

Progressive Pose-Guided 4D Animal Reconstruction from Monocular Video

Siyuan Li¹ , Weiying Chen¹ , Yilin Wang¹ ,
Xinxin Zuo² , Xingyu Li¹ , and Li Cheng¹ ✉

¹ University of Alberta, Edmonton, AB, Canada
{sli20, lcheng5}@ualberta.ca

² Concordia University, Montreal, QC, Canada

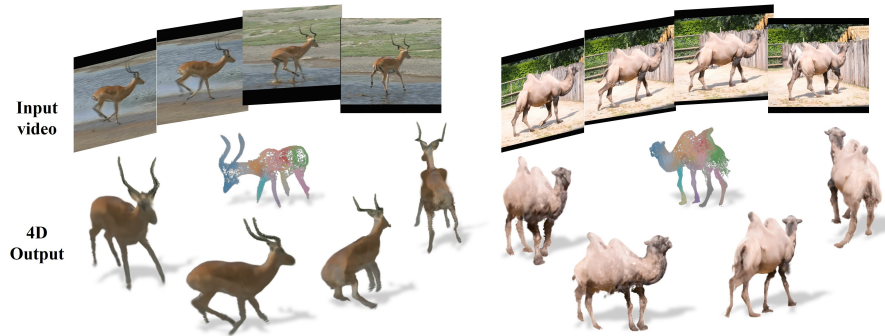


Fig. 1: Given monocular videos of animals (top), our method produces high-fidelity 4D models enabling free-viewpoint rendering across time and viewing angles (bottom). Center: canonical 3D Gaussians colored by skinning weights.

Abstract. Reconstructing 4D animals from monocular videos is challenging due to large inter-species variation, complex articulations, and the lack of reliable templates. Existing approaches typically rely on either strict category-specific priors that restrict generalization, or unconstrained generative models that sacrifice input fidelity. To bridge this gap, we present a progressive test-time optimization framework built on 3D Gaussian Splatting for high-fidelity 4D animal reconstruction from a single video. Our key insight is that a coarse shape prior suffices when coupled with a progressive strategy that disentangles articulated pose from non-rigid deformation. Specifically, we employ a symmetry-aware temporal encoding that exploits bilateral cues while absorbing camera estimation drift and a part-conditioned deformation mechanism guided by learnable part anchors and a learnable skinning field. Extensive experiments demonstrate that our approach generalizes robustly across diverse species, achieving superior geometric accuracy, temporal consistency, and visual fidelity compared to existing baselines, even under severe prior mismatch. Project page: <https://syl-322.github.io/ReWild4D/>

1 Introduction

Animals in the natural world display a stunning diversity of shapes and behaviors. Accurately reconstructing their 3D shape and motion from visual data is crucial for various applications ranging from wildlife monitoring, animal conservation and ethology research, to immersive media content creation. Despite the wide accessibility of monocular video, the task of creating realistic 4D animal models from monocular video presents a significant challenge in computer vision. This is primarily due to the inherent complexity of animal morphology and behaviors, as well as the fact that their appearance and motion are only partly observable from a monocular video.

The task of 3D animal reconstruction presents unique challenges compared to 3D human reconstruction. Human models, such as SMPL [20, 23], benefit from well-studied anatomical structures and abundant 3D motion capture datasets. In contrast, diverse animal species exhibit extreme shape and motion variations, yet very little animal motion capture benchmarks are available. The pioneer SMAL [50] is a parametric SMPL-like animal model learned from a limited collection of toy figurines; it captures a rather limited category of species and lacks realistic details of shape and motion. The dilemma of both scarcely labeled, partially observable animal data, and extraordinarily diverse shape & motion variations across animal species, forces a trade-off where, category-specific methods [2, 27, 28, 31, 33] achieve higher reconstruction quality by training on annotated dataset but struggle with generalization, while category-agnostic approaches [1, 6, 16] improve coverage at the cost of reconstruction fidelity.

Extending to 4D animal reconstruction from monocular videos reveals a fundamental tension between representation flexibility and computational efficiency. Mesh-based methods [29, 38, 39] are efficient but topologically constrained, while neural implicit methods [40, 41] offer flexibility at prohibitive computational costs. 3D Gaussian Splatting (3DGS) [10] emerges as a powerful alternative: it combines the rendering efficiency of explicit geometry with the topological flexibility of unstructured point-based models, making it uniquely suited to capture complex dynamics. However, recent 3DGS-based frameworks for animal reconstruction, such as GART [13], still rely on parametric templates. Generation-based methods [8, 25] achieve impressive synthesis but sacrifice input fidelity. Existing methods either depend on rigid templates that limit generalization or generative priors that compromise reconstruction accuracy. We argue this trade-off stems from treating shape priors as strict constraints rather than flexible initializations.

Inspired by the above observations, we present a progressive test-time optimization framework built on 3D Gaussian Splatting for high-fidelity 4D animal reconstruction from monocular in-the-wild videos, with representative results shown in Fig. 1. Our key insight is that robust 4D reconstruction does not require highly accurate shape priors; a coarse initialization suffices when allowed to continuously evolve through a principled disentanglement of articulated pose and non-rigid deformation. To achieve this, we introduce a bilateral symmetry augmentation strategy that exploits mirror cues while absorbing systematic camera drift. Furthermore, rather than isolating pose and deformation stages like prior

pipelines, our framework seamlessly bridges them using learnable part anchors. Combined with a symmetry-aware temporal encoding and part-conditioned cross-attention, these anchors provide a shared identity space that intrinsically guides local geometry updates. This progressive disentanglement enables temporally coherent reconstructions across diverse species and complex behaviors. In summary, our approach features the following key contributions:

- A progressive test-time optimization framework for 4D animal reconstruction from monocular video, requiring neither annotated training data, category-specific templates, nor multi-view generative priors. By treating the shape prior as a coarse initialization with a learnable skinning field, our method generalizes across diverse species even under severe prior mismatch.
- A two-stage pipeline that disentangles articulated motion from non-rigid deformation. For pose refinement, we introduce a symmetry-aware temporal encoding that explicitly decouples 2D visual correspondence from 3D camera estimation drift. For non-rigid deformation, we propose a part-conditioned cross-attention mechanism, where learnable part anchors serve a dual role: generating part-level temporal features for articulated pose estimation, and providing per-Gaussian identity embeddings to spatially condition local deformations.
- Extensive experiments demonstrate state-of-the-art performance in geometric accuracy, temporal consistency, and visual fidelity across diverse species on challenging in-the-wild videos.

2 Related Work

3D Animal Reconstruction. Reconstructing 3D animals is more challenging than reconstructing humans due to interspecific variation, complex articulations, and limited 3D data. Parametric models such as SMAL [50] provide the first skinned multi-animal template from toy figurines, followed by various extensions and refinement [27, 28, 48, 49]. CSM-based methods [11, 12] predict dense image-to-surface mappings but remain tied to predefined templates. Template-free methods [18, 44, 45] discover parts and skeletons from sparse images through optimizing primitive part representation. Recent learning-based approaches scale to Internet data: UMR [14], MagicPony [33], and Farm3D [6] learn category-specific models, whereas FAUNA [16] and SAOR [1] aim for category-agnostic reconstruction. The evolution in 3D animal priors has grounded a natural basis for 4D reconstruction, yet they are often treated as fixed constraints, and are severely limited in generalizing to unseen animal species. Instead, a shape prior is engaged in our approach as merely coarse initialization, which has been empirically demonstrated to contribute to flexible and faithful 4D recovery across species.

Dynamic Animal Reconstruction Extending static 3D models to capture temporal dynamics from monocular videos remains a central challenge in animal reconstruction. Deformation-based approaches [34, 38, 39] rely on deforming initial

meshes with fixed connectivity, which struggle to recover fine surface details. [40, 41] adopt NeRF-based representations for greater topological flexibility but suffer from prohibitive computational costs and lack explicit surface geometry. Hybrid explicit approaches [13, 29] have recently emerged, among which GART [13] leverages 3D Gaussian Splatting for a better efficiency-flexibility trade-off; however, these pipelines typically treat pose and deformation as isolated stages. This decoupling limits their ability to re-bind parts coherently under severe prior mismatches. In contrast, our dual-role part anchors establish a shared identity space that seamlessly couples both stages.

Gaussian Splatting for Dynamic Scenes 3D Gaussian Splatting (3DGS) [10] revolutionizes novel view synthesis by providing an explicit, flexible representation that achieves real-time performance with high fidelity. Extending it to dynamic scenes, current variants generally fall into two categories. *Time-augmented* 3DGS encodes temporality in Gaussians but scales poorly with sequence length [9, 22, 32]. *Deformation-based* variants keep a canonical 3DGS and learn warps for compact memory and smooth motion [5, 30, 43]. We adopt the deformation paradigm and drive warps with pose cues to improve temporal coherence for our task.

Video-to-4D Generation A parallel line of work leverages generative priors for video-to-4D synthesis. Zero-1-to-3 [19] pioneered this direction by leveraging diffusion models to hallucinate novel views from a single image. Building upon Zero-1-to-3, DreamMesh4D [15] generates 4D models using a hybrid mesh-3DGS representation. In contrast, PAD3R [17] reconstructs 4D content based on Zero-1-to-3 by training a personalized PoseNet. SV4D [36] and SV4D 2.0 [46] extend this concept to videos, enforcing multi-frame and multi-view consistency. However, these methods struggle with long video sequences, often losing geometric detail and requiring fixed-length inputs. Other methods that directly generate 4D models, such as Splat4D [47], L4GM [25], and GVF-Diffusion [8] usually demand large model ensembles and high memory usage. They will also produce results with significant inconsistencies in both appearance and shape compared to the input video. A fundamental limitation of these generative approaches is that they typically freeze the canonical representation once dynamic learning begins; consequently, their canonical space becomes biased toward the dataset distribution rather than the specific input subject. In contrast, we pursue reconstruction-only supervision from the input video and allow the canonical shape and skinning field to continue evolving, avoiding generative mismatch while retaining strict faithfulness to the input identity.

3 Our Approach

Our goal is to recover high-fidelity, time-varying 4D representations of animals from a monocular video sequence $\{I^t\}_{t=1}^T$. To address the inherent ambiguity of monocular observations, we employ a progressive test-time optimization framework that explicitly disentangles articulated pose from non-rigid deformation.

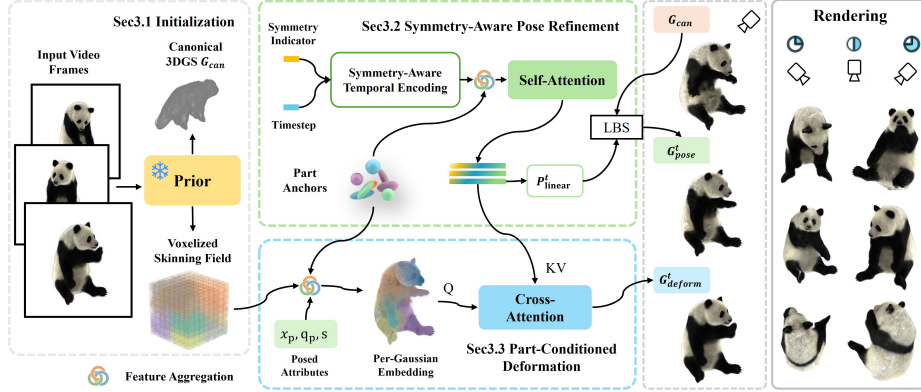


Fig. 2: Pipeline overview. From a monocular video, we initialize canonical 3D Gaussians and a learnable skinning field from the Fauna prior (Sec. 3.1). Symmetry-Aware Pose Refinement (Sec. 3.2): learnable part anchors and symmetry-aware temporal encoding are processed by self-attention to estimate per-joint transformations, yielding intermediate representations G_{pose}^t . Part-Conditioned Deformation (Sec. 3.3): part anchors are aggregated via sampled skinning weights into per-Gaussian part embeddings and, together with posed attributes, query the part-level temporal features through cross-attention to produce G_{deform}^t . Notice how the deformation module captures fine-grained dynamics (e.g., the raised front paw in G_{deform}^t) that are completely missed by the rigid LBS in G_{pose}^t . Right: rendering from arbitrary viewpoints and time steps.

Specifically, initialized from a coarse prior and organized under a bilateral symmetry augmentation strategy (Sec. 3.1), the canonical 3D Gaussian model G_{can} first undergoes pose refinement driven by learnable part anchors and symmetry-aware temporal encoding (Sec. 3.2), followed by part-conditioned deformation (Sec. 3.3). The complete pipeline is illustrated in Fig. 2.

3.1 Initialization and Bilateral Symmetry Augmentation

Initialization We bootstrap using Fauna [16], which provides a coarse estimate of animal shape and pose from a single image. Given an input frame I^t , Fauna predicts rest-pose vertices $V^t \in \mathbb{R}^{N_v \times 3}$, skinning weights $W^t \in \mathbb{R}^{N_v \times J}$, camera poses $C^t \in \mathbb{R}^{4 \times 4}$, and joint transformations $P^t \in \mathbb{R}^{J \times 4 \times 4}$. We retain the first-frame outputs (V^1, W^1) for canonical initialization, and use $\{C^t\}$ and $\{P^t\}$ as starting points for optimization.

Following [10], we represent the canonical animal as 3D Gaussians $G_{\text{can}} = \{\mathbf{x}, \mathbf{q}, \mathbf{s}, \boldsymbol{\alpha}, \mathbf{c}\}$, initialized from V^1 . We embed W^1 into a dense voxel grid $W_g \in \mathbb{R}^{J \times 32 \times 32 \times 32}$ inspired by [7, 13]. Since shape priors are often geometrically misaligned with the target video, we formulate W_g as a *learnable* skinning field optimized alongside the Gaussians, enabling continuous interpolation for dynamically spawned points and progressive correction of initial binding inaccuracies. Each Gaussian’s skinning weights $\mathbf{w}_i \in \mathbb{R}^J$ are obtained via trilinear interpolation from W_g .

Bilateral Symmetry Augmentation We assume the canonical shape is bilaterally symmetric and that $\{I^t, I_{\text{flip}}^t\}$ are mirror pairs; thus, horizontally flipping can extend viewpoint coverage. Running the prior on both yields $\{C^t, C_{\text{flip}}^t\}$ that *should* satisfy $C_{\text{flip}}^t = MC^t$ (M : reflection transformation along bilateral symmetry plane). In practice, learning-based estimators often exhibit a systematic flip-induced bias: $C_{\text{flip}}^t \neq MC^t$.

Naively fitting both as exact mirrors injects this drift as conflicting supervision, while treating them as independent samples discards their inherent geometric relationship. To leverage symmetry without introducing noise or missing geometric context, we directly apply M to C^t and C_{flip}^t , forming two internally-symmetric camera groups $\{C^t, C_{\text{sym}}^t\}$ and $\{C_{\text{flip}}^t, C_{\text{flip,sym}}^t\}$. Thus, we can construct two complementary supervision groups:

$$\mathcal{V}_{\text{orig}} = \{(I^t, C^t, P^t), (I_{\text{flip}}^t, C_{\text{sym}}^t, P_{\text{flip}}^t)\}, \quad (1)$$

$$\mathcal{V}_{\text{flip}} = \{(I_{\text{flip}}^t, C_{\text{flip}}^t, P_{\text{flip}}^t), (I^t, C_{\text{flip,sym}}^t, P^t)\}, \quad (2)$$

each containing an original-flipped frame pair under a shared camera reference. Each training iteration samples from either $\mathcal{V}_{\text{orig}}$ or $\mathcal{V}_{\text{flip}}$, enriching supervision by exposing the model to both geometric interpretations of the same scene. A symmetry-aware temporal encoding then lets the network distinguish input samples during pose refinement (Sec. 3.2).

3.2 Symmetry-Aware Pose Refinement

Initial per-frame poses P^t predicted by the prior model are often unreliable due to limited views and articulation ambiguity. Our pose refinement module estimates per-joint transformations for linear blend skinning (LBS), mapping the canonical representation G_{can} to posed states G_{pose}^t .

To incorporate bilateral symmetry and handle the systematic camera drift introduced above, we design a **symmetry-aware temporal encoding**:

$$\mathbf{e}_{m,v}^t = \text{emb}(t \cdot m) \oplus v, \quad m, v \in \{-1, 1\}, \quad (3)$$

where $\oplus$ denotes concatenation. The indicator v encodes the 2D frame state (1 for I^t and -1 for I_{flip}^t), while m identifies the 3D camera reference (1 for $\mathcal{V}_{\text{orig}}$ and -1 for $\mathcal{V}_{\text{flip}}$). This factorization decouples 2D visual symmetry from 3D camera configuration, allowing the model to exploit mirror relationships while absorbing inter-group camera drift rather than treating it as conflicting supervision. Multiplying t by m before encoding further leverages the antisymmetric property of sinusoidal embeddings to structurally relate the two camera groups.

Instead of predicting joint poses solely from a temporal embedding, we introduce randomly initialized learnable part anchors $\mathbf{A} \in \mathbb{R}^{J \times D}$ that learn joint-specific identity representations shared across all frames. We set $J = 20$ following the Fauna prior and $D = 8$ for the anchor dimension. The anchors $\mathbf{A}$ combined with $\mathbf{e}_{m,v}^t$ are processed by a self-attention block to produce joint-specific temporal features $\mathbf{F}_J^t \in \mathbb{R}^{J \times K}$. These features are projected to per-joint

transformations $P_{\text{linear}}^t \in \mathbb{R}^{J \times 7}$ (quaternion rotation and translation), and also forwarded to the subsequent deformation stage as kinematic context. Applying LBS with P_{linear}^t and interpolated skinning weights updates Gaussian centers and orientations, yielding the posed Gaussians $G_{\text{pose}}^t = \{\mathbf{x}_p, \mathbf{q}_p, \mathbf{s}, \boldsymbol{\alpha}, \mathbf{c}\}$.

To prevent appearance gradients from corrupting pose estimation, we supervise this stage with silhouette loss only, detaching non-articulation attributes during rendering (Sec. 3.4). This gradient isolation ensures that pose refinement is driven purely by geometric alignment.

3.3 Part-Conditioned Deformation

While LBS-based pose refinement captures articulated motion, real animals exhibit non-rigid effects that cannot be modeled by skeletal transformations alone. Our deformation module addresses this by predicting the final deformed geometric parameters $G_{\text{deform}}^t = \{\mathbf{x}_d, \mathbf{q}_d, \mathbf{s}_d, \boldsymbol{\alpha}, \mathbf{c}\}$, conditioned on the articulation context from pose refinement.

To provide explicit kinematic context beyond spatial position, we augment the query with a part-aware embedding $\mathbf{z}_i = \mathbf{w}_i^\top \mathbf{A} \in \mathbb{R}^D$, which encodes each Gaussian’s joint membership through the *same* learnable anchors $\mathbf{A}$ used in pose refinement (Sec. 3.2), making $\mathbf{A}$ a shared identity space that couples the two stages. This complements the positional encoding by providing a position-independent part identity signal, improving robustness when different body parts overlap in posed space.

The query for cross-attention is constructed from $\mathbf{z}_i$ and the posed geometric attributes of each Gaussian $(\mathbf{x}_{p,i}, \mathbf{q}_{p,i}, \mathbf{s}_i)$, while the joint-specific temporal features $\mathbf{F}_j^t$ from pose refinement serve as keys and values. This enables each Gaussian to selectively retrieve time-varying deformation signals from its associated joints, guided by both spatial proximity and learned part membership. The network outputs only the deformed geometry $(\mathbf{x}_d, \mathbf{q}_d, \mathbf{s}_d)$; the appearance $(\boldsymbol{\alpha}, \mathbf{c})$ is not predicted per-frame but shared from the canonical Gaussians and optimized directly, maintaining appearance consistency across time.

3.4 Progressive Optimization and Objectives

We optimize our framework through progressive test-time optimization on each monocular video. The overall objective balances geometric articulation and non-rigid shape fidelity:

$$\mathcal{L}_{\text{total}}^t = \mathcal{L}_{\text{pose}}^t + \mathcal{L}_{\text{deform}}^t + \mathcal{L}_{\text{geo}}^t, \quad (4)$$

where $\mathcal{L}_{\text{pose}}^t$ isolates articulated motion, $\mathcal{L}_{\text{deform}}^t$ captures fine details, and $\mathcal{L}_{\text{geo}}^t$ provides geometric regularization.

Stage-Aware Rendering Losses. To prevent appearance gradients from corrupting geometric articulation, the pose refinement stage relies strictly on

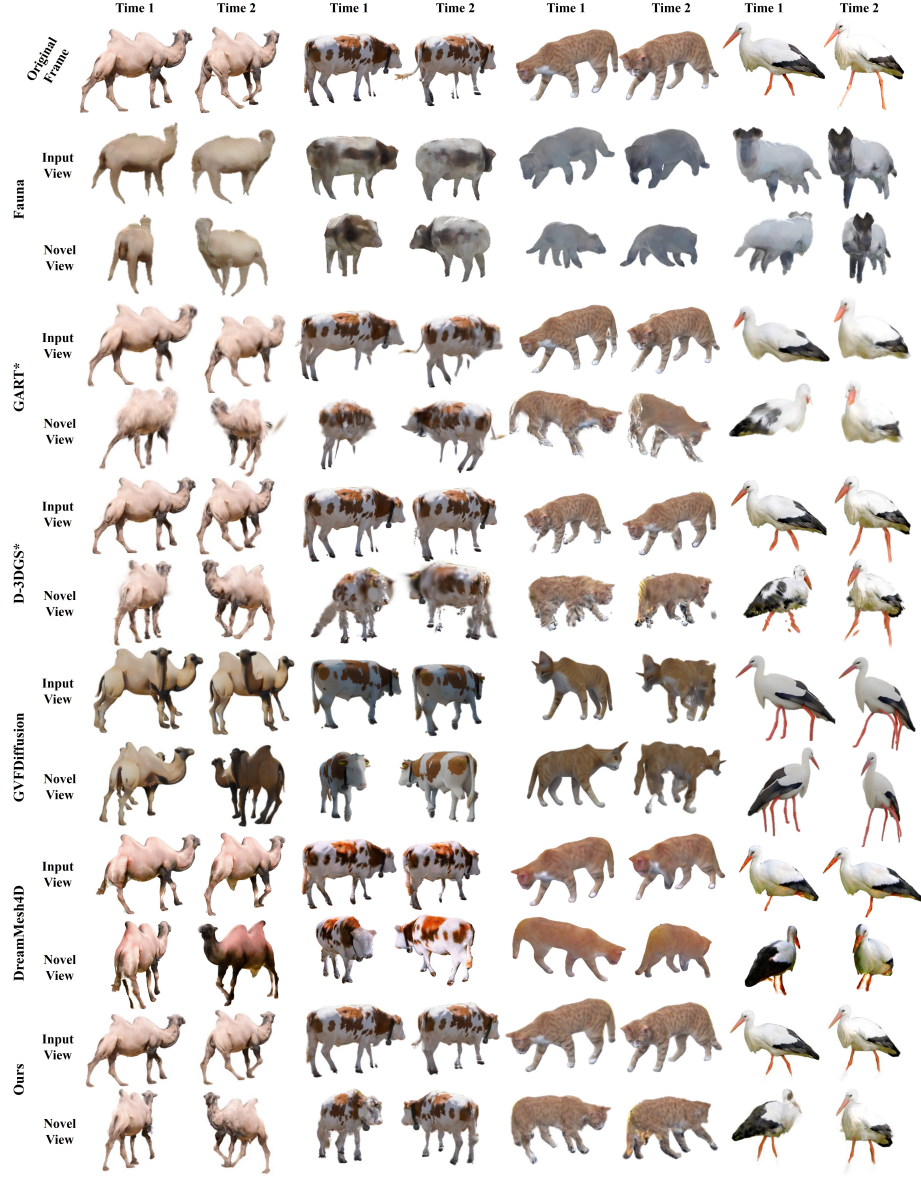


Fig. 3: Visual comparison with SOTA methods on real videos at 2 different time steps.

silhouette supervision. We block appearance-driven gradients by detaching non-articulated attributes during rasterization, yielding the posed Gaussians $G_{\text{pose}}^t = \{\mathbf{x}_p, \mathbf{q}_p, \text{sg}(\mathbf{s}, \boldsymbol{\alpha}, \mathbf{c})\}$. The pose loss is:

$$\mathcal{L}_{\text{pose}}^t = \lambda_{\text{pose}} \mathcal{L}_{\text{sil}}(\hat{S}_{\text{pose}}^t, S_{\text{SAM}}^t), \quad (5)$$

where $\mathcal{L}_{\text{sil}}$ combines standard BCE and Dice losses, $\hat{S}_{\text{pose}}^t$ is the rendered silhouette, S_{SAM}^t is the Grounded-SAM [26] mask, and $\lambda_{\text{pose}} = 0.2$.

Conversely, the deformation stage jointly optimizes all Gaussian attributes (including appearance) to capture non-rigid dynamics and realistic textures. The deformation loss combines silhouette and photometric objectives:

$$\mathcal{L}_{\text{deform}}^t = \mathcal{L}_{\text{rgb}}(\hat{I}_{\text{deform}}^t, I^t) + \lambda_{\text{sil}} \mathcal{L}_{\text{sil}}(\hat{S}_{\text{deform}}^t, S_{\text{SAM}}^t), \quad (6)$$

where $\mathcal{L}_{\text{rgb}}$ is a standard blend of $\mathcal{L}_1$ and SSIM losses ($\lambda_{\text{SSIM}} = 0.2$), $\hat{I}_{\text{deform}}^t$ is the rendered RGB, I^t is the input frame, and $\lambda_{\text{sil}} = 0.1$.

Geometric Regularization. Our geometric regularization term $\mathcal{L}_{\text{geo}}$ consists of three components: a skinning field smoothness term $\mathcal{L}_{\text{tv}}$, a part-aware compactness term $\mathcal{L}_{\text{compact}}$, and a surface normal smoothness term $\mathcal{L}_{\text{smooth}}$:

$$\mathcal{L}_{\text{geo}}^t = \lambda_{\text{tv}} \mathcal{L}_{\text{tv}} + \lambda_{\text{compact}} \mathcal{L}_{\text{compact}} + \lambda_{\text{smooth}} \mathcal{L}_{\text{smooth}}^t, \quad (7)$$

where $\lambda_{\text{tv}} = 10$, $\lambda_{\text{compact}} = 0.1$, and $\lambda_{\text{smooth}} = 0.1$. $\mathcal{L}_{\text{geo}}^t$ is only applied during the first 10k iterations. See details in the supplementary material.

First, to ensure physically plausible skinning fields, we apply a Huber total variation loss $\mathcal{L}_{\text{tv}}$ on the skinning weights in the voxel grid $W_g \in \mathbb{R}^{J \times R \times R \times R}$:

$$\mathcal{L}_{\text{tv}} = \frac{1}{3} \sum_{d \in \{x, y, z\}} \mathbb{E}[\mathcal{H}_{\delta=0.05}(\nabla_d W_g)], \quad (8)$$

where ∇_d denotes spatial gradients along axis d and $\mathcal{H}_{\delta}$ is the Huber function.

Second, to prevent floating artifacts, we introduce a part-aware compactness loss applied to both canonical and posed Gaussian positions. For each joint j , we compute the weighted centroid and covariance using the current skinning weights $\{w_{i,j}\}$, and penalize only the two smallest eigenvalues $\lambda_{j,1} \leq \lambda_{j,2}$ of the covariance, preserving elongated structures (e.g., limbs) along the principal axis while discouraging lateral spread:

$$\mathcal{L}_{\text{compact}} = \frac{1}{J} \sum_{j=1}^J \gamma_j (\lambda_{j,1} + \lambda_{j,2}), \quad (9)$$

where $\gamma_j = \exp(-\alpha \cdot \Omega_j / \max_k \Omega_k)$ down-weights large parts such as the torso, α is a scaling hyperparameter (we set $\alpha = 2.0$) and $\Omega_j = \sum_i w_{i,j}$ is the occupancy mass of joint j .

Third, to reduce surface noise in novel views, we render normal maps $\hat{N}_{\theta}^t$ from randomly sampled viewpoints θ and penalize local angular variation using absolute cosine similarity to handle normal orientation ambiguity:

$$\mathcal{L}_{\text{smooth}}^t = \sum_{d \in \{x, y\}} \frac{\sum_p \omega_{p,d}^t (1 - |\hat{N}_{\theta}^t(p) \cdot \hat{N}_{\theta}^t(p + \delta_d)|)}{\sum_p \omega_{p,d}^t + \epsilon}, \quad (10)$$

where δ_d denotes unit pixel offsets along each axis and $\omega_{p,d}^t = \hat{S}_{\theta}^t(p) \cdot \hat{S}_{\theta}^t(p + \delta_d)$ restricts computation to foreground regions.

Table 1: Input-view quality on three datasets. Best in **bold**, second best underlined.

Method	DAVIS			Online			APTv2		
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
Fauna [16]	18.281	0.761	0.280	16.669	0.774	0.279	19.561	0.760	0.267
D-3DGS* [43]	<u>25.776</u>	<u>0.905</u>	<u>0.100</u>	<u>25.659</u>	<u>0.912</u>	<u>0.096</u>	23.636	0.850	0.144
GART [13]	19.486	0.810	0.201	21.006	0.841	0.181	18.564	0.807	0.168
GART* [13]	21.347	0.859	0.171	23.128	0.883	0.158	19.167	0.834	0.150
GVFDiffusion [8]	16.419	0.836	0.174	16.835	0.857	0.167	14.820	0.778	0.256
DreamMesh4D [15]	23.150	0.881	0.116	23.859	0.889	0.131	26.000	0.899	0.102
Ours	26.321	0.920	0.089	26.206	0.923	0.091	<u>24.938</u>	<u>0.878</u>	<u>0.139</u>

Table 2: Novel-view quality on three datasets. Best in **bold**, second best underlined. A dash indicates metric not reported by the method.

Method	DAVIS			Online			APTv2		
	KID-16V $\downarrow$	FVD-F $\downarrow$	FVD-Diag $\downarrow$	KID-16V $\downarrow$	FVD-F $\downarrow$	FVD-Diag $\downarrow$	KID-16V $\downarrow$	FVD-F $\downarrow$	FVD-Diag $\downarrow$
Fauna [16]	0.279	—	—	0.334	—	—	0.247	—	—
D-3DGS* [43]	0.199	1192.015	<u>980.747</u>	0.231	<u>1244.444</u>	<u>1268.110</u>	0.262	1181.275	1112.902
GART [13]	0.216	1750.899	1675.388	0.233	1470.950	1547.561	0.230	1355.4969	992.195
GART* [13]	0.208	1680.705	1579.871	0.238	1473.927	1364.587	0.228	1134.661	948.274
GVFDiffusion [8]	<u>0.145</u>	1872.192	1270.189	<u>0.179</u>	1673.419	1387.732	0.274	1715.471	1575.461
DreamMesh4D [15]	0.148	<u>1154.038</u>	2257.034	0.185	1518.163	2911.299	0.188	629.482	720.280
Ours	0.108	895.708	696.166	0.160	972.028	1245.358	<u>0.189</u>	996.336	916.325

Progressive Stabilization. To prevent overfitting to inaccurate priors, the predicted pose $\hat{\mathbf{P}}_{\text{linear}}^t$ is blended with the Fauna prior using a weight that anneals from 1 to 0 over 7K iterations, and an L_2 regularization toward the prior is applied during the first 4K iterations. Both constraints vanish during training, ensuring the final reconstruction is not limited by prior accuracy. Following [43], we also inject decaying Gaussian noise into the temporal coordinate t during early iterations to encourage temporal smoothness.

4 Experiment

4.1 Dataset

We collect 87 in-the-wild videos from three sources: online collection (11 videos), DAVIS [24] (8 videos), and APTv2 [42] (68 videos). For APTv2, all videos contain 15 frames, except for two manually composed by concatenating similar clips. We extract animal masks using source-specific strategies (details in supplementary material), estimate animal and camera parameters for both original and horizontally flipped sequences using Fauna, and select temporally stable frames via DBSCAN clustering [4] on camera trajectories, resulting in a 4:1 train/test split.

To obtain more reliable measurements, we further evaluate on the Artemis dataset [21], which provides 9 CGI animal species with 24 motions captured by 36/72 synchronized cameras. We optimize on a single sampled monocular sequence and test on the remaining held-out views (1:71 for wolf, 1:35 for others) with dataset-provided cameras.

Table 3: Quantitative results on the Artemis dataset, covering method comparison and component ablations. Best in bold, second best underlined.**(a)** Novel-view quality and 3D geometric accuracy on the Artemis dataset.

Methods	Novel-View (2D)			Geometry (3D)			
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	CD/diag $\downarrow$	F@1% $\uparrow$	F@2% $\uparrow$	F@5% $\uparrow$
GART*	15.898	0.848	0.199	0.0256	0.237	0.519	0.879
D-3DGS*	<u>20.050</u>	<u>0.903</u>	<u>0.103</u>	<u>0.0181</u>	<u>0.358</u>	<u>0.683</u>	<u>0.952</u>
Ours	24.027	0.938	0.065	0.0156	0.439	0.760	0.959

(b) Ablation study on Artemis.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
w/o Deform	22.796	0.929	0.072
w/o Shape Prior	23.891	0.937	0.068
w/o Part Anchors	23.348	0.927	0.086
w/o Symm. Enc.	23.873	0.937	0.068
Fixed Skinning	23.934	0.938	0.067
Full	24.027	0.938	0.065

4.2 Implementation Details

We implement our method in PyTorch. Optimization runs for 20K iterations on DAVIS, Online and Artemis, and 10K on APTv2. For in-the-wild videos, we apply pose blending and prior regularization as described in Sec. 3.4; for Artemis, these are disabled. Further details are in the supplementary material. On a single NVIDIA A6000 GPU at 512×512 resolution, rendering averages 105 FPS, and per-video optimization takes approximately 13 minutes on APTv2 (10K iterations).

4.3 Results and Comparisons

Baselines We compare against five methods across two paradigms. Reconstruction-based: (1) Fauna [16], a single-image 3D model serving as our initialization prior; (2) GART [13], an articulated monocular optimization framework evaluated both standardly and with test-time refinement (denoted as GART*); and (3) D-3DGS [43], a dynamic 3DGS approach. For strict fairness, we equip D-3DGS with our Fauna initialization and silhouette supervision (denoted D-3DGS*). Additionally, we enhance both GART and D-3DGS with flip augmentation (assigning timestamp $t \cdot v$ to flipped frames). Generation-based: Video-to-4D models (4) GVFDiffusion [8] and (5) DreamMesh4D [15], powered by heavy pre-trained priors Trellis [35] and Zero-1-to-3 [19], respectively.

Evaluation Protocol. We adopt standard ground-truth metrics for evaluating input-view reconstruction. For generation-based methods (GVFDiffusion, DreamMesh4D) that do not produce camera-aligned outputs, we rotate the generated model and select the viewpoint with the highest PSNR for input-view evaluation, providing an upper bound on their input-view performance. For novel-view quality, we follow SV4D [37] and use FVD-F (temporal coherence at a fixed view) and FVD-Diag (spatio-temporal consistency). We introduce KID-V (Fig. 7a), a novel-view variant of KID [3], to evaluate novel-view quality (details in supp.), replacing FVD-V for two key reasons. First, FVD-V requires equal frame and view counts, biasing evaluation toward longer videos. Second,

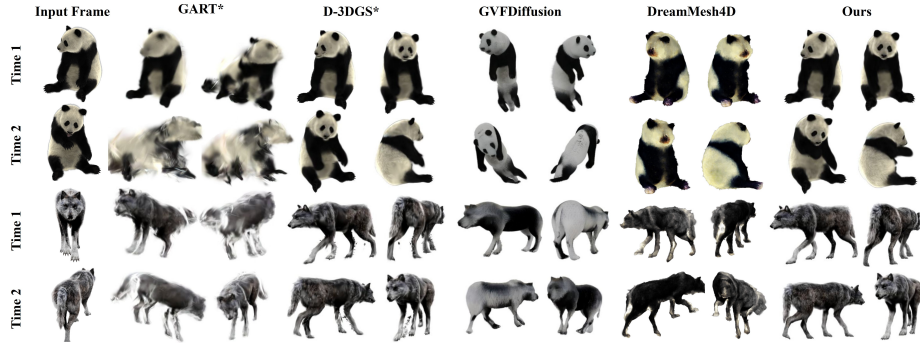


Fig. 4: Visual comparison with SOTA methods on Artemis at 2 different time steps.

FVD-V primarily measures temporal fluency; however, videos rendered by orbiting explicit 3D representations at a fixed timestep are inherently smooth by construction, making FVD-V redundant with per-frame metrics. Instead, KID-V uniformly samples novel views to provide an informative, unbiased measure that remains stable even with limited samples per timestep. We compute KID-V on novel test-set views, while FVD-F and FVD-Diag evaluate the entire generated sequence. For Fauna, a single-image method, we report only KID-16V.

Quantitative Comparison Tab. 1 and Tab. 2 report input-view and novel-view metrics. Our method consistently outperforms all baselines on DAVIS and Online across both settings. DreamMesh4D performs best on APTv2 input views, where its strong generative prior compensates for the sparse 15-frame sequences with limited camera variation, while our method ranks second across all metrics.

Tab. 3a reports quantitative results on the synthetic dataset Artemis, which provides multi-view RGBA renderings and rigged canonical voxel geometry. The 2D metrics evaluate novel-view rendering quality. For 3D geometric evaluation, we derive ground-truth surface points from the canonical voxels via 6-neighbor surface filtering and warp them per frame using the provided skeletal rig (10k points per frame). Predicted Gaussian centers are aligned to the ground truth via 7-DOF similarity ICP; we then report Chamfer distance and F-score, both normalized by the per-sequence rest-pose bounding-box diagonal. Our method outperforms both D-3DGS* and GART* by a substantial margin across all metrics. GVFDiffusion and DreamMesh4D are excluded from this table: their pipelines do not expose camera parameters, precluding the pixel-aligned 2D evaluation; moreover, as their outputs deviate substantially from the input identity, 3D geometric comparison would conflate generation fidelity with reconstruction accuracy. We instead provide a qualitative comparison with all methods in Fig. 4.

Qualitative Comparison As shown in Fig. 3, baseline methods exhibit various failure modes: incomplete limb geometry (GART*, D-3DGS*), hallucinated

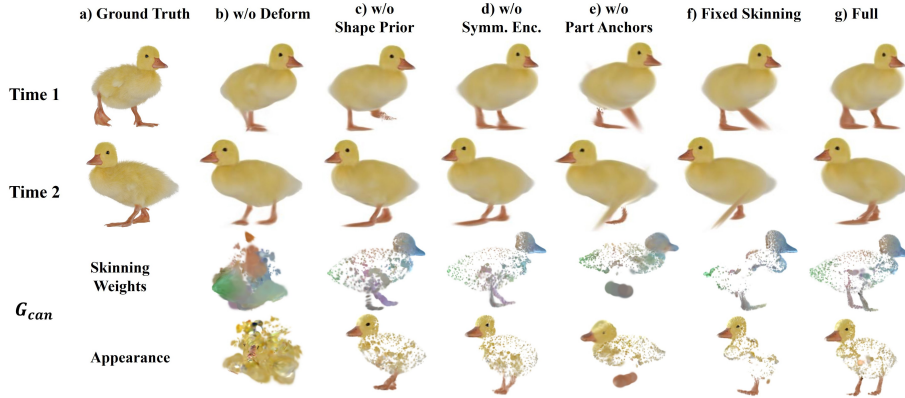


Fig. 5: Ablation study on Artemis. Top two rows: rendered frames at two time steps. Bottom two rows: canonical Gaussians G_{can} colored by skinning weights and appearance.

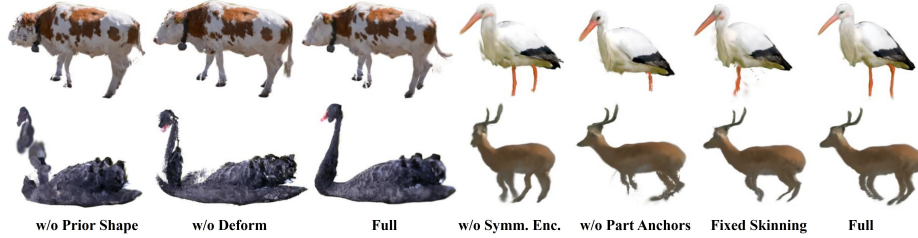


Fig. 6: Novel-view synthesis under ablations, showing degraded reconstructions without certain components, while the full model remains stable and accurate.

structures (GVFDiffusion), and appearance inconsistencies (DreamMesh4D). Our method maintains faithful reconstruction across all species. Fig. 4 shows comparisons on Artemis. GART* fails on poses that deviate from its D-SMAL template (e.g., the sitting panda) but performs reasonably when the body plan matches (wolf). D-3DGS* exhibits floating artifacts and misses fine motion. GVFDiffusion and DreamMesh4D both fail to preserve instance-specific appearance, producing outputs that deviate substantially from the input. Our method faithfully reconstructs both shape and motion across time.

Ablation Study We validate each component through quantitative metrics (Tab. 3b) and visualization (Fig. 5). Specifically: **w/o Deform** retains only the pose refinement stage; **w/o Shape Prior** initializes Gaussian positions randomly instead of from prior mesh vertices; **w/o Symm. Enc.** replaces our symmetry-aware encoding with $\text{emb}(t \cdot v) \oplus v$; **w/o Part Anchors** removes the learnable anchors and instead concatenates raw skinning weights onto posed attributes as a part identity substitute; **Fixed Skinning** freezes the skinning field at its prior-initialized values. As indicated in Tab. 3b, our full model achieves the best

performance across all metrics, with the deformation module and part anchors driving the most significant gains in perceptual quality.

The canonical-space visualizations (Fig. 5) expose the structural root causes of these performance drops. Without the deformation module (Fig. 5b), the model is forced to absorb non-rigid dynamic variations directly into the canonical space, producing a disordered $\mathcal{G}_{can}$. While the final rendering may appear plausible at a coarse glance, subtle limb articulations (note the feet orientation at Time 2) are lost. Both removing the shape prior (Fig. 5c) and disabling symmetry encoding (Fig. 5d) result in an asymmetric canonical shape where one leg is placed in front of the other. Without prior initialization, Gaussians tend to aggregate toward the most frequently observed positions rather than a neutral rest pose. Without symmetry encoding, the model lacks the bilateral cues to recover a symmetric $\mathcal{G}_{can}$. In both cases, reconstruction fails when the target pose reverses the leg ordering relative to the overfitted canonical configuration. Removing part anchors (Fig. 5e) and fixing the skinning field (Fig. 5f) produce similar artifacts in the final rendering—only one leg is reconstructed, with the other approximated by stretched Gaussians—but degrade differently in canonical space. In (Fig. 5e), the canonical shape is tilted and high-frequency moving parts are collapsed into simplified blobs; in (Fig. 5f), the canonical shape itself contains only one leg. Both variants force the deformation stage to hallucinate the missing limb, confirming that learnable part-aware modeling is essential for faithful reconstruction. In contrast, the full model (Fig. 5g) recovers a well-structured canonical representation and uniquely captures both correct articulated pose and non-rigid deformation of the duck’s feet.

We further visualize ablation results on in-the-wild videos in Fig. 6. Removing the shape prior leads to severe structural collapse, while removing the deformation module preserves overall structure but introduces artifacts and fails to capture fine motion such as the cow’s tail swing. Disabling symmetry encoding leads to blurred or distorted geometry in real videos with camera estimation noise. Without part anchors or with a fixed skinning field, limb reconstruction degrades noticeably, confirming their importance for unconstrained settings.

4.4 Discussion and Limitations

While our method performs well across diverse species, several limitations remain.

Robustness to Unstable Camera Estimation. Our filtering strategy processes the original and flipped sequences independently. A frame discarded in one sequence due to unstable camera estimation can still be optimized through its counterpart in the other sequence via our bilateral augmentation. Nevertheless, under severe camera instability where both sequences are affected, optimization may converge to a compromised solution.

Imperfect Mask Supervision. We directly use Grounded-SAM masks without manual refinement; inaccurate segmentation reduces visual supervision and degrades geometry in affected regions. Side-to-side self-occlusion is partially mitigated by our bilateral augmentation, which supplies supervision from the mirrored view, while persistent occlusion across the sequence remains a limitation.

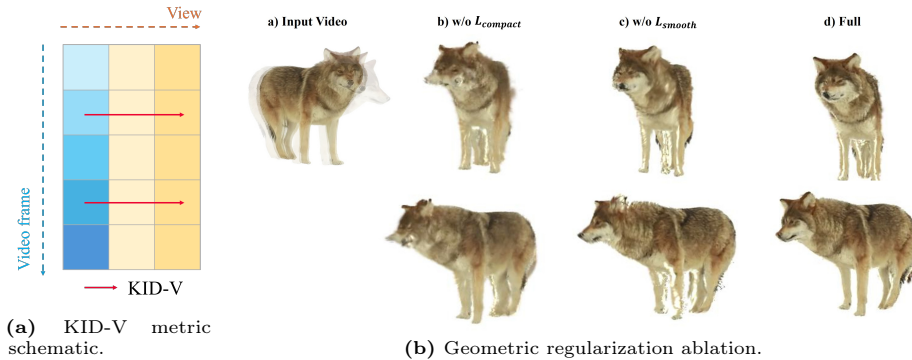


Fig. 7: (a) Conceptual illustration of our proposed KID-V metric. (b) Visual ablation of $\mathcal{L}_{compact}$ and $\mathcal{L}_{smooth}$. Without $\mathcal{L}_{compact}$, the head splits into duplicated fragments. Without $\mathcal{L}_{smooth}$, the legs and head exhibit rough, noisy surfaces. The full model maintains coherent geometry with clean surfaces across viewpoints.

Ambiguous Localized Motion. Head motion remains difficult under limited viewpoints. Unlike limbs, head rotations produce only subtle silhouette changes, making silhouette supervision insufficient. Although our geometric regularization alleviates head splitting (Fig. 7b), it cannot fully resolve this ambiguity.

Evaluation Limitations. Rendering-based metrics are indirect proxies for geometric accuracy: high rendering scores do not guarantee correct articulation. On Artemis, the geometric metrics provide a more direct assessment, but for in-the-wild videos ground-truth 3D geometry is unavailable, leaving rendering-based proxies as the only option.

Future work includes improving robustness to unreliable camera and mask estimates, incorporating stronger geometric or structural priors, and developing more reliable evaluation protocols.

5 Conclusion

We presented a progressive test-time optimization framework for 4D animal reconstruction from monocular video. By treating shape priors as coarse initialization with a learnable skinning field, introducing symmetry-aware temporal encoding to exploit bilateral cues under camera noise, and conditioning non-rigid deformation on part anchors shared with the pose estimation stage, our method disentangles articulated motion from non-rigid deformation without requiring multi-view supervision or category-specific templates. Extensive experiments across diverse species demonstrate improvements over existing baselines in reconstruction quality and temporal consistency. While developed for animals, the principle of balancing prior knowledge with optimization flexibility may extend to other articulated non-rigid objects.

References

1. Aygun, M., Mac Aodha, O.: Saor: Single-view articulated object reconstruction. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10382–10391 (2024)
2. Badger, M., Wang, Y., Modh, A., Perkes, A., Kolotouros, N., Pfrommer, B.G., Schmidt, M.F., Daniilidis, K.: 3d bird reconstruction: a dataset, model, and shape recovery from a single view. In: *European conference on computer vision*. pp. 1–17. Springer (2020)
3. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. In: *International Conference on Learning Representations* (2018)
4. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: *kdd*. vol. 96, pp. 226–231 (1996)
5. Huang, Y.H., Sun, Y.T., Yang, Z., Lyu, X., Cao, Y.P., Qi, X.: Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 4220–4230 (2024)
6. Jakab, T., Li, R., Wu, S., Rupprecht, C., Vedaldi, A.: Farm3d: Learning articulated 3d animals by distilling 2d diffusion. In: *2024 International Conference on 3D Vision (3DV)*. pp. 852–861. IEEE (2024)
7. Jiang, B., Hong, Y., Bao, H., Zhang, J.: Selfrecon: Self reconstruction your digital avatar from monocular video. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022)
8. Jiang, B., Hong, Y., Bao, H., Zhang, J.: Gaussian variation field diffusion for high-fidelity video-to-4d synthesis. In: *IEEE/CVF International Conference on Computer Vision (ICCV)* (2025)
9. Jiawei, X., Zexin, F., Jian, Y., Jin, X.: Grid4D: 4D decomposed hash encoding for high-fidelity dynamic gaussian splatting. *The Thirty-eighth Annual Conference on Neural Information Processing Systems* (2024)
10. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* **42**(4), 139–1 (2023)
11. Kulkarni, N., Gupta, A., Fouhey, D.F., Tulsiani, S.: Articulation-aware canonical surface mapping. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 452–461 (2020)
12. Kulkarni, N., Gupta, A., Tulsiani, S.: Canonical surface mapping via geometric cycle consistency. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 2202–2211 (2019)
13. Lei, J., Wang, Y., Pavlakos, G., Liu, L., Daniilidis, K.: Gart: Gaussian articulated template models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 19876–19887 (2024)
14. Li, X., Liu, S., Kim, K., De Mello, S., Jampani, V., Yang, M.H., Kautz, J.: Self-supervised single-view 3d reconstruction via semantic consistency. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV* 16. pp. 677–693. Springer (2020)
15. Li, Z., Chen, Y., Liu, P.: Dreammesh4d: Video-to-4d generation with sparse-controlled gaussian-mesh hybrid representation. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2024)
16. Li, Z., Litvak, D., Li, R., Zhang, Y., Jakab, T., Rupprecht, C., Wu, S., Vedaldi, A., Wu, J.: Learning the 3d fauna of the web. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9752–9762 (2024)

17. Liao, T.H., Liu, H., Xu, Y., Ge, S., Yang, G., Huang, J.B.: Pad3r: Pose-aware dynamic 3d reconstruction from casual videos. SIGGRAPH ASIA (2025)
18. Liu, D., Stathopoulos, A., Zhangli, Q., Gao, Y., Metaxas, D.: Leopard: Learning explicit part discovery for 3d articulated shape reconstruction. *Advances in Neural Information Processing Systems* **36**, 54187–54198 (2023)
19. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 9298–9309 (2023)
20. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. In: *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 851–866 (2023)
21. Luo, H., Xu, T., Jiang, Y., Zhou, C., Qiu, Q., Zhang, Y., Yang, W., Xu, L., Yu, J.: Artemis: articulated neural pets with appearance and motion synthesis. *ACM Transactions on Graphics (TOG)* **41**(4), 1–19 (2022)
22. Oh, S., Lee, Y., Jeon, H., Park, E.: Hybrid 3D–4D Gaussian Splatting for fast dynamic scene representation (2025), <https://arxiv.org/abs/2505.13215>, arXiv preprint arXiv:2505.13215. Accessed 29 June 2026
23. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (2019)
24. Perazzi, F., Pont-Tuset, J., McWilliams, B., Gool, L.V., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016)
25. Ren, J., Xie, K., Mirzaei, A., Liang, H., Zeng, X., Kreis, K., Liu, Z., Torralba, A., Fidler, S., Kim, S.W., Ling, H.: L4gm: Large 4d gaussian reconstruction model. In: *Proceedings of Neural Information Processing Systems(NeurIPS)* (Dec 2024)
26. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024)
27. Rueegg, N., Zuffi, S., Schindler, K., Black, M.J.: Barc: Learning to regress 3d dog shape from images by exploiting breed information. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 3876–3884 (2022)
28. Rüegg, N., Tripathi, S., Schindler, K., Black, M.J., Zuffi, S.: Bite: Beyond priors for improved three-d dog pose estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8867–8876 (2023)
29. Sabathier, R., Mitra, N.J., Novotny, D.: Animal avatars: Reconstructing animatable 3d animals from casual videos. In: *European Conference on Computer Vision*. pp. 270–287. Springer (2024)
30. Song, M., Zhang, Y., Mihajlovic, M., Tang, S., Gross, M., Aydın, T.O.: Spline deformation field. In: *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. pp. 1–10 (2025)
31. Wang, Y., Kolotouros, N., Daniilidis, K., Badger, M.: Birds of a feather: Capturing avian shape models from images. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 14739–14749 (2021)
32. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: *Proceedings of the*

- IEEE/CVF conference on computer vision and pattern recognition. pp. 20310–20320 (2024)
33. Wu, S., Li, R., Jakab, T., Rupperecht, C., Vedaldi, A.: Magicpony: Learning articulated 3d animals in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8792–8802 (2023)
 34. Wu, Y., Chen, Z., Liu, S., Ren, Z., Wang, S.: Casa: Category-agnostic skeletal animal reconstruction. *Advances in Neural Information Processing Systems* **35**, 28559–28574 (2022)
 35. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21469–21480 (2025)
 36. Xie, Y., Yao, C.H., Voleti, V., Jiang, H., Jampani, V.: Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. In: The Thirteenth International Conference on Learning Representations
 37. Xie, Y., Yao, C.H., Voleti, V., Jiang, H., Jampani, V.: Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. In: The Thirteenth International Conference on Learning Representations
 38. Yang, G., Sun, D., Jampani, V., Vlasic, D., Cole, F., Chang, H., Ramanan, D., Freeman, W.T., Liu, C.: Lasr: Learning articulated shape reconstruction from a monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15980–15989 (2021)
 39. Yang, G., Sun, D., Jampani, V., Vlasic, D., Cole, F., Liu, C., Ramanan, D.: Viser: Video-specific surface embeddings for articulated 3d shape reconstruction. *Advances in Neural Information Processing Systems* **34**, 19326–19338 (2021)
 40. Yang, G., Vo, M., Neverova, N., Ramanan, D., Vedaldi, A., Joo, H.: Banmo: Building animatable 3d neural models from many casual videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2863–2873 (2022)
 41. Yang, G., Wang, C., Reddy, N.D., Ramanan, D.: Reconstructing animatable categories from videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16995–17005 (2023)
 42. Yang, Y., Deng, Y., Xu, Y., Zhang, J.: Aptv2: Benchmarking animal pose estimation and tracking with a large-scale dataset and beyond (2023)
 43. Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20331–20341 (2024)
 44. Yao, C.H., Hung, W.C., Li, Y., Rubinstein, M., Yang, M.H., Jampani, V.: Lassie: Learning articulated shapes from sparse image ensemble via 3d part discovery. *Advances in Neural Information Processing Systems* **35**, 15296–15308 (2022)
 45. Yao, C.H., Hung, W.C., Li, Y., Rubinstein, M., Yang, M.H., Jampani, V.: Hi-lassie: High-fidelity articulated shape and skeleton discovery from sparse image ensemble. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4853–4862 (2023)
 46. Yao, C.H., Xie, Y., Voleti, V., Jiang, H., Jampani, V.: Sv4d 2.0: Enhancing spatio-temporal consistency in multi-view video diffusion for high-quality 4d generation. *arXiv preprint arXiv:2503.16396* (2025)
 47. Yin, M., Cao, Y., Peng, S., Han, K.: Splat4d: Diffusion-enhanced 4d gaussian splatting for temporally and spatially consistent content creation. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–10 (2025)

48. Zuffi, S., Kanazawa, A., Berger-Wolf, T., Black, M.J.: Three-d safari: Learning to estimate zebra pose, shape, and texture from images "in the wild". In: International Conference on Computer Vision (Oct 2019)
49. Zuffi, S., Kanazawa, A., Black, M.J.: 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. pp. 3955–3963 (2018)
50. Zuffi, S., Kanazawa, A., Jacobs, D.W., Black, M.J.: 3d menagerie: Modeling the 3d shape and pose of animals. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6365–6373 (2017)