggle the most. Our approach has fewer rendering artifacts, and the rendering quality looks higher. Novel view rendering is visualized in Fig. 4, where the human appearance looks consistent across different viewpoints.

Figure 3: Comparison of rendered images generated unconditionally.

Figure 4: Examples of novel view rendering of a 3D human appearance generated via the proposed diffusion model.

5 Conclusions

We presented a weakly supervised pipeline, where we first generate a photorealistic image with an *off-the-shelf* diffusion model and extract initial reconstructions through 3D model fitting. Next, we train the reconstruction model with efficient image features to learn the mapping to 3D point cloud features. Our training pipeline is self-contained, requiring neither external tools for multi-view augmentation nor predefined UV coordinates. Only image features containing rich human priors are extracted by Sapiens to boost the generalization of our model. The reconstructed 3DGS parameters from a single reconstruction are then used in training a point-cloud text-conditioned diffusion model with classifier-free guidance.

With extensive qualitative and performance evaluation, we show that our method outperforms the current *state-of-the-art* approaches in 3D human generation in terms of rendered quality, speed, and appearance control. Our images present high-fidelity textures, realistic proportions, with improved hand and face quality.

6 Acknowledgements

This work was supported by the SNSF project 'SMILE II' (CRSII5 193686), the Innosuisse IICT Flagship (PFFS-21 47), EPSRC grant APP24554 (SignGPT-EPZ535370/1) and through funding from Google.org via the AI for Global Goals scheme. This work reflects only the author's views and the funders are not responsible for any use that may be made of the information it contains.

References

[1] Rameen Abdal, Wang Yifan, Zifan Shi, Yinghao Xu, Ryan Po, Zhengfei Kuang, Qifeng Chen, Dit-Yan Yeung, and Gordon Wetzstein. Gaussian Shell Maps for Efficient 3D Human Generation . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9441–9451, Los Alamitos, CA, USA, June 2024. IEEE Computer Society. doi: 10.1109/CVPR52733.2024.00902. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52733.2024.00902.

[2] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. *ICCV*, 2023.

[3] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 958–968, 2024.

[4] Dan Casas and Marc Comino-Trinidad. SMPLitex: A Generative Model and Dataset for 3D Human Texture Estimation from Single Image. In *British Machine Vision Conference (BMVC)*, 2023.

[5] R. Qi Charles, Hao Su, Mo Kaichun, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 77–85, 2017. doi: 10.1109/CVPR.2017.16.

[6] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV)*, 2022.

[7] Zhikai Chen, Fuchen Long, Zhaofan Qiu, Ting Yao, Wengang Zhou, Jiebo Luo, and Tao Mei. Learning 3d shape latent for point cloud completion. *IEEE Transactions on Multimedia*, 26:8717–8729, 2024. doi: 10.1109/TMM.2024.3381814.

[8] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision (ECCV)*, 2020. URL https://expose.is.tue.mpg.de.

[9] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[12] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022.

[13] 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*, page 1–19, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-19786-4. doi: 10.1007/978-3-031-19787-1_1. URL https://doi.org/10.1007/978-3-031-19787-1_1.

[14] Tsu-Jui Fu, Wenhan Xiong, Yixin Nie, Jingyu Liu, Barlas Ouguz, and William Yang Wang. Text-guided 3d human generation from 2d collections. In *Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://api.semanticscholar.org/CorpusID:258841620.

[15] Chaoqun Gong, Yuqin Dai, Ronghui Li, Achun Bao, Jun Li, Jian Yang, Yachao Zhang, and Xiu Li. Text2avatar: text to 3d human avatar generation with codebook-driven body controllable attribute. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16–20. IEEE, 2024.

[16] 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*, pages 2672–2680, 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

[18] Fangzhou Hong, Zhaoxi Chen, Yushi LAN, Liang Pan, and Ziwei Liu. EVA3d: Compositional 3d human generation from 2d image collections. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=g7U9jD_2CUr.

[19] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. GaussianAvatar: Towards realistic human avatar modeling from a single video via animatable 3D gaussians. *arXiv preprint arXiv:2312.02134*, 2023.

[20] Tao Hu, Fangzhou Hong, and Ziwei Liu. Structldm: Structured latent diffusion for 3d human generation. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LI*, page 363–381, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72982-9. doi: 10.1007/978-3-031-72983-6_21. URL https://doi.org/10.1007/978-3-031-72983-6_21.

[21] Yukun Huang, Jianan Wang, Ailing Zeng, He Cao, Xianbiao Qi, Yukai Shi, Zheng-Jun Zha, and Lei Zhang. DreamWaltz: Make a Scene with Complex 3D Animatable Avatars. In *Advances in Neural Information Processing Systems*, pages 4566–4584, 2023.

[22] Yukun Huang, Jianan Wang, Ailing Zeng, Zheng-Jun Zha, Lei Zhang, and Xihui Liu. DreamWaltz-G: Expressive 3D Gaussian Avatars from Skeleton-Guided 2D Diffusion. 2024.

[23] Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM Transactions on Graphics (TOG)*, 42(4):1–12, 2023. doi: 10.1145/3592415. URL https://doi.org/10.1145/3592415.

[24] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. doi: 10.1109/TPAMI.2013.248.

[25] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[26] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

[27] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. *arXiv preprint arXiv:2408.12569*, 2024.

[28] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[29] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

[30] Tobias Kirschstein, Javier Romero, Artem Sevastopolsky, Matthias Nießner, and Shunsuke Saito. Avat3r: Large animatable gaussian reconstruction model for high-fidelity 3d head avatars, 2025. URL https://arxiv.org/abs/2502.20220.

[31] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: animatable 3d avatars from text. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[32] Nikos Kolotouros, Thiemo Alldieck, Enric Corona, Eduard Gabriel Bazavan, and Cristian Sminchisescu. Instant 3d human avatar generation using image diffusion models. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXXVII*, page 177–195, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-73020-7. doi: 10.1007/978-3-031-73021-4_11. URL https://doi.org/10.1007/978-3-031-73021-4_11.

[33] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

[34] Yushi Lan, Shangchen Zhou, Zhaoyang Lyu, Fangzhou Hong, Shuai Yang, Bo Dai, Xingang Pan, and Chen Change Loy. Gaussiananything: Interactive point cloud latent diffusion for 3d generation. In *ICLR*, 2025.

[35] Peng Li, Wangguandong Zheng, Yuan Liu, Tao Yu, Yangguang Li, Xingqun Qi, Mengfei Li, Xiaowei Chi, Siyu Xia, Wei Xue, et al. Pshuman: Photorealistic single-view human reconstruction using cross-scale diffusion. *arXiv preprint arXiv:2409.10141*, 2024.

[36] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J. Black. TADA! Text to Animatable Digital Avatars. In *International Conference on 3D Vision (3DV)*, 2024.

[37] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21159–21168, 2023.

[38] Xian Liu, Xiaohang Zhan, Jiaxiang Tang, Ying Shan, Gang Zeng, Dahua Lin, Xihui Liu, and Ziwei Liu. Humangaussian: Text-driven 3d human generation with gaussian splatting. *arXiv preprint arXiv:2311.17061*, 2023.

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

[40] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

[41] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. URL https://api.semanticscholar.org/CorpusID:53592270.

[42] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[43] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3D gaussian avatar. In *ECCV*, 2024.

[44] 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. doi: 10.1145/3528223.3530127. URL https://doi.org/10.1145/3528223.3530127.

[45] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts, 2022. URL https://arxiv.org/abs/2212.08751.

[46] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703, 2019. URL http://arxiv.org/abs/1912.01703.

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

[48] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[49] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024.

[50] Marco Pesavento, Yuanlu Xu, Nikolaos Sarafianos, Robert Maier, Ziteng Wang, Chun-Han Yao, Marco Volino, Edmond Boyer, Adrian Hilton, and Tony Tung. Anim: Accurate neural implicit model for human reconstruction from a single rgb-d image. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5448–5458, 2024. URL https://api.semanticscholar.org/CorpusID:268510007.

[51] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. URL https://arxiv.org/abs/2307.01952.

[52] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: deep hierarchical feature learning on point sets in a metric space. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5105–5114, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[53] Lingteng Qiu, Xiaodong Gu, Peihao Li, Qi Zuo, Weichao Shen, Junfei Zhang, Kejie Qiu, Weihao Yuan, Guanying Chen, Zilong Dong, and Liefeng Bo. Lhm: Large animatable human reconstruction model from a single image in seconds. 2025.

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231591445.

[55] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

[56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CoRR*, abs/2112.10752, 2021. URL https://arxiv.org/abs/2112.10752.

[57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, Los Alamitos, CA, USA, June 2022. IEEE Computer Society. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01042.

[58] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. URL http://arxiv.org/abs/1505.04597.

[59] Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[60] Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Julien Valentin, Jie Song, and Otmar Hilliges. X-Avatar: Expressive human avatars. In *CVPR*, 2023.

[61] Dae-Young Song, HeeKyung Lee, Jeongil Seo, and Donghyeon Cho. Difu: Depth-guided implicit function for clothed human reconstruction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8738–8747, 2023. doi: 10.1109/CVPR52729.2023.00844.

[62] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: articulated neural radiance fields for learning human shape, appearance, and pose. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713845393.

[63] Zhaoqi Su, Tao Yu, Yangang Wang, and Yebin Liu. Deepcloth: Neural garment representation for shape and style editing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1581–1593, 2023. doi: 10.1109/TPAMI.2022.3168569.

[64] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. *6th International Conference on Learning Representations*, 2017.

[65] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 2019.

[66] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *CVPR*, 2022.

[67] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, June 2022.

[68] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 9621–9630, 2019.

[69] Zhangyang Xiong, Chenghong Li, Kenkun Liu, Hongjie Liao, Jianqiao Hu, Junyi Zhu, Shuliang Ning, Lingteng Qiu, Chongjie Wang, Shijie Wang, et al. Mvhumannet: A large-scale dataset of multi-view daily dressing human captures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19801–19811, 2024.

[70] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13296–13306, June 2022.

[71] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: neural radiance fields for rendering and temporal reconstruction of humans in motion. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713845393.

[72] Yuxuan Xue, Xianghui Xie, Riccardo Marin, and Gerard. Pons-Moll. Human 3Diffusion: Realistic Avatar Creation via Explicit 3D Consistent Diffusion Models. 2024.

[73] Yuhang Yang, Fengqi Liu, Yixing Lu, Qin Zhao, Pingyu Wu, Wei Zhai, Ran Yi, Yang Cao, Lizhuang Ma, Zheng-Jun Zha, and Junting Dong. Sigman:scaling 3d human gaussian generation with millions of assets, 2025. URL https://arxiv.org/abs/2504.06982.

[74] Taoran Yi, Jiemin Fang, Junjie Wang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In *CVPR*, 2024.

[75] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*, June 2021.

[76] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. Dwnet: Dense warp-based network for pose-guided human video generation. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 51. BMVA Press, 2019. URL https://bmvc2019.org/wp-content/uploads/papers/1039-paper.pdf.

[77] xiaohui zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 10021–10039. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/40e56dabe12095a5fc44a6e4c3835948-Paper-Conference.pdf.

[78] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[79] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 586–595. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00068. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

[80] Weitian Zhang, Yichao Yan, Yunhui Liu, Xingdong Sheng, and Xiaokang Yang. E3gen: Efficient, expressive and editable avatars generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 6860–6869, 2024.

[81] Zechuan Zhang, Li Sun, Zongxin Yang, Ling Chen, and Yi Yang. Global-correlated 3d-decoupling transformer for clothed avatar reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[82] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021.

[83] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[84] Junsheng Zhou, Weiqi Zhang, and Yu-Shen Liu. Diffgs: Functional gaussian splatting diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[85] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision (ECCV)*, 2024.

[86] Yiyu Zhuang, Jiaxi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. Idol: Instant photorealistic 3d human creation from a single image, 2024. URL https://arxiv.org/abs/2412.14963.