CAMO: CATEGORY-AGNOSTIC 3D MOTION TRANSFER FROM MONOCULAR 2D VIDEOS

Taeyeon Kim* **Youngju Na*** **Jumin Lee** **Minhyuk Sung** **Sung-eui Yoon†**

Department of Computer Science, KAIST
*Equal contribution †Corresponding author

https://camo-project-page.github.io/

ABSTRACT

Motion transfer from 2D videos to 3D assets is a challenging problem, due to inherent pose ambiguities and diverse object shapes, often requiring category-specific parametric templates. We propose CAMO, a category-agnostic framework that transfers motion to diverse target meshes directly from monocular 2D videos without relying on predefined templates or explicit 3D supervision. The core of CAMO is a morphology-parameterized articulated 3D Gaussian splatting model combined with dense semantic correspondences to jointly adapt shape and pose through optimization. This approach effectively alleviates shape-pose ambiguities, enabling visually faithful motion transfer for diverse categories. Experimental results demonstrate superior motion accuracy, efficiency, and visual coherence compared to existing methods, significantly advancing motion transfer in varied object categories and casual video scenarios.

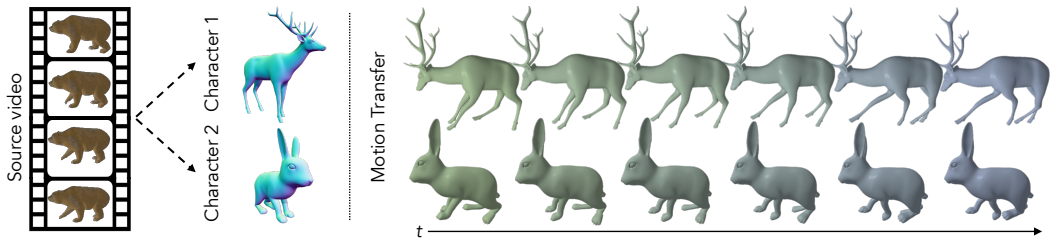

Figure 1: **Conceptual overview of CAMO.** Our method directly transfers articulated motion from 2D video to diverse target objects, without requiring 3D reconstruction of the source or any parametric templates.

1 INTRODUCTION

Efficient 3D character animation remains an important goal in both computer graphics research and content industries such as film (Bregler, 2007), interactive media (Rachmavita, 2020), and robotics (Arduengo et al., 2021). Motion transfer techniques (Aberman et al., 2020; Liao et al., 2022) provide an efficient alternative to manual keyframing or marker-based motion capture by enabling the reuse of existing animations across different characters.

However, a major limitation of many existing methods is their reliance on precomputed 3D sequences, such as articulated skeletons (Aberman et al., 2020) or sparse 3D keypoints (Chen et al., 2023). Acquiring such high-fidelity 3D data is often expensive or impractical in real-world scenarios. To address this data scarcity, recent works (Wang et al., 2023; Muralikrishnan et al., 2024) have explored extracting motion cues directly from readily accessible 2D monocular videos. A common strategy within this domain involves a two-stage reconstruct-then-retarget approach. In this process, a 3D proxy representation of the source subject is first reconstructed from the 2D video, and this intermediate representation is then fed into established 3D-to-3D motion transfer techniques.

Despite demonstrating effective retargeting performance under controlled conditions, these sequential pipelines inherently possess several limitations. A primary limitation stems from their dependence on category-specific priors, such as parametric template models (Loper et al., 2015; Zuffi

et al., 2017), which require large-scale, high-fidelity training data. Although models built on such priors (Kanazawa et al., 2018; Zhang et al., 2021; Rueegg et al., 2022) achieve robust and transferable pose estimation within the structural biases of their target domains, their ability to generalize to diverse shapes and semantic categories remains limited. Furthermore, the cascaded structure of these pipelines can lead to error propagation, where inaccuracies from the reconstruction stage detrimentally impact the fidelity of the final transferred motion.

Our category-agnostic motion transfer framework, **CAMO**, adopts an alternative strategy to conventional reconstruct-then-retarget pipelines. Rather than relying on intermediate 3D reconstructions of the source, we directly project the target character into the 2D observation space, enabling pose optimization purely through image-space supervision. Specifically, we repurpose articulated 3D Gaussian splatting (Yao et al., 2025) (articulated-GS), originally developed for reconstructing articulated animatable objects from 2D videos, to facilitate motion transfer.

CAMO extends this by explicitly modeling morphological differences between source and target characters. Structural variations are decomposed from the target's original shape and adapted to transfer the source motion while preserving topology. To complement this morphology-adaptive optimization and further mitigate shape-pose ambiguity, dense semantic correspondences are established between the 2D source frames and the 3D target mesh, providing semantic guidance for coherent pose recovery. This integration of structural modeling and semantic correspondence guides both visually plausible and semantically coherent pose optimization processes, enabling robust generalization across diverse categories and complex motions. Fig. 1 illustrates the overview of CAMO.

We comprehensively validate CAMO on synthetic benchmarks spanning diverse categories such as humanoids, quadrupeds, and other non-standard animals, as well as on real-world monocular videos. Across all these settings, CAMO consistently preserves motion fidelity and generalizes across diverse morphologies, achieving substantial improvements in both PMD (↓) and FID (↓), with reductions reaching up to 85% on the challenging categories compared to state-of-the-art methods.

2 Related Work

Motion transfer between 3D assets. Traditional techniques in motion transfer have leveraged 3D skeletal structures to enable efficient retargeting across various characters (Gleicher, 1998; Villegas et al., 2018; Aberman et al., 2020; Villegas et al., 2021; Chen et al., 2023). These approaches commonly build upon category-specific skeletal priors, which enable effective performance within their target domains but constrain their generalization to categories outside those domains.

Beyond skeleton-based approaches, skeleton-free deformation methods (Gao et al., 2018; Wang et al., 2020; Liao et al., 2022; Wang et al., 2023; Muralikrishnan et al., 2024; Yoo et al., 2024) are independent from explicit skeletal models, relaxing categorical constraints. Nevertheless, these approaches typically rely on high-quality 3D motion data, which is generally not available for objects across diverse categories. As a result, generalizing these methods to a wider variety of object categories remains a notable challenge, primarily due to the substantial cost and scarcity of such 3D data.

Shape and pose estimation from 2D videos. Another line of research focuses on capturing 3D pose from monocular video. These methods achieve impressive reconstructions within specific domains, often leveraging parametric templates. Representative works include human pose estimation (Zhang et al., 2021; Goel et al., 2023) with SMPL (Loper et al., 2015), and quadruped pose estimation (Rüegg et al., 2023; Lyu et al., 2024) with SMAL (Zuffi et al., 2017). Although effective in domains with abundant 3D scan data, these methods are constrained by their reliance on parametric templates, which limits generalization to categories without extensive 3D pose annotations.

Recent approaches (Yao et al., 2022; Wu et al., 2023a;b; Aygun & Mac Aodha, 2024; Li et al., 2024) explore parametric template-free construction of articulated models from image collections. While promising for intra-class generalization without strong parametric template priors, these methods often struggle to generalize across categories. Uzolas et al. (2023) and Yao et al. (2025) inherently avoid this limitation by employing per-scene optimization to directly decompose shape and skeletal pose from individual dynamic scene observations. However, as their focus lies in reconstruction, their ability to retarget motion to novel characters remains underexplored.

Specifically targeting character animation, auto-rigging methods (Song et al., 2025; Zhang et al., 2025a) predict the skeleton and skinning weights of a 3D asset to apply motion extracted from videos or reconstructed mesh sequences. However, these methods typically require a complete morphological (Song et al., 2025) or skeletal structural correspondence (Zhang et al., 2025a) between the motion source and the target 3D character.

2D to 3D motion transfer. Existing 3D-to-3D motion transfer frameworks (Wang et al., 2023; Muralikrishnan et al., 2024) extend to the 2D domain by combining parametric template-based pose and shape estimators (Zhang et al., 2021; Rueegg et al., 2022) with 3D pose transfer techniques. These shape estimators are typically demonstrated on humanoid or quadruped characters respectively, where the reliance on categorical templates (Loper et al., 2015; Zuffi et al., 2017) fundamentally limits their ability to generalize to novel categories. Moreover, we observe that sequentially combining independently trained components often leads to cumulative errors, ultimately degrading the fidelity of transferred motion.

Maheshwari et al. (2023) propose a category-agnostic approach that removes template priors, transferring motion from RGB-D videos to 3D meshes by estimating skeletal motion from reconstructed meshes; its performance, however, hinges on accurate depth input, limiting robustness in casual or monocular RGB settings. In contrast, Fu et al. (2024) and Zhang et al. (2024) achieve 2D-to-3D motion transfer without depth by reconstructing motion with neural bones (Yang et al., 2022) or by leveraging image-to-3D generative models (Liu et al., 2023). Despite improved generalizability, these approaches remain tied to intermediate reconstruction stages (e.g., pseudo-3D supervision or skeletonization), which makes them sensitive to reconstruction errors and less robust under large morphological variations.

In contrast, we directly leverage 2D RGB videos as motion sources through morphology-adaptive shape and pose parameter optimization. By bypassing intermediate 3D reconstruction, our approach mitigates reconstruction errors and enables robust motion transfer across diverse object categories and morphological variances without relying on category-specific templates.

3 METHODS

Our goal is to transfer articulated motion from a monocular video to arbitrary 3D characters. We take as input a static 3D target mesh $\mathcal{M}^{tgt}$ and a source monocular RGB video with paired foreground masks $\{I_t, M_t\}_{t=0}^{T}$, where I_t is a frame from time t, and M_t is obtained via off-the-shelf segmentation model (Kirillov et al., 2023). We aim to produce a temporally coherent sequence of deformed meshes $\{\mathcal{M}_t^{tgt}\}_{t=0}^{T}$ that faithfully reproduces the source motion.

We first encapsulate the target mesh with an Articulated-GS (Yao et al., 2025) representation with pose parameters (Sec. 3.1). We then parameterize morphology using learnable bone lengths, a global scale, and local Gaussian offsets (Sec. 3.2). This representation disentangles shape variation from pose dynamics. Finally, all shape and pose parameters are optimized jointly via differentiable rendering and dense semantic correspondences (Sec. 3.3–3.4), yielding semantically coherent motion aligned to the source. Fig. 2 illustrates the full pipeline.

3.1 ARTICULATED 3D GAUSSIAN SPLATTING FOR IMAGE-SPACE OPTIMIZATION

Retargeting motion from a monocular video typically requires estimating the 3D geometry of the source subject. However, inferring accurate 3D pose and shape from 2D inputs is inherently ambiguous. Reliance on these estimated 3D priors often introduces errors that propagate to the final result. We propose a direct optimization strategy to address this issue. We optimize the target character to align directly with the 2D source video observations. This approach bypasses the need for an explicit intermediate 3D representation of the source.

To this end, we employ Articulated 3D Gaussian Splatting (Articulated-GS) (Yao et al., 2025). This framework defines the target character using a single, unified canonical shape. We deform this time-invariant geometry via Linear Blend Skinning (LBS) to match the pose in each video frame. Critically, our optimization updates this single canonical shape to satisfy projection constraints across all time steps and camera views. This enforces geometric consistency throughout the entire motion sequence.

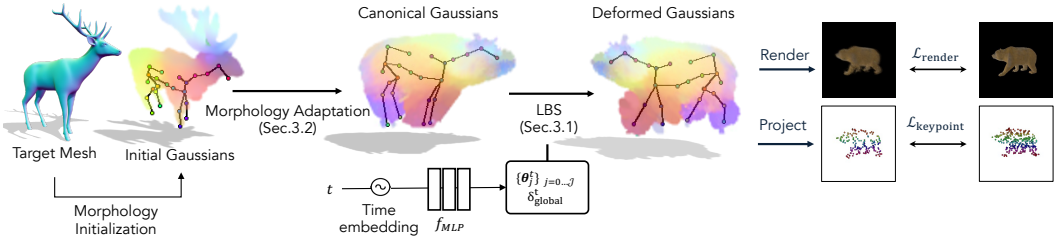

Figure 2: **Overview of the morphology-adaptive articulated Gaussian splatting pipeline.** Given a target mesh, we parameterize it with deformable 3D Gaussians. A time-conditioned MLP (f_{MLP}) predicts skeletal transformations driven by input time embeddings. Crucially, our pipeline employs morphology adaptation (Sec. 3.2) to align the target's canonical structure, followed by LBS-based deformation (Sec. 3.1) for articulation. The framework is optimized end-to-end using differentiable rendering ($\mathcal{L}_{\text{render}}$) and semantic keypoint constraints ($\mathcal{L}_{\text{keypoint}}$) consistent with the source video.

Target Representation. We represent the target character using a set of 3D Gaussians attached to a kinematic skeleton $\mathcal{T} = (\mathcal{J}, \mathcal{A})$, where $\mathcal{J}$ denotes the set of joints and $\mathcal{A} = \{A_j\}_{j \in \mathcal{J} \setminus \{j_r\}}$ maps each joint j to its parent A_j, with j_r being the root. Each Gaussian G_i is parameterized by its mean $\boldsymbol{\mu}_i \in \mathbb{R}^3$, rotation $\boldsymbol{q}_i \in \mathbb{R}^4$, scale $\boldsymbol{s}_i \in \mathbb{R}^3$, opacity $\sigma_i \in [0, 1]$, and spherical harmonic coefficients $\mathcal{SH}_i \in \mathbb{R}^K$. Unlike previous works that initialize from sparse point clouds, we leverage the explicit geometry of the target mesh to initialize these Gaussian positions $\boldsymbol{\mu}_i$ (Sec. 3.2). For unrigged meshes, we employ automatic rigging methods (Xu et al., 2020; Zhang et al., 2025b) to establish the skeletal structure.

Kinematic Deformation. To capture temporal dynamics, a time-conditioned MLP, f_{MLP}, predicts the skeletal pose for each timestamp t. Given a sinusoidal time embedding $\text{emb}(t)$, the network outputs the root translation and relative joint rotations:

$$\left\{\{\boldsymbol{\theta}_j^t\}_{j \in \mathcal{J}},\ \boldsymbol{\delta}_{\text{global}}^t\right\} = f_{\text{MLP}}(\text{emb}(t)), \tag{1}$$

where $\boldsymbol{\theta}_j^t$ is the unit quaternion for joint j and $\boldsymbol{\delta}_{\text{global}}^t$ is the global translation. These predictions drive the deformation of the canonical Gaussians. The deformed position $\boldsymbol{\mu}_i^t$ of Gaussian i is computed via LBS:

$$\boldsymbol{\mu}_i^t = \boldsymbol{\delta}_{\text{global}}^t + \sum_{j \in \mathcal{J}} w_{ij} \mathbf{T}_j^t \bar{\boldsymbol{\mu}}_i, \quad \mathbf{T}_j^t = \prod_{k \in \text{P}(\text{root}, j)} \bar{\mathbf{T}}_k^t, \quad \bar{\mathbf{T}}_k^t = \begin{pmatrix} \mathbf{R}_k^t & \mathbf{J}_{A_k} - \mathbf{R}_k^t \mathbf{J}_{A_k} \\ 0 & 1 \end{pmatrix}. \tag{2}$$

Here, $\bar{\boldsymbol{\mu}}_i$ is the canonical center, w_{ij} is the skinning weight, and $\mathbf{R}_k^t$ is the rotation matrix derived from $\boldsymbol{\theta}_k^t$. This formulation ensures that the Gaussians move coherently according to the skeletal hierarchy.

Differentiable Rendering. The deformed Gaussians are rasterized into 2D images to compute the optimization loss. For a viewpoint $\boldsymbol{v}$ and pixel $\boldsymbol{u}$, the color $\mathcal{C}(\boldsymbol{u})$ is derived via alpha compositing:

$$\mathcal{C}(\boldsymbol{u}) = \sum_{i \in \mathcal{N}} T_i\, \alpha_i\, \mathcal{SH}(\boldsymbol{sh}_i, \boldsymbol{v}), \quad T_i = \prod_{j=1}^{i-1} (1 - \alpha_j). \tag{3}$$

This differentiable rendering process allows us to backpropagate gradients from the 2D projection error directly to the 3D pose and shape parameters, bridging the domain gap between the 2D source and 3D target.

3.2 Morphology-Adaptive Shape Parameterization

Standard Articulated-GS assumes a fixed skeletal topology, which restricts its ability to transfer motion between characters with differing limb proportions. To address this, we introduce a morphology-adaptive parameterization that explicitly disentangles structural variations from pose dynamics. In this paper, we use the term *morphology* to refer to the character's limb proportion, global body scale, and local shape details. By optimizing these time-invariant parameters alongside time-variant poses, our framework enables the target character to adapt its shape to the source motion while preserving kinematic coherence (Fig. 3 (b)).

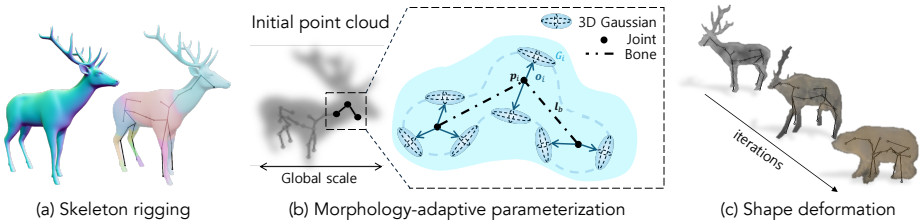

Figure 3: **Deformable morphology parameterization.** **(a)** We initialize the target character with skeleton rigging, acquiring the topological structure and skinning weights. **(b)** Morphology-adaptive parameterization of structural variations. **(c)** During optimization, shape parameters deform the target's morphological structure to align with the morphology of the source.

Learnable Bone Lengths. We first relax the fixed skeleton constraint by assigning a learnable scalar length $\ell_b \in \mathbb{R}^+$ to each bone $b \in \mathcal{B}$. Given the unit direction vector $\boldsymbol{v}_b \in \mathbb{R}^3$ from a parent to a child joint, the rest-pose position of any joint j is determined by the cumulative length of bones along the kinematic chain:

$$\mathbf{j}_{\text{rest}}(j) = \mathbf{j}_{\text{rest}}(j_{root}) + \sum_{b \in \text{P}(\text{root}, j)} \ell_b \boldsymbol{v}_b. \tag{4}$$

This allows the skeleton to stretch or shrink segments (e.g., legs or arms) to match the source subject's proportions purely through optimization.

Morphology-Aware Gaussian Initialization. Crucially, the surface geometry must adapt to these skeletal changes. Instead of treating Gaussian positions as independent variables, we parameterize the mean $\boldsymbol{\mu}_i$ of each Gaussian G_i relative to the underlying bone structure. We define $\boldsymbol{\mu}_i$ as a displacement from a skeleton-anchored reference point $\boldsymbol{p}_i$:

$$\boldsymbol{\mu}_i = \boldsymbol{p}_i + \boldsymbol{o}_i, \quad \text{where} \quad \boldsymbol{p}_i = \sum_{j \in \mathcal{J}} w_{ij}\, \mathbf{j}_{\text{rest}}(j), \tag{5}$$

where $\boldsymbol{p}_i$ represents the coarse geometry derived from joint positions $\mathbf{j}_{rest}(j)$ LBS weights w_{ij}, while the learnable offset $\boldsymbol{o}_i \in \mathbb{R}^3$ captures fine-grained local shape deviations. This formulation ensures that when bone lengths ℓ_b change, the associated Gaussians move coherently with the skeleton, preventing geometric artifacts.

Global Scale and Canonical Shape. Finally, to resolve the scale ambiguity inherent in monocular video, we introduce a global scaling factor $s_{global} \in \mathbb{R}^+$. This factor uniformly scales the entire morphology-parameterized character. The final canonical position $\bar{\boldsymbol{\mu}}_i$ used for deformation (Eq. 2) is obtained by:

$$\bar{\boldsymbol{\mu}}_i = s_{global} \cdot \boldsymbol{\mu}_i. \tag{6}$$

By jointly optimizing bone lengths (ℓ_b), local offsets ($\boldsymbol{o}_i$), and global scale (s_{global}), our parameterization allows the target mesh to conform to the source's morphology while maintaining its original topological structure (Fig. 3 (c)).

Discussion. Our morphology parameterization provides a structural basis for mitigating the shape-pose ambiguity inherent in 2D-to-3D motion transfer. By explicitly decoupling global scale, skeletal lengths, and local offsets, our formulation promotes geometric identifiability under non-degenerate motion conditions, showing that morphological changes are distinguishable from pose dynamics. This disentanglement facilitates stable optimization by reducing the solution space to physically plausible configurations. We provide a detailed discussion on theoretical analysis in the Supplementary Material.

3.3 Target-source dense semantic correspondence

While our proposed shape parameterization accounts for morphological differences, a key challenge in transferring articulated motion from 2D to 3D remains: *shape–pose ambiguity*. This refers to the inherent uncertainty in disentangling an object's underlying pose from its observation. Photometric loss provides essential low-level supervision, but relying on it alone may produce motion artifacts, as it captures only visual cues and lacks explicit semantic correspondences between characters.

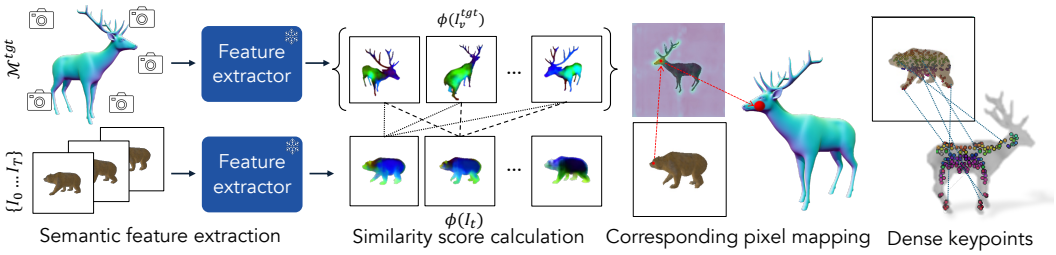

Figure 4: **Dense target-source correspondences matching.** We extract robust 2D-to-3D semantic correspondences by matching semantic features between source frames and rendered target views.

These artifacts can be mitigated by incorporating additional semantic cues, which help disambiguate overlapping projections, particularly when source and target morphologies differ.

To address this, we establish robust 2D-3D semantic correspondences by leveraging pre-trained vision foundation models. Specifically, we utilize an orientation-sensitive feature extractor (Yang et al., 2020) that produces spatially consistent descriptors across varied poses and morphologies, then obtain dense pixel-to-vertex mappings through semantic feature matching between input images and target mesh renderings (Shtedritski et al., 2024). This provides automatic correspondence estimation without requiring manual registration or additional training.

The detailed pipeline of our dense correspondence extraction module is illustrated in Fig. 4. We first compute the similarity score of the dense semantic features extracted by the feature extractor $\phi(\cdot)$ from a source video frame I_t with those from multiple rendered views $\{I_v^{\text{tgt}}\}$ of the target mesh $\mathcal{M}^{\text{tgt}}$. Then, given a source pixel $\boldsymbol{p} \in I_t$ with the extracted feature $\phi(I_t)$, we compute a pooled similarity score $\Sigma_{I_t}(\boldsymbol{p}, \boldsymbol{x}_k)$ for each vertex $\boldsymbol{x}_k \in \mathcal{M}^{\text{tgt}}$ as:

$$\Sigma_{I_t}(\boldsymbol{p}, \boldsymbol{x}_k) = \underset{v, \boldsymbol{x}_k \in \text{vis}(I_v^{\text{tgt}})}{\text{pool}} S\left(\phi(I_t)[\boldsymbol{p}],\ \phi(I_v^{\text{tgt}})[\pi_v(\boldsymbol{x}_k)]\right), \tag{7}$$

where $S(\cdot)$ denotes a cosine similarity, $\pi_v(\boldsymbol{x}_k)$ denotes the 2D projection of vertex $\boldsymbol{x}_k$ onto the rendered image I_v^{tgt}, and $\phi(I_v^{\text{tgt}})[\pi_v(\boldsymbol{x}_k)]$ is the corresponding feature vector at the 2D projected location. The operator pool aggregates similarity scores via max-pooling across all v target-rendered views where $\boldsymbol{x}_k$ is visible.

The best-matching 3D vertex $\tilde{\boldsymbol{x}}_{\boldsymbol{p},t}^{3D}$ for each pixel $\boldsymbol{p}$ in frame t is obtained by selecting the vertex with the highest pooled similarity score:

$$\tilde{\boldsymbol{x}}_{\boldsymbol{p},t}^{3D} = \arg \max_{\boldsymbol{x}_k \in \mathcal{V}(\mathcal{M}^{\text{tgt}})} \Sigma_{I_t}(\boldsymbol{p}, \boldsymbol{x}_k), \tag{8}$$

where $\mathcal{V}(\mathcal{M}^{\text{tgt}})$ denotes the set of vertices of the target mesh. These retrieved 3D points $\tilde{\boldsymbol{x}}_{\boldsymbol{p},t}^{3D}$ serve as semantic keypoints, providing supervision to guide semantic structure alignment of cross-modality during optimization, as the keypoint loss L_{keypoint} (Sec. 3.4).

3.4 Optimization

As formalized in Eq. 1 and visualized in Fig. 2, our primary objective is to recover the target mesh's time-varying skeletal pose parameters aligned with the source motion, relying solely on 2D observations without ground-truth 3D annotations or any form of pose template prior. The entire framework, composed of morphology-parameterized articulated Gaussians, is optimized end-to-end by minimizing a composite loss function. Our optimization objective combines photometric reconstruction, semantic correspondence, and multiple regularization terms: $\mathcal{L}_{\text{total}} = \lambda_{\text{render}}\mathcal{L}_{\text{render}} + \lambda_{\text{keypoint}}\mathcal{L}_{\text{keypoint}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}$, where the weights balance their respective contributions.

The render loss enforces photometric consistency between the rendered frame $\hat{I}_t$ (from Eq. 3) and the source frame I_t by combining an ℓ_1 term with a SSIM (Wang et al., 2004) term:

$$\mathcal{L}_{\text{render}} = \sum_{t=0}^{T} \left[(1 - \lambda_{\text{dSSIM}}) \left\| \hat{I}_t - I_t \right\|_1 + \lambda_{\text{dSSIM}} \left(1 - \text{SSIM}(\hat{I}_t, I_t)\right) \right]. \tag{9}$$

Table 1: **Quantitative evaluation on Mixamo and DT4D datasets.** Our method consistently outperforms all baselines across diverse categories. Results are averaged across scenes, with per-scene results in the Appendix.

	Mixamo		DT4D-Quadrupeds		DT4D-Others	
	PMD ↓	FID ↓	PMD ↓	FID ↓	PMD ↓	FID ↓
SPT$^+$	0.0029	0.0366	-	-	-	-
NPR$^+$	0.0099	0.0551	0.0032	0.0669	-	-
Transfer4D	0.0084	0.0855	0.0058	0.0505	0.0133	0.0805
Ours	**0.0028**	**0.0304**	**0.0018**	**0.0171**	**0.0023**	**0.0124**

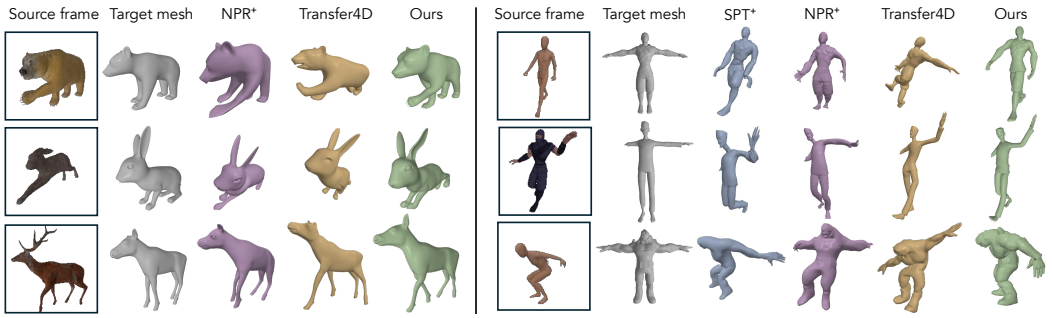

Figure 5: **Qualitative results on Mixamo and DT4D-Quadruped datasets.** Our method shows superior pose alignment compared to baselines across diverse objects. Refer to the supplementary video for full animation.

The keypoint loss supervises geometric alignment by minimizing projection error between source image pixels and their matched 3D vertices derived from dense semantic correspondences:

$$\mathcal{L}_{\text{keypoint}} = \sum_{t=0}^{T} \sum_{\boldsymbol{p} \in \mathcal{P}_t} \left\| \boldsymbol{p} - \pi_t \left(\tilde{\boldsymbol{x}}_{\boldsymbol{p},t}^{3D} \right) \right\|_2, \tag{10}$$

where $\tilde{\boldsymbol{x}}_{\boldsymbol{p},t}^{3D}$ is the best-matching 3D vertex obtained via Eq. 8, and $\mathcal{P}_t$ represents sampled foreground pixels. Finally, $\mathcal{L}_{\text{reg}}$ comprises multiple regularization terms that encourage temporal smoothness and geometric consistency (detailed formulations provided in the Supplementary Material).

4 Experiments

4.1 Datasets and Implementations

Datasets. We evaluate our approach on mesh-animation pairs sampled from DeformingThings-4D (DT4D) (Li et al., 2021) and Mixamo (Adobe, 2025). From DT4D, we select 20 animation pairs spanning diverse animal categories of quadrupeds and non-quadrupeds exhibiting varied motions. From Mixamo, we utilize 12 humanoid mesh-animation pairs across different character models and motion types. To simulate a *casually captured* monocular video scenario, we render each source animation using a single camera with constrained movement (±30° angular range), generating input frames with corresponding ground-truth 3D target mesh animations. We further conduct qualitative evaluation on real-world videos sourced from the DAVIS dataset (Perazzi et al., 2016) and two publicly available online videos (Daley, n.d.; Nicky Pe, n.d.), as well as 2D-to-2D motion transfer scenarios using additional synthetic sequences (Pumarola et al., 2021; Liu et al., 2024). Details on dataset preparation and configuration are provided in the Supplementary Material.

Implementation details. We employ a two-stage optimization strategy that first performs global alignment of scale and translation, then jointly refines local pose and shape parameters (bone length, Gaussians) to adapt morphology while preserving essential motion characteristics. All experiments use the Adam optimizer (Kingma & Ba, 2014) with adaptive learning rates over 10k iterations. Our method achieves efficient optimization, completing training in under 10 minutes on a single RTX 4090 GPU. Detailed hyperparameter specifications are provided in the Supplementary Material.

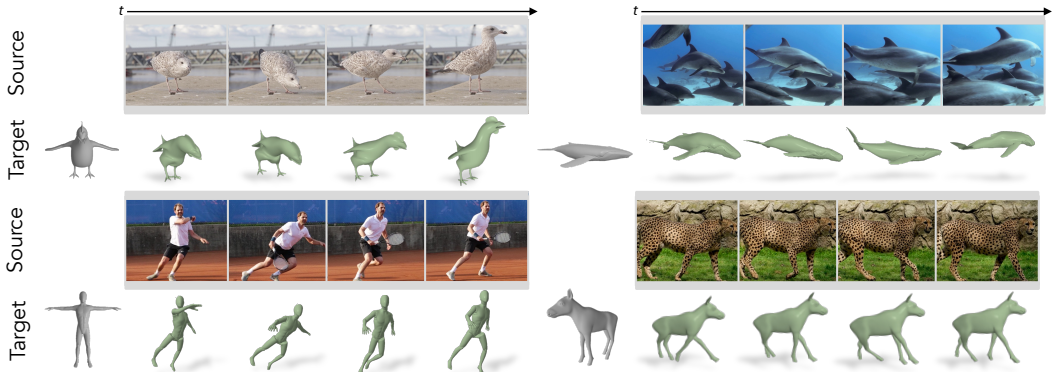

Figure 6: **Qualitative results on real-world datasets.** Our motion transfer pipeline effectively transfers motion from both synthetic and real-world videos in a category-agnostic manner.

4.2 2D-TO-3D MOTION TRANSFER

Baselines and metrics. We compare our method against two baseline categories: *composite pipelines* combining 2D-to-3D reconstruction with 3D motion transfer, and a template-free optimization-based approach (Transfer4D (Maheshwari et al., 2023)). For composite baselines, we adopt a two-stage setup with mesh reconstruction followed by motion transfer using SPT (Liao et al., 2022) and NPR (Yoo et al., 2024), denoted as SPT$^+$ and NPR$^+$ (see the Supplementary Material for baseline implementation details). SPT$^+$ is evaluated only on humanoid motion, as the original method was designed and tested on stylized human characters. Transfer4D performs motion retargeting by extracting skeletal structure from RGB-D input. On datasets with non-quadruped animals, where parametric templates of reconstruction methods are not applicable, we compare only to Transfer4D.

We quantify motion transfer by comparing the retargeted and ground-truth mesh sequences. Consistent with prior work (Liao et al., 2022; Yoo et al., 2024), we adopt Point-wise Mesh Distance (PMD) to measure per-vertex accuracy and Fréchet Inception Distance (FID) (Heusel et al., 2017) to assess perceptual fidelity. To compute FID, both ground-truth and retargeted animations are rendered from 12 viewpoints and their image distributions are compared.

Comparison results. We evaluate our method and baselines on DT4D and Mixamo datasets. As shown in Tab. 1, our approach achieves superior performance on both PMD and FID metrics. These results show that our approach achieves strong performance in a data-efficient manner, relying only on direct optimization without explicit 3D supervision. On non-quadrupeds (DT4D-Others), we significantly outperform Transfer4D even without depth input, demonstrating strong performance beyond parametric model categories.

Fig. 5 demonstrates that our method preserves the target shape and transfers motion faithfully, while baselines often produce distorted shapes by estimating incorrect transformation (Liao et al., 2022; Maheshwari et al., 2023) or relying on predicted surface Jacobians (Yoo et al., 2024). This shape fidelity is attributed to our morphology-parameterization, which we also analyze in Sec. 4.3.

Qualitative results on real-world videos. To evaluate real-world applicability, we apply our method to in-the-wild monocular videos featuring diverse animal categories with complex backgrounds and occlusions. These noisy or open-domain scenarios represent cases where obtaining corresponding 3D animations is challenging. As shown in Fig. 6, our approach successfully transfers motion across these varied scenarios while preserving target mesh structure and proportions. These results demonstrate effective motion transfer directly from monocular input without requiring 3D motion generation, highlighting the practical value of our 2D-grounded motion transfer approach.

4.3 ABLATION STUDY

We ablate key components of our framework in Tab. 2 and Fig. 7. Removing the rendering loss severely degrades performance (PMD $\uparrow \sim 5\times$), indicating it as the primary driver of motion transfer,

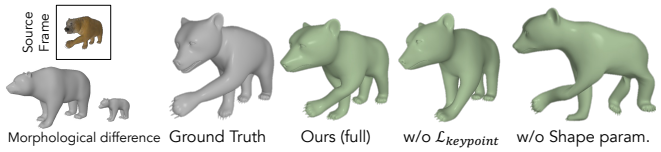

Figure 7: **Qualitative ablations.** Keypoint loss complements motion details and accuracy. Excluding shape parameters induces severe geometric artifacts for large morphological variation.

Table 2: **Quantitative evaluation of component contributions.**

Ablation	PMD (↓)	FID (↓)
Full Model	**0.0018**	**0.0171**
w/o $\mathcal{L}_{\text{render}}$	0.0090	0.0463
w/o Shape param.	0.0047	0.0747
w/o μ update	0.0039	0.0552
w/o l_b & s_{global} update	0.0040	0.0488
w/o $\mathcal{L}_{\text{keypoint}}$	0.0031	0.0252

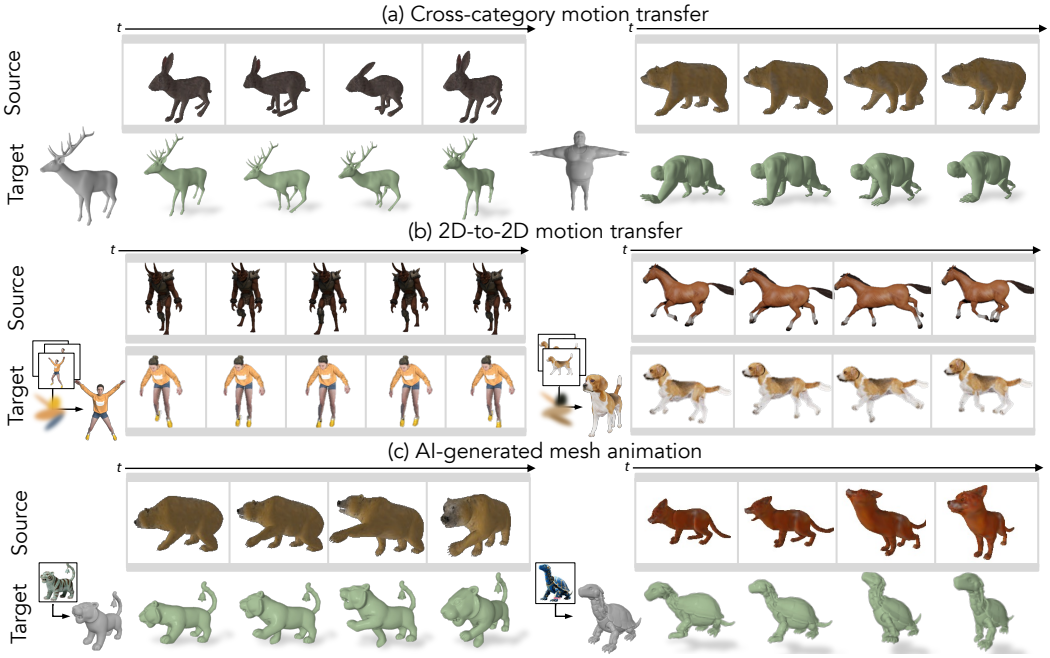

Figure 8: **Results on diverse applications.** Our method transfers motion for (a) cross-category source-target pairs, (b) 2D-to-2D videos, and (c) AI-generated mesh animations.

while the keypoint loss adds complementary semantic guidance. Fig. 7 shows that dropping the keypoint loss yields suboptimal transfers due to unresolved shape-pose ambiguities.

Excluding our shape parameterization (bone lengths l_b, Gaussian means μ, global scale s_{global}) causes distorted geometry and misaligned orientations, especially under large morphological differences. With shape parameters fixed, global translation lowers render loss by pulling the object toward the camera, partially recovering motion but distorting orientation and pose (Fig. 7; see supplementary videos). Overall, adding each component yields consistent gains (Tab. 2), confirming their complementary roles to enhance robustness. Extended ablation studies appear in the Supplementary Material.

4.4 Diverse Application Scenarios

Cross-category motion transfer. Our method demonstrates strong generalization across diverse categories, as shown in Fig. 8 (a). We successfully transfer motion between different animal species (rabbit-to-deer) and even across broader categories (animal-to-human). This flexibility stems from our universal optimization approach that does not rely on category-specific skeletal structures or explicit category matching between source and target.

2D-to-2D motion transfer. A key advantage of our method is its representation-agnostic applicability across articulated 3D assets. While primarily demonstrated on mesh targets, our framework seamlessly extends to Gaussian-based 3D representation without modification of core design.

Fig. 8(b) shows motion transfer to 3DGS reconstructed from multi-view images (Yao et al., 2025), enabling video-to-video transfer when both source and target originate from RGB sequences. Together, these results yield a single, category-agnostic framework that operates consistently across varied 3D representations.

AI-generated mesh animation. Another interesting application is animating meshes synthesized by generative models. As shown in Fig. 8 (c), we achieve effective motion transfer using meshes generated from an off-the-shelf image-to-mesh model (Zhao et al., 2025). This demonstrates the versatility of our approach to meshes from diverse sources, supporting modern content creation workflows that increasingly incorporate AI-generated assets.

5 Discussion

We introduce **CAMO**, a framework that transfers motion from monocular videos to 3D assets without relying on category-specific templates. By reformulating motion retargeting as an efficient morphology-adaptive optimization on articulated Gaussian splats, our method avoids error accumulation in traditional reconstruct-then-retarget pipelines without any 3D supervision or large datasets. The integration of morphology-adaptive modeling and semantic correspondences provides complementary cues that reduce shape-pose ambiguities and enable broad applicability across different skeletal structures and 3D representations.

Limitations and future work. While CAMO achieves robust category-agnostic motion transfer, the current morphology-adaptive formulation is limited to articulated kinematic structures. This restricts its ability to capture richer non-rigid dynamics such as soft-tissue deformation or secondary motion (e.g., hair dynamics, tail sway). Beyond these kinematic limitations, our framework currently prioritizes visual motion transfer rather than enforcing full physical plausibility. A promising direction is to augment our optimization with physically grounded constraints, such as Jacobian-space motion consistency and contact-aware regularization. Another promising avenue for future work is to enrich the framework with additional geometric cues, such as monocular depth predictors or generative 3D priors. These sources of structure-aware regularization could improve robustness in complex scenes or under limited camera motion, further extending the applicability of our approach.

References

Kfir Aberman, Peizhuo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen. Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics (TOG)*, 39(4):62–1, 2020.

Adobe. Mixamo. `https://www.mixamo.com`, 2025.

Miguel Arduengo, Ana Arduengo, Adrià Colomé, Joan Lobo-Prat, and Carme Torras. Human to robot whole-body motion transfer. In *2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)*, pp. 299–305. IEEE, 2021.

Mehmet Aygun and Oisin Mac Aodha. Saor: Single-view articulated object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10382–10391, 2024.

Chris Bregler. Motion capture technology for entertainment [in the spotlight]. *IEEE Signal Processing Magazine*, 24(6):160–158, 2007.

Jinnan Chen, Chen Li, and Gim Hee Lee. Weakly-supervised 3d pose transfer with keypoints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15156–15165, 2023.

Paul Daley. Close-up video of white seagull. `https://www.pexels.com/video/close-up-video-of-white-seagull-1536290/`, n.d. Pexels video; accessed 2025-09-23.

Zhoujie Fu, Jiacheng Wei, Wenhao Shen, Chaoyue Song, Xiaofeng Yang, Fayao Liu, Xulei Yang, and Guosheng Lin. Sync4d: Video guided controllable dynamics for physics-based 4d generation. *arXiv preprint arXiv:2405.16849*, 2024.

Lin Gao, Jie Yang, Yi-Ling Qiao, Yu-Kun Lai, Paul L Rosin, Weiwei Xu, and Shihong Xia. Automatic unpaired shape deformation transfer. *ACM Transactions on Graphics (ToG)*, 37(6):1–15, 2018.

Michael Gleicher. Retargetting motion to new characters. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pp. 33–42, 1998.

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*, pp. 14783–14794, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Yang Li, Hikari Takehara, Takafumi Taketomi, Bo Zheng, and Matthias Nießner. 4dcomplete: Non-rigid motion estimation beyond the observable surface. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12706–12716, 2021.

Zizhang Li, Dor Litvak, Ruining Li, Yunzhi Zhang, Tomas Jakab, Christian Rupprecht, Shangzhe Wu, Andrea Vedaldi, and Jiajun Wu. Learning the 3d fauna of the web. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9752–9762, 2024.

Zhouyingcheng Liao, Jimei Yang, Jun Saito, Gerard Pons-Moll, and Yang Zhou. Skeleton-free pose transfer for stylized 3d characters. In *European Conference on Computer Vision*, pp. 640–656. Springer, 2022.

Isabella Liu, Hao Su, and Xiaolong Wang. Dynamic gaussians mesh: Consistent mesh reconstruction from dynamic scenes. *arXiv preprint arXiv:2404.12379*, 2024.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 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.

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

Jin Lyu, Tianyi Zhu, Yi Gu, Li Lin, Pujin Cheng, Yebin Liu, Xiaoying Tang, and Liang An. Animer: Animal pose and shape estimation using family aware transformer. *arXiv preprint arXiv:2412.00837*, 2024.

Shubh Maheshwari, Rahul Narain, and Ramya Hebbalaguppe. 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.

Sanjeev Muralikrishnan, Niladri Dutt, Siddhartha Chaudhuri, Noam Aigerman, Vladimir Kim, Matthew Fisher, and Niloy J Mitra. Temporal residual jacobians for rig-free motion transfer. In *European Conference on Computer Vision*, pp. 93–109. Springer, 2024.

Nicky Pe. A cheetah walking and looking around. `https://www.pexels.com/video/a-cheetah-walking-and-looking-around-8451567/`, n.d. Pexels video; accessed 2025-09-23.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. 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.

Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10318–10327, 2021.

FP Rachmavita. Interactive media-based video animation and student learning motivation in mathematics. In *Journal of Physics: Conference Series*, volume 1663, pp. 012040. IOP Publishing, 2020.

Nadine Rueegg, Silvia Zuffi, Konrad Schindler, and Michael J Black. 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.

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

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. Shic: Shape-image correspondences with no keypoint supervision. In *ECCV*, 2024.

Chaoyue Song, Xiu Li, Fan Yang, Zhongcong Xu, Jiacheng Wei, Fayao Liu, Jiashi Feng, Guosheng Lin, and Jianfeng Zhang. Puppeteer: Rig and animate your 3d models. *arXiv preprint arXiv:2508.10898*, 2025.

Lukas Uzolas, Elmar Eisemann, and Petr Kellnhofer. Template-free articulated neural point clouds for reposable view synthesis. *Advances in Neural Information Processing Systems*, 36:31621–31637, 2023.

Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee. Neural kinematic networks for unsupervised motion retargetting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8639–8648, 2018.

Ruben Villegas, Duygu Ceylan, Aaron Hertzmann, Jimei Yang, and Jun Saito. Contact-aware retargeting of skinned motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9720–9729, 2021.

Jiashun Wang, Chao Wen, Yanwei Fu, Haitao Lin, Tianyun Zou, Xiangyang Xue, and Yinda Zhang. 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.

Jiashun Wang, Xueting Li, Sifei Liu, Shalini De Mello, Orazio Gallo, Xiaolong Wang, and Jan Kautz. 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.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Shangzhe Wu, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Dove: Learning deformable 3d objects by watching videos. *International Journal of Computer Vision*, 131(10):2623–2634, 2023a.

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

Zhan Xu, Yang Zhou, Evangelos Kalogerakis, Chris Landreth, and Karan Singh. Rignet: Neural rigging for articulated characters. *arXiv preprint arXiv:2005.00559*, 2020.

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

Karren Yang, Bryan Russell, and Justin Salamon. Telling left from right: Learning spatial correspondence of sight and sound. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9932–9941, 2020.

Chun-Han Yao, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. Lassie: Learning articulated shapes from sparse image ensemble via 3d part discovery. *Advances in Neural Information Processing Systems*, 35:15296–15308, 2022.

Yuxin Yao, Zhi Deng, and Junhui Hou. Riggs: Rigging of 3d gaussians for modeling articulated objects in videos. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Seungwoo Yoo, Juil Koo, Kyeongmin Yeo, and Minhyuk Sung. Neural pose representation learning for generating and transferring non-rigid object poses. *arXiv preprint arXiv:2406.09728*, 2024.

Hao Zhang, Di Chang, Fang Li, Mohammad Soleymani, and Narendra Ahuja. Magicpose4d: Crafting articulated models with appearance and motion control. *arXiv preprint arXiv:2405.14017*, 2024.

Hao Zhang, Haolan Xu, Chun Feng, Varun Jampani, and Narendra Ahuja. Physrig: Differentiable physics-based skinning and rigging framework for realistic articulated object modeling. *arXiv preprint arXiv:2506.20936*, 2025a.

Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. 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.

Jia-Peng Zhang, Cheng-Feng Pu, Meng-Hao Guo, Yan-Pei Cao, and Shi-Min Hu. One model to rig them all: Diverse skeleton rigging with unirig. *arXiv preprint arXiv:2504.12451*, 2025b.

Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025.

Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 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.