

SG-GS: Topology-aware Human Avatars with Semantically-guided Gaussian Splatting

Haoyu Zhao^{* 1,2}, Chen Yang^{* 1}, Hao Wang^{* 3}, Xingyue Zhao⁴, Wei Shen^{†1}

¹MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University

²School of Computer Science, Wuhan University

³Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology

⁴School of Software Engineering, Xi'an Jiao Tong University

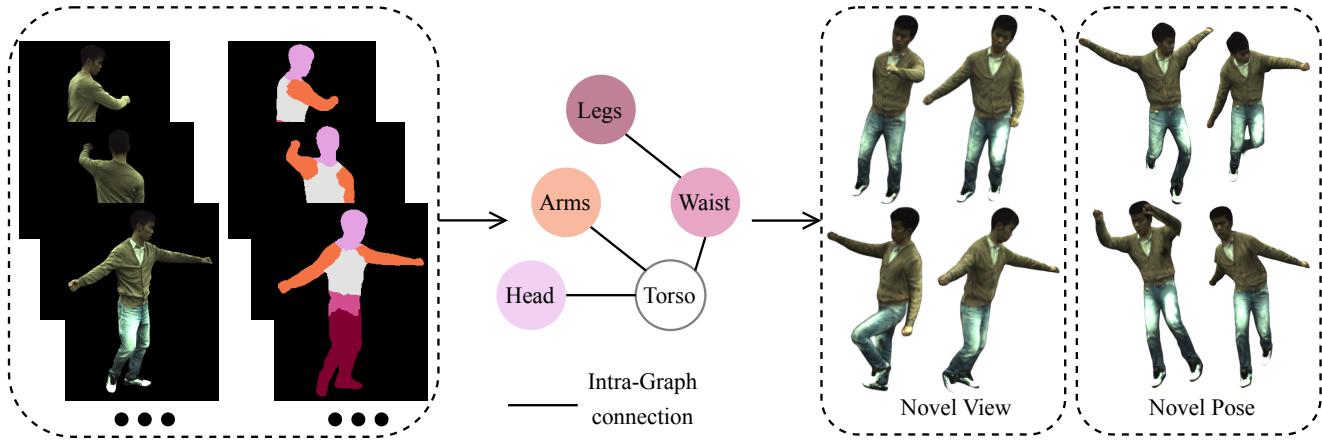


Figure 1. We propose an efficient method for creating topology-aware human avatars from just videos, ensuring both photo-realistic human appearance and accurate anatomical structure. Our method achieves better quality than the most recent state-of-the-art methods [10, 32, 39].

Abstract

Reconstructing photo-realistic and topology-aware animatable human avatars from monocular videos remains challenging in computer vision and graphics. Recently, methods using 3D Gaussians to represent the human body have emerged, offering faster optimization and real-time rendering. However, due to ignoring the crucial role of human body semantic information which represents the explicit topological and intrinsic structure within human body, they fail to achieve fine-detail reconstruction of human avatars. To address this issue, we propose SG-GS, which uses semantics-embedded 3D Gaussians, skeleton-driven rigid deformation, and non-rigid cloth dynamics deformation to create photo-realistic human avatars. We then design a Semantic Human-Body Annotator (SHA) which uti-

lizes SMPL's semantic prior for efficient body part semantic labeling. The generated labels are used to guide the optimization of semantic attributes of Gaussian. To capture the explicit topological structure of the human body, we employ a 3D network that integrates both topological and geometric associations for human avatar deformation. We further implement three key strategies to enhance the semantic accuracy of 3D Gaussians and rendering quality: semantic projection with 2D regularization, semantic-guided density regularization and semantic-aware regularization with neighborhood consistency. Extensive experiments demonstrate that SG-GS achieves state-of-the-art geometry and appearance reconstruction performance. Our project is at <https://sggs-projectpage.github.io/>.

1. Introduction

Creating photo-realistic human avatars from monocular videos has immense potential value in industries such

^{*} Equal contributions.

[†]Corresponding Author.

Haoyu Zhao completed this work during an internship at Shanghai Jiao Tong University.

as gaming [47], extended reality storytelling [7], and telepresence [8]. In this work, we are dedicated to create high-quality photo-realistic human avatars from monocular videos with semantics embedded 3D Gaussians.

Recent advances in implicit neural fields [26, 36] enable high-quality reconstruction of geometry [?, 6, 42] and appearance [13, 19, 40, 46] of clothed human bodies from sparse multi-view or monocular videos. However, they often employ large MLPs, which makes training and rendering computationally demanding and inefficient.

Point-based rendering [49] has emerged as an efficient alternative to NeRFs, offering significantly faster rendering speed. The recently proposed 3D Gaussian Splatting (3DGS) [15] achieves state-of-the-art novel view synthesis performance with significantly reduced inference time and faster training. 3DGS has inspired several recent works in human avatar creation [9, 10, 16, 17, 27, 32, 35, 37]. However, these methods often overlook crucial semantic information that represents the explicit topological structure within the human body, leading to issues in maintaining anatomical coherence during motion and preserving fine details such as muscle definition and skin folds in various poses.

To this end, we propose **SG-GS**, a Semantically-Guided 3D human model using Gaussian Splatting representation, as shown in Fig. 1. SG-GS first integrates a skeleton-driven rigid deformation, and a non-rigid cloth dynamics deformation to coordinate the movements of individual Gaussians during animation. We then introduce a Semantic Human-Body Annotator (SHA), which leverages SMPL’s [22] human semantic prior for efficient body part semantic labeling. These part labels are used to guide the optimization of 3D Gaussian’s semantic attribute. To learn topological relationships between human body parts, we propose a 3D topology- and geometry-aware network to learn body geometric and topological associations and integrate them into the avatar deformation. We further implement three key strategies to enhance semantic accuracy of 3D Gaussians and rendering quality: semantic projection with 2D regularization, semantic-guided density regularization and semantic-aware regularization with neighborhood consistency. Our experimental results demonstrate that SG-GS achieves superior performance compared to current SOTA approaches in avatar creation from monocular inputs. In summary, our work makes the following contributions:

- We propose SG-GS, which is the first to integrate semantic priors from the human body into creating animatable human avatars from monocular videos.
- We propose a 3D topology and geometry-aware network to capture topology and geometry information within the human body.
- We introduce semantic projection with 2D regularization, semantic neighborhood-consistent regularization,

and semantic-guided density regularization to enhance semantic accuracy and rendering quality.

2. Related Work

2.1. Neural Rendering for Human Avatars

Since the introduction of Neural Radiance Fields (NeRF) [26], there has been a surge of research on neural rendering for human avatars [?, 19–21, 31]. Though, NeRF is designed for static objects, HumanNeRF [40] extend the NeRF to enable capturing a dynamic moving human using just a single monocular video. Neural Body [31] associates a latent code to each SMPL [22] vertex to encode the appearance, which is transformed into observation space based on the human pose. Furthermore, Neural Actor [21] learns a deformable radiance field with SMPL [22] as guidance and utilizes a texture map to improve its final rendering quality. Posevocab [20] designs joint-structured pose embeddings to encode dynamic appearances under different key poses, enabling more effective learning of joint-related appearances. However, a major limitation of NeRF-based methods is that NeRFs are slow to train and render.

Some works focus on achieving fast inference and training times for NeRF models of human avatars, including approaches that use explicit representations such as learning a function at grid points [1], using hash encoding [28], or altogether discarding the learnable component [3]. iNGP [28] uses the underlying representation for articulated NeRFs, and enable interactive rendering speeds (15 FPS). [2] generates a pose-dependent UV volume, but its UV volume generation is not fast (20 FPS). In contrast to all these works, SG-GS achieves state-of-the-art rendering quality and speed (25 FPS) with less training time.

2.2. Dynamic 3D Gaussians for Human Avatars

Point-based rendering [34, 49] has proven to be an efficient alternative to NeRFs for fast inference and training. Extending point clouds to 3D Gaussians, 3D Gaussian Splatting (3DGS) [15] models the rendering process by splatting a set of 3D Gaussians onto the image plane via alpha blending. This approach achieves SOTA rendering quality with fast inference speed for novel views.

Given the impressive performance of 3DGS in both quality and speed, numerous works have further explored the 3D Gaussian representation for dynamic scene reconstruction. D-3DGS [24] is proposed as the first attempt to adapt 3DGS into a dynamic setup. Other works [41, 44, 48] model 3D Gaussian motions with a compact network or 4D primitives, resulting in highly efficient training and real-time rendering.

The application of 3DGS in dynamic 3D human avatar reconstruction is just beginning to unfold [10, 14, 16, 17, 32]. Human Gaussian Splatting [27] showcase 3DGS as an efficient alternative to NeRF. Splattingavatar [35] and Goma-

vatar [39] extends lifted optimization to simultaneously optimize the parameters of the Gaussians while walking on the triangle mesh. While these methods have made significant progress, they often overlook the crucial role of semantic information which is related to topological relationships between human body parts. It is a key focus of our SG-GS.

3. Preliminaries

SMPL [22]. The SMPL model is a widely-used parametric 3D human body model that efficiently represents body shape and pose variations. In our work, We utilize SMPL’s Linear Blend Skinning (LBS) algorithm to transform points from canonical space to observation space, enabling accurate body deformation across different poses. We also leverage SMPL’s body priors to enhance the model’s understanding of body structure, improving the quality and consistency of human avatar reconstruction.

3D Gaussian Splatting (3DGS) [15]. 3DGS explicitly represents scenes using point clouds, where each point is modeled as a 3D Gaussian defined by a covariance matrix Σ and a center point \mathcal{X} , the latter referred to as the mean. The value at point \mathcal{X} is: $G(\mathcal{X}) = e^{-\frac{1}{2}\mathcal{X}^T \Sigma^{-1} \mathcal{X}}$.

For differentiable optimization, the covariance matrix Σ is decomposed into a scaling matrix \mathcal{S} and a rotation matrix \mathcal{R} , such that $\Sigma = \mathcal{R} \mathcal{S} \mathcal{S}^T \mathcal{R}^T$. In practice, \mathcal{S} and \mathcal{R} are also represented by the diagonal vector $s \in \mathbb{R}^{N \times 3}$ and a quaternion vector $r \in \mathbb{R}^{N \times 4}$, respectively. In rendering novel views, differential splatting, as introduced by [45] and [50], involves applying a viewing transformation W along with the Jacobian matrix J of the affine approximation of the projective transformation. This process computes the transformed covariance matrix as: $\Sigma' = JW\Sigma W^T J^T$. The color and opacity at each pixel are computed from the Gaussian’s representation $G(\mathcal{X}) = e^{-\frac{1}{2}\mathcal{X}^T \Sigma^{-1} \mathcal{X}}$. The pixel color \mathcal{C} is computed by blending N ordered 3D Gaussian splats that overlap at the given pixel, using the formula:

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

Here, c_i, α_i represents the density and color of this point computed by a 3D Gaussian G with covariance Σ multiplied by an optimizable per-point opacity and SH color coefficients. The 3D Gaussians are optimized using a photometric loss. 3DGS adjusts their number through periodic densification and pruning, achieving an optimal density distribution that accurately represents the scene.

4. Method

In this section, we illustrate the pipeline of our SG-GS in Fig. 2. The inputs to our method include images

$X = \{\mathcal{X}_i\}_{i=1}^N$ obtained from monocular videos, fitted SMPL parameters $P = \{p_i\}_{i=1}^N$, and paired foreground masks $M = \{m_i\}_{i=1}^N$ of images. SG-GS optimizes 3D Gaussians in canonical space, which are then deformed to match the observation space and rendered from the provided camera view. For a set of 3D Gaussians, we store the following properties at each point: position $\mathcal{X} \in \mathbb{R}^3$, color defined by spherical harmonic (SH) coefficients $\mathcal{C} \in \mathbb{R}^k$ (where k is the number of SH functions), opacity $\alpha \in \mathbb{R}$, rotation factor $r \in \mathbb{R}^4$, and scaling factor $s \in \mathbb{R}^3$. To integrate semantic information about body parts into the 3D Gaussian optimization process and learn the topological structure of the human body, we divide the human body into 5 distinct parts, as shown in Fig. 1. We represent the labels using one-hot encoding, stored as semantic attribute $\mathcal{O} \in \mathbb{R}^{10}$.

4.1. Non-rigid and Rigid Deformation

Inspired by [32, 40], We decompose human deformation into two key components: 1) a non-rigid element capturing pose-dependent cloth dynamics, and 2) a rigid transformation governed by the human skeletal structure.

We employ a non-rigid deformation network, that takes the canonical position \mathcal{X}_c of the 3D Gaussians \mathcal{G}_c in canonical space and a pose latent code as input. This pose latent code encodes SMPL parameters p_i using a lightweight hierarchical pose encoder [25] into \mathcal{Z}_p . The network then outputs offsets for various parameters of \mathcal{G}_c :

$$\Delta(\mathcal{X}, \mathcal{C}, \alpha, s, r) = f_{\theta_{nr}}(\mathcal{X}_c; \mathcal{Z}_p). \quad (2)$$

This network enables efficient and detailed non-rigid deformation of the 3D Gaussians, effectively capturing the nuances of human body movement and shape. The canonical Gaussian is deformed by:

$$\mathcal{X}_d = \mathcal{X}_c + \Delta\mathcal{X}, \mathcal{C}_d = \mathcal{C}_c + \Delta\mathcal{C}, \quad (3)$$

$$\alpha_d = \alpha_c + \Delta\alpha, s_d = s_c + \Delta s, \quad (4)$$

$$r_d = r_c \cdot [1, \Delta r_1, \Delta r_2, \Delta r_3], \quad (5)$$

where the quaternion multiplication \cdot is equivalent to multiplying the corresponding rotation matrices. With $[1, 0, 0, 0]$ representing the identity rotation, $r_d = r_c$ when $\delta r = \mathbf{0}$, preserving the original orientation for zero rotation offset.

We further employ a rigid deformation network to transform the non-rigidly deformed 3D Gaussians \mathcal{G}_d to the observation space \mathcal{G}_o . This is achieved via forward Linear Blend Skinning (LBS):

$$\mathbf{T} = \sum_{b=1}^B f_{\theta_r}(\mathcal{X}_d)_b \mathbf{B}_b, \mathcal{X}_o = \mathbf{T} \mathcal{X}_d, \quad (6)$$

$$\mathcal{R}_o = \mathbf{T}_{1:3,1:3} \mathcal{R}_d, \quad (7)$$

where \mathcal{R}_d is the rotation matrix derived from quaternion r_d , and \mathbf{B}_b represents the differentiable bone transformations.

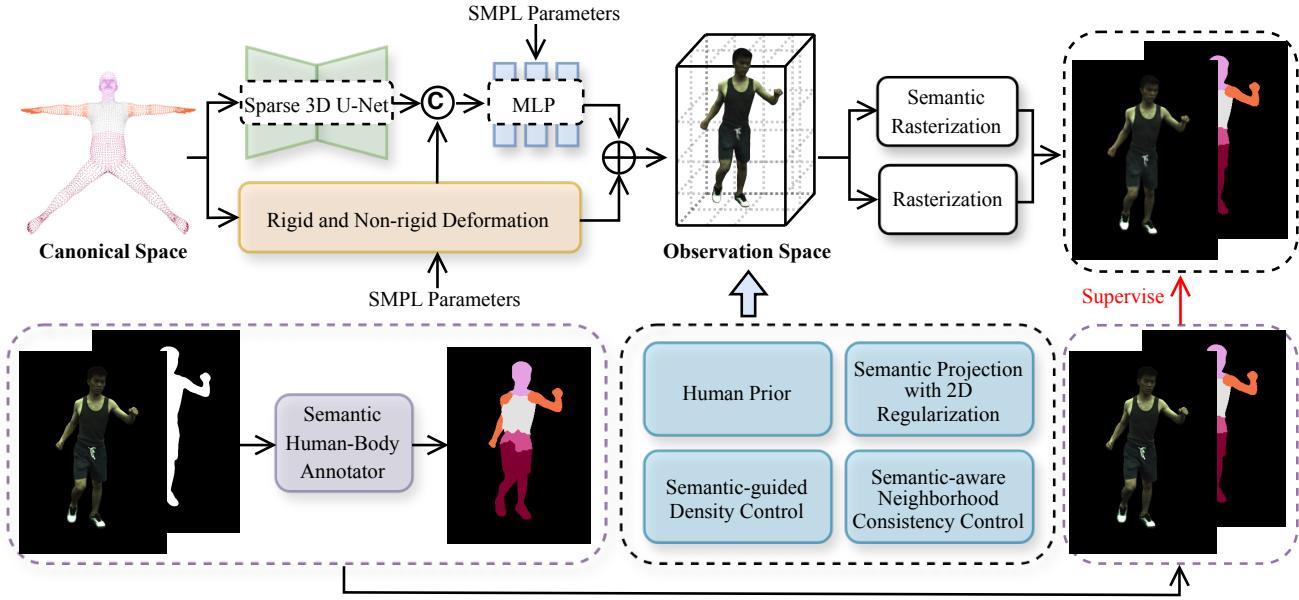


Figure 2. Our framework for creating photo-realistic animatable avatars from monocular videos. We initialize a set of 3D Gaussians in the canonical space by sampling 6,890 points from the SMPL model and assign the semantic attributes of Gaussians to each point. We first integrate a skeleton-driven rigid deformation and a non-rigid cloth dynamics deformation to deform human avatars from canonical space \mathcal{G}_c to observation space \mathcal{G}_o . Then, we introduce a Semantic Human-Body Annotator (SHA), which leverages SMPL’s human body semantic prior for efficient semantic labeling. These labels are used to guide the optimization of 3D Gaussian’s semantic attribute \mathcal{O} . We also propose a 3D topology and geometry-aware network to learn body topological and geometric associations and integrate them into learning the 3D deformation. To enhance semantic accuracy and render quality, we implement semantic projection with 2D regularization, semantic-guided density regularization and semantic-aware regularization with neighborhood consistency.

This step aligns the deformed Gaussians with the target pose in the observation space \mathcal{G}_o .

4.2. Semantic Human-Body Annotator

Most current animatable human avatar creation methods just use SMPL [22] model for its pose-aware shape priors, neglecting its inherent semantic information. We argue that semantic information contains topological relationships within human body which can improve rendering quality during complex motion deformations. We will further demonstrate this in the experimental Section. 5.3.

To achieve this, we deform the standard human body model from the SMPL model using the differentiable bone transformations \mathbf{B}_b as described in Eq. 6. Then, we use a custom point rasterizing function to render the deformed 3D SMPL model into an image m_i^p with a projection matrix from the dataset.

For each pixel in a foreground mask m_i , we employ the k-nearest neighbors (KNN) algorithm to identify the closest pixels in m_i^p . This process enables semantic-level annotation of body parts by transferring semantic labels from the SMPL model to the foreground mask m_i . The result is a semantically annotated mask m_i^s that accurately represents

the different regions of the human body. We formalize this Semantic Human Annotation (SHA) process as follows:

$$m_i^s = \mathcal{SH}\mathcal{A}(m_i, \mathbf{B}_b), \quad (8)$$

where $\mathcal{SH}\mathcal{A}$ denotes our Semantic Human-Body Annotator. We use the generated human body semantic labels m_s to supervise the Gaussian’s semantic attribute. (described in semantic projection with 2D regularization in Section. 4.4).

While there are pre-trained networks for human parsing, such as SChP [18] and Graphonomy [4], they are designed to segment both clothing and human body parts jointly. In contrast, our work focuses on leveraging semantic information to learn the topological relationships between different body parts. The objectives and tasks of these networks do not fully align with our needs, limiting their ability to model the geometric structure and topological connections of the human body. Their clothing segmentation can also introduce noise, hindering accurate body topology learning.

4.3. Topological and Geometric Feature Learning

To jointly learn and embed topology and geometry information to human avatar deformation, we propose a 3D topology- and geometry-aware network that effectively cap-

tures the human body's local topological and geometric structure in canonical space.

We treat 3D Gaussians as a point cloud. Point-level MLPs are limited by a small receptive field, which restricts their capability to capture the local geometric and topological features. Therefore, we employ sparse convolution [5] on sparse voxels to extract local topological and geometric features across varying receptive fields, following the method outlined in [23]. Given the position \mathcal{X}_c of the Gaussians \mathcal{G}_c as a point cloud, we initially convert it into voxels by partitioning the space using a fixed grid size v .

$$\mathbf{V} = \lfloor \mathcal{X}_c/v \rfloor, \quad (9)$$

where $\mathbf{V} \in \mathbb{R}^{M \times 3}$ and M is the number of voxels. We then construct a 3D sparse U-Net by stacking a series of sparse convolutions with skip connections to aggregate local features. The sparse 3D U-Net $f_{\theta_{unet}}$ takes \mathbf{V} and the semantic point-based features \mathcal{O} as input, and outputs topological and geometric features \mathbf{F}_v :

$$\mathbf{F}_v = f_{\theta_{unet}}(\mathbf{V}; \mathcal{O}). \quad (10)$$

We process the feature \mathbf{F}_v , the position \mathcal{X}_d of the deformed Gaussians \mathcal{G}_o , and pose latent code \mathcal{Z}_p in Eq. 2 through an fusion network $f_{\theta_{sr}}$:

$$\Delta(\mathcal{X}', s', r') = f_{\theta_{sr}}(\mathbf{F}_v; \mathcal{X}_d; \mathcal{Z}_p), \quad (11)$$

where $\Delta(\mathcal{X}', s', r')$ represents the final fused features. The deformed 3D Gaussian \mathcal{G}_o is then deformed by $\Delta(\mathcal{X}', s', r')$ following Eq. 3, 4, and 5.

4.4. Optimization

Unlike random initialization or Structure-from-Motion (SfM) initialization for Gaussian point clouds, we directly sample 6,890 points from the SMPL model [22] as our initial point cloud. Each Gaussian is then assigned semantic attributes based on the SMPL model's predefined semantic labels. During densification, newly created 3D Gaussian points inherit semantic attributes from their parent nodes.

Semantic projection with 2D regularization. We acquire rendered per-pixel semantic labels using the efficient Gaussian splatting algorithm following Eq. 1 as:

$$\mathcal{S} = \sum_{g \in \mathcal{N}} \mathcal{O}_g \alpha_g \prod_{j=1}^{g-1} (1 - \alpha_j), \quad (12)$$

where \mathcal{S}_k represents the 2D semantic labels of pixel k , derived from Gaussian point semantic attributes via α -blending (Eq. 1). Here, \mathcal{O}_g denotes the semantic attribute of the 3D Gaussian point g , and α_g is the influence factor of this point in rendering pixels. Upon calculating these labels, we obtain the results l_i^s and apply a BCE loss to regularize

the rendered semantic label l_i^s with semantic labels generated via SHA as follows:

$$\mathcal{L}_{semantic} = \mathcal{L}_{bce}(l_i^s, m_i^s). \quad (13)$$

Semantic-guided density regularization. Fuzzy geometric shapes often appear in local structures on the human surface, particularly in high-frequency areas like clothing wrinkles and muscle textures [37]. To improve the clarity and distribution of 3D Gaussians in these regions, we propose semantic-guided density regularization. We identify high-frequency nodes by assessing the average magnitude of structural differences between a selected node and all nodes within the same cluster. Nodes exhibiting the highest average magnitude of these differences are designated as high-frequency nodes.

$$H_m = \arg \max_{i \in C_m} \left\{ \frac{1}{|C_m| - 1} \sum_{j \in C_m \setminus \{i\}} d(A_i, A_j) \right\}, \quad (14)$$

where H_m is the high-frequency node in cluster C_m , A_i is the basic attribute of 3D Gaussian points (color, opacity, etc.), C_m represents the set of all points with semantic attribute m , $C_m \setminus \{i\}$ denotes the set of elements in C_m excluding the element i , and $d(\cdot, \cdot)$ is a dissimilarity measure between two points. To better capture and express these local structures of significant discrepancies, we perform densification operations on these 3D Gaussians, enhancing the local rendering granularity to focus on guiding the split and attribute optimization of Gaussian points in these areas.

Semantic-aware regularization with neighborhood consistency. We expect Gaussians that are in close proximity to exhibit similar semantic attributes, thereby achieving local semantic consistency in 3D space. The loss function for this semantic consistency constraint is as follows:

$$\mathcal{L}_{neighborhood} = \frac{1}{|N|} \sum_{m \in N} \sum_{n \in N_k(m)} D_{KL}(\mathcal{O}_m || \mathcal{O}_n), \quad (15)$$

where N represents the total number of Gaussian points, $N_k(m)$ contains the k nearest neighbors of 3D Gaussian point m in 3D space, \mathcal{O}_m and \mathcal{O}_n represent the predicted semantic attribute for point m and its neighbor n , respectively, and $D_{KL}(q_m || q_n)$ calculates the KL divergence between the predicted distributions of point m and its neighbor n .

Loss function. Our full loss function consists of a RGB loss \mathcal{L}_{rgb} , a mask loss \mathcal{L}_{mask} , a skinning weight regularization loss \mathcal{L}_{skin} , the as-isometric-as-possible regularization loss \mathcal{L}_{isopos} following [32], Semantic projection with 2D regularization $\mathcal{L}_{semantic}$, and Semantic-aware regularization with neighborhood consistency $\mathcal{L}_{neighborhood}$:

$$\begin{aligned} \mathcal{L}_{reconstruct} = & \mathcal{L}_{rgb} + \lambda_1 \mathcal{L}_{mask} + \lambda_2 \mathcal{L}_{SSIM} + \lambda_3 \mathcal{L}_{LPIPS} \\ & + \lambda_4 \mathcal{L}_{skin} + \lambda_5 \mathcal{L}_{isopos}. \end{aligned} \quad (16)$$

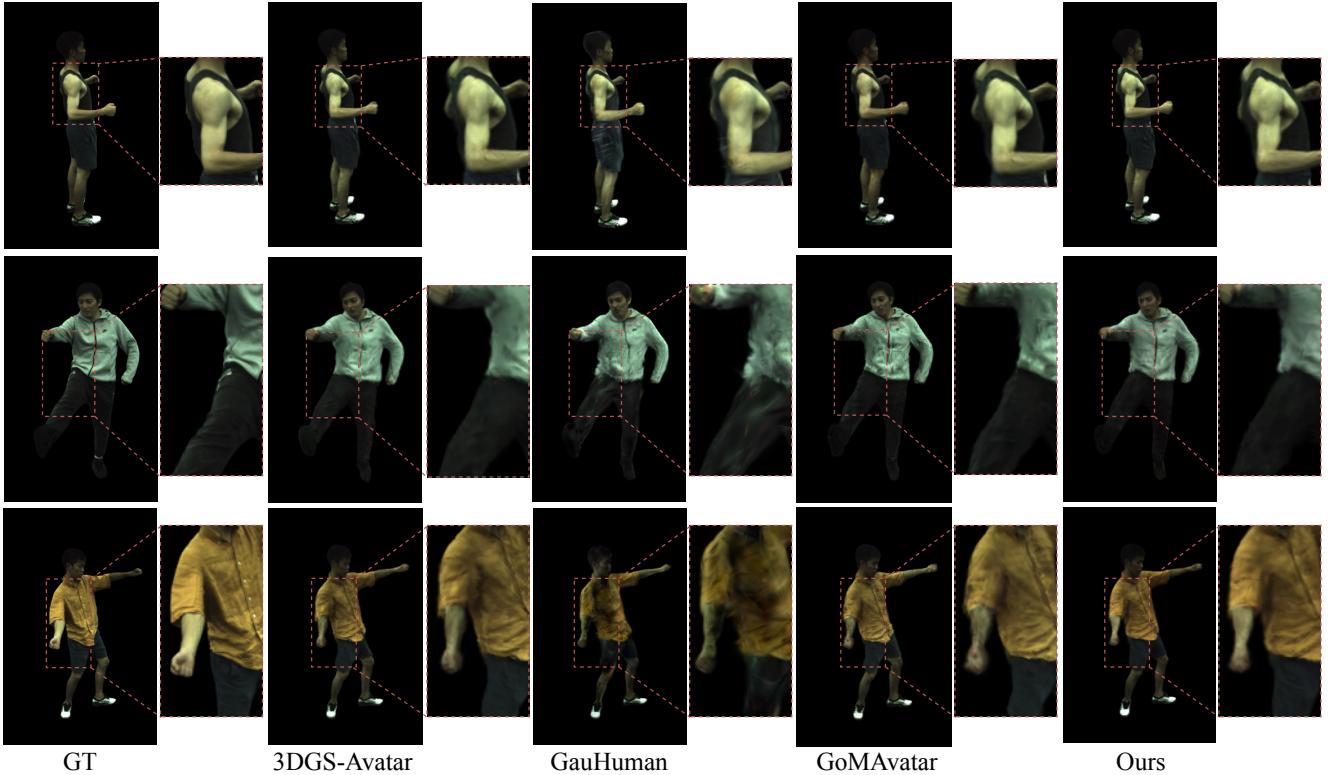


Figure 3. **Qualitative Comparison on ZJU-MoCap [31]**. We show that our SG-GS can produce realistic details in both rendered images and geometry, while other approaches struggle to generate smooth details.

The final loss function is:

$$\mathcal{L} = \mathcal{L}_{reconstruct} + \lambda_6 \mathcal{L}_{semantic} + \lambda_7 \mathcal{L}_{neighborhood}, \quad (17)$$

where λ 's are loss weights. For further details of the loss definition and respective weights, please refer to the Supp.Mat.

5. Experiment

In this section, we first compare SG-GS with recent SOTA methods [9, 10, 30–32, 38, 40, 46], demonstrating that our SG-GS achieves superior rendering quality. We then systematically ablate each component of the proposed method, showing their effectiveness in better rendering performance. All models are trained on one single NVIDIA RTX 3090 GPU. For further details of implementation, please refer to the Supp.Mat.

5.1. Dataset

ZJU-MoCap [31]. It records multi-view videos with 21 cameras and collects human poses using the marker-less motion capture system. We select six sequences (377, 386,

Table 1. **Quantitative Results on ZJU-MoCap [31]**. SG-GS achieves state-of-the-art performance across every method. The best and the second best results are denoted by pink and yellow. Frames per second (FPS) is measured on an RTX 3090. LPIPS* = LPIPS $\times 1000$.

Method:	PSNR↑	SSIM↑	LPIPS*↓	FPS
NeuralBody [31]	29.07	0.962	52.29	1.5
Ani-NeRF [30]	29.17	0.961	51.98	1.1
HumanNeRF [40]	30.24	0.968	31.73	0.3
MonoHuman [46]	29.38	0.964	37.51	0.1
DVA [33]	29.45	0.956	37.74	17
InstantAvatar [12]	29.73	0.938	64.41	4.2
3DGS-Avatar [32]	30.62	0.965	30.28	50
GauHuman [10]	30.79	0.960	32.73	180
GoMAvatar [39]	30.37	0.969	32.53	43
SG-GS	30.88	0.969	29.69	25

387, 392, 393, 394) from this dataset to conduct experiments. We also follow the same training/test split following [32, 40], i.e., one camera is used for training, while the remaining cameras are used for evaluation.

H36M [11]. It captures multi-view videos using four cameras and collects human poses with a marker-based motion

Table 2. **Quantitative Results on H36M [11]**. Our SG-GS still achieves superior performance compared to state-of-the-art methods on both training poses and novel poses.

Method:	Training Poses		Novel Poses	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
NARF [29]	23.00	0.898	22.27	0.881
NeuralBody [31]	22.89	0.896	23.09	0.891
Ani-NeRF [30]	23.00	0.890	22.55	0.880
ARAH [38]	24.79	0.918	23.42	0.896
3DGS-Avatar [32]	32.89	0.982	32.50	0.983
SG-GS	33.01	0.989	33.14	0.987

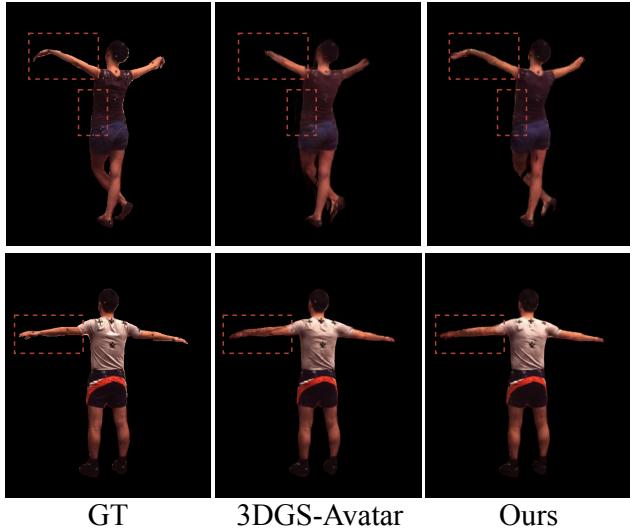


Figure 4. **Qualitative Comparison on H36M [11]**. By utilizing semantic information within human body, our SG-GS preserves better anatomical structures of the human body, producing high-quality results

Table 3. **Ablation Study on ZJU-MoCap [31]**. The proposed model achieves the lowest LPIPS, demonstrating the effectiveness of all components.

Method:	PSNR↑	SSIM↑	LPIPS*↓	FPS
Baseline	30.17	0.961	36.57	70
w/o topo-geo	30.64	0.970	32.75	70
mlp	30.54	0.967	32.06	60
w/o $\mathcal{L}_{semantic}$	30.67	0.966	29.99	25
w/o sgd	30.55	0.968	31.01	25
w/o $\mathcal{L}_{neighborhood}$	30.56	0.965	30.59	25
SG-GS	30.88	0.969	29.69	25

capture system. It includes multiple subjects performing complex actions. We select representative actions, split the videos into training and test frames, following ARAH [38], and perform experiments on sequences (S1, S5, S6, S7, S8, S9, S11). Three cameras are used for training and the remaining is selected for test.

5.2. Comparison with State-of-the-art Methods

We conduct comparative experiments against various state-of-the-art (SOTA) methods for human avatars, including NeRF-based methods such as NeuralBody [31], Ani-NeRF [30], HumanNeRF [40], and MonoHuman [46], as well as 3DGS-based methods such as 3DGS-Avatar [32], GauHuman [10], and GoMAvatar [39], under a monocular setup on ZJU-MoCap [31]. In Table 1, we evaluate the reconstruction quality using three different metrics: PSNR, SSIM, and LPIPS. Thanks to the LBS weight field and deformation field learned in HumanNeRF [40], 3DGS-Avatar [32], and GauHuman [10], these methods achieve comparable visualization results. In comparison, our proposed SG-GS achieves good performance in terms of PSNR and SSIM while significantly outperforming existing methods on LPIPS. Existing researches [32, 43] reach a consensus that *LPIPS provides more meaningful insights compared to the other metrics, given the challenges of reproducing exact ground-truth appearances for novel views*.

As shown in Fig. 3, our SG-GS method preserves sharper details compared to other methods. *Notably, our approach excels at capturing fine details in challenging areas such as clothing, where reconstruction is typically more difficult due to intricate textures.* By preserving these finer details, our method provides a more realistic and detailed reconstruction of clothing and other complex surfaces, significantly improving the overall quality and fidelity of the 3D human avatars. Please see our project website videos and supplementary material for more video visualization.

In addition, we also evaluate our SG-GS using the H36M [11] dataset. We report the quantitative results against NeRF-based methods such as NARF [29], NeuralBody [31], Ani-NeRF [30], and ARAH [38], as well as 3DGS-based methods such as 3DGS-Avatar [32] in Table 2. Our model outperforms both established NeRF-based methods and 3DGS-based methods. As shown in Fig. 4, due to the use of semantic information within human body, our SG-GS achieves better reconstruction of edge areas and preserves anatomical structures of the human body.

5.3. Ablation Study

In this section, we evaluate the effectiveness of our proposed modules through ablation experiments on the ZJU-MoCap [31] dataset. The average metrics over 6 sequences are shown in Table 3.

Topological and Geometric Feature Learning. As shown in Table 3, the proposed module significantly (topo-geo) enhances rendering performance. Though it slightly increase inference time, *the notable performance improvement justifies this additional cost*. A qualitative comparison in Fig. 5 further proves that Topological and Geometric Feature Learning maintains anatomical coherence during mo-

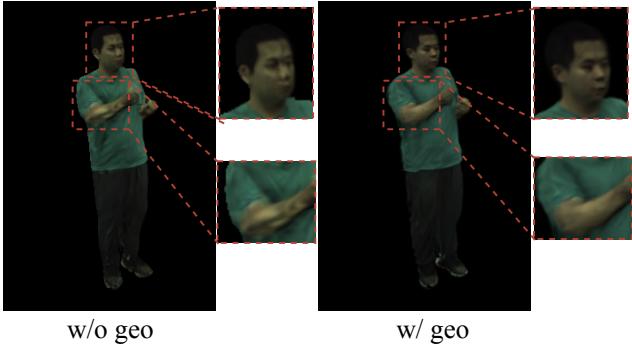


Figure 5. **Ablation Study** on Geometric and Semantic Feature Learning, which helps erase artifacts and learn fine details like cloth wrinkles and human face under novel views.

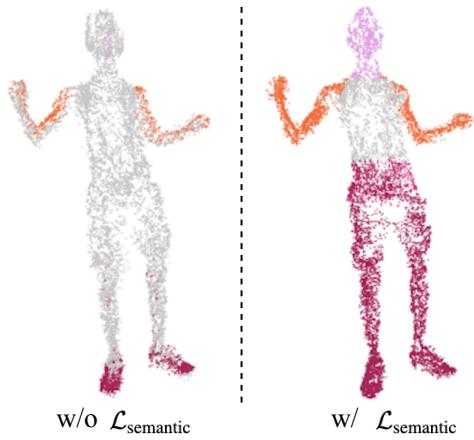


Figure 6. **Ablation Study** on semantic projection with 2D regularization, which enhances semantic accuracy. During pruning, most Gaussians are removed, leaving the remaining ones to default to torso semantics without our semantic supervision.

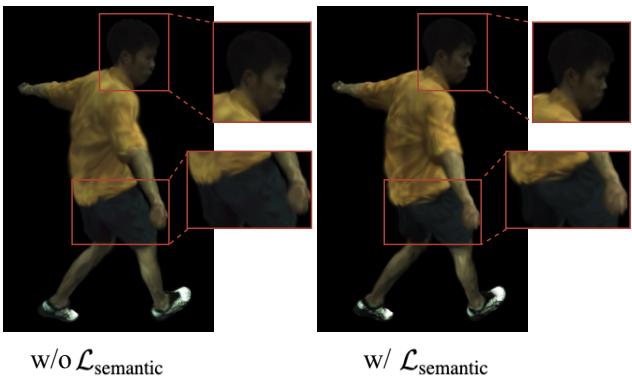


Figure 7. **Ablation Study** on semantic projection with 2D regularization, which keeps the topological consistency of the human body under novel poses.

tion and preserving fine details. We also conduct an experiment replacing the sparse 3D U-Net with an MLP (mlp in Table 3), which demonstrates point-level MLP is limited by a small receptive field, restricting capability to capture the local geometric and topological features.

Semantic Projection with 2D Regularization. This part utilizes semantic labels generated by SHA to supervise the semantic attributes of 3D Gaussians ($\mathcal{L}_{semantic}$). As shown in Fig. 6, semantic projection with 2D regularization substantially improves the semantic accuracy of 3D Gaussians. At the start of training, Gaussians are neither densified nor pruned [32], allowing their scale to grow. During pruning phase, most Gaussians are removed. As a result, most remaining Gaussians default to torso semantics without supervision. The results ($\mathcal{L}_{semantic}$ in Table 3) highlight the critical role of semantic information. This demonstrates that while the sparse 3D U-Net (introduced in Section 4.3) can capture geometric features with noisy semantic information and improve rendering quality, it still requires accurate semantic data to learn the topology to keep anatomical coherence of the human body, as shown in Fig. 7.

Semantic-Guided Density Regularization and Semantic-Aware Regularization with Neighborhood Consistency. Semantic-guided density regularization (sgd) enhances rendering quality by optimizing Gaussian density in areas with high discrepancy, while semantic-aware regularization with neighborhood consistency ($\mathcal{L}_{neighborhood}$) ensures that nearby Gaussians exhibit coherent semantic attributes, thus improving 3D semantic consistency. The improvements in rendering quality are validated by the results in Table 3.

6. Conclusion

In this paper, we propose SG-GS, which uses semantics-embedded 3D Gaussians to reconstruct photo-realistic human avatars. SG-GS first integrates a skeleton-driven rigid deformation and a non-rigid cloth dynamics deformation to deform human avatars. SG-GS then leverages SMPL’s human body semantic priors to acquire human body semantic labels, which are used to guide optimization of Gaussian’s semantic attribute. We also propose a 3D topology- and geometry-aware network to learn body geometric and topological associations and integrate them into the 3D deformation. We further implement three key strategies to enhance semantic accuracy and render quality: semantic projection with 2D regularization, semantic-guided density regularization, and semantic-aware regularization with neighborhood consistency. Extensive experiments demonstrate that SG-GS outperforms SOTA methods in creating photo-realistic avatars, further validating our hypothesis that integrating semantic priors enhances fine-detail reconstruction. We hope that our method will foster further research in high-quality clothed human avatar synthesis from monocular views.

Limitations. 1). SG-GS lacks the capability to extract 3D meshes. Developing a method to extract meshes from 3D Gaussians is an important direction for future research. 2). Topological and Geometric Feature Learning employs a sparse 3D U-Net, which is computationally intensive and may increase training and inference time to some extent.

References

- [1] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *Proc. of European Conf. on Computer Vision*, pages 333–350, 2022. [2](#)
- [2] Yue Chen, Xuan Wang, Xingyu Chen, Qi Zhang, Xiaoyu Li, Yu Guo, Jue Wang, and Fei Wang. UV volumes for real-time rendering of editable free-view human performance. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 16621–16631, 2023. [2](#)
- [3] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022. [2](#)
- [4] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. Graphonomy: Universal human parsing via graph transfer learning. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 7450–7459, 2019. [4](#)
- [5] Benjamin Graham and Laurens Van der Maaten. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307*, 2017. [5](#)
- [6] 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 *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 12858–12868, 2023. [2](#)
- [7] Jennifer Healey, Wang, and et al. A mixed-reality system to promote child engagement in remote intergenerational storytelling. In *International Symposium on Mixed and Augmented Reality Adjunct*, pages 274–279, 2021. [2](#)
- [8] Hsuan-I Ho, Lixin Xue, Jie Song, and Otmar Hilliges. Learning locally editable virtual humans. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 21024–21035, 2023. [2](#)
- [9] 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 *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 634–644, 2024. [2, 6](#)
- [10] Shoukang Hu et al. GauHuman: Articulated gaussian splatting from monocular human videos. In *cvpr*, pages 20418–20431, 2024. [1, 2, 6, 7](#)
- [11] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. on Pattern Anal. and Mach. Intell.*, 36(7):1325–1339, 2013. [6, 7](#)
- [12] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. InstantAvatar: Learning avatars from monocular video in 60 seconds. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 16922–16932, 2023. [6](#)
- [13] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *Proc. of European Conf. on Computer Vision*, pages 402–418. Springer, 2022. [2](#)
- [14] Yuheng Jiang, Zhehao Shen, Penghao Wang, Zhuo Su, Yu Hong, Yingliang Zhang, Jingyi Yu, and Lan Xu. HiFi4G: High-fidelity human performance rendering via compact gaussian splatting. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 19734–19745, 2024. [2](#)
- [15] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023. [2, 3](#)
- [16] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. HUGS: Human gaussian splats. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 505–515, 2024. [2](#)
- [17] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. GART: Gaussian articulated template models. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 19876–19887, 2024. [2](#)
- [18] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Trans. on Pattern Anal. and Mach. Intell.*, 44(6):3260–3271, 2020. [4](#)
- [19] Rui long Li, Julian Tanke, Minh Vo, Michael Zollhöfer, Jürgen Gall, Angjoo Kanazawa, and Christoph Lassner. TAVA: Template-free animatable volumetric actors. In *Proc. of European Conf. on Computer Vision*, pages 419–436, 2022. [2](#)
- [20] Zhe Li, Zerong Zheng, Yuxiao Liu, Boyao Zhou, and Yebin Liu. Posevocab: Learning joint-structured pose embeddings for human avatar modeling. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. [2](#)
- [21] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM transactions on graphics (TOG)*, 40(6):1–16, 2021. [2](#)
- [22] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *Acm Transactions on Graphics*, 34(248), 2015. [2, 3, 4, 5](#)
- [23] Zhicheng Lu, Xiang Guo, Le Hui, Tianrui Chen, Min Yang, Xiao Tang, Feng Zhu, and Yuchao Dai. 3d geometry-aware deformable gaussian splatting for dynamic view synthesis. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 8900–8910, 2024. [5](#)
- [24] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. 2024. [2](#)
- [25] Marko Mihajlovic, Yan Zhang, Michael J Black, and Siyu Tang. LEAP: Learning articulated occupancy of people. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 10461–10471, 2021. [3](#)

- [26] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2
- [27] Arthur Moreau, Jifei Song, Helisa Dhamo, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Human gaussian splatting: Real-time rendering of animatable avatars. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 788–798, 2024. 2
- [28] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 2
- [29] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *Proc. of IEEE Intl. Conf. on Computer Vision*, pages 5762–5772, 2021. 7
- [30] 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 *Proc. of IEEE Intl. Conf. on Computer Vision*, pages 14314–14323, 2021. 6, 7
- [31] 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 *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 2, 6, 7
- [32] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3DGSAvatar: Animatable avatars via deformable 3d gaussian splatting. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 5020–5030, 2024. 1, 2, 3, 5, 6, 7, 8
- [33] Edoardo Remelli, Timur Bagautdinov, Shunsuke Saito, Changlei 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. 6
- [34] Darius Rückert, Linus Franke, and Marc Stamminger. Adop: Approximate differentiable one-pixel point rendering. *ACM Transactions on Graphics (ToG)*, 41(4):1–14, 2022. 2
- [35] 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 *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1606–1616, 2024. 2
- [36] Huan Wang, Jian Ren, Zeng Huang, Kyle Olszewski, Menglei Chai, Yun Fu, and Sergey Tulyakov. R2l: Distilling neural radiance field to neural light field for efficient novel view synthesis. In *Proc. of European Conf. on Computer Vision*, pages 612–629. Springer, 2022. 2
- [37] Hongsheng Wang, Weiyue Zhang, Sihao Liu, Xinrui Zhou, Shengyu Zhang, Fei Wu, and Feng Lin. Gaussian control with hierarchical semantic graphs in 3d human recovery. *arXiv preprint arXiv:2405.12477*, 2024. 2, 5
- [38] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdf. In *Proc. of European Conf. on Computer Vision*, pages 1–19, 2022. 6, 7
- [39] 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. 1, 3, 6, 7
- [40] 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 *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 16210–16220, 2022. 2, 3, 6, 7
- [41] 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. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 20310–20320, 2024. 2
- [42] Hongyi Xu, Thimo Alldieck, and Cristian Sminchisescu. H-NeRF: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *Proc. of Advances in Neural Information Processing Systems*, 34:14955–14966, 2021. 2
- [43] Chen Yang, Sikuang Li, Jiemin Fang, Ruofan Liang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Gaussianobject: Just taking four images to get a high-quality 3d object with gaussian splatting. *arXiv preprint arXiv:2402.10259*, 2024. 7
- [44] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20331–20341, 2024. 2
- [45] Wang Yifan, Felice Serena, Shihao Wu, Cengiz Öztïrelı, and Olga Sorkine-Hornung. Differentiable surface splatting for point-based geometry processing. *ACM Transactions on Graphics*, 38(6):1–14, 2019. 3
- [46] Zhengming Yu, Wei Cheng, xian Liu, Wayne Wu, and Kwan-Yee Lin. MonoHuman: Animatable human neural field from monocular video. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 16943–16953, 2023. 2, 6, 7
- [47] Peter Zackariasson and Timothy L Wilson. The video game industry: Formation, present state, and future. 2012. 2
- [48] Haoyu Zhao, Xingyue Zhao, Lingting Zhu, Weixi Zheng, and Yongchao Xu. HFGS: 4d gaussian splatting with emphasis on spatial and temporal high-frequency components for endoscopic scene reconstruction. *arXiv preprint arXiv:2405.17872*, 2024. 2
- [49] Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J Black, and Otmar Hilliges. Pointavatar: Deformable point-based head avatars from videos. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 21057–21067, 2023. 2
- [50] 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