

# GauHuman: Articulated Gaussian Splatting from Monocular Human Videos

Shoukang Hu      Ziwei Liu  
S-Lab, Nanyang Technological University  
{shoukang.hu, ziwei.liu}@ntu.edu.sg

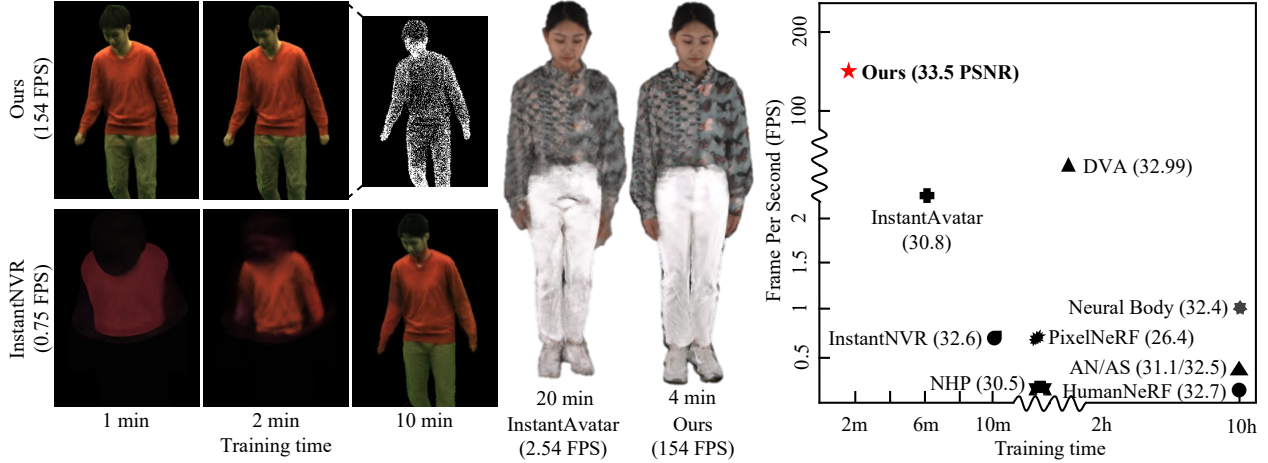


Figure 1. GauHuman models 3D humans with fast training (1~2 minutes) and real-time rendering (~189 FPS) given monocular videos. For illustration purposes, we select one sequence (1024p) from MonoCap [36, 37, 79] and one sequence (1024p) from DNA-Rendering [15].

## Abstract

We present, **GauHuman**, a 3D human model with Gaussian Splatting for both **fast training** (1~2 minutes) and **real-time rendering** (up to 189 FPS), compared with existing NeRF-based implicit representation modelling frameworks demanding hours of training and seconds of rendering per frame. Specifically, **GauHuman** encodes Gaussian Splatting in the canonical space and transforms 3D Gaussians from canonical space to posed space with linear blend skinning (LBS), in which effective pose and LBS refinement modules are designed to learn fine details of 3D humans under negligible computational cost. Moreover, to enable fast optimization of **GauHuman**, we initialize and prune 3D Gaussians with **3D human prior**, while splitting/cloning via **KL divergence** guidance, along with a novel **merge** operation for further speeding up. Extensive experiments on ZJU\_Mocap and MonoCap datasets demonstrate that **GauHuman** achieves state-of-the-art performance quantitatively and qualitatively with fast training and real-time rendering speed. Notably, without sacrificing rendering quality, **GauHuman** can fast model the 3D human performer with ~13k 3D Gaussians. Our code is available

at <https://github.com/skhul01/GauHuman>.

## 1. Introduction

Creating high-quality 3D human performers holds extensive utility across various domains, including AR/VR, telepresence, video games, and movie production. Recent methods show 3D human avatar can be learned from sparse-view videos [12, 45, 47, 75, 78, 79, 100, 111, 114, 119] or even a single image [8, 41, 61, 115] with NeRF-based implicit representation [72]. However, these methods typically require expensive time and computational costs for both training and rendering, which hinders their applications in real-world scenarios. For example, it normally takes about 10 GPU hours to learn a 3D human performer with a rendering speed of less than 1 frame per second (FPS).

To speed up 3D human modelling, generalizable Human NeRF methods [17, 28, 40, 55, 132] trade the quality for less training time. They usually fine-tune new human performers with models pre-trained on articulated human data. However, such an inefficient *pretrain-and-finetune* paradigm would normally take several hours of pre-training to obtain generalizable representations for 3D humans and an additional

one hour of finetuning for each human, while achieving low rendering quality due to the incomplete information within limited conditional input images. On the contrary, a recent line of research focuses on learning efficient 3D human representations [30, 46] to improve training speed. For example, instead of learning neural radiance field (NeRF) from coordinate-based neural networks, methods like [30, 46] incorporate 3D human modelling with multi-resolution hash encoding (MHE) [73] to accelerate the training. Although these methods can reduce the training time by a magnitude via human prior like part-based human representations, their rendering speed still limits their applications. A critical bottleneck hindering their training and rendering speed arises from the inefficiency of the backward mapping-based ray-cast rendering algorithm employed in these methods, especially at higher resolutions, *e.g.*, 1080p and 4k.

To mitigate this time-consuming dilemma, recent research [52, 130, 134] in point-based rendering can fast render both static and dynamic scenes [68, 118, 124, 125] by splatting a restricted set of 3D points. Motivated by these great successes, it is desirable to explore Gaussian Splatting for real-time rendering of articulated objects such as 3D humans with monocular videos. Yet, this promising avenue comes with two primary barriers: 1) How to properly integrate articulated human information from monocular videos in the Gaussian Splatting framework? 2) How to efficiently optimize Gaussian Splatting for fast training? In particular, one potential solution for 3D human modelling with Gaussian Splatting is to directly utilize LBS transformation to transform 3D Gaussians from canonical space into posed space. However, it is infeasible to directly apply a 3D vertex point-based LBS transformation for 3D Gaussians, with inaccurate LBS weights and pose information derived from human prior models such as SMPL [67]. Meanwhile, current Gaussian Splatting methods utilize either sparse point cloud produced from Structure-from-Motion (SfM) [95, 98] or random point cloud to initialize 3D Gaussians. This initialization is specifically designed for static scenes, rather than 3D human structures, making it unsuitable for effective and fast human modelling. Also, adaptive control of 3D Gaussians proposed in [52] mainly utilizes gradient information of 3D position to split and clone Gaussians for over-reconstruction and under-reconstruction. Despite improving the rendering performance, it trivially generates a large number of 3D Gaussians, which causes difficulty in optimization and consumes a large storage memory (up to 734 MB in a static scene reconstruction). More importantly, the Gaussian Splatting framework contains plenty of redundant closely-positioned Gaussians, which holds a potential for additional acceleration.

To tackle these two challenges, we propose **GauHuman**, a 3D human model with articulated Gaussian Splatting representation for both fast training and rendering. Our modelling

framework and fast optimization algorithm are detailed as follows. **Articulated Gaussian Splatting with integrated human information.** Inspired by previous Human NeRF methods, we encode Gaussian Splatting in a canonical space and wrap it to a posed space. Thanks to the Gaussian property, we can transform 3D Gaussians from a canonical space to a posed space via linear blend skinning (LBS). We rotate and translate the 3D position and covariance of each 3D Gaussian with the estimated LBS transformation matrix. However, directly applying multi-layer perceptron (MLP) to estimate LBS weights would be time-consuming and generate low-quality rendering results. To address this, we start with the LBS weights from SMPL [67] and employ an MLP to predict corresponding LBS weight offsets, followed by an effective pose refinement module to rectify SMPL poses for more accurate LBS transformation. For rendering an image with 1024p, our GauHuman only needs to splat a limited set (*e.g.*, 13k) of 3D Gaussians, in comparison with Human NeRF methods that apply volume rendering to millions of sampled points in the whole 3D space. This significantly reduces the computational burden in both training and rendering processes. **Fast Optimization of 3D Gaussian Splatting.** 1) *3D Gaussian Initialization.* In GauHuman, we utilize the human prior like SMPL for 3D Gaussian initialization, which enables fast optimization of articulated Gaussian Splatting as the initialization encodes human structures. 2) *Split/Clone/Merge.* Our analysis indicates that the splitting and cloning process overlooks a crucial metric, namely, the distance between 3D Gaussians, during the split and clone process. In light of this, we integrate the Kullback–Leibler (KL) divergence of 3D Gaussians to regulate the split and clone process. A novel merge operator is also proposed to merge 3D Gaussians that are close to each other for further speeding up. 3) *Prune.* As 3D human has specific structures (human prior), we prune 3D Gaussians that are far away from the human body, *e.g.*, SMPL vertices. With the above algorithm implemented in articulated 3D Gaussian Splatting, we can fast model 3D human performers with  $\sim 13k$  3D Gaussians (3.5MB memory storage) within one or two minutes. Our main contributions are as follows:

1. We present, GauHuman, a 3D model with articulated Gaussian Splatting for both fast training (1 $\sim$ 2 minutes) and real-time rendering (up to 189 FPS), in which pose and LBS refinement modules are designed to learn fine details of 3D humans.
2. To enable fast optimization of articulated 3D Gaussian Splatting, we initialize and prune 3D Gaussians with 3D human prior, while splitting/cloning via KL divergence measures, accompanied by a novel merge operation for further speeding up.
3. Experiments on two monocular data sets validate that our GauHuman achieves state-of-the-art novel view synthesis performance with fast training and real-time rendering.

## 2. Related Work

**Rendering and Radiance Fields.** Early novel view synthesis approaches focus on using light fields for densely sampled [32, 58] and unstructured [7] capture. In later years, the advent of Structure-from-Motion (SfM) [98] enables a new domain by synthesizing novel views from a collection of photos with blending techniques [9, 23, 31, 38, 53, 104]. With the development of deep learning techniques, previous research [24, 38, 87, 105, 133] propose to adopt deep learning techniques with multi-view stereo (MVS) for novel view synthesis. Recently, deep learning techniques have been used to learn implicit neural radiance fields [39, 72, 80, 97] with volume rendering [51, 69]. NeRF [72] is one of the representative work, learning high-quality novel view synthesis results, which has inspired a series of follow-up works [4, 10, 16, 19, 21, 43, 44, 50, 65, 73, 74, 83, 107, 109, 110, 120, 122, 128, 129]. To accelerate training and rendering speed, several methods [11, 25, 29, 60, 63, 73, 82, 102, 103, 127] are proposed, including replacing multi-layer perceptrons (MLPs) with more efficient representations or skipping the empty space via the occupancy grid. In contrast to NeRF, which applies volume rendering to millions of sampled points in the whole 3D space when rendering an image, point-based rendering [2, 33, 34, 54, 56, 88, 89, 130] splats a restricted set of points. This approach provides flexibility and efficiency in reconstructing 3D scenes by optimizing the opacity and position of individual points. Furthermore, point primitives can be represented as circular, elliptic discs, ellipsoids or surfels for high quality results [6, 81, 85, 117, 126, 135]. To achieve real-time rendering of large-scale scenes, 3D Gaussians with splatting techniques are used to represent static scenes [52], dynamic scenes [68, 118, 124, 125], and other tasks [86, 99, 108]. In this work, we further extend Gaussian Splatting to articulated 3D humans with both fast training and real-time rendering.

**3D Human Modelling.** Although 2D human modelling [1, 26, 27, 48, 49, 59, 70, 94] has made substantial progress, it is still challenging to reconstruct high-quality 3D humans for synthesizing free-viewpoint videos. Traditional methods either leverage multi-view stereo methods [35, 96] or depth fusion [18, 22, 101] to reconstruct 3D human geometry, which typically requires dense camera arrays or depth sensors. To alleviate this, previous research [3, 17, 28, 40, 55, 90, 91, 132] learn human priors from data sets containing either 3D ground truth human models or multi-view image collections, which enables them to synthesize novel views even with a single image. However, due to the limited diversity in pre-training data sets, these methods can not generalize well to 3D humans with complex poses. Another line of works reconstruct 3D human performers with implicit neural representations from sparse-view videos [12, 45, 47, 62, 64, 75, 78, 79, 79, 100, 111, 114, 119]

or even a single image [8, 41, 61, 115]. However, these methods usually require hours of training time and seconds of rendering time per frame, restricting their applications to real-world scenarios. To improve the training efficiency, recent research [30, 46] apply multi-hashing encoding (MHE) to reduce the training time to several minutes, while their rendering remains comparatively slow. Moreover, neural volumetric primitives [14, 66, 84] have been proposed for efficient rendering of 3D humans, but they still require hours of training time. In this work, we focus on reconstructing high-quality 3D humans from monocular videos with both fast training and real-time rendering. Our concurrent work [121] learns real-time rendering of dynamic 3D humans from dense-view videos.

## 3. Our Approach

The goal of this paper is to learn 3D human performers with fast training and rendering from monocular videos. We assume the calibrated camera parameters and the human region masks are known. We also assume the corresponding SMPL parameters  $\beta, \theta$  are given.

### 3.1. Preliminary

**SMPL** [67] is a parametric human model  $M(\beta, \theta)$ , which defines  $\beta, \theta$  to control body shapes and poses. In this work, we apply the Linear Blend Skinning (LBS) algorithm used in SMPL to transform points from a canonical space to a posed space. For example, 3D point  $p^c$  in the canonical space is transformed to a posed space defined by pose  $\theta$  as  $p^t = \sum_{k=1}^K w_k (G_k(J, \theta)p^c + b_k(J, \theta, \beta))$ , where  $J$  includes  $K$  joint locations,  $G_k(J, \theta)$  and  $b_k(J, \theta, \beta)$  are the transformation matrix and translation vector of joint  $k$  respectively,  $w_k$  is the linear blend weight.

**3D Gaussian Splatting** [52] In comparison with NeRF methods that learn continuous structured volumetric representation for the whole 3D space, Gaussian Splatting represents a 3D scene with a set of unstructured discrete 3D Gaussians. Each 3D Gaussian is defined with a point center  $p$  and a full 3D covariance matrix  $\Sigma$  in a world space, i.e.,

$$G(x) = \frac{1}{(2\pi)^{\frac{3}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-p)^T \Sigma^{-1}(x-p)}. \quad (1)$$

To ensure the positive semi-definiteness of 3D covariance matrix  $\Sigma$ ,  $\Sigma$  is decomposed into two learnable components, i.e., a quaternion  $r \in \mathcal{R}^4$  representing rotation and a vector  $s \in \mathcal{R}^3$  representing scaling. Through the transformation of quaternion  $r$  into a rotation matrix  $R$  and scaling vector  $s$  into a scaling matrix  $S$ , the associated covariance matrix  $\Sigma$  is defined as  $\Sigma = RSS^T R^T$ .

To splat 3D Gaussians, 3D Gaussian Splatting [134, 135] first transforms them from world space to ray coordinate and then projects them to a 2D screen plane. As shown in [134], utilizing a world-to-camera transformation matrix  $W$ , trans-

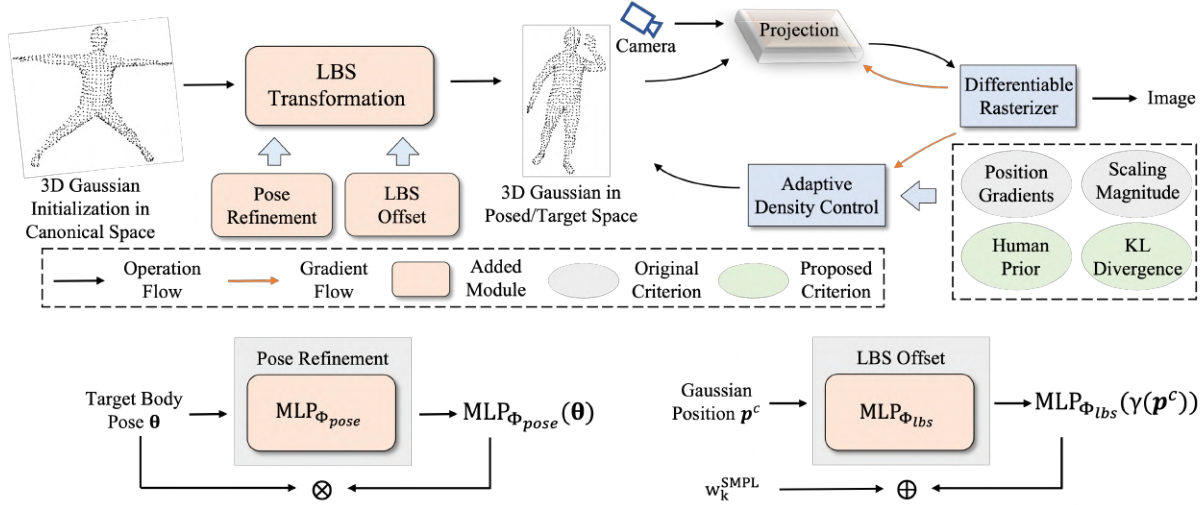


Figure 2. **GauHuman Framework.** We first initialize 3D Gaussians’ positions  $p^c$  from SMPL vertex points. Then we incorporate pose refinement module  $MLP_{\Phi_{pose}}(\cdot)$  and LBS weight field  $MLP_{\Phi_{lbs}}(\cdot)$  module to learn the LBS transformation to transform 3D Gaussians from canonical space to posed space. During optimization, a tile-based differentiable rasterizer is applied to enable fast rendering. To adaptively control the number of 3D Gaussians, we further propose to use human prior (e.g., SMPL) and KL divergence measure to regulate the split, clone, merge, and prune process.

lation vector  $d$ , and the Jacobian matrix  $J$  representing an approximated projective transformation from camera coordinate to ray coordinate, 3D Gaussians in ray coordinate assume the following form:

$$G(x) = \frac{1}{(2\pi)^{\frac{3}{2}} |\Sigma'|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-p')^T \Sigma'^{-1} (x-p')} \quad (2)$$

where  $p' = m(Wp + d)$  and  $\Sigma' = JW\Sigma W^T J^T$  are the mean vector and covariance matrix of 3D Gaussians in ray coordinate,  $m(\cdot)$  denotes the mapping from camera coordinate to ray coordinate. The projected 2D Gaussians in the 2D screen plane can then be obtained by integrating Eqn. (2) along the projection direction or utilizing a property mentioned by [134].

After mapping 3D Gaussians to a 2D screen space, we count 2D Gaussians that overlap with each pixel and calculate their color  $c_i$  and density  $\alpha_i$  contribution. Finally, the colors of each pixel can be obtained by blending  $N$  ordered Gaussians:

$$\hat{C} = \sum_{i \in N} c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j). \quad (3)$$

During optimization, 3D Gaussians are first initialized from either Structure-from-Motion (SfM) or random sampling. Then adaptive control of Gaussians is employed to improve the rendering quality, which mainly includes three operations, i.e., split, clone, and prune. The split and clone operations are performed on 3D Gaussians with large position gradients. These 3D Gaussians would be 1) split into smaller Gaussians if the magnitude of the scaling matrix is larger than a threshold or 2) cloned if the magnitude of the scaling matrix is smaller than a threshold. While the split-

ting and cloning process augments the number of Gaussians, the prune operation selectively eliminates Gaussians with either excessively small opacity ( $\alpha$ ) or overly large scaling magnitudes.

### 3.2. Articulated 3D Gaussian Splatting

Recent research can fast render static scenes [52] or dynamic scenes [68, 118, 124, 125] by splatting a restricted set of 3D points. Motivated by these great successes, it is desirable to explore Gaussian Splatting for real-time rendering of articulated objects such as 3D humans with monocular videos. When employing Gaussian Splatting to articulated 3D humans, there are two main challenges, i.e., 1) How to properly integrate articulated human information from monocular videos in the Gaussian Splatting framework? ii). how to efficiently optimize Gaussian Splatting for fast training?

Inspired by previous Human NeRF research, as shown in Fig. 2, we represent 3D humans with 3D Gaussians in a canonical space and then wrap 3D Gaussians from the canonical space to posed space through forward LBS transformation. Thanks to the Gaussian property, we can transform 3D Gaussians from a canonical space to a posed space via linear blend skinning (LBS). Specifically, we rotate and translate the 3D position and covariance of each 3D Gaussian with the estimated LBS transformation matrix.

$$\begin{aligned} p^t &= G(J^t, \theta^t) p^c + b(J^t, \theta^t, \beta^t) \\ \Sigma^t &= G(J^t, \theta^t) \Sigma^c G(J^t, \theta^t)^T, \end{aligned} \quad (4)$$

where  $p^t$ ,  $\Sigma^t$ ,  $p^c$ , and  $\Sigma^c$  are the 3D position vector and covariance matrix in the canonical and posed space respectively.  $G(J^t, \theta^t) = \sum_{k=1}^K w_k G_k(J^t, \theta^t)$ ,  $b(J^t, \theta^t, \beta^t) =$



$\sum_{k=1}^K w_k \mathbf{b}_k(\mathbf{J}^t, \boldsymbol{\theta}^t, \beta^t)$  are the rotation matrix and translation vector.  $K$  is the joint number,  $\mathbf{G}_k(\mathbf{J}^t, \boldsymbol{\theta}^t)$  and  $\mathbf{b}_k(\mathbf{J}^t, \boldsymbol{\theta}^t, \beta^t)$  are the transformation matrix and translation vector of joint  $k$  respectively,  $w_k$  is the linear blend skinning (LBS) weight we need to estimate.

**LBS Weight Field** Multi-layer perceptrons (MLPs) can be used to predict the LBS weight coefficients  $w_k$  for each 3D Gaussian in the canonical space. However, it would be time-consuming and generate low-quality rendering results. To address this, we start with the LBS weights from SMPL and employ an  $\text{MLP}_{\Phi_{\text{lbs}}}(\cdot)$  to predict corresponding LBS weight offsets. Specifically, for each 3D Gaussian, we add the LBS weight  $w_k^{\text{SMPL}}$  of nearest SMPL vertex with the predicted offsets  $\text{MLP}_{\Phi_{\text{lbs}}}(\gamma(\mathbf{p}^c))$ , *i.e.*,

$$w_k = \frac{e^{\log(w_k^{\text{SMPL}} + 10^{-8}) + \text{MLP}_{\Phi_{\text{lbs}}}(\gamma(\mathbf{p}^c))[k]}}{\sum_{k=1}^K e^{\log(w_k^{\text{SMPL}} + 10^{-8}) + \text{MLP}_{\Phi_{\text{lbs}}}(\gamma(\mathbf{p}^c))[k]}}, \quad (5)$$

where we apply the standard positional encoding  $\gamma(\cdot)$  to the input (3D position  $\mathbf{p}^c$ ) of  $\text{MLP}_{\Phi_{\text{lbs}}}(\cdot)$ . The idea of learning LBS weight field is also applied in previous human modelling works [5, 13, 20, 42, 71, 77, 92, 106, 113, 123].

**Pose Refinement** The body poses  $\boldsymbol{\theta}$  estimated from images is often inaccurate, which degrades the rendering quality. To address this, we incorporate a pose correction module  $\text{MLP}_{\Phi_{\text{pose}}}(\cdot)$  to learn relative updates to joint angles:

$$\boldsymbol{\theta} = \boldsymbol{\theta}^{\text{SMPL}} \otimes \text{MLP}_{\Phi_{\text{pose}}}(\boldsymbol{\theta}), \quad (6)$$

where  $\boldsymbol{\theta}^{\text{SMPL}}$  is the SMPL body pose parameters estimated from images. The body pose refinement is also investigated in [12, 46, 114].

During inference, thanks to the 3D Gaussian-based representation defined in a canonical space, the learned LBS weight for each 3D Gaussian can be derived from the LBS weight field module in advance, which saves computation time. Meanwhile, the rectified pose parameters can also be saved from the pose refinement module at the end of training. Considering the above, our GauHuman can model high-quality 3D humans with high rendering speed.

### 3.3. Fast Optimization of GauHuman

The current optimization scheme of Gaussian Splatting [52] is specifically designed for static scenes. To facilitate fast optimization for articulated 3D humans, we incorporate 3D human prior and KL divergence measures to speed up.

**3D Gaussian Initialization** Current Gaussian Splatting methods utilize either Structure-from-Motion (SfM) [95, 98] or random point cloud to initialize 3D Gaussians, which overlooks 3D human structures. In GauHuman, we initialize 3D Gaussians with human prior like SMPL vertex points. The above initialization encodes human structure, thus accelerating optimization.

**Split & Clone & Merge & Prune** As mentioned in Sec. 3.1, the optimization of Gaussian Splatting relies on the adaptive

control of 3D Gaussians to ensure superior rendering quality. To adaptively control the number of Gaussians, three operations, *i.e.*, split, clone, and prune, are performed based on the position gradients and magnitude of the scaling matrix. However, although split and clone operations substantially improve the performance, they trivially generate a substantial quantity of 3D Gaussians. This results in a slowdown of the optimization and a significant consumption of storage memory, reaching up to 734 MB in static scenes. Our analysis indicates that previous research overlooks a crucial metric, namely, the distance between 3D Gaussians, during the split and clone process. For example, if two Gaussians are close to each other, even if the position gradients are larger than a threshold, they should not be split or cloned as **1) these Gaussians are updating their positions to improve their performance, 2) splitting and cloning these Gaussians have negligible influence on the final performance as they are too close to each other.** To regulate the split and clone process, we propose to use **Kullback–Leibler (KL)** divergence as a measure of the distance of 3D Gaussians, *i.e.*,

$$KL(G(\mathbf{x}_0)|G(\mathbf{x}_1)) = \frac{1}{2}(\text{tr}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_0) + \ln \frac{\det \boldsymbol{\Sigma}_1}{\det \boldsymbol{\Sigma}_2}) + (\mathbf{p}_1 - \mathbf{p}_0)^T \boldsymbol{\Sigma}_1^{-1}(\mathbf{p}_1 - \mathbf{p}_0) - 3), \quad (7)$$

where  $\mathbf{p}_0$ ,  $\boldsymbol{\Sigma}_0$ ,  $\mathbf{p}_1$ ,  $\boldsymbol{\Sigma}_1$  are the position and covariance matrix of two 3D Gaussians  $G(\mathbf{x}_0)$  and  $G(\mathbf{x}_1)$ .

After obtaining the KL divergence of nearby 3D Gaussian pairs, we select 3D Gaussians with large KL divergence and positional gradients to perform the split and clone. However, the above requires calculating the KL divergence for each pair of 3D Gaussians, which consumes lots of time during the optimization. To solve this problem, we further adopt two approaches to simplify the whole calculation. 1) For each 3D Gaussian, we first identify its closest 3D Gaussian through the widely adopted k-nearest neighbor (k-NN) algorithm, which assesses the distance between their respective centers. Then we calculate a KL divergence for each pair of nearby 3D Gaussians. The aforementioned simplification effectively reduces the time complexity from  $O(U^2)$  to  $O(U)$ , where  $U$  represents the total number of 3D Gaussians. 2) As the covariance matrix is decomposed into the product of rotation and scaling matrices, we further simplify the computation of matrix inverse and determinant operations in Eqn. (7) with the diagonal and orthogonal property of rotation and scaling matrices. Details are shown in the appendix.

In addition to split and clone operations, we further propose a novel merge operation, which merges redundant 3D Gaussians with 1) large position gradients, 2) small scaling magnitude, and 3) KL divergence less than 0.1. Specifically, we merge two Gaussians by averaging their positions, opacity, and SH coefficients. Additionally, we scale the covariance (scaling) matrix of the first Gaussian by a factor of 1.25 as the initialization for the new Gaussian. For prune

Table 1. Quantitative comparison of our GauHuman and baseline methods on the ZJU\_MoCap and MonoCap data sets. We use bold and underlined text to denote the best and second-best results of each metric respectively. Our GauHuman achieves the SOTA PSNR and LPIPS with fast training (1~2 minutes) and real-time rendering (up to 189 FPS) in both data sets. Note that PixelNeRF [128] and NHP [55] are pre-trained for an additional 10 hours. <sup>†</sup> denotes the fine-tuning time for generalizable methods. LPIPS\* = 1000 × LPIPS. Frames per second (FPS) is measured on an RTX 3090. **For a fair comparison, we do not conduct test-time optimization of SMPL parameters with images from the test set on InstantAvatar [46].**

Method	ZJU_MoCap					MonoCap				
	PSNR↑	SSIM↑	LPIPS*↓	Train	FPS	PSNR↑	SSIM↑	LPIPS*↓	Train	FPS
PixelNeRF [128]	24.71	0.892	121.86	1h <sup>†</sup>	1.20	26.43	0.960	43.98	1h <sup>†</sup>	0.75
NHP [55]	28.25	0.955	64.77	1h <sup>†</sup>	0.15	30.51	0.980	27.14	1h <sup>†</sup>	0.05
NB [78]	29.03	0.964	42.47	10h	1.48	32.36	0.986	16.70	10h	0.98
AN [77]	29.77	0.965	46.89	10h	1.11	31.07	0.985	19.47	10h	0.31
AS [79]	30.38	<b>0.975</b>	37.23	10h	0.40	32.48	<b>0.988</b>	<b>13.18</b>	10h	0.29
HumanNeRF [114]	30.66	0.969	<u>33.38</u>	10h	0.30	<u>32.68</u>	<u>0.987</u>	15.52	10h	0.08
DVA [84]	29.45	0.956	37.74	1.5h	<u>16.5</u>	<u>32.99</u>	0.983	15.83	1.5h	<u>10.5</u>
InstantNVR [30]	<u>31.01</u>	<u>0.971</u>	38.45	5m	2.20	32.61	<b>0.988</b>	16.68	10m	0.75
InstantAvatar [46]	29.73	0.938	68.41	<u>3m</u>	<u>4.15</u>	30.79	0.964	39.75	<u>6m</u>	<u>2.54</u>
<b>GauHuman(Ours)</b>	<b>31.34</b>	0.965	<b>30.51</b>	<b>1m</b>	<b>189</b>	<b>33.45</b>	0.985	<u>13.35</u>	<b>2m</b>	<b>154</b>

operation, we incorporate 3D human prior like SMPL to prune 3D Gaussians far away from SMPL vertex points.

During optimization, four loss functions are used, *i.e.*,

$$\mathcal{L} = \mathcal{L}_{color} + \lambda_1 \mathcal{L}_{mask} + \lambda_2 \mathcal{L}_{SSIM} + \lambda_3 \mathcal{L}_{LPIPS}, \quad (8)$$

where  $\lambda$ 's are loss weights. Empirically, we set  $\lambda_1 = 0.5$ ,  $\lambda_2 = \lambda_3 = 0.01$  to ensure the same magnitude for each loss. Details are shown in the appendix.

## 4. Experiments

### 4.1. Experimental Setup

**Datasets.** Experiments are conducted on two widely used human modelling data sets, *i.e.*, ZJU\_MoCap [78] and MonoCap [36, 37, 79]. For ZJU\_MoCap, same as previous work [30, 114], we select 6 human subjects (377, 386, 387, 392, 393, 394) to conduct experiments. We also adopt the same training and testing setting as [30], *i.e.*, one camera is used for training, while the remaining cameras are used for evaluation. For each subject, we sample 1 frame every 5 frames and collect 100 frames for training. The foreground masks, camera, and SMPL parameters provided by the data set are used for evaluation purposes. MonoCap data set contains four multi-view videos collected by [79] from the DeepCap data set [36] and the DynaCap data set [37]. Similar to ZJU\_MoCap, it provides human masks, cameras, and SMPL parameters. We adopt the same camera views as [79] for training and testing. For each subject, we collect 100 frames for training by sampling 1 frame every 5 frames. Implementation details and evaluation metrics are shown in the appendix.

**Comparison Methods.** We compare our GauHuman with two categories of state-of-the-art human modelling methods,

*i.e.*, subject-specific optimization based methods [30, 46, 77–79, 84, 114] and generalizable methods [55, 128]. Following [30], we report the average metric values of all selected human subjects. For a fair comparison, we do not conduct test-time optimization of SMPL parameters with images from the test set on InstantAvatar [46]. More details are shown in the appendix.

### 4.2. Quantitative Results

Tab. 1 summarizes the overall novel view synthesis results of our GauHuman, NB [78], AN [77], AS [79], HumanNeRF [114], DVA [84], InstantNVR [30], InstantAvatar [46], PixelNeRF [128], and NHP [55]. Thanks to our proposed articulated 3D human Gaussian-based representation, we can fast optimize it for around 1 or 2 minutes to achieve photorealistic rendering results, while baseline methods (*i.e.*, NB, AN, AS, and HumanNeRF) take around 10 hours to finish the training and generalizable methods takes about 1 hour for fine-tuning after 10 hours' pre-training. To accelerate the training of 3D human modelling, InstantNVR [30] and InstantAvatar [46] reduce the training to several minutes with multi-hashing encoding representation. However, their rendering speed still restricts their applications in real-world applications. When compared with these two state-of-the-art methods for fast training, our GauHuman achieves slightly better performance with 3 times faster training and 45 times faster rendering speed. Among baseline methods, DVA achieves a fast rendering speed (up to 16.5 FPS) with efficient mixtures of volumetric primitives. However, in contrast to our GauHuman, DVA takes about 90x more time to converge and 11.45x more time to render one image.








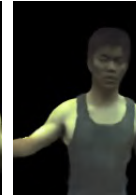
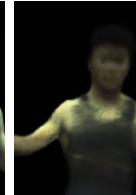


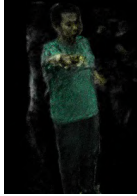














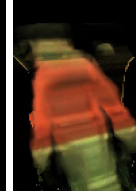









GT	Ours	InstantAvatar	InstantNVR	DVA	HumanNeRF	AS	NB	NHP
								
								
								
								
Training:	<b>1 min</b>	3 min	5 min	1.5 h	10 h	10 h	10 h	1h fine-tuning
Rendering:	<b>189 FPS</b>	4.15 FPS	2.20 FPS	16.5 FPS	0.30 FPS	0.40 FPS	1.48 FPS	0.15 FPS

Figure 3. Novel view synthesis results produced by our GauHuman and baseline methods on ZJU\_MoCap and MonoCap. The bottom lines show the training time and rendering speed of each method on ZJU\_Mocap. Zoom in for the best view.

### 4.3. Qualitative Results

As shown in Fig. 3, our GauHuman shows better appearance and geometry details than DVA [84], NB [78] and NHP [55]. Although NB [78] and NHP [55] show impressive results with 4-view videos, they do not perform well with monocular videos. NB [78] encodes human structure information with latent codes positioned in SMPL vertex’s locations, which fails to capture complex human motions. NHP [55] does not generalize well in monocular videos due to the information missing from input images. Thanks to the LBS weight field and deformation field learned in HumanNeRF [114] and AS [79], they achieve comparable visualization results as our GauHuman for 600x more time for training and 472x more time for rendering. InstantNVR [30] and InstantAvatar [46] reduce the training time to several minutes with a multi-hashing encoding representation, but they produce slightly worse visualization results than ours and require 45x more rendering time. DVA [] improves the rendering speed to 16.5 FPS with efficient mixtures of volumetric primitives. However, it fails to produce high-quality results and produces obvious artifacts on garments and faces, due to incomplete information within conditional input images and a lack of

mechanisms to adaptively control the number of voxels.

### 4.4. Ablation Study

We perform ablation studies with the sequence of 386 from the ZJU\_Mocap data set. To highlight the effectiveness of each proposed component, we project the 3D human bounding box to each camera plane to obtain the bounding box mask and report these metrics based on the masked area.

**Pose Refinement and LBS Weight Field.** Tab. 2 shows the results of ablating pose refinement and LBS weight field modules. All models are trained in around 1 minute. As shown in Tab. 2, pose refinement and LBS weight field modules significantly improve all metrics with negligible computational cost. The visualization results in Fig. 4 show that these two modules can help erase artifacts and learn fine details like cloth wrinkles. For example, we observe obvious artifacts in hands and fewer cloth wrinkles without using either one of these two modules.

**3D Gaussian Initialization.** In Gaussian Splatting, previous research [52] indicates that initialization plays an important role in the final results. For static scenes, there are two commonly adopted initialization approaches, *i.e.*, random initialization or initialization with sparse points from

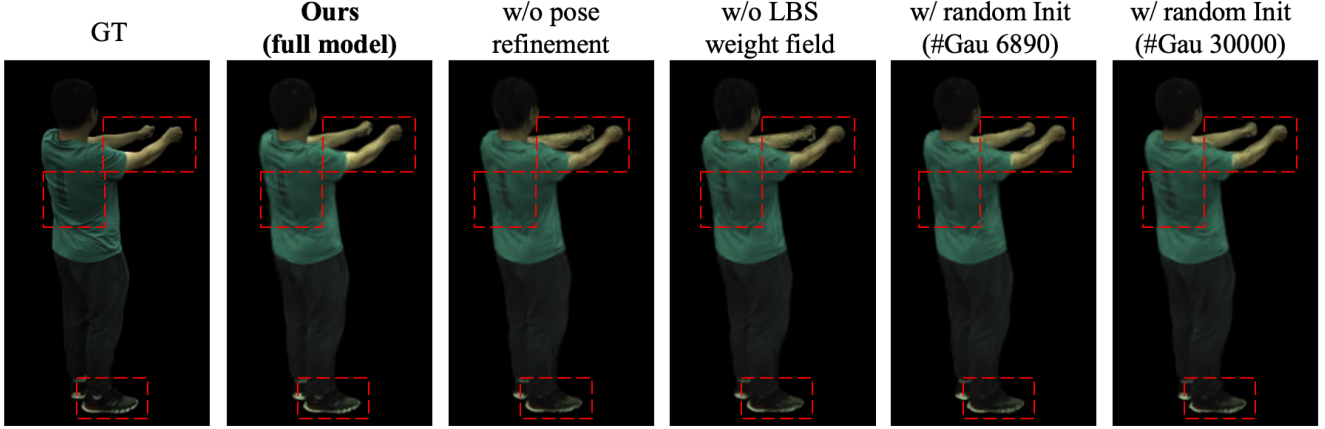


Figure 4. Qualitative results of ablation study on the 386 sequence of ZJU\_Mocap data set. "w/o pose refinement" and "w/o LBS weight field" mean that we do not adopt the pose refinement or LBS weight field module respectively. "w/ random Init (#Gau 6890)" and "w/ random Init (#Gau 30000)" mean that we use random initialization to initialize 6890 or 30000 3D Gaussians respectively.

Table 2. Quantitative Results of ablating pose refinement and LBS weight field modules.  $LPIPS^* = 1000 \times LPIPS$ .

	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS* $\downarrow$	Train	FPS
Ours (full model)	<b>28.08</b>	<b>0.886</b>	<b>103.3</b>	55s	189
Ours w/o pose refine	27.94	0.884	108.8	45s	192
Ours w/o LBS field	27.87	0.882	108.9	52s	190

Structure-from-Motion (SfM). To incorporate human articulated information into the initialization, we propose to use articulated vertex points from SMPL to initialize our 3D Gaussians in the canonical space. We compare our proposed articulated initialization with random initialization<sup>1</sup> in Tab. 3. With the same number of 3D Gaussians at initialization, our proposed articulated initialization produces both better quantitative and qualitative results than random initialization as shown in Tab. 3 and Fig. 4. Notably, our articulated initialization converges x10 times much faster than the random initialization, which illustrates the strength of encoding human articulated information into the initialization. For random initialization, we need to initialize 30k Gaussians to achieve comparable performance as our proposed articulated initialization, while at the cost of 30x times more training time.

Table 3. Quantitative Results of ablating 3D Gaussian initialization. #Gau denotes the number of 3D Gaussians at initialization.

	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS* $\downarrow$	Train
Articulated Init (w/ #Gau 6890)	<b>28.08</b>	<b>0.886</b>	<b>103.3</b>	55s
Random Init (w/ #Gau 6890)	27.98	0.882	108.9	5m
Random Init (w/ #Gau 30k)	<b>28.08</b>	0.884	104.3	30m

**Split/Clone/Merge/Prune.** Adaptive control of 3D Gaussians with split, clone, and prune operations substantially

<sup>1</sup>Estimating sparse point cloud from Structure-from-Motion (SfM) requires multi-view inputs, which is infeasible with monocular videos. So we do not compare it with our articulated initialization.

improves the rendering quality, but it trivially leads to a large number of 3D Gaussians (up to 100k), which slows down the whole optimization process. As mentioned in Sec. 3.3, we propose to use the KL divergence measure to further regulate the split and clone process. As shown in Tab. 4, without sacrificing the final performance, we learn a 3D human performer with about 13k 3D Gaussians. Without sacrificing performance, our KL-based split and clone operations lead to 60x times faster convergence speed when compared with the original split and clone operation proposed in [52]. Furthermore, our novel merge operation merges about 8.5% of 3D Gaussians without degrading the performance. Articulated prune operation with SMPL prior introduced in Sec. 3.3 also helps prune about 0.5k 3D Gaussians far away from the 3D human body.

Table 4. Quantitative Results of ablating our proposed KL-based split/clone operations and a novel merge operation. #Gau denotes the final number of 3D Gaussians during optimization.

	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS* $\downarrow$	#Gau	Train
Ours (full model)	<b>28.08</b>	<b>0.886</b>	<b>103.3</b>	<b>13162</b>	<b>55s</b>
w/o KL split/clone	28.01	0.885	103.1	100k	1h
w/o KL merge	<b>28.08</b>	<b>0.886</b>	103.6	14376	70s
w/o articulated prune	<b>28.08</b>	<b>0.886</b>	103.4	13687	57s

## 5. Discussion and Conclusion

To conclude, we propose GauHuman, a 3D human model with Gaussian Splatting representation for fast training (1~2 minutes) and real-time rendering (166 FPS) of 3D humans. Specifically, we encode this representation in a canonical space and transform 3D Gaussian from canonical space to posed space with LBS transformation, in which effective pose refinement and LBS weight field modules are also designed to learn fine details of 3D humans. To enable fast optimization, we initialize and prune 3D Gaussians with 3D



human prior, while splitting/cloning 3D Gaussians via KL divergence guidance, accompanied by a novel merge operation for further speeding up. Experiments on monocular videos suggest that our GauHuman achieves state-of-the-art performance while maintaining both fast training and real-time rendering speed. In addition, GauHuman can fast learn 3D human performers with  $\sim 13k$  Gaussians.

**Limitations.** 1). GauHuman is composed of 3D Gaussians and current pipeline does not support extracting 3D meshes. How to extract meshes from 3D Gaussians remains a future work to be investigated. 2). Recovering 3D human details like cloth wrinkles from monocular videos is still a challenging problem to tackle. One direction is to incorporate physical simulation of garments into current pipelines.

**Acknowledgment.** We would like to thank Tao Hu from NTU for discussions on the paper framework; Zhaoxi Chen from NTU for discussions on implementing the DVA baseline; Wei Li and Guangcong Wang from NTU for comments on the paper draft; Zhiyuan Yu from ZJU for discussions on the projection matrix used in Gaussian Splatting.

## References

- [1] Badour AlBahar, Jingwan Lu, Jimei Yang, Zhixin Shu, Eli Shechtman, and Jia-Bin Huang. Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan. *ACM Transactions on Graphics (TOG)*, 40(6): 1–11, 2021. 3
- [2] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 696–712. Springer, 2020. 3
- [3] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1506–1515, 2022. 3
- [4] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022. 3
- [5] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. *Advances in Neural Information Processing Systems*, 33:12909–12922, 2020. 5
- [6] Mario Botsch, Alexander Hornung, Matthias Zwicker, and Leif Kobbelt. High-quality surface splatting on today’s gpus. In *Proceedings Eurographics/IEEE VGTC Symposium Point-Based Graphics, 2005.*, pages 17–141. IEEE, 2005. 3
- [7] Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. Unstructured lumigraph rendering. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 497–504. 2023. 3
- [8] Sihun Cha, Kwanggyoon Seo, Amirsaman Ashtari, and Junyong Noh. Generating texture for 3d human avatar from a single image using sampling and refinement networks. In *Computer Graphics Forum*, pages 385–396. Wiley Online Library, 2023. 1, 3
- [9] Gaurav Chaurasia, Sylvain Duchene, Olga Sorkine-Hornung, and George Drettakis. Depth synthesis and local warps for plausible image-based navigation. *ACM Transactions on Graphics (TOG)*, 32(3):1–12, 2013. 3
- [10] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. *arXiv preprint arXiv:2103.15595*, 2021. 3
- [11] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision*, pages 333–350. Springer, 2022. 3
- [12] 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. 1, 3, 5
- [13] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11594–11604, 2021. 5
- [14] Zhaoxi Chen, Fangzhou Hong, Haiyi Mei, Guangcong Wang, Lei Yang, and Ziwei Liu. Primdifffusion: Volumetric primitives diffusion for 3d human generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3
- [15] 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. 1
- [16] Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo radiance fields (srf): Learning view synthesis for sparse views of novel scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7911–7920, 2021. 3
- [17] Hongsuk Choi, Gyeongsik Moon, Matthieu Armando, Vincent Leroy, Kyoung Mu Lee, and Grégory Rogez. Mononhr: Monocular neural human renderer. In *2022 International Conference on 3D Vision (3DV)*, pages 242–251. IEEE, 2022. 1, 3
- [18] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (ToG)*, 34(4):1–13, 2015. 3
- [19] François Darmon, Bénédicte Basclé, Jean-Clément Devaux, Pascal Monasse, and Mathieu Aubry. Improving neural implicit surfaces geometry with patch warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6260–6269, 2022. 3

- [20] Boyang Deng, John P Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Nasa neural articulated shape approximation. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 612–628. Springer, 2020. 5
- [21] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. *arXiv preprint arXiv:2107.02791*, 2021. 3
- [22] Mingsong Dou, Sameh Khamis, Yuri Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (ToG)*, 35(4): 1–13, 2016. 3
- [23] Martin Eisemann, Bert De Decker, Marcus Magnor, Philippe Bekaert, Edilson De Aguiar, Naveed Ahmed, Christian Theobalt, and Anita Sellent. Floating textures. In *Computer graphics forum*, pages 409–418. Wiley Online Library, 2008. 3
- [24] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. Deepstereo: Learning to predict new views from the world’s imagery. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5515–5524, 2016. 3
- [25] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022. 3
- [26] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. In *European Conference on Computer Vision*, pages 1–19. Springer, 2022. 3
- [27] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Wayne Wu, and Ziwei Liu. Unitedhuman: Harnessing multi-source data for high-resolution human generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7301–7311, 2023. 3
- [28] Xiangjun Gao, Jiaolong Yang, Jongyoo Kim, Sida Peng, Zicheng Liu, and Xin Tong. Mps-nerf: Generalizable 3d human rendering from multiview images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 3
- [29] Stephan J Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14346–14355, 2021. 3
- [30] Chen Geng, Sida Peng, Zhen Xu, Hujun Bao, and Xiaowei Zhou. Learning neural volumetric representations of dynamic humans in minutes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8759–8770, 2023. 2, 3, 6, 7, 15
- [31] Michael Goesele, Noah Snavely, Brian Curless, Hugues Hoppe, and Steven M Seitz. Multi-view stereo for community photo collections. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. 3
- [32] Steven J Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F Cohen. The lumigraph. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 453–464. 2023. 3
- [33] Markus Gross and Hanspeter Pfister. *Point-based graphics*. Elsevier, 2011. 3
- [34] Jeffrey P Grossman and William J Dally. Point sample rendering. In *Rendering Techniques’ 98: Proceedings of the Eurographics Workshop in Vienna, Austria, June 29–July 1, 1998 9*, pages 181–192. Springer, 1998. 3
- [35] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (ToG)*, 38(6):1–19, 2019. 3
- [36] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5052–5063, 2020. 1, 6
- [37] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. *ACM Transactions on Graphics (ToG)*, 40(4):1–16, 2021. 1, 6
- [38] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. Deep blending for free-viewpoint image-based rendering. *ACM Transactions on Graphics (ToG)*, 37(6):1–15, 2018. 3
- [39] Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. Escaping plato’s cave: 3d shape from adversarial rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9984–9993, 2019. 3
- [40] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. Sherf: Generalizable human nerf from a single image. *arXiv preprint arXiv:2303.12791*, 2023. 1, 3
- [41] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. Tech: Text-guided reconstruction of lifelike clothed humans. *arXiv preprint arXiv:2308.08545*, 2023. 1, 3
- [42] 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. 5
- [43] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5885–5894, 2021. 3
- [44] Wonbong Jang and Lourdes Agapito. Codenerf: Disentangled neural radiance fields for object categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12949–12958, 2021. 3
- [45] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *Proceedings of the IEEE/CVF Conference on*

- Computer Vision and Pattern Recognition*, pages 5605–5615, 2022. 1, 3
- [46] 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*, pages 16922–16932, 2023. 2, 3, 5, 6, 7, 15, 16
- [47] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. *arXiv preprint arXiv:2203.12575*, 2022. 1, 3
- [48] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022. 3
- [49] Yuming Jiang, Shuai Yang, Tong Liang Koh, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2performer: Text-driven human video generation. *arXiv preprint arXiv:2304.08483*, 2023. 3
- [50] Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. Geonerf: Generalizing nerf with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18365–18375, 2022. 3
- [51] James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. *ACM SIGGRAPH computer graphics*, 18(3):165–174, 1984. 3
- [52] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023. 2, 3, 4, 5, 7, 8, 15
- [53] Georgios Kopanas, Julien Philip, Thomas Leimkühler, and George Drettakis. Point-based neural rendering with per-view optimization. In *Computer Graphics Forum*, pages 29–43. Wiley Online Library, 2021. 3
- [54] Georgios Kopanas, Thomas Leimkühler, Gilles Rainer, Clément Jambon, and George Drettakis. Neural point catcaustics for novel-view synthesis of reflections. *ACM Transactions on Graphics (TOG)*, 41(6):1–15, 2022. 3
- [55] 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*, 34:24741–24752, 2021. 1, 3, 6, 7, 15
- [56] Christoph Lassner and Michael Zollhofer. Pulsar: Efficient sphere-based neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1440–1449, 2021. 3
- [57] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. *arXiv preprint arXiv:2311.16099*, 2023. 16
- [58] Marc Levoy and Pat Hanrahan. Light field rendering. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 441–452. 2023. 3
- [59] Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. Tryongan: Body-aware try-on via layered interpolation. *ACM Transactions on Graphics (TOG)*, 40(4):1–10, 2021. 3
- [60] Ruilong Li, Matthew Tancik, and Angjoo Kanazawa. Nerfacc: A general nerf acceleration toolbox. *arXiv preprint arXiv:2210.04847*, 2022. 3
- [61] Tingting Liao, Xiaomei Zhang, Yuliang Xiu, Hongwei Yi, Xudong Liu, Guo-Jun Qi, Yong Zhang, Xuan Wang, Xianguyu Zhu, and Zhen Lei. High-fidelity clothed avatar reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8662–8672, 2023. 1, 3
- [62] Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields for interactive free-viewpoint video. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 3
- [63] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33:15651–15663, 2020. 3
- [64] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM transactions on graphics (TOG)*, 40(6):1–16, 2021. 3
- [65] Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, and Wenping Wang. Neural rays for occlusion-aware image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7824–7833, 2022. 3
- [66] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 3, 15
- [67] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 2, 3
- [68] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. *arXiv preprint arXiv:2308.09713*, 2023. 2, 3, 4
- [69] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995. 3
- [70] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5084–5093, 2020. 3
- [71] 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*, pages 10461–10471, 2021. 5
- [72] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view syn-



- thesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 3
- [73] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multi-resolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 2, 3
- [74] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. *arXiv preprint arXiv:2112.00724*, 2021. 3
- [75] Atsuhiko Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5762–5772, 2021. 1, 3
- [76] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 15
- [77] 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. 5, 6, 15
- [78] 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. 1, 3, 6, 7, 15
- [79] Sida Peng, Shangzhan Zhang, Zhen Xu, Chen Geng, Boyi Jiang, Hujun Bao, and Xiaowei Zhou. Animatable neural implicit surfaces for creating avatars from videos. *arXiv preprint arXiv:2203.08133*, 4(5), 2022. 1, 3, 6, 7, 15, 16
- [80] Eric Penner and Li Zhang. Soft 3d reconstruction for view synthesis. *ACM Transactions on Graphics (TOG)*, 36(6):1–11, 2017. 3
- [81] Hanspeter Pfister, Matthias Zwicker, Jeroen Van Baar, and Markus Gross. Surfels: Surface elements as rendering primitives. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 335–342, 2000. 3
- [82] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kiloverf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14335–14345, 2021. 3
- [83] Konstantinos Rematas, Ricardo Martin-Brualla, and Vittorio Ferrari. Sharf: Shape-conditioned radiance fields from a single view. *arXiv preprint arXiv:2102.08860*, 2021. 3
- [84] 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, 6, 7
- [85] Liu Ren, Hanspeter Pfister, and Matthias Zwicker. Object space ewa surface splatting: A hardware accelerated approach to high quality point rendering. In *Computer Graphics Forum*, pages 461–470. Wiley Online Library, 2002. 3
- [86] Helge Rhodin, Nadia Robertini, Christian Richardt, Hans-Peter Seidel, and Christian Theobalt. A versatile scene model with differentiable visibility applied to generative pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 765–773, 2015. 3
- [87] Gernot Riegler and Vladlen Koltun. Free view synthesis. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pages 623–640. Springer, 2020. 3
- [88] Darius Rückert, Linus Franke, and Marc Stamminger. Adop: Approximate differentiable one-pixel point rendering. *ACM Transactions on Graphics (ToG)*, 41(4):1–14, 2022. 3
- [89] Miguel Sainz and Renato Pajarola. Point-based rendering techniques. *Computers & Graphics*, 28(6):869–879, 2004. 3
- [90] 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
- [91] 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
- [92] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black. Scanimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2886–2897, 2021. 5
- [93] Umme Sara, Morium Akter, and Mohammad Shorif Uddin. Image quality assessment through fsm, ssim, mse and psnr—a comparative study. *Journal of Computer and Communications*, 7(3):8–18, 2019. 15
- [94] Kripasindhu Sarkar, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Humangan: A generative model of human images. In *2021 International Conference on 3D Vision (3DV)*, pages 258–267. IEEE, 2021. 3
- [95] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 2, 5
- [96] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III 14*, pages 501–518. Springer, 2016. 3
- [97] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019. 3



- [98] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pages 835–846. 2006. 2, 3, 5
- [99] Carsten Stoll, Nils Hasler, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt. Fast articulated motion tracking using a sums of gaussians body model. In *2011 International Conference on Computer Vision*, pages 951–958. IEEE, 2011. 3
- [100] 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*, 34:12278–12291, 2021. 1, 3
- [101] Zhuo Su, Lan Xu, Zerong Zheng, Tao Yu, Yebin Liu, and Lu Fang. Robustfusion: Human volumetric capture with data-driven visual cues using a rgb-d camera. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 246–264. Springer, 2020. 3
- [102] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2022. 3
- [103] Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11358–11367, 2021. 3
- [104] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Wang Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. Advances in neural rendering. In *Computer Graphics Forum*, pages 703–735. Wiley Online Library, 2022. 3
- [105] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 3
- [106] Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Neural-gif: Neural generalized implicit functions for animating people in clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11708–11718, 2021. 5
- [107] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d representation and rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15182–15192, 2021. 3
- [108] Angtian Wang, Peng Wang, Jian Sun, Adam Kortylewski, and Alan Yuille. Voge: a differentiable volume renderer using gaussian ellipsoids for analysis-by-synthesis. In *The Eleventh International Conference on Learning Representations*, 2022. 3
- [109] 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. *arXiv preprint arXiv:2106.10689*, 2021. 3
- [110] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 3
- [111] Shaoifei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdf. In *European conference on computer vision*, 2022. 1, 3
- [112] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 15
- [113] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Vid2actor: Free-viewpoint animatable person synthesis from video in the wild. *arXiv preprint arXiv:2012.12884*, 2020. 5
- [114] 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. 1, 3, 5, 6, 7, 15
- [115] Zhenzhen Weng, Zeyu Wang, and Serena Yeung. Zeroavatar: Zero-shot 3d avatar generation from a single image. *arXiv preprint arXiv:2305.16411*, 2023. 1, 3
- [116] EP Wigner. Gruppentheorie und ihre anwendungen auf die quantenmechanik der atomspektr. friedr. vieweg und sohn akt. Ges., Braunschweig, 1931. 15
- [117] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7467–7477, 2020. 3
- [118] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. *arXiv preprint arXiv:2310.08528*, 2023. 2, 3, 4
- [119] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 3
- [120] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5438–5448, 2022. 3
- [121] Zhen Xu, Sida Peng, Haotong Lin, Guangzhao He, Jiaming Sun, Yujun Shen, Hujun Bao, and Xiaowei Zhou. 4k4d: Real-time 4d view synthesis at 4k resolution. *arXiv preprint arXiv:2310.11448*, 2023. 3
- [122] Jiayu Yang, Wei Mao, Jose M Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4877–4886, 2020. 3

- [123] Ze Yang, Shenlong Wang, Sivabalan Manivasagam, Zeng Huang, Wei-Chiu Ma, Xinchun Yan, Ersin Yumer, and Raquel Urtasun. S3: Neural shape, skeleton, and skinning fields for 3d human modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13284–13293, 2021. 5
- [124] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. *arXiv preprint arXiv:2309.13101*, 2023. 2, 3, 4
- [125] Zeyu Yang, Hongye Yang, Zijie Pan, Xiatian Zhu, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. *arXiv preprint arXiv:2310.10642*, 2023. 2, 3, 4
- [126] Wang Yifan, Felice Serena, Shihao Wu, Cengiz Öztireli, and Olga Sorkine-Hornung. Differentiable surface splatting for point-based geometry processing. *ACM Transactions on Graphics (TOG)*, 38(6):1–14, 2019. 3
- [127] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenotrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 3
- [128] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 3, 6, 15
- [129] Kai Zhang, Gernot Riegler, Noah Snaveley, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 3
- [130] Qiang Zhang, Seung-Hwan Baek, Szymon Rusinkiewicz, and Felix Heide. Differentiable point-based radiance fields for efficient view synthesis. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–12, 2022. 2, 3
- [131] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 15
- [132] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Generalizable neural human radiance field from sparse inputs. *arXiv preprint arXiv:2112.02789*, 3, 2021. 1, 3
- [133] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 286–301. Springer, 2016. 3
- [134] Matthias Zwicker, Hanspeter Pfister, Jeroen Van Baar, and Markus Gross. Ewa volume splatting. In *Proceedings Visualization, 2001. VIS’01.*, pages 29–538. IEEE, 2001. 2, 3, 4
- [135] Matthias Zwicker, Hanspeter Pfister, Jeroen Van Baar, and Markus Gross. Surface splatting. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 371–378, 2001. 3

## A. Implementation Details

**Implementation Details of GauHuman** Our pose refinement module consists of 4 fully connected layers, *i.e.*, an input layer, 2 hidden layers, and one output layer. Each layer is followed by a ReLU activation. The dimension of the hidden layer is 128, while the input and output dimension of the pose refinement module is 69. The LBS offset module adopts 5 fully connected layers with an input layer, 3 hidden layers, and one output layer. The ReLU activation is used after each layer. Positional encoding is applied to the 3D Gaussian positions before they are fed into the input layer. The input and output dimension of the LBS offset module is 63 and 24 (number of joints). We use Adam optimizer with a learning rate  $10^{-5}$  to optimize the above two modules. We set the threshold of KL divergence as 0.4 to perform split/clone operations. Other training details for 3D Gaussians are the same as [52].

**Evaluation Metrics.** To quantitatively evaluate the quality of rendered novel view and novel pose images, we report the peak signal-to-noise ratio (PSNR) [93], structural similarity index (SSIM) [112] and Learned Perceptual Image Patch Similarity (LPIPS) [131].

**Details of Comparable Methods.** 1). Subject-specific optimization-based methods. Neural Body (NB) [78] encodes latent codes in SMPL vertex points and uses them to learn the neural radiance fields. Animatable NeRF (AN) [77] learns a canonical human NeRF through skeleton-driven deformation and learned blend weight fields. AS[79] further extends [77] by learning a signed distance field and a pose-dependent deformation field for residual information and geometric details of dynamic 3D humans. HumanNeRF [114] incorporates a pose refinement module, LBS field, and non-rigid deformation module to optimize a volumetric representation of 3D humans in the canonical space. Unlike the above methods using coordinated neural networks to learn neural radiance fields, InstantNVR [30] and InstantAvatar [46] propose to use multi-hashing encoding for fast training of 3D humans. Given the fast rendering speed of mixtures of volumetric primitives [66], DVA further extends it to articulated 3D humans for high-quality telepresence. Specifically, DVA first attaches a fixed number (4096) of voxels in the UV space and then transforms voxels to posed space with the sparse multi-view input images as the conditioning information, along with an MLP decoder to regress the opacity and color information. 2). Generalizable methods. PixelNeRF [128] learns a neural network to infer the radiance field based on the input image. Neural Human Performer (NHP) [55] aggregates pixel-aligned features at each time step and temporally-fused features to learn generalizable neural radiance fields. For generalizable methods, we evaluate each subject (*e.g.*, one subject of MonoCap) by first pre-training the model on the other data set (*e.g.*, ZJU\_Mocap data set) and then fine-tuning it on the evaluated subject.

**Efficient Implementation of KL Divergence** As the covariance matrix is decomposed into the product of rotation and scaling matrices  $\Sigma = RSS^T R^T$ , we simplify the computation of matrix inverse and determinant operations, *i.e.*,

$$\begin{aligned}\Sigma_1^{-1} &= (RSS^T R^T)^{-1} = RS^{-1} S^{-1} R^T, \\ \det \Sigma_1 &= \det(RSS^T R^T) = \det(S) * \det(S)\end{aligned}\quad (9)$$

Since scaling matrix  $S$  is a diagonal matrix, the inverse and determinant of a diagonal matrix can be easily derived by inverting and prodding the diagonal elements respectively. Meanwhile, the inverse of the orthogonal rotation matrix is the transpose of the original matrix. The above simplification saves the computation time for matrix inverse and determinant operation.

**Rotating Spherical Harmonic coefficients** When transforming 3D Gaussians from canonical space to posed space, the SH coefficients should also be rotated for view-dependent color effects. The above is achieved by first computing a Wigner D-matrix [116] and then rotating SH coefficients with the Wigner D-matrix. In our implementation, we find that rotating SH coefficients has little effect on the final performance<sup>2</sup>, so we do not consider it in our work.

## B. Details of Loss Functions

**Photometric Loss.** Given the ground truth target image  $C$  and predicted image  $\hat{C}$ , we apply the photometric loss:

$$\mathcal{L}_{color} = \|\hat{C} - C\|_2. \quad (10)$$

**Mask Loss.** We also leverage the human region masks for Human NeRF optimization. The mask loss is defined as:

$$\mathcal{L}_{mask} = \|\hat{M} - M\|_2, \quad (11)$$

where  $\hat{M}$  is the accumulated volume density and  $M$  is the ground truth binary mask label.

**SSIM Loss.** We further employ SSIM to ensure the structural similarity between ground truth and synthesized images:

$$\mathcal{L}_{SSIM} = \text{SSIM}(\hat{C}, C). \quad (12)$$

**LPIPS Loss.** The perceptual loss LPIPS is also utilized to ensure the quality of rendered image, *i.e.*,

$$\mathcal{L}_{LPIPS} = \text{LPIPS}(\hat{C}, C). \quad (13)$$

## C. Experiments on DNA-Rendering

We further evaluate the performance of our GauHuman and two representative baseline methods on a DNA-Rendering data set. We select two sequences (0012\_09 and 0025\_11

<sup>2</sup>We also find that the implementation of Wigner D-matrix using Pytorch [76] is time-consuming due to the matrix exponential operation.

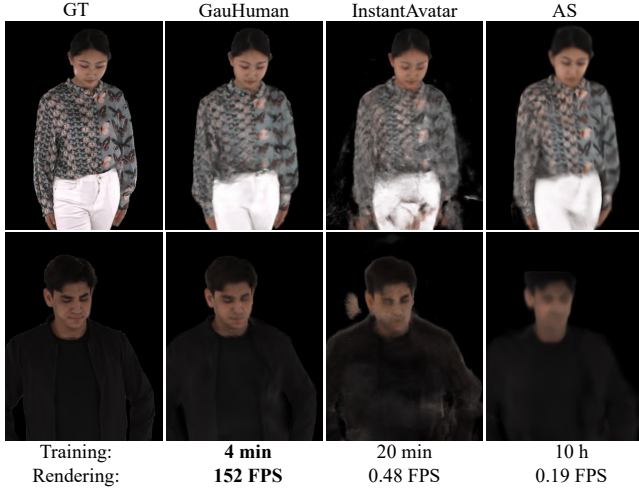


Figure 5. Novel view synthesis results produced by our GauHuman and baseline methods on DNA Rendering data set. The bottom lines show the training time and rendering speed of each method on the DNA Rendering data set. Zoom in for the best view.

from part 1) from the DNA-Rendering data set and collect 100 frames for training. Similar to ZJU\_MoCap and MonoCap, one camera is used for training. For evaluation purposes, we use four nearby camera views as testing views. As shown in Tab. 5 and Fig. AS [79] and InstantAvatar [46] struggle to produce photorealistic renderings due to the complex clothing and fast-moving human actors recorded on the DNA-Rendering data set. In comparison, our GauHuman learns high-quality 3D human performers with fast training and rendering speed, which verifies the flexibility and efficiency of 3D Gaussian Splatting.

Table 5. Quantitative comparison of our GauHuman and baseline methods on the DNA-Rendering data set. LPIPS\* =  $1000 \times \text{LPIPS}$ . Frames per second (FPS) are measured on an RTX 3090.

Method	DNA-Rendering				
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS* $\downarrow$	Train	FPS
AS [79]	27.67	0.954	50.99	10h	0.14
InstantAvatar [46]	24.77	0.922	78.55	20m	0.48
<b>GauHuman(Ours)</b>	<b>29.11</b>	<b>0.961</b>	<b>37.68</b>	<b>4m</b>	<b>152</b>

## D. Comparison with concurrent work GART [57]

Our concurrent work GART [57] also extends Gaussian Splatting to 3D human modelling with monocular videos. It achieves comparable novel view synthesis performance when compared with state-of-the-art baseline methods on ZJU\_MoCap data set while improving the rendering speed to 77 FPS with the efficient 3D Gaussian Splatting technique. We reproduce the result of GART with their released code and show the comparison results in Tab. 6. For a fair comparison, we do not conduct test-time optimization of SMPL parameters with images from the test set on GART [57]. In

comparison with GART, our GauHuman produces slightly better novel view synthesis performance with both faster training (1m vs. 3m) and rendering (189FPS vs. 77FPS) speed. Specifically, we achieve fast optimization of GauHuman by initializing and pruning 3D Gaussians with 3D human prior, while splitting/cloning via KL divergence guidance, along with a novel merge operation for further speeding up. Notably, without sacrificing rendering quality, GauHuman can fast model the 3D human performer with  $\sim 13\text{k}$  3D Gaussians.

Table 6. Quantitative comparison of our GauHuman and GART on the ZJU\_MoCap data set. LPIPS\* =  $1000 \times \text{LPIPS}$ . Frames per second (FPS) is measured on an RTX 3090. **For a fair comparison, we do not conduct test-time optimization of SMPL parameters with images from the test set on GART [57].**

Method	ZJU_MoCap (Avg)				
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS* $\downarrow$	#Gau	Train FPS
GART	30.91	0.9615	31.83	53.4k	3m 77
<b>GauHuman (Ours)</b>	<b>31.34</b>	<b>0.9647</b>	<b>30.51</b>	<b>11.8k</b>	<b>1m 189</b>
my_377					
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS* $\downarrow$	#Gau	
GART	31.90	0.9747	<b>18.8</b>	55.0k	
<b>GauHuman (Ours)</b>	<b>32.24</b>	<b>0.9757</b>	18.9	<b>12.6k</b>	
my_386					
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS* $\downarrow$	#Gau	
GART	33.50	0.9669	29.9	51.4k	
<b>GauHuman (Ours)</b>	<b>33.72</b>	<b>0.9693</b>	<b>29.0</b>	<b>13.1k</b>	
my_387					
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS* $\downarrow$	#Gau	
GART	27.74	0.9518	40.3	52.9k	
<b>GauHuman (Ours)</b>	<b>28.19</b>	<b>0.9564</b>	<b>39.3</b>	<b>9.9k</b>	
my_392					
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS* $\downarrow$	#Gau	
GART	31.92	0.9637	32.6	51.6k	
<b>GauHuman (Ours)</b>	<b>32.27</b>	<b>0.9669</b>	<b>30.2</b>	<b>11.3k</b>	
my_393					
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS* $\downarrow$	#Gau	
GART	29.34	0.9540	37.9	51.7k	
<b>GauHuman (Ours)</b>	<b>30.24</b>	<b>0.9584</b>	<b>35.2</b>	<b>11.0k</b>	
my_394					
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS* $\downarrow$	#Gau	
GART	31.08	0.9577	31.5	57.7k	
<b>GauHuman (Ours)</b>	<b>31.42</b>	<b>0.9611</b>	<b>30.6</b>	<b>12.8k</b>	