# *GART*: Gaussian Articulated Template Models

Jiahui Lei[1] Yufu Wang[1] Georgios Pavlakos[2] Lingjie Liu[1] Kostas Daniilidis[1,3]
[1] University of Pennsylvania [2] UC Berkeley [3] Archimedes, Athena RC
{leijh, yufu, lingjie.liu, kostas}@cis.upenn.edu, pavlakos@berkeley.edu

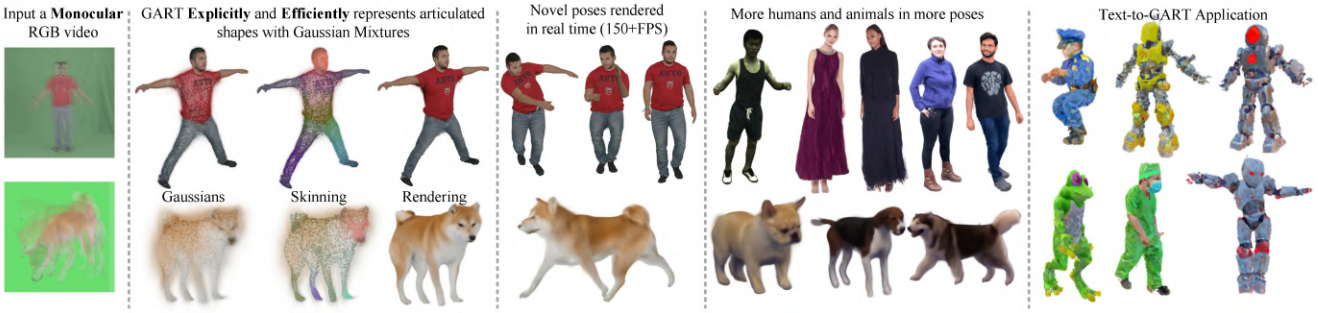


Figure 1. We propose *GART*, an explicit, efficient, and expressive model for articulated object capturing and rendering from monocular videos. Code available on the project page https://www.cis.upenn.edu/~leijh/projects/gart/.

## Abstract

*We introduce Gaussian Articulated Template Model (GART), an **explicit**, **efficient**, and **expressive** representation for non-rigid articulated subject capturing and rendering from monocular videos. GART utilizes a mixture of moving 3D Gaussians to explicitly approximate a deformable subject's geometry and appearance. It takes advantage of a categorical template model prior (SMPL, SMAL, etc.) with learnable forward skinning while further generalizing to more complex non-rigid deformations with novel latent bones. GART can be reconstructed via differentiable rendering from monocular videos in seconds or minutes and rendered in novel poses faster than 150fps.*

## 1. Introduction

Humans and animals are the most common deformable entities in real-world dynamic scenes, hence the plethora of approaches to modeling their geometry, appearance, and motion. This paper studies how to represent and reconstruct such deformable subjects from **monocular** videos. Since they share category-level structures, morphable template models are developed and widely applied, such as SMPL [42] for humans and SMAL [95] for quadrupedal animals. While they are useful for pose estimation, categorical template models cannot capture detailed appearance and geometry during a variety of deformations.

Recent studies proposed to address this problem by building additional implicit representations on templates in order to model geometry deformation and appearance. Most of these representations are based on neural fields [46, 47, 50, 75]. **Implicit** representations enhance quality but suffer from slow rendering because of costly query operations. Animating neural fields is challenging and requires specialized forward or backward skinning [8, 9]. Moreover, these methods usually depend on accurate template pose estimation since they can easily create artifacts in the empty space when the stereo is wrong. On the contrary, **explicit** representations [1, 18, 89] are efficient to render, easy to deform, and more robust to pose estimation errors because of the deformation-based optimization process. However, explicit representations often have sub-optimal quality and are restricted by fixed mesh topology [1], constrained by using too many points [89], or heavily rely on the multi-view studio camera system [18].

Our main insight into the articulation modeling was that **an explicit approximation for the implicit radiance field** would overcome the weaknesses of both worlds. We propose Gaussian Articulated Template Models (*GART*), a new renderable representation for non-rigid articulated subjects. *GART* takes advantage of classical template models by using its kinematic skeleton and models the detailed appearance via a Gaussian Mixture Model (GMM) in the canonical space that approximates the underlying radiance field (Sec. 3.2). Because a GMM does not have a fixed topology and each component can smoothly approximate a neighborhood, *GART* is as expressive as NeRFs while maintaining simplicity and interpretability.

As an explicit representation, *GART* can be animated via forward skinning similar to template meshes. However, the

predefined skeleton cannot capture the movements of loose clothes such as long dresses. We address this challenge with a novel latent bone approach, where a set of unobserved latent bones, as well as their skinning weights that drive the additional deformation, can be simultaneously learned from a monocular video (Sec. 3.3). Another challenge of the GMM approximation is the lack of local smoothness compared to neural fields, which impacts the reconstruction quality when the input views are sparse or the input human poses are noisy. We introduce smoothness priors for modeling articulated subjects to adapt *GART* for monocular reconstruction (Sec. 3.4).

To capture an articulated subject from monocular video, we initialize *GART* with the estimated template, and render the GMM with 3D Gaussian Splatting [27, 96] to reconstruct each frame. The optimization process gradually updates each Gaussian parameter and operates like a deformation-based approach that behaves more robustly under errors in the template pose estimations. With the explicit, efficient, and expressive *GART*, we are able to reconstruct a human avatar from a monocular video in 30 seconds and render it with resolution $540 \times 540$ at $150+$ FPS on a laptop, to our current knowledge, faster than any state-of-the-art NeRF-based human rendering methods. Furthermore, we use *GART* as a general framework to reconstruct animals from monocular videos in the wild with higher fidelity than previous mesh-based approaches.

In summary, our main contributions are: 1) *GART*, a general and explicit representation for non-rigid articulated subjects, which approximates the radiance field of the canonical shape and appearance with a Gaussian Mixture Model; 2) *GART* can be efficiently animated via learnable forward skinning and can capture challenging deformations such as loose clothes on humans via a novel latent bones approach; 3) Our experiments demonstrate that GART achieves SoTA performance in monocular human reconstruction and rendering on various datasets with the best training and inference efficiency and produces high-quality animal reconstruction from monocular videos in the wild.

## 2. Related Work

**3D Human Reconstruction**. Reconstructing 3D humans from monocular observations is a difficult task due to depth ambiguity. Parametric template models [39, 42, 51, 58] provide a strong prior of the human body and are key in the recent advances of monocular 3D human pose and shape reconstruction [15, 26, 29, 30, 64, 65, 69]. The explicit and predefined topology of parametric meshes, however, cannot capture personalized appearance such as texture, hair, and clothing [1, 2, 18–20]. To address this issue, recent studies [7–9, 11–13, 16, 17, 21–25, 31, 33, 38, 40, 49, 52, 53, 60, 62, 63, 68, 71, 76, 77, 86, 90, 92] propose to use neural representations, such as NeRF, to capture high-fidelity humans from multiple views or videos. To reconstruct dynamic humans, neural representations are combined with parametric models to disentangle pose and shape [40, 55, 71]. Appearance can be modeled in the canonical space and then posed by the articulated template [8, 9]. These hybrid approaches allow re-animation of the captured avatar and demonstrate high flexibility to model personalized details. However, one drawback is their inefficiency in querying and rendering. Our proposed *GART* similarly utilizes parametric templates to model the human body articulation. But unlike the above neural representations, the appearance is represented by 3D Gaussians [27, 91, 96] that are efficient to render. Additionally, the explicitness of 3D Gaussians allows us to design simple deformation and regularization rules.

**3D Animal Reconstruction**. Similar to the modeling of humans, parametric models have been proposed for different animals [3, 4, 36, 59, 95] and can be fitted to images and videos [5, 93]. Novel instances or more species can be captured with limited fidelity by deforming the template with image guidance [70, 94]. As template models are expensive to create for diverse animals, model-free approaches learn animal shapes by deforming a sphere [6, 14, 67, 73]. Recent approaches aim to build articulated models directly from videos as animatable neural fields [74, 79, 80]. High-quality neural capture of animals has been demonstrated with a multi-view RGB and Vicon camera array system [44]. However, unlike all these methods, *GART* robustly builds detailed 3D Gaussians upon D-SMAL [59] templates and can capture diverse dog species from in-the-wild monocular videos.

**3D Gaussian Splatting**. The key technique of the above-mentioned reconstruction now lies in differentiable rendering, where meshes [34, 41], points/surfels [35, 78, 84, 88] and NeRFs [37, 47, 48] have been widely used. Recent progress in differentiable rendering revives the classical EWA volume splatting [96], in which 3D Gaussians are used to approximate the underlying radiance field and achieve high efficiency and fidelity [27] via splatting-based rendering. 3D-GS [27, 28] techniques have been recently applied to modeling general dynamic scenes [43, 72, 81, 82] where the scenes do not have specific structures (i.e. articulation), and have been applied to 3D generation [10, 66, 83].

## 3. Method

### 3.1. Template Prior

Different from capturing deforming subjects from multi-view systems [19, 20] in a studio, the capture from monocular videos is extremely challenging, and many studies leverage category-level template models as a strong prior to associating and accumulating information across time for monocular reconstruction of humans and animals. These templates include the SMPL [42, 51] family for humans and

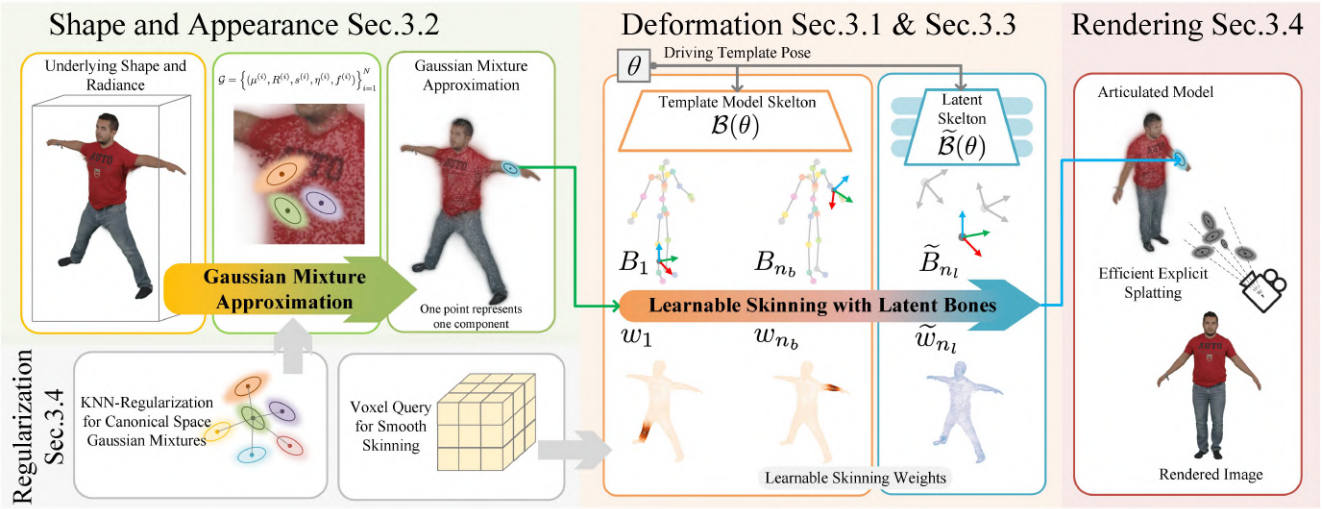


Figure 2. **Overview**: Left-top: *GART* represents the shape and appearance of articulated subjects in canonical space with Gaussian Mixtures (Sec. 3.2). Middle: Such explicit approximation can be efficiently deformed with learnable skinning and a novel latent bone approach, capturing challenging deformations (Sec. 3.3). Right: The articulated model can be efficiently rendered via Gaussian Splatting [27, 96] (Sec. 3.4). Left-bottom: Several smoothness regularizations are injected into *GART* to constrain the point-based representation (Sec. 3.4).

SMAL [59, 95] for animals. Typically, a template model consists of three components:

$$(\mathcal{M}, \mathcal{B}, \mathcal{W}). \tag{1}$$

The template mesh $\mathcal{M} = (\mathcal{V}_c, \mathcal{F})$ is defined in the canonical space to model the shape of the subjects. Predefined motion structure (skeleton) $\mathcal{B}$ of the category with $n_b$ joints can return a set of rigid bone transformations based on the driven pose $\theta$:

$$[B_1, B_2, \ldots, B_{n_b}] = \mathcal{B}(\theta), \tag{2}$$

where $B_i \in SE(3)$ represents the rigid transformation that moves the canonical joint coordinate frame to the articulated one. The surface point $x_c$ in the canonical space can be deformed to the articulated space via the linear blend skinning (LBS):

$$x = \left( \sum_{k=1}^{n_b} \mathcal{W}_k(x_c) B_k \right) x_c, \tag{3}$$

where $\mathcal{W}_k(x_c) \in \mathbb{R}$ is querying the predefined skinning weight in canonical space. Usually, $\mathcal{W}$ can be predefined in the full $\mathbb{R}^3$ space by diffusing the mesh skinning weights.

### 3.2. Shape Appearance Representation with GMM

Gaussian Articulated Template (*GART*) is a representation for deformable articulated subjects that combines the advantages of implicit and explicit representations. Inspired by recent progress in static scene rendering [27] and classical point-based graphics [96], we propose to *use 3D Gaussian Mixture Models (GMM) to explicitly approximate the implicit underlying radiance field in the canonical space*. The $i$th component in the GMM is parameterized by a 3D mean $\mu^{(i)} \in \mathbb{R}^3$, a 3D rotation $R^{(i)} \in SO(3)$, anisotropic 3-dimensional scaling factors $s^{(i)} \in \mathbb{R}^{3+}$, an opacity factor $\eta^{(i)} \in (0, 1]$, and the color radiance function encoded by Spherical harmonics $f^{(i)} \in \mathbb{R}^C$. Given a query position $x_c$ in canonical space, the density contribution of $i$th component is:

$$\sigma^{(i)}(x_c) = \eta \exp\left( -\frac{1}{2}(x_c - \mu)^T \Sigma^{-1} (x_c - \mu) \right), \tag{4}$$

where the covariance is $\Sigma = R\text{diag}(s^2)R^T$, and we omit the index $i$. The color contribution of this component is:

$$c^{(i)}(x_c, d) = \sigma(x_c)\text{sph}(R^T d, f), \tag{5}$$

where $\text{sph}(R^T d, f)$ means evaluating the spherical harmonics with coefficient $f$ at the local frame direction $R^T d$ given the global query viewing direction $d$. The total radiance field is the summation of all $N$ components:

$$\sigma(x_c) = \sum_{i=1}^{N} \sigma^{(i)}(x_c), \quad c(x_c, d) = \sum_{i=1}^{N} c^{(i)}(x_c, d). \tag{6}$$

Eq. 6 explicitly represents the canonical geometry and appearance by a list of Gaussian component parameters, which can be written as:

$$\mathcal{G} = \left\{ (\mu^{(i)}, R^{(i)}, s^{(i)}, \eta^{(i)}, f^{(i)}) \right\}_{i=1}^{N}. \tag{7}$$

In *GART*, $\mathcal{G}$ replaces the $\mathcal{M}$ in Eq. 1 triplet. Note that during optimization (Sec. 3.4), each component can move in-

dependently and is not constrained by a fixed topology like a mesh, which makes *GART* highly flexible.

### 3.3. Motion Representation with Forward Skinning

**Learnable Forward Skinning** One key advantage of using $\mathcal{G}$ is the simple and explicit deformation modeling. Since the predefined category-level skinning prior $\mathcal{W}$ from the template may not reflect the actual instance deformation, we assign each Gaussian component a learnable skinning correction:

$$\widehat{\mathcal{W}}(\mu^{(i)}) = \mathcal{W}(\mu^{(i)}) + \Delta w^{(i)}, \tag{8}$$

where $\Delta w^{(i)} \in \mathbb{R}^{n_b}$ is the learnable skinning of the $i$th Gaussian. In *GART*, learnable $\widehat{\mathcal{W}}$ replaces the $\mathcal{W}$ in Eq. 1. Given a pose $\theta$, the articulation transformation $A^{(i)}$ for the $i$th Gaussian is:

$$A^{(i)} = \sum_{k=1}^{n_b} \widehat{\mathcal{W}_k}(\mu^{(i)}) B_k, \tag{9}$$

and the Gaussian center and rotation are articulated to:

$$\mu^{(i)}_{\text{art}} = A^{(i)}_{\text{rot}} \mu^{(i)} + A^{(i)}_t, \quad R^{(i)}_{\text{art}} = A_{\text{rot}} R^{(i)}, \tag{10}$$

where $A^{(i)}_{\text{rot}}$, $A^{(i)}_t$ are the left top $3 \times 3$ and the right $3 \times 1$ block of $A^{(i)}$. Note that $A^{(i)}$ may not be an $SE(3)$ transformation anymore. But the transformed $R^{(i)}_a$ can still be used to compose a covariance as in Eq. 4, and the articulated radiance field can be directly obtained from Eq. 6 by using the articulated mean and covariance as in Eq. 10. This forward skinning enables *GART* to model motion efficiently and to avoid backward skinning root-finding [8, 9], which is used in other implicit representations.

**Latent Bones and Flexible Deformation** Person-agnostic human models such as SMPL have a predefined skeleton $\mathcal{B}$ that models the human body motion well but cannot capture the movement of loose clothing. Our goal is to find a simple approximation for the clothing motion that can be captured from a monocular video. Our insight is that the deformation of an articulated subject can be seen as driven by the $n_b$ predefined bones plus $n_l$ unknown latent bones. We can represent the latent bone transformations as a function of the pose $\theta$:

$$[\widetilde{B}_1, \ldots, \widetilde{B}_{n_l}] = \widetilde{\mathcal{B}}(\theta) \tag{11}$$

where $\widetilde{B}_i \in SE(3)$ and $\widetilde{\mathcal{B}}(\theta)$ can be parameterized with an MLP or a per-frame optimizable table. Similarly, we can learn the latent bone skinning weight for each Gaussian $\widetilde{\mathcal{W}}(\mu) \in \mathbb{R}^{n_l}$ during training. With the addition of latent bones, the forward skinning from Eq. 9 became

$$A^{(i)} = \sum_{k=1}^{n_b} \widehat{\mathcal{W}_k}(\mu^{(i)}) B_k + \sum_{q=1}^{n_l} \widetilde{\mathcal{W}_q}(\mu^{(i)}) \widetilde{B}_q. \tag{12}$$

Note this deformation model is computationally efficient and compact since the transformations $\mathcal{B}$, $\widetilde{\mathcal{B}}$ are globally shared across all Gaussians.

**Summary** Now, we fully introduced *GART*:

$$(\mathcal{G}, \mathcal{B}, \widehat{\mathcal{W}}, \widetilde{\mathcal{B}}, \widetilde{\mathcal{W}}), \tag{13}$$

which explicitly approximates the canonical shape and appearance with learnable GMM $\mathcal{G}$, and compactly represents the forward deformation with prior skeleton and learnable latent bones $\mathcal{B}, \widetilde{\mathcal{B}}$, and their learnable skinning weights $\widehat{\mathcal{W}}, \widetilde{\mathcal{W}}$. Given a pose $\theta$, using Eq. 10,12, the articulated radiance field approximation is:

$$\begin{aligned} \mathcal{G}_{\text{art}}(\theta) &= \left\{ (\mu^{(i)}_{\text{art}}, R^{(i)}_{\text{art}}, s^{(i)}, \eta^{(i)}, f^{(i)}) \right\}_{i=1}^{N} \\ \mu^{(i)}_{\text{art}} &= A^{(i)}_{\text{rot}} \mu^{(i)} + A^{(i)}_t, \quad R^{(i)}_{\text{art}} = A_{\text{rot}} R^{(i)} \\ A^{(i)} &= \sum_{k=1}^{n_b} \widehat{\mathcal{W}_k}(\mu^{(i)}) B_k(\theta) + \sum_{q=1}^{n_l} \widetilde{\mathcal{W}_q}(\mu^{(i)}) \widetilde{B}_q(\theta) \end{aligned} \tag{14}$$

### 3.4. Reconstruct *GART* from Monocular Videos

**Differentiable Rendering with Splatting** Given a perspective projection $\pi(x; E, K)$ where $E$ is the camera extrinsics and $K$ the intrinsics matrix, the projection of a 3D Gaussian can be approximately treated as a 2D Gaussian with mean and covariance:

$$\mu_{2D} = \pi(\mu; E, K); \ \Sigma_{2D} = JE\Sigma E^T J^T, \tag{15}$$

where $J$ is the Jacobian of the perspective projection, see equations (26-31) in [96]. With the preservation of the Gaussians through projection, we can efficiently splat and approximate the volume rendering with sorted color accumulation [96]:

$$\begin{aligned} I(u, d) &= \sum_{i=1}^{N} \alpha^{(i)} \text{sph}(R^T d, f^{(i)}) \prod_{j=1}^{i-1} (1 - \alpha^{(j)}) \\ \alpha^{(i)} &= G_{2D}(u | \eta^{(i)}, \mu^{(i)}_{2D}, \Sigma^{(i)}_{2D}), \end{aligned} \tag{16}$$

where the index is sorted along the depth direction, the querying pixel coordinate is $u$, the viewing direction in the world frame is $d$, and $G_{2D}$ is evaluating the 2D Gaussian density similar to Eq. 4. Eq. 16 is differentiable [27] and can provide supervision from 2D observations to update all Gaussian parameters, and we refer the readers to [27, 96] for more details.

**Optimization** Given a set of $M$ images from a monocular video and the estimated poses of the template $\{(I^*_1, \theta_1), \ldots, (I^*_M, \theta_M)\}$, we optimize $\mathcal{G}, \widehat{\mathcal{W}}, \widetilde{\mathcal{B}}, \widetilde{\mathcal{W}}$ as well as refine $\theta$ by comparing the rendered image of $\mathcal{G}_{\text{art}}(\theta)$ with the ground truth images. We initialize $\mathcal{G}$ on the template

mesh and follow the densify-and-prune strategy from 3D-GS [27] during optimization. Denote the rendered image of *GART* as $\hat{I}(\mathcal{G}_{\text{art}}(\theta))$, the training loss is:

$$L = L_1(\hat{I}, I^*) + \lambda_{\text{SSIM}} L_{\text{SSIM}}(\hat{I}, I^*) + L_{\text{reg}}, \tag{17}$$

where $\lambda$ is the loss weight, and $L_{\text{reg}}$ is introduced below.

**Regularization** The flexible nature of the 3D Gaussians in *GART* can be under-constrained when the 2D observation is sparse. 3D Gaussian mixture does not have the smoothness induced by MLPs as in NeRFs, which often leads to artifacts in unobserved spaces. Inspired by [8, 9], the learnable skinning weights $\widehat{\mathcal{W}}, \widetilde{\mathcal{W}}$ should be spatially smooth, so we distill them to a coarse voxel grid, and the per-Gaussian skinning $\Delta w^{(i)}, \widetilde{\mathcal{W}}(\mu^{(i)})$ in Eq. 8,12 are tri-linearly interpolated at $\mu^{(i)}$ from the voxel grid. We further regularize the variation of the Gaussian attributes in the KNN neighborhood of $\mu$, which leads to:

$$L_{STD}^{(i)} = \sum_{attr \in \{R,s,\eta,f,\widehat{\mathcal{W}},\widetilde{\mathcal{W}}\}} \lambda_{attr} \text{STD}_{i \in KNN(\mu^{(i)})}(attr^{(i)}), \tag{18}$$

where $STD$ is the standard deviation. Additionally, we encourage the fitting to make small changes from the original motion structure to further exploit the template model prior knowledge and to encourage small Gaussians since the non-rigid subject is approximated by piece-wise rigid moving Gaussians, which leads to:

$$L_{norm}^{(i)} = \lambda_{\widehat{\mathcal{W}}} \|\Delta w^{(i)}\|_2 + \lambda_{\widetilde{\mathcal{W}}} \|\widetilde{\mathcal{W}}(\mu^{(i)})\|_2 + \lambda_s \|s^{(i)}\|_\infty. \tag{19}$$

The total regularization loss is:

$$L_{reg} = \frac{1}{N} \sum_{i=1}^{N} L_{STD}^{(i)} + L_{norm}^{(i)}. \tag{20}$$

**Inference** During inference, all the attributes of *GART* are explicitly stored per Gaussian (no voxel grid query is needed). With our efficient modeling of appearance and motion, rendering an articulated subject is as fast as rendering a static scene. The inference FPS is more than 150 on People-Snapshot [1] at resolution $540 \times 540$.

# 4. Experiments

## 4.1. Comparison on Human Rendering

In this section, we verify *GART*'s effectiveness, efficiency, and expressiveness in monocular human reconstruction and view synthesis. We use SMPL [42] as the template, and the input during training is a monocular RGB video with an estimated SMPL pose at each video frame. The evaluation during testing is the novel view synthesis with the PSNR, SSIM, and LPIPS metrics.

The SoTA baselines in this task are the recent efficient NeRF-based human rendering methods Instant-Avatar [24] and Instant-NVR [13], which demonstrate better fidelity than classical mesh-based representations [1]. Instant-Avatar uses instant-NGP [48] in the canonical space and utilizes Fast-SNARF [8], a highly tailored GPU solver for fast backward skinning root finding, to model the deformation. It also proposes a special opacity caching strategy to accelerate the volume rendering. Instant-NVR models the appearance of each body part with separate NeRFs and utilizes a carefully designed Chart-based backward skinning to model the deformation. We conduct comparisons on three datasets: ZJU-MoCap [55], People-Snapshot [1], and UBC-Fashion [87]. Similar to InstantAvatar [24], we also conduct test-time refinement of the SMPL pose.

| Methods | Training time | PSNR | SSIM | LPIPS* |
|---|---|---|---|---|
| HumanNeRF [71] | ~10h | 30.66 | 0.969 | 33.38 |
| AS [54] | ~10h | 30.38 | 0.975 | 37.23 |
| NB [55] | ~10h | 29.03 | 0.964 | 42.47 |
| AN [52] | ~10h | 29.77 | 0.965 | 46.89 |
| NHP [32] | ~1h tuning | 28.25 | 0.955 | 64.77 |
| PixelNeRF [85] | ~1h tuning | 24.71 | 0.892 | 121.86 |
| Instant-NVR [13] | ~5min | 31.01 | 0.971 | 38.45 |
| GART | ~2.5min | **32.22** | **0.977** | **29.21** |
| GART | ~30s | 31.76 | 0.976 | 34.01 |

Table 1. Comparison of view synthesis on ZJU-MoCap [55].

**ZJU-MoCap [55]** We compare with Instant-NVR [13] and other human rendering methods on the ZJU-MoCap dataset [55] with the same setup as [13]. The average results are shown in Tab. 1. *GART* surpasses other methods in terms of synthesis results with less training time. Thanks to its efficient rendering [27] and forward skinning (Sec. 3.3), *GART* can achieve similar quantitative performance after less than 30 seconds of training. Qualitative results in Fig. 3 show that our results capture more details than Instant-NVR [13].

**People-Snapshot [1]** Another commonly compared human avatar dataset is People-Snapshot [1], and we compare *GART* to Instant-Avatar [24] with the same experimental setup. The results are shown in Tab. 2 and Fig. 4. Our method achieves comparable performance with shorter training time. Besides the training efficiency, *GART* has a unique advantage over Instant-Avatar during inference. At the resolution of $540 \times 540$, Instant-Avatar can be rendered at 15FPS [24], but *GART* can be rendered at more than 150FPS on a single RTX-3080-Laptop GPU.

**UBC-Fashion [87]** While the ZJU-MoCap [55] and People-Snapshot [1] datasets are widely benchmarked, the clothing in these datasets is all tight and does not differ much from the SMPL body model. We take a step forward towards modeling more challenging clothing, such as long dresses with highly dynamic motion and deformation. We use six videos from the UBC-Fashion [87] dataset that contains dynamic dresses and different skin colors. As shown in Fig. 5, each monocular video captures a model wearing

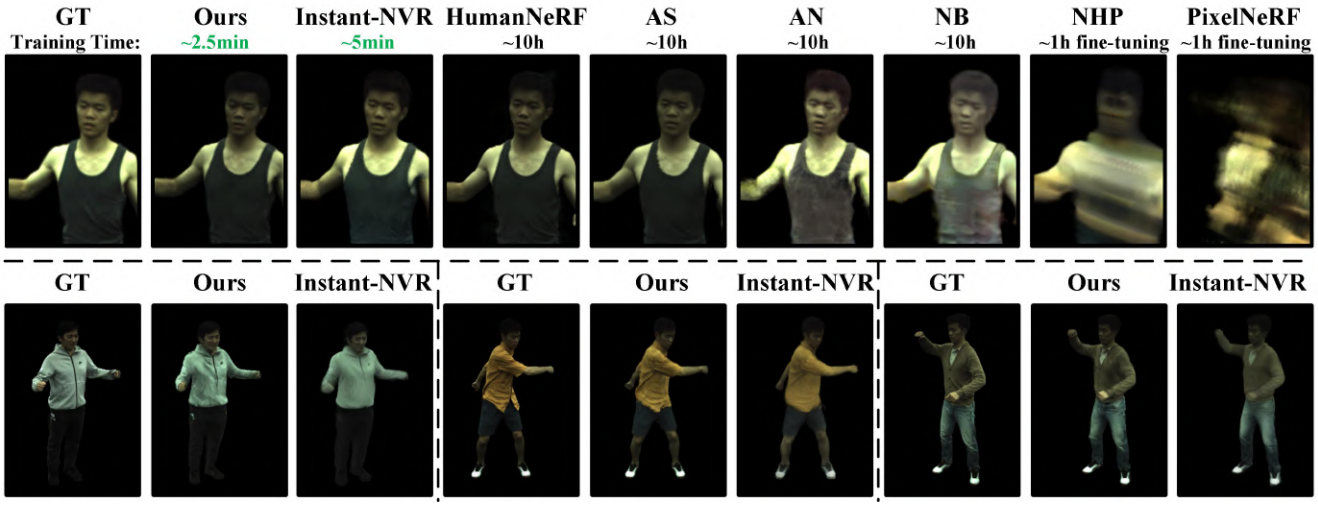


Figure 3. Comparison on ZJU-MoCap [55]. The training time is highlighted under the method names.

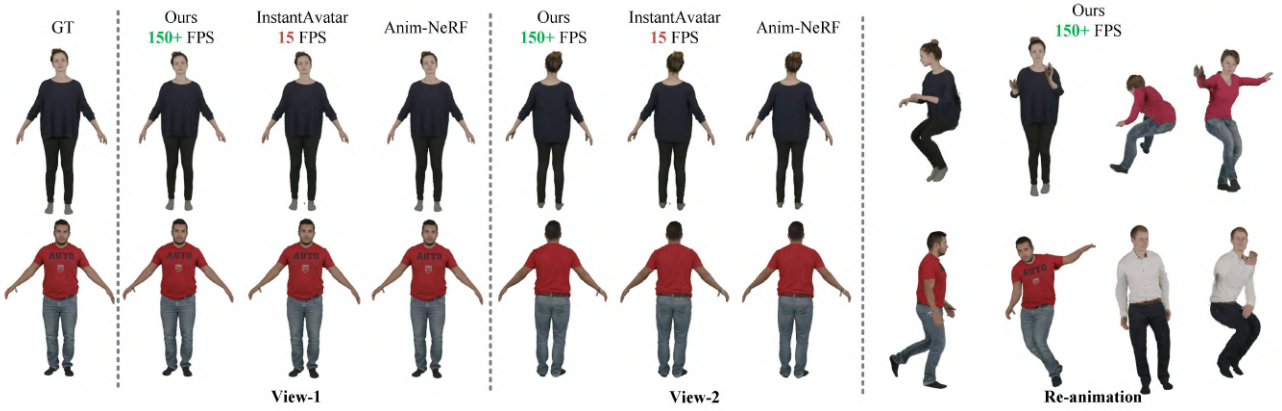


Figure 4. Comparison on People-Snapshot [1]. Note our method achieves similar quality via shorter training time and $10\times$ inference FPS.

loose clothing in front of the camera and turning around. Since these sequences have very limited variation of poses and only capture one view, we use the frames starting from 0 and pick frames with an interval of 4 for training and use the frames starting from 2 and pick frames with the same interval for testing. The SMPL poses are obtained via the SoTA human pose estimator ReFit [69]. Because both *GART* and Instant-Avatar [24] can optimize the SMPL pose during testing, and the pose estimation is noisy since the long skirts also lead to challenges for pose estimators, we found that the results of both methods are better if we use the nearest training pose and optimize it during testing.

The quantitative comparison is shown in Tab. 3, where we evaluate two variants of *GART*: *GART*-MLP uses an MLP to represent the latent bones $\widetilde{\mathcal{B}}(\theta)$ in Eq. 11, where the input to the MLP is the SMPL Pose; *GART*-T-Table directly optimize a list of rigid transformations per-time-frame that represents the latent bones $[\widetilde{B}_1, \ldots, \widetilde{B}_{n_l}]$. As shown in Fig 5, Instant-Avatar successfully captures the upper body but fails to capture the dynamic clothing. There are three potential reasons: 1) Because of the implicit modeling, Fast-SNARF [8] is utilized to solve the backward skinning, leading to multiple ambiguous correspondences in the highly dynamic skirt area. So we observe the wrong skinning that attaches skirts to the arm. 2) Using 24 SMPL bones and learnable skinning weights is insufficient to capture the complex deformation; 3) Because of the limited expressiveness of the deformation but the flexible nature of NeRF and the noisy pose estimation, many artifacts are created in the empty space due to their photometric significance at some wrong poses. On the contrary, our deformation is modeled via simple forward skinning, which can further capture flexible deformation via latent bones as in Sec. 3.3, and is optimized with 3D-GS [27] in a deformation-based process, which leads to our better performance.

| | male-3-casual | | | male-4-casual | | | female-3-casual | | | female-4-casual | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR | SSIM | LPIPS | PSNR | SSIM | LPIPS | PSNR | SSIM | LPIPS | PSNR | SSIM | LPIPS |
| Neural Body [55] (∼14h) | 24.94 | 0.9428 | 0.0326 | 24.71 | 0.9469 | 0.0423 | 23.87 | 0.9504 | 0.0346 | 24.37 | 0.9451 | 0.0382 |
| Anim-NeRF [7] (∼13h) | 29.37 | 0.9703 | **0.0168** | 28.37 | 0.9605 | **0.0268** | **28.91** | **0.9743** | **0.0215** | 28.90 | 0.9678 | **0.0174** |
| Anim-NeRF [7] (1m) | 12.39 | 0.7929 | 0.3393 | 13.10 | 0.7705 | 0.3460 | 11.71 | 0.7797 | 0.3321 | 12.31 | 0.8089 | 0.3344 |
| InstantAvatar [24] (1m) | 29.65 | 0.9730 | 0.0192 | **27.97** | 0.9649 | 0.0346 | 27.90 | 0.9722 | 0.0249 | 28.92 | 0.9692 | 0.0180 |
| GART (30s) | **30.40** | **0.9769** | 0.0377 | 27.57 | **0.9657** | 0.0607 | 26.26 | 0.9656 | 0.0498 | **29.23** | **0.9720** | 0.0378 |

Table 2. Comparison on People-Snapshot [1]. InstantAvatar [24] can inference at 15 FPS while *GART* achieves 150+ FPS.

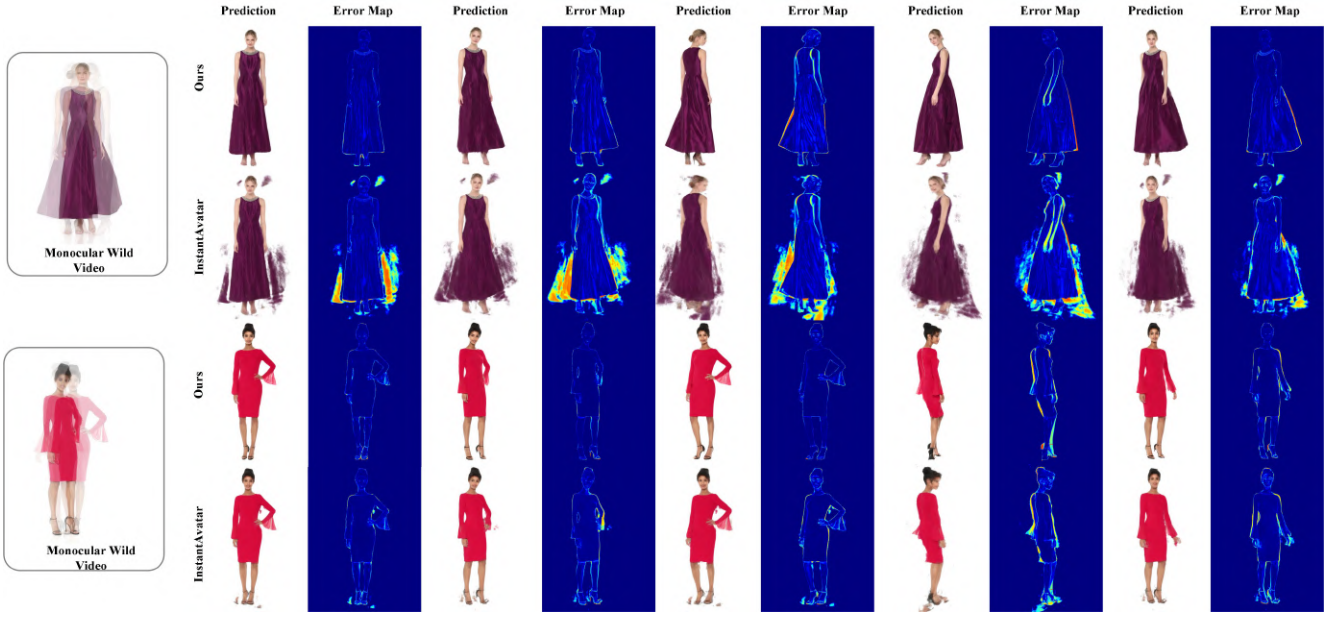


Figure 5. Comparison on UBC-Fashion [87] challenging sequences. Note how the SoTA method [24] makes artifacts around the long skirts (top) and feet (bottom).

| Methods | PSNR | SSIM | LPIPS* |
|---|---|---|---|
| InstantAvatar [24] Test Pose | 19.87 | 0.880 | 157.70 |
| InstantAvatar [24] Training Pose | 19.97 | 0.882 | 157.05 |
| GART MLP | 25.65 | 0.934 | 81.88 |
| GART T-Table | **25.96** | **0.935** | **80.57** |

Table 3. Quantitative comparison of view synthesis on UBC-Fashion [87] sequences.

| Data | Method | PSNR | SSIM | LPIPS |
|---|---|---|---|---|
| National Dog Show (6 seq) | InsAvat-Dog | 16.13 | 0.759 | 0.318 |
| | *GART* | **17.86** | **0.825** | **0.238** |
| Adobe Stock (2seq) | InsAvat-Dog | 20.62 | 0.834 | 0.227 |
| | *GART* | **24.50** | **0.921** | **0.114** |

Table 4. Quantitative evaluation of view synthesis on ITW dogs.

### 4.2. Application on Dog Rendering

In this section, we demonstrate *GART* as a general framework to capture and render animals from monocular in-the-wild videos. Specifically, we utilize the new D-SMAL [59] model that is proposed for diverse dog breeds as the base template. We conduct experiments on a total of 8 new sequences: 6 sequences from the 2022 National Dog Show (the 6 best-in-show participants), and 2 sequences captured with a green screen obtained from Adobe Stock Videos. Compared to humans, pose estimation for dogs is more challenging because of the scarcity of training data and occlusions in the environment. Therefore, we select sections where the poses are estimated corrected by BITE [59], and there are few occlusions. As shown in Fig. 6, *GART* captures different dog breeds well. Compared to D-SMAL, *GART* better reconstructs breed-specific appearance such as tails, ears, and textured fur. We also adapt InstantAvatar [24] to the D-SMAL template, which we call InsAvat-Dog for comparison. Compared to *GART*, InsAvat-Dog may create ghost artifacts under such a challenging setting, potentially due to the inaccurate and highly dynamic dog poses during training. We include a small set of test frames for each sequence and report the metrics in Tab. 4 as a baseline for neural animal reconstruction in the wild.

### 4.3. Ablation Study

To verify the effectiveness of our deformation modeling, we compare the full model with 1) removing the la-

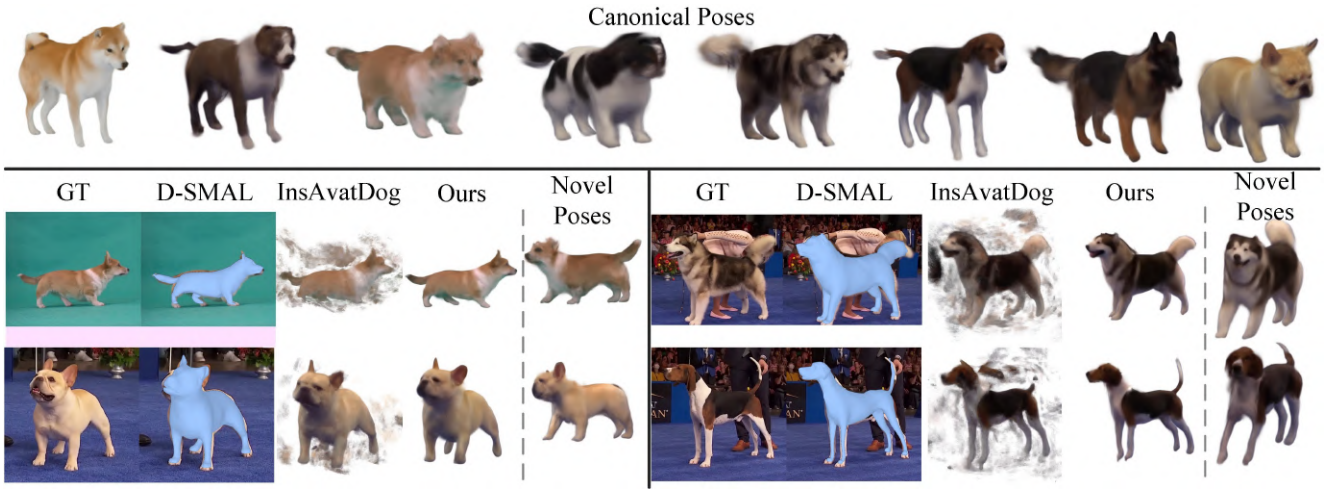


Figure 6. Qualitative results for in-the-wild dogs in canonical pose (Top) and in novel poses and views (Bottom).

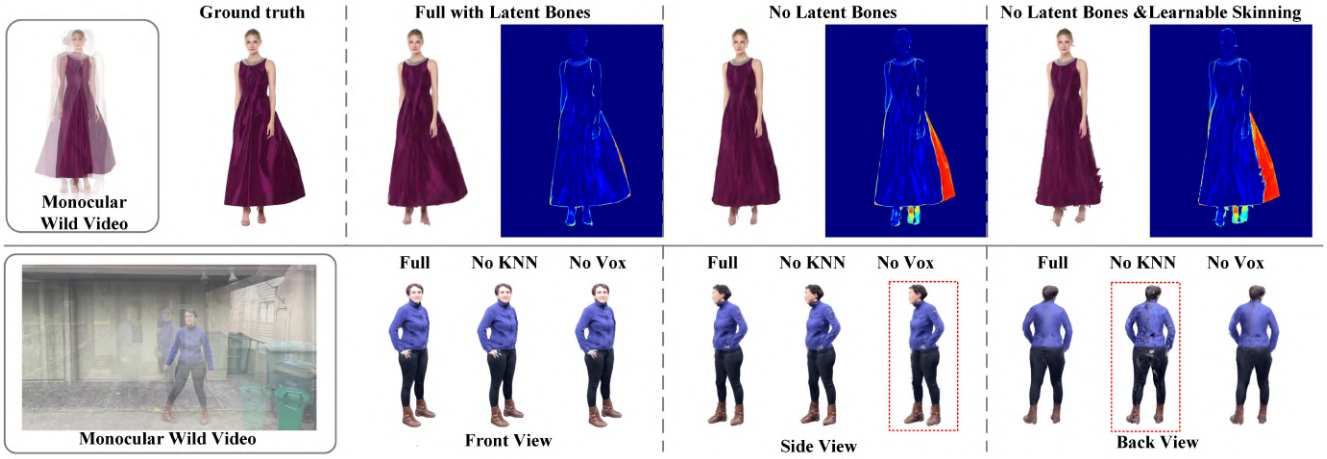


Figure 7. **Ablation**: (Top) Learnable skinning and latent bones help *GART* to capture highly dynamic skirts. (Bottom) KNN regularization in Eq. 18 helps smooth the results in the back view, and voxel-distilled skinning helps prevent noisy artifacts on the side view.

| Methods | PSNR | SSIM | LPIPS* |
|---|---|---|---|
| No Learnable Skinning | 23.76 | 0.925 | 88.76 |
| No Latent Bones | 25.00 | 0.932 | 82.03 |
| Full | 25.65 | 0.934 | 81.88 |

Table 5. Ablation of learned deformation on UBC-Fashion [87].

tent bones and 2) removing the learnable skinning on the UBC-Fashion sequences. The results are shown in Tab. 5 and Fig. 7. We observe that the full model works the best and note that the model without latent bones can still reconstruct the dress with fewer artifacts than Instant-Avatar [24], showing our robustness and effectiveness under noisy poses and large deformations. Visually, we observe from Fig. 7 that when adding the latent bones, the independent motion of the skirts can be captured better than the ablated ones. We also verify the effectiveness of our injected smoothness by 1) removing the voxel distilled skinning weight but storing the skinning weight for each Gaussian as a list, and 2) removing the KNN regularization as in Eq. 18. The qualitative comparisons of in-the-wild video are shown in Fig. 7. We note that the No-KNN version results in strong artifacts on the back, while the No-Vox version produces noisy artifacts around the body in the side view.

### 4.4. Application: Text-to-GART

*GART* is a general representation of articulated subjects and is not restricted to real monocular video reconstruction. In this section, we further demonstrate an application – Text-to-GART, by simply changing the rendering $L1$ loss and SSIM loss in Eq. 17 to an SDS loss [56]. The input is a text describing the content the user aims to generate, and the output is an optimized *GART* representing this subject. The optimization loss becomes $L = L_{\text{SDS}} + L_{\text{reg}}$, where $L_{\text{SDS}}$ is computed via forwarding a fine-tuned Stable-Diffusion [57] model MVDream [61]. For more details on $L_{\text{SDS}}$, please see Stable-Diffusion [57] and DreamGaussian [66]. Since


Figure 8. **Additional application**: Text-to-GART

there are no real poses estimated from video frames, we randomly sample some reasonable SMPL [42] template poses from AMASS [45] to augment *GART* during distillation. The generation results are shown in Fig. 8. Thanks to the efficiency of *GART*, the computation bottleneck of this application is mainly in the 2D diffusion forwarding, and the typical generation time is around 10 minutes per subject on a single GPU.

## 5. Conclusions

This paper proposes a simple and general representation, *GART*, for non-rigid articulated subjects via Gaussian Mixtures and novel skinning deformation. *GART* achieves SoTA performance on monocular human and animal reconstruction and rendering while maintaining high training and inference efficiency.

**Limitations and future work** Our proposed method has two main limitations, which could be explored in the future: 1) Our method relies on a template pose estimator, which may not exist for more general animal species. 2) *GART* can fit a single monocular video efficiently, and it's an interesting next step to explore how to capture the category-level prior knowledge of articulated subjects from in-the-wild video collections.

**Acknowledgements** The authors appreciate the support of the following grants: NSF NCS-FO 2124355, NSF FRR 2220868.

## References


[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *2018 International Conference on 3D Vision (3DV)*, pages 98–109. IEEE, 2018. 1, 2, 5, 6, 7

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018. 2

[3] Marc Badger, Yufu Wang, Adarsh Modh, Ammon Perkes, Nikos Kolotouros, Bernd G Pfrommer, Marc F Schmidt, and Kostas Daniilidis. 3d bird reconstruction: a dataset, model, and shape recovery from a single view. In *European Conference on Computer Vision*, pages 1–17. Springer, 2020. 2

[4] Benjamin Biggs, Oliver Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. Who left the dogs out? 3d animal reconstruction with expectation maximization in the loop. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 195–211. Springer, 2020. 2

[5] Benjamin Biggs, Thomas Roddick, Andrew Fitzgibbon, and Roberto Cipolla. Creatures great and smal: Recovering the shape and motion of animals from video. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14*, pages 3–19. Springer, 2019. 2

[6] Thomas J Cashman and Andrew W Fitzgibbon. What shape are dolphins? building 3d morphable models from 2d images. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):232–244, 2012. 2

[7] Jianchuan Chen, Ying Zhang, Di Kang, Xuefei Zhe, Linchao Bao, Xu Jia, and Huchuan Lu. Animatable neural radiance fields from monocular rgb videos. *arXiv preprint arXiv:2106.13629*, 2021. 2, 7

[8] Xu Chen, Tianjian Jiang, Jie Song, Max Rietmann, Andreas Geiger, Michael J Black, and Otmar Hilliges. Fast-snarf: A fast deformer for articulated neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 2, 4, 5, 6

[9] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11594–11604, 2021. 1, 2, 4, 5

[10] Zilong Chen, Feng Wang, and Huaping Liu. Text-to-3d using gaussian splatting. *arXiv preprint arXiv:2309.16585*, 2023. 2

[11] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021. 2

[12] Qingzhe Gao, Yiming Wang, Libin Liu, Lingjie Liu, Christian Theobalt, and Baoquan Chen. Neural novel actor: Learning a generalized animatable neural representation for human actors. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[13] Chen Geng, Sida Peng, Zhen Xu, Hujun Bao, and Xiaowei Zhou. Learning neural volumetric representations of dynamic humans in minutes. In *CVPR*, 2023. 2, 5

[14] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoint without keypoints. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 88–104. Springer, 2020. 2

[15] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. *arXiv preprint arXiv:2305.20091*, 2023. 2

[16] 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*, pages 12858–12868, 2023. 2

[17] Marc Habermann, Lingjie Liu, Weipeng Xu, Gerard Pons-Moll, Michael Zollhoefer, and Christian Theobalt. Hdhumans: A hybrid approach for high-fidelity digital humans. *Proc. ACM Comput. Graph. Interact. Tech.*, 6(3), aug 2023. 2

[18] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. *ACM Transactions on Graphics (ToG)*, 40(4):1–16, 2021. 1, 2

[19] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Transactions On Graphics (TOG)*, 38(2):1–17, 2019. 2

[20] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5052–5063, 2020. 2

[21] 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*, pages 11046–11056, 2021. 2

[22] Zhang Jiakai, Liu Xinhang, Ye Xinyi, Zhao Fuqiang, Zhang Yanshun, Wu Minye, Zhang Yingliang, Xu Lan, and Yu Jingyi. Editable free-viewpoint video using a layered neural representation. In *ACM SIGGRAPH*, 2021.

[23] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5605–5615, 2022.

[24] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16922–16932, 2023. 5, 6, 7, 8

[25] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *European Conference on Computer Vision*, pages 402–418. Springer, 2022. 2

[26] Angjoo Kanazawa, Michael J Black, David W Jacobs, and

Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 2

[27] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4):1–14, 2023. 2, 3, 4, 5, 6

[28] Leonid Keselman and Martial Hebert. Flexible techniques for differentiable rendering with 3d gaussians. *arXiv preprint arXiv:2308.14737*, 2023. 2

[29] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020. 2

[30] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2252–2261, 2019. 2

[31] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems*, 34:24741–24752, 2021. 2

[32] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems*, 34:24741–24752, 2021. 5

[33] Youngjoong Kwon, Lingjie Liu, Henry Fuchs, Marc Habermann, and Christian Theobalt. Deliffas: Deformable light fields for fast avatar synthesis. *Advances in Neural Information Processing Systems*, 2023. 2

[34] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 2

[35] Christoph Lassner and Michael Zollhofer. Pulsar: Efficient sphere-based neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1440–1449, 2021. 2

[36] Ci Li, Nima Ghorbani, Sofia Broomé, Maheen Rashid, Michael J Black, Elin Hernlund, Hedvig Kjellström, and Silvia Zuffi. hsmal: Detailed horse shape and pose reconstruction for motion pattern recognition. *arXiv preprint arXiv:2106.10102*, 2021. 2

[37] Ruilong Li, Matthew Tancik, and Angjoo Kanazawa. Nerfacc: A general nerf acceleration toolbox. *arXiv preprint arXiv:2210.04847*, 2022. 2

[38] 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. 2

[39] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 2

[40] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM transactions on graphics (TOG)*, 40(6):1–16, 2021. 2

[41] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7708–7717, 2019. 2

[42] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 1, 2, 5, 9

[43] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. *arXiv preprint arXiv:2308.09713*, 2023. 2

[44] Haimin Luo, Teng Xu, Yuheng Jiang, Chenglin Zhou, Qiwei Qiu, Yingliang Zhang, Wei Yang, Lan Xu, and Jingyi Yu. Artemis: articulated neural pets with appearance and motion synthesis. *arXiv preprint arXiv:2202.05628*, 2022. 2

[45] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, Oct. 2019. 9

[46] 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*, pages 4460–4470, 2019. 1

[47] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2

[48] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 2, 5

[49] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *International Conference on Computer Vision*, 2021. 2

[50] 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*, pages 165–174, 2019. 1

[51] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 2

[52] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021. 2, 5

[53] Sida Peng, Chen Geng, Yuanqing Zhang, Yinghao Xu, Qian-

qian Wang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Implicit neural representations with structured latent codes for human body modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[54] Sida Peng, Shangzhan Zhang, Zhen Xu, Chen Geng, Boyi Jiang, Hujun Bao, and Xiaowei Zhou. Animatable neural implict surfaces for creating avatars from videos. *arXiv preprint arXiv:2203.08133*, 4(5), 2022. 5

[55] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 2, 5, 6, 7

[56] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 8

[57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 8

[58] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), Nov. 2017. 2

[59] Nadine Rüegg, Shashank Tripathi, Konrad Schindler, Michael J Black, and Silvia Zuffi. Bite: Beyond priors for improved three-d dog pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8867–8876, 2023. 2, 3, 7

[60] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 2

[61] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv:2308.16512*, 2023. 8

[62] Shih-Yang Su, Timur Bagautdinov, and Helge Rhodin. Danbo: Disentangled articulated neural body representations via graph neural networks. In *European Conference on Computer Vision*, 2022. 2

[63] Shih-Yang Su, Frank Yu, Michael Zollhoefer, and Helge Rhodin. A-nerf: Surface-free human 3d pose refinement via neural rendering. *arXiv preprint arXiv:2102.06199*, 2021. 2

[64] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11179–11188, 2021. 2

[65] Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J Black. Trace: 5d temporal regression of avatars with dynamic cameras in 3d environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8856–8866, 2023. 2

[66] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 2, 8

[67] Shubham Tulsiani, Nilesh Kulkarni, and Abhinav Gupta. Implicit mesh reconstruction from unannotated image collections. *arXiv preprint arXiv:2007.08504*, 2020. 2

[68] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdfs. In *European Conference on Computer Vision*, 2022. 2

[69] Yufu Wang and Kostas Daniilidis. Refit: Recurrent fitting network for 3d human recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14644–14654, 2023. 2, 6

[70] Yufu Wang, Nikos Kolotouros, Kostas Daniilidis, and Marc Badger. Birds of a feather: Capturing avian shape models from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14739–14749, 2021. 2

[71] 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*, pages 16210–16220, 2022. 2, 5

[72] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. *arXiv preprint arXiv:2310.08528*, 2023. 2

[73] Shangzhe Wu, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Dove: Learning deformable 3d objects by watching videos. *International Journal of Computer Vision*, pages 1–12, 2023. 2

[74] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Magicpony: Learning articulated 3d animals in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8792–8802, 2023. 2

[75] 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. In *Computer Graphics Forum*, volume 41, pages 641–676. Wiley Online Library, 2022. 1

[76] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022. 2

[77] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion, 2021. 2

[78] Zhen Xu, Sida Peng, Haotong Lin, Guangzhao He, Jiaming Sun, Yujun Shen, Hujun Bao, and Xiaowei Zhou. 4k4d: Real-time 4d view synthesis at 4k resolution. *arXiv preprint arXiv:2310.11448*, 2023. 2

[79] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2863–2873, 2022. 2

[80] Gengshan Yang, Chaoyang Wang, N Dinesh Reddy, and Deva Ramanan. Reconstructing animatable categories from videos. In *Proceedings of the IEEE/CVF Conference on*

*Computer Vision and Pattern Recognition*, pages 16995–17005, 2023. 2

[81] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. *arXiv preprint arXiv:2309.13101*, 2023. 2

[82] Zeyu Yang, Hongye Yang, Zijie Pan, Xiatian Zhu, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. *arXiv preprint arXiv:2310.10642*, 2023. 2

[83] Taoran Yi, Jiemin Fang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussian splatting with point cloud priors. *arXiv preprint arXiv:2310.08529*, 2023. 2

[84] Wang Yifan, Felice Serena, Shihao Wu, Cengiz Öztireli, and Olga Sorkine-Hornung. Differentiable surface splatting for point-based geometry processing. *ACM Transactions on Graphics (TOG)*, 38(6):1–14, 2019. 2

[85] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 5

[86] Zhengming Yu, Wei Cheng, Xian Liu, Wayne Wu, and Kwan-Yee Lin. Monohuman: Animatable human neural field from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16943–16953, 2023. 2

[87] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. Dwnet: Dense warp-based network for pose-guided human video generation. *arXiv preprint arXiv:1910.09139*, 2019. 5, 7, 8

[88] Qiang Zhang, Seung-Hwan Baek, Szymon Rusinkiewicz, and Felix Heide. Differentiable point-based radiance fields for efficient view synthesis. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–12, 2022. 2

[89] Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J Black, and Otmar Hilliges. Pointavatar: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21057–21067, 2023. 1

[90] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15893–15903, 2022. 2

[91] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3d gaussian avatars. 2023. 2

[92] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4574–4584, 2023. 2

[93] Silvia Zuffi, Angjoo Kanazawa, Tanya Berger-Wolf, and Michael J Black. Three-d safari: Learning to estimate zebra pose, shape, and texture from images" in the wild". In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5359–5368, 2019. 2

[94] Silvia Zuffi, Angjoo Kanazawa, and Michael J Black. Lions and tigers and bears: Capturing non-rigid, 3d, articulated shape from images. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3955–3963, 2018. 2

[95] Silvia Zuffi, Angjoo Kanazawa, David Jacobs, and Michael J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2, 3

[96] Matthias Zwicker, Hanspeter Pfister, Jeroen Van Baar, and Markus Gross. Ewa splatting. *IEEE Transactions on Visualization and Computer Graphics*, 8(3):223–238, 2002. 2, 3, 4