GIGA: Generalizable Sparse Image-Driven Gaussian Humans

Anton Zubekhin[1,2] Heming Zhu[1] Paulo Gotardo[3] Thabo Beeler[3]
Marc Habermann[1,2] Christian Theobalt[1,2]
[1] Max Planck Institute for Informatics, Saarland Informatics Campus [2] VIA Research Center [3] Google

Abstract

Driving a high-quality and photorealistic full-body virtual human from a few RGB cameras is a challenging problem that has become increasingly relevant with emerging virtual reality technologies. A promising solution to democratize such technology would be a generalizable method that takes sparse multi-view images of any person and then generates photoreal free-view renderings of them. However, the state-of-the-art approaches are not scalable to very large datasets and, thus, lack diversity and photorealism. To address this problem, we propose GIGA, a novel, generalizable full-body model for rendering photoreal humans in free viewpoint, driven by a single-view or sparse multi-view video. Notably, GIGA can scale training to a few thousand subjects while maintaining high photorealism and synthesizing dynamic appearance. At the core, we introduce a MultiHeadUNet architecture, which takes an approximate RGB texture accumulated from a single or multiple sparse views and predicts 3D Gaussian primitives represented as 2D texels on top of a human body mesh. At test time, our method performs novel view synthesis of a virtual 3D Gaussian-based human from 1 to 4 input views and a tracked body template for unseen identities. Our method excels over prior works by a significant margin in terms of identity generalization capability and photorealism.

1. Introduction

Driving your own full-body and photoreal virtual double from a scarce and affordable single or sparse-view (up to 4) camera setup has the potential to revolutionize communication, gaming, and remote collaboration. However, there are to date modeling challenges that remain unsolved: 1) Achieving photorealism and fidelity despite sensor scarcity and limited input data. 2) Generalization ability to novel identities unseen during training. In this work, we attempt to jointly solve them by leveraging recent large-scale data capture efforts. Notably, this requires a generalizable method to synthesize digital humans at test time in a simple feed-forward manner, which is the subject of this work.

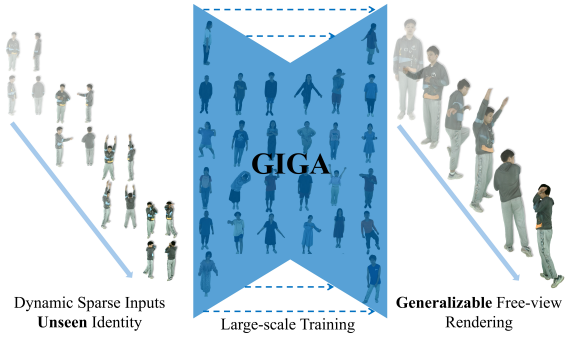


Figure 1. GIGA is trained efficiently on a large-scale human dataset. Given sparse views of an *unseen* identity, and respective skeletal poses, GIGA generates photorealistic dynamic renderings in free viewpoint.

Recent related works have most focused on *person-specific* avatars, i.e., a learned representation that is trained per subject on dense dome-like camera setups, which can be later articulated with arbitrary skeletal motion. These representations may involve meshes [1, 6, 59], neural radiance fields [12, 37, 53, 69], points [50], or volumetric primitives [40] such as 3D Gaussians [22, 31, 34]. Democratizing such technology is difficult since, for high-quality results, a dense camera dome is required *before* the character can be driven at inference time. Some methods aim to utilize simpler capture setups, e.g., monocular images [61, 62] or videos [9, 54]. Nevertheless, due to the much scarcer input, their visual quality often falls short compared to multi-view methods. Single-image to 3D reconstruction methods [11, 44, 45, 61, 62, 71, 73] also primarily focus on adequate geometry reconstruction and produce a 3D asset, drivable with skeletal motion only, which typically lacks the skeletal motion-dependent geometry and appearance changes.

Designing a high-quality approach for lightweight human reconstruction and rendering, thus, remains an open problem: the model must accurately represent diverse hu-

man appearances, body types, and clothing configurations, and it must correctly derive pose-dependent appearance changes from the scarce input signal. While previous works [18, 19, 33, 56] have shown promising first steps towards identity generalization, those methods were trained and evaluated on small-scale datasets [38] with few training subjects. Moreover, the quality and efficiency of these works are constrained by the limitations of implicit neural representations and non-scalable network architectures. More recently, large-scale datasets [3, 60] helped to pave the way for truly generalizable human modeling methods. As was shown by Guo et al. [5], learning from a large data collection enables construction of a rich human avatar prior, albeit requiring preparation of personalized body templates beforehand. Nonetheless, training and evaluating such models at scale demands particular consideration: First, a generalizable model has to learn meaningful feature representations to scale, both to the training set and outside of it. Second, the model architecture must be computationally and memory-efficient to capture fine detail and enable large-scale training. Third, the underlying 3D representation has to yield high-quality, efficient rendering.

To address these challenges, we propose GIGA, a feed-forward method to synthesize virtual 3D-Gaussian-based humans from single or sparse input views and a tracked body template at inference for free-view rendering. Notably, for unseen subjects, GIGA *requires no personalized training on dense dome data* and no fitting of personalized mesh templates. At the core, we project sparse-view image features into the UV space of the SMPL-X model [35] while the virtual human predicted by GIGA is represented as a set of texel-aligned 3D Gaussians [14, 19, 63, 71]. We propose a MultiHeadUNet, a UNet with multiple encoding and decoding heads, which takes the texture containing the projected image information as well as shape and motion codes and regresses per-texel Gaussian appearance and geometry parameters. In detail, we employ cross attention to inject motion information into the model, and skip-connections to propagate the learning signal at different spatial scales within the network. Our architecture design choices ensure reliable learning of intrinsic feature statistics from the training data while maintaining person-specific information contained in the model inputs. The predicted 3D Gaussians can be posed into 3D space using the respective SMPL-X body pose and further rendered onto novel views. In summary, our contributions are:

- We introduce GIGA, a novel generalizable method for dynamic novel view synthesis of virtual humans operating in a single or sparse view setup in a feed-forward regime at test time.
- We design a MultiHeadUNet with motion conditioning to synthesize high-quality texel-aligned 3D Gaussians with dynamic appearance changes.
- We train GIGA on a large-scale collection of multi-view videos, after which it generalizes to unseen identities within and across datasets, producing dynamic novel view renders even with as few as a single input video.

We demonstrate the generalization capabilities of GIGA through a comprehensive evaluation on large-scale multi-view [3, 60, 65] and monocular videos [13] datasets. Experimental results show that GIGA significantly outperforms prior works in visual quality as well as identity and pose generalization.

2. Related Work

Person-specific Human Capture and Rendering. Neural implicit approaches for novel view synthesis [30, 32, 52] were primarily designed for per-scene optimization. This paradigm naturally extended to human modeling through various implicit representations: radiance fields [12, 38, 49], signed distance functions [53, 57, 69], and occupancy fields [29, 44, 45]. These methods typically rely on parametric human body models [27, 35] as geometric proxies to initialize neural implicit representations. The inherent limitations of generic parametric templates in capturing person-specific geometry led to methods utilizing personalized mesh templates [6, 23, 40], which demonstrate superior reconstruction quality, mimicking pose-dependent appearance changes. Drivable Volumetric Avatars (DVA) [40] target telepresence applications by representing human avatars as mixtures of volumetric primitives [26], regressed from texel-aligned image and pose features extracted from 3 input views. Holoported Characters [47] combines a personalized mesh template with dynamic feature textures predicted from partial texture and normal maps, operating with 4 input views, which yields 4K rendering resolution. 3D Gaussian Splatting [15] marked a significant advancement in both rendering quality and computational efficiency. 3D Gaussians have been successfully adapted for free-viewpoint human avatar rendering in multi-view [14, 31, 34, 72] and monocular [8, 9, 39] setups, mostly constrained to person-specific optimization without cross-identity training.

Generalizable Human Capture and Rendering. Recent research has addressed the challenge of synthesizing free-viewpoint videos of human performances from sparse multi-view captures with generalization across subjects [7, 10, 70]. Neural Human Performer (NHP) [18] integrates pixel-aligned visual features with skeletal pose information extracted from multi-view video sequences through cross-attention. TransHuman [33] addresses performance degradation of NHP in occluded regions by learning to synthesize a NeRF-based representation in the canonical pose space with a transformer architecture that processes individual body parts as tokens. Neural Novel Actor (NNA) [56] disentangles appearance and geometry by processing

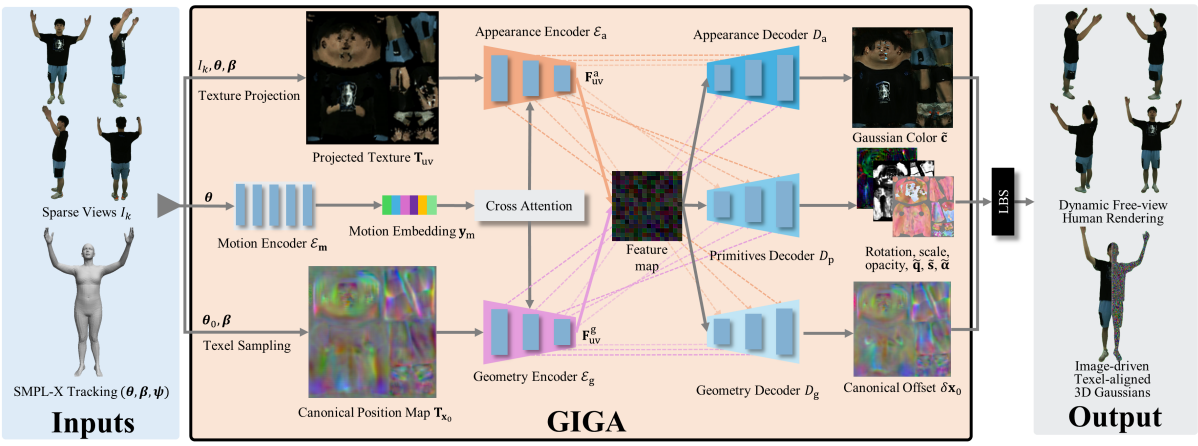


Figure 2. **Method Overview.** GIGA generates dynamic textures of 3D Gaussians for photoreal avatars from sparse input views $\mathcal{I}_k$ and a tracked body template $(\boldsymbol{\theta}, \boldsymbol{\beta})$. An initial RGB texture $\mathbf{T}_{\text{uv}}$ is gathered from the input images and passed to the appearance encoder $\mathcal{E}_{\text{a}}$ to extract appearance features $\mathbf{F}^{\text{a}}_{\text{uv}}$. A canonical position map $\mathbf{T}_{\mathbf{x}_0}$ is processed by the geometry encoder $\mathcal{E}_{\text{g}}$ into geometry features $\mathbf{F}^{\text{g}}_{\text{uv}}$. Both intermediate features are motion-dependent via cross-attention conditioning on the observed pose $\boldsymbol{\theta}$. Gaussian texel maps are regressed by separate decoders, each for an individual group of parameters. Skip-connections (dashed lines) propagate intermediate features from encoders to decoders. Output of GIGA is articulated with linear blend skinning.

spatial point features and SMPL surface features through a graph CNN, enabling separate prediction of person-specific appearance and pose-dependent deformation, as proposed by Liu et al. [24].

More recently, GPS-Gaussian [68] achieved real-time novel view synthesis of human performances by predicting per-pixel 3D Gaussians from 2 source views, which are unprojected to 3D space and aggregated for final rendering. Notably, GPS-Gaussian allows for view interpolation when source views have at most a 60-degree angular difference, thus requiring at least 6 input views in total. Generalizable Human Gaussians (GHG) [19] introduces an alternative approach for reconstructing static human models from sparse input views. By leveraging a large-scale dataset of textured human 3D scans [65], GHG demonstrates generalization to unseen subjects by representing virtual humans as multiple scaffolds of 3D Gaussians in the UV space. The most recent Vid2Avatar-pro [5] achieved animatable avatar generation from a monocular video by training on a large corpus of multi-view data. However, Vid2Avatar-pro requires a high-quality personalized body template at test time, which can only be reconstructed in a few hours for each novel identity.

Generative Methods. Another line of work focuses on single-image to animatable avatar reconstruction through generative modeling. The most common technique here is to rely on SMPL-X and its UV mapping to anchor 3D Gaussians. IDOL [71] employs a transformer-based encoder to combine image features with a learnable UV token to later decode 3D Gaussians with a convolutional network. Following IDOL, SIGMAN [63] learns a variational autoencoder [16] to map sparse views to 3D Gaussians, which is followed up by a diffusion transformer [36] to generate latents for novel avatars, conditioned on an image or a text prompt. Both IDOL and SIGMAN combine real-world datasets with synthetic data, generated with custom pipelines. MoGa [73] jointly optimizes a diffusion model to learn per-avatar latents and an autodecoder to convert latents to 3D avatars. At test time, MoGa optimizes a latent code for the input image and uses a multi-view diffusion model to generate additional pseudo ground truth views for this latent inversion process.

3. Method

In this section, we first formalize the problem setting and relevant background knowledge (Sec. 3.1). Then we introduce our generalizable human representation (Sec. 3.2), our training strategy (Sec. 3.3), and finally the implementation details (Sec. 3.4).

3.1. Problem Setting and Background

Problem Setting. GIGA aims at mapping single/sparse image observations (1-4 views) of an *unseen* human to a 3D Gaussian representation that can be rendered from a free viewpoint, faithfully preserving the appearance and clothing dynamics from these driving videos, even for unseen identities. Fig. 2 presents an overview of our method. To render dynamic sequences, GIGA processes each frame independently. Following the common settings [19, 40, 47], we assume that a motion tracking and a parametric body template are given to roughly describe the body shape and

the pose of the human. Importantly, our method requires *no personalized training*.

For training, GIGA assumes a collection of multi-view videos of several hundred subjects with per-frame subject segmentation masks. Each subject is captured by at least $\hat{K} = 16$ calibrated cameras, with $\pi_{\hat{k}}$ denoting the projection matrix for camera $\hat{k}$. Each video frame is annotated with SMPL-X [35] parameters released as additional labels in most multi-view human performance capture datasets [3, 60, 65]. During training, we randomly sample 3 or 4 input views and use estimated SMPL-X parameters to construct the input for the encoder network. Images from the remaining views serve as ground truth to supervise the outputs decoded by the network.

3D Gaussian Splatting (3DGS). 3DGS [15] models a scene using 3D Gaussian primitives. Each primitive $\mathcal{G} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha, \mathbf{c}\}$ is paramterized by its position $\boldsymbol{\mu} \in \mathbb{R}^3$, covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{3\times3}$, opacity $\alpha \in \mathbb{R}$, and RGB colors $\mathbf{c} \in \mathbb{R}^3$. The covariance matrix $\boldsymbol{\Sigma}$ is factorized as $\boldsymbol{\Sigma} = \mathbf{RSS}^T\mathbf{R}^T$, where the rotation $\mathbf{R}$ matrix is obtained from the quaternion $\mathbf{q} \in \mathbb{R}^4$ and the diagonal scaling matrix $\mathbf{S} = \mathrm{diag}\,(\mathbf{s})$, with per-axis scales $\mathbf{s} \in \mathbb{R}^3$. The Gaussians $\mathcal{G}$ is then rendered from the target camera π to an image I and an accumulated density image A using a Gaussian rasterizer

$$I, A = \mathcal{R}(\mathcal{G}, \pi). \tag{1}$$

SMPL-X Human Body Template. SMPL-X [35] is a parametric human body model with articulated limbs, hands, and an expressive face. SMPL-X has $N = 10475$ vertices, $J = 54$ joints, and can be articulated using linear blend skinning [21] to obtain a set of posed vertices $\mathcal{V} \in \mathbb{R}^{N\times3}$. Formally, SMPL-X is defined as a function $\mathcal{V} = \mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\psi})$, with $\boldsymbol{\theta} \in \mathbb{R}^{(J+1)\times3}$ describing J joint angles and a rigid root transformation, body shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$, and face expression $\boldsymbol{\psi} \in \mathbb{R}^{10}$. As we tested GIGA without expression tracking, we fixed $\boldsymbol{\psi}$ to the neutral expression.

3.2. Generalizable Human Representation

Inspired by previous *personalized*, animatable Gaussian avatar methods [14, 34, 55], we represent virtual humans as texel-aligned 3D Gaussian maps within the texture space $\mathcal{M}_{\mathrm{uv}}$ of the SMPL-X template mesh, displacing each 3D Gaussian in the canonical pose with a motion-dependent offset for better geometry modeling. Texel-aligned 3D Gaussians are semantically similar across various characters and 2D textures can be efficiently processed by convolutional neural networks. Thus, the predicted Gaussians from the same region of the texture space will be entangled thanks to the locality bias of convolutions. As such, GIGA can learn cross-subject information and predict the final texel-aligned 3D Gaussians with a single pass of a 2D convolutional neural network.

Input Appearance Encoding. We use information from the input views to capture identity-specific appearance with pose-dependent variations. To transform this information from image to texture space, we first compute partial textures $\mathbf{T}_{\mathrm{uv},k}$ for each view π_k by projecting each pixel onto the 3D surface of SMPL-X. This 3D point can be mapped into texture space using the UV mapping of SMPL-X such that the final image pixel color is projected onto the respective 2D location in the texture map. We then fuse those textures according to visibility, resulting in the final texture $\mathbf{T}_{\mathrm{uv}} \in \mathbb{R}^{T\times T\times3}, T = 512$ in the UV space $\mathcal{M}_{\mathrm{uv}}$.

This texture is the input to our appearance encoder $\mathcal{E}_{\mathrm{a}}$:

$$\mathbf{F}^{\mathrm{a}}_{\mathrm{uv}}, \mathbf{H}^{\mathrm{a}} = \mathcal{E}_{\mathrm{a}}\left(\mathbf{T}_{\mathrm{uv}}; \mathbf{y}_m\right), \tag{2}$$

which extracts appearance features $\mathbf{F}^{a}_{\mathrm{uv}} \in \mathbb{R}^{T_f\times T_f\times d}$, that encode character-specific appearance and identity information. The encoder $\mathcal{E}_{\mathrm{a}}$ consists of 2D convolutional downsampling residual blocks. Feature maps $\mathbf{H}^{\mathrm{a}}$ from also taken from the downsampling levels for later usage in decoding. $\mathcal{E}_{\mathrm{a}}$ is conditioned on the motion embedding $\mathbf{y}_{\mathrm{m}}$. Following Rombach et al. [41], this conditioning is implemented using cross-attention blocks as final layers of the encoder.

Motion Embeddings. Since the input texture $\mathbf{T}_{\mathrm{uv}}$ contains only a localized pose-dependent learning signal from the template surface. We additionally use an MLP-based motion encoder $\mathcal{E}_{\mathrm{m}}$ to construct motion embeddings for SMPL-X poses $\boldsymbol{\theta}$, which adds global body motion awareness at the encoding stage:

$$\mathbf{y}_{\mathrm{m}} = \mathcal{E}_{\mathrm{m}}\left(\boldsymbol{\theta}\right). \tag{3}$$

Input Geometry Encoding. Even though appearance information is texel-aligned, it is insufficient to infer correct human shapes. To address this, we employ a geometry encoder $\mathcal{E}_{\mathrm{g}}$ to extract approximate geometry information from the body template. Starting with a T-posed SMPL-X mesh $\mathcal{V}(\boldsymbol{\theta}_0, \boldsymbol{\beta})$, we compute a canonical position map $\mathbf{T}_{\mathbf{x}_0} \in \mathbb{R}^{T\times T\times3}$ from canonical vertex coordinates $\mathbf{x}_0\left(\boldsymbol{\beta}\right) \in \mathbb{R}^{N\times3}$ using the UV map $\mathcal{M}_{\mathrm{uv}}$. The geometry encoder $\mathcal{E}_{\mathrm{g}}$ has the same architecture as the appearance encoder $\mathcal{E}_{\mathrm{a}}$ and produces geometry features $\mathbf{F}^{\mathrm{g}}_{\mathrm{uv}} \in \mathbb{R}^{T_f\times T_f\times d}$ with a stack of feature maps $\mathbf{H}^{\mathrm{g}}$:

$$\mathbf{F}^{\mathrm{g}}_{\mathrm{uv}}, \mathbf{H}^{\mathrm{g}} = \mathcal{E}_{\mathrm{g}}\left(\mathbf{T}_{\mathbf{x}_0}; \mathbf{y}_{\mathrm{m}}\right). \tag{4}$$

To handle dynamically changing details in the final shape, the geometry encoder $\mathcal{E}_{\mathrm{g}}$ is also conditioned on the motion embedding $\mathbf{y}_{\mathrm{m}}$.

Gaussian Primitives Regression. For the decoding stage, we design three separate decoders (Fig. 2): $\mathcal{D}_{\mathrm{a}}$ for the appearance, $\mathcal{D}_{\mathrm{p}}$ for scales, quaternions and opacities of 3D Gaussians, and $\mathcal{D}_{\mathrm{g}}$ for per-Gaussian offsets. All three share the same convolutional architecture and receive appearance and geometry features $\mathbf{F}^{\mathrm{a}}_{\mathrm{uv}}, \mathbf{F}^{\mathrm{g}}_{\mathrm{uv}}$ as inputs:

$$\mathcal{G}'_{\mathrm{uv}} = \mathcal{D}_{\{\mathrm{a}|\mathrm{p}|\mathrm{g}\}}\left([\mathbf{F}^{\mathrm{a}}_{\mathrm{uv}}, \mathbf{F}^{\mathrm{g}}_{\mathrm{uv}}]; [\mathbf{H}^{\mathrm{a}}, \mathbf{H}^{\mathrm{g}}]\right). \tag{5}$$

Figure 3. **Qualitative Comparison against Dynamic Methods.** We show results for identity generalization. GIGA (**Ours**) achieves significantly higher quality, while also being able to faithfully synthesize a virtual avatar after training on a large-scale dataset [60].

The output map $\mathcal{G}'_{\text{uv}}$ contains RGB color channels $\mathbf{c}$, quaternions $\mathbf{q}_0$, normalized scales $\mathbf{s}'$, opacities α, and offsets $\delta\mathbf{x}_0$, which are defined in the canonical T-pose. To utilize the representational power of the shared texel space and to propagate semantic information from encoders to decoders at different spatial scales, we use UNet-like [43] skip-connections, building a feature propagation bridge from each encoder to each decoder. We stack intermediate feature maps $\mathbf{H}^{\text{a}}$ and $\mathbf{H}^{\text{g}}$ along the feature dimension and feed them to every decoder at each corresponding upsampling layer. To convert a texel-aligned 3D Gaussian map to the observed pose space $\boldsymbol{\theta}$, we use linear blend skinning applied to both offsets $\delta\mathbf{x}_0$ and quaternions $\mathbf{q}_0$: $\delta\mathbf{x} = \text{lbs}\left(\delta\mathbf{x}_0, \boldsymbol{\theta}\right); \mathbf{q} = \text{lbs}\left(\mathbf{q}_0, \boldsymbol{\theta}\right)$. We also multiply normalized Gaussian scales $\mathbf{s}'$ by a hyperparameter ρ: $\mathbf{s} = \rho\mathbf{s}'$, value $\rho = 5e^{-3}$ was empirically found adequate to provide sufficient coverage of the template with Gaussians.

The resulting 3D Gaussian map $\mathcal{G}_{\text{uv}}$ is rendered for each camera view π_k using differentiable Gaussian rasterizer [64], yielding an RGB and accumulated opacity image pair, $I_k, A_k = \mathcal{R}\left(\mathcal{G}_{\text{uv}}, \pi_k\right)$, as in Eq. (1).

3.3. Training

GIGA generates texel-aligned 3D Gaussian texture maps at 512×512 resolution. As some texels in these textures are not associated with any triangle, they are assigned zero opacity during rendering.

Training Objective. GIGA is trained to minimize a combination of the reconstruction $\mathcal{L}_{\text{rec}}$ and geometry-related regularization terms $\mathcal{L}_{\text{reg}}$:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{reg}}\,, \tag{6}$$

Reconstruction Term. Following Kerbl et al. [15], we compute the mean absolute error $\mathcal{L}_{\text{L1}}$ and the structural similarity measure $\mathcal{L}_{\text{ssim}}$ [58] between the rendered image I_k and the ground-truth image $I_{\text{gt},k}$. We additionally evaluate the VGG-based [48] perceptual loss [67] $\mathcal{L}_{\text{LPIPS}}$ on randomly sampled patches with centers within the ground truth segmentation mask $A_{\text{gt},k}$. To improve outlines of Gaussian primitives, we also compute $\mathcal{L}_{\text{mask}}$, the mean squared error between the rendered opacity images A_k and ground truth character segmentation masks $A_{\text{gt},k}$. The final reconstruction term is averaged over all images in the batch and defined as:

$$\mathcal{L}_{\text{rec}} = \lambda_1\mathcal{L}_{\text{L1}} + \lambda_2\mathcal{L}_{\text{ssim}} + \lambda_3\mathcal{L}_{\text{LPIPS}} + \lambda_4\mathcal{L}_{\text{mask}}\,, \tag{7}$$

with $\lambda_1 = \lambda_2 = \lambda_3 = 0.5$ and $\lambda_4 = 0.1$ in all experiments.

Regularization Term. Modeling various dynamically changing geometrical shapes with Gaussian offsets $\delta\mathbf{x}_0$ is particularly challenging during the early stages of optimization. To this end, we introduce additional penalties to con-

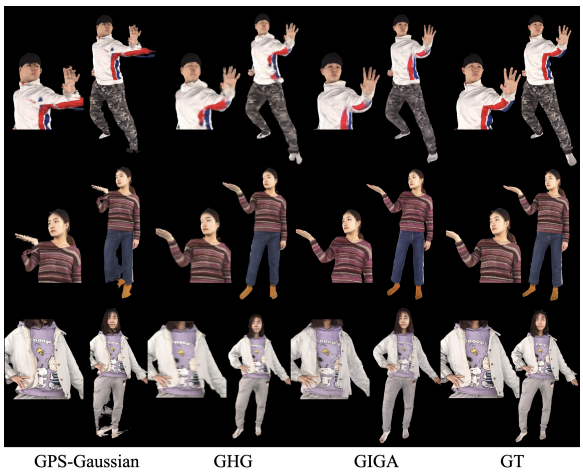


Figure 4. **Qualitative Comparison against 3D Gaussian-based methods.** GIGA achieves higher rendering quality than previous GHG [19] and GPS-Gaussian [68] with fewer artifacts. Results reported on the validation split of THuman2.1 dataset [65].

strain some of the predicted Gaussian parameters:

$$\begin{aligned}\mathcal{L}_{\text{reg}} &= \lambda_5 \mathcal{L}_{\delta \mathbf{x}_0} + \lambda_6 \left(\mathcal{L}_{\delta \mathbf{x}_0;\alpha} + \mathcal{L}_{\mathbf{s}';\alpha}\right) + \lambda_7 \mathcal{L}_{\alpha} = \\ &= \lambda_5 \left\|\delta \mathbf{x}_0\right\|_2 + \lambda_6 \left(\left\|\mathbb{1}_{[\alpha<\epsilon]} \mathbf{x}_0\right\|_2 + \left\|\mathbb{1}_{[\alpha<\varepsilon]} \mathbf{s}'\right\|_2\right) + \\ &+ \lambda_7 \text{Beta}\left(\alpha\right) ,\end{aligned} \tag{8}$$

with $\lambda_5 = 0.15, \lambda_6 = \lambda_7 = 0.1$. The offset penalty $\mathcal{L}_{\delta \mathbf{x}_0}$ prevents Gaussians from drifting far away from the mesh.

Due to UV mapping distortions, some mesh regions may be oversampled with Gaussians. The poor initial model state will encourage some Gaussians to become transparent and inadequately large and move away from the mesh. To reclaim those Gaussians for actual modeling, we additionally penalize offsets of low-opacity primitives with $\mathcal{L}_{\delta \mathbf{x}_0;\alpha}$, where the indicator function $\mathbb{1}_{[\alpha<\epsilon]}$ triggers the extra penalty only when Gaussian opacity α is below a threshold ϵ. Importantly, we disable gradient flow through the indicator function. We similarly penalize scales $\mathbf{s}$ when opacity is below ε with $\mathcal{L}_{\mathbf{s}';\alpha}$. Following Lombardi et al. [25], we encourage all Gaussians to be either completely opaque or completely transparent by computing the negative log-likelihood of the beta distribution $\text{Beta}\left(0.5, 0.5\right)$ as $\mathcal{L}_{\alpha}$.

3.4. Implementation Details

We train GIGA for 250k iterations on a single NVIDIA H100 GPU using 4 subjects per batch. For each subject, we rasterize 4 views to compute image-space losses. All models are optimized with the AdamW optimizer [28], the learning rate is linearly increased from 0 to $1e^{-4}$ during the first 25k training steps. The weight decay parameter is set to $1e^{-4}$.

Dataset	Method	PSNR	SSIM	LPIPS
MVHumanNet	NHP	17.72	0.5678	402.7
	TH	17.46	0.5808	392.1
	GIGA	**22.19**	**0.7526**	**70.2**
THuman2.1	GPS-G	18.22	0.7768	207.9
	GHG	18.86	0.7394	76.5
	GIGA	**20.19**	**0.7944**	**58.7**

Table 1. **Quantitative Comparison.** We evaluate generalization to unseen identities of Neural Human Performer (NHP) and TransHuman (TH) on MVHumanNet; GPS-Gaussian (GPS-G) and Generalizable Human Gaussians (GHG) on THuman2.1.

At the warmup stage, we also apply linear annealing to the predicted offsets $\delta \mathbf{x}_0$ in addition to other penalties, Eq.(8). We empirically find it crucial, as it allows the model to focus on approximate appearance reconstruction before refining person-specific geometry, effectively preventing unstable offset predictions.

4. Results

We first explain the datasets and metrics (Sec. 4.1). Then, we show qualitative results (Sec. 4.2) and comparisons (Sec. 4.3). Lastly, we ablate our design choices (Sec. 4.4). Supplementary materials contain additional qualitative examples.

4.1. Dataset and Metrics

Datasets. To evaluate GIGA, we use the multi-view human performance capture datasets MVHumanNet [60] and DNA-Rendering [3], and also the 3D human scan dataset THuman2.1 [65]. Details on the data usage protocol for training can be found in the supplementary material.

Metrics. To evaluate rendering quality, we compute the following metrics: PSNR, SSIM [58], and the perceptual LPIPS [67] using AlexNet [17] features (scaled by 1000).

4.2. Qualitative Results

Fig. 3, 4, and 5 (and the supplementary video) present GIGA's novel view rendering results of unseen subjects performing unseen poses. Notably, GIGA achieves photorealistic and view-consistent rendering and effectively captures fine details such as clothing wrinkles and intricate textures. We highlight that for novel subjects, GIGA achieves a rendering quality comparable to that of the training subjects, demonstrating its generalization ability to novel identities. Please see the in-the-wild examples from Neuman dataset in the supplementary materials.

4.3. Comparison

Competing Methods. We compare GIGA to other generalizable dynamic image-driven methods discussed in Sec. 2,

	Training set → Testing set	
Metric	DNA → DNA	MVH → DNA
↑ PSNR	19.63	19.49
↑ SSIM	0.7255	0.7138
↓ LPIPS	106.4	114.9

Table 2. **Generalization to Novel Data.** Training on MVHumanNet allows GIGA to generalize to DNA-Rendering with minor quantitaive performance drop in comparison with GIGA trained on DNA-Rendering.

i.e., NHP [18], TransHuman (TH) [33], GPS-Gaussian [68], and GHG [19]. We exclude concurrent human avatar reconstruction methods IDOL [71], SIGMAN [63], and MoGa [73] from comparison, since they focus on reconstructing animatable avatars from single images.

Tab. 1 provides quantitative comparisons of GIGA against state-of-the-art baselines for generalizable human rendering. In all cases, we evaluate the generalization ability to unseen identities from held-out validation sequences. Fig. 3 provides an overview of qualitative results. NHP relies on sparse 3D convolutions to process volumetric features in the observed pose space and suffers from missing input signals due to occlusions, ultimately failing to generalize to unseen identities. Despite operating in a canonical template pose space and tokenizing the template for processing with a transformer-based network, TransHuman also cannot learn meaningful priors from the large data collection. GIGA, on the other hand, utilizes the power of the shared texel space to a maximum degree: all feature representations for digital humans are defined in the same texel space, and intermediate features enhance the quality of the final prediction through skip-connections. MultiHead-UNet is also significantly more computationally efficient, which explains both qualitative and quantitative improvements over baselines.

GPS-Gaussian heavily relies on depth estimation from two adjacent source views to unproject pixel-aligned 3D Gaussians for novel view rendering. As such, it is limited by the quality of the depth estimation, which might fail if source views are too far from each other. Artifacts of erroneous depth-based unprojection are visible in Fig. 4.

GHG targets only static reconstruction from sparse input views; it models humans as a set of 3D Gaussian scaffolds in the observed pose space and cannot be easily extended to dynamic scenarios. GHG handles Gaussian color prediction by pretraining a separate texture inpainting network. GIGA learns to operate with appearance and geometry features simultaneously, yielding results of higher quality, both qualitatively (Fig. 4) and quantitatively (Tab. 1).

Cross-Dataset Validation. To demonstrate the cross-dataset generalization ability, we train two variants of GIGA, where one is trained and tested on the DNA-Rendering [60] dataset, while the other variant is trained on MVHumanNet [3] and tested on the DNA-Rendering dataset. As seen in Tab. 2 and Fig. 5, our method is capable of generalizing across datasets with minimal quality drop, as the model trained on the MVHumanNet dataset performs comparably to the counterpart trained on the DNA-Rendering dataset, which clearly shows that our method effectively learns the prior from large datasets.

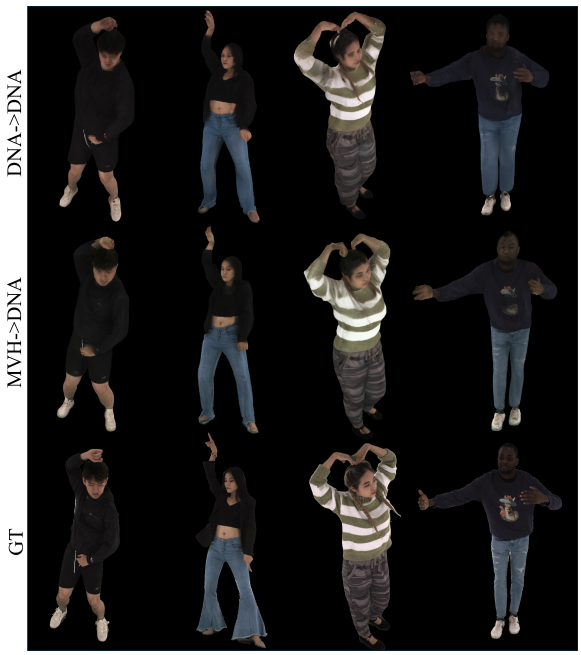
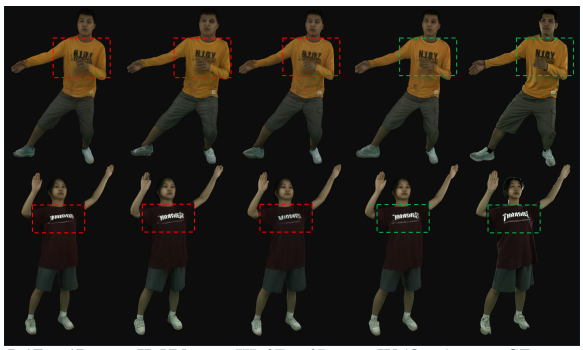


Figure 5. **Cross-Dataset Examples.** After training on MVHumanNet, GIGA successfully generalizes to novel subjects from DNA-Rendering (MVH→DNA) and performs similarly to GIGA, specifically trained on DNA-Rendering (DNA→DNA).

Single/Sparse-View Inputs. To enable a single-input-view regime at test-time, for each character we precompute a static RGB texture from multiple views and concatenate it with $\mathbf{T}_{\mathbf{x}_0}$ from Eq. (2) along channel dimension. This serves as "mean" appearance information to be used for unobserved parts of the template by the network. We retrain GIGA with this additional input and evaluate the effect of the number of views at test-time in Tab. 3. The 4-view setup (front, back, left, and right; **0c**) provides a full coverage of the character, whereas the 2-view setup (**0b**) only uses the front and the back views, resulting in slightly better metric values. In the single view setup (**0a**), we only provide the frontal view, which contains information about the dynamic appearance, while the backside of the character will be reconstructed mainly based on the precomputed static texture - thus leading to reduced quantitative performance. We report qualitative results in the supplementary material.



I, 1Enc-1Dec **II**, UNet **III**, 2Enc-3Dec **IV (Ours)** GT

Figure 6. **Impact of Different GIGA Network Architectures.** While all proposed GIGA configurations produce reasonable results at coarse level, multiple encoding and decoding heads with additional skip-connections (**IV**, MultiHeadUNet) manages fine-grained appearance better.

Configuration	↑ PSNR	↑ SSIM	↓ LPIPS
0a, 1 view	21.32	0.7795	99.3
0b, 2 views	**22.55**	**0.7992**	**85.4**
0c, 4 views	22.28	0.7968	89.4
I, Enc-Dec	21.93	0.7450	75.4
II, UNet	21.94	0.7449	73.4
III, 2Enc-3Dec	21.88	0.7441	77.1
IV (**GIGA**) MultiHead-UNet	**22.19**	**0.7526**	**70.2**
IVa, w/o CrossAttn	21.69	0.7377	76.5
IVb, w/o offset annealing	—-	—-	—-
IVc, w/o offset penalty	21.73	0.7454	71.0
IVd, w/o opacity penalties	21.78	0.7465	71.1

Table 3. **Ablation Results for GIGA.** We vary the number of input views (**0a–0c**), compare architecture backbones (**I–IV**), and ablate motion conditioning and regularization terms (**IVa–IVd**). See detailed analysis in Sec. 4.4.

4.4. Ablation Study

In this section, we discuss our design choices for the architecture and training objective of GIGA. We also examine the effects of different sizes of training datasets and the influence of motion tracking from sparse input views in the supplementary material.

Model Architecture. To benchmark MultiHeadUNet (**IV**) at the core of GIGA, we propose 3 alternative architectures with approximately the same number of trainable parameters ($\simeq$90M): a simple encoder-decoder model (**I**), a conventional UNet with skip-connections between corresponding up- and downsampling blocks of the encoder and decoder (**II**), and a model with 2 encoders $\mathcal{E}_{\mathrm{a}}, \mathcal{E}_{\mathrm{g}}$ and 3 decoders $\mathcal{D}_{\mathrm{a}}, \mathcal{D}_{\mathrm{p}}, \mathcal{D}_{\mathrm{g}}$, but without skip-connections (**III**). We observe, both quantitatively (Tab. 3) and qualitatively (Fig. 6), that the configuration **IV** leads to a higher quality overall, being particularly helpful at preservation of fine appearance details, observed in the input signal.

Motion Conditioning. Configurations **I-IV** use an MLP-based motion encoder $\mathcal{E}_{\mathrm{m}}$ and cross-attention motion conditioning by default. We additionally remove cross attention blocks from **IV**, leaving intermediate features $\mathbf{F}^{\mathrm{a}}_{\mathrm{uv}}, \mathbf{F}^{\mathrm{g}}_{\mathrm{uv}}$ without motion conditioning (**IVa**). As it is evident from the quantitative evaluation (3), enabling motion conditioning via cross-attention improves the visual quality of GIGA's predictions.

Regularizations. Additional geometry penalties also contribute to the quality of GIGA. Removing opacity-related penalties $\mathcal{L}_{\alpha}, \mathcal{L}_{\delta\mathbf{x}_0;\alpha}, \mathcal{L}_{\mathbf{s};\alpha}$ (**IVd**), and then the offset penalty $\mathcal{L}_{\delta\mathbf{x}_0}$ (**IVc**) from the complete training objective (7) leads to quality degradation (Tab. 3), as some of the predicted Gaussians become underconstrained. The model fails to converge without offset annealing during the warmup stage (**IVb**).

5. Discussion and Conclusion

Limitations. GIGA shows scalability to hundreds of multi-view videos, thanks to our efficient representation and efficient architecture, and respective generalization without sacrificing rendering quality. Future work could explore the construction of a latent space for virtual humans to improve temporal consistency. SMPL-X as a body template forms the basis for generalization, but it does not allow for handling non-rigid dynamics (e.g., hair and loose clothing) without additional assumptions or physics-based priors. More advanced human shape prior that includes clothing geometry might alleviate some of these limitations [4, 42, 66]. To mitigate motion tracking issues, an end-to-end optimization of body shape and pose parameters could be a promising next step, which has already proven to be successful for face-only rendering approaches [51].

Conclusion. This work presented GIGA, a generalizable, sparse image-driven 3D Gaussian human model. Trained on a large-scale multi-view dataset, GIGA dynamically synthesizes texel-aligned 3D Gaussians from sparse input videos in a feed-forward manner. Our approach achieves state-of-the-art generalization to unseen identities while preserving person-specific pose-dependent appearance changes thanks to our scalable architecture and efficient representation. We believe our proposed model could benefit future research in this domain and take another step towards enabling more accessible and immersive remote collaboration.

Acknowledgements. This work was funded by the Saarbrücken Research Center for Visual Computing and Artificial Intelligence (VIA).

References

[1] Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabián Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. Driving-signal aware full-body avatars. *ACM Trans. Graph.*, 40(4), 2021. 1

[2] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, Yaser Sheikh, and Jason Saragih. Authentic volumetric avatars from a phone scan. *ACM Trans. Graph.*, 41(4), 2022. 13

[3] Wei Cheng, Ruixiang Chen, Siming Fan, Wanqi Yin, Keyu Chen, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, Daxuan Ren, Lei Yang, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, Bo Dai, and Kwan-Yee Lin. DNA-Rendering: A Diverse Neural Actor Repository for High-Fidelity Human-centric Rendering. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 4, 6, 7

[4] Artur Grigorev, Bernhard Thomaszewski, Michael J. Black, and Otmar Hilliges. HOOD: Hierarchical graphs for generalized modelling of clothing dynamics. 2023. 8

[5] Chen Guo, Junxuan Li, Yash Kant, Yaser Sheikh, Shunsuke Saito, and Chen Cao. Vid2avatar-pro: Authentic avatar from videos in the wild via universal prior. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025. 2, 3

[6] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. *ACM Trans. Graph.*, 40(4), 2021. 1, 2

[7] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[8] 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 (CVPR)*, 2024. 2

[9] 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 (CVPR)*, 2024. 1, 2

[10] Jiancheng Huang, Mingfu Yan, Songyan Chen, Yi Huang, and Shifeng Chen. Magicfight: Personalized martial arts combat video generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024. 2

[11] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. TeCH: Text-Guided Reconstruction of Lifelike Clothed Humans. In *2024 International Conference on 3D Vision (3DV)*, 2024. 1

[12] Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM Transactions on Graphics (TOG)*, 42(4), 2023. 1, 2

[13] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *Proceedings of the European conference on computer vision (ECCV)*, 2022. 2, 15

[14] Yujiao Jiang, Qingmin Liao, Xiaoyu Li, Li Ma, Qi Zhang, Chaopeng Zhang, Zongqing Lu, and Ying Shan. UV Gaussians: Joint Learning of Mesh Deformation and Gaussian Textures for Human Avatar Modeling. *arXiv preprint arXiv:2403.11589*, 2024. 2, 4

[15] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4), 2023. 2, 4, 5

[16] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Commun. ACM*, 60(6), 2017. 6

[18] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 7, 13

[19] Youngjoong Kwon, Baole Fang, Yixing Lu, Haoye Dong, Cheng Zhang, Francisco Vicente Carrasco, Albert Mosella-Montoro, Jianjin Xu, Shingo Takagi, Daeil Kim, Aayush Prakash, and Fernando De la Torre. Generalizable Human Gaussians for Sparse View Synthesis. *European Conference on Computer Vision*, 2024. 2, 3, 6, 7, 13

[20] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular Primitives for High-Performance Differentiable Rendering. *ACM Transactions on Graphics*, 39(6), 2020. 12

[21] John P Lewis, 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*. 2023. 4

[22] 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. 1

[23] Lingjie Liu, Weipeng Xu, Marc Habermann, Michael Zollhöfer, Florian Bernard, Hyeongwoo Kim, Wenping Wang, and Christian Theobalt. Neural Human Video Rendering by Learning Dynamic Textures and Rendering-to-Video Translation. *IEEE Transactions on Visualization and Computer Graphics*, PP, 2020. 2

[24] 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 Trans. Graph.(ACM SIGGRAPH Asia)*, 2021. 3

[25] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: learning dynamic renderable volumes from images. *ACM Trans. Graph.*, 38(4), 2019. 6

[26] 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), 2021. 2

[27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6), 2015. 2

[28] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2017. 6

[29] Marko Mihajlovic, Yan Zhang, Michael J Black, and Siyu Tang. LEAP: Learning articulated occupancy of people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

[30] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*, 2020. 2

[31] Arthur Moreau, Jifei Song, Helisa Dhamo, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Human Gaussian Splatting: Real-time Rendering of Animatable Avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2

[32] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4), 2022. 2

[33] Xiao Pan, Zongxin Yang, Jianxin Ma, Chang Zhou, and Yi Yang. TransHuman: A Transformer-based Human Representation for Generalizable Neural Human Rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 7, 13

[34] 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 (CVPR)*, 2024. 1, 2, 4

[35] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 4

[36] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023. 3

[37] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable Neural Radiance Fields for Modeling Dynamic Human Bodies. In *ICCV*, 2021. 1

[38] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans. In *CVPR*, 2021. 2

[39] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[40] Edoardo Remelli, Timur Bagautdinov, Shunsuke Saito, Chenglei Wu, Tomas Simon, Shih-En Wei, Kaiwen Guo, Zhe Cao, Fabian Prada, Jason Saragih, and Yaser Sheikh. Drivable Volumetric Avatars using Texel-Aligned Features. In *ACM SIGGRAPH 2022 Conference Proceedings*. Association for Computing Machinery, 2022. 1, 2, 3

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 4

[42] Boxiang Rong, Artur Grigorev, Wenbo Wang, Michael J. Black, Bernhard Thomaszewski, Christina Tsalicoglou, and Otmar Hilliges. Gaussian Garments: Reconstructing simulation-ready clothing with photorealistic appearance from multi-view video, 2025. 8

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Cham, 2015. Springer International Publishing. 5

[44] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019. 1, 2

[45] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2

[46] István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3d human pose and shape estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 15

[47] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3

[48] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition, 2014. 5

[49] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *Advances in neural information processing systems*, 34, 2021. 2

[50] Shih-Yang Su, Timur Bagautdinov, and Helge Rhodin. NPC: Neural Point Characters from Video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1

[51] 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 Trans. Graph.*, 43(6), 2024. 8

[52] 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*. Curran Associates, Inc., 2021. 2

[53] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. ARAH: Animatable Volume Rendering of Articulated Human SDFs. In *European Conference on Computer Vision*, 2022. 1, 2

[54] Shaofei Wang, Bozidar Antic, Andreas Geiger, and Siyu Tang. IntrinsicAvatar: Physically Based Inverse Rendering of Dynamic Humans from Monocular Videos via Explicit Ray Tracing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[55] Shaofei Wang, Tomas Simon, Igor Santesteban, Timur Bagautdinov, Junxuan Li, Vasu Agrawal, Fabian Prada, Shoou-I Yu, Pace Nalbone, Matt Gramlich, Roman Lubachersky, Chenglei Wu, Javier Romero, Jason Saragih, Michael Zollhoefer, Andreas Geiger, Siyu Tang, and Shunsuke Saito. Relightable full-body gaussian codec avatars. *arXiv.org*, 2501.14726, 2025. 4

[56] Yiming Wang, Qingzhe Gao, Libin Liu, Lingjie Liu, Christian Theobalt, and Bao Xin Chen. Neural Novel Actor: Learning a Generalized Animatable Neural Representation for Human Actors. *IEEE Transactions on Visualization and Computer Graphics*, 30, 2022. 2

[57] 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*, 2023. 2

[58] Zhou Wang, Alan Bovik, Hamid Sheikh, and Eero Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *Image Processing, IEEE Transactions on*, 13, 2004. 5, 6

[59] Donglai Xiang, Timur Bagautdinov, Tuur Stuyck, Fabian Prada, Javier Romero, Weipeng Xu, Shunsuke Saito, Jingfan Guo, Breannan Smith, Takaaki Shiratori, et al. Dressing avatars: Deep photorealistic appearance for physically simulated clothing. *ACM Transactions on Graphics (TOG)*, 41 (6), 2022. 1

[60] 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. 2, 4, 5, 6, 7, 15

[61] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[62] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[63] Yuhang Yang, Fengqi Liu, Yixing Lu, Qin Zhao, Pingyu Wu, Wei Zhai, Ran Yi, Yang Cao, Lizhuang Ma, Zheng-Jun Zha, et al. SIGMAN: Scaling 3d human gaussian generation with millions of assets. *arXiv preprint arXiv:2504.06982*, 2025. 2, 3, 7

[64] Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. gsplat: An Open-Source Library for Gaussian Splatting. *arXiv preprint arXiv:2409.06765*, 2024. 5

[65] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4D: Real-time Human Volumetric Capture from Very Sparse Consumer RGBD Sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*, 2021. 2, 3, 4, 6

[66] Ilya Zakharkin, Kirill Mazur, Artur Grigorev, and Victor Lempitsky. Point-based modeling of human clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 8

[67] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*, 2018. 5, 6

[68] 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. 3, 6, 7, 13

[69] 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*, 44(1), 2024. 1, 2

[70] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision*. Springer, 2024. 2

[71] Yiyu Zhuang, Jiaxi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. IDOL: Instant photorealistic 3d human creation from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025. 1, 2, 3, 7

[72] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3d gaussian avatars. *arXiv preprint arXiv:2311.08581*, 2023. 2

[73] Dong Zijian, Duan Longteng, Song Jie, J.Black Michael, and Geiger Andreas. MoGA: 3d Generative Avatar Prior for Monocular Gaussian Avatar Reconstruction. In *International Conference on Computer Vision (ICCV)*, 2025. 1, 3, 7