# RAT4D: Rig and Animate any object without Templates in 4D

**Mosam Dabhi**[1] **Simon Lucey**[2*] **László A. Jeni**[1*]

[1] **Carnegie Mellon University** [2] **The University of Adelaide**

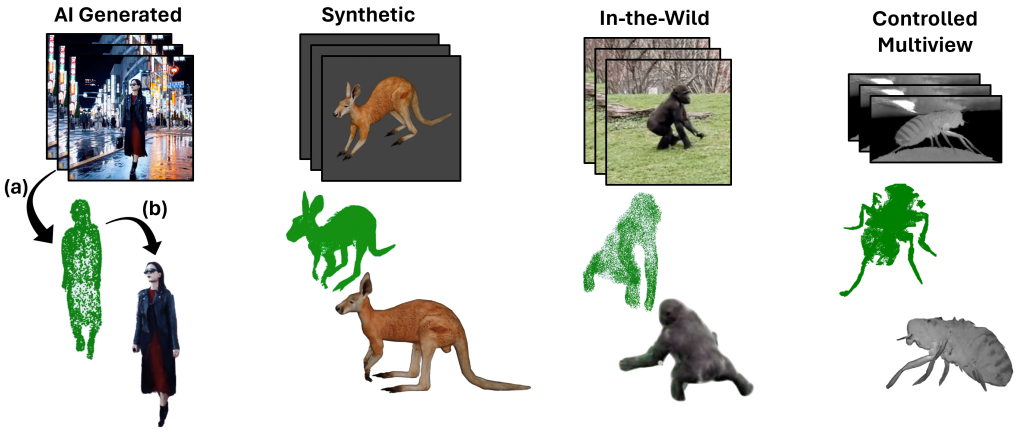


Figure 1: Diverse input sources and object categories processed by our method. From left to right: AI-generated frames, synthetic data, in-the-wild captures, and controlled multiview setups. Each column shows the input frames, rigged 3D Gaussian splatting model, and novel re-posing rendered results.

## Abstract

Achieving high-fidelity reconstructions of dynamic and deformable objects typically relies on costly, calibrated camera rigs or proprietary 3D templates like SMPL, making reconstructions in uncontrolled environments challenging. We propose a method to create re-animatable 3D models from monocular images, bypassing pre-built templates and introducing three key innovations. First, we use noisy 2D landmark detectors and 3D lifters, often limited by missing camera intrinsics and inconsistent bone lengths. We introduce a novel nonlinear optimization process that refines landmark detection and skeletal rigidity from monocular sequences, enhancing 3D accuracy. Second, we show that 3D parametric templates, limited to quadrupeds and humans, are not required. We use a new heuristic based on the 3D lifted skeleton, doubling as an animation rig, and a Gaussian-splatting differentiable renderer that handles a wide array of objects, including non-human and inanimate entities. Finally, we demonstrate that these innovations enable robust re-animations and reconstructions of dense 3D geometry from uncalibrated monocular sequences, without prior 3D template knowledge. We evaluated our method with controlled and in-the-wild datasets featuring a diverse array of deformable objects. It achieved comparable or better results in novel view and pose synthesis than leading template-based methods, as measured by PSNR and SSIM.

[*] indicates the authors advised equally

Preprint. Under review.

# 1 Introduction

In the rapidly evolving field of 3D asset generation, creating high-quality 3D models from simple inputs such as RGB videos remains a formidable challenge. Existing methods often rely heavily on predefined 3D surface templates like SMPL [15] and SMAL [3]. While these templates provide a structured approach to modeling, they come with significant limitations. They restrict the applicability of methods to only those categories for which templates exist and bind users to the intellectual property constraints of the template owners. Additionally, these methods assume the availability of accurate 3D poses and camera information, which is rarely the case in unconstrained, real-world scenarios.

The quest for a versatile, template-free solution that can handle a broad spectrum of dynamic and deformable objects has been a long-standing goal in the research community. Traditional approaches fall short in addressing the complexities of various object categories, especially non-humans and inanimate entities. This has created a substantial gap in the ability to generate high-fidelity 3D models in a flexible and widely applicable manner. For instance GART [13], a leading method that utilizes Gaussian splatting, still depends on SMPL and SMAL templates – limiting its versatility and applicability across diverse object categories.

Our work addresses these fundamental limitations by introducing a novel, geometry-based pipeline that operates effectively with minimal peripheral input other than the RGB sequence itself. Central to our innovation is the replacement of object category specific 3D surface templates with agnostic geometry-based initialization of Gaussian parameters and skinning weights for the deformable object. Unlike [13], our approach does not need the templates to initialize the Gaussian splats, enabling us to handle a diverse range of non-traditional object categories. To get the 3D pose and camera positions we leverage off-the-shelf 2D pose detections [21] and state-of-the-art 3D lifting models [22, 8]. In practice, these estimates are often noisy and inadequate for high-quality 3D reconstruction. We also propose a novel non-linear optimization step that not only refines these landmarks by optimizing for consistent geometric bone proportions (bone lengths) but also refines camera parameters. This step ensures accurate 3D pose estimation and significantly improves MPJPE performance – especially for 3D lifting models that have only been trained on atemporal data.

Our extensive evaluations demonstrate that our method achieves comparable or superior performance in terms of PSNR and SSIM for novel view and pose synthesis compared to leading template-based methods. Additionally, we showcase a novel application of our approach: detecting and correcting inaccuracies in AI-generated videos. This highlights the importance of geometric accuracy in AI-based video creation frameworks and opens new avenues for ensuring the quality and reliability of AI-generated content. Our contributions are as follows:-

- **Challenging the Need for 3D Surface Templates:** We eliminate the dependency on predefined 3D surface templates, using a geometry-based initialization of Gaussian parameters. This approach broadens the applicability of 3D asset generation methods to a wide variety of object categories, including non-humans and inanimate objects.

- **Nonlinear Refinement for Enhanced 3D Accuracy:** Our non-linear optimization step not only improves MPJPE performance but also serves as a general enhancement technique that can be integrated into any 3D lifting method. This step is particularly beneficial for methods trained on spatial supervision, bringing their performance closer to spatiotemporal-based approaches.

- **Extensive Experimentation and Novel Applications:** Through rigorous experimentation, we demonstrate superior or comparable results in 3D reconstruction, novel view synthesis, and novel pose synthesis. Furthermore, we present a unique application of our method in detecting and correcting inaccuracies in AI-generated videos, underscoring the practical relevance and innovation of our approach.

By addressing these critical challenges and demonstrating significant improvements, our work sets a new benchmark in 3D asset generation, pushing the boundaries of what is possible in dynamic and deformable object reconstruction.

## 2 Related Works

**Differentiable Rendering for Deformable Models.** Recent advancements in differentiable rendering, exemplified by methods such as GART (Gaussian Articulated Template Models) [13] and Instant-NVR (Neural Radiance Fields) [9], have demonstrated significant success in generating high-quality 3D models. GART utilizes Gaussian splatting for rendering, whereas Instant-NVR employs Neural Radiance Fields (NeRF). Despite their effectiveness, both approaches heavily depend on predefined 3D surface templates, such as SMPL [15] for humans and SMAL [3] for animals. This reliance restricts their utility to object categories for which templates are available. This limitation poses a substantial challenge, especially in accurately capturing 3D poses and camera parameters for diverse and less common categories, like fish, kangaroos, and insects. In contrast, our work introduces a template-free method that employs geometric initialization. This approach uses our novel optimization framework to learn dense, deformable rigs and skinning weights directly from 'in the wild' images, thus expanding the scope of differentiable rendering to include a broader range of objects.

**3D Poses and Camera Parameters for Rendering Pipelines** Supervised methods such as HybrIK [14], SMPLify [17], and SMALify [2] are designed to predict 3D poses directly from images. However, their applicability is limited to specific categories like humans or quadruped animals, and they struggle to generalize to long-tail categories that lack available templates. Furthermore, these methods do not explicitly enforce constraints on geometric bone proportions or ensure that the generated 3D poses accurately align with the corresponding 2D landmark data. This omission can lead to significant inaccuracies in the reconstruction of dynamic and deformable objects.

**2D to 3D Landmark Lifting** To address the challenges in predicting 3D poses (joint angles), several Non-Rigid Structure from Motion (NRSfM) based approaches for lifting 2D landmarks to 3D locations have been developed, such as MotionBERT [22], PAUL [20], C3DPO [16], and 3D-LFM [8]. MotionBERT utilizes a spatio-temporal transformer but is limited to human categories, specifically formatted for the Human3.6M dataset (17-joint humans) [11]. While it successfully generates 3D locations, it does not produce joint angles and fails to enforce consistency in geometric bone proportions, resulting in implausible 3D reconstructions in 'in the wild' sequences. Conversely, 3D-LFM [8] marks a significant advancement as it aims to lift 2D landmarks for any deformable object category. However, it still primarily generates 3D locations without explicit temporal modeling and lacks enforcement of consistent geometric bone proportions.

In this paper, we introduce a runtime optimization step for pose optimization that enhances existing 2D to 3D lifting methods, coupled with a template-free deformable 3D modeling approach based on Gaussian splatting and differentiable rendering. The proposed optimization step refines 3D poses by transforming noisy 3D locations into accurate joint angles, enforces consistent geometric bone proportions across sequences, and optimizes camera parameters to achieve precise calibrations aligned with 2D landmark data. This enhancement not only boosts the performance of spatial-only trained methods, making them competitive with spatio-temporal approaches, but also facilitates large-scale 3D data collection in uncontrolled environments. Additionally, our template-free modeling technique leverages adaptive Gaussian parameters to accurately render a wide range of dynamic and deformable objects without the need for pre-existing 3D templates, promoting broader applicability and innovation in 3D reconstruction.

## 3 Methodology

In this section, we outline our methodology for generating high-fidelity 3D assets from simple RGB videos. Our approach consists of two main components: a novel nonlinear optimization for refining 3D poses and camera parameters, and a template-free, geometry-based method for initializing and refining Gaussian parameters for differentiable rendering.

### 3.1 Nonlinear Optimization for 3D Poses and Camera Parameters

Given 2D pose detections $\mathbf{P}_{2D}$ from off-the-shelf detectors, we use state-of-the-art 3D lifting models to obtain initial 3D landmarks $\mathbf{J}_{3D}$. However, these landmarks are often noisy, leading to inaccurate

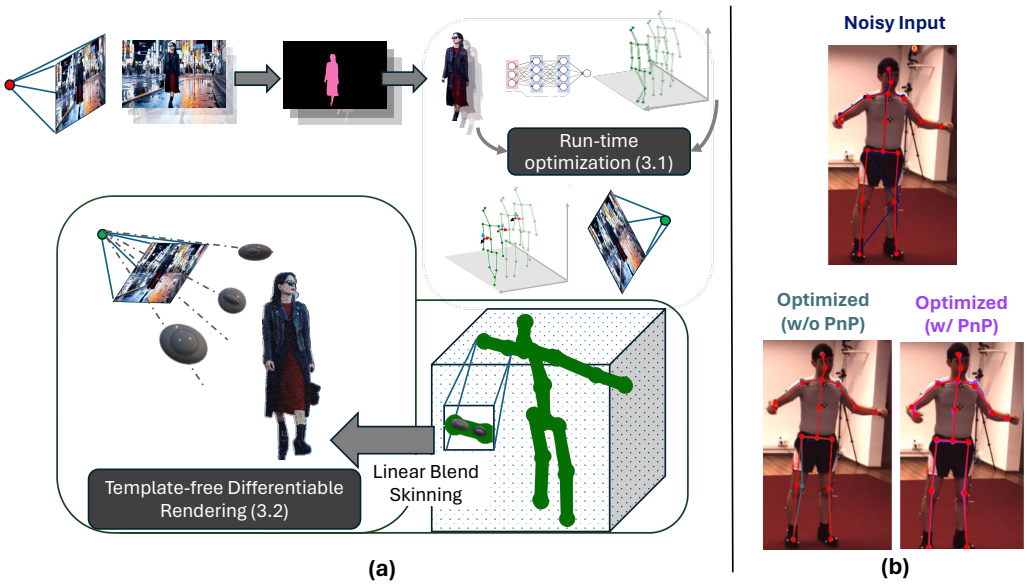


Figure 2: **(a)** *Method overview*: Overview of our template-free 3D asset generation pipeline. Starting from 2D images, the process involves segmentation [6], initial 3D pose estimation, nonlinear optimization for pose refinement, and template-free Gaussian splatting for differentiable rendering. The final model is capable of high-fidelity re-rendering in novel poses. **(b)** Initial noisy input and our optimized solutions. Optimization refines poses, but using 2D information (PnP) achieves final accuracy and consistency.

poses and unreliable camera parameters. To address this, we introduce the following variables and parameters:

Table 1: Optimization Parameters

| | | | |
|---|---|---|---|
| $\mathbf{J}$ | 3D joint locations (landmarks) | $\mathbf{L}$ | Bone lengths |
| $\mathbf{P}$ | Parent-child relationships between joints | $\mathbf{K}$ | Camera intrinsic matrix |
| $\theta$ | Joint angles | $\mathbf{P}^f_{2D,i}$ | 2D projected points |
| $\mathbf{t}$ | Global translations | $\mathbf{J}^f_{2D,i}$ | Observed 2D points |
| $\mathbf{R}$ | Global rotation | $\mathbf{P}^f_i$ | Optimized 3D joint positions |

Our goal is to refine the 3D poses and camera parameters by ensuring consistent bone lengths and accurate projections. To do this, we introduce a preprocessing step where we scale the incoming noisy frames to match the initial bone lengths. Then, we define the optimization problem as follows:

**Inverse Kinematics update for $\theta$** The optimization starts by initializing the parameters $\theta$, $\mathbf{R}$, $\mathbf{t}$, and $\mathbf{L}$ from the initial 3D landmarks. We then proceed with the nonlinear update of the transformation matrix $G$ for each joint $i$ based on its parent joint $p$ as follows:

$$G_i^f = G_p^f \cdot \begin{bmatrix} \mathcal{R}(\theta_i^f) & \mathcal{T}_i(\mathbf{L}) \\ \mathbf{0} & 1 \end{bmatrix}$$

where $G_i^f$ is the transformation matrix for joint $i$ at frame $f$, and $\mathcal{R}_i$, $\mathcal{T}_i$ are rodrigues rotation matrix, translation vector, respectively from the parent joint $p$. Translation vector is a function of bone lengths. Final transformation results in a complete forward pass of kinematic chain, giving $\mathbf{P}_i^f = G_i^f$.

**Perspective-N-Point update for $\mathbf{R}, \mathbf{t}$** We integrate the PnP approach [7, 20] to jointly optimize the pose parameters $\theta$, rotation $\mathbf{R}$, and translation $\mathbf{t}$. This ensures the projected 2D points $\mathbf{P}^f_{2D,i}$ match the observed 2D points $\mathbf{J}^f_{2D,i}$:

$$\mathbf{P}^f_{2D,i} = \mathbf{K}\left(\mathbf{R}\mathbf{P}^f_i + \mathbf{t}\right)$$

**Temporal smoothness for $\theta$** To ensure temporal consistency and smoothness of the poses, we incorporate regularization terms in the loss function:

$$\min_{\theta,\mathbf{R},\mathbf{t},\mathbf{L}} \sum_{f=1}^{F}\sum_{i=1}^{N}\left\|\mathbf{P}^f_{2D,i} - \mathbf{J}^f_{2D,i}\right\|^2 + \lambda_1 \sum_{f=1}^{F}\sum_{i=1}^{N}\left\|\mathbf{P}^f_i - \mathbf{J}^f_i\right\|^2 + \lambda_2 \sum_{f=2}^{F}\left\|\theta^f - \theta^{f-1}\right\|^2 + \lambda_3 \sum_{f=2}^{F}\left\|\theta^f - 2\theta^{f-1} + \theta^{f-2}\right\|^2 \quad (1)$$

where the terms $\lambda_2$ and $\lambda_3$ enforce the smoothness and acceleration constraints respectively, ensuring the poses are consistent and smooth over time.

In summary, our approach refines the initial noisy 3D landmarks by jointly optimizing the pose parameters, global transformations, and ensuring consistency in 2D projections, ultimately leading to more accurate and reliable 3D poses and camera parameters.

### 3.2 Template-Free Gaussian Initialization and Differentiable Rendering

To replace the dependency on predefined 3D surface templates, we employ a geometry-based method to initialize and refine Gaussian parameters for differentiable rendering. We introduce the following additional variables and parameters: $\mu$ and $\mathbf{\Sigma}$ are the means and covariances of Gaussian parameters, $\mathbf{W}$ are the skinning weights, $\mathbf{G}$ is the voxel grid, and $\mathbf{V}$ are the vertices of cylindrical mesh.

**Geometry-Based Initialization and Refinement** Given the 3D joint locations $\mathbf{J}$ and parent-child relationships $\mathbf{P}$, we initialize vertices on the surface of cylinders connecting each joint $i$ to its parent $p(i)$:

$$\mathbf{V}_i = \{\mathbf{v}_{ij} \mid j = 1, \ldots, M_i\}$$

where $\mathbf{v}_{ij}$ are points sampled on the cylinder's surface, and $M_i$ is the number of sampled points for joint $i$. The initial skinning weights for each sampled point $\mathbf{v}_{ij}$ are computed based on proximity to the joints as shown in Eq. 2, where $d(\mathbf{v}_{ij}, \mathbf{J}_k)$

$$w_{ijk} = \frac{1}{d(\mathbf{v}_{ij}, \mathbf{J}_k)} \qquad \mathbf{W}_{ij} = \frac{\mathbf{w}_{ij}}{\sum_{k=1}^{K} w_{ijk}} \quad (2)$$

is the distance between the point $\mathbf{v}_{ij}$ and joint $\mathbf{J}_k$. The weights are then normalized according to Eq. 2. To refine these initial weights, we use a voxel grid $\mathbf{G}$ and perform trilinear interpolation:

$$\mathbf{W}_{\text{voxel}} = \text{TrilinearInterpolate}(\mathbf{W}_{\text{init}}, \mathbf{G})$$

**Gaussian Parameter Initialization and Transformation** The vertices $\mathbf{V}$ are used to initialize the means $\mu$ of the Gaussians, while the refined weights determine the covariances $\mathbf{\Sigma}$:

$$\mu_i = \mathbf{v}_i, \quad \mathbf{\Sigma}_i = \mathbf{W}_{\text{voxel},i}$$

To transform the Gaussians from the canonical space to the posed space, we use forward kinematics:

$$\mathbf{T}_i = \mathbf{R}_i(\theta)\mathbf{\Sigma}_i\mathbf{R}_i(\theta)^\top + \mathbf{t}_i \quad (3)$$

$$\mu'_i = \mathbf{R}_i(\theta)\mu_i + \mathbf{t}_i \quad (4)$$

$$\mathbf{\Sigma}'_i = \mathbf{R}_i(\theta)\mathbf{\Sigma}_i\mathbf{R}_i(\theta)^\top \quad (5)$$

**Differentiable Rendering with Gaussian Splatting** Using the refined Gaussian parameters, we perform Gaussian splatting for differentiable rendering. The density $\rho$ at a pixel $\mathbf{p}$ is computed by accumulating the contributions from all Gaussians:

$$\rho(\mathbf{p}) = \sum_i \exp\left(-\frac{1}{2}(\mathbf{p} - \mu_i'')^\top \mathbf{\Sigma}_i''^{-1}(\mathbf{p} - \mu_i'')\right)$$

**Differentiable rendering losses** To ensure high-fidelity reconstruction, we incorporate both image and silhouette losses in our optimization:

$$\mathcal{L}_{\text{rendering}} = \sum_{\mathbf{p}} \|\mathbf{I}(\mathbf{p}) - \mathbf{I}_{\text{rendered}}(\mathbf{p})\|^2 + \sum_{\mathbf{p}} \|\mathbf{S}(\mathbf{p}) - \mathbf{S}_{\text{rendered}}(\mathbf{p})\|^2$$

where $\mathbf{I}(\mathbf{p})$ and $\mathbf{S}(\mathbf{p})$ are the ground truth image and silhouette values at pixel $\mathbf{p}$, and $\mathbf{I}_{\text{rendered}}(\mathbf{p})$ and $\mathbf{S}_{\text{rendered}}(\mathbf{p})$ are the corresponding rendered values.

This section now sets the stage and ensures a robust and flexible approach to 3D asset generation, enabling high-fidelity 3D models from simple RGB videos. Our extensive evaluations in the next section validate the effectiveness of our method across various applications, demonstrating significant improvements over traditional approaches.

# 4 Results

This section presents the experimental validation of our approach, divided into four subsections. We begin by demonstrating the effectiveness of our novel non-linear optimization step in improving 3D pose performance across various methods. We then justify the removal of 3D surface templates and show how our method performs in their absence. Next, we highlight the generalizability of our approach to diverse object categories, novel re-posing, as well as novel applications(see Sec. 4.3), including AI-generated video correction. Finally, we provide an ablation study of different design choices.

To evaluate the performance of our method, we will be using following metrics: **3D Pose Accuracy**: Measured by Mean Per Joint Position Error (MPJPE). **Reconstruction Quality**: Assessed using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). We delegate the discussion of datasets to the relevant subsections that follow.

## 4.1 Pose Optimization

The primary objective of our runtime optimization is to ensure high-quality 3D poses and camera parameters for template-free deformable modeling applications. By enhancing spatial-only lifters like [8], our approach aims to match the performance of spatiotemporal methods [22], demonstrating significant improvements through optimized kinematic chain constraints, geometric bone proportions, and camera parameters. We begin by validating our approach with synthetic experiments, paving the way to assess improvements on real-world lifters and showing how spatial-only lifters can boost their performance to match spatiotemporal ones.

**Synthetic Noise Experiments** To validate the robustness of our approach, we conducted experiments using sequences from subject #8 of the Human3.6M dataset with synthetic noise, including jitter and random flips (Fig. 2(b)), to simulate real-world conditions encountered by 3D lifters. We utilized three metrics: Mean Per Joint Position Error in 3D (**MPJPE 3D** in mm) for accuracy of 3D joint locations, Mean Square Successive Difference (**MSSD** [19]) for temporal smoothness, and Mean Per Joint Position Error in 2D (**MPJPE 2D** in pixels) for 2D projection accuracy and camera parameter quality. MPJPE 3D and MPJPE 2D are standard in evaluating 3D pose accuracy and projection fidelity, while MSSD is particularly chosen to highlight improvements in temporal consistency and reduction of jitter, which are not fully captured by MPJPE alone.

Our approach significantly reduces MPJPE 3D, MSSD, and MPJPE 2D across all noise levels as shown in Fig. 3. These results highlight our method's ability to enhance the accuracy and smoothness of 3D poses, even with substantial noise, thus enabling high-quality 3D poses and camera parameters

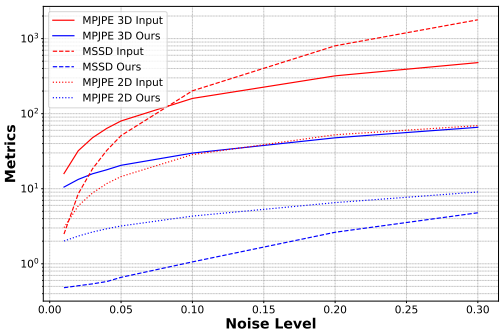


Figure 3: Log-log comparison of MPJPE 3D, MSSD, and MPJPE 2D metrics for input noisy 3D data and our optimized approach across varying noise levels. The red lines represent the input metrics, while the blue lines represent our approach.

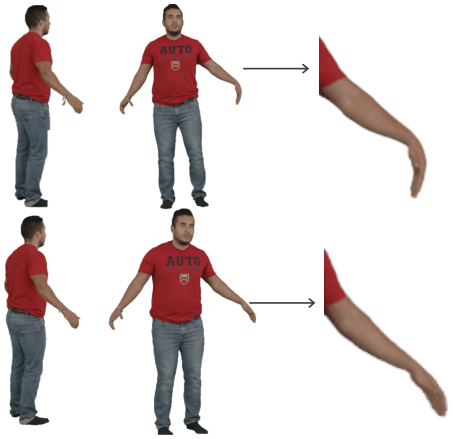

Figure 4: Visual comparison of GART (top) and our approach (bottom) on novel view synthesis, emphasizing geometric correction for hand pose.

Table 2: Performance comparison of methods

| Method | Approach | MPJPE 3D (mm) | MSSD | MPJPE 2D (px) |
|---|---|---|---|---|
| MotionBert | Original | 39.2 | 0.43 | 3.2 |
| | Optimized (ours) | **35.98** | **0.52** | **2.56** |
| 3DLFM | Original | 44.5 | 0.69 | 3.8 |
| | Optimized (ours) | **37.02** | **0.47** | **3.07** |

crucial for template-free deformable modeling applications. Furthermore, as illustrated in Fig. 2(b), our method effectively handles the flipping noise introduced in the input by taking advantage of smoothness constraints and refining it further by leveraging additional source of information from 2D reprojection errors (via PnP). This ensures that even with noisy 3D input, our method can perform runtime optimization as a post-process on existing 3D data from 3D lifters to achieve better overall performance and ensure high-quality 3D poses and camera parameters, which are crucial for our template-free deformable modeling application.

**Real-World Lifter Experiments** We evaluate the performance on real-world 3D lifters using the same metrics and sequences from the Human3.6M dataset. In this setup, 2D poses were obtained from off-the-shelf 2D detectors [21]. Our findings confirm that our optimization step not only boosts the performance of spatial-only lifters, allowing them to outperform temporal baselines by enforcing temporal smoothness, 2D reprojection loss, and kinematic chain constraints, but also enhances the performance of spatiotemporal lifters through PnP refinement and bone length consistency.

### 4.2 Challenging the Need for 3D Surface Templates

Our approach eliminates the dependency on predefined 3D surface templates like SMPL [15] and SMAL [3] by leveraging a novel initialization technique for Gaussian parameters and skinning weights, based on simple geometric methods (see Sec. 3.2). This allows us to handle a diverse range of object categories, including those previously unattainable with template-based methods.

**Why do we need 3D-surface templates in the first place**? The major contribution of these templates [15, 3] is the initialization of Gaussian points for Gaussian splatting and initial skinning weights. We propose that even with coarse approximations derived purely from geometric sense, the end-to-end differentiable rendering pipeline is powerful enough to learn accurate models without a dedicated rig. This observation is pivotal, enabling us to create general, deformable, and animatable

Table 3: Challenging the need for 3D surface template models. Models highlighted with red rely on parametric templates.

| Method | **ZJU-Mocap** [18] | | **People** [1] | | **DogsShow** [13] | | **Kangaroo** | | **Drosophilia** [10] | |
|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR (dB) ↑ | SSIM ↑ | PSNR (dB) ↑ | SSIM ↑ | PSNR (dB) ↑ | SSIM ↑ | PSNR (dB) ↑ | SSIM ↑ | PSNR (dB) ↑ | SSIM ↑ |
| InsAvat-Dog [13] | - | - | - | - | 18.37 | 0.79 | - | - | - | - |
| Instant-NVR [9] | 31.01 | 0.971 | - | - | - | - | - | - | - | - |
| InstantAvatar [12] | - | - | 29.65 | 0.973 | - | - | - | - | - | - |
| GART [13] | **31.76** | 0.976 | 30.4 | 0.976 | **21.18** | **0.87** | - | - | - | - |
| Ours | 31.48 | **0.981** | **30.72** | **0.979** | 21.12 | 0.86 | **35.87** | **0.984** | **31.65** | **0.971** |

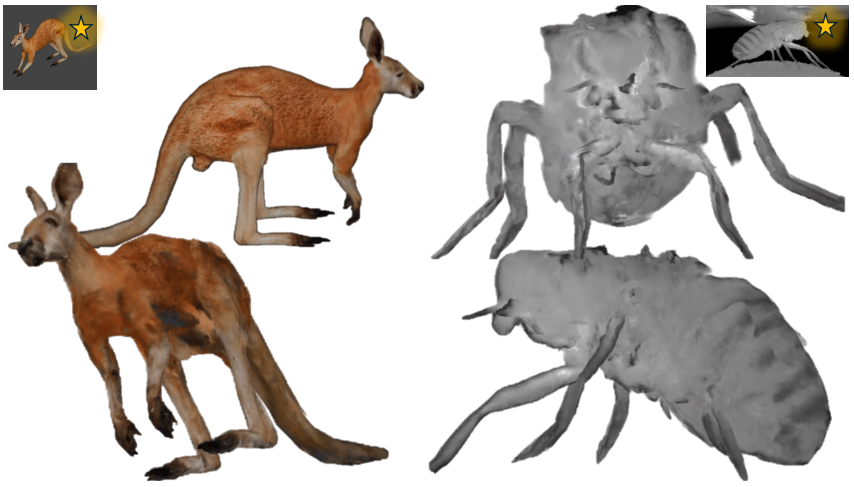

Figure 5: *Novel view synthesis and novel posing for diverse categories [10]*: Original views (stars) are shown in insets, with larger images depicting synthesized views by our method. Direct baseline comparisons are not feasible due to the novelty of our method. Our method does not rely on pre-existing 3D surface templates, making it versatile for 3D asset generation.

rigs, especially for non-humans, non-quadrupeds, as well as inanimate entitites. We show results for kangaroos and drosophila [10], which were not possible before.

To validate this, we conduct ablation-style experiments in this subsection assuming accurate 3D pose and camera information is already available. This setup allows for a fair comparison by isolating the effect of not using 3D surface templates. Using our method, we achieve comparable or superior performance in various metrics, as shown in Table 3. Notably, for the newly introduced datasets of kangaroos and fruit flies (drosophilia), our approach achieves high PSNR and SSIM scores for novel view synthesis (we do not report quantitative numbers for novel pose synthesis) even without predefined 3D surface templates.

### 4.3 Real-world applications

In this subsection, we show versatility of our approach by running our full pipeline on People-snapshot dataset [1] as well as the videos captured in the wild, including AI-generated videos from state-of-the-art models like SORA [5] from OpenAI.

We replicate the experiment performed in the previous subsection on People-snapshot [1] dataset, using 3D poses obtained via our methodology described in Sec. 3. As shown in Figure 4, our optimization step effectively refines initial estimates in the presence of noisy 3D data, leading to high-quality 3D reconstructions. This is in contrast to GART [13], which relies on off-the-shelf SMPL parameter estimators [17] and lacks the constraints introduced by our runtime optimization framework.

We introduce a novel application of detecting and correcting inaccuracies in AI-generated videos [5, 4]. This addresses a unique problem not tackled by current methods, improving the quality and reliability of AI-generated content by correcting geometric and anatomical inconsistencies. Applying our method to AI-generated videos from SORA, we initialize our pipeline with optimized poses and cameras from Sec.3. We do not let gradients pass through the pose parameters during the differentiable

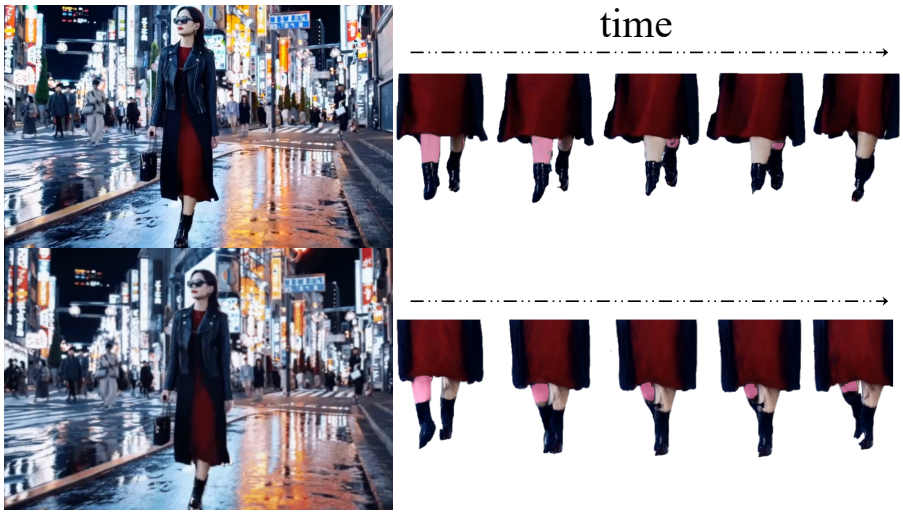


Figure 6: *AI-generated video correction*: Right sequences highlight the right leg's movement over time. The top row shows the original video with the right leg flipping from front to back, revealing flaws in AI-generated videos [5]. The bottom row demonstrates our correction by reconstructing the 3D model, optimizing with proposed approach, and rendering, resulting in smoother, consistent, and physically plausible leg motion.

rendering pipeline to detect glaring geometric inconsistencies. Figure6 illustrates this correction process, with the top row showing the original AI-generated video with noticeable geometric flaws and the bottom row demonstrating the corrected version with physically plausible motion.

## 5 Limitations

While our method provides significant advancements in template-free 3D asset generation, it does have some limitations. Firstly, the reliance on off-the-shelf 2D pose detectors and 3D lifting models means that our results are contingent on the quality of these initial estimates. In scenarios where these models fail to provide accurate initial poses, our optimization may struggle to achieve high fidelity reconstructions. Secondly, although our geometry-based initialization and Gaussian splatting techniques are versatile, they may not fully capture the fine-grained details of highly intricate objects, potentially limiting the resolution of the final 3D models.

## 6 Conclusion

In conclusion, our work presents a novel approach to 3D asset generation that eliminates the dependency on predefined 3D surface templates, thereby broadening the applicability to a wide array of object categories. Through our novel run-time pose optimization and template-free geometric initialization, we demonstrate significant improvements in 3D pose accuracy and overall model fidelity, especially boosting the performance of spatial as well as spatiotemporal 3D lifting models. Moreover, through extensive experiments we show that our method achieves comparable or superior performance to leading template-based techniques, validating its effectiveness in both controlled and in-the-wild environments and can create 3D rig animatable models like those for non-human or non-quadrupeds, previously not possible with as little as 50-100 frames. Furthermore, our unique application of correcting inaccuracies in AI-generated videos highlights the practical relevance and robustness of our approach. Through this work, we challenge some existing notions of how much 3D dense surface models actually help if a differentiable rendering is provided with a rough initialization through geometric-based approaches. This work could open interesting directions in what is possible in dynamic and deformable object reconstruction, and opens new avenues for future research and applications in 3D modeling and animation.

## 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.

[2] 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.

[3] 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.

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[5] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. Technical report, OpenAI, 2024. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

[6] Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. Segment and track anything. *arXiv preprint arXiv:2305.06558*, 2023.

[7] Mosam Dabhi, Chaoyang Wang, Tim Clifford, László Jeni, Ian Fasel, and Simon Lucey. Mbw: Multi-view bootstrapping in the wild. *Advances in neural information processing systems*, 35:3039–3051, 2022.

[8] Mosam Dabhi, Laszlo A Jeni, and Simon Lucey. 3d-lfm: Lifting foundation model. *arXiv preprint arXiv:2312.11894*, 2023.

[9] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794, 2023.

[10] Semih Günel, Helge Rhodin, Daniel Morales, João Campagnolo, Pavan Ramdya, and Pascal Fua. Deepfly3d, a deep learning-based approach for 3d limb and appendage tracking in tethered, adult drosophila. *Elife*, 8:e48571, 2019.

[11] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.

[12] 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.

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

[14] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3383–3393, 2021.

[15] 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. Springer, 2023.

[16] David Novotny, Nikhila Ravi, Benjamin Graham, Natalia Neverova, and Andrea Vedaldi. C3dpo: Canonical 3d pose networks for non-rigid structure from motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7688–7697, 2019.

[17] 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.

[18] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021.

[19] John Von Neumann. Distribution of the ratio of the mean square successive difference to the variance. *The Annals of Mathematical Statistics*, 12(4):367–395, 1941.

[20] Chaoyang Wang and Simon Lucey. Paul: Procrustean autoencoder for unsupervised lifting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 434–443, 2021.

[21] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose++: Vision transformer for generic body pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[22] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15085–15099, 2023.