StructLDM: Structured Latent Diffusion for 3D Human Generation

Tao Hu, Fangzhou Hong, and Ziwei Liu

S-Lab, Nanyang Technological University, Singapore

Fig. 1: StructLDM generates diverse view-consistent humans, and supports different levels of controllable generations and editings, such as compositional generations by blending the five selected parts from a), and part-aware editings such as identity swapping, local clothing editing, 3D virtual try-on, etc. Note that the generations and editing are **clothing-agnostic** without clothing types or masks.

Abstract. Recent 3D human generative models have achieved remarkable progress by learning 3D-aware GANs from 2D images. However, existing 3D human generative methods model humans in a compact 1D latent space, ignoring the articulated structure and semantics of human body topology. In this paper, we explore more expressive and higher-dimensional latent space for 3D human modeling and propose **StructLDM**, a diffusion-based unconditional 3D human generative model, which is learned from 2D images. **StructLDM** solves the challenges imposed due to the high-dimensional growth of latent space with three key designs: **1) A semantic structured latent space** defined on the dense surface manifold of a statistical human body template. **2) A structured 3D-aware auto-decoder** that factorizes the global latent space into several semantic body parts parameterized by a set of conditional structured local NeRFs anchored to the body template, which embeds the properties learned from the 2D training data and can be decoded to render view-consistent humans under different poses and clothing styles. **3) A structured latent diffusion model** for generative human appearance sampling. Extensive experiments validate StructLDM's state-of-the-art generation performance and illustrate the expressiveness of the structured latent space over the well-adopted 1D latent space. Notably,

StructLDM enables different levels of controllable 3D human generation and editing, including pose/view/shape control, and high-level tasks including compositional generations, part-aware clothing editing, 3D virtual try-on, etc. Project page: taohuumd.github.io/projects/StructLDM.

Keywords: 3D Human Generation · Latent Diffusion Model

1 Introduction

Generating and editing high-quality 3D digital humans have been long-studied topics. It empowers many downstream applications, *e.g.*, virtual try-on, and telepresence. Existing works use 3D-aware GAN to learn 3D human generation from 2D images [12, 21, 50, 71], which suffer from low-fidelity generation. In this paper, we contribute a new paradigm of 3D human generation by proposing **StructLDM**, a diffusion-based 3D human generation model with structured human representation that learns from 2D multi-view images or monocular videos.

The problems of 3D-aware human GANs are arguably two parts. Firstly, existing works overlook the semantics and structure of the human body. They sample humans in a compact 1D space, which severely limits their controlling ability. Instead, we propose to explore higher-dimensional semantic latent space for human representation, which allows better capturing of fine details of 3D humans and easy local editing. Secondly, although the 3D-aware GAN has been a success in generating 3D faces and single-class instances [7, 53], its application in 3D human generation has not achieved comparable generation quality. It shows the complexity of 3D human modeling over other subjects and indicates the need for a more powerful model to advance the field. Following the recent success in diffusion models [20, 57], we bring the power of diffusion model to 3D human generation.

Combining the powerful diffusion model and the structured latent representation, we achieve diverse and high-quality 3D human generation, as shown in Fig. 1 a). We have also demonstrated novel editing results without using clothing masks, such as 3D compositional generation, clothing editing, and 3D virtual in Fig. 1 b) and c).

However, the extension over 1D latent space is non-trivial, which imposes more challenges on the generative models due to the high-dimensional growth of the global latent space. To tackle these challenges, in contrast to the well-adopted global 1D latent learning, we propose to model the human latent space locally by proposing three key designs: **1)** a structured and dense human representation; **2)** an auto-decoder architecture to embed latent features; **3)** a latent diffusion model with structure-specific normalization.

Firstly, to fully explore the rich semantics and articulations of the human body, we propose a structured and dense human representation semantically corresponding to human body mesh [42]. It preserves the articulated nature of the human body, and enables detailed appearance capture and editing, as shown in Fig. 2. In contrast to [3, 7, 10, 16, 51, 63] that rely on an implicit mapping

network (*e.g.*, StyleGAN [33]) to map 1D embedding to latent space while sample humans in 1D space, we explicitly model the structured latent space with explainable differentiable rendering **without relying on mapping networks**, which faithfully preserves the fidelity and semantic structures of the latent or embedding space.

Secondly, we design a structured 3D-aware auto-decoder that embeds the structured latent from the 2D training dataset to a shared latent space. We propose to divide the human body into several parts for rendering. At the core of the auto-decoder is a set of structured NeRFs that are locally conditioned on the structured latent space to render a specific body part. Both reconstruction and adversarial supervision are employed to encourage high-fidelity and high-quality image synthesis with high robustness to estimation errors of the training dataset (*e.g.*, human pose or camera estimation errors).

Thirdly, with the structured latents prepared, we learn a latent diffusion model to sample in that space. Since the latents are structured and semantically aligned, we further tailor the diffusion process by using a structure-aligned normalization, which helps to better capture the distribution of our data. Together with the structured latent, the diffusion model enables different levels of controllable 3D human generation and editing, as shown in Fig. 1.

Quantitative and qualitative experiments are performed on three datasets with different setup: monocular videos of UBCFashion [68], multi-view images of RenderPeople [64] and THUman2.0 [67]. They illustrate the versatility and scalability of StructLDM. To conclude, our contributions are listed as follows.
1) Different from the widely adopted 1D latent, we explore the higher-dimensional latent space **without latent mapping** for 3D human generation and editing.
2) We propose StructLDM, a diffusion-based 3D human generative model, which achieves state-of-the-art results in 3D human generation.
3) Emerging from our design choices, we show novel controllable generation and editing tasks, *e.g.*, 3D compositional generations, part-aware 3D editing, 3D virtual try-on.

2 Related Work

2D Human Generation. Generative adversarial networks (GAN) [15] have been a great success in human faces generation [32, 34, 35]. However, due to the complexity of human poses and appearances, it is still challenging for GANs to generate realistic human images [24, 28, 38, 58, 59]. Scaling up the dataset has been proven effective in improving human generation quality [13, 14]. Recent advancements in diffusion models [20, 48, 57] have inspired its application in human image generation [39, 72]. [31, 65] both take human image and pose as input and utilize a pre-trained diffusion model for conditional generation, whereas we train a 3D diffusion model from scratch for unconditional generation.
3D Human Generation. Using real-scanned 3D human dataset, gDNA learns to generate detailed geometry [9]. With the success of 3D-aware GANs [7, 8, 49], the second line of work proposes to learn 3D human generation from 2D human

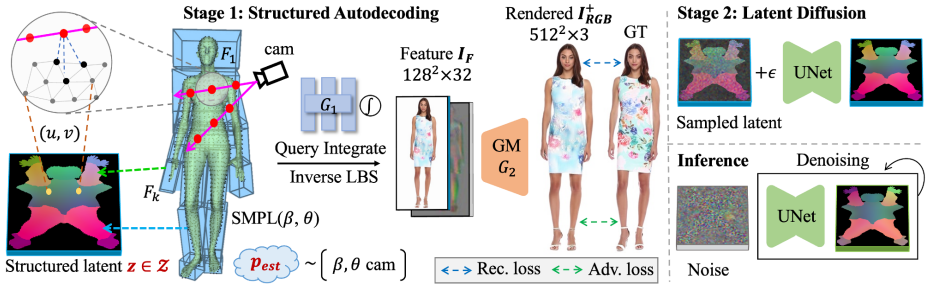

Fig. 2: Two-stage framework. In Stage 1, given a training dataset containing various human subject images with estimated SMPL and camera parameters distribution p_{est}, an auto-decoder is learned to optimize the structured latent $z \in \mathcal{Z}$ for each training subject. Each latent is rendered into a pose- and view-dependent image by a structured volumetric renderer G_1 and a global style mixer module (GM) G_2. In Stage 2, the auto-decoder parameters are frozen and the learned structured latent $\mathcal{Z}$ are then used to train the latent diffusion model. At inference time, latents are randomly sampled and decoded by $G_2 \circ G_1$ for human rendering.

image collections. ENARF-GAN [50] is the first to combine human neural radiance fields with 3D-aware adversarial training. Super-resolution modules can be further used to increase the generation resolution [3,71]. EVA3D [21] proposes to use compositional human representation to increase the raw resolution of neural rendering. AG3D [12] further improves the result with face and normal discriminators. [1] utilizes Gaussian Splatting [36,76] for fast rendering. PrimDiff [11] parameterizes humans with volumetric primitives [41], and learns a diffusion model for unconditional generation. Recent advances in text-image joint distribution learning [55,57] have enabled text-driven 3D generation [27,54]. Combined with 3D human representations, open-vocabulary text-driven 3D human generation can be achieved [4,5,22].

Diffusion Models for 3D Generation. Diffusion models have shown great ability in modeling complex distributions. Many 3D diffusion models have been explored in recent years based on different 3D representations, such as explicit 3D representations of point clouds [43,47,70] and voxels [46,75], implicit functions [29], triplanes [7,17,18,61,66], volumetric primitives [11,41].

High-Dimensional Structured Representation. Existing methods [3,7,10, 16,50,51,63] all rely on an implicit mapping network (*e.g.*, StyleGAN [33]) to map 1D embedding to a triplane or UV latent in a high-dimensional space, while they still sample 1D noises in 1D space for object or human generations. Instead, we explicitly model the high-dimensional structured latent space without relying on mapping networks, which faithfully preserves the fidelity and semantic structures of the latent or embedding space. We illustrate that though the latent mapping enables a smoother latent space for diffusion learning, it degrades the reconstruction fidelity in the supp. mat.

3 Our Approach

StructLDM is a two-stage approach. In the first stage, we learn an auto-decoder containing a set of structured embeddings $\mathcal{Z}$ corresponding to the human subjects in the training dataset, and both the auto-decoder and embeddings are optimized to render pose- and view-conditioned human images with 2D supervision from the training images. The embeddings $\mathcal{Z}$ are then employed to train a latent diffusion model operating in the compact structured latent space in the second stage, which enables diverse and realistic human generations. The full pipeline is depicted in Fig. 2.

We first present the structured latent representation (Sec. 3.1), and then describe the auto-decoding architecture and training procedure (Sec. 3.2, 3.3). Finally, we provide details for the training and sampling of diffusion in the structured latent space (Sec. 3.4).

3.1 Structured 3D Human Representation

Most existing 2D/3D human generative approaches [2,12,14,21,51] model human appearances (*e.g.*, clothing style) in a compact 1D latent space, ignoring the articulated structure and the semantics of the human body. Instead, we explore the articulated structure of the human body and propose to model humans on the dense surface manifold of a parametric human body. The 3D human is recorded on a 2D latent map $\mathbb{R}^{U\times V\times C}$ in the UV space of SMPL mesh [42]. The 2D latent preserves the rich semantics and structures of human body. It is proven to capture the fine details of human appearance better than the existing 1D latent (Sec. 4.3). Note that different from [6, 12, 16] that rely on an implicit mapping network of StyleGAN [33], we explicitly model the structured latent space with explainable differentiable rendering, which faithfully preserves the fidelity and semantic structures of the sampling space. Besides, ours is distinguished from StylePeople [16] by learning high-quality 3D view-consistent generations.

In addition, instead of using the discontinuous SMPL UV mapping [23,25,26, 42,44,45,56] where each body patch is placed discretely, we employ a continuous boundary-free UV mapping that maintains most of the neighboring relations on the original mesh surface and preserves the clothing structures globally with a boundary-free arrangement. The UV mapping has also been proven to be more friendly for CNN-based architecture in [69] than the standard discontinuous UV mapping. Besides, in contrast to existing work employing the UV-aligned feature encoding for subject-specific reconstructions [25, 26, 44, 45, 56], ours is distinguished by encoding various subjects for 3D generation tasks.

3.2 Structured Auto-decoder

Structured Local NeRFs. The extension from 1D to 2D latent introduces more challenges for generative models due to the square growth of the global latent space. Therefore, we propose to divide the latent space into several body parts for local human modeling, and further present a structured auto-decoder

$\mathbb{F}_{\Phi}$ that consists of a set of structured NeRFs $\{F_1, ..., F_k, ...F_K\}$ locally conditioned on a corresponding body patch in the structured latent space as shown in Fig. 2. Specifically, each body part is parameterized by a local NeRF F_k, which model body in a local volume box $\{b^k_{min}, b^k_{max}\}$.

To render the observation space with the estimated posed SMPL mesh $M(\beta, \theta)$ and camera parameters of a specific identity, we query the corresponding identity latent $z \in \mathcal{Z}$ for all sampled points along camera rays and integrate it into a 2D feature map. To be more specific, given a 3D query point x_i in the posed space, we first transform it to a canonical space using inverse linear blend skinning (LBS) which yields $\hat{x}_i$. Then in the canonical space, we find the nearest face f_i of the SMPL mesh for the query point $\hat{x}_i$ and (u_i, v_i) are the barycentric coordinates of the nearest point on f_i. We then obtain the local UV-aligned feature $z_i = B_{u_i,v_i}(z)$ of the query point x_i, where B is the barycentric interpolation function and z is the 2D structured latent. Given camera direction d_i, the appearance features c_i and density σ_i of point x_i are predicted by

$$\{c_i, \sigma_i\} = \frac{1}{\sum \omega_a} \sum_{a \in \mathbb{A}_i} \omega_a F_k(\hat{x}^k_i, d_i, z_i), \tag{1}$$

where $\mathbb{A}_i$ indicates the NeRF sets where point x_i falls in, $\hat{x}^k_i$ is the local coordinate of x_i in NeRF F_k, and $\omega_a = \exp(-m(\hat{x}^k_i(x)^n + \hat{x}^k_i(y)^n + \hat{x}^k_i(z)^n))$ is blending weight when x_i falls in multiple volume boxes. m and n are chosen empirically.

Note that each local NeRF F_k is conditioned on a local UV-aligned feature z_i by $F_k(\hat{x}^k_i, d_i, z_i)$. We integrate all the radiance features of sampled points into a 2D feature map $I_F \in \mathbb{R}^{H \times W \times \bar{C}}$ through volume renderer G_1 [30]:

$$I_F = G_1(z, \beta, \theta, cam; \mathbb{F}_{\Phi}). \tag{2}$$

The structured representation inherits human priors and renders human images in a canonical space to disentangle pose and appearance learning, and enables adaptive allocation of network parameters for efficient training and rendering. In contrast to [74] that parameterizes the human body with hundreds of vertex-level NeRFs, we adopt the part-level structure of EVA3D [21] with 16 local NeRFs for smooth clothing style mixing. In contrast to EVA3D, we use a lightweight auto-decoder architecture, conditioned on the structured 2D latent.

Efficient Geometry-Aware Global Style Mixer. The compositional NeRFs model each body part separately in a canonical space, while they struggle to learn the full-body appearance style as a whole, especially for dress. In addition, the compositional volume boxes with fixed sizes are not effective in reconstructing loose clothing, *i.e.*, predicting hemline or between-leg offsets of dress, as observed in [12,21]. To solve these issues, we propose a global style mixer G_2 that learns to mix the compositional features of each body part to learn full-body appearance style in the compact image space.

$$I^+_{RGB} = G_2 \circ G_1(z, \beta, \theta, cam; \mathbb{F}_{\Phi}). \tag{3}$$

The style mixer is built upon Transposed CNN [60]. G_2 mixes neighbor 4 pixels in the feature map I_F by a receptive field of 4, and upsamples the feature

Fig. 3: Qualitative results on UBCFashion. We generate diverse view-consistent humans under different poses/views for different clothing styles (*e.g.* dress) and hairstyles.

map with 4× super-resolution in image space, which enables efficient geometry-aware rendering as used in [6]. Refer to more details in the supp. mat.

3.3 Joint Learning of Auto-decoder

StructLDM is trained to optimize renderers G_1, G_2 and structured embeddings $\mathcal{Z}$. We employ 1) Reconstruction Loss including Pixel Loss, Perceptual Loss, Face Identity Loss, and Volume Rendering Loss, 2) Adversarial Loss, and 3) Regularization Loss for the learning of geometry and embeddings $\mathcal{Z}$ in training, with Adam [37] as the optimizer. Refer to the supp. mat. for more details.

3.4 Structured Latent Diffusion Model

After embedding training subjects in a structure 2D latent space, we trained a latent diffusion model [57] to learn to sample in this space. Diffusion models [20] are probabilistic models that learn a data distribution $z \sim p_z$ by gradually removing noises from random Gaussian noises. Specifically, the noise-removing process corresponds to a reverse Markov Chain of length T. For each step t, the model, parameterized by U-Net ϵ_θ, learns to predict the noise ϵ from the input noised sample z_t. The training objective can be formulated as

$$L_D = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),t\sim\mathcal{U}(1,T)}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\right]. \tag{4}$$

The structured 2D latents are spatially aligned with the UV space, and we further improve the diffusion model learning by using structure-aligned normalization. Specifically, we calculate the mean and variance of each UV pixel across the whole dataset and normalize latent z for each pixel independently.

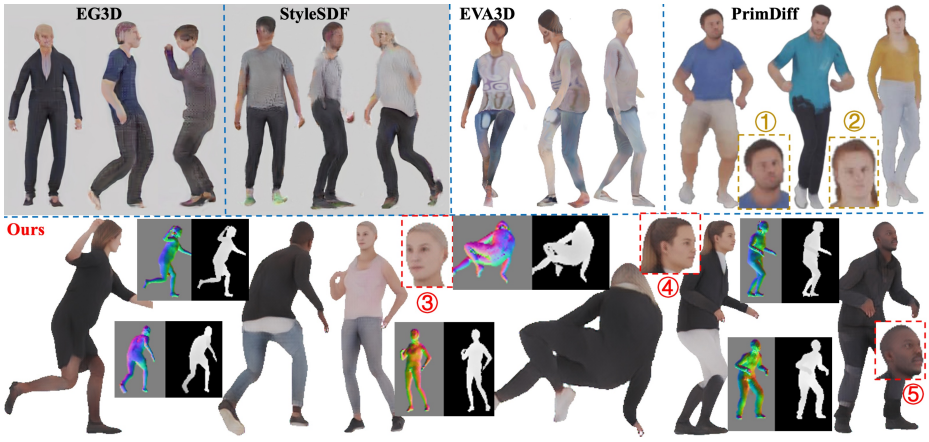

Fig. 4: Qualitative comparisons on RenderPeople [64]. The geometry is visualized as normal/depth maps at 128×64 resolution. The rendered images are cropped to 512×256 in visualization. We synthesize high-quality faces ③④⑤ vs. ①② PrimDiff [11].

4 Experiments

4.1 Experimental Setup

Datasets. We perform experiments on three datasets: UBCFashion [68] (500 monocular human videos with natural fashion motions), THuman 2.0 [67] (526 human scans), and RenderPeople [64] (796 high-quality 3D human models with diverse identities and clothes). We render each 3D scan/model of THuman 2.0 and RenderPeople into multi-view images for training.
Metrics. We measure the diversity and quality of generated human images using the Frechet Inception Distance (FID) [19] between 50k generated images and real images at the resolution of 512×512. In addition, we conduct a perceptual user study and report how often the generated images by our method are preferred over other methods in terms of both overall appearance quality and face quality.
Baseline Methods. EG3D [7] and StyleSDF [53], EVA3D [21], AG3D [12] and PrimDiff [11] are the state-of-the-art methods for 3D-aware generation of static objects and articulated 3D humans respectively.

4.2 Comparisons to SOTA Methods

Quantitative Comparisons. Table. 1 summarizes the quantitative comparisons against SOTA methods. Our approach achieves better results than all four 3D-aware GANs on all datasets in terms of FID. Our method outperforms others by a significant margin on RenderPeople and THUman2.0, and the performance of EVA3D on THUman2.0 is much worse partly because of the small pose variance of THUman2.0. The improvement is also confirmed by our user study in

Fig. 5: Comparisons with PrimDiff (PD) on texture transfer. Ours achieves better results ③④ by mixing the source ①②.

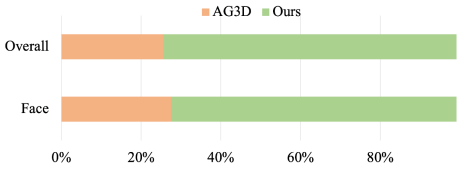

Fig. 6: User study. We report how often the generated images by ours are preferred over AG3D in terms of overall appearance and face quality.

Table 1: Quantitative comparisons with SOTA methods on FID@50K ↓. For reference, we report the quantitative results from the EVA3D [ICLR'23], AG3D [ICCV'23] and PrimDiff [NeurIPS'23] paper.

Methods	UBCFashion	RenderPeople	THUman2.0
StyleSDF [53]	18.52	51.27	-
EG3D [7]	23.95	24.32	-
EVA3D [21]	12.61	44.37	124.54
AG3D [12]	11.04	-	-
PrimDiff [11]	-	17.95	-
Ours	**9.56**	**13.98**	**25.22**

Table 2: Ablation of normalization in diffusion on FID@4K on UBCFashion. Part-aware normalization normalizes each body part latent locally by segmentations (see Fig. 11).

Norm. Method	FID ↓
None	20.13
Standard	24.86
Part-aware	25.02
Struct-aligned	19.58

Fig. 6. More than 20 participants are asked to select the images with better overall quality or face quality from the random generations of AG3D and ours. It was observed that about 74.68% and 72.5% of generated images by our method are considered to be more realistic in terms of overall quality and face quality respectively. More metrics discussion can be found in the supp. mat.

Qualitative Comparisons. The improvements over baselines are further confirmed qualitatively in Fig. 3, where we show the renderings of each approach on UBCFashion. Compared to 3D-aware GANs, StructLDM is capable of generating diverse humans with various clothing styles and different skin colors with powerful diffusion-based sampling. While GAN-based sampling lacks such diversity. In addition, Fig. 3 shows that our method generates view-consistent humans with high-quality appearances and details (*e.g.*, high heels) under different poses and views in different clothing styles including dresses and even for challenging hairstyles, where the consistency is well-preserved in the structured latent space.

We can even learn to generate realistic human images with reasonable geometry reconstructions (*e.g.*, normal, depth) from **single-view images** on DeepFashion [40]. Refer to more details in the supp. mat.

The qualitative comparisons on RenderPeople [64] are shown in Fig. 4, where the learned geometries are visualized as normal and depth maps. Besides consecutive monocular sequences, our method is also capable of learning 3D diffusion models from multi-view images with several static poses, and ours significantly

outperforms existing SOTA 3D-aware GANs in diverse generations. We synthesize high-quality faces ③④⑤ with adversarial training while PrimDiff [11] cannot produce the same level of realism. Ours also generalizes well to different poses, including some extreme poses.

Texture Transfer. Both PrimDiff (PD) and ours support texture transfer by editing UV latents, *i.e.*, combining ① and ② shown in Fig. 5. **PrimDiff renders the combined latent directly in a decoder-free way**, which leads to artifacts. However, ours has an auto-decoder with a global style mixer (G_2) to decode the combined latent, which enables better style mixing ③④.

Efficiency. In inference, our rendering network runs at 9.17 FPS to render 512^2 resolution images on a V100 GPU, 1.94× faster than AG3D. Refer to the supp. mat. for the time cost of training and sampling.

4.3 Ablation Study

Auto-decoder: 2D vs. 1D Latent for Human Reconstruction. We compare the performance of our structured 2D latent with the widely adopted global 1D latent for human reconstructions in auto-decoder both quantitatively and qualitatively in Fig. 7 and Fig. 9. The reconstruction quality measured by LPIPS [73] on 4K samples of UBCFashion is reported in Fig. 7. No improvement is observed for 1D latent when increasing the latent size from 128 to 4096. Whereas, our structured 2D latent outperforms the 1D latent by a large margin in reconstruction. And the quality can be significantly improved by increasing the resolutions or channels of the 2D latent. Note that our structured 2D latent also works at extremely low resolutions, *e.g.*, $16 \times 16 \times 8$, which outperforms the 1D latent of 2048 with the same amount of parameters. Note that though $64 \times 64 \times 32$ achieves comparable performances with $128 \times 128 \times 16$, the latter is more friendly for editing tasks with higher resolutions.

In addition, the qualitative results in Fig. 9 suggest that 1D latent fails to capture detailed high-quality details for face or cloth patterns, and often generates blurry rendering results. This is because the global 1D latent neither encodes semantics nor structure features for the articulated human body. In contrast, our 2D latent captures the face structures and clothing patterns under the same reconstruction supervision (more details in the supp. mat). The comparisons in Fig. 9 are based on the best performances of 1D latent (256) and 2D latent ($128 \times 128 \times 16$) according to the test results in Fig. 7.

Auto-decoder: Adversarial Training. Different from the training on rigid objects with only reconstruction losses as supervision [52], rendering articulated humans with complicated clothing styles is far more challenging. In addition, per-pixel reconstruction losses are often sensitive to misalignment caused by human pose or camera estimation errors, especially for single-view setup. Instead, a discriminator with adversarial training is employed in our auto-decoder framework to enforce realistic rendering with high robustness to pose or camera estimation errors. The adversarial training enables both high-fidelity and high-quality image reconstruction, as confirmed in Fig. 9.

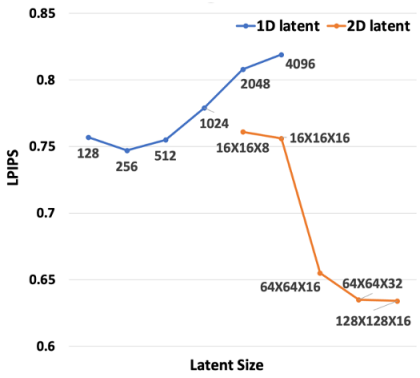

Fig. 7: Ablation of latent size in auto-decoder.

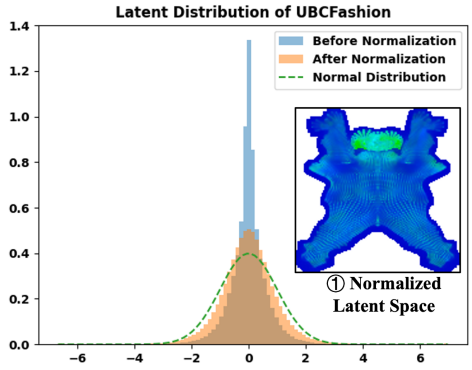

Fig. 8: Ablation of structure-aligned normalization.

Fig. 9: Ablation study of learning auto-decoder (1D vs. 2D latent representation).

Diffusion: Structure-aligned Normalization. We also analyze the normalization methods for our structured latent learning in diffusion quantitatively and qualitatively in Tab. 2 and Fig. 8. Thanks to the structured alignment of the latent space and the auto-decoder, the learned latent space is well-structured and enables the training of diffusion even without normalization ('None'). However, Tab. 2 illustrates that this is non-trivial since the well-adopted standard normalization even prohibits 3D structured diffusion. Instead, a unique structure-aligned normalization (see Sec. 3.4) further improves the generation results quantitatively. Moreover, Fig. 8 suggests that the proposed structure-aligned normalization reduces the distance between the latent distribution learned in auto-decoder and normal distribution. Furthermore, an illustration of the structure-aligned normalization is shown as ① where the standard deviation and mean for each pixel are shown after normalization, which illustrates that our 2D latent encodes the structure information (*e.g.*, symmetry) of the human body in differentiable rendering without explicit latent structure supervision.

Latent Diffusion. We also analyze the effect of latent diffusion in a challenging compositional generation task in Fig. 12. Refer to Sec. 4.4 for more details.

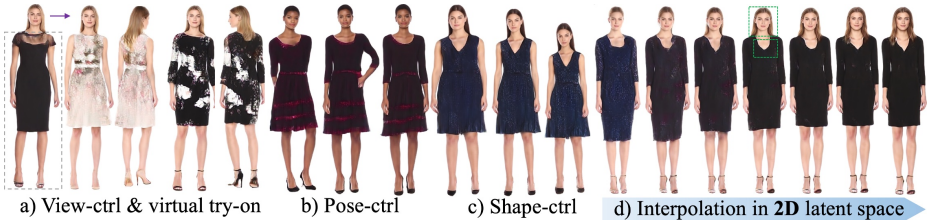

Fig. 10: Controllable generations. The 3D-aware architecture with inherent human body priors enables explicit control over rendering views, human poses, and shapes. We can also smoothly interpolate two samples on the 2D latent space in a similar way to the interpolation on the 1D latent space.

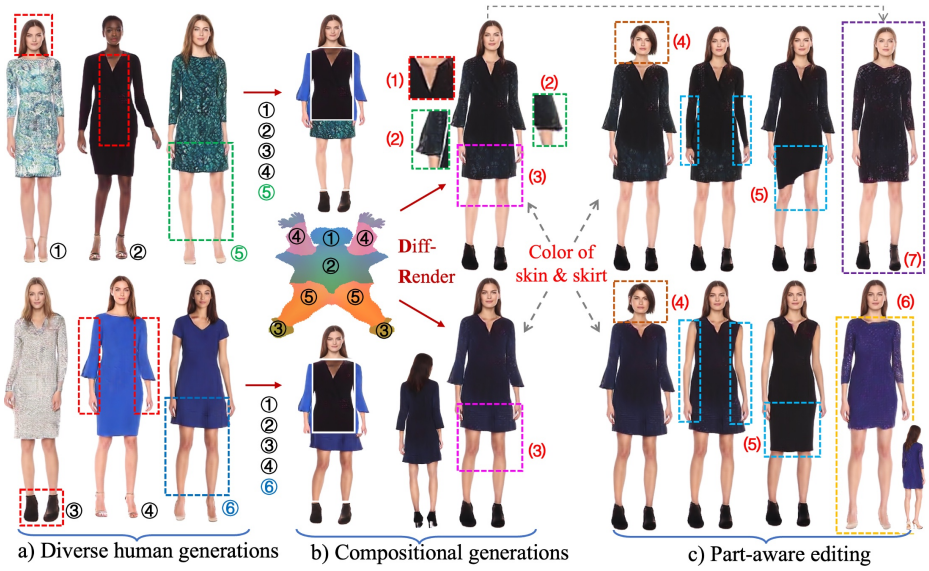

Fig. 11: StructLDM enables compositional 3D human generation and part-aware editing. Taking six body parts from a), coherent composition and blending results can be achieved in b). Using the Diff-Render procedure, part-aware editing enables lots of downstream tasks in c).

4.4 Controllable Human Generation and Editing

Emerging from the technical choice of 2D latent and diffusion model, we show various human generation and editing applications below, which would potentially boost the productivity of fashion designers.

Pose-view-shape Control. Benefiting from the articulated human representation, Struct-LDM enables designers to freely control the generations under different pose, view, and shape conditions as shown in Fig. 10.

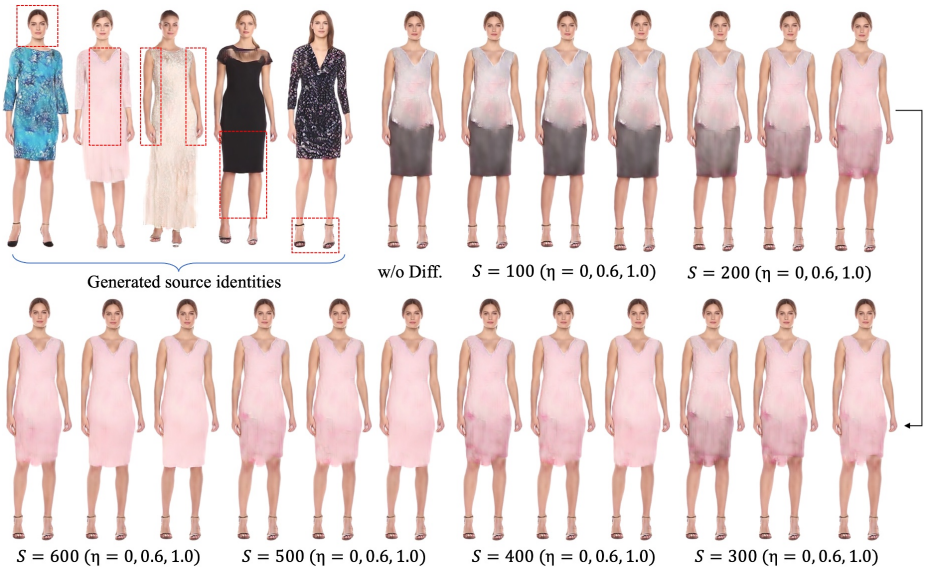

Fig. 12: The effect of latent diffusion in compositional generations.

Interpolation. As shown in Fig. 10 d), we interpolate two latent codes in diffusion [62] to generate a smooth transition. Though only 500 identities of UBC-Fashion are used for training, semantically meaningful 2D latent space can be learned by our auto-decoder and diffusion model.

Compositional Generations. As shown in Fig. 11 b), taking six body parts from six generated source identities in a), StructLDM is capable of blending these parts in the unified structure-aligned latent space and using a Diff-Render procedure for decent style mixing. It includes latent noising and denoising steps by part-aware diffusion, and a rendering step for decent style mixing. Geometry styles (*e.g.*, neckline (1), cuff (2), hemline (3)) are well transferred with high fidelity, and different skin colors are also decently blended, *i.e.*, ② to (1). Note that the skin and skirt colors are different for the compositional generations of ①②③④⑤ and ①②③④⑥ because of our full-body style mixing strategy.

The Effect of Diff-Render. In Fig. 12, we analyze the effect of latent diffusion in Diff-Render in a more challenging compositional generation task, which suggests that the auto-decoder alone fails to blend the pink skirt and the black skirt, *e.g.*, 'w/o Diff'. In contrast, a procedure of latent noising and denoising in diffusion enables plausible color blending with a proper setup of steps S or noise factor η. The diffusion is based on DDIM [62] sampling.

We further detail the effect of Diff-Render in Fig. 13. When we mix different styles, including clothing geometry style and skin or clothing colors, we would like to faithfully preserve the geometry style of each source part and instead blend different colors in a decent way. Yet directly blending the colors, *e.g.*, ①,

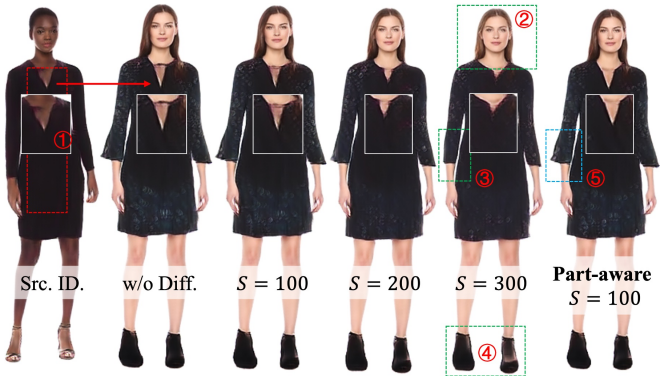

Fig. 13: Part-aware diffusion (supp.) for local enhancement.

Fig. 14: Inversion.

often leads to artifacts shown by 'w/o Diff' in Fig. 13. Though the effect can be alleviated by latent noising and denoising that yields a better color blending with more denoising steps, the fidelity of clothing shape also degrades, *e.g.*, ②③④. Instead, we employ part-aware diffusion (supp. mat.) to solve the local inconsistencies in skin tone by utilizing the learned diffusion prior. It allows users to locally edit or enhance the generations without losing geometry fidelity, and it also requires fewer steps to generate a desired enhancement.

Part-aware Editing. StructLDM allows users to locally edit the generations, as shown in Fig. 11 c). It enables applications such as identity swapping, and local clothing editing, which is achieved by the above-mentioned Diff-Render.

3D Virtual Try-on. As a byproduct of part-aware editing, StructLDM supports 3D virtual try-on, *i.e.*, rendering view-consistent humans wearing different clothes while preserving the identity, as shown in Fig. 11 c) (6) and Fig. 10 a).

Full-body Style Transfer. In addition to part-aware editing, users are also allowed to transfer the color match of a human asset to a new identity (Fig. 11 c) (7)), which is achieved by applying Diff-Render in the full-body latent space.

Inversion. Fig. 14 shows the inversion of in-the-wild images: ① target image, ② inversion, ③ a new pose rendering, via a pre-trained model on DeepFashion.

5 Discussion

We propose a new paradigm for 3D human generation from 2D image collections. The key is the structured 2D latent, which allows better human modeling and editing. A structural auto-decoder and a latent diffusion model are utilized to embed and sample the structured latent space. Experiments on three human datasets show the state-of-the-art performance, and qualitative generation and editing results further demonstrate the advantages of the structured latent.

Limitations. We train models from scratch, and due to the limited scale and dataset bias, diversity is limited. See the supp. mat. for more discussions.

Acknowledgement

This study is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOET2EP20221- 0012), NTU NAP, and 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).

References

1. Abdal, R., Yifan, W., Shi, Z., Xu, Y., Po, R., Kuang, Z., Chen, Q., Yeung, D.Y., Wetzstein, G.: Gaussian shell maps for efficient 3d human generation (2023)
2. Bergman, A.W., Kellnhofer, P., Wang, Y., Chan, E., Lindell, D.B., Wetzstein, G.: Generative neural articulated radiance fields. ArXiv **abs/2206.14314** (2022), https://api.semanticscholar.org/CorpusID:250113850
3. Bergman, A.W., Kellnhofer, P., Wang, Y., Chan, E.R., Lindell, D.B., Wetzstein, G.: Generative neural articulated radiance fields. arXiv preprint arXiv:2206.14314 (2022)
4. Cao, Y., Cao, Y.P., Han, K., Shan, Y., Wong, K.Y.K.: Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. arXiv preprint arXiv:2304.00916 (2023)
5. Cao, Y., Cao, Y.P., Han, K., Shan, Y., Wong, K.Y.K.: Guide3d: Create 3d avatars from text and image guidance. arXiv preprint arXiv:2308.09705 (2023)
6. Chan, E., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., Mello, S.D., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., Karras, T., Wetzstein, G.: Efficient geometry-aware 3d generative adversarial networks. ArXiv **abs/2112.07945** (2021)
7. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16123–16133 (2022)
8. Chan, E.R., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G.: 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. pp. 5799–5809 (2021)
9. Chen, X., Jiang, T., Song, J., Yang, J., Black, M.J., Geiger, A., Hilliges, O.: gdna: Towards generative detailed neural avatars. arXiv (2022)
10. Chen, Y., Wang, X., Zhang, Q., Li, X., Chen, X., Guo, Y., Wang, J., Wang, F.: Uv volumes for real-time rendering of editable free-view human performance. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 16621–16631 (2022), https://api.semanticscholar.org/CorpusID:247762811
11. Chen, Z., Hong, F., Mei, H., Wang, G., Yang, L., Liu, Z.: Primdiffusion: Volumetric primitives diffusion for 3d human generation. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
12. Dong, Z., Chen, X., Yang, J., Black, M.J., Hilliges, O., Geiger, A.: Ag3d: Learning to generate 3d avatars from 2d image collections. ArXiv **abs/2305.02312** (2023), https://api.semanticscholar.org/CorpusID:258461509
13. Frühstück, A., Singh, K.K., Shechtman, E., Mitra, N.J., Wonka, P., Lu, J.: Insetgan for full-body image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7723–7732 (2022)

14. Fu, J., Li, S., Jiang, Y., Lin, K.Y., Qian, C., Loy, C.C., Wu, W., Liu, Z.: Stylegan-human: A data-centric odyssey of human generation. In: European Conference on Computer Vision (2022), https://api.semanticscholar.org/CorpusID:248377018
15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
16. Grigorev, A., Iskakov, K., Ianina, A., Bashirov, R., Zakharkin, I., Vakhitov, A., Lempitsky, V.S.: Stylepeople: A generative model of fullbody human avatars. 2021 (CVPR) pp. 5147–5156 (2021)
17. Gu, J., Trevithick, A., Lin, K.E., Susskind, J.M., Theobalt, C., Liu, L., Ramamoorthi, R.: Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In: International Conference on Machine Learning. pp. 11808–11826. PMLR (2023)
18. Gupta, A., Xiong, W., Nie, Y., Jones, I., Oğuz, B.: 3dgen: Triplane latent diffusion for textured mesh generation. arXiv preprint arXiv:2303.05371 (2023)
19. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
20. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
21. Hong, F., Chen, Z., Lan, Y., Pan, L., Liu, Z.: Eva3d: Compositional 3d human generation from 2d image collections. ArXiv **abs/2210.04888** (2022), https://api.semanticscholar.org/CorpusID:252780848
22. Hong, F., Zhang, M., Pan, L., Cai, Z., Yang, L., Liu, Z.: Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. arXiv preprint arXiv:2205.08535 (2022)
23. Hu, T., Hong, F., Liu, Z.: Surmo: Surface-based 4d motion modeling for dynamic human rendering. In: Computer Vision and Pattern Recognition (CVPR) (2024)
24. Hu, T., Sarkar, K., Liu, L., Zwicker, M., Theobalt, C.: Egorenderer: Rendering human avatars from egocentric camera images. In: ICCV (2021)
25. Hu, T., Xu, H., Luo, L., Yu, T., Zheng, Z., Zhang, H., Liu, Y., Zwicker, M.: Hvtr++: Image and pose driven human avatars using hybrid volumetric-textural rendering. IEEE Transactions on Visualization and Computer Graphics pp. 1–15 (2023). https://doi.org/10.1109/TVCG.2023.3297721
26. Hu, T., Yu, T., Zheng, Z., Zhang, H., Liu, Y., Zwicker, M.: Hvtr: Hybrid volumetric-textural rendering for human avatars. 3DV (2022)
27. Jain, A., Mildenhall, B., Barron, J.T., Abbeel, P., Poole, B.: Zero-shot text-guided object generation with dream fields (2022)
28. Jiang, Y., Yang, S., Qiu, H., Wu, W., Loy, C.C., Liu, Z.: Text2human: Text-driven controllable human image generation. ACM Transactions on Graphics (TOG) **41**(4), 1–11 (2022). https://doi.org/10.1145/3528223.3530104
29. Jun, H., Nichol, A.: Shap-e: Generating conditional 3d implicit functions. arXiv preprint arXiv:2305.02463 (2023)
30. Kajiya, J.T., Herzen, B.V.: Ray tracing volume densities. Proceedings of the 11th annual conference on Computer graphics and interactive techniques (1984)
31. Karras, J., Holynski, A., Wang, T.C., Kemelmacher-Shlizerman, I.: Dreampose: Fashion image-to-video synthesis via stable diffusion (2023)
32. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. In: Proc. NeurIPS (2021)

33. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR. pp. 4401–4410 (2019)
34. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
35. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020)
36. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023), https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/
37. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
38. Lewis, K.M., Varadharajan, S., Kemelmacher-Shlizerman, I.: Tryongan: Body-aware try-on via layered interpolation. ACM Transactions on Graphics (TOG) **40**(4), 1–10 (2021)
39. Liu, X., Ren, J., Siarohin, A., Skorokhodov, I., Li, Y., Lin, D., Liu, X., Liu, Z., Tulyakov, S.: Hyperhuman: Hyper-realistic human generation with latent structural diffusion. arXiv preprint arXiv:2310.08579 (2023)
40. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: CVPR. pp. 1096–1104 (2016)
41. Lombardi, S., Simon, T., Schwartz, G., Zollhoefer, M., Sheikh, Y., Saragih, J.M.: Mixture of volumetric primitives for efficient neural rendering. ACM Transactions on Graphics (TOG) **40**, 1 – 13 (2021)
42. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: a skinned multi-person linear model. ACM Trans. Graph. **34**, 248:1–16 (2015)
43. Luo, S., Hu, W.: Diffusion probabilistic models for 3d point cloud generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2837–2845 (2021)
44. Ma, Q., Saito, S., Yang, J., Tang, S., Black, M.J.: Scale: Modeling clothed humans with a surface codec of articulated local elements. In: CVPR (2021)
45. Ma, Q., Yang, J., Tang, S., Black, M.J.: The power of points for modeling humans in clothing. In: ICCV (2021)
46. Müller, N., Siddiqui, Y., Porzi, L., Bulò, S.R., Kontschieder, P., Nießner, M.: Diffrf: Rendering-guided 3d radiance field diffusion. arXiv preprint arXiv:2212.01206 (2022)
47. Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., Chen, M.: Point-e: A system for generating 3d point clouds from complex prompts. arXiv preprint arXiv:2212.08751 (2022)
48. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171. PMLR (2021)
49. Niemeyer, M., Geiger, A.: Giraffe: Representing scenes as compositional generative neural feature fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11453–11464 (2021)
50. Noguchi, A., Sun, X., Lin, S., Harada, T.: Unsupervised learning of efficient geometry-aware neural articulated representations. arXiv preprint arXiv:2204.08839 (2022)
51. Noguchi, A., Sun, X., Lin, S., Harada, T.: Unsupervised learning of efficient geometry-aware neural articulated representations. In: European Conference on Computer Vision (2022), https://api.semanticscholar.org/CorpusID:248239659

52. Ntavelis, E., Siarohin, A., Olszewski, K., Wang, C., Gool, L.V., Tulyakov, S.: Autodecoding latent 3d diffusion models (2023)
53. Or-El, R., Luo, X., Shan, M., Shechtman, E., Park, J.J., Kemelmacher-Shlizerman, I.: Stylesdf: High-resolution 3d-consistent image and geometry generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13503–13513 (2022)
54. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv (2022)
55. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
56. Remelli, E., Bagautdinov, T.M., Saito, S., Wu, C., Simon, T., Wei, S.E., Guo, K., Cao, Z., Prada, F., Saragih, J.M., Sheikh, Y.: Drivable volumetric avatars using texel-aligned features. ACM SIGGRAPH (2022)
57. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
58. Sarkar, K., Golyanik, V., Liu, L., Theobalt, C.: Style and pose control for image synthesis of humans from a single monocular view. arXiv preprint arXiv:2102.11263 (2021)
59. Sarkar, K., Liu, L., Golyanik, V., Theobalt, C.: Humangan: A generative model of humans images. arXiv preprint arXiv:2103.06902 (2021)
60. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3431–3440 (2014), https://api.semanticscholar.org/CorpusID:1629541
61. Shue, J.R., Chan, E.R., Po, R., Ankner, Z., Wu, J., Wetzstein, G.: 3d neural field generation using triplane diffusion. arXiv preprint arXiv:2211.16677 (2022)
62. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
63. Sun, J., Wang, X., Wang, L., Li, X., Zhang, Y., Zhang, H., Liu, Y.: Next3d: Generative neural texture rasterization for 3d-aware head avatars. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 20991–21002 (2022), https://api.semanticscholar.org/CorpusID:253735045
64. https://renderpeople.com/3d-people/: Renderpeople (2018), https://renderpeople.com/3d-people/
65. Wang, T., Li, L., Lin, K., Lin, C.C., Yang, Z., Zhang, H., Liu, Z., Wang, L.: Disco: Disentangled control for referring human dance generation in real world. arXiv preprint arXiv:2307.00040 (2023)
66. Wang, T., Zhang, B., Zhang, T., Gu, S., Bao, J., Baltrusaitis, T., Shen, J., Chen, D., Wen, F., Chen, Q., et al.: Rodin: A generative model for sculpting 3d digital avatars using diffusion. arXiv preprint arXiv:2212.06135 (2022)
67. Yu, T., Zheng, Z., Guo, K., Liu, P., Dai, Q., Liu, Y.: Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021) (June 2021)
68. Zablotskaia, P., Siarohin, A., Zhao, B., Sigal, L.: Dwnet: Dense warp-based network for pose-guided human video generation. arXiv preprint arXiv:1910.09139 (2019)
69. Zeng, W., Ouyang, W., Luo, P., Liu, W., Wang, X.: 3d human mesh regression with dense correspondence. 2020 IEEE/CVF Conference on Computer Vision and Pat-

tern Recognition (CVPR) pp. 7052–7061 (2020), https://api.semanticscholar.org/CorpusID:219558352
70. Zeng, X., Vahdat, A., Williams, F., Gojcic, Z., Litany, O., Fidler, S., Kreis, K.: Lion: Latent point diffusion models for 3d shape generation. arXiv preprint arXiv:2210.06978 (2022)
71. Zhang, J., Jiang, Z., Yang, D., Xu, H., Shi, Y., Song, G., Xu, Z., Wang, X., Feng, J.: Avatargen: a 3d generative model for animatable human avatars. arXiv preprint arXiv:2208.00561 (2022)
72. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023)
73. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. CVPR pp. 586–595 (2018)
74. Zheng, Z., Huang, H., Yu, T., Zhang, H., Guo, Y., Liu, Y.: Structured local radiance fields for human avatar modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2022)
75. Zhou, L., Du, Y., Wu, J.: 3d shape generation and completion through point-voxel diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5826–5835 (2021)
76. Zwicker, M., Pfister, H., van Baar, J., Gross, M.H.: Ewa splatting. IEEE Trans. Vis. Comput. Graph. **8**, 223–238 (2002), https://api.semanticscholar.org/CorpusID:9389692