Hyper Diffusion Avatars: Dynamic Human Avatar Generation using Network Weight Space Diffusion

Dongliang Cao
University of Bonn

Guoxing Sun
Max Planck Institute for Informatics

Marc Habermann
Max Planck Institute for Informatics

Florian Bernard
University of Bonn

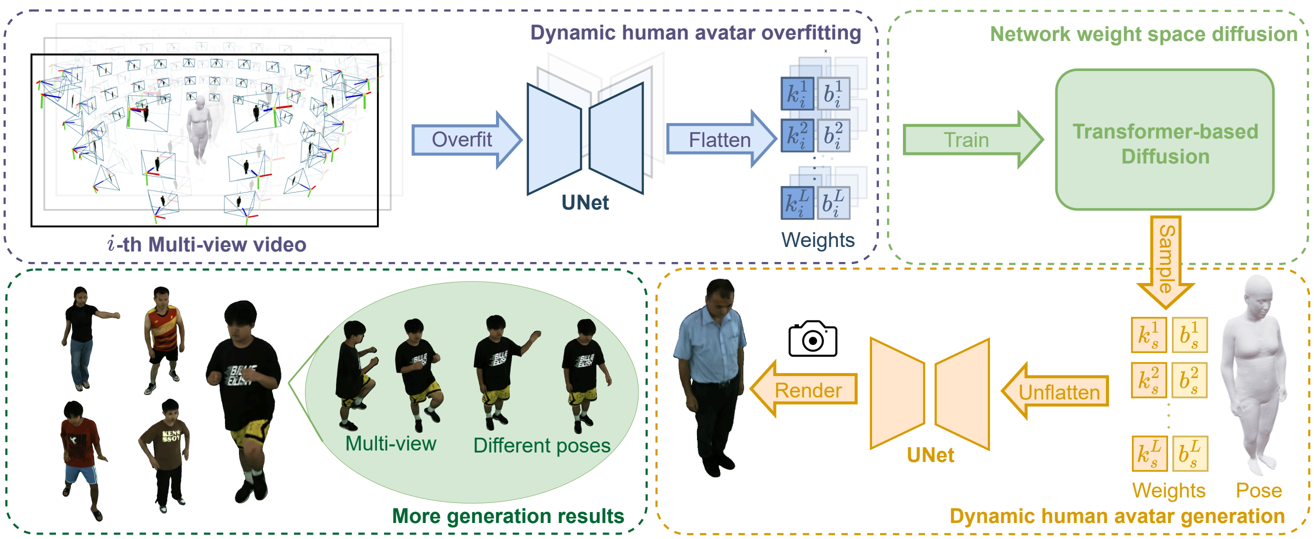

Figure 1. **Our method enables dynamic human avatar generation via diffusion in network weight space.** First, we optimize a set of UNets, each representing an individual dynamic human avatar (top left). Next, we train a transformer network to model a diffusion process over these optimized network weights (top right). At inference time, our approach samples new network weights for real-time, controllable dynamic human avatar rendering by predicting pose-dependent 3D Gaussian Splatting based on a given pose (bottom).

Abstract

Creating human avatars is a highly desirable yet challenging task. Recent advancements in radiance field rendering have achieved unprecedented photorealism and real-time performance for personalized dynamic human avatars. However, these approaches are typically limited to person-specific rendering models trained on multi-view video data for a single individual, limiting their ability to generalize across different identities. On the other hand, generative approaches leveraging prior knowledge from pre-trained 2D diffusion models can produce cartoonish, static human avatars, which are animated through simple skeleton-based articulation. Therefore, the avatars generated by these methods suffer from lower rendering quality compared to person-specific rendering methods and fail to capture pose-dependent deformations such as cloth wrinkles. In this paper, we propose a novel approach that unites the strengths of person-specific rendering and diffusion-based generative modeling to enable ***dynamic human avatar*** *generation with both high photorealism and realistic pose-dependent deformations. Our method follows a two-stage pipeline: first, we optimize a set of person-specific UNets, with each network representing a dynamic human avatar that captures intricate pose-dependent deformations. In the second stage, we train a hyper diffusion model over the optimized network*

weights. During inference, our method generates network weights for real-time, controllable rendering of dynamic human avatars. Using a large-scale, cross-identity, multi-view video dataset, we demonstrate that our approach outperforms state-of-the-art human avatar generation methods.

1. Introduction

Generating high-quality renderings of humans is a crucial challenge in computer vision and computer graphics, with numerous real-world applications in remote communication, movies, gaming, and immersive experiences in augmented as well as virtual reality. Traditionally, generating digital avatars from real-world data requires complicated hardware setups, manual efforts from skilled artists, and advanced physical-based rendering techniques to synthesize the final image [2, 18].

With the advancement of neural radiance fields and subsequent works [28, 48, 79, 80], recent methods [19, 33, 40, 50, 96] have focused on learning photorealistic and controllable human avatars directly from calibrated multi-view videos. Although these approaches achieve unprecedented levels of photorealism, they are still person-specific, meaning that for each individual human a dense multi-view video has to be captured, data has to be processed and annotated, and a dedicated neural model has to be trained from scratch. This process is neither scalable nor fast and resource-efficient as these steps can easily take multiple days [19, 40].

Meanwhile, recent generative methods [16, 54, 61] have made significant progress in generalization quality and scalability, driven by advances in generative diffusion models [21, 39, 68]. To this end, recent avatar generation methods [7, 26, 30, 38, 41] attempted to distill prior knowledge from 2D image diffusion models through score distillation sampling [55]. Despite their compelling results, the rendering quality remains significantly lower than that of person-specific rendering methods. Notably, their rendered videos are unable to capture skeletal pose-dependent deformations like clothing wrinkles and appearance variations, e.g. cast shadows, due to the limitations of simple skeleton-based articulation. To address the limitations mentioned above, for the first time we aim to unify the person-specific rendering method and the diffusion-based generation model to generate photorealistic real-time renderings across different individuals, which faithfully captures pose-dependent deformations.

To this end, we represent the digital human as 3D Gaussians [28] that are parameterized in UV space [23, 50, 63, 75]. In contrast to person-specific rendering methods relying on individual mesh templates [37, 50], we use a parametric human body model (i.e. SMPL-X [44, 51]) to offer a canonical template and a consistent UV space across individuals [93]. However, instead of directly optimizing the 3D Gaussian parameters defined in UV space for each individual, we optimize a UNet [62] that maps the human pose into the Gaussian parameters defined in UV space. To this end, our method is capable of capturing pose-dependent deformation by predicting motion-aware 3D Gaussian parameters. After optimizing the person-specific network for all individuals, we propose a hyper diffusion model, which generates network weights of the optimized UNet rather than 3D Gaussian parameters directly. The motivation for training a diffusion model in this network weight space is two fold: (1) the single network encodes comprehensive pose-dependent 3D Gaussian parameters, as opposed to a static UV Gaussian map; (2) the network weights provide a shared canonical representation across different individuals, as opposed to person-specific rendering methods. During inference, we can directly use our diffusion model to sample network weights and use the generated network to render dynamic digital avatars with the skeletal pose as the input. Fig. 1 provides an overview of our method. We summarize our main contributions as follows:

- For the first time, we unify person-specific rendering and diffusion-based generation to enable dynamic human avatar generation with *pose-dependent deformations*.
- To this end, we encode a dynamic human avatar into a motion-aware network and learn a hyper diffusion model that generates the network weights representing a dynamic avatar.
- To train our hyper diffusion model on network weights, we leverage a transformer-based diffusion model that effectively learns the complex structure of these weights.

2. Related work

In this work, we focus on unconditional dynamic human avatar generation. As a result, human reconstruction methods that rely on multi-view images and the corresponding human pose as inputs at inference time [58, 65, 71, 82, 94, 98] are out of the scope of our paper.

2.1. Personalized 3D human rendering

Recent advancements in neural rendering, such as NeRF [48] and 3DGS [28], have made it possible to learn human avatars directly from calibrated multi-view video inputs. Starting from NeRF [48], various approaches have been proposed to reconstruct the dynamic appearance of 3D humans [1, 17, 24, 36, 40, 81]. The key idea behind these methods is to introduce deformable human NeRFs that deform the posed space to a shared pose canonical space. Despite producing high-quality renderings, these methods inherit the limitations of NeRF-based approaches, resulting in significantly longer rendering times. To overcome this limitation, more recent methods [23, 25, 50, 63, 94] replace NeRF by 3DGS to enable real-time rendering speed while also improving photorealism. Nevertheless, the aforementioned methods primarily focus on achieving photorealistic renderings of a *single personalized human*. In contrast, our method aims at building a generative and dynamic 3D human avatar model by training on large, cross-identity, and multi-view datasets.

2.2. 3D human generation

Recent diffusion-based image generation models have demonstrated unprecedented progress in the context of quality, diversity, and controllability [16, 46, 54, 61, 64]. To this end, numerous efforts [47, 56, 66, 73, 84] have been made to leverage the rich 2D prior knowledge for 3D generation through score distillation sampling [55, 78]. Similarly, recent 3D human generation methods [7, 26, 38, 41, 91] also utilize the idea to optimize the underlying 3D representation, i.e. NeRF or 3DGS, given text or image conditions. Despite their compelling results, these methods suffer from computational inefficiency, due to the involved per-instance optimization [43]. To improve efficiency, most recent works [10, 93] directly train a diffusion model in the underlying 3D representation space, e.g. volumetric primitives [42] or 3DGS UV maps from multi-view human data [77, 87]. Nevertheless, they fail to model pose-dependent deformations by learning a static representation and using solely simple skeleton-based articulation, i.e. linear blend skinning [11]. In contrast, our method trains a diffusion model directly in the network weight space, where the network captures pose-dependent deformations while also achieving real-time rendering speed once the weights have been initialized.

2.3. Diffusion models for generative 3D Gaussian Splatting

In comparison to 2D image generation, generating 3D objects is much more difficult due to the additional dimension and the scarcity of high-quality 3D data [9, 13, 14, 89]. Among all 3D generation methods, there is a line of work that utilizes diffusion models to generate 3DGS [49, 60, 74, 88, 90, 97]. The unstructured nature of 3DGS poses a significant challenge in finding a shared canonical space to train diffusion model. GSD [49] constraints the number of 3D Gaussians, while L3DG [60] embeds 3D Gaussians into a dense latent grid. TriplaneGaussian [97] and DiffGS [95] directly decode 3D Gaussian attributes from generated Triplanes [8]. Omegas [85] trains a 2D diffusion model to predict 2D UV maps of the geometry and materials of 3D objects. To generate dynamic 3D objects, a recent method [59] explicitly introduces the time dimension by leveraging HexPlanes [5]. Nevertheless, none of the existing methods is capable of generating articulated 3D humans due to their inherent complexity and large deformations. To address this, our method introduces the diffusion process in the network weight space, which encapsulates the information of dynamic human avatars. The concept of diffusion in network weight space has been explored in areas such as shape generation [15] and transfer learning [70]. However, our approach is the first to leverage hyper diffusion for dynamic human avatar generation.

3. Background

3.1. SMPL-X

SMPL-X [51] is a 3D parametric human model that represents the human shape (without cloth) consisting of body, hands, and face. This model consists of 10,475 vertices and 54 joints, allowing for control over body shape, body pose, and face expression. The deformation process can be defined as

$$\mathcal{M}(\beta, \theta, \psi) = LBS(T_P(\beta, \theta, \psi), J(\beta), \theta, W), \quad (1)$$

where β, θ and ψ represent shape, pose and expression parameters, respectively. The linear blend skinning (LBS) function [35], denoted as $LBS(\cdot)$, is used to transform the canonical template T_P to the given pose θ based on the skinning weight W and joint locations $J(\beta)$. The canonical template T_P can be computed as

$$T_P(\beta, \theta, \psi) = T_C + B_S(\beta) + B_P(\theta) + B_E(\psi), \quad (2)$$

where T_C is the mean shape and $B_S(\beta), B_P(\theta), B_E(\psi)$ represent per-vertex displacements calculated by the blend shapes S, P, E with their corresponding shape, pose, expression parameters.

3.2. 3D Gaussian Splatting

3D Gaussian Splatting [28] is an explicit point-based representation for novel view synthesis and 3D reconstruction that models static scenes using a collection of 3D Gaussian primitives. These primitives enable real-time rendering through differentiable rasterization. Each Gaussian primitive is parametrized by their center position $\mu \in \mathbb{R}^3$, covariance $\Sigma \in \mathbb{R}^{3\times3}$, color $c \in \mathbb{R}^3$, and opacity $\alpha \in \mathbb{R}$. By projecting 3D Gaussians onto the camera's imaging plane, the 2D Gaussians are assigned and sorted to the corresponding tiles for point-based rendering [99], i.e.

$$\mathrm{c}(p) = \sum_{i \in N} c_i \sigma_i \prod_{j=1}^{i-1} (1 - \sigma_j), \quad (3)$$

where $\sigma_i = \alpha_i \exp(-\frac{1}{2}(p - \mu_i)^\intercal \Sigma_i^{-1}(p - \mu_i))$, and p is the location of queried point and $\mu_i, \Sigma_i, c_i, \alpha_i$ and σ_i are the center position, covariance, color, opacity, and density of the i-th Gaussian primitive, respectively. To model view-dependent appearance, the color c is represented via coefficients of spherical harmonics (SH) [28]. In practice, each

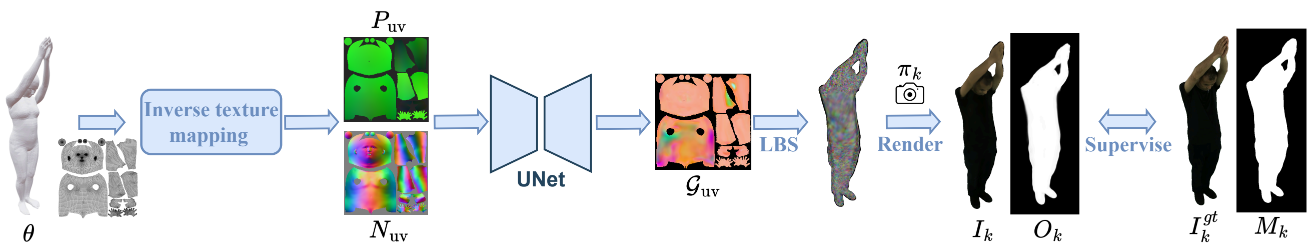

Figure 2. **Dynamic human representation learning based on UNet.** Given a specific human pose, the pose-dependent position and normal maps are generated via inverse texture mapping. These maps serve as inputs to the UNet, which predicts pose-dependent 3D Gaussians for rendering. During training, the UNet is optimized using multi-view RGB image sequences along with their corresponding segmentation masks.

Gaussian is parametrized as $\mathcal{G}_i = \{p_i, s_i, q_i, \alpha_i, h_i\} \in \mathbb{R}^{59}$, including 3D center position $p_i \in \mathbb{R}^3$, scaling $s_i \in \mathbb{R}^3$, quaternion $q_i \in \mathbb{R}^4$, opacity $\alpha_i \in \mathbb{R}$, and spherical harmonics $h_i \in \mathbb{R}^{48}$.

3.3. Denoising diffusion models

Given a dataset of examples drawn independently from a real data distribution $q(x)$, diffusion models aim to learn the data distribution by sequentially denoising random noise samples [21, 67, 69]. During training, the diffusion model defines a forward diffusion process in which a small amount of Gaussian noise is added in T steps, producing a sequence of noisy samples $x_1, ..., x_T$. The step sizes are controlled by a variance schedule $\{\beta_t \in (0, 1)\}_{t=1}^{T}$, i.e.

$$q(x_t|x_{t-1}) = N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I), \tag{4}$$

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}). \tag{5}$$

During inference, the reverse process iteratively removes noise from an input x_T drawn from the Gaussian distribution using the learned denoiser and obtains a clean sample x_0 in the end [21, 68].

4. Our method

In this work, we present an unconditional generative model for synthesizing dynamic human avatars trained on a large, cross-identity, and multi-view human video dataset. Our approach involves two stages. In the first stage, we train a UNet [62] to map 3D skeletal human poses to the corresponding pose-dependent 3DGS for each human avatar individually. In the second stage, we propose a transformer-based [57] hyper diffusion model for generative and photorealistic human modeling, which is trained on the collection of network weights obtained from the first stage. At inference, our model can generate network weights corresponding to valid dynamic human avatars by performing the reverse diffusion process on randomly sampled noise.

4.1. Dynamic human representation

The overall pipeline of dynamic human representation learning is depicted in Fig. 2. Inspired by recent advances in person-specific dynamic human rendering [34, 37, 50], we model each individual human avatar with a dedicated lightweight UNet with network weight w, denoted as $\mathcal{U}_w$. The UNet takes pose-dependent texture as input, specifically the normal texture $N_{uv}(\theta) \in \mathbb{R}^{T\times T\times 3}$ and position texture $P_{uv}(\theta) \in \mathbb{R}^{T\times T\times 3}$, which together encode the body pose θ in the 2D UV space. These textures are derived from the posed template T_P (see Eq.2) via inverse texture mapping [50]. Notably, unlike previous person-specific methods [37, 50], we utilize the mean SMPL-X template (i.e. $\beta = 0, \psi = 0$) instead of a person-specific template mesh, enabling a unified input motion representation across different individuals (i.e. shared UV space and mesh template). The UNet outputs 3D Gaussians parameterized in the same 2D UV space (i.e. $\mathcal{G}_{uv}(\theta) \in \mathbb{R}^{T\times T\times 59}$), such that each texel of the template mesh encodes the parameters of a corresponding 3D Gaussian. This approach effectively binds the Gaussians to the template, enabling accurate and flexible avatar representation. To this end, the UNet learns the pose-dependent 3DGS, i.e.

$$\mathcal{G}_{uv}(\theta) = \mathcal{U}_w(N_{uv}(\theta), P_{uv}(\theta)). \tag{6}$$

After obtaining the Gaussians, we use LBS [35] to transform the positions of Gaussians from canonical pose space to world space:

$$p_{uv} = LBS((a_a \cdot \bar{v}_a + a_b \cdot \bar{v}_b + a_c \cdot \bar{v}_c) + d_{uv}), \tag{7}$$

where $d_{uv} \in \mathbb{R}^{T\times T\times 3}$ denotes the learned offset of Gaussians, $p_{uv} \in \mathbb{R}^{T\times T\times 3}$ represents the final positions of Gaussians in world space, $a_\bullet$ is the barycentric weight on each texel and $\bar{v}_\bullet$ is the corresponding canonical vertex position of the template mesh. To this end, our method models the pose-dependent deformations by learning the pose-dependent offset of Gaussians d_{uv}.

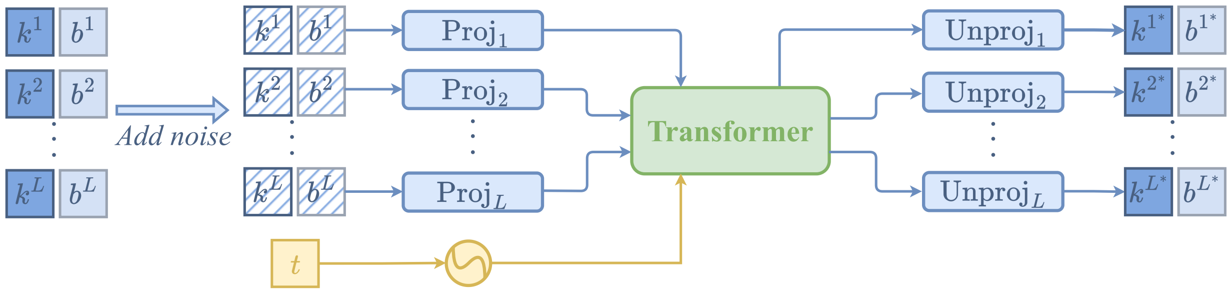

Figure 3. **Diffusion process on network weight space.** During the forward diffusion process, the standard Gaussian noise at time step t is added to the network weights and the transformer take the noisy weights as well as the time step t to predict the denoised weights.

For each camera view k with projection matrix π_k, the resulting 3D Gaussians $\mathcal{G}_{\text{uv}}$ are rendered using a differentiable Gaussian rasterizer $\mathcal{R}$, producing a RGB image $I_k \in \mathbb{R}^{H\times W\times 3}$ and an opacity image $O_k \in \mathbb{R}^{H\times W\times 1}$, i.e.

$$(I_k, O_k) = \mathcal{R}(\mathcal{G}_{\text{uv}}(\theta), \pi_k). \tag{8}$$

To train the UNet $\mathcal{U}_w$, we compute the mean absolute error $\mathcal{L}_{\text{L1}}$ and the structural similarity $\mathcal{L}_{\text{SSIM}}$ between the rendered RGB image I_k and the ground-truth image I_k^{gt}, following prior works [28, 50]. Additionally, we compute the AlexNet-based [31] perceptual loss [92] $\mathcal{L}_{\text{LPIPS}}$ for better visual appearance and the mean absolute error $\mathcal{L}_{\text{mask}}$ between the rendered opacity image O_k and the ground-truth human segmentation mask M_k for better outlines of Gaussian primitives [6]. The overall training loss is a weighted sum of the individual losses, i.e.

$$\begin{aligned}\mathcal{L}_{\text{total}} = \lambda_{\text{pix}}\mathcal{L}_{\text{L1}}(I_k, I_k^{gt}) + \lambda_{\text{str}}\mathcal{L}_{\text{SSIM}}(I_k, I_k^{gt}) + \\ \lambda_{\text{per}}\mathcal{L}_{\text{LPIPS}}(I_k, I_k^{gt}) + \lambda_{\text{m}}\mathcal{L}_{\text{mask}}(O_k, M_k).\end{aligned} \tag{9}$$

In this manner, each dynamic human avatar is represented by its corresponding neural network weights w_i, providing a unified canonical space that accommodates variations in shape and appearance across different individuals. Thus, the per-instance optimization leads to a collection of network weights $\mathcal{W} = \{w_i\}_{i=1}^N$, where N is the number of human individuals. To constrain the network weight distribution, we use a consistent weight initialization [15] instead of random weight initialization.

4.2. Network weight space diffusion

Once we obtain the collection of network weights $\mathcal{W}$, we train a diffusion model to learn the underlying distribution of the network weights as shown in Fig. 3. We consider the set of weights of a given UNet w_i as a sequence of convolutional kernels and biases, i.e.

$$w_i = \{k_i^l, b_i^l\}_{l=1}^L, \tag{10}$$

where $k_i^l \in \mathbb{R}^{C_{\text{out}}^l \times C_{\text{in}}^l \times K_h^l \times K_w^l}$ and $b_i^l \in \mathbb{R}^{C_{\text{out}}^l}$ are the kernel and bias of l-th convolutional layer, respectively. During the forward diffusion process, standard Gaussian noise at step t is added to the network weights w_i and we employ a transformer architecture $\mathcal{T}$ as our diffusion model, following recent approaches [15, 53]. Specifically, for each layer, we add Gaussian noise and flatten the kernel weights and concatenate the corresponding biases, treating this combined vector as a distinct token for the transformer input. This process partitions the entire set of weights w_i into L separate tokens, one for each layer. The layer-wise partitioning preserves the hierarchical structure of the network, which helps the transformer to more effectively capture and learn the complex network weight space. This is in contrast to previous methods [53, 70], which flatten all network weights into a single 1D vector and chunk it into tokens, potentially losing important structural information. Before passing these tokens to the transformer, we project each one into a shared feature space using a linear layer for each token, i.e.

$$t_{\text{in}}^i = \text{Proj}_i(k^i \oplus b^i), \text{ for } i = \{1, \dots, L\}, \tag{11}$$

where t_{in}^i is the projected input token and $\oplus$ indicates concatenation. As a result, tokens with the same dimension can be directly used as input for the transformer. The resulting noisy tokens, along with the sinusoidal embedding of t, are then fed into the transformer $\mathcal{T}$, i.e.

$$(t_{\text{out}}^1, t_{\text{out}}^2, \dots, t_{\text{out}}^L) = \mathcal{T}(t_{\text{in}}^1, t_{\text{in}}^2, \dots, t_{\text{in}}^L, \text{emb}(t)). \tag{12}$$

The transformer $\mathcal{T}$ consists of multiple self-attention and feed-forward layers, facilitating effective information exchange, both within and across the tokens representing each layers. After processing through the transformer, we unproject each token back to its original dimension using a separate linear layer for each token, mirroring the projection performed at the input, i.e.

$$(k^{i^*} \oplus b^{i^*}) = \text{Unproj}_i(t_{\text{out}}^i), \text{ for } i = \{1, \dots, L\}. \tag{13}$$

This yields the denoised network weights w^*.

Following prior approaches [15, 21], we train the a Mean Squared Error (MSE) loss between the denoised weights w^* and the input weights w. During inference we utilize DDIM [68] to sample network weights from the diffusion

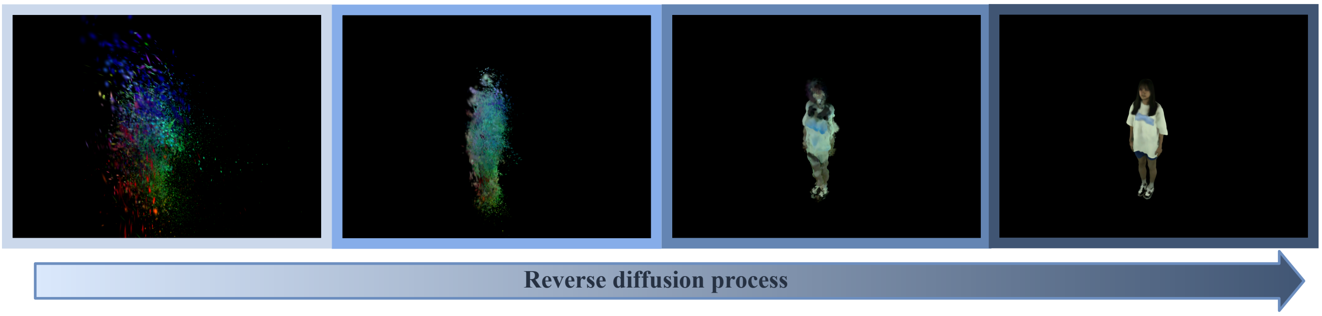

Figure 4. **Denoising network weights at various time steps.** Here the network weights are visualized based on the rendering images. The rendering images show the reverse diffusion process based on DDIM sampling. We observe that UNet weights corrupted by noise fail to represent a valid human avatar. However, the iterative denoising process yields a high-quality human avatar.

model. Fig. 4 shows an example of the denoising process to generate a valid UNet that represents a dynamic human avatar.

4.3. Implementation details

Rather than using the original UNet [62], which contains approximately 30 million learnable parameters and thus poses significant challenges for the diffusion process, we empirically reduce the number of hidden channels to 64, resulting in a lightweight network with only 0.6 million parameters. Each avatar-specific UNet is trained using the AdamW optimizer [29] with a batch size of 1 and a learning rate of 1×10^{-4}. In terms of the training loss, we empirically set $\lambda_{\text{pix}} = 1.0, \lambda_{\text{str}} = 0.1, \lambda_{\text{per}} = 0.01$, and $\lambda_{\text{m}} = 0.1$ in Eq.9. During training, the images are downsampled to 1024×750 and cropped using the segmentation mask, while the UV map resolution is set to 256 for efficiency. Following prior work [50], we employ a 30k-step warm-up and train the UNet for a total of 700k iterations. In the context of network weight space diffusion, we utilize a transformer architecture comprising 12 blocks, each equipped with multi-head self-attention (16 heads) and a feed-forward layer with a hidden dimension of 2048. For training the diffusion model, network weights are standardized to zero mean and unit variance. We use the AdamW optimizer [29] with a batch size of 16 and a learning rate of 2×10^{-4}. The learning rate is reduced by 10% every 200 epochs. We train the transformer for 6000 epochs until convergence.

5. Experimental results

5.1. Datasets and metrics

For evaluation, we utilize the multi-view human dataset MVHumanNet [83], which contains a large number of diverse identities with everyday clothing. To be comparable with baseline methods [10, 93], we manually select 500 video sequences from the first 1500 sequences based on the SMPL-X pose parameter estimation accuracy. Evaluation of unconditional generation of dynamic human avatar can be challenging due to the lack of direct correspondence to ground truth data [15]. Following prior unconditional generation methods [15, 45, 86], we evaluate the methods based on Minimum Matching Distance (MMD), Coverage (COV), and 1-Nearest-Neighbor Accuracy (1-NNA), i.e.

$$\text{MMD}(S_g, S_r) = \frac{1}{|S_r|} \sum_{Y \in S_r} \min_{X \in S_g} D(X, Y),$$

$$\text{COV}(S_g, S_r) = \frac{|\{\arg\min_{Y \in S_r} D(X, Y) \mid X \in S_g\}|}{|S_r|},$$

$$\text{1-NNA}(S_g, S_r) = \frac{\sum_{X \in S_g} \mathbf{1}[N_X \in S_g] + \sum_{Y \in S_r} \mathbf{1}[N_Y \in S_r]}{|S_g| + |S_r|},$$

where S_g, S_r are the set of generated data and reference data respectively, $D(X, Y)$ is the distance function between data sample X and Y, in the 1-NNA metric N_X is a data sample that is closest to X in both generated and reference dataset, i.e.,

$$N_X = \text{argmin}_{K \in S_g \cup S_r / X} D(X, K).$$

Here, we use the PSNR, and the LPIPS [92] with AlexNet features [31] (scaled by 1000) as the distance functions between the rendered images from generated human avatars and the corresponding ground-truth images. Following baseline methods [10, 93], we also adopt Fréchet Inception Distance (FID) [20] and Kernel Inception Distance (KID) [3] to evaluate the quality of rendered images based on the Inception-V3 model [72].

5.2. Comparison

We compare our method to other human avatar generation methods training on multi-view human dataset: PrimDiffusion [10], E3Gen [93]. Notably, both methods can only generate static human avatars, which are animated based on simple skeleton-based articulation (i.e. LBS [11]). We follow the same experiment settings to train them using the 500 multi-view human video sequences from MVHumanNet [83]. To evaluate unconditional generation performance, we generate 500 samples for each method. The

Figure 5. **Qualitative results on unconditional human avatar generation.** Compared to baseline methods, our method is able to generate more photorealistic human avatars.

Methods	$\text{MMD}_{\text{PSNR}} \uparrow$	$\text{MMD}_{\text{LPIPS}} \downarrow$	$\text{COV}_{\text{PSNR}}(\%) \uparrow$	$\text{1-NNA}_{\text{PSNR}}(\%) \downarrow$	FID $\downarrow$	KID $\downarrow$
PrimDiffusion [10]	22.23	26.45	52.3	26.8	41.97	328.46
E3Gen [93]	21.14	32.28	58.2	21.3	32.17	284.31
Ours	**27.52**	**12.13**	**63.8**	**15.7**	**12.68**	**123.26**

Table 1. **Quantitative results on unconditional human avatar generation.** Our method outperforms the prior state-of-the-art methods in the context of rendering quality as well as generation diversity.

quantitative results are summarized in Tab. 1. Compared to baseline methods, our approach generates more photorealistic renderings. Additionally, it outperforms existing techniques in terms of generation diversity. Fig. 5 provides a qualitative comparison demonstrating our method's ability to render more photorealistic dynamic human avatars. Moreover, our method is capable of generating dynamic human avatar with pose-dependent deformations as shown in Fig. 6.

6. Ablation study

In this section, we examine the different choices of the diffusion model. In contrast to images, which have a well-defined grid-like structure and can leverage specialized network architectures such as UNet [61, 62] or DiT [52] for diffusion models, network weights exhibit a more complex and less regular structure. As a result, selecting an appropriate representation for network weights, as well as an effective network architecture for the diffusion process, becomes crucial. Here we ablate different choices for network weight representation and network architectures to identify the most effective configuration. The experiment setting is the same as Sec. 5. Specifically, we compare our layer-wise partitioning to latent diffusion [61] and 1D vector flatten. Fig. 7 illustrates the process of latent diffusion on network weight space. Following recent work [70], the network weights are first reshaped into a 2D feature map. This feature map is then treated as an input image and processed using the standard latent diffusion model [61]. To this end, the training contains two stages. In the first stage, the encoder and decoder are trained to reconstruct the input feature map, with a KL-penalty applied to encourage the latent features to follow a standard normal distribution. In the second stage, a diffusion model is trained on the latent space representation. In the context of 1D vector flattening, the network weights are first flattened into a single vector. This vector is then partitioned into N_{chunk} equal-sized chunks, each with dimension C_{chunk}. If needed, zero-padding is

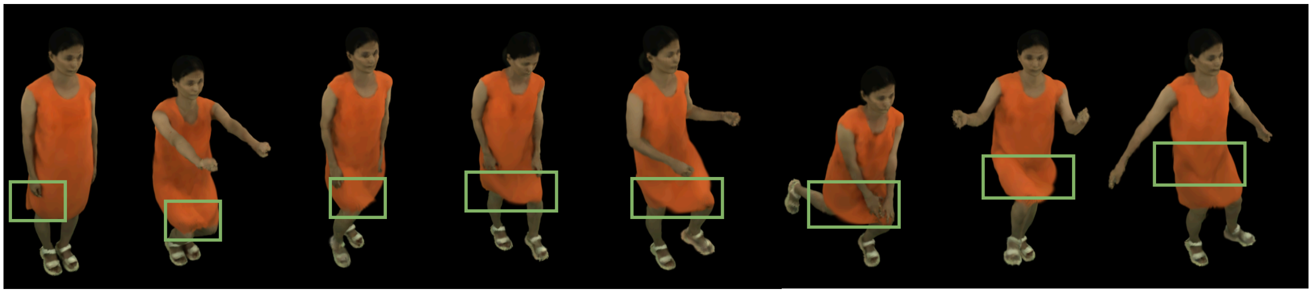

Figure 6. **An example of unconditional human avatar generation of our method.** Rendering sequence demonstrates our method's ability to generate dynamic human avatars. Pose-dependent deformations are emphasized with green rectangles.

	$\text{MMD}_{\text{PSNR}} \uparrow$	$\text{MMD}_{\text{LPIPS}} \downarrow$	$\text{COV}_{\text{PSNR}}(\%) \uparrow$	$\text{1-NNA}_{\text{PSNR}}(\%) \downarrow$	FID $\downarrow$	KID $\downarrow$
Latent diffusion	21.23	27.28	0.4	98.0	58.52	480.34
1D vector flatten	27.12	12.30	54.2	27.6	14.73	134.65
Ours	**27.52**	**12.13**	**63.8**	**15.7**	**12.68**	**123.26**

Table 2. **Ablation study on the choice of the diffusion model.** Our layer-wise partition achieves the best performance in comparison to other choices.

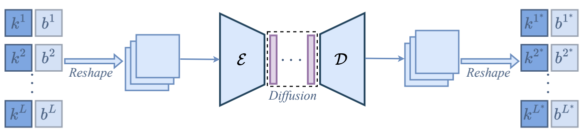

Figure 7. **Latent diffusion model on network weight space.** The network weights are first reshaped into a 2D feature map. An encoder then converts this 2D feature map into a latent space representation. The diffusion process takes place on the latent space. Afterward, a decoder transforms the latent features back into a 2D feature map. Finally, the network weights are recovered by reshaping the 2D feature map.

used to ensure all chunks are the same size. Each chunk is subsequently treated as an input token for the transformer-based diffusion model. Tab. 2 summarizes the results of our ablation study and highlights that our layer-wise partition achieves the best performance. We observe that latent diffusion is prone to mode collapse, resulting in highly similar or nearly identical generations across samples.

7. Limitation and future work

We introduce, for the first time, an unconditional generative model for synthesizing dynamic human avatars through network weight space diffusion. Unlike prior approaches that rely solely on simple skeleton-based articulation, our method enables the generation of photorealistic human avatars with complex, pose-dependent deformations. Despite these advancements, some limitations warrant further investigation. In the first training stage, our method optimizes a UNet to learn motion-aware 3D Gaussians. However, we observe that the UNet shows limited generalization to unseen poses, underscoring the need to enhance its ability to handle novel poses. Additionally, each human avatar is currently represented by a separate UNet, without addressing the entanglement between geometry and appearance. In future work, it would be valuable to explore methods for disentangling geometry and appearance by leveraging relationships across different human avatars [76]. In the context of hyper diffusion, the current method attempts to directly learn the complex, high-dimensional distribution of the network weight space, which poses significant challenges (e.g. neural permutation symmetry [12, 32]) for training and limits generative performance. To address this, it would be valuable to explore approaches such as low-rank adaptation (e.g. LoRA [22]) or network basis learning [4, 27], which could simplify the learning process and enhance generation capabilities.

8. Conclusion

In this work, we present the first method for dynamic human avatar generation that incorporates pose-dependent deformations. Our approach uniquely combines the strengths of person-specific rendering and diffusion-based generative modeling to achieve highly photorealistic results. Specifically, we optimize a set of UNets, each corresponding to an individual human avatar, and leverage a diffusion model trained over the network weights to enable avatar generation. Experimental results demonstrate that our method outperforms existing approaches by producing more photorealistic avatars with accurately learned pose-dependent deformations by evaluating on a large-scale, cross-identity, multi-view video dataset. This contribution paves the way for more realistic human avatar generation in a variety of applications.

References

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

[2] Bernd Bickel, Mario Botsch, Roland Angst, Wojciech Matusik, Miguel Otaduy, Hanspeter Pfister, and Markus Gross. Multi-scale capture of facial geometry and motion. *ACM Transactions on Graphics (ToG)*, 26(3):33–es, 2007. 2

[3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018. 6

[4] Erik L Bolager, Iryna Burak, Chinmay Datar, Qing Sun, and Felix Dietrich. Sampling weights of deep neural networks. In *Advances in Neural Information Processing Systems*, 2023. 8

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

[6] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, et al. Authentic volumetric avatars from a phone scan. *ACM Transactions on Graphics (ToG)*, 41(4):1–19, 2022. 5

[7] 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*, 2024. 2, 3

[8] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022. 3

[9] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 3

[10] Zhaoxi Chen, Fangzhou Hong, Haiyi Mei, Guangcong Wang, Lei Yang, and Ziwei Liu. Primdiffusion: Volumetric primitives diffusion for 3d human generation. In *Advances in Neural Information Processing Systems*, 2023. 3, 6, 7

[11] Matt Cordner and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 811–818. Association for Computing Machinery, 2023. 3, 6

[12] Donato Crisostomi, Marco Fumero, Daniele Baieri, Florian Bernard, and Emanuele Rodola. Cycle-consistent multi-model merging. In *Advances in Neural Information Processing Systems*, 2024. 8

[13] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. *arXiv preprint arXiv:2212.08051*, 2022. 3

[14] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023. 3

[15] Ziya Erkoç, Fangchang Ma, Qi Shan, Matthias Nießner, and Angela Dai. Hyperdiffusion: Generating implicit neural fields with weight-space diffusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023. 3, 5, 6

[16] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024. 2, 3

[17] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J Black, and Timo Bolkart. Capturing and animation of body and clothing from monocular video. In *SIGGRAPH Asia*, 2022. 2

[18] Pablo Garrido, Levi Valgaerts, Chenglei Wu, and Christian Theobalt. Reconstructing detailed dynamic face geometry from monocular video. *ACM Transactions on Graphics (ToG)*, 32(6):158–1, 2013. 2

[19] Marc Habermann, Lingjie Liu, Weipeng Xu, Gerard Pons-Moll, Michael Zollhoefer, and Christian Theobalt. Hdhumans: A hybrid approach for high-fidelity digital humans. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 6(3):1–23, 2023. 2

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017. 6

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020. 2, 4, 5

[22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 8

[23] 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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 2

[24] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. Sherf: Generalizable human nerf from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2

[25] Shoukang Hu, Tao Hu, and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 2

[26] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. Tech: Text-guided reconstruction of lifelike clothed humans. In *International Conference on 3D Vision (3DV)*, 2024. 2, 3

[27] Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. Speeding up convolutional neural networks with low rank expansions. *arXiv preprint arXiv:1405.3866*, 2014. 8

[28] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4):139–1, 2023. 2, 3, 5

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[30] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. In *Advances in Neural Information Processing Systems*, 2023. 2

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012. 5, 6

[32] Daniel Kunin, Javier Sagastuy-Brena, Surya Ganguli, Daniel LK Yamins, and Hidenori Tanaka. Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics. In *International Conference on Learning Representations*, 2020. 8

[33] Youngjoong Kwon, Lingjie Liu, Henry Fuchs, Marc Habermann, and Christian Theobalt. Deliffas: Deformable light fields for fast avatar synthesis. In *Advances in Neural Information Processing Systems*, 2023. 2

[34] Youngjoong Kwon, Baole Fang, Yixing Lu, Haoye Dong, Cheng Zhang, Francisco Vicente Carrasco, Albert Mosella-Montoro, Jianjin Xu, Shingo Takagi, Daeil Kim, et al. Generalizable human gaussians for sparse view synthesis. In *European Conference on Computer Vision*. Springer, 2024. 4

[35] JP Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 165–172, 2000. 3, 4

[36] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhöfer, Jürgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. In *European Conference on Computer Vision*, 2022. 2

[37] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 2, 4

[38] 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. 2, 3

[39] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2022. 2

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

[41] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 3

[42] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021. 3

[43] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 3

[44] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (ToG)*, 34(6), 2015. 2

[45] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021. 6

[46] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[47] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023. 3

[48] 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. *Communications of the ACM*, 65(1):99–106, 2021. 2

[49] Yuxuan Mu, Xinxin Zuo, Chuan Guo, Yilin Wang, Juwei Lu, Xiaofeng Wu, Songcen Xu, Peng Dai, Youliang Yan, and Li Cheng. Gsd: View-guided gaussian splatting diffusion for 3d reconstruction. In *European Conference on Computer Vision*, 2024. 3

[50] Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. Ash: Animatable gaussian splats for efficient and photoreal human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 4, 5, 6

[51] 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 of the IEEE/CVF international conference on computer vision*, 2019. 2, 3

[52] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023. 7

[53] William Peebles, Ilija Radosavovic, Tim Brooks, Alexei A Efros, and Jitendra Malik. Learning to learn with generative models of neural network checkpoints. *arXiv preprint arXiv:2209.12892*, 2022. 5

[54] 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. *arXiv preprint arXiv:2307.01952*, 2023. 2, 3

[55] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations*, 2022. 2, 3

[56] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. In *International Conference on Learning Representations*, 2024. 3

[57] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 4

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

[59] Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023. 3

[60] Barbara Roessle, Norman Müller, Lorenzo Porzi, Samuel Rota Bulò, Peter Kontschieder, Angela Dai, and Matthias Nießner. L3dg: Latent 3d gaussian diffusion. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 3

[61] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022. 2, 3, 7

[62] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention*. Springer, 2015. 2, 4, 6, 7

[63] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable gaussian codec avatars. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 2

[64] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, 2024. 3

[65] Ashwath Shetty, Marc Habermann, Guoxing Sun, Diogo Luvizon, Vladislav Golyanik, and Christian Theobalt. Holoported characters: Real-time free-viewpoint rendering of humans from sparse rgb cameras. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2024. 2

[66] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 3

[67] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015. 4

[68] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 2, 4, 5

[69] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, 2019. 4

[70] Bedionita Soro, Bruno Andreis, Hayeon Lee, Wonyong Jeong, Song Chong, Frank Hutter, and Sung Ju Hwang. Diffusion-based neural network weights generation. In *International Conference on Learning Representations*, 2025. 3, 5, 7

[71] Guoxing Sun, Rishabh Dabral, Heming Zhu, Pascal Fua, Christian Theobalt, and Marc Habermann. Real-time free-view human rendering from sparse-view rgb videos using double unprojected textures. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2025. 2

[72] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2016. 6

[73] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023. 3

[74] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In *International Conference on Learning Representations*, 2024. 3

[75] Kartik Teotia, Hyeongwoo Kim, Pablo Garrido, Marc Habermann, Mohamed Elgharib, and Christian Theobalt. Gaussianheads: End-to-end learning of drivable gaussian head avatars from coarse-to-fine representations. *ACM Transactions on Graphics (ToG)*, 43(6):1–12, 2024. 2

[76] Ayush Tewari, Xingang Pan, Ohad Fried, Maneesh Agrawala, and Christian Theobalt. Disentangled3d: Learning a 3d generative model with disentangled geometry and appearance from monocular images. In *Proceedings of*

the IEEE/CVF conference on computer vision and pattern recognition, 2022. 8

[77] https://renderpeople.com/3d-people/. Renderpeople, 2018. 3

[78] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023. 3

[79] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Advances in Neural Information Processing Systems*, 2021. 2

[80] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[81] 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*, 2022. 2

[82] Junjin Xiao, Qing Zhang, Yongwei Nie, Lei Zhu, and Wei-Shi Zheng. RoGSplat: Learning robust generalizable human gaussian splatting from sparse multi-view images. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2025. 2

[83] 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*, 2024. 6

[84] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360deg views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[85] Xingguang Yan, Han-Hung Lee, Ziyu Wan, and Angel X. Chang. An object is worth 64x64 pixels: Generating 3d object via image diffusion, 2024. 3

[86] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019. 6

[87] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021. 3

[88] LAN Yushi, Shangchen Zhou, Zhaoyang Lyu, Fangzhou Hong, Shuai Yang, Bo Dai, Xingang Pan, and Chen Change Loy. Gaussiananything: Interactive point cloud flow matching for 3d generation. In *International Conference on Learning Representations*, 2025. 3

[89] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions on Graphics (ToG)*, 42(4):1–16, 2023. 3

[90] Bowen Zhang, Yiji Cheng, Jiaolong Yang, Chunyu Wang, Feng Zhao, Yansong Tang, Dong Chen, and Baining Guo. Gaussiancube: Structuring gaussian splatting using optimal transport for 3d generative modeling, 2024. 3

[91] Jingbo Zhang, Xiaoyu Li, Qi Zhang, Yanpei Cao, Ying Shan, and Jing Liao. Humanref: Single image to 3d human generation via reference-guided diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[92] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 5, 6

[93] Weitian Zhang, Yichao Yan, Yunhui Liu, Xingdong Sheng, and Xiaokang Yang. E 3gen: Efficient, expressive and editable avatars generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024. 2, 3, 6, 7

[94] Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 2

[95] Junsheng Zhou, Weiqi Zhang, and Yu-Shen Liu. Diffgs: Functional gaussian splatting diffusion. In *Advances in Neural Information Processing Systems*, 2024. 3

[96] Heming Zhu, Fangneng Zhan, Christian Theobalt, and Marc Habermann. Trihuman: a real-time and controllable tri-plane representation for detailed human geometry and appearance synthesis. *ACM Transactions on Graphics (ToG)*, 44(1):1–17, 2024. 2

[97] Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 3

[98] Anton Zubekhin, Heming Zhu, Paulo Gotardo, Thabo Beeler, Marc Habermann, and Christian Theobalt. Giga: Generalizable sparse image-driven gaussian avatars. *arXiv*, 2025. 2

[99] Matthias Zwicker, Hanspeter Pfister, Jeroen Van Baar, and Markus Gross. Ewa splatting. *IEEE Transactions on Visualization and Computer Graphics*, 8(3):223–238, 2002. 3