Parametric Gaussian Human Model: Generalizable Prior for Efficient and Realistic Human Avatar Modeling

Cheng Peng[1,*], Jingxiang Sun[1,*], Yushuo Chen[1], Zhaoqi Su[1], Zhuo Su[2], Yebin Liu[1,†]
[1]Tsinghua University [2]ByteDance

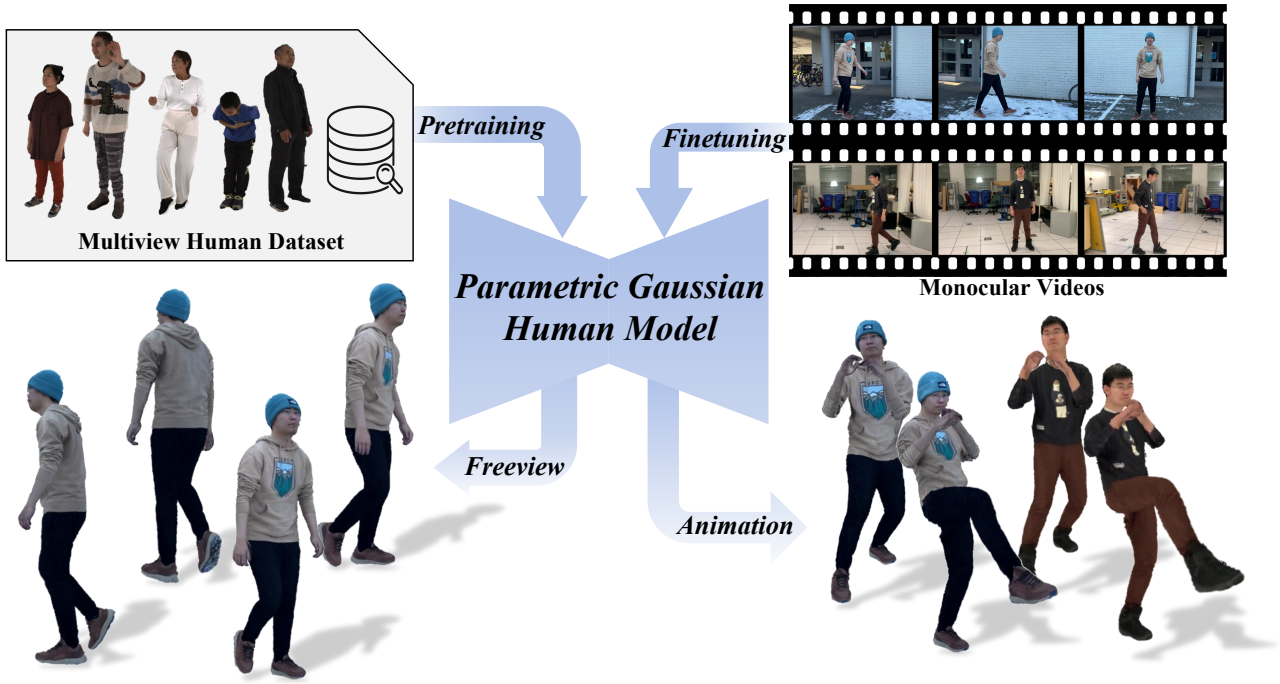

Figure 1. We introduce the Parametric Gaussian Human Model (PGHM), a generalizable prior for efficient and realistic human avatar modeling. After being trained on a large-scale, high-quality multiview human dataset, PGHM can be efficiently fine-tuned using monocular single-person videos. This enables accurate avatar reconstruction and supports both free-viewpoint rendering and animation.

Abstract

Photorealistic and animatable human avatars are a key enabler for virtual/augmented reality, telepresence, and digital entertainment. While recent advances in 3D Gaussian Splatting (3DGS) have greatly improved rendering quality and efficiency, existing methods still face fundamental challenges, including time-consuming per-subject optimization and poor generalization under sparse monocular inputs. In this work, we present the Parametric Gaussian Human Model (PGHM), a generalizable and efficient framework that integrates human priors into 3DGS for fast and high-fidelity avatar reconstruction from monocular videos. PGHM introduces two core components: (1) a UV-aligned latent identity map that compactly encodes subject-specific geometry and appearance into a learnable feature tensor; and (2) a disentangled Multi-Head U-Net that predicts Gaussian attributes by decomposing static, pose-dependent, and view-dependent components via conditioned decoders. This design enables robust rendering quality under challenging poses and viewpoints, while allowing efficient subject adaptation without requiring multi-view capture or long optimization time. Experiments show that PGHM is significantly more efficient than optimization-from-scratch methods, requiring only approximately 20 minutes per subject to produce avatars with comparable visual quality, thereby demonstrating its practical applicability for real-world monocular avatar creation.

1. Introduction

Photorealistic and animatable human avatars represent a crucial research direction in 2D and 3D vision, enabling ap-

plications in virtual/augmented reality, telepresence, digital entertainment, etc.. Traditional mesh-based or point-based human avatars suffer from fundamental limitations - predefined topologies and unstructured representations, respectively - that hinder realistic avatar creation. In recent years, while the emergence of NeRF has significantly advanced human avatar quality, NeRF's intrinsic slow rendering and expensive optimization remain fundamental bottlenecks.

Recently, the explicit 3D Gaussian Splatting representation has emerged as a breakthrough technology that combines accelerated rendering speeds with superior visual quality, thereby significantly benefiting photorealistic human avatar creation. However, current Gaussian avatar approaches still face fundamental challenges. On one hand, human avatar from multi-view video inputs, e.g., Animatable Gaussians [45], achieve high-quality rendering results and pose-dependent dynamics by introducing pose-dependent Gaussian maps, yet require 1-2 days of training per subject. Besides, multi-view inputs require complicated data capture setups. On the other hand, avatars from monocular inputs [24, 49] combine 3DGS with SMPL-UV or SMPL-X geometric models to learn the pose-dependent effects, yet suffer from blurriness and lack appearance details, as monocular inputs tend to struggle with generalizing to diverse or unseen poses due to incomplete observations and inherent limitations in the training data. Therefore, to achieve efficient and high-quality human avatars from monocular inputs, the key lies in developing a generalizable parametric model that can learn human priors from large-scale data while maintaining the representational advantages of 3D Gaussians.

We argue that incorporating parametric human priors into the Gaussian-based human avatar is essential, as it enables rapid subject-specific adaptation through the learned prior, and more robust performance under challenging input conditions. In this paper, we propose Parametric Gaussian Human Model, which learns generalizable avatar priors from large-scale data while enabling fast adaptation to novel subjects. Our Parametric Gaussian Human Model achieves this through two key designs. Firstly, we introduce the UV-Aligned Latent Identity Map, which encodes identity-specific attributes such as facial features and clothing geometry into a compact, learnable feature tensor. This approach differs from the GaussianAvatar's [24] reliance solely on UV position maps for pose information. Our design enhances the original framework by utilizing the UV-Aligned Latent Identity Map as a control signal for identity. During the fine-tuning phase, we can optimize this map alone to effectively capture personalized characteristics. This method contributes to rapid fine-tuning and improved identity control. Secondly, for better learning Gaussian attributes from the Latent Identity Map, we propose the disentangled Multi-Head U-Net that explicitly models static, dynamic, and view-dependent effects through pose/view-conditioned decoders, achieving consistent performance across unseen identities, diverse poses and challenging viewpoints. Together, we achieve robust and high-fidelity human avatars through parametric Gaussian priors learned from large human datasets combined of selected MVHumanNet [81] and DNA-Rendering[13] datasets, enabling fast adaptation for a personalized avatar from a monocular video. Compared with the concurrent work Vid2Avatar-Pro [18], our method achieves comparable optimization time for monocular-input personalized avatars, while Vid2Avatar-Pro requires an additional 36–48 hours for per-identity mesh template reconstruction. Experiments demonstrate that our method achieves SOTA avatar rendering quality and avatar training efficiency. The contributions of our paper are summarized as follows:

- We propose the Parametric Gaussian Human Model, a framework that integrates parametric human priors with 3D Gaussian Splatting, enabling fast and high-fidelity avatar adaptation to novel subjects.
- A UV-aligned latent identity map that compactly encodes subject-specific attributes into a learnable feature tensor, allowing memory-efficient personalization while preserving fine details.
- A disentangled Multi-Head U-Net architecture that dynamically decomposes Gaussian properties into static, pose-dependent, and view-aware components through conditioned decoders, ensuring dynamic details under challenging poses and viewpoints.

2. Related Work

2.1. 3D Human Avatar Modeling

Avatar Modeling from Monocular Video. The advent of neural radiance fields (NeRF) [48] has spurred a wave of *implicit* avatar reconstruction approaches that fit articulated NeRFs to monocular video[8, 16, 30, 32, 68–70, 74, 78]. Given the inherent noise in monocular pose estimation, many of these methods jointly refine motion trajectories during inverse rendering [17, 31, 32, 78, 87]. Although effective, the learned deformation fields often overfit to training sequences, resulting in artifacts under novel or out-of-distribution poses.

Recent advances replace implicit volumetric representations with *explicit* 3D Gaussian Splatting (3DGS), significantly improving rendering speed and simplifying optimization. Some approaches optimize Gaussian parameters per subject [39, 42, 66, 73], while others utilize neural networks to predict Gaussian attributes for more efficient personalization [24, 25, 36, 41, 49, 60, 77**?**]. Despite the efficiency gains, current per-subject fittings often suffer from blurry textures and limited generalization to novel motions. Our method addresses these limitations by pretrain-

ing a 3DGS human prior on a large corpus of dynamic human sequences, then adapting it to short monocular videos for sharper textures and improved motion robustness.

Avatar Modeling from Multi-View Videos. Calibrated multi-camera systems enable high-fidelity avatar reconstruction by jointly modeling geometry and appearance across views [2, 11, 20, 26, 29, 43, 44, 46, 52, 57, 59, 62, 65, 67, 75, 80, 84, 86, 95–97]. Early works often rely on canonical implicit fields to support non-rigid deformation, but they come at a high computational cost. With the introduction of 3DGS [35], recent methods replace volumetric rendering with Gaussian splatting and attach Gaussians to skeletal joints [34, 45, 51, 54, 94, 99]. To further enhance the representation capability for human dynamics and regularize surface reconstruction, 2D-map parameterizations are employed to guide geometry and texture learning [24, 45, 54]. Nevertheless, current designs are either manually engineered or limited to a fixed identity, restricting generalization and scalability.

2.2. Avatar Prior Models

Statistical and Regression Priors. Full-body statistical models such as SMPL and SMPL-X, along with part-specific 3DMMs for facial modeling, offer low-dimensional shape spaces governed by joint angles [33, 47, 56]. However, their minimalist templates lack garment wrinkles and fine-scale facial details. Pixel-aligned regressors—including PIFu, PIFuHD, ICON, ECON, SITH, ARCH/ARCH-H, and MIGS [7, 21, 22, 28, 63, 64, 82, 83]—predict high-resolution surface detail from single images, but they are trained on limited-scale static scan datasets. As a result, the reconstructed avatars are typically non-rigged or exhibit unnatural deformations during animation. Generalizable human rendering methods [9, 10, 37, 38, 53, 71, 90, 91] synthesize novel views from sparse camera inputs but offer limited articulation control. In contrast, part-specific priors such as HeadGap, SEGA, URAvatar, URHand, OHTA, CAFCA, and Lisa [4–6, 12, 14, 19, 40, 50, 92, 93] reconstruct highly detailed and realistically deformable heads or hands, yet they do not generalize to diverse full-body clothing. Our concurrent work [18] introduces a unified clothed-human prior that enables high-fidelity reconstruction across varied identities and garments, with robust generalization to novel poses.

Generative Avatar Priors Recent progress in generative 3D avatar modeling can be broadly categorized into GAN-based and diffusion-based approaches. GAN-based methods have demonstrated strong capabilities in generating 3D-aware avatars from single-view image datasets [1, 3, 15, 23, 72, 85, 89]. GSM [1] proposes a surface-constrained Gaussian representation based on shell maps, which enhances training efficiency while preserving visual quality. On the diffusion side, latent diffusion models (LDMs) have enabled more flexible and semantically grounded 3D avatar generation. StructLDM [27] introduces a structured, 3D-aware LDM trained on a semantically aligned latent space to facilitate controllable generation and editing. Recently, IDOL [98] and LHM [61] employ transformer-based architectures in a feed-forward manner with single image input, enabling fast and scalable 3D human avatar modeling, while limited in capturing dynamic deformation details.

3. Parametric Gaussian Avatar Prior

In this section, we present the Parametric Gaussian Avatar Prior. In contrast to person-specific training tasks, initializing and training Gaussian-based multi-human prior models pose distinct challenges. This section introduces the Gaussian human representation, the carefully designed Gaussian Parametric Human Model, and how to utilize our model for avatar creation when given an input monocular video.

3.1. Preliminaries

3D Gaussian Splatting [35] is a 3D point-based representation for efficient and realistic rendering. It is represented by a set of 3D Gaussians, each of which is parameterized by its 3D center position $\boldsymbol{\mu} \in \mathbb{R}^3$, scale $\mathbf{s} \in \mathbb{R}^3$, rotation parameterized as a quaternion $\mathbf{q} \in \mathbb{R}^4$, color $\mathbf{c} \in \mathbb{R}^3$, and opacity $\sigma \in \mathbb{R}$, and distributed as:

$$f(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right),$$

where the covariance matrix $\Sigma = \mathbf{RSS}^\top\mathbf{R}^\top$ is factorized into a scaling matrix S and a rotation matrix R given by the quaternions q and scaling $\mathbf{s}$.

To enhance the adaptability of Gaussian representations for generating and driving digital humans, inspired by GaussianAvatar[24] and ExAvatar[49], we constrain all Gaussian assets to be isotropic. This is achieved by fixing the scale degree of freedom to 1, while setting the rotation to the identity matrix and the opacity to 1.

3.2. UV-aligned Gaussian Human Representation

To learn a Gaussian attribute representation that incorporates human semantics, we define a UV-position map to represent the posed SMPL-X mesh vertex positions and a UV-Gaussian map to encode the attributes of the generated Gaussian points and their correspondence to the initial SMPL-X mesh, instead of using an orthogonal projection position map as in Animatable Gaussians [45]. The use of UV maps enables the placement of a greater number of effective Gaussian points per unit area, thereby improving model efficiency. Furthermore, by allocating a larger proportion of the UV map to the facial region, our approach captures richer facial details.

UV Position Map. We adjust the SMPL-X human model to a predefined neutral A-pose. Then, the mesh position attributes are unwrapped and mapped to their corresponding positions on a UV map, constructing a UV map representation of the Gaussian digital human. This strategy enables efficient 2D representation for storing 3D position information. Finally, based on the initialized SMPL-X mesh information, we can obtain the UV map $\mathcal{U}_\theta \in \mathbb{R}^{L\times L\times 3}$ corresponding to the digital human in pose θ.

UV Gaussian Map. Building upon the UV Position Map as the initial position map for the Gaussian points, we further predict the pixelwise offsets for position (δ_{position}), scale (δ_{scale}), and color (δ_{color}) through the model. Similar to the UV Position Map, we organize these attributes into corresponding UV Gaussian maps: the position offset map $\mathcal{U}_p \in \mathbb{R}^{L\times L\times 3}$, the scale map $\mathcal{U}_s \in \mathbb{R}^{L\times L\times 1}$, and the color map $\mathcal{U}_c \in \mathbb{R}^{L\times L\times 3}$. In this way, each Gaussian point is aligned with the UV points sampled from the mesh.

3.3. Parametric Prior Model Training

To effectively learn a generalizable Gaussian avatar prior from diverse human subjects, we propose a unified parametric model that encodes large-scale human data into compact latent representations while preserving distinctive individual characteristics. Specifically, we formalize our Gaussian avatar prior as a pipeline that first embeds each subject into a compact latent map, then expands this map through a lightweight decoder into a dense identity tensor, and finally feeds the tensor—together with pose and view cues—into a disentangled Multi-Head U-Net whose static, dynamic, and view branches jointly predict the canonical geometry/texture, pose-driven deformations, and view-dependent color of the UV-aligned Gaussian atlas. In the following, we provide detailed descriptions of each component.

UV-Aligned Latent Identity Map. To parametrize the identity information of different characters (such as appearance, clothing, etc.), we employ a UV-aligned per-identity feature map to encode these attributes. The feature map contains n features $\mathcal{F}' \in \mathbb{R}^{C\times L'\times L'}$, where each feature represents the characteristic information of a distinct character identity. To maintain training efficiency and memory economy, we utilize a Feature Decoder to decode each feature into a tensor $\mathcal{F} \in \mathbb{R}^{C\times L\times L}$, where L is twice the size of L'. This approach ensures that the feature maps remain consistent while containing richer information, which is beneficial for subsequent single-character fine-tuning.

Disentangled Multi-Head U-Net. For a specific character, there is a static geometric offset and static appearance relative to the SMPL-X body model. As the character moves, dynamic changes in geometry and appearance are introduced. Furthermore, varying viewpoints and lighting conditions induce view-dependent color shifts. Modeling such complex requirements therefore calls for a carefully designed architecture. To address this, we propose a Multi-Head U-Net to predict the Gaussian properties of the character. This design enables the network to model static geometric and appearance offsets (*w.r.t.* SMPL-X), dynamic deformations, as well as view-dependent appearance variations in a unified manner.

To accurately model the Gaussian representations, we designed a MultiHead U-Net architecture, which consists of a feature code encoder head $\mathcal{E}_{unet}$ and three decoder heads $\mathcal{D}_{\text{static}}$, $\mathcal{D}_{\text{pose}}$, and $\mathcal{D}_{\text{view}}$. The output from each layer of the encoder is connected in a U-shaped manner to all three decoders. For decoders $\mathcal{D}_{\text{pose}}$ and $\mathcal{D}_{\text{view}}$, we inject pose information and view information, respectively.

For the pose and view injection, we employ two lightweight convolutional encoders $\mathcal{E}_p$ and $\mathcal{E}_v$ to extract feature information. To obtain pose-dependent information, we use $\mathcal{E}_p$ to encode the UV position map, yielding pose-dependent feature $\mathcal{F}_p$:

$$\mathcal{F}_p = \mathcal{E}_p\left(\mathcal{U}_\theta\right)$$

For view-dependent information, we first construct a view direction map $\mathcal{V}$ based on the normal map to model view-dependent variance, similar to NeRF-based approaches [58]. We then use $\mathcal{E}_v$ to extract view features $\mathcal{F}_v$:

$$\mathcal{F}_v = \mathcal{E}_v\left([\mathcal{U}_\theta, \mathcal{V}]\right)$$

To extract the identity information of the subject, we employ the U-Net encoder $\mathcal{E}_u$ to extract appearance features $\mathcal{F}_u$ from the previously mentioned UV-Aligned Latent Identity Map $\mathcal{F}$:

$$\mathcal{F}_u = \mathcal{E}_u\left(\mathcal{F}\right)$$

After obtaining the three types of features described above ($\mathcal{F}_u$, $\mathcal{F}_p$, and $\mathcal{F}_v$), we further utilize the three decoder heads of the U-Net to extract Gaussian representation information.

For the first decoder head $\mathcal{D}_{static}$, we predict the static Gaussian representations, including the Gaussian position UV map ($\mathcal{U}_p$), Gaussian scale UV map ($\mathcal{U}_s$), and Gaussian color UV map ($\mathcal{U}_c$), as follows:

$$\mathcal{U}_p,\, \mathcal{U}_s,\, \mathcal{U}_c = \mathcal{D}_{\text{static}}\left(\mathcal{F}_u\right)$$

For the second decoder head $\mathcal{D}_{\text{pose}}$, we predict dynamic, pose-dependent Gaussian position offset UV map ($\mathcal{U}_p^o$) and Gaussian scale offset UV map ($\mathcal{U}_s^o$), conditioned on both the identity features $\mathcal{F}_u$ and the pose features $\mathcal{F}_p$:

$$\mathcal{U}_p^o,\, \mathcal{U}_s^o = \mathcal{D}_{\text{pose}}\left(\mathcal{F}_u, \mathcal{F}_p\right)$$

For the third decoder head $\mathcal{D}_{\text{view}}$, we predict the Gaussian color offset UV map ($\mathcal{U}_c^o$), which is both pose and view-dependent, by conditioning on the identity features $\mathcal{F}_u$ and the view features $\mathcal{F}_v$:

$$\mathcal{U}_c^o = \mathcal{D}_{\text{view}}\left(\mathcal{F}_u, \mathcal{F}_v\right)$$

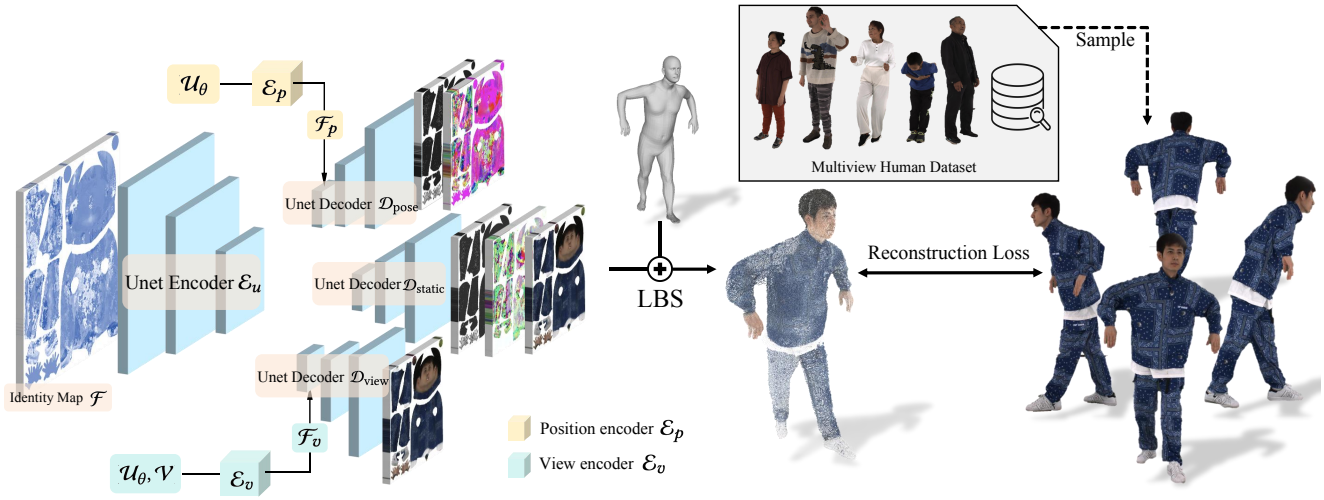

Figure 2. The overall pipeline of the parametric model training involves pre-training our model on a large-scale human dataset to obtain a robust human prior. This process consists of two key components: 1) a UV-aligned identity map to extract the appearance feature information of individuals, and 2) a Disentangled Multi-Head U-Net to decouple pose-dependent and view-dependent Gaussian attributes.

Deformation. After all, we obtain the Gaussian position UV map ($\mathcal{U}_p$), Gaussian scale UV map ($\mathcal{U}_s$), Gaussian color UV map ($\mathcal{U}_c$), Gaussian position offset UV map ($\mathcal{U}_p^o$), Gaussian scale offset UV map ($\mathcal{U}_s^o$), and Gaussian color offset UV map ($\mathcal{U}_c^o$). Due to our UV map design, we can use these Gaussian UV maps to index various Gaussian attributes, including $\mathbf{V}_p$, $\mathbf{V}_s$, $\mathbf{V}_c$, $\mathbf{V}_p^o$, $\mathbf{V}_s^o$, and $\mathbf{V}_c^o$. Based on these attributes, we separately construct pose-independent (static) and pose-dependent (dynamic) Gaussian deformations. For the pose-independent Gaussian deformation and the pose-dependent Gaussian deformation, we compute the canonical space pose-independent Gaussian locations $\mathbf{V}_{\text{sta}}$ and canonical space pose-dependent Gaussian locations $\mathbf{V}_{\text{dyn}}$ as follows:

$$\overline{\mathbf{V}}_{\text{sta}} = \overline{\mathbf{V}} + \mathbf{V}_p$$
$$\overline{\mathbf{V}}_{\text{dyn}} = \overline{\mathbf{V}} + \mathbf{V}_p + \mathbf{V}_p^o$$

The final vertex positions are then computed using Linear Blend Skinning (LBS) as:

$$\mathbf{V}_{\text{sta}} = \text{LBS}\left(\overline{\mathbf{V}}_{\text{sta}}, \theta\right)$$
$$\mathbf{V}_{\text{dyn}} = \text{LBS}\left(\overline{\mathbf{V}}_{\text{dyn}}, \theta\right)$$

where $\overline{\mathbf{V}}$ denotes the canonical vertex positions, $\mathbf{V}_p$ and $\mathbf{V}_p^o$ are the pose-independent and pose-dependent Gaussian position offsets indexed from the corresponding UV maps, θ denotes the pose parameters.

Rasterization. To render the animated 3D geometry, we utilize the 3D Gaussian Splatting (3DGS) rendering pipeline [22], defined as follows:

$$I_s = f\left(\mathbf{V}_{sta}, \mathbf{V_c}, \mathbf{K}, \mathbf{E}\right),$$
$$I_d = f\left(\mathbf{V}_{dyn}, \mathbf{V_c} + \mathbf{V_c}^o, \mathbf{K}, \mathbf{E}\right),$$

where f denotes the 3DGS rendering function, while $\mathbf{K}$ and $\mathbf{E}$ are the camera intrinsic and extrinsic matrices, respectively.

As previously mentioned, all Gaussian primitives are constrained to be isotropic to enhance generalization capability. Therefore, both the rotation and the opacity of each Gaussian are fixed to the identity and unity, respectively, and are omitted from the above equations for clarity.

Training objectives. During the training process of the parametric model, we jointly optimize the per-identity feature map, Feature Encoder, and Multihead-Unet modules. Additionally, since the pose annotations in the training data are not entirely accurate, we further optimize the human poses in the training data during the training process. To ensure the reliability of both pose-independent and pose-dependent information, we employ both types of rendered results as supervision. The final training objective is defined as follows:

$$\begin{aligned}\mathcal{L} = \lambda_1 &\left[\, \mathcal{L}_1\left(I_s, I_{gt}\right) + \mathcal{L}_1\left(I_d, I_{gt}\right) \right] \\ &+ \lambda_2 \left[\, \mathcal{L}_{\text{ssim}}\left(I_s, I_{gt}\right) + \mathcal{L}_{\text{ssim}}\left(I_d, I_{gt}\right) \right] \\ &+ \lambda_3 \left[\, \mathcal{L}_{\text{lpips}}\left(I_s, I_{gt}\right) + \mathcal{L}_{\text{lpips}}\left(I_d, I_{gt}\right) \right] \\ &+ \lambda_4 \, \mathcal{L}_{\text{reg}}\end{aligned}$$

In our loss function, the regularization term $\mathcal{L}_{\text{reg}}$ is designed to prevent unreasonable scale and position values. Specifically, we apply an (L_2) penalty to both the scale offset and position offset parameters. This encourages the optimized scale and position to remain close to their initial values, thus avoiding implausible transformations during training.

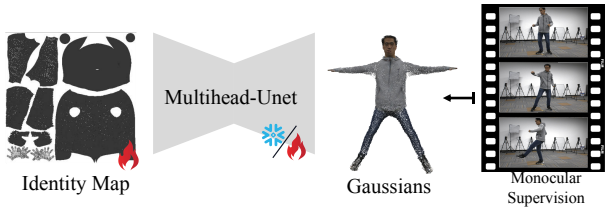

Figure 3. The personalization stage consists of two steps: 1) Identity Map adaptation and 2) Joint tuning of the Identity Map and Multi-Head U-Net.

4. Personalization with Parametric Prior

As shown in Fig 3, we adopt a simple two-stage strategy to personalize our parametric avatar prior to a specific subject from a monocular video.

Identity Map adaptation. As discussed in Section 3.3, we usually expand discrete identity map $\mathcal{F}$ from a smaller feature $\mathcal{F}'$ using a feature encoder. During personalization, we skip the encoder and instead initialize a new learnable feature tensor to directly replace the expanded map. This approach makes the tuning process more efficient and helps retain subject-specific details that might otherwise be lost through the encoder.

Joint tuning of Identity Map and Multi-Head U-Net. Because monocular videos provide only limited supervision, adapting the feature code alone isn't enough. To better capture subject-specific appearance and geometry, we further fine-tune both the feature code and the Multi-Head U-Net together. This joint optimization allows the model to better generalize to novel inputs while keeping the number of trainable parameters small.

5. Experiment

5.1. Dataset

We trained our prior model using the MVHumanNet [81] and DNA-Rendering [13] datasets, both of which are large-scale multi-view human datasets. Given the limited availability of accurate pose annotations in DNA-Rendering, we selectively fit poses for 300 samples from parts 3 to 6 of this dataset. To further increase the diversity of identities, we additionally sampled 300 identities from MVHumanNet, resulting in a mixed training set. It is worth noting that the DNA-Rendering dataset exhibits superior temporal continuity, whereas MVHumanNet consists of frames sparsely sampled from original videos, leading to lower temporal consistency.

For our evaluation, we conducted both qualitative and quantitative experiments on the NeuMan [32] and THuman4.0 [95] datasets. The NeuMan dataset consists of monocular video sequences captured in natural environments, featuring human subjects walking through various settings. We followed the methodologies outlined in [32] for splitting the training and test sets, with pose initialization based on [49]. In contrast, the THuman4.0 dataset is a high-resolution multi-view collection (1330 × 1150 pixels) known for its richly textured and dynamically detailed subjects. Here, we selected 500 frames as the training set and the subsequent 50 frames for testing, utilizing the dataset's original poses for initialization. Our analyses of these two datasets thoroughly assess the effectiveness of our learned priors.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	time
iHuman [55]	25.09	0.961	0.0277	15 min
HumanNeRF [79]	27.06	0.967	0.0252	72 hour
InstantAvatar [31]	28.74	0.972	0.0277	5 min
GaussianAvatar [24]	28.90	0.969	0.0242	6 hour
NeuMan [32]	29.32	0.972	0.0201	7 day
ExAvatar [49]	31.42	0.981	0.0190	5 hour
Vid2AvatarPro [18]	31.18	0.982	**0.0150**	48 hour
Ours	**31.85**	**0.987**	0.0171	20 min

Table 1. **Quantitative comparisons on Neuman test dataset.** We outperform other methods and achieve a significant improvement in efficiency compared to other Gaussian-based approaches.

5.2. Experiment Settings

During the prior training phase, our model was trained on a combined dataset consisting of MVHumanNet and DNA-Rendering samples. We utilized a total batch size of 32, distributed across eight Nvidia V100 GPUs, for a total of 50k iterations. This extensive training procedure lasted approximately seven days. The loss function was composed of several terms with their respective weights: L1 with $\lambda_1 = 0.8$, SSIM with $\lambda_2 = 0.2$, LPIPS with $\lambda_3 = 0.2$, and regularization set to $\lambda_4 = 0.1$.

For the reconstruction process, we typically execute 1k iterations dedicated to feature map alignment, followed by another 1k iterations for model fine-tuning on each sequence from the NeuMan dataset. This entire procedure requires around 20 minutes per sequence when run on a single Nvidia V100 GPU.

5.3. Comparison

We mainly compared our method with GaussianAvatar [24] and ExAvatar [49], both of which are 3DGS-based approaches for constructing human avatars. To quantitatively evaluate performance, we employed established image quality metrics, including Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [76], and Learned Perceptual Image Patch Similarity (LPIPS) [88]. All metrics were computed over the entire image, with backgrounds set to white.

We present the quantitative evaluation results on two benchmark datasets. As summarized in Tab 1 and Tab **??** (Tab **??** see Supplementary Material), our proposed method consistently surpasses all existing baselines across a comprehensive set of metrics, demonstrating superior capability in recovering fine-grained dynamic appearance details and more plausible cloth movement.

Figure 4. Qualitative comparisons on NeuMan test dataset. Our method outperforms other approaches in both color appearance and details.

Qualitative comparisons, illustrated in Figures 4 and **??** (Fig **??** see Supplementary Material), demonstrate the advantages of our approach over baseline methods, producing reconstructions with significantly enhanced detail fidelity and accurate pose-dependent dynamic texture information. On the NeuMan dataset, ExAvatar maintains a reasonable level of detail but struggles on the THuman4.0 dataset, where its performance is limited by complex regularization constraints that oversmooth results and blur fine details. In contrast, the GaussianAvatar method fails to effectively disentangle lighting and motion, resulting in unrealistic lighting artifacts. Our approach addresses these limitations through two key innovations: the introduction of learned priors to reduce reliance on heavy regularization, thus preserving intricate details, and the implementation of a multi-head U-Net architecture that effectively disentangles pose and illumination information, leading to more accurate and expressive reconstructions.

5.4. Ablation Study

To further validate the effectiveness of our proposed priors and the multi-head U-Net architecture, we conducted a series of ablation experiments. These experiments are designed to systematically analyze the contributions of each component to the overall performance of our framework.

(a) (b) (c) (d) (e)

Figure 5. Ablation on ID number effect on feature map tuning stage: (a)-(d) 0, 6, 60 and 600 IDs pretrained (e) Jointly finetune both feature map and unet

Effectiveness of the Learned Priors Figure 5 presents a comprehensive analysis of the impact of our learned priors. Subfigures (a)–(d) correspond to models fine-tuned on identity maps pre-trained with increasing numbers of identities: specifically, 0, 6, 60, and 600 IDs, respectively. Subfigure (e) shows the results after jointly fine-tuning both the identity map and the multi-head U-Net.

The results clearly demonstrate that scaling up the number of identities used during prior training significantly enhances the representational capacity of the identity map. As the number of pre-trained identities increases, the identity map exhibits greater generalization ability across unseen individuals and achieves faster convergence during fine-tuning. Moreover, even when only the identity map is fine-tuned, the reconstructed geometry and texture details become noticeably richer and more plausible, highlighting the benefit of our prior learning strategy.

However, due to the inherent limitations in the diversity and size of available training data, further fine-tuning of both the identity map and the multi-head U-Net is necessary to achieve higher fidelity in both texture and geometry. Importantly, we observe that our two-stage fine-tuning approach dramatically improves training efficiency: compared to training the model from scratch, our method achieves superior results using an order of magnitude fewer optimization steps (e.g., 2,000 steps versus 20,000 steps).

(a) (b) (c) (d) (e) (f)

Figure 6. Ablation on multihead-unet: (a) (d) Ground Truth; (b) (e)Single-head U-Net Result; (c) (f) Multi-head U-Net Result.

Effectiveness of the Multi-Head U-Net Architecture Figure 6 evaluates the efficacy of our multi-head U-Net design in disentangling pose and illumination. In each set of results, subfigures (a) and (d) depict the ground truth images; (b) and (e) show the single-head U-Net results; while (c) and (f) present our multi-head U-Net model results.

The comparison demonstrates that our multi-head U-Net significantly enhances both the realism and accuracy of synthesized outputs. Specifically, disentangling appearance and lighting yields more plausible illumination in (c), while finer dynamic details such as wrinkles and folds in clothing are better preserved in (f), closely resembling the ground truth. In contrast, the single-head U-Net results (b, e) are noticeably darker and lack critical dynamic texture details, highlighting its limitations in modeling lighting and fine appearance. These results underscore the superiority of our multi-head U-Net architecture in generating detailed, physically consistent reconstructions.

Overall, the ablation studies demonstrate that both the learned priors and the multi-head U-Net architecture play critical roles in the success of our approach. The learned priors enhance identity generalization and accelerate convergence, while the multi-head U-Net enables effective disentanglement of pose and lighting, leading to more realistic and detailed human reconstructions.

6. Conculsion

We present *Parametric Gaussian Human Model (PGHM)*, a novel framework that integrates parametric human priors into 3D Gaussian Splatting for efficient and high-fidelity monocular human avatar reconstruction. By introducing a UV-aligned latent identity map and a disentangled Multi-Head U-Net, PGHM enables fast subject-specific adaptation and robust rendering under challenging poses and viewpoints. Compared to existing methods, our approach achieves efficient training with only $\sim$20 minutes per subject, while maintaining competitive visual quality. We believe our method paves the way for scalable, efficient human avatar generation in immersive applications.

Our method currently presents two main limitations. First, it depends on input video sequences for optimizing the identity map, which restricts its use in cases with limited subject data. Future work will explore end-to-end feed-forward architectures that can infer identity features directly from fewer images, reducing per-subject optimization requirements.Second, the approach is less effective for subjects wearing loose or flowing clothing, like skirts or robes, due to challenges in modeling large non-rigid deformations. Addressing these issues will make our system more robust and widely applicable.

Acknowledgment The work is supported by NationalKey R&D Program of China (2022YFF0902200) and National Natural Science Foundation of China (NSFC) under Grant Number 62125107, 62402274.

References

[1] Rameen Abdal, Wang Yifan, Zifan Shi, Yinghao Xu, Ryan Po, Zhengfei Kuang, Qifeng Chen, Dit-Yan Yeung, and Gordon Wetzstein. Gaussian shell maps for efficient 3d human generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9441–9451, 2024. 3

[2] 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 Transactions on Graphics (TOG)*, 40(4):1–17, 2021. 3

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

[4] Marcel C. Buehler, Gengyan Li, Erroll Wood, Leonhard Helminger, Xu Chen, Tanmay Shah, Daoye Wang, Stephan Garbin, Sergio Orts-Escolano, Otmar Hilliges, Dmitry Lagun, Jérémy Riviere, Paulo Gotardo, Thabo Beeler, Abhimitra Meka, and Kripasindhu Sarkar. Cafca: High-quality novel view synthesis of expressive faces from casual few-shot captures. In *ACM SIGGRAPH Asia 2024 Conference Paper*, 2024. 3

[5] Marcel C Bühler, Kripasindhu Sarkar, Tanmay Shah, Gengyan Li, Daoye Wang, Leonhard Helminger, Sergio Orts-Escolano, Dmitry Lagun, Otmar Hilliges, Thabo Beeler, et al. Preface: A data-driven volumetric prior for few-shot ultra high-resolution face synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3402–3413, 2023.

[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, Yaser Sheikh, and Jason Saragih. Authentic volumetric avatars from a phone scan. *ACM Trans. Graph.*, 41(4), 2022. 3

[7] Aggelina Chatziagapi, Grigorios G. Chrysos, and Dimitris Samaras. Migs: Multi-identity gaussian splatting via tensor decomposition. In *ECCV*, 2024. 3

[8] Jianchuan Chen, Ying Zhang, Di Kang, Xuefei Zhe, Linchao Bao, Xu Jia, and Huchuan Lu. Animatable neural radiance fields from monocular rgb videos. *arXiv preprint arXiv:2106.13629*, 2021. 2

[9] Jianchuan Chen, Wentao Yi, Liqian Ma, Xu Jia, and Huchuan Lu. Gm-nerf: Learning generalizable model-based neural radiance fields from multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20648–20658, 2023. 3

[10] Mingfei Chen, Jianfeng Zhang, Xiangyu Xu, Lijuan Liu, Yujun Cai, Jiashi Feng, and Shuicheng Yan. Geometry-guided progressive nerf for generalizable and efficient neural human rendering. In *ECCV*, 2022. 3

[11] Yushuo Chen, Zerong Zheng, Zhe Li, Chao Xu, and Yebin Liu. Meshavatar: Learning high-quality triangular human avatars from multi-view videos. In *European Conference on Computer Vision*, pages 250–269. Springer, 2024. 3

[12] Zhaoxi Chen, Gyeongsik Moon, Kaiwen Guo, Chen Cao, Stanislav Pidhorskyi, Tomas Simon, Rohan Joshi, Yuan Dong, Yichen Xu, Bernardo Pires, He Wen, Lucas Evans, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, Shoou-I Yu, Javier Romero, Michael Zollhöfer, Yaser Sheikh, Ziwei Liu, and Shunsuke Saito. URhand: Universal relightable hands. In *CVPR*, 2024. 3

[13] Wei Cheng, Ruixiang Chen, Siming Fan, Wanqi Yin, Keyu Chen, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, et al. Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19982–19993, 2023. 2, 6

[14] Enric Corona, Tomas Hodan, Minh Vo, Francesc Moreno-Noguer, Chris Sweeney, Richard Newcombe, and Lingni Ma. Lisa: Learning implicit shape and appearance of hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20533–20543, 2022. 3

[15] Zijian Dong, Xu Chen, Jinlong Yang, Michael J Black, Otmar Hilliges, and Andreas Geiger. Ag3d: Learning to generate 3d avatars from 2d image collections. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14916–14927, 2023. 3

[16] 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 Conference Papers*, 2022. 2

[17] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[18] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2, 3, 6

[19] Chen Guo, Zhuo Su, Jian Wang, Shuang Li, Xu Chang, Zhaohu Li, Yang Zhao, Guidong Wang, and Ruqi Huang. Sega: Drivable 3d gaussian head avatar from a single image, 2025. 3

[20] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. *ACM Transactions on Graphics*, 40(4), 2021. 3

[21] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11046–11056, 2021. 3

[22] Hsuan-I Ho, Jie Song, and Otmar Hilliges. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[23] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. Eva3d: Compositional 3d human generation from

2d image collections. *arXiv preprint arXiv:2210.04888*, 2022. 3

[24] 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*, pages 634–644, 2024. 2, 3, 6

[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 (CVPR)*, pages 20418–20431, 2024. 2

[26] Tao Hu, Tao Yu, Zerong Zheng, He Zhang, Yebin Liu, and Matthias Zwicker. Hvtr: Hybrid volumetric-textural rendering for human avatars. In *2022 International Conference on 3D Vision (3DV)*, 2022. 3

[27] Tao Hu, Fangzhou Hong, and Ziwei Liu. Structldm: Structured latent diffusion for 3d human generation. In *European Conference on Computer Vision*, pages 363–381. Springer, 2025. 3

[28] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2020. 3

[29] Zhang Jiakai, Liu Xinhang, Ye Xinyi, Zhao Fuqiang, Zhang Yanshun, Wu Minye, Zhang Yingliang, Xu Lan, and Yu Jingyi. Editable free-viewpoint video using a layered neural representation. In *ACM SIGGRAPH*, 2021. 3

[30] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[31] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 6

[32] 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, 6

[33] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[34] HyunJun Jung, Nikolas Brasch, Jifei Song, Eduardo Perez-Pellitero, Yiren Zhou, Zhihao Li, Nassir Navab, and Benjamin Busam. Deformable 3d gaussian splatting for animatable human avatars, 2023. 3

[35] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 2023. 3

[36] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. HUGS: Human gaussian splatting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[37] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. In *Advances in Neural Information Processing Systems*, 2021. 3

[38] 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. In *European Conference on Computer Vision*, 2024. 3

[39] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19876–19887, 2024. 2

[40] Junxuan Li, Chen Cao, Gabriel Schwartz, Rawal Khirodkar, Christian Richardt, Tomas Simon, Yaser Sheikh, and Shunsuke Saito. Uravatar: Universal relightable gaussian codec avatars. In *ACM SIGGRAPH 2024 Conference Papers*, 2024. 3

[41] Mingwei Li, Jiachen Tao, Zongxin Yang, and Yi Yang. Human101: Training 100+fps human gaussians in 100s from 1 view, 2023. 2

[42] Mengtian Li, Shengxiang Yao, Zhifeng Xie, and Keyu Chen. Gaussianbody: Clothed human reconstruction via 3d gaussian splatting, 2024. 2

[43] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jurgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[44] Zhe Li, Zerong Zheng, Yuxiao Liu, Boyao Zhou, and Yebin Liu. Posevocab: Learning joint-structured pose embeddings for human avatar modeling. *ACM SIGGRAPH Conference Proceedings*, 2023. 3

[45] 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 (CVPR)*, 2024. 2, 3

[46] 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

[47] 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):1–16, 2015. 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. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[49] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3D gaussian avatar. In *ECCV*, 2024. 2, 3, 6

[50] Gyeongsik Moon, Weipeng Xu, Rohan Joshi, Chenglei Wu, and Takaaki Shiratori. Authentic hand avatar from a phone scan via universal hand model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2029–2038, 2024. 3

[51] Arthur Moreau, Jifei Song, Helisa Dhamo, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Human gaussian splatting: Real-time rendering of animatable avatars. In *CVPR*, 2024. 3

[52] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *International Conference on Computer Vision*, 2021. 3

[53] Panwang Pan, Zhuo Su, Chenguo Lin, Zhen Fan, Yongjie Zhang, Zeming Li, Tingting Shen, Yadong Mu, and Yebin Liu. Humansplat: Generalizable single-image human gaussian splatting with structure priors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[54] 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)*, pages 1165–1175, 2024. 3

[55] Pramish Paudel, Anubhav Khanal, Danda Pani Paudel, Jyoti Tandukar, and Ajad Chhatkuli. ihuman: Instant animatable digital humans from monocular videos. In *European Conference on Computer Vision*, pages 304–323. Springer, 2024. 6

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

[57] 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. 3

[58] Sida Peng, Shangzhan Zhang, Zhen Xu, Chen Geng, Boyi Jiang, Hujun Bao, and Xiaowei Zhou. Animatable neural implicit surfaces for creating avatars from videos. In *ICCV*, 2021. 4

[59] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 3

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

[61] Lingteng Qiu, Xiaodong Gu, Peihao Li, Qi Zuo, Weichao Shen, Junfei Zhang, Kejie Qiu, Weihao Yuan, Guanying Chen, Zilong Dong, et al. Lhm: Large animatable human reconstruction model from a single image in seconds. *arXiv preprint arXiv:2503.10625*, 2025. 3

[62] 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. 3

[63] 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*, pages 2304–2314, 2019. 3

[64] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 3

[65] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable gaussian codec avatars. In *CVPR*, 2024. 3

[66] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1606–1616, 2024. 2

[67] Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Zarate, Julien Valentin, Jie Song, and Otmar Hilliges. X-avatar: Expressive human avatars. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[68] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In *NeurIPS*, 2021. 2

[69] Shih-Yang Su, Timur Bagautdinov, and Helge Rhodin. Danbo: Disentangled articulated neural body representations via graph neural networks. In *European Conference on Computer Vision*, 2022.

[70] Shih-Yang Su, Timur Bagautdinov, and Helge Rhodin. Npc: Neural point characters from video. In *ICCV*, 2023. 2

[71] Guoxing Sun, Rishabh Dabral, Pascal Fua, Christian Theobalt, and Marc Habermann. Metacap: Meta-learning priors from multi-view imagery for sparse-view human performance capture and rendering. In *ECCV*, 2024. 3

[72] Jingxiang Sun, Xuan Wang, Lizhen Wang, Xiaoyu Li, Yong Zhang, Hongwen Zhang, and Yebin Liu. Next3d: Generative neural texture rasterization for 3d-aware head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[73] David Svitov, Pietro Morerio, Lourdes Agapito, and Alessio Del Bue. Haha: Highly articulated gaussian human avatars with textured mesh prior, 2024. 2

[74] Gusi Te, Xiu Li, Xiao Li, Jinglu Wang, Wei Hu, and Yan Lu. Neural capture of animatable 3d human from monocular video. In *ECCV*, 2022. 2

[75] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdfs. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[76] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004. 6

[77] Jing Wen, Xiaoming Zhao, Zhongzheng Ren, Alexander G Schwing, and Shenlong Wang. Gomavatar: Efficient animatable human modeling from monocular video using gaussians-on-mesh. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2059–2069, 2024. 2

[78] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *CVPR*, 2022. 2

[79] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, 2022. 6

[80] Donglai Xiang, Fabian Prada, Timur Bagautdinov, Weipeng Xu, Yuan Dong, He Wen, Jessica Hodgins, and Chenglei Wu. Modeling clothing as a separate layer for an animatable human avatar. *ACM Trans. Graph.*, 40(6), 2021. 3

[81] 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 *CVPR*, 2024. 2, 6

[82] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13296–13306, 2022. 3

[83] 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. 3

[84] Tianhan Xu, Yasuhiro Fujita, and Eiichi Matsumoto. Surface-aligned neural radiance fields for controllable 3d human synthesis. In *CVPR*, 2022. 3

[85] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Jiashi Feng, and Mike Zheng Shou. Xagen: 3d expressive human avatars generation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[86] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[87] Zhengming Yu, Wei Cheng, Xian Liu, Wayne Wu, and Kwan-Yee Lin. Monohuman: Animatable human neural field from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16943–16953, 2023. 2

[88] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 6

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

[90] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Efficiently generated human radiance field from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7743–7753, 2022. 3

[91] 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 (CVPR)*, 2024. 3

[92] Xiaozheng Zheng, Chao Wen, Zhaohu Li, Weiyi Zhang, Zhuo Su, Xu Chang, Yang Zhao, Zheng Lv, Xiaoyuan Zhang, Yongjie Zhang, Guidong Wang, and Xu Lan. Headgap: Few-shot 3d head avatar via generalizable gaussian priors. *arXiv preprint arXiv:2408.06019*, 2024. 3

[93] Xiaozheng Zheng, Chao Wen, Su Zhuo, Zeran Xu, Zhaohu Li, Yang Zhao, and Zhou Xue. Ohta: One-shot hand avatar via data-driven implicit priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[94] Yang Zheng, Qingqing Zhao, Guandao Yang, Wang Yifan, Donglai Xiang, Florian Dubost, Dmitry Lagun, Thabo Beeler, Federico Tombari, Leonidas Guibas, and Gordon Wetzstein. Physavatar: Learning the physics of dressed 3d avatars from visual observations. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[95] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 6

[96] Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. Avatarrex: Real-time expressive full-body avatars. *ACM Transactions on Graphics (TOG)*, 42(4): 1–19, 2023.

[97] 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 Trans. Graph.*, 2024. 3

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

[99] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3d gaussian avatars, 2023. 3