REACTO: Reconstructing Articulated Objects from a Single Video

Chaoyue Song[1,2], Jiacheng Wei[1, †], Chuan Sheng Foo[2,3], Guosheng Lin[1, †], Fayao Liu[2, †]
[1]Nanyang Technological University, [2]Institute for Inforcomm Research, A*STAR
[3]Centre for Frontier AI Research, A*STAR
[†]Corresponding Authors
{chaoyue002@e., jiacheng.wei@, gslin@}ntu.edu.sg, {foo_chuan_sheng, liu_fayao}@i2r.a-star.edu.sg

Abstract

In this paper, we address the challenge of reconstructing general articulated 3D objects from a single video. Existing works employing dynamic neural radiance fields have advanced the modeling of articulated objects like humans and animals from videos, but face challenges with piece-wise rigid general articulated objects due to limitations in their deformation models. To tackle this, we propose Quasi-Rigid Blend Skinning, a novel deformation model that enhances the rigidity of each part while maintaining flexible deformation of the joints. Our primary insight combines three distinct approaches: 1) an enhanced bone rigging system for improved component modeling, 2) the use of quasi-sparse skinning weights to boost part rigidity and reconstruction fidelity, and 3) the application of geodesic point assignment for precise motion and seamless deformation. Our method outperforms previous works in producing higher-fidelity 3D reconstructions of general articulated objects, as demonstrated on both real and synthetic datasets. Project page: https://chaoyuesong.github.io/REACTO.

1. Introduction

We focus on reconstructing general articulated objects from a casually captured monocular video, a challenging task that involves creating 3D models from everyday footage and dealing with the complexity of objects with movable parts. Understanding and recognizing the structure of general articulated objects from videos plays a crucial role in various fields, such as robotics, animation, 3D generation [4, 5, 60] virtual reality, and augmented reality.

Recently, NASAM [59] introduced a method to learn categories of articulated objects from multi-view images across various articulations. However, this approach necessitates training on several objects within the same category. Another method, PARIS [27], was proposed to learn articulation in a self-supervised manner but relies on multi-view images to provide complete views of the object at different articulations. Consequently, both of these methods face limitations when applied to casually captured everyday videos.

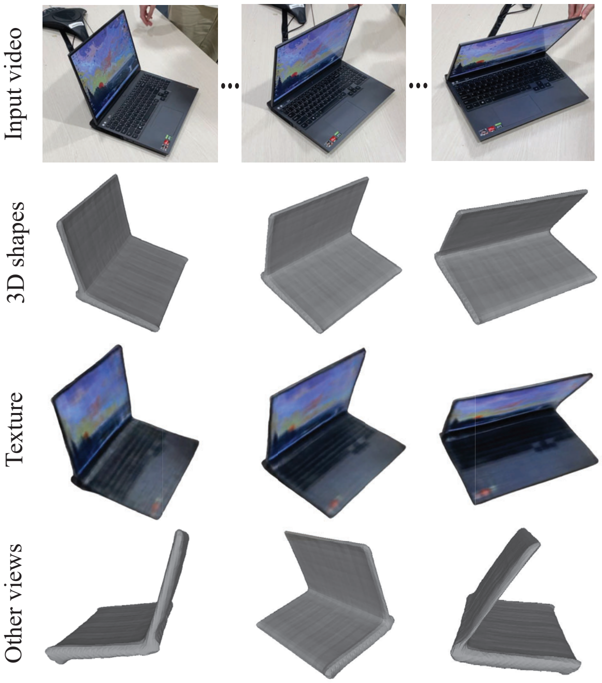

Figure 1. Given a single casual video capturing a piece-wise rigid general articulated object, REACTO can model the 3D shape, texture, and motion. The second row presents shape reconstruction results from reference views, the third row showcases the reconstructed texture, and the fourth row displays the shapes from another view.

Previous research [11, 14, 15, 61] on reconstructing articulated objects from monocular videos has primarily focused on humans and quadrupeds, utilizing readily available parametric models like SMPL [29] and SMAL [80], while neglecting the diverse range of everyday objects we commonly encounter and use. Non-parametric methods, like BANMo [70], MoDA [52], and PPR[73], utilizing dynamic volumetric neural radiance fields to model deformable objects. These methods are predominantly optimized for non-rigid, deformable subjects such as humans and animals, whose movable parts such as arms and legs are distinctly separated. However, the rigid movable components of general objects often sit adjacent to each other in their usual poses. For instance, consider the blades of a pair of scissors, which come into close proximity during use. This presents a considerable challenge to the previously mentioned methods with blend skinning techniques used for motion modeling, often leading to incorrect motions and artifacts.

Specifically, BANMo [70] utilizes neural Linear Blend Skinning (LBS) as the deformation model, while LBS is efficient and straightforward, it can sometimes lead to unrealistic deformations, resulting in substantial defects like candy wrapper artifacts and volume loss. MoDA [52] proposes to use Neural Dual Quaternion Blend Skinning (NeuDBS) to relieve these issues. Although NeuDBS offers improvements in handling rotations and preserving volume, it can still lead to the generation of unsmooth, less refined surfaces. PPR [73] incorporates Dual Quaternion Blend Skinning (DBS) along with novel skin losses and a more stable eikonal loss [10] to enhance the overall surface smoothness. However, it is observed that surfaces of one moving part tend to tear and get drawn towards another, with visible seam artifacts. Furthermore, the joints appear over-smoothed, leading to a loss of geometric precision. These defects likely arise from inaccurately assigned skinning weights.

In this work, we present **REACTO** to REconstruct general ArtiCulaTed Objects from a single casually captured monocular video. Methods with conventional blend skinning techniques, like SMPL [29], define their rig on the joints, note that some methods like BANMo [70] refer to joints as bones. In this case, it has been observed that the reconstructed shape of each rigid component can be bent by two joints, sometimes leading to seam artifacts. To address this, we define the rig on the bones. As depicted in Figure 2, our method optimizes the placement of bones to be near the centroid of each component, effectively enhancing the rigidity and motion integrity of these components.

As discussed previously, the defects also stem from inaccurately assigned skinning weights. For each rigid component, these problems can be addressed by implementing Rigid Skinning (RS), where each vertex is exclusively linked to a single bone. However, RS fails in modeling deformations near joints and can also lead to unwanted discontinuities. To overcome this, we propose Quasi-Rigid Blend Skinning, which merges the rigidity of RS with the flexibility of DBS. Specifically, we optimize the skinning weights on rigid components to be quasi-sparse, minimizing the influence from other bones and ensuring a strong association with their corresponding bone, thus displaying characteristics of rigid skinning. Concurrently, points near joints retain the adaptability inherent in DBS. The accuracy of the commonly used Mahalanobis distance [68, 70] is often compromised because its calculation relies on the precision of bone properties, including center, orientation, and scale, all of which are optimized during training. Consequently, we utilize geodesic distance as a more effective measure to jointly ascertain the appropriate corresponding bone for each point or to determine if the point is part of a joint. We demonstrate through experiments that REACTO consistently produces 3D shapes with higher-fidelity details compared to previous state-of-the-art approaches [52, 70, 73].

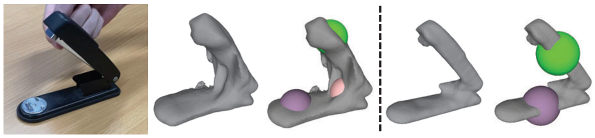

Figure 2. **Rig on joints vs. rig on bones.** A straightforward approach to control the motion of general articulated objects is to adopt methods [73] used for modeling humans or animals, which typically define the rig based on joints. This design can lead to bending shapes and corrupted motion. In contrast, we propose a novel approach by defining the rig based on bones, enhancing the rigidity and motion integrity of each component.

We summarize our contributions as:

- We present REACTO, a novel approach for modeling general articulated 3D objects from single casual videos, without complete views of the objects and any 3D supervision. REACTO demonstrates superior performance over current methods on both real and synthetic datasets.
- We redefine the rigging structure in our approach by placing rigs on the bones instead of joints, enhancing the rigidity and motion integrity of each component in general articulated objects.
- We propose Quasi-Rigid Blend Skinning (QRBS), a hybrid technique that harmonizes the rigidity of Rigid Skinning with the flexibility of Dual Quaternion Blend Skinning, empowered by quasi-sparse skinning weights, and geodesic point assignment for precise motion reconstruction of general articulated objects.

2. Related Work

Modeling articulated objects. In the field of computer vision, previous research on articulated deformations has predominantly concentrated on human and animal subjects [6, 18, 29, 32, 37, 48, 51, 53, 65, 66, 80], with less attention given to the modeling of general articulated objects that exhibit piece-wise rigidity. Building on the advancements in implicit representations like DeepSDF [41], A-SDF [36] models category-level articulation by introducing distinct shape and articulation codes. It integrates joint angles into the shape code, thereby learning to map these angles to their corresponding deformed shapes. ANCSH [25] introduces normalized articulated object coordinate space to model the canonical representations of articulated objects at the category level. CAPTRA [62] presents a unified framework for online pose tracking of both rigid and articulated objects from point cloud sequences. Ditto [17] predicts motion and geometry across object categories using two 3D point clouds, but it struggles with generalizing to new categories and detailed appearance reconstruction. StrobeNet [79] extends the previous works to reconstruct articulated objects from multi-view images. However, these methods require ground-truth 3D data for processing or training. CARTO [13] and NASAM [59] can model articulated objects without 3D ground truth but require category-specific training data with multiple objects. PARIS [27], designed for self-supervised learning of articulation, can generalize to new objects but relies on complete multi-view images, limiting its applicability to casually captured videos.

Shape reconstruction from images or videos. Various methods have been developed to learn 3D reconstruction from images or videos, guided by annotations like 3D key points [19, 23], optical flow [63], and semantic mask [9, 24, 77]. However, the models suffer in generalization since they heavily rely on prior shape templates. Neural implicit surface representations [39, 49, 50, 58, 75] have found extensive use in reconstructing images or videos. Works like [12, 38] have focused on reconstructing rigid objects from videos, however, they fall short in modeling articulated and deformable objects. Recent advancements, including LASR [68] and ViSER [69], have made strides in optimizing a single 3D deformable model from a monocular video guided by mask and optical flow, yet the reconstructed motion often presents unrealistic artifacts. Several studies [3, 16, 21, 22, 28, 37, 45, 55, 56, 78] have explored reconstructing shape and appearance from images or videos relied on neural radiance fields (NeRF) [33]. In this study, we model general articulated objects from a single video, employing a canonical Neural Radiance Field (NeRF) for shape and appearance, coupled with a deformation model that facilitates the transformation of 3D points between observation and canonical spaces.

Neural representations for dynamic scenes. Several recent studies have focused on developing deformation models that characterize dynamic scenes by transforming 3D points between the observation space and the canonical space. NR-NeRF [57] depicts deformations on non-rigid objects by learning a rigidity network. D-NeRF [46] is designed to transform points to the canonical space by learning a displacement, while NSFF [26] displaces 3D points utilizing scaled scene flow. Additionally, Nerfies [42] and HyperNeRF [43] define deformation by employing a learned dense SE(3) field. These approaches, however, tend to struggle with large motions between foreground objects and their backgrounds. To address these challenges, several works [16, 37, 44, 61] employ the parametric 3D human models, such as SMPL [29], while other methods [28, 44, 45] utilize synchronized multi-view video inputs. BANMo [70], MoDA [52], RAC [72], Total-Recon [54] and PPR[73] can reconstruct 3D shapes from casual videos without relying on human or animal models, by adopting linear blend skinning or dual quaternion blend skinning to learn the deformation model. However, these methods often result in notable artifacts when applied to general articulated objects. To solve this problem, we propose quasi-rigid blend skinning (QRBS) to model the motion of general articulated objects.

3. Method

The overview of our approach is illustrated in Figure 3. In this work, we undertake the task of modeling a 3D articulated object from a single video, employing a canonical Neural Radiance Field (NeRF) as the basis for our shape and appearance model (Section 3.1). Additionally, our approach includes a deformation model (Section 3.2) that transforms 3D points between observation and canonical spaces. Traditional methods like linear blend skinning or dual quaternion blend skinning, typically used for human or animal motion modeling, are inadequate for capturing motion in general articulated objects with multiple rigid components. To overcome this, we introduce Quasi-Rigid Blend Skinning (QRBS) as our deformation model, providing a more apt solution for modeling the motion of such objects. The models are then optimized using volume rendering (Section 3.3).

3.1. Canonical NeRF for shape and appearance

We first define the canonical NeRF [33] to model the shape and appearance of an articulated object. As in BANMo [70], we learn the color and density of a 3D point $\mathbf{X}^* \in \mathbb{R}^3$ in the canonical space,

$$\mathbf{c}^t = \mathbf{MLP}_{\text{color}}(\mathbf{X}^*, \mathbf{D}^t, \boldsymbol{\psi}_a^t), \tag{1}$$

$$\sigma = \Phi_\beta(\mathbf{MLP}_{\text{SDF}}(\mathbf{X}^*)), \tag{2}$$

where $\mathbf{MLP}_{\text{color}}$ and $\mathbf{MLP}_{\text{SDF}}$ are multi-layer perceptron (MLP) networks, $\mathbf{D}^t = (\phi^t, \theta^t)$ is the time-varying view direction and $\boldsymbol{\psi}_a^t$ is a 64-dimensional latent appearance code, serving to encode variations in appearance [31]. To perform volume rendering as [33], we follow [58, 76] to use the Cumulative Distribution Function $\Phi_\beta(\cdot)$ of the Laplace distribution with zero mean and β scale to convert signed distances into density. Here, β is a learnable parameter that controls the solidness of the object.

3.2. Quasi-rigid blend skinning for deformation

With the 3D point $\mathbf{X}^*$ in the canonical space and $\mathbf{X}^t$ in the observation space, we achieve 3D deformation between them via the deformation model. The canonical-to-observation and observation-to-canonical deformation at time t are denoted as $\mathcal{D}^{t,c\longrightarrow o}$ and $\mathcal{D}^{t,o\longrightarrow c}$ respectively.

Motion representation. For the motion of articulated objects, it encompasses global-level transformations $\mathbf{T}_{\text{global}} \in \text{SE}(3)$ and object-level articulation $\mathbf{T}_{\text{obj}} \in \mathbb{R}^8$ represented

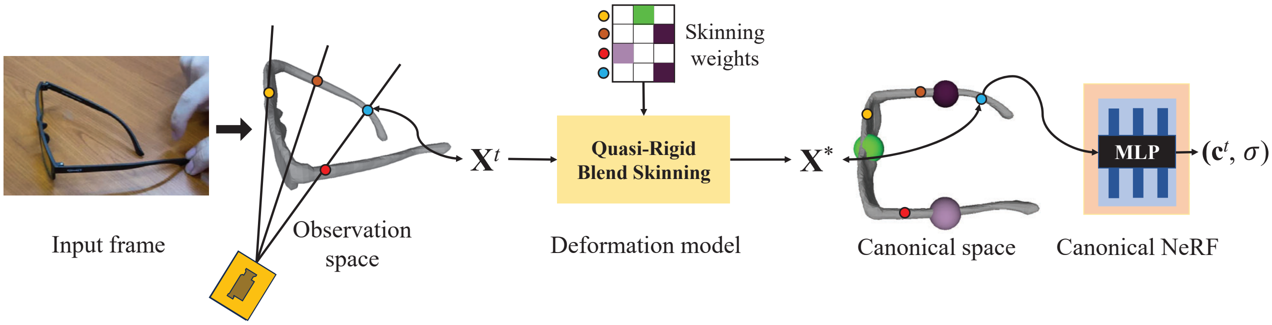

Figure 3. **The overview of REACTO.** We model an articulated 3D object from a single video using a shape and appearance model based on a canonical Neural Radiance Field (NeRF) and a deformation model for transforming 3D points between the observation space and the canonical space. Instead of linear blend skinning or dual quaternion blend skinning designed for human or animal motion modeling, we propose Quasi-Rigid Blend Skinning (QRBS) as our deformation model, with the learned quasi-sparse skinning weights, to accurately transform $\mathbf{X}^t$ from the observation space to $\mathbf{X}^*$ in the canonical space. We visualize the 3 bones for glasses in the canonical space. The colors in skinning weights signify the assigned bone for each point.

by a unit dual quaternion. Given the 3D point $\mathbf{X}^*$ in the canonical space and $\mathbf{X}^t$ in the observation space, we can deform one to the other via

$$\mathbf{X}^t = \mathcal{D}^{t,c\rightarrow o}(\mathbf{X}^*) = \mathbf{T}^t_{\text{global}}\mathbf{T}^{t,c\rightarrow o}_{\text{obj}}\mathbf{X}^*, \tag{3}$$

$$\mathbf{X}^* = \mathcal{D}^{t,o\rightarrow c}(\mathbf{X}^t) = \mathbf{T}^{t,o\rightarrow c}_{\text{obj}}(\mathbf{T}^t_{\text{global}})^{-1}\mathbf{X}^t, \tag{4}$$

where $\mathbf{T}_{\text{global}}$ comprises camera pose transformations $\mathbf{T}_{\text{cam}}$ and root body transformations $\mathbf{T}_{\text{root}}$, both modeled as per-frame SE(3) transformations represented by MLP networks. A detailed introduction to the object-level articulation of general articulated objects will be provided in the following.

Bone definition. When modeling the object-level motion of general articulated objects, such as a stapler in Figure 2, a straightforward design is to follow the previous methods that model the motion of humans and animals from videos. In this design, the motion of a stapler is considered analogous to the arm of a human, and the rig is defined on three joints to control the stapler's motion. The joints will be optimized to align with the positions at the ends of the object's parts to minimize energy as illustrated in the middle of Figure 2 (PPR).

Applying PPR [73] in such a design results in noticeable artifacts like bending shapes and corrupted motion, which is unacceptable for objects characterized by multiple rigid components. To address this limitation, we propose to define the rig on the bones, ideally the part centroids as illustrated in Figure 2 (Ours). Consequently, each rigid part is strongly associated with one bone, effectively defining the motion of the articulated objects. The number of bones, denoted as B, depends on the number of rigid components in an articulated object.

Skinning weights. We define the skinning weights as $\mathbf{W} = \{w_0, ..., w_{B-1}\} \in \mathbb{R}^B$. Given a 3D point $\mathbf{X}$, we compute the Gaussian skinning weights [52, 70] based on the Mahalanobis distance $d_M(\mathbf{X})$ between 3D points and the Gaussian bones,

$$d_M(\mathbf{X}) = (\mathbf{X} - \mathbf{O})^T\mathbf{V}^T\mathbf{\Lambda}^0\mathbf{V}(\mathbf{X} - \mathbf{O}), \tag{5}$$

where $\mathbf{O} \in \mathbb{R}^{B\times 3}$ are bone centers, $\mathbf{V} \in \mathbb{R}^{B\times 3\times 3}$ are bone orientations and $\mathbf{\Lambda}^0 \in \mathbb{R}^{B\times 3\times 3}$ are diagonal scale matrices. Each Gaussian bone has three parameters for center, orientation, and scale respectively, which are all optimized during training. To further refine the Gaussian skinning weights, we incorporate delta skinning weights learned by an MLP,

$$\mathbf{W} = \text{softmax}(d_M(\mathbf{X}) + \mathbf{W}_\Delta), \tag{6}$$

where $\mathbf{W}_\Delta = \mathbf{MLP}_{\text{skin}}(\mathbf{X}_{\text{bone}})$ is the delta skinning weights. $\mathbf{X}_{\text{bone}} \in \mathbb{R}^{B\times 3}$ denotes the relative positions of point $\mathbf{X}$ in the bone coordinates.

However, given that general articulated objects are typically piece-wise rigid, the refined Gaussian skinning weights $\mathbf{W}$ may introduce redundant associations to multiple bones for each point, thus hampering the rigidity of the parts. Therefore, we aim to make the skinning weights quasi-sparse to minimize the influence of other bones and ensure a strong association with their corresponding bone for 3D points. We first introduce a temperature factor γ to the calculation of $\mathbf{W}$ to stimulate the sparsity,

$$\mathbf{W}^s = \text{softmax}(\frac{d_M(\mathbf{X}) + \mathbf{W}_\Delta}{\gamma}). \tag{7}$$

Geodesic point assignment. We further propose a geodesic point assignment process to help correctly assign each point to the corresponding bone or joint, hence preventing surface tearing and corrupted motion. We can further enhance the sparsity of the skinning weights for the points in the rigid parts while keeping the weights unchanged

Algorithm 1 Geodesic point assignment

```
Input: Point assignment $\mathbf{M} = \mathbf{0} \in \mathbb{R}^B$, Mahalanobis distance $d_M^i$ and $d_M^j$, geodesic distance $d_G^i$ and $d_G^j$, bone index $i$ and $j$, hyperparameters $\eta$, $\zeta$.
Output: Updated assignment $\mathbf{M}$
if $d_M^i / d_M^j < 1 - \eta$ then
    $\mathbf{M}[i] \leftarrow 1$
else if $\frac{|d_G^i - d_G^j|}{\min(d_G^i, d_G^j)} < \zeta$ then
    $\mathbf{M}[i], \mathbf{M}[j] \leftarrow \mathbf{1}$        ▷ Assigning to joints
else
    $\mathbf{M}[\text{argmin}(d_G)] \leftarrow 1$
end if
```

for points in the joints based on the assignment, therefore, achieving quasi-rigid blend skinning.

The accuracy of Mahalanobis distance calculation depends on the precision of bone properties, including center, orientation, and scale. However, since these properties are all optimized during training, uncertainties are introduced into the calculation process. This can lead to inaccurate point assignments, as observed in our experiments. For instance, a point near the surface of one component may have a shorter Mahalanobis distance to a bone belonging to a different component.

Additionally, in cases where a point exhibits similar Mahalanobis distances to multiple bones, determining whether the point is associated with joints or rigid components remains challenging. To address these issues, we employ geodesic distance as depicted in Figure 4, which provides a measure of the shortest path between two points along a mesh surface.

To elaborate, for a point $\mathbf{X}$, we initially set up a point assignment vector $\mathbf{M} = \mathbf{0} \in \mathbb{R}^B$, representing the assignment of the point to its corresponding bone. We proceed by identifying the nearest bone b_i and the second nearest bone b_j to $\mathbf{X}$, based on their Mahalanobis distances d_M^i and d_M^j. Given that geodesic distance calculations require a mesh surface, we first extract a canonical mesh using the marching cubes algorithm [30]. Following this, we employ the KNN algorithm to locate the nearest vertices $\hat{\mathbf{X}}$, $\hat{b}_i$, and $\hat{b}_j$ relative to the point $\mathbf{X}$ and the centers of bones b_i and b_j. Finally, we compute the geodesic distances d_G^i and d_G^j from $\hat{\mathbf{X}}$ to $\hat{b}_i$ and $\hat{b}_j$, respectively, utilizing the exact geodesic algorithm as described in [34].

As illustrated in Algorithm 1, if the ratio of Mahalanobis distances d_M^i / d_M^j is less than $1 - \eta$, which means the point is obviously closer to b_i, we assign a value of 1 to the i^{th} element of $\mathbf{M}$. If the point is close to both bones, we check if $\frac{|d_G^i - d_G^j|}{\min(d_G^i, d_G^j)} < \zeta$, which means the geodesic distances are close. If the Mahalanobis distance and geodesic distance from the points to the bone i and j are both similar, we assign the value 1 to the i^{th} and the j^{th} elements of $\mathbf{M}$, as assigning the point to joints. If neither of the previous conditions is satisfied, we assign the point to the bone with the shortest geodesic distance. The distances are all passed through a softmax layer before being input into Algorithm 1.

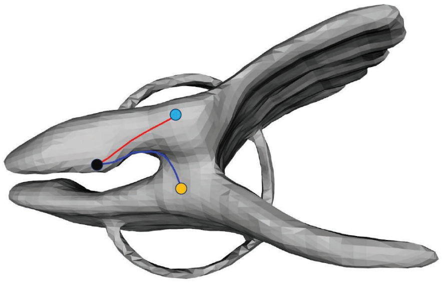

Figure 4. **Geodesic distances between 3D point and bones.** Geodesic distance can correctly associate the 3D point (black) with the top bone (blue) rather than the bottom bone (yellow) by following the shortest path on the mesh surface. Shorter distances indicate stronger associations.

As the mesh evolves during training, we refrain from applying the point assignment directly to the skinning weights as a mask. Instead, we impose penalties on the weights associated with bones not corresponding to the targeted point with a sparse skinning loss,

$$\mathcal{L}_{sparse} = \frac{\sum \left\|\mathbf{W}^s \odot \bar{\mathbf{M}}\right\|^2}{\sum \bar{\mathbf{M}}}, \tag{8}$$

where $\odot$ denotes Hadamard product and $\bar{\mathbf{M}} = 1 - \mathbf{M}$, since $\mathbf{M}$ indicates the correct assignment and we want to penalize the weights everywhere else. For points that have been assigned to joints, the skinning weights are not penalized.

Quasi-rigid blend skinning. With the learned quasi-sparse skinning weights, the articulated motion of 3D points under pose $\boldsymbol{\psi}_p$ can be obtained using our quasi-rigid blend skinning (QRBS):

$$\mathbf{X}(\boldsymbol{\psi}_\mathbf{p}) = \mathbf{T}_{\text{obj}}\mathbf{X} = (\sum_{b=0}^{B-1} w_b^S \mathbf{T}_b)\mathbf{X}, \tag{9}$$

where $\{\mathbf{T}_0, ..., \mathbf{T}_{B-1}\} = \mathbf{MLP}_{\text{pose}}(\boldsymbol{\psi}_p)$ and are represented by dual quaternions. This articulated motion is invertible by inverting $\mathbf{T}_b$ in Equation (9) and recomputing the skinning weights in Equation (7). We utilize a 3D cycle loss [26, 70] to supervise this invertible process.

3.3. Volume rendering and optimization

Volume rendering. We use the volume rendering in NeRF [33] to synthesize images. With the pixel location $\mathbf{x}^t \in \mathbb{R}^2$, the n-th sampled point along the ray that originates from $\mathbf{x}^t$ is $\mathbf{X}_n^t$. The color and opacity are given by:

$$\mathbf{c}(\mathbf{x}^t) = \sum_{n=1}^{N} \tau_n \mathbf{c}_n^t, \quad \mathbf{o}(\mathbf{x}^t) = \sum_{n=1}^{N} \tau_n, \tag{10}$$

where $\tau_n = \alpha_n \prod_{m=1}^{n-1}(1-\alpha_m)$, $\alpha_n = 1 - \exp(-\sigma_n \delta_n)$, N is the number of sampled points, δ_n is the distance between the n-th point and the next, and σ_n is the density in Equation (2).

Optimization. Except for the sparse skinning loss in Equation (8), we optimize our models with multiple reconstruction losses (color, object mask, optical flow, pixel features) that are similar to existing methods [52, 70, 73]. These losses are employed to minimize the difference between the predicted results and the observed ones, alongside regularization terms.

$$\mathcal{L} = \mathcal{L}_{rgb} + \mathcal{L}_{mask} + \mathcal{L}_{flow} + \mathcal{L}_{feature} + \mathcal{L}_{sparse} + \mathcal{L}_{reg}. \tag{11}$$

The predicted mask, optical flow, and pixel features are obtained from off-the-shelf methods [40, 67, 74]. Please refer to the supplementary materials for the regularization terms.

4. Experiments

4.1. Dataset, metrics, and implementation details

Real-world videos. To demonstrate the effectiveness of REACTO, we conducted evaluations on real-world videos with only partial views of different articulated objects such as laptops, staplers, scissors, faucets, nail clippers, glasses, and more. These videos were captured using a phone camera with no control over camera movements. For detailed information about the videos, please refer to the supplementary materials. In the preprocessing stage, we employed methodologies outlined in Lab4D [71]. Specifically, we utilized Track Anything [74] for predicting object silhouettes, VCN-robust [67] for optical flow estimation, and DINOv2 [40] for extracting pixel features. Additionally, we also annotate sparse camera poses (approximately 4 annotations per video) for camera estimation. These annotations serve as initialization and will be further optimized during the training process. To distinguish from synthetic data, we prepend *real-* to articulated objects (e.g., *real-laptop*).

Synthetic video. To evaluate our method quantitatively, we render videos using PartNet-Mobility dataset [2, 35, 64] that provide ground truth meshes. We chose 3 categories for evaluation in this paper, namely *USB*, *stapler*, and *scissors*. For more results from other categories, please refer to the supplementary material. For each articulated object, we render 100 frames with the camera moving through a 120-degree azimuthal angle and a 30-degree polar angle using Blender [1]. The sequence consists of 50 frames corresponding to 50 consecutive articulations, followed by another 50 frames in reverse order. We train the synthetic dataset with ground truth object silhouettes. Similar to the process for real-world videos, we utilize VCN-robust [67] and DINOv2 [40] for predicting optical flow and pixel features, respectively. The initial camera poses are obtained in the same manner as those for real-world videos.

Metrics. To quantitatively evaluate various methods, we employ Chamfer distance (CD) [8] and F-scores as our metrics. For CD, lower values indicate better performance. F-scores are compared across different methods at distance thresholds $d = 10\%$ and $d = 5\%$. A higher F-score is better. As the ground truth meshes from PartNet-Mobility dataset [64] exhibit limited vertices and uneven distribution, we uniformly sample $10,000$ points using PyTorch3D [47] from both predicted and ground truth meshes to compute Chamfer Distance (CD) and F-scores, which ensures a fair and robust evaluation.

Implementation details. We employ the AdamW optimizer to optimize the model for 4,000 iterations. For all objects, we start with the same shape of a unit sphere as PPR [73]. The reconstructed meshes are extracted using marching cubes on a 128^3 grid. For additional implementation details, please refer to the supplementary materials.

4.2. Comparison with state-of-art methods

Baselines. We compare our method with BANMo [70], MoDA [52] and PPR [73]. These methods were originally designed for modeling humans or animals from videos. For the deformation model, BANMo employs linear blend skinning, while MoDA and PPR utilize dual quaternion blend skinning (note that the learning of dual quaternion differs between MoDA and PPR).

To ensure fair comparisons, we report the results of BANMo, MoDA, and PPR with rigging on bones in this section. Each method utilizes 64 sampled points per ray to ensure consistent evaluation conditions. We supply BANMo, MoDA, and PPR with the same initial camera poses.

Results. The qualitative and quantitative results are presented in Figure 5 and Table 1, respectively. In Figure 5, both BANMo and MoDA struggle to reconstruct the complete shape of articulated objects. This is evident in instances such as both methods facing difficulties with *real-faucet* and BANMo also encountering challenges with *real-scissors*. These two methods often yield non-smooth surfaces, as observed with BANMo on *real-stapler*, MoDA on *real-scissors*, and both methods on *real-laptop*. Although PPR generates smoother surfaces compared to BANMo and MoDA, it still encounters challenges in accurately modeling the motion of articulated objects. Notably, it introduces surface tearing artifacts in cases such as *real-stapler* and *real-scissors*. Also, we observe over-smoothed joints in *real-faucet*, *real-stapler*, and *real-laptop*. Furthermore, when applied to *real-faucet* and *real-laptop*, PPR demonstrates inaccuracies in modeling motions, such as the rotation of the handle in *real-faucet* and the folding motion of *real-laptop*. In contrast, our REACTO

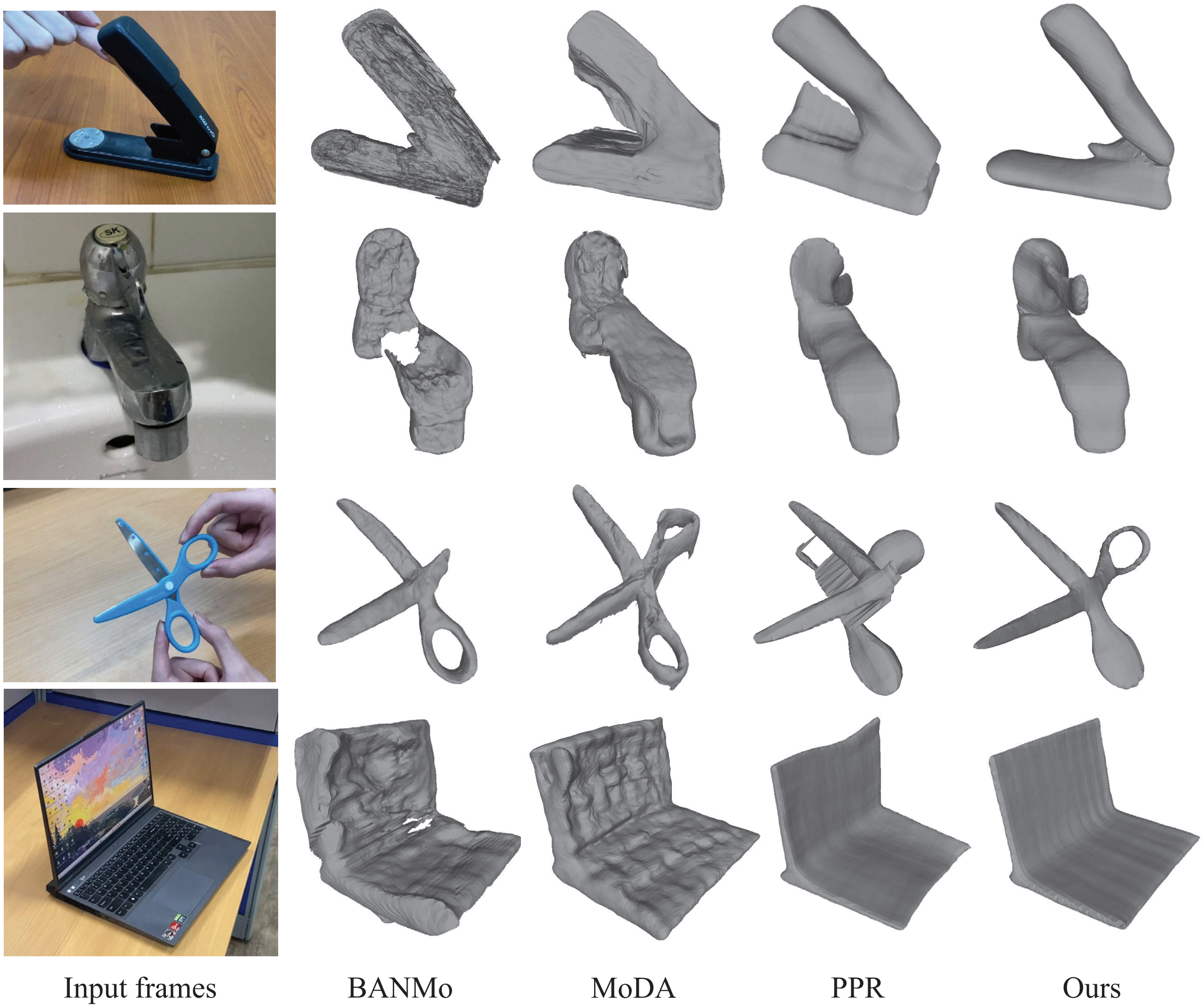

Figure 5. **Qualitative comparison of our method with BANMo [70], MoDA [52] and PPR [73].** BANMo and MoDA struggle with complete shape reconstruction (*real-faucet*, *real-scissors*). Non-smooth surfaces (BANMo on *real-stapler*, MoDA on *real-scissors*, BANMo and MoDA on *real-laptop*) are also observed. The results of PPR are smoother but with surface tearing (*real-stapler*, *real-scissors*), over-smoothed joints (*real-faucet*, *real-laptop*, *real-stapler*), and inaccuracies in motion modeling (*real-faucet*, *real-laptop*). In contrast, REACTO outperforms these methods, excelling in the shape and deformation reconstruction of articulated objects. Please find the video results in the supplementary material.

consistently outperforms these methods, with superior capabilities in modeling the shape and deformation of various articulated objects.

Our quantitative results support qualitative observations, demonstrating that REACTO outperforms all baselines across all metrics on the synthetic data.

4.3. Ablation study on deformation models

In this section, we compare our quasi-rigid blend skinning with other deformation models employed for articulated object motion, such as displacement field in NASAM [59] and invertible Real-NVP [7] in CaDeX [20].

The qualitative and quantitative results are presented in Figure 6 and Table 2, respectively. For the synthetic *USB*, both displacement field and Real-NVP struggle to accurately distinguish the motion of the two rigid parts. In contrast, our method successfully models the motion with the optimized rigging system. For *real-nail clipper*, the displacement field still fails to separate the two rigid parts. Real-NVP introduces non-smoothness during motion, while our method maintains a consistently smooth mesh surface. The quantitative results on synthetic data further confirm that our quasi-rigid blend skinning offers a more reasonable approach than other deformation models for modeling the motion of general articulated objects.

Table 1. **Quantitative comparison between different methods.** Our method has better performance than BANMo [70], MoDA [52], and PPR [73] across all metrics.

Method	USB			stapler			scissors		
	CD(↓)	F(10%, ↑)	F(5%, ↑)	CD(↓)	F(10%, ↑)	F(5%, ↑)	CD(↓)	F(10%, ↑)	F(5%, ↑)
BANMo	20.3	65.1	45.0	19.1	57.8	32.8	19.9	66.8	41.4
MoDA	17.1	74.9	49.5	18.8	64.2	40.3	14.8	77.7	42.3
PPR	20.7	65.7	38.9	16.8	67.5	40.0	16.1	71.4	39.9
Ours	**15.3**	**78.6**	**51.5**	**14.3**	**75.5**	**42.7**	**14.0**	**78.2**	**43.9**

Table 2. **Quantitative ablation studies on deformation models.** Our method outperforms the displacement field, Real-NVP, and rigid skinning across various data.

Method	USB			stapler			scissors		
	CD(↓)	F(10%, ↑)	F(5%, ↑)	CD(↓)	F(10%, ↑)	F(5%, ↑)	CD(↓)	F(10%, ↑)	F(5%, ↑)
Displacement	19.7	59.5	28.2	17.9	63.8	30.8	19.1	58.6	30.9
Real-NVP	17.6	70.7	47.2	16.0	70.4	32.6	19.6	63.6	32.7
Rigid	16.3	73.5	49.3	15.1	72.8	41.1	14.8	76.4	43.7
Ours	**15.3**	**78.6**	**51.5**	**14.3**	**75.5**	**42.7**	**14.0**	**78.2**	**43.9**

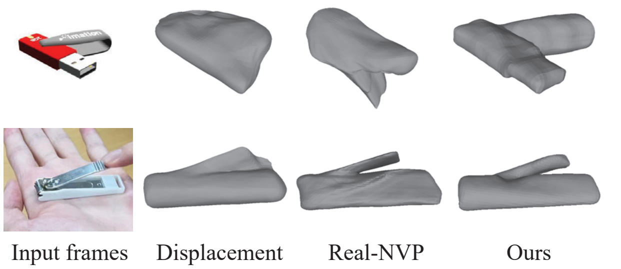

Figure 6. **Ablation study on deformation models.** We compare displacement field [59] and Real-NVP [7, 20] with our QRBS on synthetic *USB* and *real-nail clipper*. The displacement field struggles to accurately separate the motion of the two rigid parts in both *USB* and *real-nail clipper*. Real-NVP also fails to separate the two rigid parts of *USB* and produces non-smoothness when modeling the motion of *real-nail clipper*. In contrast, our QRBS consistently outperforms both methods in both cases.

Besides, we also propose a straightforward design for rigid skinning. For the skinning weights $w_b, b \in [0, B-1]$ ($w_b \in [0, 1], \sum_b w_b = 1$), we binaryize them by setting the largest w_b to 1 and all others to 0. As illustrated in Table 2, rigid skinning exhibits comparable performance with our method when evaluated on synthetic data. However, it may lead to seam artifacts, as exemplified in Figure 7, particularly noticeable in the leg of the glasses.

5. Conclusion

In this paper, we introduce REACTO, a groundbreaking method for reconstructing general articulated 3D objects from single casual videos, achieving enhanced modeling and precision by redefining rigging structures and employing Quasi-Rigid Blend Skinning. QRBS ensures the rigidity of each component while retaining smooth deformation on the joints by utilizing quasi-sparse skinning weights and geodesic point assignment. Extensive experiments show that REACTO outperforms existing methods in fidelity and detail on both real and synthetic datasets.

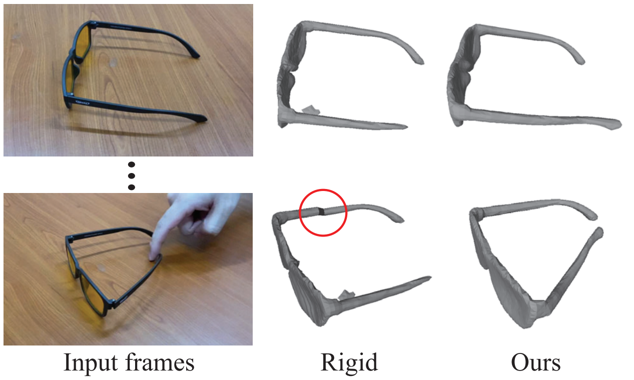

Figure 7. **Rigid skinning vs. Quasi-rigid blend skinning.** For rigid skinning, we binaryize skinning weights by setting the largest w_b to 1 and all others to 0, which fails to model the articulation while causing seam artifacts on the leg of *real-glasses* (in the red circle).

Limitations: As casual videos typically offer only partial views of objects, the quality of surface reconstruction may suffer on the unseen side. The limitations of this approach will be further detailed in the supplementary materials.

Acknowledgements

This research work is supported by the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021).

References

[1] Blender Online Community. Blender - a 3D modelling and rendering package, 2018. Blender Foundation, Stichting Blender Foundation, Amsterdam,. 6

[2] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 6

[3] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 3

[4] Cheng Chen, Xiaofeng Yang, Fan Yang, Chengzeng Feng, Zhoujie Fu, Chuan-Sheng Foo, Guosheng Lin, and Fayao Liu. Sculpt3d: Multi-view consistent text-to-3d generation with sparse 3d prior. *arXiv preprint arXiv:2403.09140*, 2024. 1

[5] Yiwen Chen, Zilong Chen, Chi Zhang, Feng Wang, Xiaofeng Yang, Yikai Wang, Zhongang Cai, Lei Yang, Huaping Liu, and Guosheng Lin. Gaussianeditor: Swift and controllable 3d editing with gaussian splatting. *arXiv preprint arXiv:2311.14521*, 2023. 1

[6] Boyang Deng, John P Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Nasa neural articulated shape approximation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 612–628. Springer, 2020. 2

[7] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016. 7, 8

[8] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017. 6

[9] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoint without keypoints. In *European Conference on Computer Vision*, pages 88–104. Springer, 2020. 3

[10] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 2

[11] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12858–12868, 2023. 1

[12] Philipp Henzler, Jeremy Reizenstein, Patrick Labatut, Roman Shapovalov, Tobias Ritschel, Andrea Vedaldi, and David Novotny. Unsupervised learning of 3d object categories from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4700–4709, 2021. 3

[13] Nick Heppert, Muhammad Zubair Irshad, Sergey Zakharov, Katherine Liu, Rares Andrei Ambrus, Jeannette Bohg, Abhinav Valada, and Thomas Kollar. Carto: Category and joint agnostic reconstruction of articulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21201–21210, 2023. 3

[14] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5605–5615, 2022. 1

[15] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *European Conference on Computer Vision*, pages 402–418. Springer, 2022. 1

[16] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video, 2022. 3

[17] Zhenyu Jiang, Cheng-Chun Hsu, and Yuke Zhu. Ditto: Building digital twins of articulated objects from interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5616–5626, 2022. 3

[18] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8320–8329, 2018. 2

[19] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 371–386, 2018. 3

[20] Jiahui Lei and Kostas Daniilidis. Cadex: Learning canonical deformation coordinate space for dynamic surface representation via neural homeomorphism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6624–6634, 2022. 7, 8

[21] Lingzhi Li, Zhen Shen, Li Shen, Ping Tan, et al. Streaming radiance fields for 3d video synthesis. In *Advances in Neural Information Processing Systems*. 3

[22] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhöfer, Jürgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. In *European Conference on Computer Vision*, pages 419–436. Springer, 2022. 3

[23] Xueting Li, Sifei Liu, Shalini De Mello, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. Online adaptation for consistent mesh reconstruction in the wild. *Advances in Neural Information Processing Systems*, 33:15009–15019, 2020. 3

[24] Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Self-supervised single-view 3d reconstruction via semantic consistency. In *European Conference on Computer Vision*, pages 677–693. Springer, 2020. 3

[25] Xiaolong Li, He Wang, Li Yi, Leonidas J Guibas, A Lynn Abbott, and Shuran Song. Category-level articulated object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3706–3715, 2020. 2

[26] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of

dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021. 3, 5

[27] Jiayi Liu, Ali Mahdavi-Amiri, and Manolis Savva. Paris: Part-level reconstruction and motion analysis for articulated objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 352–363, 2023. 1, 3

[28] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Transactions on Graphics (TOG)*, 40(6):1–16, 2021. 3

[29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 1, 2, 3

[30] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. page 163–169, New York, NY, USA, 1987. Association for Computing Machinery. 5

[31] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021. 3

[32] Marko Mihajlovic, Yan Zhang, Michael J Black, and Siyu Tang. Leap: Learning articulated occupancy of people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10461–10471, 2021. 2

[33] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3, 5

[34] Joseph SB Mitchell, David M Mount, and Christos H Papadimitriou. The discrete geodesic problem. *SIAM Journal on Computing*, 16(4):647–668, 1987. 5

[35] Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[36] Jiteng Mu, Weichao Qiu, Adam Kortylewski, Alan Yuille, Nuno Vasconcelos, and Xiaolong Wang. A-sdf: Learning disentangled signed distance functions for articulated shape representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13001–13011, 2021. 2

[37] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5762–5772, 2021. 2, 3

[38] David Novotny, Diane Larlus, and Andrea Vedaldi. Learning 3d object categories by looking around them. In *Proceedings of the IEEE international conference on computer vision*, pages 5218–5227, 2017. 3

[39] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021. 3

[40] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 6

[41] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2

[42] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 3

[43] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6), 2021. 3

[44] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021. 3

[45] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 3

[46] 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*, pages 10318–10327, 2021. 3

[47] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 6

[48] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 2

[49] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 3

[50] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for

high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 3

[51] Chaoyue Song, Jiacheng Wei, Ruibo Li, Fayao Liu, and Guosheng Lin. 3d pose transfer with correspondence learning and mesh refinement. *Advances in Neural Information Processing Systems*, 34:3108–3120, 2021. 2

[52] Chaoyue Song, Tianyi Chen, Yiwen Chen, Jiacheng Wei, Chuan Sheng Foo, Fayao Liu, and Guosheng Lin. Moda: Modeling deformable 3d objects from casual videos. *arXiv preprint arXiv:2304.08279*, 2023. 1, 2, 3, 4, 6, 7, 8

[53] Chaoyue Song, Jiacheng Wei, Ruibo Li, Fayao Liu, and Guosheng Lin. Unsupervised 3d pose transfer with cross consistency and dual reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[54] Chonghyuk Song, Gengshan Yang, Kangle Deng, Jun-Yan Zhu, and Deva Ramanan. Total-recon: Deformable scene reconstruction for embodied view synthesis. *arXiv preprint arXiv:2304.12317*, 2023. 3

[55] Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *arXiv preprint arXiv:2210.15947*, 2022. 3

[56] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *Advances in Neural Information Processing Systems*, 34:12278–12291, 2021. 3

[57] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12970, 2021. 3

[58] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 3

[59] Fangyin Wei, Rohan Chabra, Lingni Ma, Christoph Lassner, Michael Zollhöfer, Szymon Rusinkiewicz, Chris Sweeney, Richard Newcombe, and Mira Slavcheva. Self-supervised neural articulated shape and appearance models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15816–15826, 2022. 1, 3, 7, 8

[60] Jiacheng Wei, Hao Wang, Jiashi Feng, Guosheng Lin, and Kim-Hui Yap. Taps3d: Text-guided 3d textured shape generation from pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16805–16815, 2023. 1

[61] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pages 16210–16220, 2022. 1, 3

[62] Yijia Weng, He Wang, Qiang Zhou, Yuzhe Qin, Yueqi Duan, Qingnan Fan, Baoquan Chen, Hao Su, and Leonidas J Guibas. Captra: Category-level pose tracking for rigid and articulated objects from point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13209–13218, 2021. 2

[63] Shangzhe Wu, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Dove: Learning deformable 3d objects by watching videos. *arXiv preprint arXiv:2107.10844*, 2021. 3

[64] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. SAPIEN: A simulated part-based interactive environment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

[65] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6184–6193, 2020. 2

[66] Fan Yang, Tianyi Chen, Xiaosheng He, Zhongang Cai, Lei Yang, Si Wu, and Guosheng Lin. Attrihuman-3d: Editable 3d human avatar generation with attribute decomposition and indexing. *arXiv preprint arXiv:2312.02209*, 2023. 2

[67] Gengshan Yang and Deva Ramanan. Volumetric correspondence networks for optical flow. *Advances in neural information processing systems*, 32, 2019. 6

[68] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T Freeman, and Ce Liu. Lasr: Learning articulated shape reconstruction from a monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15980–15989, 2021. 2, 3

[69] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Ce Liu, and Deva Ramanan. Viser: Video-specific surface embeddings for articulated 3d shape reconstruction. *Advances in Neural Information Processing Systems*, 34:19326–19338, 2021. 3

[70] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2863–2873, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[71] Gengshan Yang, Jeff Tan, Alex Lyons, and Neehar Perci. Lab4d. `https://github.com/lab4d-org/lab4d`, 2023. 6

[72] Gengshan Yang, Chaoyang Wang, N Dinesh Reddy, and Deva Ramanan. Reconstructing animatable categories from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16995–17005, 2023. 3

[73] Gengshan Yang, Shuo Yang, John Z Zhang, Zachary Manchester, and Deva Ramanan. Ppr: Physically plausible reconstruction from monocular videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3914–3924, 2023. 1, 2, 3, 4, 6, 7, 8

[74] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos, 2023. 6

[75] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33:2492–2502, 2020. 3

[76] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 3

[77] Yufei Ye, Shubham Tulsiani, and Abhinav Gupta. Shelf-supervised mesh prediction in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8843–8852, 2021. 3

[78] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 3

[79] Ge Zhang, Or Litany, Srinath Sridhar, and Leonidas Guibas. Strobenet: Category-level multiview reconstruction of articulated objects. *arXiv preprint arXiv:2105.08016*, 2021. 3

[80] 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*, pages 6365–6373, 2017. 1, 2