HandSCS: Structural Coordinate Space for Animatable Hand Gaussian Splatting

Yilan Dong[1], Wenqing Wang[2], Qing Wang[1], Jiahao Yang[1], Haohe Liu[2], Xiatian Zhu[2], Gregory Slabaugh[1], and Shanxin Yuan[1⋆]

[1] Queen Mary University of London, London, UK
[2] University of Surrey, Guildford, UK

Abstract. Photorealistic and animatable hand avatars are essential for applications such as AR/VR, gaming, and telepresence. Recent 3D Gaussian Splatting (3DGS) based avatar methods enable real-time rendering of articulated humans, but modeling hands remains challenging due to their compact structure, frequent self-occlusions, and complex finger interactions. Existing approaches primarily rely on pose-driven transformations while representing Gaussians in Euclidean space, lacking an explicit structural association with the underlying skeleton, which makes preserving fine-grained hand structures under complex articulations difficult. In this work, we introduce the Structural Coordinate Space (SCS), a skeleton-relative representation that assigns each Gaussian primitive an explicit structural coordinate with respect to the articulated hand skeleton. SCS is constructed using a hybrid static–virtual bone basis together with a distance–rotation structural descriptor that encodes the geometric relationship between Gaussians and bones. Based on SCS, we enforce both intra-pose and cross-pose structural consistency by combining per-Gaussian residual embeddings for local appearance modeling with structural correspondence across poses. Experiments demonstrate that our approach significantly improves structural consistency and preserves fine geometric details under challenging hand articulations compared with existing 3DGS-based avatar methods.

Keywords: Hand Avatar · 3D Gaussian Splatting · Structural Representation · Real-time Rendering

1 Introduction

Rendering photorealistic, animatable hand avatars is essential for applications in AR/VR, gaming, and telepresence. Recent advances in 3D Gaussian Splatting (3DGS) [16] enable real-time rendering of articulated avatars by representing subjects as collections of 3D Gaussian primitives deformed through skeleton-driven transformations [8, 13, 20, 29, 32, 35, 36, 38, 43]. While this paradigm has shown promising results for human bodies and heads, hands remain particularly challenging. Hand articulation is concentrated in compact structured regions,

⋆ Corresponding author.

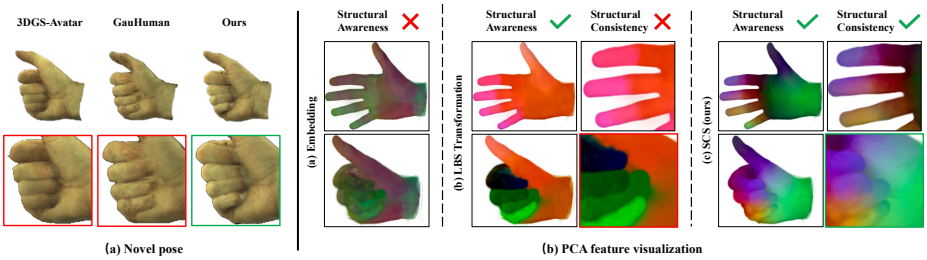


Fig. 1: (a) Compared to the 3DGS-Avatar [36] and GauHuman [8], our approach can maintain consistency and recover details under novel views and novel poses. (b) Each representation is projected to three principal components mapped to RGB, similar colors indicate similar roles. Color distinction across the hand reflects structural awareness, while consistent coloring of the same region across poses reflects structural consistency. Embedding, nearly uniform, lacking structural awareness. LBS transform, spatially varied but inconsistent across poses. Ours, both structurally aware and pose-consistent.

where visually similar fingers frequently interact and occlude each other. Such tightly coupled, fine-grained motions place stronger demands on maintaining structural consistency across poses in Gaussian-based representations.

Despite the progress of 3DGS-based avatar models, current approaches deform Gaussian primitives primarily through pose-driven transformations, such as global pose conditioning [13, 36, 43] or LBS refinement [8, 18, 38]. While these mechanisms propagate articulated motion, the Gaussian representation itself is formulated in Euclidean positions and learned attributes, without a disentangled geometric mapping to the underlying skeleton. As illustrated in Fig. 1, although the LBS transformation matrices provide pose-dependent cues for guiding deformation, these signals vary across poses and do not establish a consistent structural association between a Gaussian and a specific anatomical region. Consequently, maintaining consistent structural behavior under complex articulations becomes challenging. In particular, fine geometric cues, such as fingernail structures and clear separations between adjacent fingers, may degrade in highly articulated configurations where fingers interact closely.

These observations suggest that each Gaussian primitive should carry a structural descriptor relative to the underlying skeleton. However, constructing such a representation is non-trivial. The hand skeleton provides a sparse kinematic scaffold composed of only a small number of joints, whereas 3DGS represents the hand using a large set of spatially distributed Gaussian primitives. Some methods introduce per-Gaussian latent codes to enhance identity modeling capacity [1, 29]. However, as shown in Fig. 1, the embedding itself does not show any semantic region-related distribution. Then, without the underlying structure guidance, a Gaussian can drift and lie in different semantic roles under different poses, while the embedding is fixed across these semantic roles, leading to smoothed-out representations and loss of fine-grained details.

To address this challenge, we introduce the Structural Coordinate Space (SCS), which consists of two components: a hybrid static–virtual bone basis

that defines a structural coordinate system for Gaussian primitives, and a distance–rotation transformation that describes the geometric relationship between a Gaussian and the bone basis. We construct the bone basis by combining static bones in posed space derived from the MANO [37] kinematic topology with additional cross-finger connections and removing redundant connections. To further enrich the structural representation, we introduce pose-related virtual bones in canonical space that provide complementary structural cues in highly articulated regions. While static bones provide stable global anatomical coverage, virtual bones expand the basis to better capture pose-dependent structural variations. The second component is a *structural coordinate descriptor*, the transformation that maps each Gaussian's physical position into its coordinate in SCS, defined through angular and radial terms for each bone in the basis. Together, the basis and descriptor assign every Gaussian an explicit, interpretable structural coordinate within the hand anatomy.

With this representation, each Gaussian can be associated with a structural attribute relative to the articulated skeleton. We then promote consistency from two aspects: (1) Within a given pose, SCS assigns each Gaussian a distinct structural identity. Conditioned on this structural reference, we introduce per-Gaussian embeddings [1, 29] to capture local geometric and appearance variations. These embeddings parameterize residual offsets for standard Gaussian attributes, enabling flexible detail modeling while preserving the structural identity provided by SCS. (2) Across poses, Gaussians with similar SCS structural coordinates exhibit consistent behavior under similar articulation. To enforce this, we establish cross-pose correspondences by matching Gaussians through their SCS structural coordinates, and penalize attribute divergence between corresponding Gaussians, modulated by pose similarity.

Our contributions are summarized as follows:

1. We introduce the Structural Coordinate Space (SCS), a skeleton-relative representation that assigns each Gaussian primitive an explicit structural coordinate with respect to the articulated hand skeleton.
2. Built upon SCS, we design a structure-aware learning scheme that enforces both intra-pose and cross-pose consistency through per-Gaussian residual embeddings and structural correspondence across poses.
3. Experiments on the InterHand2.6M dataset demonstrate that HandSCS achieves state-of-the-art in hand avatar animation, effectively preserves fine-grained geometric details under complex hand articulations while maintaining real-time rendering speed.

2 Related Work

2.1 Mesh based Hand Avatars

Parametric meshes are widely used for hand avatar modeling [3, 5, 6, 14, 15, 21, 26, 27, 33, 34, 37] due to their explicit geometry and compatibility with animation pipelines. Early works [3, 26] estimate colored meshes, but lack fine-grained

appearance. Recent methods enhance realism by incorporating appearance modeling: AMVUR [14] introduces attention-based vertices and occlusion-aware texture regression; NIMBLE [21] builds a non-rigid model with bones, muscles, and physically-based skin; HTML [34] extends MANO [37] with a PCA-based texture model; and HARP [15] adds explicit albedo and normal maps. However, these methods often rely on dense surface capture, non-rigid tracking, or subject-specific templates, increasing cost and limiting scalability.

2.2 Neural Rendering for Hand Avatars

Neural implicit representations [11, 24, 25] have been widely adopted for animatable hand avatar. Many methods [2, 4, 7, 10, 17, 30, 45] render hands via inverse LBS and volume rendering of predicted color and density fields. LISA [4] predicts per-bone color and signed distance fields; HandAvatar [2] adds explicit albedo and illumination modeling; HandNeRF [7] introduces hand-specific deformation fields. While producing high-quality results, these methods suffer from slow rendering due to dense ray sampling. To improve efficiency, mesh-guided acceleration has been explored [10,30], though PHRIT [10] remains slow (10 FPS) and LiveHand [30] sacrifices realism for speed. OHTA [45] and BiTT [17] reconstruct relightable, personalized appearance from a single image, but struggle to maintain consistency across poses and viewpoints.

2.3 3DGS based Hand Avatars

3D Gaussian Splatting (3DGS [16]) provides a new paradigm for real-time scene reconstruction, and has been adept at representing hand avatars by modeling geometry and appearance through continuous and differentiable Gaussian primitives [9, 13, 32, 39, 43]. 3D-PSHR [13] incorporates albedo and shading into the Gaussian representation, but lacks the ability to model pose-dependent appearance. GaussianHand [43] improves rendering quality by learning Gaussian offsets from pose-dependent and pose-independent properties with a UNet-based appearance decoder. However, existing approaches do not explicitly model the underlying hand structure, limiting their ability to maintain stable geometry and appearance under challenging cases involving large articulations.

2.4 3DGS based Human Avatars

Recent works have explored animatable human avatars using 3D Gaussian Splatting [8, 12, 18, 19, 22, 29, 31, 36, 38, 41, 44, 46]. Some methods [8, 18, 38, 41] reconstruct avatars from monocular or multi-view videos, but do not explicitly model pose-dependent appearance. To address this, several approaches [29, 36] learn Gaussian property offsets via a MLP, conditioned on the pose along with either positions [36] or per-Gaussian latent codes [29]. These methods have demonstrated promising results for clothed human bodies, which often involve challenges such as loose garments and topology variation. However, hand modeling

introduces different difficulties, including high articulation, self-occlusion, and inter-part contact, highlighting the need for structure-aware modeling, which prior Gaussian-splatting avatars have not explored.

3 Method

Fig. 2 shows the framework of HandSCS. We model a 3D hand in canonical space, where each MANO mesh vertex is initialized as a 3D Gaussian with attributes $\mathbf{\Lambda} = \{\mathbf{x}, \mathbf{c}, \mathbf{\Sigma}, \mathbf{o}\}$ representing position, color, covariance, and opacity. The core of our approach is the Structural Coordinate Space (SCS), which reparameterizes each Gaussian by its geometric relationship to the hand skeleton (Sec. 3.1). We define SCS through its coordinate basis, the set of bones forming the reference frame (Sec. 3.1.1), and a coordinate descriptor, the transformation that maps each Gaussian into this space (Sec. 3.1.2). Non-rigid deformation is formulated within SCS, with each Gaussian's attribute offsets determined by its structural coordinate and per-Gaussian embeddings (Sec. 3.1.3). To promote consistency across poses, we introduce an inter-pose consistency loss that establishes correspondences through structural coordinates in SCS (Sec. 3.2).

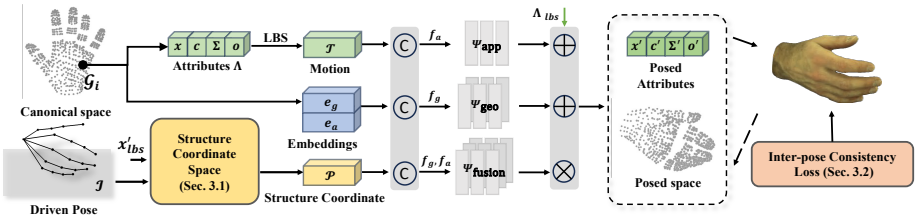
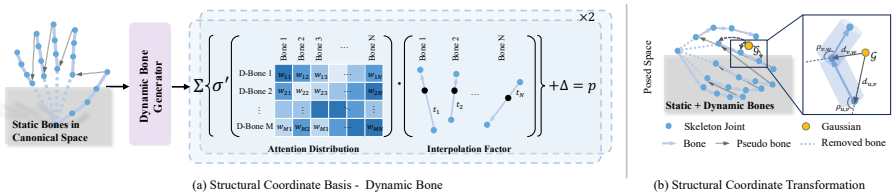


Fig. 2: Overview of HandSCS. Each 3D Gaussian is modeled in canonical space and reparameterized in SCS. The structural coordinate $\mathcal{P}$ is computed from the LBS-posed position x'_{lbs} and the skeleton joints via angular-radial descriptors over the hybrid static-virtual bone basis. Together with per-Gaussian geometry embedding e_g and appearance embedding e_a, the structural coordinate determines non-rigid attribute offsets. Cross-pose consistency is enforced by matching Gaussians through their structural coordinates in SCS and penalizing attribute divergence.

3.1 Structural Coordinate Space

3.1.1 Structural Coordinate Basis We construct the coordinate basis from a unified set of static and virtual bone vectors. The static bones capture the unified physical hand structure and are deformed through the standard LBS in MANO model, while preserving a consistent canonical skeletal topology across poses. In contrast, the virtual bones adaptively model pose-dependent structural variations. Together, they form a flexible structural coordinate basis.

Static Bone Basis. The static part defines the physical hand bones derived from the MANO kinematic model [37], augmented with cross-finger connections



(a) Structural Coordinate Basis - Dynamic Bone (b) Structural Coordinate Transformation

Fig. 3: Illustration of **Structural Coordinate Space** (SCS). The left part shows the extended kinematic topology $\mathcal{E}$ based on MANO model, with added cross-finger pseudo-bones and removal of redundant connections. The right part illustrates how each Gaussian is encoded with its cosine of angle $\rho_{u,v}$ and distance $d_{u,v}$ relative to each bone in the posed space.

to capture anatomical relationships between adjacent fingers. To maintain a compact topology, we remove redundant wrist-to-finger edges for all fingers except the thumb, as illustrated in Fig. 3. The resulting static bone topology is formulated as,

$$\mathcal{E}_{\mathrm{s}} = \left(\mathcal{E}_{\mathrm{MANO}} \setminus \mathcal{E}_{\mathrm{removed}}\right) \cup \mathcal{E}_{\mathrm{cross}}, \tag{1}$$

Given the pose parameters $\boldsymbol{\theta}$ and shape parameters $\boldsymbol{\beta}$, the MANO model computes the joint positions in observed space $\mathcal{J}_o = \{\boldsymbol{j}_o^i\}_{i=0}^{K-1}$, where $\mathcal{J}_o \in \mathbb{R}^{K\times 3}$ and K is the number of joints. Using the defined topology $\mathcal{E}_{\mathrm{static}} = \{(u,v) \mid v \in \mathcal{J}\}$, where u denotes the parent joint of v. The set of static bones is formulated as,

$$\mathcal{B}_{\mathrm{s}} = \{\mathbf{b}_{u,v} = (\boldsymbol{j}_o^u, \boldsymbol{j}_o^v) \mid (u,v) \in \mathcal{E}_{\mathrm{s}}\}, \tag{2}$$

Here, "static" refers to the fixed skeletal topology.

Virtual Bone Basis. The predefined static bone alone cannot fully capture the structural variations induced by highly articulated or interacting poses. To address this limitation, we introduce virtual bones, constructed by referencing the static bones and interpolating their endpoints along bone segments.

Formally, given the static bone topology $\mathcal{E}_{\mathrm{static}}$ and the canonical joint locations $\mathcal{J} = \{\boldsymbol{j}^i\}_{i=0}^{K-1}$, we can obtain the set of canonical static bones as $\mathcal{B}_{\mathrm{s}}^c = \{(\boldsymbol{j}^u, \boldsymbol{j}^v) \mid (u,v) \in \mathcal{E}_{\mathrm{static}}\}$. The virtual bone $\mathcal{B}_{\mathrm{d}}$ is represented by two endpoints (p,q). For virtual bone i, the model (i) predicts an attention distribution w over all static bones to identify relevant reference regions, and (ii) learns interpolation factors t and offsets Δ to determine precise start and end positions. The endpoints are computed as,

$$\begin{aligned} p &= \sum_{b}^{\mathcal{B}} \tilde{w}_b^s \left[(1-t_b^s)\mathbf{j}^u + t_b^s \mathbf{j}^v\right] + \Delta p_b, \\ q &= \sum_{b}^{\mathcal{B}} \tilde{w}_b^e \left[(1-t_b^e)\mathbf{j}^u + t_b^e \mathbf{j}^v\right] + \Delta q_b, \end{aligned} \tag{3}$$

where w_b^s and w_b^e are the attention weights over canonical static bones, and $t^s, t^e \in [0,1]$ denote the interpolation factors along each bone, and $\Delta\mathbf{p}$ and

$\Delta\mathbf{q}$ are pose-dependent offsets for fine-grained adjustments. The final structural coordinate basis is obtained as $\mathcal{B} = \mathcal{B}_s \cup \mathcal{B}_d$ and the topology is $\mathcal{E} = \mathcal{E}_s \cup \mathcal{E}_d$.

To encourage smooth transitions between adjacent bones, we apply a smoothing kernel to the attention distribution. Given the attention logits $\mathcal{W} = \{w_b | b \in \mathcal{E}_{\text{static}}\}$, the smoothed weights are computed as,

$$\tilde{\mathcal{W}} \;=\; \frac{K\,\sigma(\mathcal{W})}{\left\|K\,\sigma(\mathcal{W})\right\|_1} \tag{4}$$

where $\sigma(\cdot)$ denotes the softmax operation, and $K \in \mathbb{R}^{B\times B}$ is a fixed topology-aware smoothing kernel defined over the static hand bone, assigning a weight of 0.5 to each bone itself, 0.25 to its directly connected neighbors and 0 to all other bones.

In practice, we generate M virtual bones using a lightweight MLP as the virtual bone generator. The network takes the pose parameters $\boldsymbol{\theta}$ and canonical joint positions $\mathcal{J}$ as input, and outputs the parameters $w^s, w^e, t^s, t^e, \Delta\mathbf{p}, \Delta\mathbf{q}$. The resulting virtual bones are then combined with the static ones to form an enriched structural prior, providing a more flexible coordinate basis. The static bones define a stable anatomical reference frame; the virtual bones expand this basis to capture pose-dependent structural variations, enabling the coordinate space to adapt to the current articulation.

3.1.2 Structural Coordinate Descriptor The coordinate descriptor maps each Gaussian's physical position into its structural coordinate in SCS, encoding its geometric relationship to each bone in the basis.

Per-bone structural relations. Let $\mathbf{x}'_{\text{lbs}}$ denote the LBS-posed center of a Gaussian. The structural relation between $\mathbf{x}'_{\text{lbs}}$ and the coordinate basis $\mathcal{B}$ is decomposed into an *angular* term and a *radial* term. The angular term measures how the Gaussian is oriented with respect to the bone. For each bone $\mathbf{b}_{u,v} = (j^u, j^v)$, we define the Gaussian-to-basis vector $\mathbf{r}_{u,v} = \mathbf{j}_u - \mathbf{x}'_{\text{lbs}}$. We use the cosine similarity between $\mathbf{r}_{u,v}$ and $\mathbf{b}_{u,v}$,

$$\rho_{u,v} = \frac{\mathbf{r}_{u,v} \cdot \mathbf{b}_{u,v}}{\|\mathbf{r}_{u,v}\|_2 \, \|\mathbf{b}_{u,v}\|_2}, \quad (u,v) \in \mathcal{E}, \tag{5}$$

where $\|\cdot\|_2$ denotes L2 norm. The radial term measures the spatial affinity between the Gaussian and the bone. For static bones in posed space, we apply an exponential decay,

$$d_{u,v} = \exp\!\left(-\frac{\|\mathbf{r}_{u,v}\|_2^2}{\tau^2}\right), \quad (u,v) \in \mathcal{E}_{\text{s}}, \tag{6}$$

where τ is the decay rate, set to 0.08, restricting each bone's influence to its local neighborhood. For virtual bones in canonical space, we use the raw distance,

$$d_{u,v} = \|\mathbf{r}_{u,v}\|_2, \quad (u,v) \in \mathcal{E}_{\text{d}}, \tag{7}$$

which preserves a larger spatial dynamic range, providing complementary structural cues to the bounded static bone descriptors.

Structural coordinates. By aggregating these relations across all bones, the structural coordinate descriptor for each Gaussian is defined as,

$$\mathcal{P} = [d_{u,v} \cdot \rho_{u,v}]_{(u,v)\in\mathcal{E}}. \tag{8}$$

where $[\cdot]$ denotes concatenation. Each dimension of $\mathcal{P} \in \mathbb{R}^{|\mathcal{E}|}$ encodes the Gaussian's structural relationship to one bone in the basis. The resulting structural coordinate is spatially smooth, as nearby Gaussians obtain similar coordinates, yet structurally discriminative: Gaussians from different structural regions maintain distinct coordinates even when spatially close, such as during self-contact between fingers.

3.1.3 Intra-pose Consistency in SCS. In SCS, each Gaussian's identity consists of two complementary aspects: its structural coordinate $\mathcal{P}$, which encodes its geometric relationship to the skeleton, and per-Gaussian learnable embeddings that capture fine-grained variations specific to each Gaussian. We assign each Gaussian a geometry embedding $\mathbf{e}_g \in \mathbb{R}^d$ and an appearance embedding $\mathbf{e}_a \in \mathbb{R}^d$, jointly optimized through differentiable rendering. While $\mathcal{P}$ captures the Gaussian's structural role within the hand anatomy, the embeddings model subject-specific surface details that vary across individual Gaussians at the same structural position.

Non-rigid deformation is determined jointly by the structural coordinate $\mathcal{P}$, per-Gaussian embeddings, and the LBS transformation matrix $\mathcal{T}$. Geometry and appearance offsets are decoded from these components through MLPs. The final attributes combine canonical values with the decoded offsets,

$$\begin{aligned} \mathbf{x}' &= \mathbf{x} + \Delta\mathbf{x} + \Delta\mathbf{x}^f, \\ c' &= c + \Delta c + \Delta c^f, \\ \boldsymbol{\Sigma}' &= (\Delta\boldsymbol{\Sigma}^f \Delta\boldsymbol{\Sigma})\, \boldsymbol{\Sigma}\, (\Delta\boldsymbol{\Sigma}\Delta\boldsymbol{\Sigma}^f). \end{aligned} \tag{9}$$

where $\Delta\mathbf{x}, \Delta\boldsymbol{\Sigma}, \Delta c$ are decoded from geometry or appearance representations and $\Delta\mathbf{x}^f, \Delta\boldsymbol{\Sigma}^f, \Delta c^f$ from their fusion (details in supplementary material).

Each component contributes a distinct aspect: $\mathcal{T}$ captures rigid bone-driven motion, the embeddings capture per-Gaussian identity, and $\mathcal{P}$ encodes each Gaussian's position within the hand's skeletal structure. This enables the deformation to reflect structural role: Gaussians near joints, at fingertips, or in contact regions each receive structurally appropriate offsets determined by their position in SCS.

3.2 Inter-Pose Consistency in SCS.

The structural coordinate assigned to each Gaussian encodes its relationship to the skeleton rather than its absolute spatial position, making it a natural basis

for establishing correspondences across different poses. We leverage this property to enforce cross-pose consistency: Gaussians occupying the same structural role should exhibit coherent attributes across similar articulations.

At training iteration t, each Gaussian $\mathcal{G}_t^i$ is associated with pose parameter $\boldsymbol{\theta}_t$, structural coordinate $\mathcal{P}_t^i$ in SCS, and attributes $\mathbf{M}_t^i = (\mathbf{x}_t^i, \boldsymbol{\Sigma}_t^i, \mathbf{c}_t^i, \mathbf{e}_{g,t}^i, \mathbf{e}_{a,t}^i)$. To modulate the loss, we first compute the pose similarity as weighting factor,

$$\omega_{t,t-1} = \exp\left(-\frac{\|\Phi_\theta(\boldsymbol{\theta}_t) - \Phi_\theta(\boldsymbol{\theta}_{t-1})\|_2^2}{2\delta^2}\right), \tag{10}$$

where Φ_θ is a pose encoder and δ is set to 1.3.

Correspondences between the current and previous frames are established through structural coordinates. For each Gaussian $\mathcal{G}_t^i$, we find its structurally nearest counterpart in the previous frame,

$$\pi(i) = \arg\min_{j=1}^{N_{t-1}} \|\mathcal{P}_t^i - \mathcal{P}_{t-1}^j\|_2^2, \tag{11}$$

The consistency loss penalizes attribute divergence between structurally corresponding Gaussians,

$$\mathcal{L}_{\text{con}} = \omega_{t,t-1} \sum_{i=1}^{N_t} \|\mathbf{M}_t^i - \mathbf{M}_{t-1}^{\pi(i)}\|_2^2, \tag{12}$$

where $\mathbf{M}_{t-1}^{\pi(i)}$ denotes attributes from last iteration. For initialization, $\mathbf{M}_0^i = \mathbf{M}_1^i$.

By including $\mathbf{e}_g, \mathbf{e}_a$, the consistency loss also regularizes the non-rigid deformation to be coherent across structurally corresponding Gaussians. Also, when $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1}$, the loss naturally enforces cross-view alignment for the same pose, improving robustness under self-occlusion and viewpoint variation.

3.3 Training Objective Function

Following other 3DGS-based methods [8, 13, 32, 36], we apply a standard RGB loss, mask loss, SSIM loss, and LPIPS loss to supervise the 3D model,

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{rgb}} + \lambda_1 \mathcal{L}_{\text{mask}} + \lambda_2 \mathcal{L}_{\text{SSIM}} + \lambda_3 \mathcal{L}_{\text{LPIPS}}, \tag{13}$$

where λ are loss weights. Empirically, we set $\lambda_1 = 0.1$, $\lambda_2 = \lambda_3 = 0.01$. Additionally, to prevent the adaptive attribute offsets from producing excessively large deformations, we apply an offset regularization loss,

$$\mathcal{L}_{\text{reg}} = \|\Delta\mathbf{x}\|_2 + \|\Delta\boldsymbol{\Sigma} - \mathbf{I}\|_F + \|\Delta c\|_2, \tag{14}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. This term penalizes the position and color offset magnitudes while encouraging the covariance offset to remain close to the identity matrix, preventing unstable shape distortions. The overall loss is,

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \lambda_{\text{con}} \mathcal{L}_{\text{con}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}, \tag{15}$$

where λ_{con} is set to 0.01, and λ_{reg} is set to 0.1.

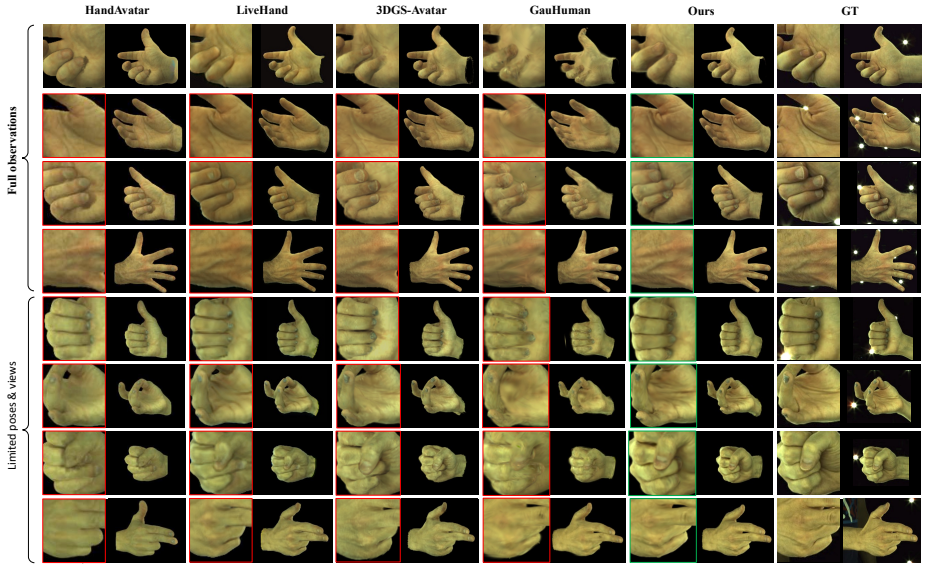


Fig. 4: Qualitative Comparison of Novel Pose Synthesis. We compare results on InterHand2.6M [28] across five subjects using HandAvatar [2], LiveHand [30], GauHuman [8], and our method. HandSCS produces sharper structures, cleaner boundaries, and fewer artifacts under challenging articulations.

4 Experiments

4.1 Datasets and Metrics

We conduct experiments on the widely used InterHand2.6M [28] dataset at 5 FPS, which provides large-scale multi-view sequences of diverse hand poses. To fairly assess our method, we select right-hand sequences from five subjects as shown in Tab. 1, where the last two subjects are sparse-view and sparse-pose tasks. For novel pose rendering, we allocate the last 50 poses for testing and use the remaining poses for training. For novel view synthesis, 10 views are selected for training, 15 views are used as testing set. Detailed specifications of the selected camera views are provided in the supplementary material.

We evaluate Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [40] and Learned Perceptual Image Patch Similarity (LPIPS) [42] as metrics for the rendering results on full image. We use FPS to compare the rendering speed. Rendering speeds are evaluated on a single NVIDIA A100 GPU.

4.2 Implementation details

All three offset decoders, geometry, appearance, and fusion share the same structure. We adopt the same architecture, consisting of a fully connected multilayer perceptron with 2 hidden layers, each containing 256 neurons, a learnable gate (initialized to zero) is included to avoid very large outputs at the beginning of

Table 1: Novel pose animation on InterHand2.6M [28]. HandSCS achieves the best overall performance. Bold numbers indicate the best results.

Method	test/Capture0			test/Capture1			val/Capture0			train/Capture5			train/Capture6		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
LiveHand [30]	29.17	0.949	0.0528	26.85	0.776	0.0619	29.33	0.893	0.0606	31.35	**0.975**	0.0238	30.67	**0.969**	0.0274
HandAvatar [2]	31.36	0.959	0.0453	30.37	0.961	0.0448	30.35	0.954	0.0499	30.36	0.966	0.0315	29.33	0.958	0.0390
3dgs-avatar [36]	31.35	0.953	0.0477	31.07	0.957	0.0475	31.12	0.950	0.0498	30.80	0.965	0.0304	29.59	0.956	0.0356
GauHuman [8]	31.71	0.958	0.0424	31.72	0.962	0.0445	31.15	0.955	0.0488	31.06	0.965	0.0304	30.53	0.960	0.0354
HandSCS	**32.91**	**0.966**	**0.0291**	**32.99**	**0.970**	**0.0282**	**32.40**	**0.963**	**0.0328**	**32.09**	0.972	**0.0228**	**31.31**	0.966	**0.0273**

Table 2: Sparse-view novel view synthesis on InterHand2.6M [28]. Models are trained on 10 views and evaluated on 15 held-out views. While LiveHand [30] attains slightly higher SSIM due to smoother NeRF interpolation under sparse viewpoints, HandSCS produces more consistent structures, clearer boundaries, and fewer artifacts under challenging viewing angles. Bold numbers indicate the best results.

Method	test/Capture0			test/Capture1			val/Capture0		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
LiveHand [30]	31.95	**0.984**	0.0335	31.03	**0.979**	0.0367	30.21	0.969	0.0449
HandAvatar [2]	32.06	0.961	0.0440	31.62	0.963	0.0440	32.74	0.963	0.0432
3dgs-avatar [36]	31.77	0.953	0.0512	31.56	0.956	0.0479	32.20	0.953	0.0498
GauHuman [8]	32.10	0.958	0.0446	31.37	0.958	0.0462	31.73	0.955	0.0479
HandSCS	**35.18**	0.973	**0.0297**	**35.41**	0.975	**0.0283**	**35.74**	**0.972**	**0.0300**

training. For Gaussian initialization, we use the MANO model with 778 vertices as the initial state. The color and opacity of each Gaussian are randomly initialized. More details are provided in our supplementary materials and code[3].

4.3 Baselines and Result Comparison

Baselines. To evaluate the performance of HandSCS, we conduct comparisons with state-of-the-art NeRF-based method HandAvatar [2] and LiveHand [30] and Gaussian Splatting-based methods 3DGS-Avatar [36] and GauHuman [8]. To ensure fairness, we include only open-source methods and use their official implementations with default configurations.

Novel Pose Synthesis. Tab. 1 summarizes the novel pose results across five subjects. HandSCS achieves the best performance on all metrics, with average gains of +1.24 dB PSNR, +0.008 SSIM, and a 33.4% relative LPIPS reduction over the strongest baseline. These improvements are comparable to or larger than those reported in recent works [2, 8, 30, 36]. Qualitative examples in Fig. 4 show that HandSCS produces sharper details, clearer boundaries, and more stable appearance in challenging articulations. In contrast, existing methods often struggle with fine structures such as fingernails and skin textures or exhibit noticeable artifacts due to limited structural modeling.

Novel View Synthesis. As shown in Tab. 2 and Fig. 5, HandSCS shows the highest rendering quality under novel viewpoints. Averaged over three subjects, HandSCS improves PSNR by +3.78 dB and SSIM by +0.017, and reduces LPIPS by 0.017 (37% relative reduction) compared with the strongest baseline

[3] Code: `https://github.com/yilan120/HandSCS`

GauHuman [8]. Our method preserves fine appearance details such as skin folds, fingernails, and finger boundaries, and introduces fewer artifacts in challenging regions, including occlusions and self-shadowed areas. LiveHand [30] obtains slightly higher SSIM on two subjects. This is expected, as NeRF-based methods interpolate smoothly across sparse viewpoints, while Gaussian-based models may exhibit mild color inconsistencies under difficult viewing angles. Nevertheless, HandSCS consistently produces sharper details, higher-fidelity textures, and significantly fewer rendering artifacts across views. Additional qualitative comparisons are provided in the supplementary material.

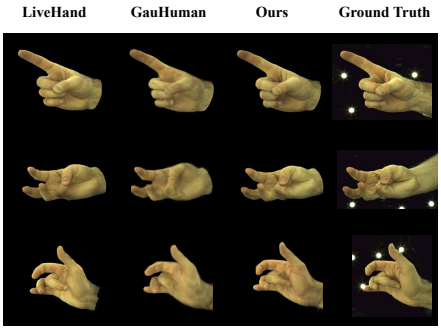


Fig. 5: Visual comparison of Novel View Synthesis.

Models	PSNR ↑	SSIM ↑	LPIPS ↓
(a) baseline	31.71	0.958	0.0428
(b) Embedding+(a)	32.23	0.961	0.0337
$\mathcal{L}_{\text{reg}}$ + (b)	32.32	0.963	0.0339
(c) Inter-pose+(b)	32.42	0.964	0.0335
(d) SCS(Intra-pose)+(b)	32.78	0.966	0.0301
Full model	**32.91**	**0.966**	**0.0291**

Table 3: Ablation study on multi-component structural guidance. (b) introduces disentangled embeddings, and (c)–(d) further incorporate inter-pose and intra-pose (SCS) structural guidance. All components contribute measurable gains, and the full model achieves the best overall performance.

4.4 Ablation Study

We conduct ablation experiments on the InterHand2.6M dataset using subject test/Capture0 to assess the contribution of each component in HandSCS, including the structural coordinate (Sec. 3.1), intra-pose consistency (Sec. 3.1.3), and inter-pose consistency (Sec. 3.2).

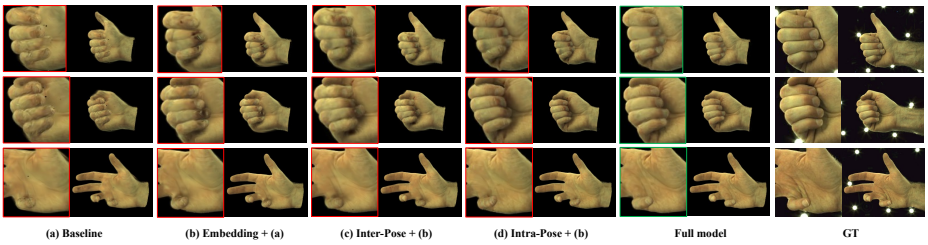


Fig. 6: Visual comparison of component ablations. The full model reduces artifacts in challenging deformation and self-contact regions, producing clearer boundaries and more consistent details.

Component Ablation. As shown in Tab. 3, intra-pose consistency via per-Gaussian embeddings (b) improves over the baseline by providing each Gaussian with a learnable identity for non-rigid deformation. Inter-pose consistency

(c) further enhances global coherence by aligning attributes across structurally corresponding Gaussians. The SCS coordinate (d) yields the largest single gain by grounding each Gaussian with an explicit structural representation. The full model, combining all components, achieves the best overall performance. Visually, as shown in Fig. 6, the baseline shows noise, blurred boundaries, and strong artifacts in contact regions. Embeddings improve stability but still fail under complex deformations and close contact. The consistency loss sharpens global structure, while the SCS coordinate maintains structure consistency such as wrinkles and fingernails. The full model yields the cleanest results with clear boundaries and minimal artifacts.

Coordinate Basis. To understand the role of the coordinate basis, we analyze static and virtual bones individually. Virtual bones provide complementary structural cues beyond the physical skeleton, suppressing artifacts such as surface glitches and broken boundaries around highly bent fingers, but may occasionally introduce pose inconsistencies due to the lack of strict kinematic constraints. Static bones, based on the MANO topology, maintain globally coherent structure and correct overall pose, but their limited diversity makes them less effective in highly deformed regions. As shown in Tab. 4, removing either component degrades quantitative performance. The visual comparison in Fig. 7 further shows that combining both provides stable global structure from static bones and expressive local cues from virtual bones, leading to the most robust reconstruction across both rigid and highly articulated poses.

Models	PSNR ↑	SSIM ↑	LPIPS ↓
w/o Intra-pose	32.42	0.9636	0.0335
w/o Static Bones	32.64	0.965	0.0302
w/o Dynamic Bones	32.70	0.965	0.0311
Full model	**32.91**	**0.966**	**0.0291**

Table 4: Coordinate basis ablation. Virtual bones reduce artifacts in high curvature deformed regions but may introduce pose inaccuracies, while static bones provide a stable structural prior but are less expressive. Combining both provides the most robust reconstruction.

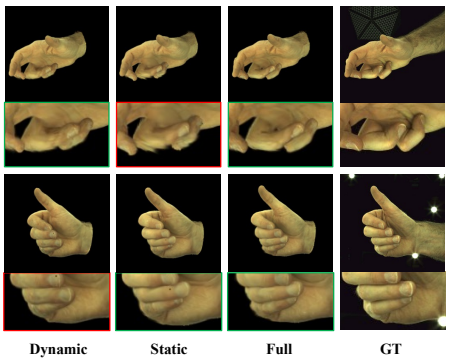


Fig. 7: Visual comparison of Coordinate Basis.

Choice of Space for Virtual Bones. We compare constructing virtual bones in canonical space and posed space. As shown in Tab. 5, canonical space achieves better performance, particularly in LPIPS. This is because static bones already operate in the deformed space, capturing each Gaussian's relationship to the posed skeleton, making virtual bones in the same space partially redundant. In contrast, canonical-space virtual bones encode the displacement of each Gaussian from its rest-pose configuration, providing complementary structural cues.

Virtual Bone Component Ablation. We analyze the two factors determining the endpoint positions of virtual bones. Removing the interpolation parameter t degrades performance, as t enables smooth positioning along bone segments. Removing the offset term Δ reduces local deformation flexibility and detail fidelity. As shown in Tab. 6, using both provides complementary benefits and yields the best reconstruction results.

Models	PSNR ↑	SSIM ↑	LPIPS ↓
Posed Space	32.87	0.9657	0.0303
Canonical Space	32.91	0.9663	0.0291

Table 5: Canonical vs. posed space for virtual bones.

Models	PSNR ↑	SSIM ↑	LPIPS ↓
w/o Intra-pose	32.42	0.9636	0.0335
w/o t	32.67	0.9656	0.0300
w/o Δ	32.69	0.9657	0.0300
Full model	**32.91**	**0.9663**	**0.0291**

Table 6: Ablation of virtual bone components.

Visualization of Deformation Features. To qualitatively assess the effect of SCS, we visualize the learned per-Gaussian deformation descriptors used by our non-rigid deformation MLP. Specifically, we embed these descriptors with t-SNE [23] (Fig. 8) and color each point by its phalange label (LBS weight argmax). Compared with the baseline without SCS, descriptors from the same anatomical region form more compact clusters, indicating that SCS injects meaningful structural cues into the deformation representation.

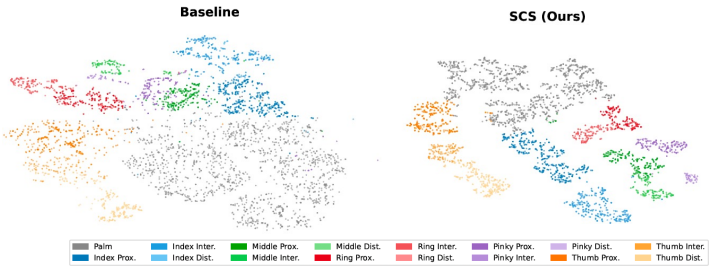


Fig. 8: t-SNE visualization. Without SCS, per-Gaussian features show weak structural separation, while adding SCS leads to clearer clustering by anatomical region (phalange label from LBS argmax).

5 Conclusions

We presented HandSCS, a structure-aware framework for animatable hand avatars based on 3D Gaussian Splatting. Our method introduces the Structural Coordinate Space (SCS), which associates each Gaussian primitive with an explicit structural coordinate relative to the articulated hand skeleton. By formulating non-rigid deformation and cross-pose consistency within this structural representation, the model better preserves fine geometric details and maintains stable appearance under complex articulations. Extensive experiments on the InterHand2.6M dataset demonstrate that HandSCS consistently outperforms existing 3DGS-based and NeRF-based approaches in both quantitative metrics and

visual quality. Ablation studies further confirm the effectiveness of the structural coordinate and its hybrid static–virtual bone basis, highlighting the importance of explicit structural representations for Gaussian-based avatar modeling and their potential for broader articulated avatars.

6 Acknowledgement

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/Y009800/1], through funding from Responsible Ai UK (KP0016). This research utilised Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT.

References

1. Bae, J., Kim, S., Yun, Y., Lee, H., Bang, G., Uh, Y.: Per-gaussian embedding-based deformation for deformable 3D gaussian splatting. In: European Conference on Computer Vision. pp. 321–335. Springer (2024)
2. Chen, X., Wang, B., Shum, H.Y.: Hand Avatar: Free-pose hand animation and rendering from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8683–8693 (2023)
3. Chen, Y., Tu, Z., Kang, D., Bao, L., Zhang, Y., Zhe, X., Chen, R., Yuan, J.: Model-based 3d hand reconstruction via self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10451–10460 (2021)
4. Corona, E., Hodan, T., Vo, M., Moreno-Noguer, F., Sweeney, C., Newcombe, R., Ma, L.: Lisa: Learning implicit shape and appearance of hands. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20533–20543 (2022)
5. Gan, Q., Li, W., Ren, J., Zhu, J.: Fine-grained multi-view hand reconstruction using inverse rendering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 1779–1787 (2024)
6. Gan, Q., Zhou, Z., Zhu, J.: Xhand: Real-time expressive hand avatar. arXiv preprint arXiv:2407.21002 (2024)
7. Guo, Z., Zhou, W., Wang, M., Li, L., Li, H.: Handnerf: Neural radiance fields for animatable interacting hands. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21078–21087 (2023)
8. Hu, S., Hu, T., Liu, Z.: GauHuman: Articulated gaussian splatting from monocular human videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20418–20431 (2024)
9. Huang, X., Li, H., Liu, W., Liang, X., Yan, Y., Cheng, Y., Gao, C.: Learning interaction-aware 3D gaussian splatting for one-shot hand avatars. Advances in Neural Information Processing Systems **37**, 14127–14147 (2025)
10. Huang, Z., Chen, Y., Kang, D., Zhang, J., Tu, Z.: Phrit: Parametric hand representation with implicit template. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14974–14984 (2023)
11. Jiang, B., Hong, Y., Bao, H., Zhang, J.: Selfrecon: Self reconstruction your digital avatar from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5605–5615 (2022)

12. Jiang, Y., Shen, Z., Wang, P., Su, Z., Hong, Y., Zhang, Y., Yu, J., Xu, L.: HiFi4G: High-fidelity human performance rendering via compact gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19734–19745 (2024)
13. Jiang, Z., Rahmani, H., Black, S., Williams, B.: 3D points splatting for real-time dynamic hand reconstruction. Pattern Recognition p. 111426 (2025)
14. Jiang, Z., Rahmani, H., Black, S., Williams, B.M.: A probabilistic attention model with occlusion-aware texture regression for 3D hand reconstruction from a single rgb image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 758–767 (2023)
15. Karunratanakul, K., Prokudin, S., Hilliges, O., Tang, S.: HARP: Personalized hand reconstruction from a monocular rgb video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12802–12813 (2023)
16. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3D gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023), `https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/`, accessed: 2026-06-29
17. Kim, M., Kim, T.K.: BiTT: Bi-directional texture reconstruction of interacting two hands from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10726–10735 (2024)
18. Kocabas, M., Chang, J.H.R., Gabriel, J., Tuzel, O., Ranjan, A.: Hugs: Human gaussian splats. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 505–515 (2024)
19. Kwon, Y., Fang, B., Lu, Y., Dong, H., Zhang, C., Carrasco, F.V., Mosella-Montoro, A., Xu, J., Takagi, S., Kim, D., et al.: Generalizable human gaussians for sparse view synthesis. In: European Conference on Computer Vision. pp. 451–468. Springer (2024)
20. Li, M., Yao, S., Xie, Z., Chen, K.: Gaussianbody: Clothed human reconstruction via 3d gaussian splatting. arXiv preprint arXiv:2401.09720 (2024)
21. Li, Y., Zhang, L., Qiu, Z., Jiang, Y., Li, N., Ma, Y., Zhang, Y., Xu, L., Yu, J.: NIMBLE: a non-rigid hand model with bones and muscles. ACM Transactions on Graphics **41**(4), 1–16 (2022)
22. Li, Z., Zheng, Z., Wang, L., Liu, Y.: Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19711–19722 (2024)
23. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(Nov), 2579–2605 (2008)
24. Mihajlovic, M., Zhang, Y., Black, M.J., Tang, S.: Leap: Learning articulated occupancy of people. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10461–10471 (2021)
25. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
26. Moon, G., Shiratori, T., Lee, K.M.: Deephandmesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In: European Conference on Computer Vision. pp. 440–455. Springer (2020)
27. Moon, G., Xu, W., Joshi, R., Wu, C., Shiratori, T.: Authentic hand avatar from a phone scan via universal hand model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2029–2038 (2024)

28. Moon, G., Yu, S.I., Wen, H., Shiratori, T., Lee, K.M.: InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In: European Conference on Computer Vision. pp. 548–564. Springer (2020)
29. Moreau, A., Song, J., Dhamo, H., Shaw, R., Zhou, Y., Pérez-Pellitero, E.: Human gaussian splatting: Real-time rendering of animatable avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 788–798 (2024)
30. Mundra, A., Wang, J., Habermann, M., Theobalt, C., Elgharib, M., et al.: Livehand: Real-time and photorealistic neural hand rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18035–18045 (2023)
31. Pang, H., Zhu, H., Kortylewski, A., Theobalt, C., Habermann, M.: ASH: Animatable gaussian splats for efficient and photoreal human rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1165–1175 (2024)
32. Pokhariya, C., Shah, I.N., Xing, A., Li, Z., Chen, K., Sharma, A., Sridhar, S.: MANUS: Markerless grasp capture using articulated 3D gaussians. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2197–2208 (2024)
33. Potamias, R.A., Ploumpis, S., Moschoglou, S., Triantafyllou, V., Zafeiriou, S.: Handy: Towards a high fidelity 3D hand shape and appearance model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4670–4680 (2023)
34. Qian, N., Wang, J., Mueller, F., Bernard, F., Golyanik, V., Theobalt, C.: HTML: A parametric hand texture model for 3D hand reconstruction and personalization. In: European Conference on Computer Vision. pp. 54–71. Springer (2020)
35. Qian, S., Kirschstein, T., Schoneveld, L., Davoli, D., Giebenhain, S., Nießner, M.: Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20299–20309 (2024)
36. Qian, Z., Wang, S., Mihajlovic, M., Geiger, A., Tang, S.: 3DGS-Avatar: Animatable avatars via deformable 3D gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5020–5030 (2024)
37. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: modeling and capturing hands and bodies together. ACM Transactions on Graphics **36**(6) (Nov 2017)
38. Shao, Z., Wang, Z., Li, Z., Wang, D., Lin, X., Zhang, Y., Fan, M., Wang, Z.: Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1606–1616 (2024)
39. Sun, Z., Shen, X., Hu, Y., Zhong, Y., Zhou, X.: JGHand: Joint-driven animatable hand avater via 3D gaussian splatting. arXiv preprint arXiv:2501.19088 (2025)
40. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
41. Wen, J., Zhao, X., Ren, Z., Schwing, A.G., Wang, S.: 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. pp. 2059–2069 (2024)
42. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)

43. Zhao, L., Lu, X., Fan, R., Im, S.K., Wang, L.: Gaussianhand: Real-time 3D gaussian rendering for hand avatar animation. IEEE Transactions on Visualization and Computer Graphics (2024)
44. Zheng, S., Zhou, B., Shao, R., Liu, B., Zhang, S., Nie, L., Liu, Y.: GPS-Gaussian: Generalizable pixel-wise 3D gaussian splatting for real-time human novel view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19680–19690 (2024)
45. Zheng, X., Wen, C., Su, Z., Xu, Z., Li, Z., Zhao, Y., Xue, Z.: OHTA: One-shot hand avatar via data-driven implicit priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 799–810 (2024)
46. Zielonka, W., Bagautdinov, T., Saito, S., Zollhöfer, M., Thies, J., Romero, J.: Drivable 3D gaussian avatars. arXiv preprint arXiv:2311.08581 (2023)