VGA: Reconstructing Vivid 3D Gaussian Avatars from Monocular Videos

Xinqi Liu
School of Robotics, Hunan University, China
liuxinqi@zju.edu.cn

Chenming Wu
Baidu Inc., China
wcm94@live.com

Abstract

In this paper, we present VGA, an innovative framework designed for the reconstruction of vivid and high-fidelity 3D Gaussian avatars, incorporating comprehensive body and fine-grained finger control derived from monocular video inputs. Our contributions are twofold, focusing on the enhancement of pose alignment precision and the refinement of 3D Gaussian representation. First, we introduce a pose refinement methodology that augments the accuracy of hand and foot poses through the utilization of normal maps and silhouette alignment, thereby facilitating accurate shape and appearance modeling. Second, we tackle the challenges of unbalanced aggregation and initialization bias inherent in 3D Gaussian representation by proposing a surface-guided re-initialization strategy. This approach guarantees a more homogeneous distribution of 3D Gaussians, ensuring their effective alignment with the avatar's potential surface, which in turn enhances rendering quality and stability under novel pose conditions. Extensive experimental evaluations demonstrate that our method achieves state-of-the-art performance in photo-realistic novel view synthesis, offering fine-grained control over body and finger movements. Both qualitative and quantitative analyses corroborate the robustness and expressiveness of our methodology, marking a substantial progression in the domain of 3D avatar reconstruction from monocular video.

Keywords: 3D Gaussian, Avatar, Human Reconstruction, Monocular Video.

1. Introduction

Reconstructing drivable and photorealistic avatars from monocular video sequences has attracted considerable attention owing to its extensive potential for transformative applications across diverse industries. This technology holds particular promise in domains such as e-commerce marketing (e.g., virtual try-on systems), live broadcasting, film production, and gaming, where it facilitates the creation of highly realistic and personalized avatars. By eliminating the reliance on expensive multi-camera setups or labor-intensive manual digital modeling, this approach offers a cost-effective and scalable solution for generating immersive digital representations, thereby driving innovation and efficiency in these fields.

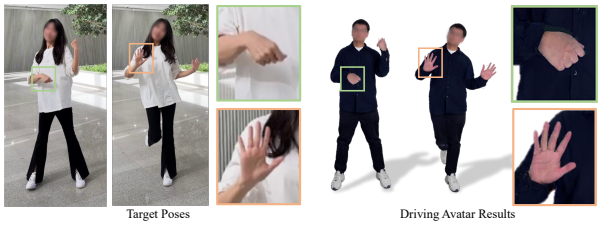
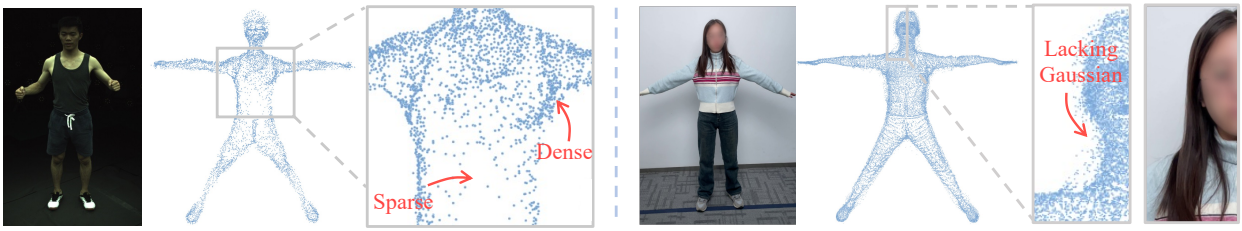


Figure 1. Our proposed VGA framework enables the robust reconstruction of 3D Gaussian avatars with comprehensive body and fine-grained finger control from monocular video inputs. This approach facilitates seamless animation of novel whole-body poses, making it highly suitable for applications in entertainment and live broadcasting. We blurred all faces for anonymity.

Traditionally, avatar reconstruction methods have relied on RGB-D cameras [9, 55, 56], multi-view dome acquisition systems [5, 8], or the labor-intensive process of digitally modeling human subjects. However, these approaches face challenges, including high costs in data acquisition and production, as well as difficulties in achieving photo-realistic rendering. The recent advent of Neural Radiance Fields (NeRF) [33, 2] has made it possible to generate cost-effective and high-quality 3D avatars [38, 36, 48, 15] using volume rendering techniques. By incorporating pose-conditioned MLP (Multi-Layer Perceptron) deformation fields, NeRF-based methods allow avatars to be controlled according to specific poses. Despite these advances, NeRF suffers from long training times and limited pose generalization, especially when dealing with significant pose deformations due to the implicit nature of the representation.

Recently, 3D Gaussian Splatting (3DGS) [20] has gained significant attention due to its explicit representation, which enables rapid convergence, real-time rendering, and high-fidelity expressiveness. Since its inception, numerous studies [64, 57, 40, 43, 12, 39] have utilized 3DGS to generate highly detailed 3D avatars by integrating it with parametric human models. Despite the impressive results achieved by 3DGS, existing approaches still encounter two major limi-



(a) Unbalanced Aggregation (b) Initialization Bias

Figure 2. The illustration of the widespread phenomena of unbalanced aggregation and initialization bias within the 3D Gaussian avatar reconstruction algorithms.

tations.

First, current 3D Gaussian avatars primarily emphasize body control and seldom support fine-grained finger interactions. This limitation largely stems from the significant alignment errors in finger regions observed in existing whole-body pose estimation methods [58, 28, 26], which fail to provide accurate shape and appearance guidance for the fingers. Second, existing 3D Gaussian avatar representations suffer from suboptimal Gaussian distributions, such as unbalanced aggregation and initialization bias (illustrated in Fig. 2). Specifically, unbalanced aggregation manifests as an excessive concentration of Gaussians in high-frequency texture regions, while texture-less areas are sparsely populated. Simultaneously, initialization bias arises when regions such as accessories or hair deviate from the initial shape and consequently receive insufficient Gaussian allocation. These issues lead to an uneven distribution that may perform adequately in static scenarios but introduce artifacts when avatars are driven into novel poses. Even minor deformations in the Gaussian distribution can significantly compromise visual quality during pose manipulation.

In this paper, we introduce VGA, a novel framework designed to address these challenges. To tackle the first issue, we incorporate normal priors and silhouette supervision to enhance the pose alignment accuracy of fingers and feet. For the second issue, we propose a surface-guided re-initialization mechanism that iteratively redistributes Gaussians near the explicit surface, ensuring a balanced and accurate Gaussian distribution. As a result, our method enables the reconstruction of avatars with both body and fine-grained finger control from monocular video inputs, as illustrated in Fig. 1. The key contributions of this paper are as follows:

- We present VGA, a novel framework for reconstructing 3D Gaussian avatars directly from monocular video. This approach advances beyond existing methods by eliminating the reliance on detailed annotations and demonstrating superior performance in reconstructing avatars across a diverse range of settings.

- We propose a pose refinement technique for avatar reconstruction, which significantly enhances the alignment accuracy of both body and finger poses, alongside a surface-guided Gaussian re-initialization mechanism that effectively mitigates issues of unbalanced aggregation and initialization bias.

- Extensive experiments have been conducted to validate the efficacy of our proposed method, demonstrating its capability to reconstruct avatars with both body and fine-grained finger control.

2. Related Work

2.1. Human Avatar Reconstruction

The task of reconstructing avatar models with accurate shapes and realistic appearances has been a long-standing focus in computer vision and graphics research. Early approaches primarily utilized RGB-D sensors [14, 34, 9, 55, 56, 5, 4] to capture the 3D shape of the target subject. These methods typically involved manually binding the reconstructed surface to a predefined skeleton to create an avatar model. However, the high cost of scanning equipment and the labor-intensive process of manual skin binding have hindered the widespread adoption of these techniques.

The advent of parametric human models such as SMPL [30] and SMPL-X [35] offered a more cost-effective solution for avatar reconstruction. These models allow for the creation of avatars using only RGB images, eliminating the need for expensive scan data. Many works [19, 22, 23, 29, 58, 28, 26, 63] have focused on estimating shape and pose parameters from images to drive parametric human body models, supporting both novel view rendering and novel pose generation. However, these methods tend to focus primarily on basic body shapes, often lacking fine user-specific details such as clothing and accessories.

More recently, a new wave of avatar reconstruction methods has emerged, taking parametric human body models as priors and enhancing them with additional details through techniques such as vertex offsets [31, 49], signed distance fields (SDF) [46, 42, 44, 61, 10, 51, 50], neural radiance fields (NeRF) [24, 38, 36, 48, 15], and 3D Gaussian points [64, 57, 40, 43, 12, 39, 18, 27]. These approaches

significantly improve the avatars realism by capturing user-specific details, such as clothing and facial features, resulting in more expressive and lifelike reconstructions.

Despite these advancements, the quality of these reconstructions is highly dependent on the accuracy of the estimated poses. Current end-to-end pose estimation methods [19, 22, 23, 29, 58, 28, 26, 63] excel in estimating body poses, but often struggle with finer details, such as finger and foot alignment. This limitation has led to avatar reconstruction methods [24, 38, 36, 64, 57, 40] that primarily support body-controllable avatars, while finer-grained controls, such as finger movements, remain challenging.

In contrast, our method introduces a pose refinement technique that utilizes predicted surface normals and silhouette cues to guide the reconstruction process. This significantly mitigates misalignment issues, particularly in the finger and foot regions, enabling the creation of highly expressive avatars with controllable body and finger movements from monocular videos. By addressing these limitations, our approach advances the field of avatar reconstruction, facilitating the generation of detailed and expressive avatars from simple video inputs.

2.2. Human Avatar Representation

The choice of human avatar representation plays a crucial role in determining both the fidelity and usability of reconstructed avatars. Historically, mesh-based approaches [30, 35, 31, 49, 13, 11] and point-cloud-based methods [32] have been widely favored for their simplicity and ease of use. However, these representations often lead to avatars that lack high-frequency geometric and texture details, limiting their realism.

The introduction of Neural Radiance Fields (NeRF) [33] has inspired a new wave of research due to the capability of photorealistic renderings. NeRF-based representations [24, 38, 36, 48, 15, 17] have achieved groundbreaking rendering quality in novel view synthesis. However, these methods typically require extended training times, often taking hours, and their rendering speed is far from real-time, limiting practical usability in interactive applications. Recently, there has been growing interest in 3D Gaussian Splatting (3DGS) [20], which offers a compelling balance between real-time rendering performance and photorealistic quality. This has sparked rapid progress in the field of 3D Gaussian-based avatar reconstruction [64, 57, 40, 43, 12, 39, 18, 27, 60], making it an active research topic. While 3DGS-based methods effectively leverage the power of Gaussians representation for avatar construction, they also inherit certain limitations, such as unbalanced aggregation and initialization bias, which can lead to noticeable artifacts during novel pose driving.

Our work builds on the strengths of the 3D Gaussian representation for avatar reconstruction but introduces a surface-guided Gaussian re-initialization mechanism to address these limitations. By mitigating unbalanced distribution and initialization bias, our approach enhances the driving ability and expressiveness of reconstructed avatars, resulting in higher-quality performance during pose manipulation.

3. Preliminary

3DGS [20] utilizes explicit 3D Gaussian as the fundamental rendering primitives. A 3D Gaussian is defined mathematically as the function $G(\boldsymbol{x})$, which can be expressed as:

$$G(\boldsymbol{x}) = \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\intercal}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \quad (1)$$

where $\boldsymbol{\mu}$ represents the spatial mean, and $\boldsymbol{\Sigma}$ is the covariance matrix. Each Gaussian is also associated with an opacity η and a view-dependent color $\boldsymbol{c}$, parameterized as spherical harmonic coefficients $\boldsymbol{f}$.

During rendering, the 3D Gaussians are projected onto the view plane via splatting. The 2D projection of each Gaussian is computed by transforming its 3D mean using the projection matrix, while the 2D covariance matrix is approximated as:

$$\boldsymbol{\Sigma}' = \boldsymbol{J}_g \boldsymbol{W}_g \boldsymbol{\Sigma} \boldsymbol{W}_g^{\top} \boldsymbol{J}_g^{\top}, \quad (2)$$

where $\boldsymbol{W}_g$ denotes the viewing transformation, and $\boldsymbol{J}_g$ is the Jacobian of the affine approximation of the perspective projection applied to the Gaussian.

To compute the final pixel color, alpha-blending is performed on the projected 2D Gaussians, layered sequentially from front to back. The pixel color C is given by:

$$C = \sum_{i \in N} T_i \alpha_i \boldsymbol{c}_i, \quad \text{where} \quad T_i = \prod_{j=1}^{i}(1-\alpha_j). \quad (3)$$

In this process, the opacity α_i for each Gaussian is computed by multiplying η (the opacity of the Gaussian) with its contribution based on the 2D covariance $\boldsymbol{\Sigma}'$ and the pixel coordinates in image space. The covariance matrix $\boldsymbol{\Sigma}$ is parameterized using a unit quaternion $\boldsymbol{q}$ and a 3D scaling vector $\boldsymbol{s}$, which ensures that the optimization process maintains a meaningful interpretation of the Gaussian's shape and orientation.

SMPL-X [35] is an extension of the original SMPL body model [30], designed to capture more detailed and expressive human deformations. SMPL-X extends the joint set of SMPL to include the face, fingers, and toes, enabling a more accurate representation of intricate body movements.

The SMPL-X model is defined by the function $M(\theta, \beta, \psi) : \mathbb{R}^{|\theta| \times |\beta| \times |\psi|} \longrightarrow \mathbb{R}^{3\mathrm{N}}$, where the parameters are: $\theta \in \mathbb{R}^{3K}$: Pose parameters (with K representing the

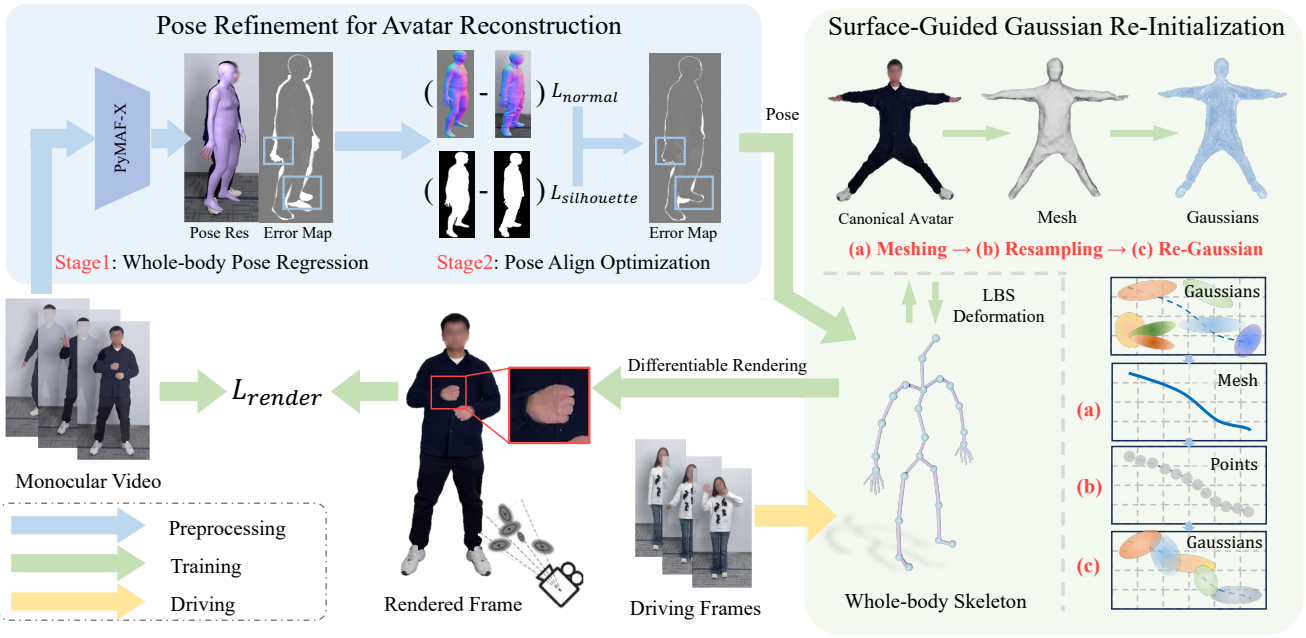


Figure 3. This framework constructs a body and finger controllable Gaussian avatar from a monocular video. It includes a pose refinement module to provide more accurate body and finger pose results to guide the correct shape and appearance reconstruction. The 3D Gaussian avatar is driven by a whole-body skeleton and obtains a more uniform and accurate Gaussian distribution through a surface-guided reinitialization mechanism. The avatar can adapt to new poses from videos or generated pose sequences.

number of body joints), $\beta \in \mathbb{R}^{|\beta|}$: Shape parameters for the face and hands, $\psi \in \mathbb{R}^{|\psi|}$: Facial expression parameters. The function $M(\beta, \theta, \psi)$ is defined as:

$$M(\beta, \theta, \psi) = W\left(T_p(\beta, \theta, \psi), \mathcal{J}(\beta), \theta, \mathcal{W}\right), \quad (4)$$

where $T_p(\beta, \theta, \psi)$ represents the human body mesh in a canonical pose, $\mathcal{J}(\beta)$ is a regression matrix predicting joint locations, W is the function responsible for applying pose-dependent transformations, $\mathcal{W}$ denotes the precomputed skinning weights. For further details, refer to [35].

4. Proposed Method

We illustrate the avatar reconstruction pipeline in Fig. 3, which consists of three main components: (1) a body and finger drivable Gaussian avatar representation (Sec. 4.1), (2) pose refinement for improved whole-body pose accuracy and reconstruction quality (Sec. 4.2), and (3) surface-guided Gaussian re-initialization to enhance the rationality of Gaussian distribution (Sec. 4.3).

4.1. The 3D Gaussian Avatars Representation

We represent 3D Gaussian avatar model using two key components, denoted as $\{G, B\}$. The first component, G, is a collection of 3D Gaussian that captures the shape and appearance characteristics of the target subject. The second component, B, is a skeleton model that enables avatar manipulation.

The Gaussian G are initialized in the canonical pose space (i.e., T-pose) by utilizing the vertices of the SMPL-X model. To facilitate deformations and pose variations, we use the SMPL-X skeleton structure [35], which consists of $K = 55$ joints. These joints are divided as follows: 22 joints control the body pose, 15 joints for each hand (left and right), and 3 joints for the head. Pose transformations are applied using the joint hierarchy, and for each pose θ, we compute the pose transformation matrix $\mathcal{T}(\theta)$ of each joint.

For Gaussian, the pose transformation $\mathcal{A}$ is computed based on the nearest $P = 4$ joints using the following equation:

$$\mathcal{A}(\theta) = \sum_{p=1}^{P} \mathcal{W}_p(\boldsymbol{\mu}) \mathcal{T}(\theta), \quad (5)$$

where $\mathcal{W}_p(\boldsymbol{\mu})$ is the skinning weight of the Gaussian $\boldsymbol{\mu}$, obtained by referencing the skinning of the nearest vertex in the SMPL-X model. The deformation of a Gaussian from the canonical pose to the target pose θ is expressed as:

$$\boldsymbol{\mu}_\theta = \mathcal{A}_{\text{rot}}(\theta)\boldsymbol{\mu}' + \mathcal{A}_t, \quad \boldsymbol{R}_\theta = \mathcal{A}_{\text{rot}}(\theta)\boldsymbol{R}, \quad (6)$$

where $\mathcal{A}_{\text{rot}}(\theta)$ is the rotation component, and $\mathcal{A}_t$ is the translation component. The rotation matrix $\boldsymbol{R}_\theta$ of the Gaussian is calculated based on its quaternion $\boldsymbol{q}$. To handle non-rigid local deformations, such as dynamic wrinkles in clothing, we introduce an adjusted Gaussian position $\boldsymbol{\mu}'$. This adjustment incorporates a pose-conditioned residual added to the

original Gaussian position, expressed as:

$$\boldsymbol{\mu}' = \boldsymbol{\mu} + \text{MLP}(\theta). \tag{7}$$

4.2. Pose Refinement for Avatar Reconstruction

Creating a high-quality 3D Gaussian avatar relies heavily on the precision of pose estimation from input images. Accurate pose is essential for properly aligning the 3D Gaussian avatar with the subject in images. However, current whole-body pose estimation methods [58, 28, 26] struggle to consistently align finger and foot areas, limiting existing 3D Gaussian-based avatar methods [64, 57, 40, 43] to body-controllable reconstructions without fine-grained finger control. To address this limitation, we propose a two-stage method that improves the whole-body pose accuracy.

In the first stage, we obtain an initial pose estimation by applying an off-the-shelf whole-body pose estimation network, $\mathcal{E}$ [58], to the input video sequences I. This process yields the SMPL-X pose parameters θ, shape parameters β and camera parameters Π, providing a coarse whole-body pose estimation:

$$\theta^{\text{stage1}}, \Pi = \mathcal{E}(I). \tag{8}$$

However, these initial poses often exhibit noticeable misalignment, particularly in the finger regions, as shown in Fig. 11.

In the second stage, we refine the pose estimation by incorporating constraints from normal maps and silhouettes, aiming to improve the alignment of the SMPL-X model with the subject in the images. The key insights are twofold: (1) normal maps effectively guide the alignment of the entire body, particularly the fingers and feet, and (2) silhouettes provide boundary conditions that ensure accurate positioning of the fingers and feet based on the observed image data.

Here, it is worth mentioning that we did not use 2D key points because the mainstream 2D keypoint detector [52] will have obvious estimation errors when fingers interact, as shown in Fig. 15, which deteriorates the final pose results.

For a given input image, we use the Segment Anything Model (SAM) [21] to obtain the subject's mask, which serves as the predicted silhouette S^{pred}, and we use ICON [51] to predict the normal map N^{pred}. We define the following loss function to optimize the pose:

$$\mathcal{L}_{\text{pose}} = \underbrace{\left|N - N^{\text{pred}}\right|}_{\mathcal{L}_{\text{normal}}} + \lambda_1 \underbrace{\left|S - S^{\text{pred}}\right|}_{\mathcal{L}_{\text{silhouette}}} + \lambda_2 \underbrace{\sum_{i=1}^{K} \omega_i \left(\theta_i - \theta_i^{\text{stage1}}\right)}_{\mathcal{L}_{\text{regular}}}, \tag{9}$$

where λ_1 and λ_2 are weights for the different loss terms. In our experiments, we empirically set $\lambda_1 = 5.0$ and $\lambda_2 = 0.5$.

The loss function consists of three terms: The first term $\mathcal{L}_{\text{normal}}$ enforces consistency between the normal map N rendered from the SMPL-X model using the current pose parameters θ and the predicted normal map N^{pred} from the image. The second term $\mathcal{L}_{\text{silhouette}}$ ensures the alignment between the rendered silhouette S and the predicted silhouette S^{pred} of the subject. The third term $\mathcal{L}_{\text{regular}}$ regularizes the pose θ to remain close to the initial estimation θ^{stage1}, with a weighting mechanism ω_i applied to each joint based on its distance from the root joint, giving lower weights to joints further away.

4.3. Surface-Guided Gaussian Re-Initialization

To address the issues of unbalanced aggregation and initialization bias that degrade the performance of 3D Gaussian avatars, we introduce a surface-guided Gaussian re-initialization method. The problem of unbalanced aggregation arises due to the cloning and splitting operations in 3D Gaussian Splatting (3DGS), which tend to over-propagate Gaussian in high-frequency texture areas, leading to local clustering. Additionally, the initialization of 3D Gaussian is prone to bias, which further exacerbates artifacts in the avatar model.

Existing 3D Gaussian avatar methods typically initialize Gaussian using the SMPL model [30, 35]. While this approach works well for subjects with tight clothing, it faces challenges when dealing with subjects wearing loose garments or with long hair. In this case, a limited number of Gaussian are propagated outside the body to describe these extra-body features. When faced with deformation, these sparse Gaussian will suffer from noticeable blurring and artifacts in the rendered result.

Our key insight is to impose additional constraints on the Gaussian, ensuring they are uniformly distributed near the surface of the subject. To achieve this, we propose a surface-guided Gaussian re-initialization method (illustrated in Fig. 4). This method consists of three iterative operations: *Meshing, Resampling, and Re-Gaussian*. These three steps are applied iteratively 2-3 times to gradually guide the Gaussian avatar toward the real surface of the human body.

Meshing. We perform spherical shell surface reconstruction [6] to generate the surface mesh of the avatar, using the outermost Gaussian to represent the surface of the subject.

Resampling. To refine the mesh, we apply Laplacian smoothing, introducing surface smoothness as a prior. Following this, we perform curvature-based uniform sampling on the mesh, generating new Gaussian that are evenly distributed.

Re-Gaussian. For each resampled new Gaussian, we identify its K-nearest raw Gaussian and inherit their opacity η and spherical harmonic coefficients $\boldsymbol{f}$. The rotation $\boldsymbol{R}$ and scaling $\boldsymbol{s}$ attributes are reinitialized based on the reconstructed surface normal vectors and average vertex dis-

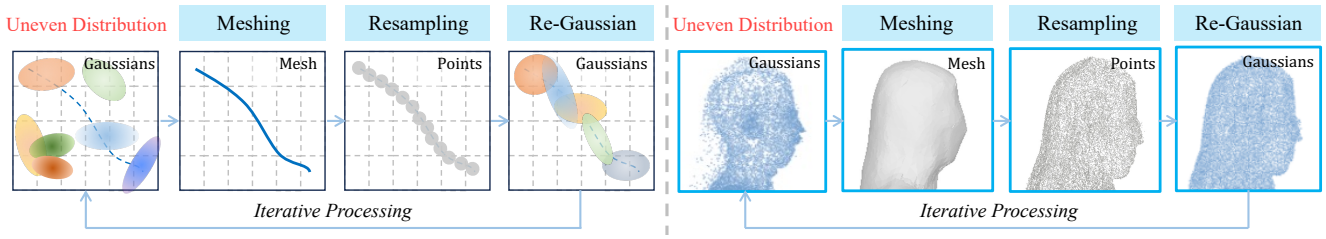


Figure 4. The surface-guided re-initialization mechanism uses the three operations of *Meshing, Resampling, and Re-Gaussian* to redistribute unevenly Gaussian points near the real surface, thereby enhancing the stability of the avatar in novel poses.

Figure 5. Rendered frames of our reconstructed Gaussian avatar from novel views.

Figure 6. Multiple reconstructed avatars demonstrate pose-driven movements using videos.

tances to avoid falling into local minima.

By applying this surface-guided re-initialization, the iteratively updated Gaussian is progressively aligned with the surface of the human body, leading to a more accurate and artifact-free reconstruction.

4.4. Differentiable Rendering Loss Function

We use the SMPL-X skeleton transformation (as outlined in Eq. 5 and Eq. 6) to drive the Gaussian avatar from the canonical pose space to the image pose space, optimizing it via differentiable rendering. Given the rendered image C and the input image I, we compute the following loss terms: reconstruction loss $\mathcal{L}_{\text{recon}}$, perceptual loss $\mathcal{L}_{\text{perceptual}}$, and residual regularization $\mathcal{L}_{\text{residual}}$. The total loss function is defined as:

$$\mathcal{L}_{\text{render}} = \underbrace{|C - I|}_{\mathcal{L}_{\text{recon}}} + \lambda_3 \underbrace{|\text{VGG}(C) - \text{VGG}(I)|}_{\mathcal{L}_{\text{perceptual}}} + \lambda_4 \underbrace{|\text{MLP}(\theta)|}_{\mathcal{L}_{\text{residual}}}, \quad (10)$$

where λ_3 and λ_4 are weights for the perceptual and residual regularization terms, respectively. In our experiments, they are empirically set $\lambda_3 = 0.1$ and $\lambda_4 = 0.5$.

The reconstruction loss $\mathcal{L}_{\text{recon}}$ ensures that the rendered image C is consistent with the input image I. The perceptual loss $\mathcal{L}_{\text{perceptual}}$ enforces consistency between the encoded features of C and I, which helps to capture high-frequency appearance details. Here, VGG$(*)$ represents the high-dimensional image features extracted using a pre-trained VGG network [45]. The residual regularization term $\mathcal{L}_{\text{residual}}$ regularizes the pose-conditioned residual to remain close to small values, preventing excessive interference with the Gaussian avatar.

This combination of losses allows for effective optimization of the avatar to match the input image both visually and in terms of detailed feature representation.

Table 1. Quantitative comparison on the ZJU-MoCap [38] dataset. LPIPS* = $10^3 \times$ LPIPS. Pink highlights the best, and orange highlights the second best.

Methods	PSNR↑	SSIM↑	LPIPS*↓	Training time
HumanNeRF [48]	30.66	0.9690	33.38	∼ 10 h
AS [37]	30.38	0.9750	37.23	∼ 10 h
AN [36]	29.77	0.9652	46.89	∼ 10 h
Neural Body [38]	29.03	0.9641	42.47	∼ 10 h
DVA [41]	29.45	0.9564	37.74	∼ 1.5 h
NHP [24]	28.25	0.9551	64.77	∼ 1 h tuning
PixelNeRF [54]	24.71	0.8920	121.86	∼ 1 h tuning
Instant-NVR [7]	31.01	0.9710	38.45	∼ 5 min
Instant-Avatar [16]	29.73	0.9384	68.41	∼ 3 min
GauHuman [12]	31.34	0.9650	30.51	∼ 1 min
GART [25]	32.22	0.9771	29.21	∼ 2.5 min
Ours	32.45	0.9773	26.94	∼ 1 min

5. Experiments

5.1. Setup and Datasets

Our approach is based on the PyTorch framework and utilizes the Adam optimizer. The model is optimized for $3,000$ steps, with the learning rate for the Gaussian's position, rotation, scale, opacity, and spherical harmonic coefficient all set similarly to [25]. The experiment is conducted on an NVIDIA A100 GPU, with pose refinement requiring 10 seconds per frame.

People-Snapshot [1] is a monocular video dataset, which contains 8 subjects wearing various clothing and performing self-rotation motions in front of a fixed camera,

Figure 7. Qualitative comparison on the ZJU-MoCap [38] dataset.

maintaining an A-pose during the recording.

ZJU-MoCap [38] is a multi-view dataset that includes dynamic videos of 6 subjects captured by over 20 simultaneous cameras.

ZJU-MoCap and People-Snapshot lack diversity in finger pose, therefore, we introduce the VGA-Snapshot dataset.

VGA-Snapshot dataset is intended for evaluating body and finger reconstruction from monocular videos. It includes self-rotation videos and carefully designed finger movement videos of 7 subjects. Each data frame provides 4K resolution RGB images, precise masks, and corresponding refined SMPL-X pose parameters. Additionally, our subjects exhibit challenging features such as shawl-length hair, which are absent in current public datasets. More details are presented in the supplementary materials.

5.2. Baselines and Evaluation Metrics

According to the differences in avatar representation, baseline methods can be categorized into NeRF-based and 3D Gaussian-based approaches. NeRF-based methods such as HumanNeRF [48], AS [37], AN [36], Neural Body [38], DVA [41], NHP [24], PixelNeRF [54], Instant-NVR [7], and Instant-Avatar [16] employ different variations of the NeRF representation for avatar reconstruction. HumanNeRF, AS, AN, Neural Body, and DVA utilize a naive NeRF representation combined with locally encoded human body features. NHP and PixelNeRF use a generalizable NeRF representation, reducing training time through finetuning. Instant-NVR and Instant-Avatar enable NeRF representation for minute-level training and real-time rendering using grid hashing. Gaussian-based methods, including GauHuman [12] and GART [25], represent the current state-of-the-art approaches for Gaussian avatar reconstruction.

For quantitative evaluation, we use three metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) [59]. PSNR is used to evaluate pixel-level errors between avatar-rendered images and ground-truth images. SSIM is employed to assess structure-level errors, while LPIPS evaluates perceptual errors.

5.3. Qualitative Experiments

Three qualitative experiments are conducted to demonstrate the effectiveness of our proposed method as follows.

First, we showcase the capability of our method to render reconstructed avatars from various novel viewpoints, as shown in Fig. 5. This demonstrates the ability to reconstruct complete and visually accurate avatar models from monocular videos, capturing photorealistic effects from different viewpoints. Additionally, we utilize a video captured in natural settings to estimate its SMPL-X pose as a driving sequence, enabling whole-body pose control and motion reproduction for the avatar, as depicted in Fig. 6. Our reconstructed avatar maintains fidelity in details and accurately represents finger movements when driven to unseen poses, highlighting the strong generalization ability.

Second, we evaluate our method against multiple baseline methods on the ZJU-MoCap and People-Snapshot [1] dataset, as shown in Fig. 7 and Fig. 8. Compared to AS [37], NB [38], NHP [24], PixelNeRF [54], and Instant-NVR [7], our method demonstrates superior accuracy in capturing

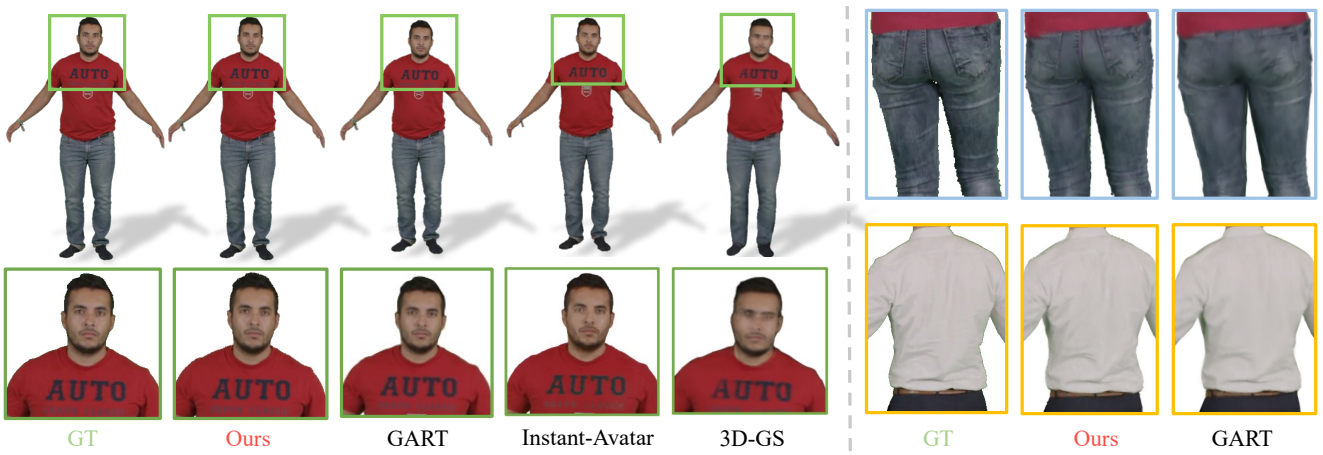


Figure 8. Qualitative evaluation on the People-Snapshot [1] dataset, comparing our method with multiple baseline approaches.

Table 2. Quantitative comparison on the People-Snapshot [1] dataset.

Methods	male-3-casual			male-4-casual			female-4-casual		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
3D-GS [20]	26.60	0.9393	0.0820	24.54	0.9469	0.0880	25.74	0.9364	0.0750
Neural Body [38]	24.94	0.9428	0.0326	24.71	0.9469	0.0423	24.37	0.9451	0.0382
Anim-NeRF [3]	12.39	0.7929	0.3393	13.10	0.7705	0.3460	12.31	0.8089	0.3344
Instant-Avatar [16]	29.65	0.9730	0.0192	27.97	0.9649	0.0346	28.92	0.9692	0.0180
GART [25]	30.40	0.9769	0.0377	27.57	0.9657	0.0607	29.23	0.9721	0.0378
Ours	30.82	0.9808	0.0199	27.62	0.9742	0.0351	29.27	0.9743	0.0213

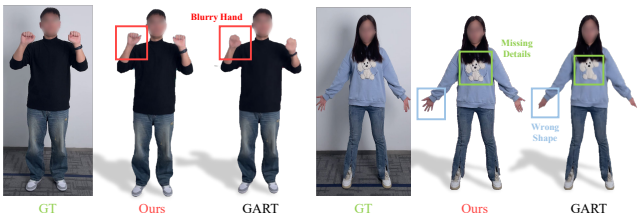


Figure 9. Comparison of results between ours and GART [25] on VGA-Snapshot dataset.

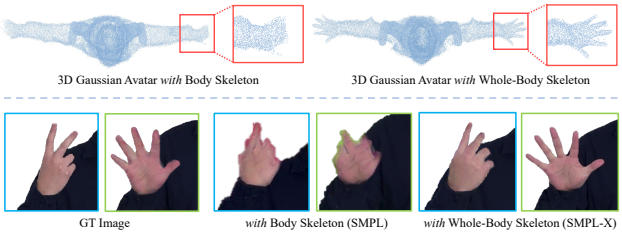


Figure 10. Ablation study investigating the impact of the body or whole-body skeleton. Top: 3D Gaussian avatar visualization; Bottom: Zoom-in rendered images.

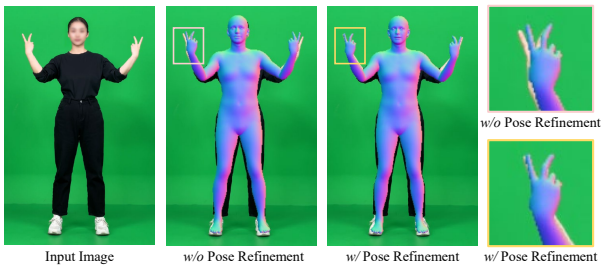


Figure 11. The qualitative ablation study evaluating the effectiveness of the pose refinement in improving pose accuracy.

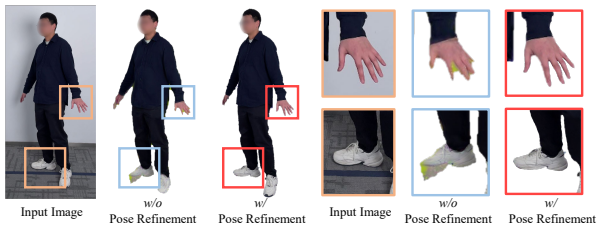


Figure 12. The qualitative ablation study evaluating the effectiveness of the pose refinement in improving avatar accuracy.

shape and appearance from novel views. Compared to HumanNeRF [48] in Fig. 7, our method achieves a visually comparable performance with significantly reduced time consumption. Compared to GART [25] and Instant-Avatar [16] in Fig. 8, our method captures more details. These results highlight our method's advantages in realism and efficiency.

Third, we compare our approach with GART [25] on the VGA-Snapshot dataset, as depicted in Fig. 9. GART [25], which uses SMPL as the skeleton without finger pose guidance, shows incorrect shapes and blurred finger. In contrast, our method incorporates the SMPL-X skeleton and incorpo-

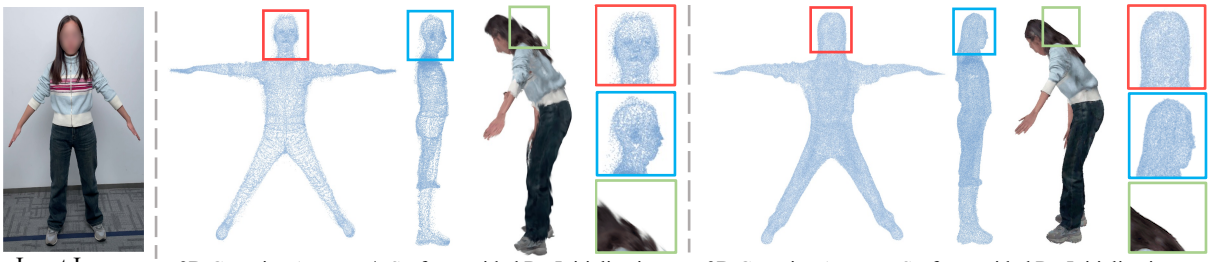


Figure 13. Ablation study on the utilization of the surface-guided re-initialization.

Table 3. Quantitative comparison between ours and GART [25] on VGA-Snapshot.

Methods	PSNR↑	SSIM↑	LPIPS*↓
GART [25]	31.61	0.9907	38.52
Ours	**32.36**	**0.9912**	**27.24**

Table 4. Quantitative ablation study on main technical components.

Methods	PSNR↑	SSIM↑	LPIPS*↓
w/o Pose Refinement	26.76	0.978	51.20
w/o Finger Skeleton	28.79	0.982	34.45
w/o Surface-guided Re-Initialization	30.48	0.989	35.30
Ours (Full)	**32.22**	**0.989**	**31.80**

Table 5. A quantitative study evaluating the effectiveness of the pose refinement module in improving pose accuracy, conducted on the 3DPW[47] dataset.

Methods	PVE↓	MPJPE↓
w/o Pose Refinement	91.3	78.0
w/ Pose Refinement (Ours)	90.6	77.3

rates finger guidance, enabling whole-body pose control for the avatar and providing more precise details.

5.4. Quantitative Results

In Tabel 1 and Tabel 2, we compare our method with baseline methods on the ZJU-MoCap and Peopel-Snapshot datasets. Our method notably outperforms various NeRF-based methods, is on par with GART [25] in terms of PSNR and SSIM, and significantly outperforms in LPIPS. These results align with qualitative observations. Given the absence of finger pose changes in the above two datasets, we compare our method with GART [25] on the VGA-Snapshot dataset. The comparative results are detailed in Table 3. These results indicate that our method outperforms GART [25] across all metrics, consistent with the qualitative assessment. These observations indicate that our method attains superior avatar reconstruction performance.

5.5. Ablation Study

This section examines the influence of key technical components, namely the whole-body skeleton, pose refinement, and surface-guided re-initialization.

Fig. 10 illustrates the effect of incorporating whole-body skeleton on the reconstructed avatar. Without the whole-body skeleton (body skeleton), Gaussian struggles to capture the finger shape accurately, leading to blurred images.

For the pose refinement, Fig. 11 illustrates its impact on whole-body pose accuracy, and Fig. 12 demonstrates its influence on the final avatar results. The comparison primarily focuses on the avatar results obtained through solely one-stage pose estimation. The findings reveal that relying solely on the existing whole-body pose estimation (w/o pose refinement) fails to completely align the subject's pose in the image, particularly in finger region. This inadequacy leads to significant artifacts in the learned avatar. However, with increased pose refinement, the avatar acquires more accurate pose guidance, effectively mitigating this issue.

Fig. 13 illustrates the impact of employing surface-guided re-initialization. Without surface-guided re-initialization, Gaussian are only sparsely allocated in the external areas of the naked body (such as hair), making the avatar susceptible to noticeable artifacts undergoing new pose drives. Conversely, utilizing surface-guided re-initialization effectively redistributes the avatar's Gaussian, ensuring a more even distribution across the real human body surface, thus enhancing the stability of new pose results.

Table 5 investigates the pose refinement module from the perspective of 3D pose estimation accuracy. By conducting comparisons on the 3DPW [47] dataset using Per Vertex Error (PVE) and Mean Per Joint Position Error (MPJPE) metrics, it quantitatively demonstrates the effectiveness of the pose refinement. Table 4 conducts a quantitative ablation analysis of the main technical modules from the perspective of avatar rendering quality. In alignment with the qualitative analysis, it demonstrates that each technical component contributes positively to the final body-finger avatar reconstruction results.

Fig. 14 showcases the Gaussian avatar models and recon-

Figure 14. Reconstructed Gaussian Avatar and mesh.

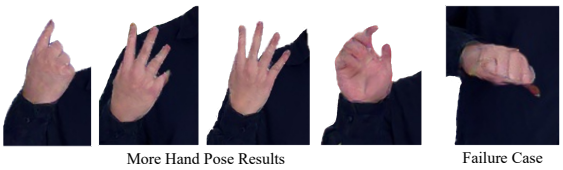


Figure 15. More finger-driven results at new target finger poses.

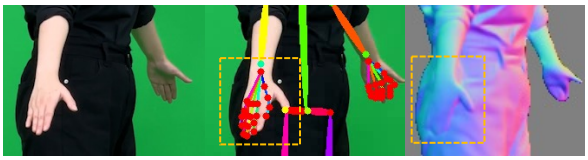


Figure 16. Comparison of DWPose [52] 2D keypoints and normal map during pose refinement.

structed meshes, highlighting the variations in body shapes, textures, and clothing. The results demonstrate the robustness of our method in handling a wide range of body shapes and accurately reconstructing the corresponding shapes and appearances that align with the observed images. This highlights the exceptional resilience of our approach, showcasing its ability to consistently produce accurate and visually coherent reconstructions across diverse body variations.

Fig. 15 shows more finger-driven results under the new target finger pose, which shows that our Avatar supports fine-grained finger pose control and stronger expressiveness.

Fig. 16 compares the difference between using DWPose [52] 2D keypoints or normal maps as supervision in the pose refinement. The former is adopted by InstantAvatar [16] and AvatarReX [62]. It can be seen that 2D keypoints are prone to misestimation in the finger region, which will cause the optimized pose misalignment. In contrast, using normals as targets is more stable.

Finally, we compared the time cost of the pose refinement process with ICON [51], as shown in Table 6. We used a distributed design and improve calculation process to achieve 10s per frame optimization and 10min of total optimization time. However, ICON took 25s per frame to optimize the pose and 330min of total optimization time. Obviously, our pose optimization process has better performance.

6. Conclusion

In this paper, we presented a method for body- and finger-drivable 3D Gaussian avatar reconstruction from monocular videos. Our approach incorporates pose refinement to enhance the accuracy of finger and foot alignment, enabling the avatar to better capture the subject's shape and appearance. Additionally, we introduced a surface-guided Gaussian re-initialization mechanism to mitigate issues related to unbalanced aggregation and initialization bias derived from Gaussian representation. We hope that this work will contribute to more lifelike and accurate avatar reconstructions in future developments.

Limitation. Although our method has successfully achieved body- and finger-controllable avatar reconstruction, further increasing facial expression controllability remains a challenge. Introducing learnable blendshapes may be a feasible way. Second, our pose refinement module relies on the accuracy and robustness of normal map estimation methods. Adopting state-of-the-art techniques such as StableNormal [53] may better address this challenge. Third, our current Avatar reconstruction method primarily focuses on monocular human video collected from controllable indoor scenes. For monocular video inputs captured in outdoor scenes with complex lighting, self-shadows, and dynamic backgrounds, the reconstructed Avatar may suffer from artifacts. Exploring material decomposition-based Avatar reconstruction could solve this problem. Fourth, for extremely loose clothing, our method may fail during animation due to the lack of physics-aware modeling. Incorporating physics-based deformation models to handle clothing dynamics is a promising direction for future research.

Potential Negative Impact. Our methods may invade privacy or be used by criminals for improper purposes. Therefore, watermarking technology and related regulations need to be improved to ensure that the technology can be used safely and serve society.

Table 6. Comparison of the time cost of the pose refinement process.

Methods	Time per frame	Total time
ICON [51]	25s	330min
Ours	**10s**	**10min**

Acknowledgments

This work was supported by the Startup Foundation for Young Faculty at School of Robotics, Hunan University. We would like to express our sincere gratitude to our colleagues at Baidu for their valuable support and assistance in this work.

References

[1] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In *CVPR*, pages 8387–8397, 2018. 6, 7, 8

[2] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *ICCV*, pages 5855–5864, 2021. 1

[3] J. Chen, Y. Zhang, D. Kang, X. Zhe, L. Bao, X. Jia, and H. Lu. Animatable neural radiance fields from monocular rgb videos. *arXiv preprint arXiv:2106.13629*, 2021. 8

[4] M. Dou, P. Davidson, S. R. Fanello, S. Khamis, A. Kowdle, C. Rhemann, V. Tankovich, and S. Izadi. Motion2fusion: Real-time volumetric performance capture. *ACM TOG*, 36(6):1–16, 2017. 2

[5] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, S. O. Escolano, C. Rhemann, D. Kim, J. Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM TOG*, 35(4):1–13, 2016. 1, 2

[6] H. Edelsbrunner, D. Kirkpatrick, and R. Seidel. On the shape of a set of points in the plane. *TIT*, 29(4):551–559, 1983. 5

[7] C. Geng, S. Peng, Z. Xu, H. Bao, and X. Zhou. Learning neural volumetric representations of dynamic humans in minutes. In *CVPR*, 2023. 6, 7

[8] K. Guo, P. Lincoln, P. Davidson, J. Busch, X. Yu, M. Whalen, G. Harvey, S. Orts-Escolano, R. Pandey, J. Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM TOG*, 38(6):1–19, 2019. 1

[9] K. Guo, F. Xu, T. Yu, X. Liu, Q. Dai, and Y. Liu. Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera. *ACM TOG*, 36(4):1, 2017. 1, 2

[10] T. He, J. Collomosse, H. Jin, and S. Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *NeurIPS*, 33:9276–9287, 2020. 2

[11] T. He, Y. Xu, S. Saito, S. Soatto, and T. Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *ICCV*, pages 11046–11056, 2021. 3

[12] S. Hu, T. Hu, and Z. Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. In *CVPR*, pages 20418–20431, 2024. 1, 2, 3, 6, 7

[13] Z. Huang, Y. Xu, C. Lassner, H. Li, and T. Tung. Arch: Animatable reconstruction of clothed humans. In *CVPR*, pages 3093–3102, 2020. 3

[14] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *UIST*, pages 559–568, 2011. 2

[15] B. Jiang, Y. Hong, H. Bao, and J. Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *CVPR*, pages 5605–5615, 2022. 1, 2, 3

[16] T. Jiang, X. Chen, J. Song, and O. Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *CVPR*, pages 16922–16932, 2023. 6, 7, 8, 10

[17] W. Jiang, K. M. Yi, G. Samei, O. Tuzel, and A. Ranjan. Neuman: Neural human radiance field from a single video. In *ECCV*, pages 402–418. Springer, 2022. 3

[18] H. Jung, N. Brasch, J. Song, E. Perez-Pellitero, Y. Zhou, Z. Li, N. Navab, and B. Busam. Deformable 3d gaussian splatting for animatable human avatars. *arXiv preprint arXiv:2312.15059*, 2023. 2, 3

[19] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2018. 2, 3

[20] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4), 2023. 1, 3, 8

[21] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 5

[22] M. Kocabas, N. Athanasiou, and M. J. Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, pages 5253–5263, 2020. 2, 3

[23] N. Kolotouros, G. Pavlakos, and K. Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, pages 4501–4510, 2019. 2, 3

[24] Y. Kwon, D. Kim, D. Ceylan, and H. Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *NeurIPS*, 34:24741–24752, 2021. 2, 3, 6, 7

[25] J. Lei, Y. Wang, G. Pavlakos, L. Liu, and K. Daniilidis. Gart: Gaussian articulated template models. In *CVPR*, 2024. 6, 7, 8, 9

[26] J. Li, S. Bian, C. Xu, Z. Chen, L. Yang, and C. Lu. Hybrik-x: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery. *IEEE TPAMI*, 2025. 2, 3, 5

[27] M. Li, J. Tao, Z. Yang, and Y. Yang. Human101: Training 100+ fps human gaussians in 100s from 1 view. *arXiv preprint arXiv:2312.15258*, 2023. 2, 3

[28] J. Lin, A. Zeng, H. Wang, L. Zhang, and Y. Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *CVPR*, pages 21159–21168, 2023. 2, 3, 5

[29] K. Lin, L. Wang, and Z. Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, pages 1954–1963, 2021. 2, 3

[30] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 2, 3, 5

[31] Q. Ma, J. Yang, A. Ranjan, S. Pujades, G. Pons-Moll, S. Tang, and M. J. Black. Learning to dress 3d people in generative clothing. In *CVPR*, pages 6469–6478, 2020. 2, 3

[32] Q. Ma, J. Yang, S. Tang, and M. J. Black. The power of points for modeling humans in clothing. In *ICCV*, pages 10974–10984, 2021. 3

[33] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 3

[34] R. A. Newcombe, D. Fox, and S. M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR*, pages 343–352, 2015. 2

[35] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, pages 10975–10985, 2019. 2, 3, 4, 5

[36] S. Peng, J. Dong, Q. Wang, S. Zhang, Q. Shuai, X. Zhou, and H. Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, pages 14314–14323, 2021. 1, 2, 3, 6, 7

[37] S. Peng, Z. Xu, J. Dong, Q. Wang, S. Zhang, Q. Shuai, H. Bao, and X. Zhou. Animatable implicit neural representations for creating realistic avatars from videos. *IEEE TPAMI*, 2024. 6, 7

[38] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, pages 9054–9063, 2021. 1, 2, 3, 6, 7, 8

[39] S. Qian, T. Kirschstein, L. Schoneveld, D. Davoli, S. Giebenhain, and M. Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20299–20309, 2024. 1, 2, 3

[40] Z. Qian, S. Wang, M. Mihajlovic, A. Geiger, and S. Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *CVPR*, pages 5020–5030, 2024. 1, 2, 3, 5

[41] E. Remelli, T. Bagautdinov, S. Saito, C. Wu, T. Simon, S.-E. Wei, K. Guo, Z. Cao, F. Prada, J. Saragih, et al. Drivable volumetric avatars using texel-aligned features. In *ACM SIGGRAPH*, pages 1–9, 2022. 6, 7

[42] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, pages 2304–2314, 2019. 2

[43] S. Saito, G. Schwartz, T. Simon, J. Li, and G. Nam. Relightable gaussian codec avatars. In *CVPR*, 2024. 1, 2, 3, 5

[44] S. Saito, T. Simon, J. Saragih, and H. Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, pages 84–93, 2020. 2

[45] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 6

[46] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *ECCV*, pages 20–36, 2018. 2

[47] T. von Marcard, R. Henschel, M. Black, B. Rosenhahn, and G. Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, sep 2018. 9

[48] C.-Y. Weng, B. Curless, P. P. Srinivasan, J. T. Barron, and I. Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *CVPR*, pages 16210–16220, 2022. 1, 2, 3, 6, 7, 8

[49] D. Xiang, F. Prada, C. Wu, and J. Hodgins. Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In *3DV*, pages 322–332. IEEE, 2020. 2, 3

[50] Y. Xiu, J. Yang, X. Cao, D. Tzionas, and M. J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *CVPR*, 2023. 2

[51] Y. Xiu, J. Yang, D. Tzionas, and M. J. Black. Icon: Implicit clothed humans obtained from normals. In *CVPR*, pages 13286–13296. IEEE, 2022. 2, 5, 10, 11

[52] Z. Yang, A. Zeng, C. Yuan, and Y. Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 5, 10

[53] C. Ye, L. Qiu, X. Gu, Q. Zuo, Y. Wu, Z. Dong, L. Bo, Y. Xiu, and X. Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM Transactions on Graphics (TOG)*, 2024. 10

[54] A. Yu, V. Ye, M. Tancik, and A. Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 6, 7

[55] T. Yu, K. Guo, F. Xu, Y. Dong, Z. Su, J. Zhao, J. Li, Q. Dai, and Y. Liu. Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera. In *ICCV*, pages 910–919, 2017. 1, 2

[56] T. Yu, Z. Zheng, K. Guo, J. Zhao, Q. Dai, H. Li, G. Pons-Moll, and Y. Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *CVPR*, pages 7287–7296, 2018. 1, 2

[57] Y. Yuan, X. Li, Y. Huang, S. De Mello, K. Nagano, J. Kautz, and U. Iqbal. Gavatar: Animatable 3d gaussian avatars with implicit mesh learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 896–905, 2024. 1, 2, 3, 5

[58] H. Zhang, Y. Tian, Y. Zhang, M. Li, L. An, Z. Sun, and Y. Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE TPAMI*, 2023. 2, 3, 5

[59] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7

[60] Y. Zhao, C. Wu, B. Huang, Y. Zhi, C. Zhao, J. Wang, and S. Gao. Surfel-based gaussian inverse rendering for fast and relightable dynamic human reconstruction from monocular video. *arXiv preprint arXiv:2407.15212*, 2024. 3

[61] Z. Zheng, T. Yu, Y. Liu, and Q. Dai. Pamir: Parametric model-conditioned implicit representation for image-based

human reconstruction. *IEEE TPAMI*, 44(6):3170–3184, 2021. 2

[62] Z. Zheng, X. Zhao, H. Zhang, B. Liu, and Y. Liu. Avatarrex: Real-time expressive full-body avatars. *ACM Transactions on Graphics (TOG)*, 42(4), 2023. 10

[63] Y. Zhou, M. Habermann, I. Habibie, A. Tewari, C. Theobalt, and F. Xu. Monocular real-time full body capture with inter-part correlations. In *CVPR*, pages 4811–4822, 2021. 2, 3

[64] W. Zielonka, T. Bagautdinov, S. Saito, M. Zollhöfer, J. Thies, and J. Romero. Drivable 3d gaussian avatars. In *3DV*, March 2025. 1, 2, 3, 5