MorphGS: Morphology-Adaptive Articulated 3D Motion Transfer from Videos

Taeyeon Kim*, Youngju Na*, Jumin Lee, Sebin Lee, Minhyuk Sung, and Sung-Eui Yoon†

KAIST, Republic of Korea
{tykim5931, yjna2907, jmlee, seb.lee, mhsung, sungeui}@kaist.ac.kr

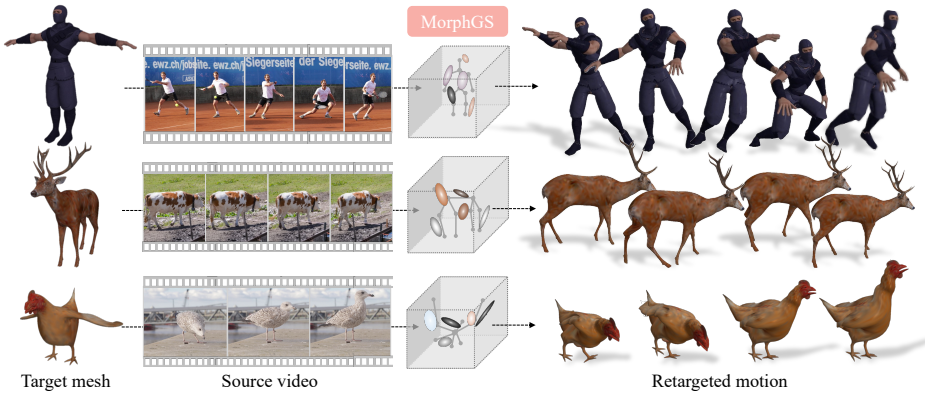


Fig. 1: Given a monocular video and a rigged 3D target character, MorphGS directly optimizes target morphology and pose parameters to reproduce the observed motion without explicit source 3D reconstruction or category-specific parametric templates.

Abstract. Transferring articulated motion from monocular videos to rigged 3D characters is challenging due to pose ambiguity in 2D observations and morphological differences between source and target. Existing approaches often follow a reconstruct-then-retarget paradigm, tying transfer quality to intermediate 3D reconstruction and limiting applicability to categories with parametric templates. We propose MorphGS, a framework that formulates motion retargeting as a target-driven analysis-by-synthesis problem, directly optimizing target morphology and pose through image-space supervision. A rig-coupled morphology parameterization factorizes character identity from time-varying joint rotations, while dense 2D-3D correspondences and synthesized views provide complementary structural and multi-view guidance. Experiments on synthetic benchmarks and real-world videos show consistent improvements over baselines. Project page: https://xodus777.github.io/MorphGS/

*Equal contribution †Corresponding author

1 Introduction

Animating 3D characters from videos has long been a goal in vision and graphics. At its core, the task requires retargeting articulated motion observed in a monocular video to a rigged 3D target character. Despite progress in video-based 3D motion reconstruction, robust motion retargeting from monocular videos remains challenging to deploy broadly, as source videos and target characters often differ substantially in morphology and appearance [1, 41].

A common paradigm follows a *reconstruct-then-retarget* pipeline [8,28,30,43, 56]: first reconstruct an intermediate 3D representation from the source video, then transfer the recovered motion to a target character. These approaches often rely on parametric templates [26, 62], category-specific models [15, 48, 52], or large-scale training data [23], limiting their applicability to well-studied categories such as humans and quadrupeds. Furthermore, the cascaded nature of the pipeline causes errors to propagate into the retargeting stage, where they are further amplified by morphological mismatches between the source and target.

We propose **MorphGS**, a video-to-3D motion transfer framework (Fig. 1) that casts retargeting as a *target-driven analysis-by-synthesis* problem [5,20,42]. Instead of reconstructing the source in 3D, we directly optimize the target character's morphology and motion so that its differentiable rendering [17] reproduces the source observations, while its structural identity remains anchored to the target. This avoids both failure modes of the cascade, as no intermediate reconstruction propagates errors under morphological mismatch, and shape changes are grounded in the target rig rather than in category-specific priors.

The key to making this target-driven optimization tractable is to prevent pose updates from entangling with shape changes. MorphGS introduces an explicit morphology parameterization that factorizes time-invariant character identity from time-varying motion. Concretely, we model shape with structured morphology parameters—bone lengths, scale, and local rest-pose offsets anchored to the rig—while optimizing per-frame pose parameters via image-space supervision. For robust structural guidance under monocular ambiguity, we combine two complementary signals. Dense 2D-3D semantic correspondences supply part-level anchors, while diffusion-based novel-view synthesis adds multi-view supervision.

Together, these components let MorphGS retarget motion without source reconstruction, motion priors, or category-specific training, while a rig-coupled parameterization keeps morphology and pose identifiable up to a global scale. We validate this design across synthetic and real-world benchmarks, showing that target-driven morphology-and-pose optimization provides an effective alternative to reconstruct-then-retarget pipelines.

2 Related Work

Our study addresses video-to-3D motion transfer, where the goal is to recover a pose sequence that animates a rigged 3D target mesh to reproduce articulated motion observed in videos. This task builds upon pose estimation from 2D obser-

vations and 3D motion transfer. We review each area below, including existing 2D-to-3D motion transfer, and position our approach within this landscape.

Pose and shape estimation from videos Recovering 3D pose from video often relies on category-specific parametric templates such as SMPL [26] for humans [11, 57] and SMAL [62] for quadrupeds [27, 35]. These methods achieve strong results but remain confined to categories with dense 3D annotations and parametric models. Template-free methods [3, 23, 47, 48, 52] predict 3D shape and pose directly from image collections, yet their category-specific training limits generalization beyond seen categories and does not extend to cross-character retargeting. Recent 4D reconstruction enables per-scene articulated modeling from monocular or sparse observations, leveraging representations such as 3D Gaussian splatting [13, 14, 17, 46, 54] and mesh-Gaussian hybrids [22, 49], while others animate static meshes via automatic rigging or video diffusion priors [38, 45]. However, these methods remain subject-specific, as the recovered motion is tied to the source identity and does not directly transfer to other characters.

3D-to-3D motion transfer While the above methods focus on recovering motion for a single subject, motion transfer addresses re-applying articulation across different characters. Skeletal motion transfer methods operate on explicit skeleton representations [1, 7, 10, 40, 41], yielding robust results when two characters share compatible rigs. Skeleton-free deformation methods [9, 24, 30, 43, 44, 55] relax the need for structural alignment, but both lines still require high-quality 3D motion sequences (*e.g.*, skeletal trajectories or mesh animations), which remain scarce beyond well-studied categories such as humans.

2D-to-3D motion transfer Video-driven 3D motion transfer combines pose estimation and retargeting into a single pipeline. In the human and quadruped domains, parametric templates [26, 62] are paired with pose estimators [34, 57] and 3D transfer techniques [4, 30, 43], but remain inherently limited to categories that such templates support. Meanwhile, category-agnostic methods lift this restriction by first reconstructing source 3D motion with a generic prior—RGB-D videos [28], novel view synthesis [8], keypoint annotation [31], or image-to-3D generative models [12, 56]—and then transferring it to the target via algorithmic rigging or learned deformation. Despite their generality, these approaches follow a *reconstruct-then-retarget* paradigm whose performance is tied to the intermediate source reconstruction. While recent works [12, 31] reduce the need for dense reconstruction by learning deformation transfer from sparse 3D keypoints or skeletal joints, their generalization still depends on the shapes and motions seen during training.

In contrast, our method casts video-to-3D retargeting as a target-driven analysis-by-synthesis problem, directly optimizing the target's morphology and pose through image-space supervision via differentiable Gaussian rendering. A skeletal-anchored morphology parameterization accounts for the shape differences between source and target while preserving the target's rig, enabling trans-

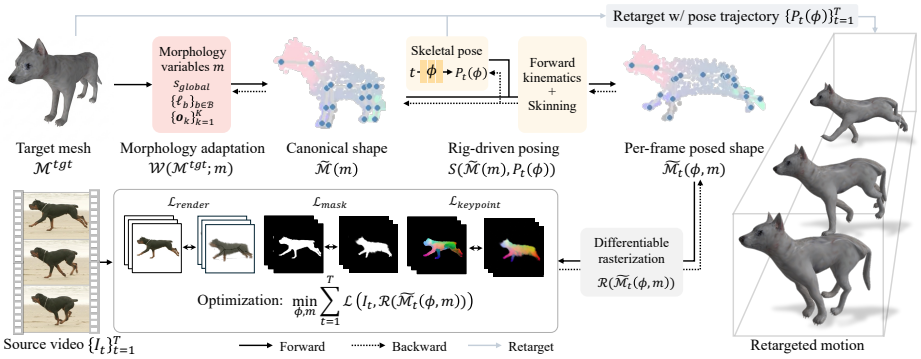


Fig. 2: Overview of MorphGS. Given a source video and a target character, MorphGS first represents the target as a morphology-adaptive articulated 3D mesh (Sec. 3.2). 3D Gaussian representation is wrapped to deform pose and render differentiably (Sec. 3.3). The morphology and pose parameters are jointly optimized through image-space supervision (Sec. 3.4). The recovered motion is then retargeted to the target character, yielding an animation that closely follows the source video.

fer between morphologically different characters without parametric templates, motion priors, or explicit 3D supervision.

3 Methods

3.1 Problem Formulation

Consider a source monocular RGB video $\{I_t \in \mathbb{R}^{H\times W\times 3}\}_{t=1}^T$ depicting an articulated subject. Given a rigged target mesh $\mathcal{M}^{\text{tgt}}$ whose morphology may differ from the source, our goal is to estimate a temporally coherent pose trajectory $\{P_t\}_{t=1}^T$ that retargets the observed motion onto $\mathcal{M}^{\text{tgt}}$ while preserving the target's structural identity (i.e., skeleton topology and mesh connectivity).

Target-driven analysis-by-synthesis We formulate video-to-3D motion retargeting as a target-driven analysis-by-synthesis problem. Instead of reconstructing the source in 3D and transferring motion afterward, we directly optimize target-specific parameters so that the rendering of the animated target closely reproduces the input frames. At each time step t, the posed shape $\tilde{\mathcal{M}}_t$ is obtained in two stages: *(i) morphology adaptation* in the target's canonical space, which adjusts limb proportions, global scale, and local geometry to bridge the shape gap between the source and target, followed by *(ii) rig-driven posing* via forward kinematics and linear blend skinning:

$$\tilde{\mathcal{M}}(m) = \mathcal{W}(\mathcal{M}^{\text{tgt}}; m), \qquad \tilde{\mathcal{M}}_t(\phi, m) = \mathcal{S}\big(\tilde{\mathcal{M}}(m),\, P_t(\phi)\big), \tag{1}$$

where $\mathcal{W}(\cdot\,; m)$ is the morphology adaptation function parameterized by m, producing the adjusted canonical shape $\tilde{\mathcal{M}}$, and $\mathcal{S}$ denotes the skinning operator

that applies skeletal pose $P_t(\phi)$ to $\tilde{\mathcal{M}}$, yielding the per-frame posed shape $\tilde{\mathcal{M}}_t$. Here ϕ denotes the parameters of a temporally conditioned pose network.

A differentiable renderer $\mathcal{R}$ maps the posed shape $\tilde{\mathcal{M}}_t$ to a predicted frame $\hat{I}_t = \mathcal{R}(\tilde{\mathcal{M}}_t)$, where we adopt the 3D Gaussian rasterizer [17] for efficient rendering. We estimate target-side parameters (ϕ, m) by minimizing objectives in the projection space over the foreground mask Ω_t:

$$\min_{\phi,\, m} \sum_{t=1}^{T} \mathcal{L}\left(I_t,\ \mathcal{R}\big(\tilde{\mathcal{M}}_t(\phi, m)\big);\ \Omega_t\right). \tag{2}$$

Once optimization is complete, only the recovered pose trajectory $\{P_t(\phi)\}_{t=1}^{T}$ is applied back to the target rig $\mathcal{M}^{\text{tgt}}$, preserving its original geometry. Note that we fix intrinsics and do not estimate per-frame extrinsics in our renderer. Global source–target alignment in each frame is approximated by a per-frame root transform in $SE(3)$.

Building on the above formulation, we first introduce a morphology-adaptive parameterization for articulated 3D modeling (Sec. 3.2), then describe pose-conditioned Gaussian rendering (Sec. 3.3), and finally present the full objective and optimization procedure (Sec. 3.4). The overall pipeline is illustrated in Fig. 2.

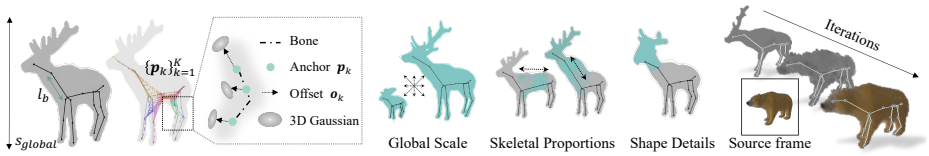
3.2 Morphology-Parameterized Articulated 3D Modeling

The goal of the morphology adaptation function $\mathcal{W}$ is to account for time-invariant shape differences between the source subject and the target character (*e.g.*, differences in limb proportions, body size, or local geometry), so that the per-frame source observations can be explained primarily by the pose trajectory $\{P_t(\phi)\}_{t=1}^{T}$ through the skinning operator $\mathcal{S}$ (Eq. (1)).

However, jointly optimizing shape and pose from monocular observations is challenging due to the well-known shape-pose ambiguity, where shape changes can compensate for pose updates in the image plane [16, 50]. If the canonical shape is optimized without structural and temporal constraints, it can absorb per-frame pose errors and lead to geometric drift. We mitigate this by restricting all shape changes to structured *time-invariant* morphology parameters coupled to the target rig, thereby reducing the degrees of freedom that can explain the same 2D observations. Accordingly, we parameterize morphology as:

$$m \;=\; \Big(s_{\text{global}},\ \{\ell_b\}_{b\in\mathcal{B}},\ \{\mathbf{o}_k\}_{k=1}^{K}\Big), \tag{3}$$

where $s_{\text{global}} \in \mathbb{R}_+$ resolves monocular depth–scale ambiguity, $\ell_b \in \mathbb{R}_+$ are learnable bone lengths controlling skeletal proportions, and $\mathbf{o}_k \in \mathbb{R}^3$ are local geometric offsets that capture residual shape details (*e.g.*, limb thickness, silhouette) beyond what the skeleton alone can express, as illustrated in Fig. 3. We assume the target mesh $\mathcal{M}^{\text{tgt}}$ is rigged with a fixed kinematic tree and skinning weights $\{w_{kj}\}$. For unrigged targets (*e.g.*, generated mesh), we obtain these using off-the-shelf auto-rigging methods [4, 58].



(a) Learnable morphology parameterization (b) Deformation effects of morphology parameters (c) Morphology adaptation

Fig. 3: Morphology parameterization. (a) Skeleton anchor $\mathbf{p}_k$, local offset $\mathbf{o}_k$, and 3D Gaussian primitives on the target rig. (b) Distinct deformation induced by global scale, bone lengths, and offset Gaussians. (c) An example of morphology adaptation.

Learnable bone lengths Let $\mathcal{J}$ denote the set of joints with root joint j_{root}. For each bone $b \in \mathcal{B}$, $\mathbf{d}_b$ denotes the unit direction from its parent to child joint. We define the rest-pose joint locations $\mathbf{j}_{\text{rest}} : \mathcal{J} \to \mathbb{R}^3$ as:

$$\mathbf{j}_{\text{rest}}(j;\ell) = \mathbf{j}_{\text{rest}}(j_{\text{root}}) + \sum_{b \in \mathcal{P}(j_{\text{root}},\, j)} \ell_b\, \mathbf{d}_b, \tag{4}$$

where $\mathcal{P}(j_{\text{root}}, j)$ denotes the kinematic-chain path from the root to joint j. The kinematic tree $\mathcal{B}$ and directions $\mathbf{d}_b$ are initialized from the target rig and kept fixed, while only the bone lengths ℓ_b are optimized. This preserves the skeletal topology while allowing bone proportions to adapt via ℓ_b for all $b \in \mathcal{B}$.

Canonical target shape Given the adjusted skeleton, we define each canonical vertex position relative to the rig. Specifically, we compute a skeleton-dependent anchor $\mathbf{p}_k$ as the skinning-weighted average of the rest joints:

$$\mathbf{p}_k = \sum_{j \in \mathcal{J}} w_{kj}\, \mathbf{j}_{\text{rest}}(j;\ell). \tag{5}$$

With normalized skinning weights ($w_{kj} \geq 0$, $\sum_j w_{kj} = 1$), $\mathbf{p}_k$ is a convex combination of joint locations and therefore lies in the convex hull of $\{\mathbf{j}_{\text{rest}}(j)\}$. We then model the canonical vertex as a residual around this anchor, $\mathbf{v}_k = \mathbf{p}_k + \mathbf{o}_k$, where $\mathbf{o}_k \in \mathbb{R}^3$ is a skeleton-anchored rest-space offset relative to $\mathbf{p}_k$ that captures local residual geometry (*e.g.*, thickness, silhouette) beyond skeletal proportions.

Finally, we apply a uniform global scale $s_{\text{global}} \in \mathbb{R}_+$ to the canonical shape, including both the skeleton and surface geometry, to resolve the monocular depth-scale ambiguity:

$$\bar{\mathbf{v}}_k = s_{\text{global}}\, \mathbf{v}_k. \tag{6}$$

Discussion Theoretically, recovering both 3D shape and motion from monocular 2D trajectories with unconstrained non-rigid structure is ill-posed [50]. Under weak-perspective projection, our formulation instead imposes a piecewise-rigid kinematic model in which bone lengths $\{\ell_b\}$ and local rest-space offsets $\{o_k\}$ are strictly time-invariant [19]. This factorization confines temporal variation to articulated joint rotations, which discourages morphology parameters from compensating for pose errors and reduces shape–pose entanglement. With sufficiently

non-degenerate motion, the resulting kinematic constraints make morphology and pose identifiable up to a global scale factor. While real videos often deviate from these ideal assumptions (e.g., perspective effects and soft-tissue deformation), the analysis provides a useful explanation for the empirical stability of our target-driven optimization. We include a proof sketch and further discussion in the appendix.

3.3 Pose-conditioned Gaussian Rendering

Our target-driven formulation jointly optimizes both the morphology parameters m and the pose parameters ϕ through image-space gradients. We realize this by representing the morphology-adapted canonical shape as an articulated 3D Gaussian set, posing it through forward kinematics and Linear Blend Skinning (LBS), and rendering via 3D Gaussian splatting [17].

Articulated 3D Gaussians We represent the target mesh with K 3D Gaussian primitives, initialized from the mesh vertices in the rest pose so that each Gaussian inherits its vertex's skinning weights $\{w_{kj}\}_{j\in\mathcal{J}}$ directly from the rig. Each primitive k carries a canonical center $\bar{\mathbf{v}}_k \in \mathbb{R}^3$ (Eq. 6) together with the standard splatting attributes: anisotropic scale, rotation vector, opacity, and view-dependent color [17]. This initialization ensures that skeletal deformations and surface geometry remain tightly coupled, preventing the Gaussians from drifting away from the underlying rig during optimization.

Pose-conditioned deformation To encourage temporal coherence, we parameterize the pose trajectory with a single time-conditioned network rather than optimizing independent per-frame pose:

$$P_t(\phi) = \big(\{\boldsymbol{\theta}_j^t\}_{j\in\mathcal{J}},\, \boldsymbol{\delta}_{\text{root}}^t\big) = f_\phi(\text{emb}(t))\,, \tag{7}$$

where $\text{emb}(t)$ is a sinusoidal time embedding [29,39] and f_ϕ is a lightweight MLP that outputs per-joint relative rotations $\boldsymbol{\theta}_j^t$ (axis-angle) and root translation $\boldsymbol{\delta}_{\text{root}}^t$. Forward kinematics converts the local rotations into global joint transforms $\{\mathbf{T}_j^t \in \text{SE}(3)\}_{j\in\mathcal{J}}$, which then deform the canonical centers via LBS:

$$\mathbf{v}_k^t = \sum_{j\in\mathcal{J}} w_{kj}\, \mathbf{T}_j^t\, \bar{\mathbf{v}}_k. \tag{8}$$

Rasterization The posed Gaussians are projected onto the image plane and composited via depth-sorted alpha blending [17], producing a rendered frame $\hat{I}_t = \mathcal{R}(\tilde{\mathcal{M}}_t(\phi, m))$. This closes the differentiable chain from (m, ϕ) to the pixel-level objective in Eq. 2, enabling joint morphology–pose optimization through image-space supervision alone.

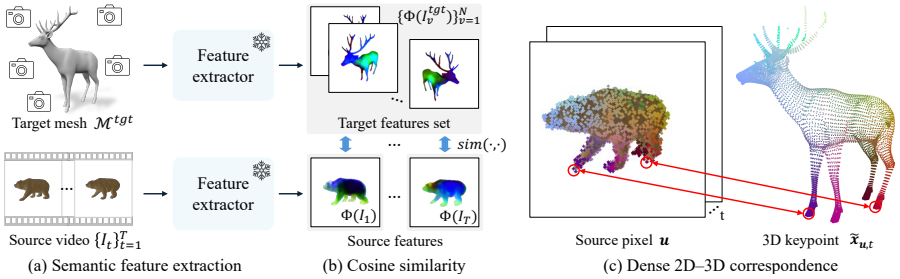


Fig. 4: Dense 2D–3D correspondence matching. (a) Extracting semantic features from source frames and rendered target views. (b) Calculating similarity between them. (c) 3D keypoint mapping corresponding to the source pixel.

3.4 Differentiable Optimization Framework

During optimization, the Gaussian representation serves as a differentiable proxy, enabling image-space gradients to flow back to both shape and pose through the rendering pipeline. Unlike prior pipelines that assume aligned source–target [13, 36,54] or require canonical pose initialization [38], our approach jointly optimizes morphology and pose from a single monocular video without such assumptions. Our objective combines rendering losses, keypoint loss, and regularization:

$$\mathcal{L}_t = \underbrace{\mathcal{L}_{\text{rgb}} + \mathcal{L}_{\text{mask}}}_{\mathcal{L}_{\text{render}}} + \mathcal{L}_{\text{keyp}} + \mathcal{L}_{\text{mv}} + \mathcal{L}_{\text{reg}}. \tag{9}$$

Rendering losses The rendering terms provide the primary gradient signal for motion and appearance alignment by directly comparing the rendered target $\hat{I}_t$ with the source frame I_t in the image plane. Specifically, $\mathcal{L}_{\text{rgb}}$ enforces photometric consistency within the foreground region, and $\mathcal{L}_{\text{mask}}$ aligns the rendered alpha mask $\hat{\Omega}_t$ with a foreground segmentation Ω_t to stabilize global outline.

Keypoint loss Photometric and mask supervision provide robust appearance and shape signals, but can be insufficient early in optimization or under large source–target morphological differences, where pixel-level losses alone struggle to establish reliable part-to-part correspondences. To provide explicit structural guidance, we establish dense 2D–3D correspondences between the source frame and the target mesh and use them as keypoint supervision.

Dense 2D–3D correspondence. The detailed pipeline of our dense correspondence extraction module is illustrated in Fig. 4. We first extract dense semantic features from the source video frames $\{I_t\}_{t=1}^T$ and from the rendered images of the target mesh across N views, $\{I_v^{\text{tgt}}\}_{v=1}^N$, using a geometry-aware feature extractor $\Phi(\cdot)$ [59]. For each frame t, we define $\mathbf{u} \in \mathbb{R}^2$ as a foreground pixel in the source frame I_t, and $\mathbf{x} \in \mathbb{R}^3$ as a vertex from the target mesh vertices $\mathcal{V}(\mathcal{M}^{\text{tgt}})$. Following SHIC [37], we compute the pooled similarity score between $\mathbf{u}$ and $\mathbf{x}$:

$$\Sigma_t(\mathbf{u}, \mathbf{x}) = \underset{v,\ \mathbf{x} \in \mathcal{V}(\mathcal{M}^{\text{tgt}})}{\text{pool}} \ \text{sim}\left(\Phi(I_t)[\mathbf{u}],\ \Phi(I_v^{\text{tgt}})\left[\pi_v(\mathbf{x})\right]\right), \tag{10}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity, $\pi_v(\mathbf{x})$ projects $\mathbf{x}$ onto rendered view I_v^{tgt}, and pool operator aggregates scores across all N rendered views. Each source pixel is then assigned to its best-matching 3D keypoint $\tilde{\mathbf{x}}_{\mathbf{u},t}$:

$$\tilde{\mathbf{x}}_{\mathbf{u},t} = \arg \max_{\mathbf{x} \in \mathcal{V}(\mathcal{M}^{\mathrm{tgt}})} \Sigma_t(\mathbf{u}, \mathbf{x}). \tag{11}$$

Keypoint supervision. Let $\mathcal{P}_t$ denote the set of sampled foreground pixels in frame t. For each frame t, we select source pixels $\mathbf{u}_i \in \mathcal{P}_t$, whose similarity scores exceed a confidence threshold, and penalize the reprojection error between their observed locations and the projected positions of the matched 3D keypoints:

$$\mathcal{L}_{\mathrm{keyp}} = \frac{1}{\sum_t |\mathcal{P}_t|} \sum_t \sum_{\mathbf{u}_i \in \mathcal{P}_t} \rho\left(\mathbf{u}_i - \pi\left(\tilde{\mathbf{x}}_{\mathbf{u}_i,t}\right)\right), \tag{12}$$

where $\pi(\cdot)$ denotes a fixed camera projection, and $\rho(\cdot)$ is a smooth-ℓ_1 penalty.

Synthesized-view guidance Monocular input leaves occluded body parts under-constrained, making pose ambiguous under self-occlusion. To provide complementary viewpoints, we leverage a diffusion-based novel-view synthesis module [53] to generate temporally consistent pseudo-views of the source sequence. Our target-driven formulation naturally accommodates multi-view supervision, as the synthesized views simply provide additional image-space objectives over the same target parameters. We extend the rendering losses to all viewpoints $\mathcal{C}$:

$$\mathcal{L}_{\mathrm{mv}} = \sum_{t=1}^{T} \sum_{c \in \mathcal{C}} \mathcal{L}_{\mathrm{render},t}^{(c)}. \tag{13}$$

This additional supervision regularizes pose estimation under self-occlusion, where the monocular input alone provides insufficient constraints.

Regularizations Finally, $\mathcal{L}_{\mathrm{reg}}$ enforces frame-to-frame temporal motion smoothness. Additional details on regularizations are provided in the Appendix.

4 Experiments

4.1 Implementation details

We employ a staged optimization strategy that first resolves global alignment (root translation, root orientation, s_{global}), then we enable updating ℓ_b and local joint rotation estimation from 500 to 1.5K iterations. Lastly, from 1.5K iterations, we jointly optimize all parameters, including Gaussian splatting attributes (offsets $\mathbf{o}_k$, color, scale, rotation). All experiments are run on a single NVIDIA RTX 4090 GPU. For a target mesh with 10K vertices, the optimization process takes approximately 5 minutes for 5K iterations using Adam [18] with adaptive learning rates. Further details are provided in the Appendix.

Table 1: Quantitative evaluation on Mixamo and DT4D datasets. Our method consistently outperforms all baselines across evaluated datasets on average. Results are averaged across scenes, with per-scene results in the appendix. PMD ($\times 10^3$, $\downarrow$) measures geometric accuracy and FVMD ($\times 10^{-3}$, $\downarrow$) measures perceptual motion fidelity.

	Mixamo-Humanoids		DT4D-Quadrupeds		DT4D-Others	
Method	PMD	FVMD	PMD	FVMD	PMD	FVMD
SPT$^+$	2.96	12.56	–	–	–	–
NPR$^+$	3.98	16.67	3.28	16.32	–	–
Transfer4D	7.65	16.16	4.67	15.34	7.16	17.52
Pinocchio$^+$	4.55	14.03	5.56	15.24	7.59	14.23
Ours	**1.91**	**8.82**	**1.33**	**8.81**	**1.95**	**12.83**

4.2 Experimental Setup

Datasets We evaluate MorphGS on both real-world and synthetic benchmarks to assess motion transfer performance across different categories. For synthetic evaluation, we test on Mixamo [2] and DeformingThings-4D (DT4D) [21], following prior motion-transfer protocols [24, 55].

In the Mixamo dataset, we utilize 12 humanoid source–target pairs with different motion types including diverse human motions (*e.g.*, jumping, running, and dancing), where rigged target meshes are directly available from the dataset. For DT4D, we include 20 animal source–target pairs with paired posed mesh sequences for the same motion, split into two groups by skeleton topology: *DT4D-Quadrupeds* (four-legged animals), and *DT4D-Others*, which includes 5 categories beyond the standard quadruped skeleton topology. Since DT4D does not provide rigged meshes, we obtain target rigs using off-the-shelf auto-rigging tools [4, 58]. To generate source videos, we animate the source mesh with the ground-truth pose sequence and render a monocular video from a fixed viewpoint using the differentiable mesh renderer in PyTorch3D [33], yielding 1,505 source–target mesh pairs in total across the three splits.

For real-world evaluation, we curate 8 object-centric clips from DAVIS [32], comprising 4 human and 4 animal sequences. These clips span challenging capture conditions (e.g., motion blur, occlusion) and varying motion complexity (e.g., fast or non-rigid motion), as illustrated in Fig. 6. We use the provided object masks when available and otherwise extract them with SAM3 [6], cropping each frame to the mask region. Extended real-world evaluations on 16 additional clips from SoloDance [51] and AiM [60] are provided in the appendix.

Evaluation metrics For synthetic datasets, we measure Point-wise Mesh Distance (PMD) [44,61] across all frames to evaluate motion transfer accuracy as the average per-vertex ℓ_2 distance between the retargeted and ground-truth meshes following [24, 43, 55]. For both synthetic and real datasets, we measure Fréchet Video Motion Distance (FVMD) [25], which measures perceptual motion plausibility by comparing distributions of velocity and acceleration features extracted from keypoint trajectories in rendered videos.

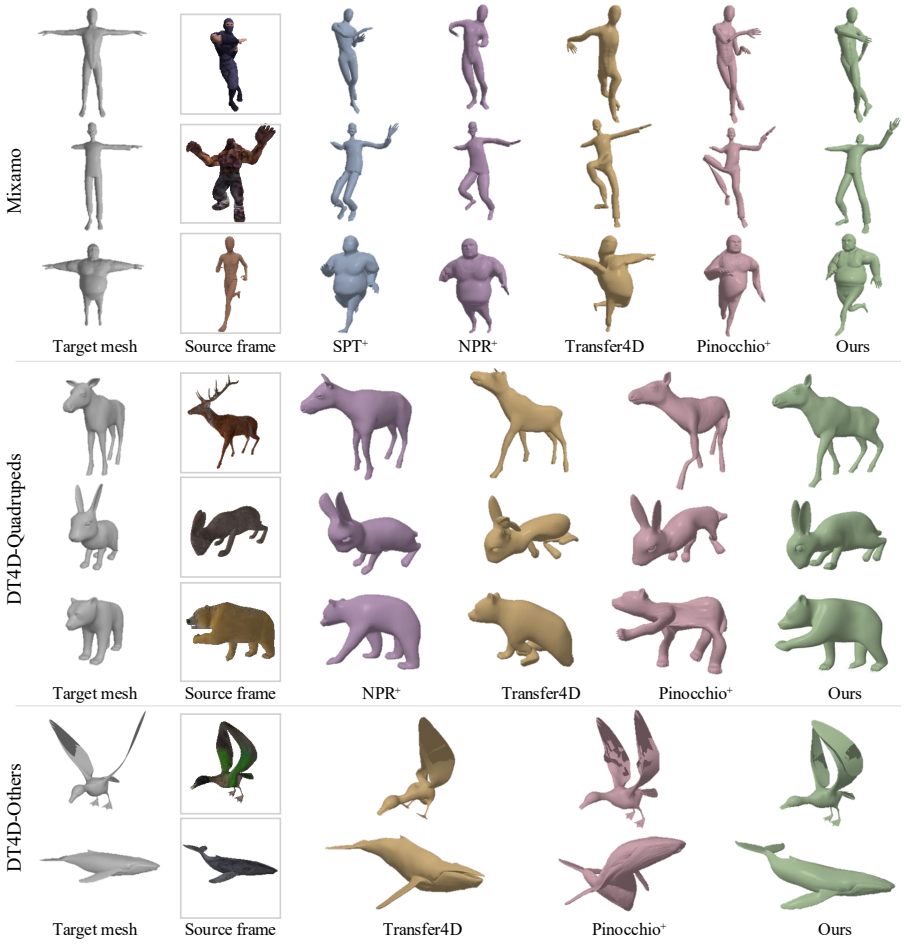


Fig. 5: Qualitative results on Mixamo and DT4D datasets. Our method shows superior pose alignment compared to baselines across diverse objects. Refer to the supplementary video for full animations.

Baselines We compare against baselines from two groups. (i) The first follows a *reconstruct-then-retarget* paradigm, where a 3D mesh sequence is first reconstructed from the source RGB video using off-the-shelf parametric template pose estimators [11, 27], and then retargeted to the target character via SPT [24] (denoted SPT$^+$), NPR [55] (denoted NPR$^+$), or Pinocchio [4] (denoted Pinocchio$^+$). (ii) The second is Transfer4D [28], which estimates a motion skeleton from RGB-D input and transfers it to the target via automatic rigging.

4.3 Motion Transfer Evaluation

We evaluate motion transfer on synthetic benchmarks spanning humanoid and animal targets with different topologies. Tab. 1 summarizes the results.

Table 2: Quantitative evaluation on real-world videos. We use FVMD ($\times 10^{-3}$, $\downarrow$) as the quantitative metric for motion fidelity.

	Human scenes					Animal scenes				
Method	tennis	rollerblade	hockey	snowboard	avg	bear	camel	cows	dog	avg
SPT$^+$	13.95	21.89	5.53	19.30	15.17	–	–	–	–	–
NPR$^+$	21.03	23.95	11.96	29.54	21.62	18.26	17.46	14.62	25.66	19.00
Pinocchio$^+$	16.91	18.63	7.64	18.38	15.39	18.27	13.80	12.58	18.67	15.83
Ours	**13.87**	**13.90**	**5.13**	**18.07**	**12.74**	**13.31**	**12.28**	**6.43**	**17.94**	**12.49**

Humanoid motion transfer comparisons For reconstruct-then-retarget baselines, we first estimate SMPL [26] pose sequences from the source video using an off-the-shelf monocular human pose estimator [11], and then retarget the recovered 3D motion to the target character using SPT [24], NPR [55], and Pinocchio [4], respectively. For Transfer4D [28], which requires RGB-D input, we additionally provide rendered metric depth maps.

As shown in the first column of Tab. 1, MorphGS shows lower PMD and FVMD than the compared baselines, consistent with the qualitative results in Fig. 5. Baseline methods often exhibit incomplete motion transfer (*e.g.*, third row of Fig. 5) or over-smoothed deformations. This is because they rely on learned pose priors or latent manifold constraints, and the retargeting stage may prioritize globally plausible motion over fine-grained vertex-level alignment. In addition, preprocessing targets into manifold mesh representations [24, 55] can remove high-frequency surface details. MorphGS instead optimizes target morphology and pose via differentiable rendering in the target parameter space, which maintains target mesh structure while achieving accurate pose alignment to source motions as in Fig. 5.

Animal motion transfer comparisons We further report animal motion transfer results in the DT4D-Quadrupeds and DT4D-Others columns of Tab. 1. For reconstruct-and-retarget baselines, we use [27] to reconstruct SMAL [62] template pose sequences from the source videos. We omit SPT$^+$ for animals because pretrained weights are not available for non-human shapes. NPR$^+$ is omitted on DT4D-Others since it requires pretrained pose manifolds, which are not provided for non-quadruped animal categories.

For DT4D-Others, which includes categories beyond common parametric templates such as bird or whale, we compare our method against Transfer4D and Pinocchio$^+$. For Pinocchio$^+$ on these categories, we provide a ground-truth source mesh sequence as input, bypassing the 2D video-to-3D reconstruction stage to avoid severe error propagation from an intermediate reconstruction pipeline. As shown in Tab. 1, MorphGS shows lower PMD and FVMD than both baselines, demonstrating effective motion transfer performance on non-standard categories that remain relatively underexplored in the motion retargeting literature.

Real-world motion transfer comparisons We evaluate on the real-world monocular videos curated from DAVIS [32], using them as motion sources. Since

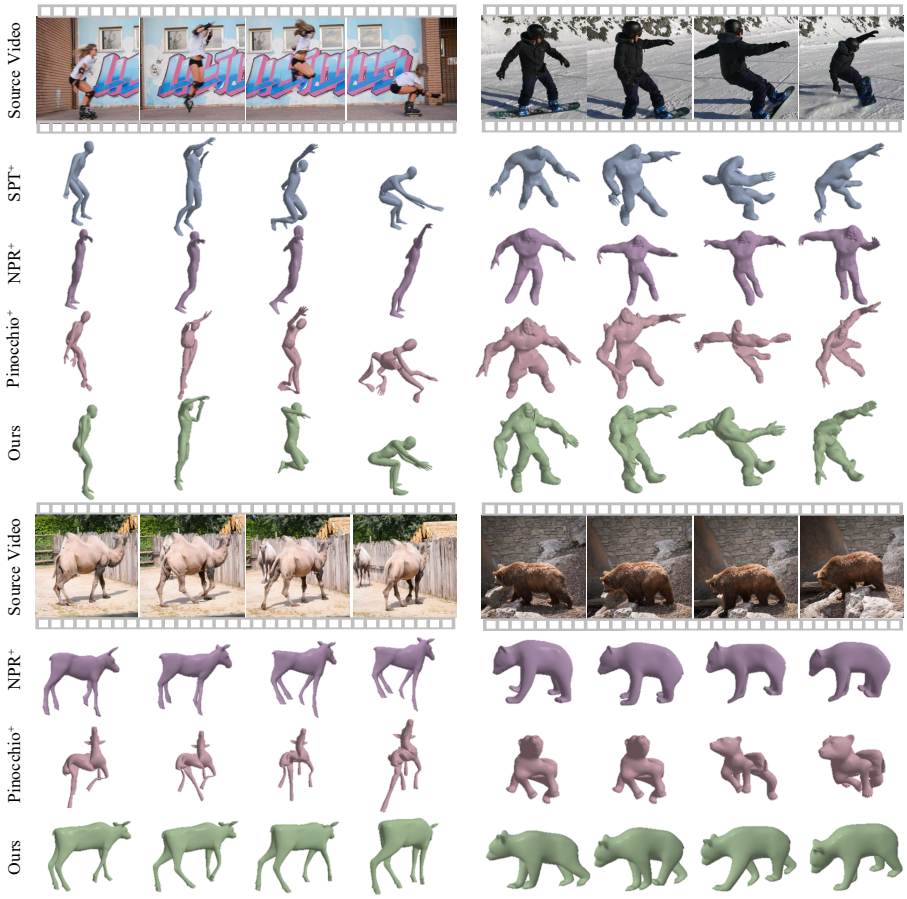


Fig. 6: Qualitative results on real-world videos. We show motion transfer results on real-world videos. Refer to the supplementary video for full animations.

ground-truth 3D mesh sequences are unavailable for such footage, we report FVMD [25] as a perceptual measure of motion consistency rather than direct geometric error, and complement it with qualitative comparisons in Fig. 6 and the supplementary video. We use target meshes from Mixamo [2] for humanoids and DT4D [21] for animals. As summarized in Tab. 2, MorphGS achieves a lower average FVMD than the compared baselines on both human and animal videos. Qualitative results in Fig. 6 show that the posed targets track the motion in the input frames while maintaining the target's original structure. These results indicate that our target-driven, morphology-adaptive optimization is effective on unconstrained monocular videos.

Cross-category motion transfer Figure 7 further demonstrates cross-category transfer, where a single rabbit running motion is retargeted to a fox, a deer, and a quadruped robot with substantially different body proportions and local

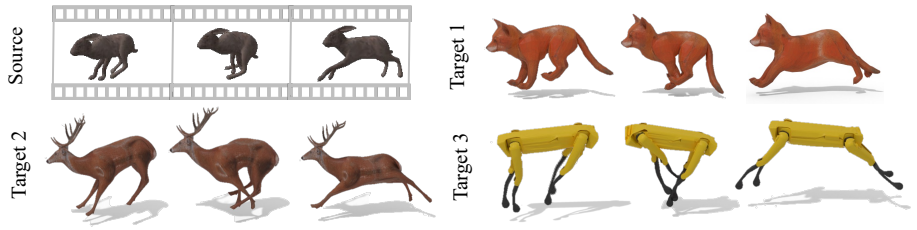


Fig. 7: Cross-category motion transfer. A single rabbit source motion is retargeted onto structurally distinct targets. Each target keeps its own identity while following the source motion. Refer to the supplementary video for full animations.

Table 3: Ablation study. PMD ($\times 10^3$, $\downarrow$) and FVMD ($\times 10^{-3}$, $\downarrow$) are averaged over 12 scenes from the Mixamo [2] dataset.

Model	PMD	FVMD
Morphology parameterization		
Naïve model	8.45	17.54
Fixed morphology	3.72	15.30
+ ℓ_b	2.87	12.45
+ ℓ_b, s_{global}	1.99	9.55
+ ℓ_b, s_{global}, $\mathbf{o}_k$	1.91	8.82
Supervision losses		
$\mathcal{L}_{\text{render}}$	4.52	13.57
+ $\mathcal{L}_{\text{keyp}}$	2.96	11.93
+ $\mathcal{L}_{\text{keyp}}$, $\mathcal{L}_{\text{mv}}$	1.91	8.82

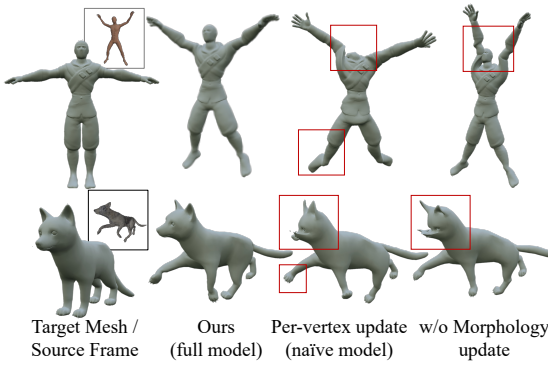


Fig. 8: Effect of morphology parameterization. Skeleton anchoring couples vertices to the rig, preventing structure-violating deformation.

geometries. This factorization of body proportions and local geometry enables MorphGS to reproduce the source running motion across targets while preserving target-specific morphology.

4.4 Ablation Studies

In this section, we evaluate the key components of our framework by ablating the morphology parameterization and supervision signals.

Effects of morphology parameterization We first analyze the effect of our morphology parameterization by comparing it against a *naïve model* that defines each canonical vertex $\mathbf{v}_k \in \mathbb{R}^3$ as a free variable in the global coordinate, rather than as a local offset $\mathbf{o}_k$ relative to the skeleton-dependent anchor $\mathbf{p}_k$ (Eq. 5). Without the anchor $\mathbf{p}_k$ coupling $\mathbf{v}_k$ to the bone lengths ℓ_b, individual vertices can shift independently of the kinematic structure during optimization. As shown in Tab. 3 and Fig. 8, this leads to significant performance degradation (PMD 8.45, FVMD 17.54).

We further ablate individual morphology parameters by progressively enabling them. Without morphology adaptation (fixed morphology), the optimizer compensates for the shape gap through incorrect joint rotations, degrading pose accuracy (PMD 3.72). Enabling bone lengths ℓ_b and global scale s_{global} resolves skeletal proportion and scale mismatches, respectively, yielding substantial gains. Offsets $\mathbf{o}_k$ provide further refinement by capturing residual surface details.

Effects of supervision losses Rendering losses alone recover coarse motion but lack part-level precision, yielding a relatively high PMD of 4.52, since photometric and silhouette cues constrain the projected appearance and outline without resolving part-to-part correspondence. Adding dense 2D–3D keypoint supervision introduces explicit part-wise anchors, leading to a substantial reduction in PMD from 4.52 to 2.96. Synthesized-view guidance further improves robustness under self-occlusion by reducing monocular ambiguity, yielding the best performance with 1.91 PMD and 8.82 FVMD. Additional ablations on supervision losses are provided in the appendix.

5 Conclusion and Future Work

In this study, we have presented MorphGS, a video-to-3D motion transfer framework. By formulating motion retargeting as an inverse-graphics optimization problem, our approach explicitly factorizes morphology and rig-driven pose parameters. This parameterization effectively addresses source-target shape misalignments, enabling motion transfer without relying on predefined parametric templates or intermediate 3D source reconstruction.

Although our morphology adaptation facilitates cross-identity motion transfer, the current framework assumes a compatible articulated topology between the source and target. This assumption inherently limits transfer across subjects with significantly different kinematic structures, such as differences in branch count or joint hierarchy. For future work, integrating physics-based motion priors could provide stronger constraints to resolve ambiguities under severe self-occlusions and reduce temporal artifacts. Moreover, explicitly modeling non-rigid soft-tissue deformations could further improve the expressiveness of the transferred motion. We hope our approach provides a useful step toward more flexible 3D content creation and cross-category motion understanding.

Acknowledgements

Sung-Eui Yoon is the corresponding author. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00208506), and in part by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2025-25443318, Physically-grounded Intelligence: A Dual Competency Approach to Embodied AGI through Constructing and Reasoning in the Real World).

References

1. Aberman, K., Li, P., Lischinski, D., Sorkine-Hornung, O., Cohen-Or, D., Chen, B.: Skeleton-aware networks for deep motion retargeting. ACM Transactions on Graphics (TOG) **39**(4), 62–1 (2020)
2. Adobe: Mixamo. `https://www.mixamo.com`, accessed: 2026-03-04
3. 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)
4. Baran, I., Popović, J.: Automatic rigging and animation of 3d characters. ACM Transactions on graphics (TOG) **26**(3), 72–es (2007)
5. Beker, D., Kato, H., Morariu, M.A., Ando, T., Matsuoka, T., Kehl, W., Gaidon, A.: Monocular differentiable rendering for self-supervised 3d object detection. In: European conference on computer vision. pp. 514–529. Springer (2020)
6. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
7. Chen, J., Li, C., Lee, G.H.: Weakly-supervised 3d pose transfer with keypoints. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15156–15165 (2023)
8. Fu, Z., Wei, J., Shen, W., Song, C., Yang, X., Liu, F., Yang, X., Lin, G.: Sync4d: Video guided controllable dynamics for physics-based 4d generation. arXiv preprint arXiv:2405.16849 (2024)
9. Gao, L., Yang, J., Qiao, Y.L., Lai, Y.K., Rosin, P.L., Xu, W., Xia, S.: Automatic unpaired shape deformation transfer. ACM Transactions on Graphics (ToG) **37**(6), 1–15 (2018)
10. Gleicher, M.: Retargetting motion to new characters. In: Proceedings of the 25th annual conference on Computer graphics and interactive techniques. pp. 33–42 (1998)
11. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4d: Reconstructing and tracking humans with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14783–14794 (2023)
12. Gong, K., Wen, Z., He, W., Xu, M., Wang, Q., Zhang, N., Li, Z., Lian, D., Zhao, W., He, X., et al.: Mocapanything: Unified 3d motion capture for arbitrary skeletons from monocular videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7089–7099 (2026)
13. Hu, L., Zhang, H., Zhang, Y., Zhou, B., Liu, B., Zhang, S., Nie, L.: Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 634–644 (June 2024)
14. 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 (CVPR). pp. 4220–4230 (June 2024)
15. 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)
16. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Computer Vision and Pattern Recognition (CVPR) (2018)

17. 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)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
19. Kovalenko, O., Golyanik, V., Malik, J., Elhayek, A., Stricker, D.: Structure from articulated motion: accurate and stable monocular 3d reconstruction without training data. Sensors **19**(20), 4603 (2019)
20. Krull, A., Brachmann, E., Michel, F., Yang, M.Y., Gumhold, S., Rother, C.: Learning analysis-by-synthesis for 6d pose estimation in rgb-d images. In: Proceedings of the IEEE international conference on computer vision. pp. 954–962 (2015)
21. Li, Y., Takehara, H., Taketomi, T., Zheng, B., Nießner, M.: 4dcomplete: Non-rigid motion estimation beyond the observable surface. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12706–12716 (2021)
22. Li, Z., Chen, Y., Liu, P.: Dreammesh4d: Video-to-4d generation with sparse-controlled gaussian-mesh hybrid representation. Advances in Neural Information Processing Systems **37**, 21377–21400 (2024)
23. 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)
24. Liao, Z., Yang, J., Saito, J., Pons-Moll, G., Zhou, Y.: Skeleton-free pose transfer for stylized 3d characters. In: European Conference on Computer Vision. pp. 640–656. Springer (2022)
25. Liu, J., Qu, Y., Yan, Q., Zeng, X., Wang, L., Liao, R.: Fréchet video motion distance: A metric for evaluating motion consistency in videos. arXiv preprint arXiv:2407.16124 (2024)
26. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015)
27. Lyu, J., Zhu, T., Gu, Y., Lin, L., Cheng, P., Liu, Y., Tang, X., An, L.: Animer: Animal pose and shape estimation using family aware transformer. arXiv preprint arXiv:2412.00837 (2024)
28. Maheshwari, S., Narain, R., Hebbalaguppe, R.: Transfer4d: A framework for frugal motion capture and deformation transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12836–12846 (2023)
29. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
30. Muralikrishnan, S., Dutt, N., Chaudhuri, S., Aigerman, N., Kim, V., Fisher, M., Mitra, N.J.: Temporal residual jacobians for rig-free motion transfer. In: European Conference on Computer Vision. pp. 93–109. Springer (2024)
31. Muralikrishnan, S., Dutt, N.S., Mitra, N.J.: Smf: Template-free and rig-free animation transfer using kinetic codes. ACM Transactions on Graphics (TOG) **44**(6), 1–11 (2025)
32. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 724–732 (2016)
33. Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.Y., Johnson, J., Gkioxari, G.: Accelerating 3d deep learning with pytorch3d. arXiv:2007.08501 (2020)

34. 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)
35. 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)
36. Sabathier, R., Mitra, N.J., Novotny, D.: Animal avatars: Reconstructing animatable 3d animals from casual videos. In: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXIX. p. 270–287 (2024). https://doi.org/10.1007/978-3-031-72986-7_16
37. Shtedritski, A., Rupprecht, C., Vedaldi, A.: Shic: Shape-image correspondences with no keypoint supervision. In: ECCV (2024)
38. Song, C., Li, X., Yang, F., Xu, Z., Wei, J., Liu, F., Feng, J., Lin, G., Zhang, J.: Puppeteer: Rig and animate your 3d models. arXiv preprint arXiv:2508.10898 (2025)
39. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
40. Villegas, R., Ceylan, D., Hertzmann, A., Yang, J., Saito, J.: Contact-aware retargeting of skinned motion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9720–9729 (2021)
41. Villegas, R., Yang, J., Ceylan, D., Lee, H.: Neural kinematic networks for unsupervised motion retargetting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8639–8648 (2018)
42. Wang, A., Ma, W., Yuille, A., Kortylewski, A.: Neural textured deformable meshes for robust analysis-by-synthesis. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3108–3117 (2024)
43. Wang, J., Li, X., Liu, S., De Mello, S., Gallo, O., Wang, X., Kautz, J.: Zero-shot pose transfer for unrigged stylized 3d characters. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8704–8714 (2023)
44. Wang, J., Wen, C., Fu, Y., Lin, H., Zou, T., Xue, X., Zhang, Y.: Neural pose transfer by spatially adaptive instance normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5831–5839 (2020)
45. Wang, Y., Guo, C., Mu, Y., Javed, M.G., Zuo, X., Lu, J., Jiang, H., Cheng, L.: Motiondreamer: One-to-many motion synthesis with localized generative masked transformer. arXiv preprint arXiv:2504.08959 (2025)
46. 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 (CVPR). pp. 20310–20320 (June 2024)
47. Wu, S., Jakab, T., Rupprecht, C., Vedaldi, A.: Dove: Learning deformable 3d objects by watching videos. International Journal of Computer Vision **131**(10), 2623–2634 (2023)
48. Wu, S., Li, R., Jakab, T., Rupprecht, 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)
49. Wu, Z., Yu, C., Jiang, Y., Cao, C., Wang, F., Bai, X.: Sc4d: Sparse-controlled video-to-4d generation and motion transfer. In: European Conference on Computer Vision. pp. 361–379. Springer (2024)

50. Xiao, J., Chai, J.x., Kanade, T.: A closed-form solution to non-rigid shape and motion recovery. In: European conference on computer vision. pp. 573–587. Springer (2004)
51. Yang, Z., Zhu, W., Wu, W., Qian, C., Zhou, Q., Zhou, B., Loy, C.C.: Transmomo: Invariance-driven unsupervised video motion retargeting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5306–5315 (2020)
52. 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)
53. 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. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13248–13258 (2025)
54. Yao, Y., Deng, Z., Hou, J.: Riggs: Rigging of 3d gaussians for modeling articulated objects in videos. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
55. Yoo, S., Koo, J., Yeo, K., Sung, M.: Neural pose representation learning for generating and transferring non-rigid object poses. arXiv preprint arXiv:2406.09728 (2024)
56. Zhang, H., Chang, D., Li, F., Soleymani, M., Ahuja, N.: Magicpose4d: Crafting articulated models with appearance and motion control. arXiv preprint arXiv:2405.14017 (2024)
57. Zhang, H., Tian, Y., Zhou, X., Ouyang, W., Liu, Y., Wang, L., Sun, Z.: Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11446–11456 (2021)
58. Zhang, J.P., Pu, C.F., Guo, M.H., Cao, Y.P., Hu, S.M.: One model to rig them all: Diverse skeleton rigging with unirig. arXiv preprint arXiv:2504.12451 (2025)
59. Zhang, J., Herrmann, C., Hur, J., Chen, E., Jampani, V., Sun, D., Yang, M.H.: Telling left from right: Identifying geometry-aware semantic correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3076–3085 (2024)
60. Zhao, B.N., Wu, J., Wu, S.: Web-scale collection of video data for 4d animal reconstruction. Advances in Neural Information Processing Systems **38** (2026)
61. Zhou, K., Bhatnagar, B.L., Pons-Moll, G.: Unsupervised shape and pose disentanglement for 3d meshes. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16. pp. 341–357. Springer (2020)
62. 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)