GASPACHO: GAUSSIAN SPLATTING FOR CONTROLLABLE HUMANS AND OBJECTS

Anonymous authors
Paper under double-blind review

Figure 1. **Left**. Using multi-view RGB images of human-object interactions, our method jointly reconstructs animatable models of the human and the object. **Right** The reconstructed model can be driven by control signals to synthesize photo-real *novel pairs* of human-object interactions with *free camera viewpoint control*.

ABSTRACT

We present GASPACHO, a method for generating photorealistic, controllable renderings of human–object interactions from multi-view RGB video. Unlike prior work that reconstructs only the human and treats objects as background, GASPACHO simultaneously recovers animatable templates for both the human and the interacting object as distinct sets of Gaussians, thereby allowing for controllable renderings of novel human object interactions in different poses from novel-camera viewpoints. We introduce a novel formulation that learns object Gaussians on an underlying 2D surface manifold rather than in 3D volume, yielding sharper, fine-grained object details for dynamic object reconstruction. We further propose a contact constraint in Gaussian space that regularizes human–object relations and enables natural, physically plausible animation. Across three benchmarks—BEHAVE, NeuralDome, and DNA-Rendering—GASPACHO achieves high-quality reconstructions under heavy occlusion and supports controllable synthesis of novel human–object interactions. We also demonstrate that our method allows for composition of humans and objects in 3D scenes and for the first time showcase that neural rendering can be used for the controllable generation of photoreal humans interacting with dynamic objects in diverse scenes.

1 INTRODUCTION

Photorealistic novel view synthesis of human avatars is essential for several applications ranging from telepresence, AR/VR, special effects to e-commerce. The recent emergence of Gaussian Splatting as a 3D representation (Kerbl et al., 2023) has enabled a new wave of efficient methods that reconstruct high-quality animatable human avatars (Moreau et al., 2024; Wen et al., 2024; Kocabas et al., 2023; Qian et al., 2023; Li et al., 2024b; Pang et al., 2023). These articulated human models can be animated by controllable human pose signals to deform the 3D Gaussians into a novel pose and then rendered from any camera viewpoint in real-time. However, these methods assume that humans are *recorded in isolation*, without any occlusions, and when they are animated, their actions are generated in *isolation*.

In this work, we use multi-view (sparse and dense) captures of human interacting with objects and aim to create animatable 3D Gaussians models for both the human and the object. The reconstructed models can be used to render the captured scene from novel viewpoints, but our main goal is to generate novel human-object interactions that can be integrated in 3D scenes and rendered in real-time. While there exists a body of work on human-object interaction (Guzov et al., 2021; Zhang et al., 2022; Hassan et al., 2019; Zhu et al., 2024; Hassan et al., 2021a), it focuses only on modeling human behaviour in environments and *neglects human appearance*. A few recent papers study the photorealistic rendering/reconstruction of human-object interaction, but assume the objects to be static (Kocabas et al., 2023; Xue et al., 2024), focus only on human-object reconstruction and are unable to synthesize *novel human-object interactions* (Sun et al., 2021; Jiang et al., 2022b; Wang et al., 2025a) or assume the object to be a small extension of one human hand (Zhan et al., 2024; Liu et al., 2023). The new problem we define is challenging in several ways:

First, Human-Object interactions inevitably creates significant occlusions of the human body and/or the object. We observe that these occlusions happen to be an important problem for existing human reconstruction pipelines that fail on these occluded regions and propose a composition and occlusion guided loss to address it.

Then, animatable human reconstruction methods are facilitated by pre-computed pose estimation of the SMPL (Loper et al., 2015) body template to initialize and track the human geometry. However, we do not assume that object geometry and tracking is known beforehand. Because objects are often small and occluded, obtaining an accurate template often leads to degenerate solutions with common techniques. Instead, we propose to reconstruct objects from scratch in a coarse-to-fine manner, that includes learning a template in features space, tracking it across frames with photometric alignment and finally use *pose-independent* Gaussian maps to learn high-fidelity textures.

Finally, beyond reconstruction, generating novel interactions from avatars and objects while maintaining a physically plausible contact is not straightforward. Having separated animatable models potentially enables interesting scenarios: reanimate the observed human and object with a novel interaction, replace one human by another to execute the same interaction, or both at the same time. However, determining the correct control parameters for these scenarios is difficult because they depend on the body shape of the human and the geometry of the object. We address this problem by introducing a novel human-object contact constraint in Gaussian space to generate interactions with natural human-object contacts.

We evaluate our method on DNA rendering (Cheng et al., 2023a), NeuralDome (Zhang et al., 2023), and BEHAVE datasets (Bhatnagar et al., 2022), showing significant improvement over existing methods for joint human-object reconstruction. DNA-Rendering provides dense camera setups, while evaluating on BEHAVE demonstrates the utility of our method for data collected with sparse (only four cameras) low-cost mobile, commodity sensors. We also show photorealistic demonstrations of novel interactions between avatars and objects reconstructed from different datasets and integrated in 3D Gaussians scenes. To the best of our knowledge, such demonstrations were not achieved before and are possible thanks to the multiple contributions presented in this paper.

To summarize, our most important contributions are: 1) We present a novel method to jointly reconstruct humans and objects models under occlusion, which can then be animated to synthesize novel human-object interactions. 2) We introduce a coarse-to-fine pipeline for reconstructing *dynamic* objects under interaction-induced occlusion, with features space template, 3D tracking and refinement with *pose-independent* Gaussian maps. 3) We introduce human-object contact constraints in Gaussian space to ensure proper contacts when animating 3DGS humans interacting with novel objects. 4) We demonstrate that our method allows for photorealistic and controllable composition of humans, objects and 3D scenes coming from diverse captures, an application that was not seen before to the best of our knowledge.

2 Related Work

Neural Rendering: Since the publication of NeRF (Mildenhall et al., 2020), there has been an explosion of interest in Neural Rendering (Xie et al., 2022b). Despite appealing image quality, NeRF is limited by its computational complexity. Though there have been several follow-up improvements (Müller et al., 2022; Barron et al., 2022; 2023; Tancik et al., 2023), the high computational cost of NeRF remains. 3DGS introduced by (Kerbl et al., 2023) addresses this limitation by repre-

senting scenes with an explicit set of primitives shaped as 3D Gaussians, extending previous work using spheres (Lassner & Zollhöfer, 2021). 3DGS rasterizes Gaussian primitives into images using a splatting algorithm (Westover, 1992). 3DGS originally designed for static scenes has been extended to dynamic scenes (Shaw et al., 2023; Luiten et al., 2024; Wu et al., 2024; Lee et al., 2024; Li et al., 2023a), slam-based reconstruction, (Keetha et al., 2024), mesh reconstruction (Huang et al., 2024; Guédon & Lepetit, 2024) and NVS from sparse cameras (Mihajlovic et al., 2024) and embodied views (Wang et al., 2025b).

Human Reconstruction and Neural Rendering: Mesh-based templates (Pavlakos et al., 2019; Loper et al., 2015) have been used to recover 3D human shape and pose from images and video (Bogo et al., 2016; Kanazawa et al., 2018). However, despite recent efforts and improvements (Alldieck et al., 2018; 2019; Shin et al., 2024), mesh-based representation does not often allow for photorealistic renderings. Implicit functions (Mescheder et al., 2019; Park et al., 2019) have also been utilized to reconstruct detailed 3D clothed humans (Chen et al., 2021; Alldieck et al., 2021; Saito et al., 2020; He et al., 2021; Huang et al., 2020; Deng et al., 2020). However, they are also unable to generate photorealistic renderings and are often not reposable. Several works (Peng et al., 2021; Guo et al., 2023; Weng et al., 2022; Jiang et al., 2022a; Habermann et al., 2023; Zhu et al., 2024; Li et al., 2022; Liu et al., 2021; Xu et al., 2021) build a controllable NeRF that produces photorealistic images of humans from input videos. However they inherit the speed and quality restrictions of the original NeRF formulation. Furthermore, unlike us, they do not model human-object interactions. With the advent of 3DGS, several recent papers use a 3DGS formulation (Kocabas et al., 2023; Qian et al., 2024; Moreau et al., 2024; Abdal et al., 2024; Zielonka et al., 2023; Moon et al., 2024; Li et al., 2024b; Pang et al., 2024; Lei et al., 2023; Hu et al., 2024; Li et al., 2024a; Zheng et al., 2024; Jiang et al., 2024b; Dhamo et al., 2024; Qian et al., 2023; Xu et al., 2024; Guo et al., 2025; Qiu et al., 2025; Junkawitsch et al., 2025; Moreau et al., 2025) to build controllable human or face avatars. Unlike our method, they do not model human-object interactions or reconstruct humans under occlusion. Prior works have also extended the 3DGS formulation to model humans along with their environment, (Xue et al., 2024; Zhan et al., 2024), but unlike us, they either assume that the 3D scene remains static or that its constituent parts are an extension of the human hand.

Human Object Interaction: Human Object Interaction is another recurrent topic of study in computer vision and graphics. Early works (Fouhey et al., 2014; Wang et al., 2017; Gupta et al., 2011) model affordances and human-object interactions using monocular RGB. The collection of several recent human-object interaction datasets (Hassan et al., 2021a; Guzov et al., 2021; Hassan et al., 2019; Savva et al., 2016; Taheri et al., 2020; Bhatnagar et al., 2022; Jiang et al., 2024a; Cheng et al., 2023b; Zhang et al., 2022) has allowed the computer vision community to make significant progress in joint 3D reconstruction of human-object interactions (Xie et al., 2022a; 2023; 2024; Zhang et al., 2020). These datasets have also led to the development of methods that synthesize object conditioned controllable human motion (Zhang et al., 2022; Starke et al., 2019; Hassan et al., 2021b; Diller & Dai, 2024). All these methods represent humans and objects as 3D meshes and as such inherit all the limitations of mesh-based representations including their inability to generate photorealistic images, while our method allows for photorealistic renderings of humans and objects.

3 METHOD

Our method receives a set of multi-view RGB images of humans interacting with objects, captured from N cameras at T temporal frames, $\{\mathbf{I}_t^c\}_{t=1,c=1}^{t=T,c=N}$, along with estimates of camera parameters $\{\gamma_t^c\}_{t=1,c=1}^{t=T,c=N}$ and human pose $\{\theta_t\}_{t=1}^{T}$. We aim to learn a deformable function $f(\theta, \phi; \gamma)$ that maps new human poses θ and object poses ϕ to RGB images rendered from novel viewpoints γ. **Intuitively**, we are interested in a function that allows us to map human and object poses to 3D Gaussians, which can be used to render novel human-object interactions with free camera viewpoint control.

We first use a feature-based representation to construct coarse object templates (Sec. 3.1) and track the template across the temporal frames (Sec. 3.2) to obtain estimates of object pose parameters. Using the estimated object template and the SMPL mesh we define 2D canonical human and object Gaussian maps to learn Gaussian properties (Sec. 3.3) of two sets of Gaussians $\mathcal{G}_O$ and $\mathcal{G}_H$ anchored to the canonical object and SMPL templates respectively. These are deformed using the provided human and estimated object pose parameters to posed space (Sec. 3.4) and learnt using an occlusion

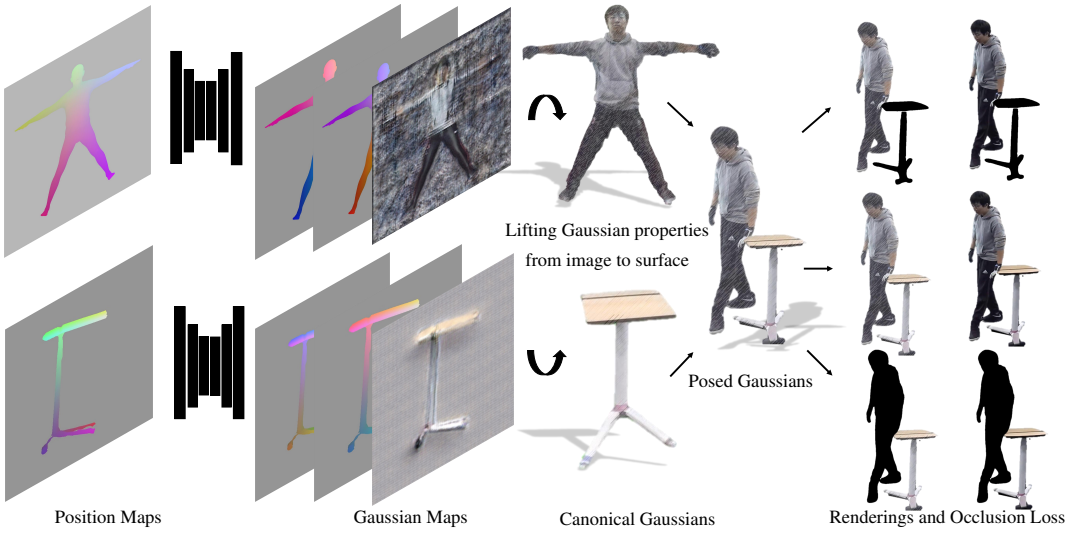

Figure 2: **Method Overview:** Using position maps as input, we learn Gaussian parameters for Gaussians anchored to canonical human and object templates. Gaussian properties - orientation, scale, opacity, color - are learnt as 2D maps structured according to a UV unwrapping of canonical human and object templates. Each pixel thus corresponds to a Gaussian in the canonical template. Once mapped to canonical Gaussians, these Gaussians are posed using LBS for the human and a rigid transform for the object. We render the posed object and human Gaussians separately and compare against the segmented portions of the target image. We further guide the reconstruction using known occlusion information. In the figure above, the black regions of the rendered images are masked out thereby allowing our method to deal with occlusions. In the 2D maps, the gray regions indicate pixels which don't map to any 3D Gaussians.

guided photometric loss (Sec. 3.5). Finally at test time, we introduce a human-object contact (Sec. 3.6) constraints when humans are animated interacting with novel objects.

3.1 Canonical Object Template

To model object motion, we first learn a canonicalised 3D object template. Direct optimization of 3DGS Kerbl et al. (2023) fails under naturally occurring human-object occlusions (Tab. 3), so we adopt feature-based planes Mihajlovic et al. (2024). The features act as a regulizer and smoothness prior which allows for better convergence under occlusion.

For feature-based object Gaussians reconstruction, we learn 3 - $M \times W$, l-dimensional planes $\mathbf{F} \in \mathbb{R}^{3\times M\times W\times l}$ - where l is the feature dimension, for the object. Each Gaussian location for the object $\mathbf{x}_p$ (where p indexes the Gaussian in the set of Object Gaussians) is projected onto the three feature planes to obtain feature vectors - using standard orthographic projection. This idea is analogous to the feature based representation introduced in Chan et al. (2021) and also used in Kocabas et al. (2023), but we show its utility for template reconstruction of occluded objects. These features are then concatenated along the feature dimension. We denote this operation - that maps object Gaussian location $\mathbf{x}_p$ to its corresponding sampled feature $\mathbf{f}_p$ - as $\mathbf{f}_p = \pi(\mathbf{x}_p; \mathbf{F}) \in \mathbb{R}^{3l}$. Using multiple small MLPs we map these feature vectors onto Gaussian parameters - color, covariance, scale, opacity. We learn the positions of the Gaussians directly; query the features at these positions and map these queried features to the rest of the Gaussian parameters. Hence, though we represent the canonical Gaussians using standard 3DGS parameters, we do not learn these parameters directly. Instead, we learn the network weights, feature grids and Gaussian locations $\mathbf{x}_p$ that are mapped onto these 3DGS parameters. Once the features are learnt, they can be discarded after querying them at gaussian positions to obtain other parameters. For an illustration of occlusion and feature based learning, we refer to the supplementary.

We pick a frame where the object is minimally occluded. To do this, we compute the number of object pixels (using object segmentation masks across all cameras) in the images at timestep t; of all the temporal frames, the frame with the most number of object pixels is considered to be the one with minimal occlusion. Using this frame t^* with minimal occlusion, we optimize $\mathcal{G}_O^* = \arg\min_{\mathcal{G}_O} \sum_{c=1}^{N} \|\mathcal{R}(\mathcal{G}_O; \gamma_c) - \mathbf{O}_{t^*}^c \cdot \mathbf{I}_{t^*}^c\|_2$, where $\mathcal{R}(.;\gamma)$ is a 3DGS rasterizer with camera parameters γ, $\mathbf{O}_t^c$ the object mask, and $\mathbf{I}_t^c$ the RGB image, i.e we optimize the 3DGS parameters

which match the segmented object pixels in one set of multiview images. This setup closely matches standard 3DGS reconstruction but with a feature based parameterization and segmented object images as targets. We want to highlight that we use the feature based parameterization detailed above - i.e we do learn the Gaussian properties - just not directly but parameterized by Gaussian positions and the features. This yields a coarse template $\mathcal{G}_O^*$.

3.2 Object Tracking

Once we have learnt a set of 3D Gaussians describing one frame $\mathcal{G}_O^*$, we want to optimize the transformations that map this canonical object to the renderings in subsequent timesteps t. Thus, we apply an optimizable rigid 6D pose transform ϕ to the object template $\mathcal{G}_O^*$, and minimize the per-pixel error between renderings and target images

$$\arg\min_{\phi_t} \sum_{c=1}^{N} ||\mathcal{R}(\mathcal{T}(\mathcal{G}_O^*; \phi_t); \gamma_c) - \mathbf{O}_t^c \cdot \mathbf{I}_t^c||_2 \,. \tag{1}$$

Here we define $\mathcal{T}(\mathcal{G}_O; \phi)$ to be a rigid transformation on the position and covariance of canonical object Gaussians. Specifically, for every canonical 3D Gaussian in the set $\mathcal{G}_O^*$, we transform its position $\mathbf{x}$ and covariance $\mathbf{\Sigma}$ attributes:

$$\mathbf{x}' = \mathbf{R}_\phi \mathbf{x} + \mathbf{t}_\phi, \quad \mathbf{\Sigma}' = \mathbf{R}_\phi \mathbf{\Sigma} \mathbf{R}_\phi^\top, \tag{2}$$

where $\mathbf{R}_\phi$ and $\mathbf{t}_\phi$ are the rotation and translation components of the 6D object pose ϕ. To ensure convergence, ϕ_t is initialized from ϕ_{t-1}.

3.3 Gaussian Maps

We use the SMPL body model to first define a canonical human Gaussian template $\mathcal{G}_H^*$ by densely sampling points on the SMPL mesh and setting initial color to zero, opacity and covariance to a fixed constant. Given the canonical templates $\mathcal{G}_H^*$ and $\mathcal{G}_O^*$, we now learn their Gaussian properties (position offsets, covariance, opacity, color) in image space using two StyleUNet models S_H for the human and S_O for the object. This 2D formulation leverages strong inductive biases of CNNs (Pang et al., 2024; Li et al., 2024b; Hu et al., 2023) while maintaining a one-to-one correspondence between pixels and 3D Gaussians.

Human maps (pose-dependent): For the human we project the SMPL body model in canonical pose from two front and back views (similiar to Fig. 3) to obtain canonical UV maps. Each pixel in the 2D map is assigned to a canonical 3D human Gaussian in $\mathcal{G}_H^*$. At timestep t with pose θ_t, to provide pose and shape information to a 2D CNN (following (Li et al., 2023b)), we first create pose-dependent position maps for the human by using LBS to deform the canonical SMPL mesh vertices without any global orientation or translation. Each deformed vertex is stored in the corresponding location of the UV map to define a posed position map (similiar to Fig. 3). A StyleUNet (Wang et al., 2023) S_H takes these maps as input (Fig. 2) and predicts per-pixel Gaussian properties: opacity, color, position, covariance as offsets from canonical human template Gaussian.

Object maps (pose-independent): For the object, we project the coarse canonical template $\mathcal{G}_O^*$ from front and back views (Fig. 3) to define mappings between a 2D image and 3D object Gaussians. To define fixed pose-independent position map we directly store the location of the Gaussian in the template at the 2D projected pixel location (Fig. 3). Unlike humans, this map does not vary with time and is reused at every timestep. To our knowledge, such pose-independent Gaussian maps for rigid objects are novel, and we find that they stabilize 6D pose optimization and yield more consistent renderings. A second StyleUNet S_O processes these maps and predicts offsets of Gaussian properties from the canonical template.

3.4 Deformation

The canonical Gaussians $\mathcal{G}_H^*$ and $\mathcal{G}_O^*$ are first updated by the StyleUNets (Sec. 3.3). For both humans and object, S_H and S_O predict offsets from the canonical templates: position, covariance, opacities, color; additionally S_H also predicts per-gaussian (human Gaussians) skinning weights $\mathbf{w}_k^H$. We use

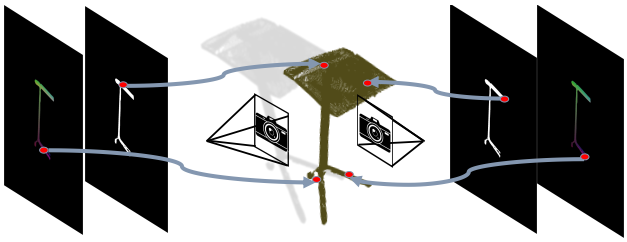

Figure 3: Object Gaussian Maps. Each pixel corresponds to a canonical Gaussian. Object maps are pose-independent, a novel design that yields high-fidelity textures. Arrows indicate correspondence. To define position-maps we store the 3D location of Object Gaussians corresponding to a pixel in the 2D map.

these offsets to update canonical Gaussians and refer to these updated Gaussians as $\mathcal{G}_H$ and $\mathcal{G}_O$. In practice the human canonical Gaussian change per frame.

We then deform these StyleUNet-modified Gaussians into posed space. For humans, Linear Blend Skinning (LBS) (Fig. 2) is applied using the learned weights $\mathbf{w}_k^H$, giving per-Gaussian transformations $\mathbf{R}_k^\theta = \sum_{j=1}^{J} \mathbf{w}_{kj}^H \mathbf{R}_j^{SMPL}(\theta)$. Here we use $\mathbf{R}_j^{SMPL}(\theta)$ to denote the j^{th} joint-transformation matrix of the SMPL body pose - we use k to denote the k^{th} human Gaussian.

For every canonical 3D human Gaussian in the set of canonical Gaussians $\mathcal{G}_H$, we transform its position $\mathbf{x}_k$ and covariance $\mathbf{\Sigma}_k$ attributes as:

$$\mathbf{x}'_k = \mathbf{R}_k^\theta \mathbf{x}_k + \mathbf{t}_k^\theta, \quad \mathbf{\Sigma}'_k = \mathbf{R}_k^\theta \mathbf{\Sigma}_k \mathbf{R}_k^{\theta\top}, \tag{3}$$

Here we use $\mathbf{R}_k^\theta$ and $\mathbf{t}_k^\theta$ to denote the rotational and translational components of the per-gaussian transformation matrix. For objects, rigidity is assumed and a global 6D transform with pose ϕ is applied - using $\mathcal{T}(\mathcal{G}_O; \phi)$ (defined in Sec. 3.2). This process yields posed human and object Gaussians (Fig. 2). In both cases - object and human - only positions and covariances are deformed, while colors and opacities remain as predicted by the StyleUNets.

3.5 Composition and Occlusion Guided Loss

Given input images, $\mathbf{I}_t^c$ with human and object masks $\mathbf{H}_t^c, \mathbf{O}_t^c$, and human and object poses: θ_t and ϕ_t - with camera index c, timestep t - we render posed human and object Gaussians $\mathcal{G}_H^t = \mathcal{T}(\mathcal{G}_H; \theta_t)$ and $\mathcal{G}_O^t = \mathcal{T}(\mathcal{G}_O; \phi_t)$ to obtain $\mathbf{I}_H^{ct} = \mathcal{R}(\mathcal{G}_H^t; \gamma_c)$, $\mathbf{I}_O^{ct} = \mathcal{R}(\mathcal{G}_O^t; \gamma_c)$, and $\mathbf{I}_A^{ct} = \mathcal{R}(\mathcal{G}_H^t \cup \mathcal{G}_O^t; \gamma_c)$. We use the input images and segmentation masks to supervise these renderings.

To avoid penalizing occluded regions, we ignore pixels where the other entity is visible. The per-frame, per-camera losses are $\mathcal{L}_H^{ct} = \|\mathbf{I}_t^c \mathbf{H}_t^c (1 - \mathbf{O}_t^c) - \mathbf{I}_H^{ct}(1 - \mathbf{O}_t^c)\|_1$, $\mathcal{L}_O^{ct} = \|\mathbf{I}_t^c \mathbf{O}_t^c (1 - \mathbf{H}_t^c) - \mathbf{I}_O^{ct}(1 - \mathbf{H}_t^c)\|_1$, and $\mathcal{L}_A^{ct} = \|\mathbf{I}_t^c (\mathbf{H}_t^c + \mathbf{O}_t^c) - \mathbf{I}_A^{ct}\|_1$.

The overall image loss is $\mathcal{L}_1 = \sum_{c,t} (\mathcal{L}_H^{ct} + \mathcal{L}_O^{ct} + \mathcal{L}_A^{ct})$. We additionally use a perceptual loss $\mathcal{L}_{per}$ (Zhang et al., 2018) with the same masking strategy and a regularization loss $\mathcal{L}_{reg}$ that encourages predicted skinning weights to remain close to SMPL-derived weights. The final objective is $\mathcal{L} = \lambda_{L1}\mathcal{L}_1 + \lambda_{per}\mathcal{L}_{per} + \lambda_{reg}\mathcal{L}_{reg}$.

This formulation prevents erroneous supervision in occluded regions, while multi-view consistency ensures that ignored Gaussians are still supervised in other frames or views. Furthermore our *pose-independent* Gaussian map formulation allows us to optimize the 6D object pose during training as well and hence refine the initially tracked object 6D parameters.

3.6 Gaussian Human-Object Contact Refinement

A key novelty of our framework is that it enables reanimating a reconstructed human identity with novel motions and novel objects. Given new motion parameters θ_t and ϕ_t, the human StyleUNet S_H and object StyleUNet S_O are used to predict canonical Gaussian $\mathcal{G}_O$ and $\mathcal{G}_H$ These are deformed with LBS for humans and a rigid transform for objects to obtain posed Gaussians $\mathcal{G}_H^t = \mathcal{T}(\mathcal{G}_H; \theta_t)$ and $\mathcal{G}_O^t = \mathcal{T}(\mathcal{G}_O; \phi_t)$. The naive composition of these two sets of Gaussians often causes penetrations or missing human-object contacts, so we further regularize these two sets of Gaussians by optimizing displacements $\mathbf{\Delta}$ for human contact Gaussians. However unlike SMPL, Gaussians Avatars

Figure 4: **Top Row:** Qualitative Results of our method on the DNA-Rendering (left), BEHAVE (middle) and Neural Dome (right) Datasets using the reconstructed humans and objects in *novel human and object poses*. **Bottom Row:** Reconstructed humans animated interacting with *novel objects* in *novel human poses* Note this is a novel application not possible with existing methods.

are not in correspondence to a fixed template. Therefore, we cannot manually define contact Gaussian primitives on the human avatar. Instead, to identify contact Gaussians, we define SMPL contact vertices V_c^{SMPL} (feet, hips, hands) and assign them to canonical Gaussians by nearest-neighbour search $i^\star = \arg\min_k \|\mathbf{x}_k^{\mathsf{C}} - \mathbf{u}\|_2$ with $\mathbf{u} \in V_c^{\text{SMPL}}$, yielding the index set $\mathcal{I}_c$ - which is a subset of the human Gaussians. Contact frames t are detected from SMPL motion cues (low vertical velocity and downward acceleration). We find displacements $\mathbf{\Delta}$ for contact Gaussians. with the objective

$$\mathcal{L} = \lambda_c \sum_{k \in \mathcal{I}_c} \|\mathbf{x}_{k,t}^H - \mathbf{x}_{p(k,t)}^O\|_2^2 + \lambda_p \sum_{k \notin \mathcal{I}_c} h_r(d_\beta(\mathbf{x}_{k,t}^H, \mathcal{G}_O^t))^2 + \lambda_r \|\mathbf{\Delta}\|_2^2 + \lambda_t \|\mathbf{\Delta}_t - \mathbf{\Delta}_{t-1}\|_2^2 \quad (4)$$

where $\mathbf{x}_{p(k,t)}^O$ is the nearest object Gaussian center, $d_\beta(\mathbf{x}) = -\frac{1}{\beta} \log \sum_{p=1}^{N_O} \exp(-\beta \|\mathbf{x} - \mathbf{x}_p^O\|)$ is the soft nearest-neighbour distance to object Gaussians, and $h_r(d) = \max(0, r - d)$ is a hinge margin enforcing separation for non-contact Gaussians. Here we use the sub-index t to denote the timestep.

Intuitively, this optimization forces contact Gaussians to snap onto the object surface when contact is expected ($\delta = 1$), pushes non-contact Gaussians at least a margin r away when contact is absent ($\delta = 0$), penalizes large displacements through $\|\mathbf{\Delta}\|_2^2$, and encourages temporal smoothness with $\|\mathbf{\Delta}_t - \mathbf{\Delta}_{t-1}\|_2^2$. The human Gaussian positions are moved as $\mathbf{x}_{k,t}^H + \mathbf{\Delta}_{k,t}$ yielding a refined set of Gaussians $\tilde{\mathcal{G}}_H^t$ These are then composited with $\mathcal{G}_O^t = \mathcal{T}(\mathcal{G}_O; \phi_t)$ and jointly rendered, producing photorealistic human–object interactions with coherent contacts.

3.7 Learning and Network Initialization

As all our StyleUnets learn offsets from canonical template, we first train the human StyleUnet for approximately 2000 iterations to map randomly sampled human-position maps to predict zero offsets. We also train the object StyleUnet for a similiar number of iterations to map the pose-independent position map to zero offsets. This ensures that the network predictions start off from zero and all deformation is learnt on top of the Canonical templates. We use the Adam optimizer for all optimization. For learning features (for the canonical Object template) we use two separate MLPs - one maps to geometry parameters - scale, covariance and the other MLP to appearance parameters - color, opacity.

4 Experiments

In this section we compare our method with recent methods that reconstruct *only* animatable 3D humans from RGB images (Li et al., 2024b; Zhan et al., 2024; Moon et al., 2024; Kocabas et al.,

Figure 5: Quantiative comparison with unmodified existing methods for animatable human+object reconstruction on Existing methods are unable to reconstruct animatable objects.

Figure 6: Gaussian map ablation. Naive 3DGS fails due to occlusion. See Supp Mat for illustration

2023). All existing methods are designed to work only with humans recorded in isolation with *no occlusion* and are *not designed to handle objects*. We conduct two sets of experiments (Sec. 4.1): in the first we use the unmodifed baselines to reconstruct both humans and objects and demonstrate that they fail without significant modification. In the second experiment, we add the object template to the strongest Human reconstruction baseline ((Li et al., 2023b)) without our *Gaussian Map* object formulation. In this experiment, while the human reconstruction quality is comparable to ours, the object is clearly blurry. We then ablate (Sec. 4.3) the various components of our method. In Sec. 4.2, we demonstrate the key novelty of our framework and show how avatars from different datasets can be animated with novel objects to synthesize novel human-object interactions and integrated in 3D scenes to create renderings of avatars interacting with dynamic objects in varied environments.

Evaluation Dataset: We use two datasets of human-object interaction: **BEHAVE dataset (Bhatnagar et al., 2022).** It is captured using four Kinect cameras. We use a subset of each sequence for training and test our method on the remaining held-out frames. **DNA-Rendering (Cheng et al., 2023b).** A subset of the DNA-Rendering dataset records people interacting with objects in a studio with 60 cameras. Of the 60 cameras, we use 48 for training and evaluate on the rest.

4.1 BASELINES

Unmodified Baselines: We compare our approach with ExAvatar (Moon et al., 2024), Animatable-Gaussians (Li et al., 2024b), HUGS (Kocabas et al., 2023) and ToMiE (Zhan et al., 2024). As these methods are primarily designed to reconstruct and animate humans recorded in isolation, they need to be modified to deal with objects for a fair comparison. In this experiment, to reconstruct both humans and objects together, we modify the segmentation masks provided in the datasets to include the object along with the humans. All compared methods reconstruct a human using 3DGS, with a canonical space deformed via LBS to posed space and supervised against ground truth pixels. Since

Subject:	Human			Object			Full		
Metric:	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
HUGS (Kocabas et al., 2023)	22.13	0.7674	43.71	17.54	0.6878	45.21	21.13	0.7180	42.47
ExAvatar (Moon et al., 2024)	25.41	0.7743	37.06	16.20	0.5752	37.91	23.18	0.7632	35.58
ToMiE (Zhan et al., 2024)	26.85	0.7700	36.60	20.30	0.6581	35.30	25.49	0.7790	32.33
AnimGau (Li et al., 2024b)	27.85	0.8900	32.60	23.21	0.7550	37.40	26.49	0.8656	31.43
AnimGaus+Obj (mod)	28.16	0.9090	30.60	25.81	0.8430	31.40	27.49	0.8956	30.43
Ours	28.14	0.9174	28.88	28.63	0.8973	29.77	28.31	0.9242	29.24

Table 1: Quantitative Results on DNA-Rendering evaluated on held-out images. We outperform existing baselines that reconstruct animatable 3D humans — including baselines modified for object reconstruction.

(a) w/o occl., with occl. (b) i) Ours ii) AnimGaus+Obj (c) w/o and w contact constraint

Figure 7: Ablation and modified baseline comparison.

Subject:	Human			Object			Full		
Metric:	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
HUGS (Kocabas et al., 2023)	23.11	0.8274	33.95	19.34	0.6378	47.43	22.12	0.7434	39.47
ExAvatar (Moon et al., 2024)	25.41	0.8130	34.06	18.35	0.6752	38.99	22.18	0.7820	36.58
ToMiE (Zhan et al., 2024)	23.85	0.7840	34.60	18.51	0.6781	39.40	21.49	0.7260	37.43
AnimGau (Li et al., 2024b)	26.85	0.8700	26.60	21.50	0.7581	34.40	24.99	0.8160	29.43
AnimGaus+Obj (mod)	27.65	0.8710	28.70	24.50	0.8431	33.40	26.11	0.8130	29.93
Ours	27.64	0.8741	28.46	26.39	0.8732	32.47	26.92	0.8724	29.24

Table 2: Quantitative Results on the BEHAVE Dataset, evaluated on held-out images not used in training (all human and object poses are unseen in training) We outperform existing baselines that reconstruct animatable 3D humans under occlusion and in presence of objects - including baselines modified for object reconstruction

we force all these methods to reconstruct the object as well, they are forced to modify the canonical Gaussians even to explain pixels detached from the human body. This results in significant artifacts. The object either disappears or is reconstructed as a blob-like surface, as shown in Fig. 5.

We report standard metrics computed on the BEHAVE and DNA-Rendering datasets in Tab. 2 and Tab. 1. To evaluate the reconstruction quality of the human and object, we also report separate standard image metrics for the human and object. Please note that all metrics are computed for *novel poses* on held out *test images unseen during training*. For the DNA-Rendering dataset we report metrics on novel views while for BEHAVE the training and test camera are same as only 4 cameras are available. In Fig. 4 we show results of our method on BEHAVE, DNA-Rendering and NeuralDome.

Why no comparison with DynamicGaussians (Luiten et al., 2024)? Note that we focus on animatable humans and objects - while DynamicGaussians and follow ups do reconstruct human-object interaction, these are not controllable. Existing interactions can be replayed from novel views while our method allows for synthesizing humans interacting with novel objects in varied environments.

Modified Baseline for Objects: We also conduct a comparison with AnimGaussians (the strongest human reconstruction baseline) where the human and object are reconstructed separately - creating a stronger version of an animatable human baseline. Since we need an object template for object reconstruction, we use the one reconstructed and tracked using the feature representation, without the *pose-independent* Gaussian Map formulation from our method. In this experiment, while the human reconstruction quality is comparable to ours, without our Gausian Map formulation, the object is clearly blurry (Fig.7b). We report PSNR for this experiment on BEHAVE and DNA for novel-pose synthesis in Tab 2 and 1 under the row **AnimGaus+Obj (mod)**.

Subject:	DNA-01			DNA-02			BEH-01		
Metric:	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
w/o Features	26.11	0.9674	40.95	27.54	0.9678	46.43	25.00	0.9218	59.47
w/o Obj Map	30.41	0.9743	24.06	33.20	0.9752	28.99	26.18	0.9232	35.58
w/o Occ Loss	29.12	0.9727	26.58	32.94	0.9695	36.04	26.63	0.9141	41.76
w/o Contact	32.34	0.9664	20.91	32.63	0.9673	25.87	27.44	0.9674	28.86
Full	32.64	0.9774	20.88	33.63	0.9773	25.77	27.64	0.9574	28.46

Table 3: Ablation Study regarding the various components of our method on the DNA-Rendering and BEHAVE datset. As the table indicates, each component in our method adds to the performance and our method cumulatively yields best results. Each line the table corresponds to an experiment detailed in Sec. 4.3

Figure 8: Our method allows for combination of dynamic object movements with avatars reconstructed without objects (here from AvatarX) and placed in 3D scenes.

4.2 Animating Humans with Novel Objects and Composition with Scenes

Our framework allows reanimating a reconstructed human identity reconstructed from one set of videos with object Gaussians, human poses and object poses of another sequence. This allows, *for the first time* in the context of photo-real neural rendering, retargeting of human object interaction from one sequence to another. We find it necessary to optimize for contacts to ensure human-object interaction remains plausible (see Sec. 3.6) In Fig. 4, we qualitatively demonstrate this ability of our method by animating a reconstructed avatar to interact with novel objects. In Fig. 8, we demonstrate that these dyanmic interactions can be placed in diverse 3D environments and thus for the first time creating photoreal virtual Avatars that interact with dynamic objects in their environments.

4.3 Ablation Studies

In Tab. 3, we evaluate our design choices. Each method component improves the results. **Contact Constraint:** Here we evaluate contact constraint utility on same human+object using GT images but for novel poses. See **w/o Contact** in Tab. 3. The improvement for same human+object is minimal but pronounced for novel human-object interaction. See Fig. 7c. **Obj Gaussian Map:** In this setting, we do not refine the final object reconstructions using Gaussian maps but generate final renderings using the coarse object template. As shown in Tab. 3 (**w/o Obj Map**) and Fig. 6, the Gaussian Map based Gaussian parameter prediction allows for the reconstruction of high-frequency details. **Features:** Here we optimize the locations of the 3DGS parameters directly. Naive 3DGS fails due to significant occlusion patterns in the captured videos. See **w/o Features** experiment in Tab. 3. Without features the template reconstruction does not converge. Please see Supp Mat for illustration. **Occlusion Loss:** We switch off the occlusion-aware loss. This forces the occluded regions of the human and object in the rendered image to be part of the background, which degrades reconstruction quality as shown in Fig. 7a and the **w/o Occ Loss** experiment reported in Tab. 3.

5 Conclusion

We have presented GASPACHO, a method to reconstruct photorealistic human and object interactions from multi-view 2D images, which is capable of human, object and camera pose control. Different from prior art, which only generates photo-real *human actions in empty space* our approach completely decouples the object from the human, enables independent object and human pose control and *for the first time allows for the synthesis of photo-real human animations interacting with novel, previously unseen objects* and *placement of such interactions in diverse 3D scenes* and thus for the first time creating photoreal virtual Avatars that interact with dynamic objects in their environments. We evaluate GASPACHO quantitative and qualitatively on BEHAVE and DNA-Rendering datasets. Our method does make assumptions about the nature of the 3D scenes: i.e we only model one dynamic object and assume that its motion can be explained using only a rigid transform. Future extensions include the reconstruction of humans and objects not only in a lab setting but also from monocular RGB videos collected in the wild. We also hope to extend our work to build autonomous agents that interact not only with small objects but also with large 3D scenes.

REFERENCES

Rameen Abdal, Wang Yifan, Zifan Shi, Yinghao Xu, Ryan Po, Zhengfei Kuang, Qifeng Chen, Dit-Yan Yeung, and Gordon Wetzstein. Gaussian shell maps for efficient 3d human generation. In *Proceedings of CVPR*, 2024.

Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2019.

Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. imghum: Implicit generative models of 3d human shape and articulated pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5461–5470, 2021.

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5470–5479, 2022.

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. *ICCV*, 2023.

Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15935–15946, 2022.

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science. Springer International Publishing, October 2016.

Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *arXiv*, 2021.

Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *International Conference on Computer Vision (ICCV)*, 2021.

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

Wei Cheng, Ruixiang Chen, Wanqi Yin, Siming Fan, Keyu Chen, Honglin He, Huiwen Luo, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, Daxuan Ren, Lei Yang, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, Bo Dai, and Kwan-Yee Lin. Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. *arXiv preprint*, arXiv:2307.10173, 2023b.

Boyang Deng, John P Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Nasa neural articulated shape approximation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pp. 612–628. Springer, 2020.

Helisa Dhamo, Yinyu Nie, Arthur Moreau, Jifei Song, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Headgas: Real-time animatable head avatars via 3d gaussian splatting. *ECCV*, 2024.

Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2024.

David F Fouhey, Vincent Delaitre, Abhinav Gupta, Alexei A Efros, Ivan Laptev, and Josef Sivic. People watching: Human actions as a cue for single view geometry. *International journal of computer vision*, 110(3):259–274, 2014.

Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. *CVPR*, 2024.

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

Chen Guo, Junxuan Li, Yash Kant, Yaser Sheikh, Shunsuke Saito, and Chen Cao. Vid2avatar-pro: Authentic avatar from videos in the wild via universal prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025.

Abhinav Gupta, Scott Satkin, Alexei A Efros, and Martial Hebert. From 3d scene geometry to human workspace. In *CVPR 2011*, pp. 1961–1968. IEEE, 2011.

Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2021.

Marc Habermann, Lingjie Liu, Weipeng Xu, Gerard Pons-Moll, Michael Zollhoefer, and Christian Theobalt. Hdhumans: A hybrid approach for high-fidelity digital humans. 6(3), aug 2023. doi: 10.1145/3606927. URL https://doi.org/10.1145/3606927.

Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *Proceedings International Conference on Computer Vision*, pp. 2282–2292. IEEE, October 2019. URL https://prox.is.tue.mpg.de.

Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. Stochastic Scene-Aware motion prediction. August 2021a.

Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3D scenes by learning human-scene interaction. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 14708–14718, June 2021b.

Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11046–11056, 2021.

Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. *arXiv preprint arXiv:2312.02134*, 2023.

Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 634–644, 2024.

Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery, 2024. doi: 10.1145/3641519.3657428.

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

Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1737–1747, 2024a.

Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video, 2022a. URL https://arxiv.org/abs/2203.12575.

Yuheng Jiang, Suyi Jiang, Guoxing Sun, Zhuo Su, Kaiwen Guo, Minye Wu, Jingyi Yu, and Lan Xu. Neuralhofusion: Neural volumetric rendering under human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6155–6165, 2022b.

Yuheng Jiang, Zhehao Shen, Yu Hong, Chengcheng Guo, Yize Wu, Yingliang Zhang, Jingyi Yu, and Lan Xu. Robust dual gaussian splatting for immersive human-centric volumetric videos. *arXiv preprint arXiv:2409.08353*, 2024b.

Hendrik Junkawitsch, Guoxing Sun, Heming Zhu, Christian Theobalt, and Marc Habermann. Eva: Expressive virtual avatars from multi-view videos. pp. 1–11, 2025.

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

Nikhil Keetha, Jay Karhade, Krishna Murthy Jatavallabhula, Gengshan Yang, Sebastian Scherer, Deva Ramanan, and Jonathon Luiten. Splatam: Splat, track & map 3d gaussians for dense rgb-d slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats. *arXiv preprint arXiv:2311.17910*, 2023.

Christoph Lassner and Michael Zollhöfer. Pulsar: Efficient sphere-based neural rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

Junoh Lee, ChangYeon Won, Hyunjun Jung, Inhwan Bae, and Hae-Gon Jeon. Fully explicit dynamic guassian splatting. In *Proceedings of the Neural Information Processing Systems*, 2024.

Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. *arXiv preprint arXiv:2311.16099*, 2023.

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

Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. Spacetime gaussian feature splatting for real-time dynamic view synthesis. *arXiv preprint arXiv:2312.16812*, 2023a.

Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. *arXiv preprint arXiv:2311.16096*, 2023b.

Zhe Li, Yipengjing Sun, Zerong Zheng, Lizhen Wang, Shengping Zhang, and Yebin Liu. Animatable and relightable gaussians for high-fidelity human avatar modeling. *arXiv preprint arXiv:2311.16096v4*, 2024a.

Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Jia-Wei Liu, Yan-Pei Cao, Tianyuan Yang, Zhongcong Xu, Jussi Keppo, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Hosnerf: Dynamic human-object-scene neural radiance fields from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18483–18494, 2023.

Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Trans. Graph.(ACM SIGGRAPH Asia)*, 2021.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015. doi: 10.1145/2816795.2818013.

Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *3DV*, 2024.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4460–4470, 2019.

Marko Mihajlovic, Sergey Prokudin, Siyu Tang, Robert Maier, Federica Bogo, Tony Tung, and Edmond Boyer. SplatFields: Neural gaussian splats for sparse 3d and 4d reconstruction. In *European Conference on Computer Vision (ECCV)*. Springer, 2024.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

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

Arthur Moreau, Jifei Song, Helisa Dhamo, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. 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.

Arthur Moreau, Mohammed Brahimi, Richard Shaw, Athanasios Papaioannou, Thomas Tanay, Zhensong Zhang, and Eduardo Pérez-Pellitero. Better together: Unified motion capture and 3d avatar reconstruction. *arXiv preprint arXiv:2503.09293*, 2025.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. doi: 10.1145/3528223.3530127. URL https://doi.org/10.1145/3528223.3530127.

Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. Ash: Animatable gaussian splats for efficient and photoreal human rendering. *arXiv preprint arXiv:2312.05941*, 2023.

Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. Ash: Animatable gaussian splats for efficient and photoreal human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1165–1175, 2024.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 165–174, 2019.

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

Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, 2021.

Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. *arXiv preprint arXiv:2312.09228*, 2023.

Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. 2024.

Lingteng Qiu, Shenhao Zhu, Qi Zuo, Xiaodong Gu, Yuan Dong, Junfei Zhang, Chao Xu, Zhe Li, Weihao Yuan, Liefeng Bo, et al. Anigs: Animatable gaussian avatar from a single image with inconsistent gaussian reconstruction. In *CVPR*, 2025.

Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, 2020.

Manolis Savva, Angel X. Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. PiGraphs: Learning Interaction Snapshots from Observations. *ACM Transactions on Graphics (TOG)*, 35 (4), 2016.

Richard Shaw, Jifei Song, Arthur Moreau, Michal Nazarczuk, Sibi Catley-Chandar, Helisa Dhamo, and Eduardo Perez-Pellitero. Swags: Sampling windows adaptively for dynamic 3d gaussian splatting. *arXiv preprint arXiv:2312.13308*, 2023.

Jisu Shin, Junmyeong Lee, Seongmin Lee, Min-Gyu Park, Ju-Mi Kang, Ju Hong Yoon, and Hae-Gon Jeon. Canonicalfusion: Generating drivable 3d human avatars from multiple images. In *European Conference on Computer Vision*, pp. 38–56. Springer, 2024.

Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Trans. Graph.*, 38(6), November 2019. ISSN 0730-0301. doi: 10.1145/3355089.3356505. URL https://doi.org/10.1145/3355089.3356505.

Guoxing Sun, Xin Chen, Yizhang Chen, Anqi Pang, Pei Lin, Yuheng Jiang, Lan Xu, Jingyi Yu, and Jingya Wang. Neural free-viewpoint performance rendering under complex human-object interactions. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4651–4660, 2021.

Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020. URL https://grab.is.tue.mpg.de.

Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, SIGGRAPH '23, 2023.

Lizhen Wang, Xiaochen Zhao, Jingxiang Sun, Yuxiang Zhang, Hongwen Zhang, Tao Yu, and Yebin Liu. Styleavatar: Real-time photo-realistic portrait avatar from a single video. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023.

Weiquan Wang, Jun Xiao, Yueting Zhuang, and Long Chen. Physics-aware human-object rendering from sparse views via 3d gaussian splatting, 2025a. URL https://arxiv.org/abs/2503.09640.

Xiaolong Wang, Rohit Girdhar, and Abhinav Gupta. Binge watching: Scaling affordance learning from sitcoms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2596–2605, 2017.

Xiaoyuan Wang, Yizhou Zhao, Botao Ye, Xiaojun Shan, Weijie Lyu, Lu Qi, Kelvin C. K. Chan, Yinxiao Li, and Ming-Hsuan Yang. Holigs: Holistic gaussian splatting for embodied view synthesis, 2025b. URL https://arxiv.org/abs/2506.19291.

Jing Wen, Xiaoming Zhao, Zhongzheng Ren, Alex Schwing, and Shenlong Wang. GoMAvatar: Efficient Animatable Human Modeling from Monocular Video Using Gaussians-on-Mesh. In *CVPR*, 2024.

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

Lee Alan Westover. *Splatting: a parallel, feed-forward volume rendering algorithm*. PhD thesis, USA, 1992.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20310–20320, June 2024.

Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *European Conference on Computer Vision (ECCV)*. Springer, October 2022a.

Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single rgb camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

Xianghui Xie, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Template free reconstruction of human-object interaction with procedural interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10003–10015, 2024.

Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. *Computer Graphics Forum*, 2022b. ISSN 1467-8659. doi: 10.1111/cgf.14505.

Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *Advances in Neural Information Processing Systems*, 34:14955–14966, 2021.

Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1931–1941, 2024.

Lixin Xue, Chen Guo, Chengwei Zheng, Fangjinhua Wang, Tianjian Jiang, Hsuan-I Ho, Manuel Kaufmann, Jie Song, and Hilliges Otmar. HSR: holistic 3d human-scene reconstruction from monocular videos. In *European Conference on Computer Vision (ECCV)*, 2024.

Yifan Zhan, Qingtian Zhu, Muyao Niu, Mingze Ma, Jiancheng Zhao, Zhihang Zhong, Xiao Sun, Yu Qiao, and Yinqiang Zheng. Tomie: Towards modular growth in enhanced smpl skeleton for 3d human with animatable garments, 2024. URL `https://arxiv.org/abs/2410.08082`.

Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *European Conference on Computer Vision (ECCV)*, 2020.

Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In *CVPR*, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. *European Conference on Computer Vision (ECCV)*, October 2022.

Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Heming Zhu, Fangneng Zhan, Christian Theobalt, and Marc Habermann. Trihuman: A real-time and controllable tri-plane representation for detailed human geometry and appearance synthesis. *ACM Trans. Graph.*, September 2024. ISSN 0730-0301. doi: 10.1145/3697140. URL https://doi.org/10.1145/3697140.

Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3d gaussian avatars. *arXiv preprint arXiv:2311.08581*, 2023.