Building Rearticulable Models for Arbitrary 3D Objects from 4D Point Clouds

Shaowei Liu[1] Saurabh Gupta[1*] Shenlong Wang[1*]
[1]University of Illinois Urbana-Champaign
https://stevenlsw.github.io/reart/

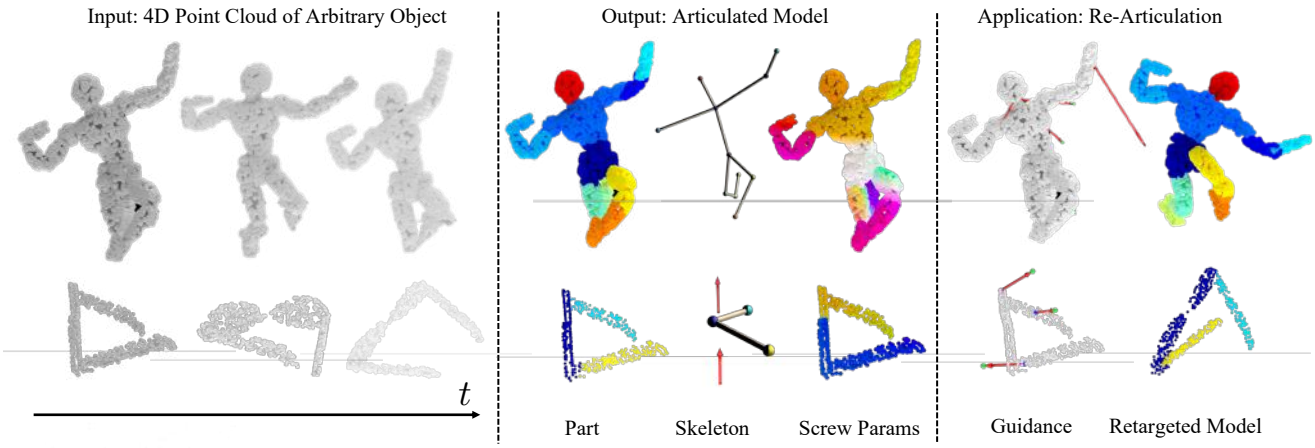


Figure 1. Given a short point cloud sequence of arbitrary articulated object **(left)**, our method outputs an animatable 3D model **(middle)**, which can be retargeted to novel poses with only a few sparse point correspondences **(right)**.

Abstract

We build rearticulable models for arbitrary everyday man-made objects containing an arbitrary number of parts that are connected together in arbitrary ways via 1 degree-of-freedom joints. Given point cloud videos of such everyday objects, our method identifies the distinct object parts, what parts are connected to what other parts, and the properties of the joints connecting each part pair. We do this by jointly optimizing the part segmentation, transformation, and kinematics using a novel energy minimization framework. Our inferred animatable models, enables retargeting to novel poses with sparse point correspondences guidance. We test our method on a new articulating robot dataset, and the Sapiens dataset with common daily objects. Experiments show that our method outperforms two leading prior works on various metrics.

1. Introduction

Consider the sequence of points clouds observations of articulating everyday objects shown in Figure 1. As humans, we can readily infer the *kinematic structure* of the underlying object, *i.e.* the different object parts and their connectivity and articulation relative with one another [21]. This paper develops computational techniques with similar abilities. Given point cloud videos of *arbitrary* everyday objects (with an *arbitrary* number of parts) undergoing articulation, we develop techniques to build animatable 3D reconstructions of the underlying object by a) identifying the distinct object parts, b) inferring what parts are connected to what other parts, and c) the properties of the joint between each connected part pair. Success at this task enables

Figure 2. Many man-made everyday objects can be explained with rigid parts connected in a kinematic tree with 1DOF joints.

*equal advising, alphabetic order

	Arbitrary Parts	Realistic Joint Constraints	Arbitrary Kinematics
Category-specific *e.g.*			
people [31]	no	yes	no
quadrupeds [51, 61]	no	yes	no
cartoons [53]	no	yes	no
DeepPart [57]	yes	no	no
NPP [11]	yes	no	no
ScrewNet [17]	yes	yes	no
UnsupMotion [42]	no	yes	no
Ditto [20]	yes	yes	no
MultiBodySync [15]	yes	no	no
WatchItMove [35]	yes	no	yes
Ours	yes	yes	yes

Table 1. Most past work on inferring rearticulable models is category specific. Building rearticulable models for arbitrary everyday man-made objects requires reasoning about arbitrary part geometries, arbitrary part connectivity, and realistic joint constraints (1DOF w.r.t. parent part). We situate past work along these 3 dimensions, and discuss major trends in Sec. 2.

rearticulation of objects. Given just a few user clicks specifying what point goes where, we can fill in the remaining geometry as shown on the right side in Figure 1.

Most past work on inferring how objects articulate tackles it in a *category-specific* manner, be it for people [31, 36, 40], quadrupeds [51, 61], or even everyday objects [34]. Category-specific treatment allows the use of specialized shape models (such as the SMPL model [31] for people), or defines the output space (*e.g.* location of 2 hinge joints for the category eye-glasses). This limits applicability of such methods to categories that have a canonical topology, leaving out categories with large intra-class variation (*e.g.* boxes that can have 1-4 hinge joints), or in-the-wild objects which may have an arbitrary number of parts connected in arbitrary ways (*e.g.* robots).

Only a very few past works tackle the problem of inferring rearticulable models in a category-agnostic manner. Huang *et al*. [15] only infer part segmentations, which by itself, is insufficient for rearticulation. Jiang *et al*. [20] only consider a single 1-DOF joint per object, dramatically restricting its application (think about a humanoid robot with four limbs, but the articulable model can only articulate one). Noguchi *et al*. [35] present the most complete solution but instead work with visual input and don't incorporate the 1DOF constraint, *i.e*. a part can only rotate or translate about a fixed axis on the parent part, common to a large number of man-made objects as can be seen in Fig. 2. Inferring 3DOF / 6DOF joints leads to unrealistic rearticulation and is thus undesirable (consider the leg of eyeglasses can freely move or rotate). Our work fills this gap in the literature. Our method extracts 3D rearticulable models for arbitrary everyday objects (containing an arbitrary number of parts that are connected together in arbitrary ways via 1DOF joint) from point cloud videos. To the best of our knowledge, this is the first work to tackle this specific problem.

Our proposed method jointly reasons about part geometries and their 1-DOF inter-connectivity with one another. At the heart of our approach is a novel continuous-discrete energy formulation that seeks to jointly learn parameters of the object model (*i.e*. assignments of points in the canonical view to parts, and the connectivity of parts to one another) by minimizing shape and motion reconstruction error (after appropriate articulation of the inferred model) on the given views. As it is difficult to directly optimize in the presence of continuous and discrete variables with structured constraints, we first estimate a *relaxed* model that infers parts that are free to follow an arbitrary 6DOF trajectory over time (*i.e*. doesn't require parts to be connected in a kinematic tree with 1DOF joints). We *project* the estimated relaxed model to a kinematic model and continue to optimize with the reconstruction error to further finetune the estimated joint parameters. Our joint approach leads to better models and improved rearticulation as compared to adaptations of past methods [15, 35] to this task.

2. Related Work

Building rearticulable models for arbitrary objects requires a) identifying the distinct parts, b) the kinematic topology that connects the different parts, and c) any constraints that there may be on the joints. This makes inferring articulable models from raw observation inputs challenging. Consequently, there are very few past works that tackle all these challenges to produce an end-to-end system for this task. A large body of work has focused on tackling subsets of these problems, or they operate in settings where some of these aspects are simplified. We overview past works with respect to these aspects in Tab. 1 and discuss major lines of work in more detail below.

A large body of work tackles this problem in a **category-specific** manner. This leads to a well specified definition of parts and their kinematic connectivity, and allows the use of sophisticated modeling. A good example is modeling for humans [31], human hands [36, 40], cartoon characters [53], and quadrupeds [51, 56, 61]. However, doing this for each new category requires manual effort and this approach doesn't scale to arbitrary objects that we consider in our work. Many recent works have pursued extending such approaches to other categories making assumptions about the number of parts [1, 34, 48], kinematic topology [2, 10, 54], articulation type or common geometric features [12, 27, 49]. Researchers have also tackled intermediate problems that are useful for eventual rearticulation in a category-specific manner, *e.g*. articulated pose estimation [9, 25, 27], motion prediction [14, 19, 26, 41], and 3D reconstruction [16, 23, 34].

Category-agnostic modeling, that is necessary to enable modelling of arbitrary objects, is considerably more

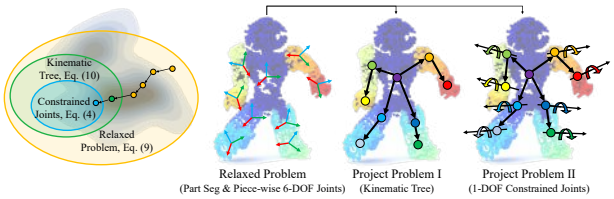


Figure 3. Overview: We formulate the problem of rearticulable object modeling from 4D point cloud as an energy minimization problem. The optimization is divided into a *relaxation* stage that reasons about a 6-DOF piece-wise rigid model without kinematic constraints and a *project* stage that casts the solution to a valid kinematic tree (all joints satisfy 1-DoF constraints).

challenging, and most past work only tackles subsets of the problem. A body of work focuses on unsupervised part segmentation, *i.e*. parsing articulated objects into multiple rigid moving parts. Early work addressed the problem by clustering and co-segmentation [13, 42, 44, 45, 58] by relying on motion and geometry, while more recent approaches [5, 15, 29, 46, 47, 55, 57, 59] use more expressive features extracted using a neural network. Recent work has also tackled pose estimation for arbitrary 1DOF joints [17, 20, 39] or for given kinematic trees [28].

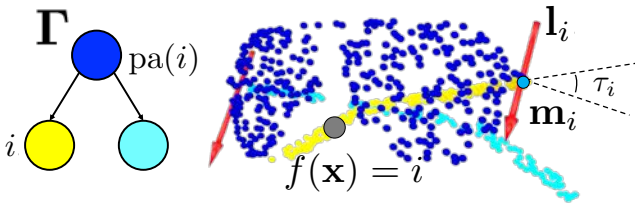
Closest to us are works from Huang *et al*. [15], Jiang *et al*. [20], and Noguchi *et al*. [35]. Huang *et al*. [15] build upon part segmentation work from Yi *et al*. [57] to output temporally-consistent part segmentation by solving a joint synchronization problem. This works very well for part segmentation. However, Huang *et al*. do not infer the kinematic tree connecting these parts together and parts can undergo 6DOF transformation relative to one another. Thus, their output, as is, falls short of the rearticulation goal. Jiang *et al*. [20] tackle rearticulation of generic objects and study the 1DOF nature of the joint. However, their formulation is limited to only infering a *single* 1DOF joint (*e.g*. laptops), and is thus incapable of analysing more complex objects that our approach is able to handle. Lastly, Noguchi *et al*. [35] fit articulated objects models to multi-view RGB video sequences and demonstrate impressive reanimation results. In contrast, we a) work with point-cloud input, b) model fine-grained shape for parts (as opposed to approximating them as ellipsoids), and c) additionally incorporate 1DOF constraint for joints as is common to everyday objects (as opposed to a general 3DOF spherical joints for [35]). To the best of our knowledge, our work is the first to simultaneously achieves all three desiderata for reanimation: inferring parts, kinematic connectivity, and joint constraints.

3. Approach

Our goal is to infer the articulated structures, parts, and joint parameters jointly, given a point cloud video. Given as input a point cloud sequence $\mathcal{P} = \{\mathbf{P}^t\}_{t\in 1,\dots T}$, our goal is to produce an animatable kinematic model with part definitions, part connectivity, joint parameters, and joint states at each time step.

We first introduce the parameterized articulated model $\mathbf{\Gamma}$ in Sec. 3.1. We then cast estimation of this articulated model as an energy minimization problem (Sec. 3.2). It turns out that directly optimizing this energy is difficult, thus we design a two-stage *relax and project* approach to optimizing this energy (Sec. 3.3), as shown in Fig. 3. We first solve a more tractable relaxed problem (estimating a 6DOf piece-wise rigid model without any kinematic constraints, Sec. 3.3.1), project the solution to a valid kinematic tree (Sec. 3.3.2) and further optimize the 1DOF joint parameters (Sec. 3.3.3).

3.1. Articulated Model



Our kinematic model $\mathbf{\Gamma}$ comprises of n parts that are connected to one another. Each part $i \in [1 \dots n]$ (except the root joint) is connected to its parent part $\mathrm{pa}(i) \in \mathbf{\Gamma}$ via a 1-DOF joint. Each joint, is either be a revolute joint or a prismatic joint parameterized by screw parameters [32], $\mathbf{s}_i = \{\mathbf{l}_i, \mathbf{m}_i\}$, specifying where the joint is located ($\mathbf{m}_i$) and its axis ($\mathbf{l}_i$, rotation axis for revolute joints or translation axis for prismatic joints) with respect to the parent part.

The articulation is controlled by a set of joint state parameters $\boldsymbol{\theta}^t = \{\boldsymbol{\theta}_i^t\}_{i\in 1,\dots n}$ where each joint parameter $\boldsymbol{\theta}_i^t = (\tau_i^t, d_i^t)$ represents the rotation τ or translation d along its axis for joint i at time step t.

Parts are represented as a partition of Euclidean space, $\mathbb{R}^3$, at a *canonical* time step c, *i.e*. a coordinate-based semantic field $f(\mathbf{x})$ that maps points $\mathbf{x} \in \mathbb{R}^3$ to a part label in $[1 \dots n]$. We use a coordinate-based neural network to parameterize the part segmentation field. This completes the definition of the articulable model.

Given a set of state parameters $\boldsymbol{\theta}^t$, we can deform a point cloud in canononical frame to the target pose through piece-wise rigid transform:

$$M(\boldsymbol{\theta}^t; \mathbf{\Gamma}, f) = \{\mathbf{T}^t_{f(\mathbf{x})} \cdot \mathbf{x}\}_{\mathbf{x}\in\mathbf{P}^c} \tag{1}$$

where the rigid transformation of each part is computed through rigid pose composition along the kinematic chain from the part to root:

$$\mathbf{T}_i^t = \prod_{i'\in \mathrm{ans}(i)} \mathbf{T}(\mathbf{s}_{i'}, \boldsymbol{\theta}_{i'}^t) \tag{2}$$

and $\mathrm{ans} = \{i, \mathrm{pa}(i), \mathrm{pa}(\mathrm{pa}(i)), \dots\}$ denotes the ordered set of ancestor nodes of i, which forms a kinematic chain.

Each joint's rigid transformation to its parent $\mathbf{T}(\mathbf{s}_i, \boldsymbol{\theta}_i^T)$

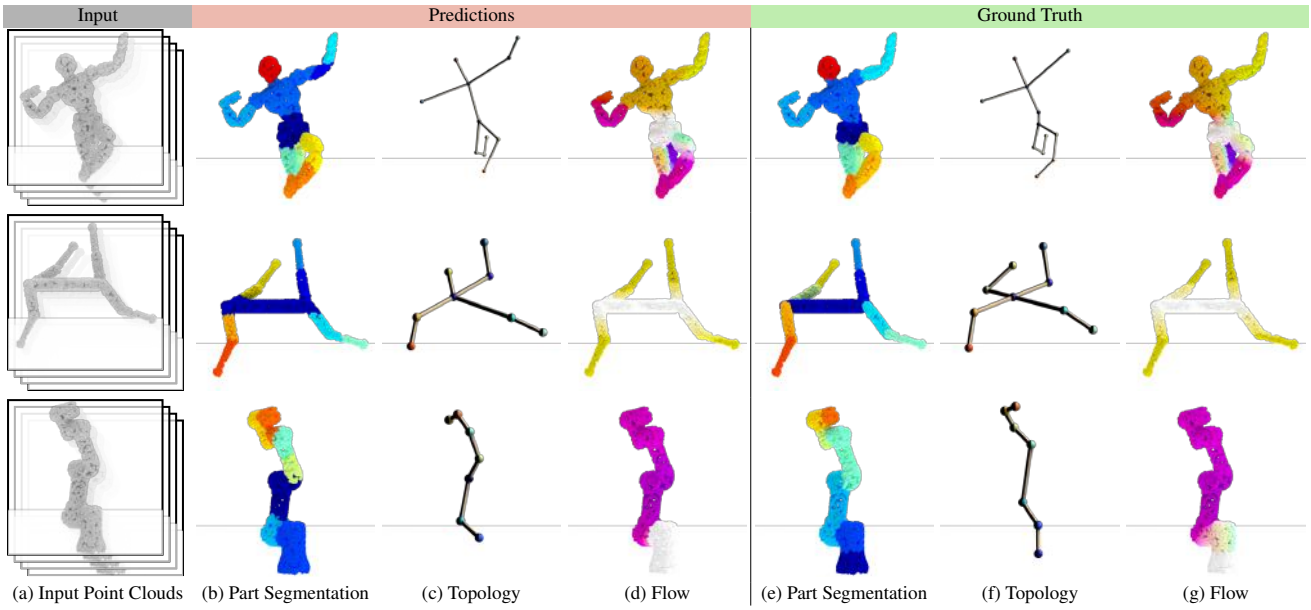


Figure 4. Qualitative Results on RoboArt Dataset. Given the input point cloud sequence (shown in (a)), we show the part segmentation, part connectivity, and implied flow using our inferred articulated model (in (b, c, d)) and the ground truth articulated model (in (e, f, g)) across the robots from the test set. Our method can successfully deal with various parts connected in complex ways.

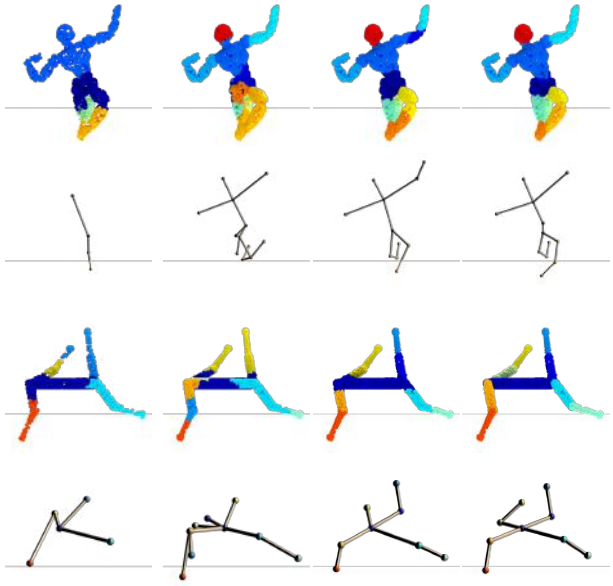


Figure 5. Qualitative comparison against MultiBodySync [15] and WatchItMove [35] on the RoboArt dataset. Note, a) MultiBodySync by itself doesn't produce a kinematic tree, we use our method on top of their output to generate one, and b) we provide WatchItMove [35] with the ground truth SDFs and number of parts (which are not used by our method). Even after these modifications, the past methods cannot solve the task as well as ours.

can be computed in closed-form using its screw parameters $(\mathbf{s}_i, \boldsymbol{\theta}_i^T)$ and Rodriguez formula [8]:

$$\mathbf{T}(\mathbf{s}, \boldsymbol{\theta}) = \text{Exp}([\boldsymbol{\omega}, \mathbf{v}]) = \text{Exp}([\tau\mathbf{l}; \tau\mathbf{m} \times \mathbf{l} + d\mathbf{l}]^T) \quad (3)$$

where $\boldsymbol{\omega} = \tau\mathbf{l} \in \mathbb{R}^3$ is the Euler vector representing rotation; $\mathbf{v} = \tau\mathbf{l} \times \mathbf{m} + d\mathbf{l}$ the translational vector; $\text{Exp} : \mathbb{R}^6 \mapsto \mathbb{SE}(3)$ maps the minimal 6-DoF representation to an $\mathbb{SE}(3)$ rigid transform; $\times$ is the cross product between vectors. There exists two special cases. A joint is called **prismatic joint** if its rotation angle is always zero, *i.e.* $\tau = 0$. In this case, the rigid body can only slide along the axis $\mathbf{l}$ without rotation (think about a drawer). A joint is called **revolute joint** if translational component always equals to 0, *i.e.* $d = 0$. In this case, the rigid body can only rotate along the rotation axis $\mathbf{l}$ (think about a door);

3.2. Energy Formulation

Our approach takes as input a point cloud sequence $\mathcal{P} = \{\mathbf{P}^t\}_{t \in 1,\ldots T}$. The output of our approach is the number of parts n, the part labeling function f, a kinematic tree connecting the different parts $\boldsymbol{\Gamma}$, and screw parameters $\{\mathbf{s}_i\}$ for each non-root joint.

We estimate these parameters using an analysis-by-synthesis approach and find model parameters M, such that after appropriate per-time articulation $\boldsymbol{\theta}^t$, $M(\boldsymbol{\theta}^t; \boldsymbol{\Gamma}, f)$ matches the given point cloud $\mathbf{P}^t$ well.

$$\min_{n, f, \boldsymbol{\Gamma}, \{\mathbf{s}_i\}} \sum_t \min_{\boldsymbol{\theta}_t} E_{\text{recons}} \left(M(\boldsymbol{\theta}^t; \boldsymbol{\Gamma}, f), \mathbf{P}^t \right), \quad (4)$$

where $M(\boldsymbol{\theta}^t; \boldsymbol{\Gamma}, f)$ denotes the canonical point cloud $\mathbf{P}^c$ transformed using the model parameters and the articulation $\mathbf{x}^t$ at time t.

E_{recons} measures the compatibility between the point cloud articulated according to the inferred model and the observed point cloud at each step. Specifically, the energy

consists of three terms:

$$E_{\text{recons}} = \lambda_{\text{CD}} E_{\text{CD}} + \lambda_{\text{EMD}} E_{\text{EMD}} + \lambda_{\text{flow}} E_{\text{flow}} \quad (5)$$

Chamfer distance term: E_{CD} measures agreement between two point clouds using the Chamfer distance:

$$E_{\text{CD}}(\mathbf{X}, \mathbf{Y}) = \sum_{\mathbf{x} \in \mathbf{X}} \min_{\mathbf{y}} \|\mathbf{x} - \mathbf{y}\|_2^2 + \sum_{\mathbf{y} \in \mathbf{Y}} \min_{\mathbf{x}} \|\mathbf{x} - \mathbf{y}\|_2^2. \quad (6)$$

Earth-mover distance term: The earth-mover distance E_{EMD} also measures geometric compatibility. It finds the optimal assignment between the two points sets by solving a bipartite matching problem [22] and measure the residual loss:

$$E_{\text{EMD}}(\mathbf{X}, \mathbf{Y}) = \min_{\mathbf{S}} \|\mathbf{S}\mathbf{X} - \mathbf{Y}\|_F^2, \quad (7)$$

where $\mathbf{S}$ is a permutation matrix that represents the bipartite assignment. Compare against E_{CD}, E_{EMD} could better capture details while E_{CD} only focus on point coverage. We compute the assignment via linear assignment solver [7].

Flow energy term: This term encourages the estimated 3D motion to be similar to observed point-wise 3D motion obtained from a scene flow network,

$$E_{\text{flow}} = \sum_{\mathbf{x}} \|\mathbf{x}^t - \mathbf{x}^{t-1} - g(\mathbf{P}^t, \mathbf{P}^{t-1}; \mathbf{x}^t)\|_2^2. \quad (8)$$

where $\mathbf{x}^t - \mathbf{x}^{t-1} = (\mathbf{T}^t_{f(\mathbf{x})} - \mathbf{T}^{t-1}_{f(\mathbf{x})}) \cdot \mathbf{x}$ is the predicted motion flow using the state for point $\mathbf{x}$ at time t and $t-1$. $g(\mathbf{P}^t, \mathbf{P}^{t-1}; \mathbf{x}^t)$ is the observed 3D motion flow between two point clouds $\mathbf{P}^t$ and $\mathbf{P}^{t-1}$, at the query point location $\mathbf{x}^t$. Since there is no guaranteed one-to-one mapping between $M(\boldsymbol{\theta}^t)$ and $\mathbf{P}^t$, we use trilinear interpolation to estimate the inferred flow at an arbitrary point $\mathbf{x}^t$:

$$g(\mathbf{P}^t, \mathbf{P}^{t-1}; \mathbf{x}^t) = \frac{\sum_{\mathbf{x}^t_k \in \text{knn}(\mathbf{x}^t; \mathbf{P}^t)} \|\mathbf{x}^t - \mathbf{x}^t_k\|^{-1} F(\mathbf{x}^t_k)}{\sum_{\mathbf{x}^t_k \in \text{knn}(\mathbf{x}^t; \mathbf{P}^t)} \|\mathbf{x}^t - \mathbf{x}^t_k\|^{-1}},$$

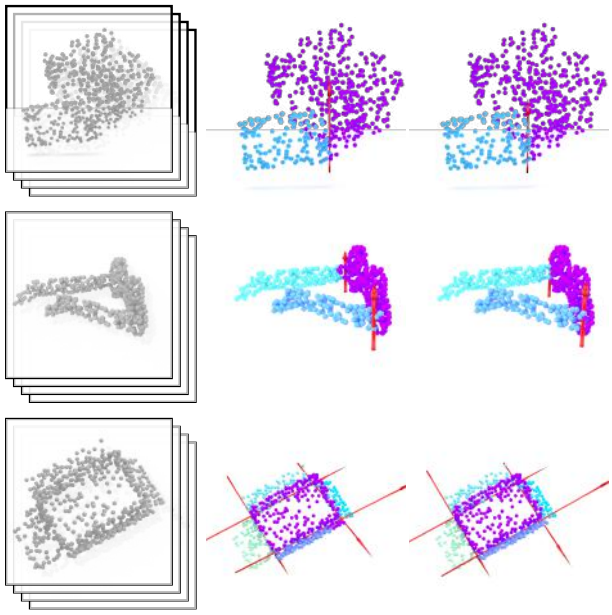
where $\mathbf{F}^t$ is the predicted flow between observation $\mathbf{P}^t$ and $\mathbf{P}^{t-1}$. Given the established flow $\mathbf{F}^t$ at locations in $\mathbf{P}^t$, $g(\cdot)$ takes query prediction $\mathbf{x}^t$ as input and returns an interpolated motion estimation at point $\mathbf{x}^t$. $\text{knn}(\mathbf{x}^t; \mathbf{P}^t)$ retrieves the K-nearest neighbors of $\mathbf{x}^t$ from $\mathbf{P}^t$.

3.3. Inference via Relax and Project

Directly optimizing over the set of model parameters as defined in Eq. (4) is difficult. It involves both discrete and continuous optimization variables as well as structured constraints such as the tree-structure of $\boldsymbol{\Gamma}$, making it hard for both numerical approaches and combinatorial methods. Hence we pursue an alternate "relax-and-project" approach.

3.3.1 Fitting a Relaxed Model

Our method first fits a *relaxed* model $\hat{M}$. This relaxed model doesn't constrain the parts to form a kinematic tree and thus lets individual parts to follow their own independent 6DOF trajectory over time. This relaxed model is parameterized via the number of parts n (same as for M), the part labeling function f (also same as for M). This model is articulated via a 6DOF pose for each part at each time step t: $\hat{\mathbf{T}}^t = \{\hat{\mathbf{T}}^t_i\}_{i \in [1 \ldots n]}$.



(a) Input (b) Predicted Model (c) Ground Truth

Figure 6. Qualitative results on Sapiens dataset from [15]. We visualize the predicted and ground truth articulated models. Different parts are in different colors, and we also show the screw parameters (in red) for the inferred joints.

We first optimize this relaxed model using the same cost function as in Eq. (4) but evaluated for reposing under the relaxed model at each time step t via $\hat{\mathbf{T}}^t$:

$$\min_{n,f} \sum_t \min_{\mathbf{T}^t} E_{\text{recons}} \left(\hat{M}(\mathbf{T}^t), \mathbf{P}^t \right), \quad (9)$$

where $\hat{M}(\mathbf{T}^t)$ denotes the canonical point cloud $\mathbf{P}^c$ transformed using the model parameters and the per-time step part transformations $\mathbf{T}^t$ at time t.

The energy function defined in Eq. (5) involves a joint optimization over 6-DOF transformations $\hat{\mathbf{T}}^t_i$ and a neural segmentation field f. For each point, the part segmentation field outputs a discrete-valued index $f(\mathbf{x})$, which is later used to query the corresponding rigid transform to warp the point cloud as defined in Eq. (1). Optimizing through this discrete index $f(\mathbf{x})$ is hard. We leverage Gumbel-softmax trick [18] along with the straight-through gardient estimator [3] to overcome this challenge. This makes the energy function end-to-end differentiable w.r.t. f and $\hat{\mathbf{T}}^t$, allowing us to use gradient descent to minimize the energy.

3.3.2 Projecting to the Kinematic Model

The estimated relaxed model parameters are projected onto a valid kinematic model by inducing a tree structured con-

nectivity between the parts and projecting absolute 6DOF transformations into child-parent screw parameters. This is done using a cost function that measures the discrepancy between kinematic parameters $\mathbf{\Gamma}, \{\mathbf{s}_i\}, \{\boldsymbol{\theta}^t\}$ and relaxed 6DOF transformations $\{\hat{\mathbf{T}}^t\}$:

$$\min_{\mathbf{\Gamma}, \{\mathbf{s}_i\}, \{\boldsymbol{\theta}^t\}} E_{\text{project}} \left(\left(\mathbf{\Gamma}, \{\mathbf{s}_i\}, \{\boldsymbol{\theta}^t\} \right), \{\hat{\mathbf{T}}^t\} \right). \quad (10)$$

Given individual 6DOF part trajectories for all parts, we want to infer a kinematic model that is as similar to the relaxed model while obeying the kinematic constraints. The kinematic constraints consists of two aspects: the tree structure (*i.e.* the connectivity of parts with one another) and the 1DOF constraint for the joint between each pair of connected parts. We tackle the projection problem defined in Eq. (10) by minimizing the cost over different trees topologies $\mathbf{\Gamma}$ and associated screw parameters $\mathbf{s}_i = \{\mathbf{l}_i, \mathbf{m}_i\}$:

$$E_{\text{project}} = \lambda_{\text{spatial}} E_{\text{spatial}} + \lambda_{\text{1-DOF}} E_{\text{1-DOF}}. \quad (11)$$

This loss function evaluates the total fitness of parent-child pairs in $\mathbf{\Gamma}$. E_{spatial} measures the spatial proximity of parent-child pairs, $E_{\text{spatial}}(\mathbf{\Gamma}) = \sum_i \min_{\mathbf{x} \in \mathbf{p}_i} \min_{\mathbf{y} \in \mathbf{p}_{\text{pa}(i)}} \|\mathbf{x} - \mathbf{y}\|_2^2$, where $\mathbf{p}_i = \{\mathbf{x} \in \mathbf{P}^c | f(\mathbf{x}) = i\}$ is the set of points corresponding to part i. The $E_{\text{1-DOF}}$ term measures how close to a 1DOF motion is the motion of the child part relative to the parent part. $E_{\text{1-DOF}}$ is computed as the error in approximating the temporal sequence of relative transformation of the child part with respect to the parent part as a 1DOF transformation:

$$E_{\text{1-DOF}} = \sum_i \sum_t \text{trace}((\hat{\mathbf{T}}^t_{pa(i)} \ominus \hat{\mathbf{T}}^t_i) \ominus \mathbf{T}(\mathbf{s}_i, \boldsymbol{\theta}^t_i)), \quad (12)$$

where $\oplus$ is the rigid pose composition operator: $\mathbf{T}_a \oplus \mathbf{T}_b = \mathbf{T}_b \cdot \mathbf{T}_a$ and $\ominus$ is the inverse rigid pose composition operator $\mathbf{T}_a \ominus \mathbf{T}_b = \text{inv}(\mathbf{T}_b) \cdot \mathbf{T}_a$.

Solving the projection problem defined in Eq. (11) is also challenging: it's a joint optimization between discrete tree structure $\mathbf{\Gamma}$ and continuous screw parameters $\mathbf{s}_i, \boldsymbol{\theta}^t_i$; the desire for a valid tree topology $\mathbf{\Gamma}$ introduces a complicated structured constraint. We observe that screw parameters $\mathbf{s}_i, \boldsymbol{\theta}^t_i$ are only involved in exactly 1 1-DOF energy term under any valid tree. Thus, they can be independently optimized. For all part pairs (i, j), we compute screw parameters from $\hat{\mathbf{T}}_i$ and $\hat{\mathbf{T}}_j$ under the assumption that $j = \text{pa}(i)$ using screw theory [43]. This let us compute the 1DOF energy for all part pairs (i, j). The spatial term can also be similarly computed for all part pairs. The minimization in Eq. (11) then reduces to finding the minimum spanning tree which can be done efficiently [6].

Merging. Before the projection, we iteratively merge parts that are spatially close and don't have relative motion. The latter falls out directly from Eq. (12) if the approximation against the identity transform is below ϵ_m.

Table 2. Comparison to past methods on RoboArt test set. We outperform past methods on all metrics that measure input reconstruction quality (shape reconstruction error, flow error, flow accuracy), model accuracy (part segmentation rand index, and kinematic tree edit distance), and reanimiation error.

Method	Recons. error ↓	Flow error ↓	Flow acc. ↑	Rand Index ↑	Tree Edit Distance ↓	Reanimate Error ↓
MultiBodySync [15]	4.76	3.42	0.26	0.70	6.6	9.72
WatchItMove [35]	12.77	8.42	0.06	0.78	6.6	7.43
Ours	**1.26**	**0.57**	**0.59**	**0.86**	**2.9**	**3.66**

Table 3. Comparison to past methods on Sapiens dataset [15]. We do better in both flow estimation and segmentation under two different evaluation settings.

Method	Flow Error ↓	Multi-scan RI↑	Per-scan RI↑
PWC-Net [50]	6.20	-	-
PointNet++ [38]	-	0.62	0.65
MeteorNet [30]	-	0.59	0.60
Deep Part [57]	5.95	0.64	0.67
NPP [11]	21.22	0.63	0.66
MultiBodySync [15]	5.03	0.76	0.77
Ours	**4.79**	**0.79**	**0.79**

Table 4. Role of different energy terms in Eq. (5) evaluated on the RoboArt validation set. To isolate the direct impact of the energy terms, we conduct this experiment w/o the canonical frame selection (we always use the middle frame) and w/o screw parameter finetuing after projection to a kinematic model. All three terms are important with the flow term being the most important.

E_{CD}	E_{flow}	E_{EMD}	Recons. error ↓	Flow error ↓	Flow acc. ↑	Rand Index ↑	Tree Edit Distance ↓
✓			6.97	8.05	0.31	0.76	6.50
	✓		2.97	1.64	0.31	0.83	5.00
		✓	3.49	3.03	0.13	0.28	6.75
✓	✓		1.47	0.88	**0.48**	0.83	4.00
✓		✓	1.93	1.99	0.17	**0.86**	4.25
	✓	✓	1.54	0.97	0.33	0.83	5.25
✓	✓	✓	**1.31**	**0.86**	0.40	**0.86**	**3.25**

3.3.3 Final Fitting

We further finetune $\{\mathbf{s}_i\}, \{\boldsymbol{\theta}^t_i\}$ over the original problem in Eq. (4) with gradient descent while keeping f and $\mathbf{\Gamma}$ fixed. This gives the final estimation of tree structure $\mathbf{\Gamma}$, part segmentation f, screw parameters $\{\mathbf{s}_i\}$, and joint states $\{\boldsymbol{\theta}^t_i\}$.

Canonical frame selection. Given the non-convexity of the optimization, we run the optimization multiple times by choosing different time steps c as *canonical* frame to build neural part field. We pick the best c that has the lowest E_{project} in Eq. (10).

4. Experiments

We performing experiments on two 3D asset datasets, demonstrates our method could be applied on arbitrary articulated objects both qualitatively and quantitatively.

Table 5. Evaluation of other design choices on the RoboArt validation set. Neural segmentation field (**f**), Gumbel-softmax (**Gumbel**), projection to the kinematic model (**Project**), canonical frame selection (**Cano**) all contribute to the final performance.

Designs	Recons. error ↓	Flow error ↓	Flow acc. ↑	Rand Index ↑	Tree Edit Distance ↓	Reanimate Error ↓
w/o **f**	2.63	1.26	0.40	0.76	7.25	11.68
w/o **Gumbel**	3.54	1.65	0.34	0.74	7.25	11.69
w/o **Project**	1.25	0.75	0.45	**0.88**	**3.00**	8.86
w/o **Cano**	1.54	0.74	0.47	0.85	3.33	6.83
ours	**1.19**	**0.64**	**0.49**	**0.88**	**3.00**	6.50

4.1. Experimental Setup

Datasets. We conduct experiments on two datasets: the *RoboArt* dataset that we introduce in this work, and the *Sapiens* dataset from [15]. The RoboArt dataset consists of 18 popular robots which span manipulator like panda robot to humanoid robots like nao. These robots have large variation in number of parts (7–15), part geometries (barrett robot's fingers to atlas robot's trunk), and part connectivity (linear chains to complex trees). We split the dataset in training (4 robots), validation (4 robots), and test (10 robots). For each robot we articulate them using [33] to record a 10 time-step points cloud sequence, each containing 4096 points. Crucially, we make sure that these 4096 points are randomly resampled at each time step to prevent any leakage of correspondence information between time steps. Supplementary material shows visualizations for different robots and summarizes more statistics. The Sapiens dataset from [15] contains 720 test sequences across 20 different object categories from Sapien [52]. The dataset provides 4 point cloud frames as input for each object. The 4 frames all have a different global coordinate frame. More results (include real-world) are shown in supplementary.

Metrics. Our experiments measure the final reanimation error, intermediate metrics that evaluate part segmentation and inferred kinematics, and reconstruction metrics that evaluate how well our inferred models explain the input point clouds. We detail them next. **Reconstruction metrics** include a) Reconstruction Error that measures the mean end-point-error between the object articulated using the inferred articulation parameters *vs*. with the ground truth parameters, b) Flow Error that measures the error between the flow between consecutive frames implied by the predicted model *vs*. the ground truth model, and c) Flow Accuracy that measures the %age of points for which flow implied by the predictions is within a δ threshold from ground truth flow. **Intermediate metrics** include a) Random Index (RI) [4] that measures overlap between the partition induced by predicted part labels *vs*. the partition induced by ground truth part labels and b) Tree Edit Distance [37,60] that measures the similarity between predicted kinematic tree and the ground truth kinematic tree. These metrics have been used in past works, [15] and [53] respectively. **Reanimataion Metric:** Our inferred articulated model can be reanimated given new locations for a sparse set of points (1 per part). We measure the quality of such reanimation via the mean end-point-error between the ground truth reanimated model and the predicted reanimated model. Precise metric definition is provided in the supplementary.

Implementation Details. Our training requires us to compute the flow between pairs of frames. We train a Siamese correspondence network using PointNet++ [38]. At test time, we compute the similarity between each pair of points in the two point clouds and use correspondences that are mutually the best. The correspondence network is trained on the Sapiens and the RoboArt datasets and does not include any objects or robots that we evaluate upon (neither for validation nor for testing). **Part Segmentation Field** function f is realized using an MLP with one hidden layer of 128 dimensions. All optimizations are done using PyTorch. We use the standard Adam optimizer [24] for all optimization with a learning rate of 1e-3 for the MLP and 1e-2 for transformations. During **relaxed model estimation stage**, we set maximum number of parts to 20 and all 6DOF transformations are initialized to be identity. We found it helpful to only optimize with E_{CD} and E_{flow} for the first 5000 iterations, and replace E_{CD} by E_{EMD} and optimizing for another 10000 iterations. For efficiency reasons, we apply E_{EMD} on $4\times$ downsampled point clouds and only update the optimal assignment every 5 iterations. During **projection stage**, we first merge parts that are spatially close and don't move relative to one another with $\epsilon_m = 3e - 2$. During the final optimization stage, we only optimize the screw parameters and their states. We keep the number of parts and part segmentation fixed. Merging details and visualization can be found in supplementary.

4.2. Articulation Object Modeling Results

Comparison on the RoboArt Dataset. As discussed in Sec. 2, to the best of our knowledge, there isn't a past work that exactly solves the entire task we tackle. We adapt and extend MultiBodySync [15] and WatchItMove [35] for our task and compare them on the RoboArt dataset. MultiBodySync [15] uses pairwise flow prediction and sets up a joint synchronization problem to obtain a temporally-consistent part segmentation. We use the part segmentation and pairwise flows to produce a kinematic tree and associated screw parameters using the projection part of our algorithm. WatchItMove [35] was designed for building articulable models from RGB videos. We modify it to work with point clouds. For point clouds, there is no rendering loss but only a loss on the predicted SDF. We provide ground truth SDFs and ground truth numbers of parts to their method.

Tab. 2 presents the quantitative results on the RoboArt. We outperform both these past methods by a large margin across all metrics. We looked into the poor performance

(a) Input (b) w/o kinematics (c) w/ kinematics (d) GT

Figure 7. Reanimation Results on the RoboArt Dataset. Given new locations for a sparse set of points on the object(shown in (a)), our full method (shown in (c)) is able to generate a reasonable re-animation to match the specified points. (b) shows reanimation results from an ablated version of our model where we don't restrict the parts to form a kinematic tree. This results in poor reanimation. Thus, correctly inferring the connectivity of different parts with one another is crucial for high-quality reanimation.

of these past methods. MultiBodySync relies on accurate *pairwise* flow predictions. For objects with a long chain of articulation, *e.g*. a robot arm, over time, the object deforms quite a bit limiting the performance of the correspondence network. For WatchItMove, we believe poor performance comes from some their use of *soft* assignment of points to segments during training. This leads to incorrect assignment of points to parts at test time. In our own ablations (presented in Sec. 4.3) we note that our hard assignment of points to segments during training is crucial to the final performance. Fig. 5 show qualitative comparisons.

Comparison on Sapiens Dataset. We also compare performance on specific sub-tasks (part segmentation and flow prediction) useful for re-animation as studied in past work [15] on their Sapiens dataset in Tab. 3. Here, we also compare a number of past methods, including PWC-Net [50], PointNet++ [38], *etc*. Please refer to [15] for details about these methods. We outperform all these past methods. We believe our strong performance on this dataset is because of the additional joint and kinematic constraints enforced by our method, which are not used by these past methods. See qualitative results in Fig. 6. For a fair comparison, we use the flow network from [15] for our method.

4.3. Ablations

Tab. 4 evaluates the importance of the different energy terms in our formulation on RoboArt validation set. We note that all terms are important, but particularly E_{flow} has the largest effect. We also find that just the chamfer distance term, E_{CD}, is insufficient and works poorly. We believe this is because E_{CD} doesn't provide good gradients (it tries to snap a point to the closest point in the shape, which may not necessarily be the correct location for it).

Tab. 5 reports the impact of our other design choices. Representing the part segmentation as a neural field induces a smoothness prior on the parts. Removing this prior by optimizing each point' segmentation logits independently works worse (Row 1: w/o **f**). Taking Gumbel-softmax as hard assignment during training is important. Soft assignment (softmax) of points to transformations provide a lot more freedom for optimization. This prevents learning of the correct transformations, which don't work when used with hard assignments at test time (Row 2, w/o **Gumbel**). Projecting relaxed model to valid kinematic one leads to better reconstruction quality and lower re-animation error by further constraint the transformations and regularize the motions (Row 3, w/o **Project**). We also found that the choice of frame for building the neural part field is important. Sometimes two otherwise far away parts can be close to one another in some frame (*e.g*. left and right arms of a humanoid) making it difficult for the part segmentation field to separate them. Searching for which frame to use as the canonical frame helps in over-coming the non-convexity in the optimization (Row 4, w/o **Cano**).

Re-animation Results. We now showcase how the reconstructed rearticulatable model can be retargeted to new poses. Given an inferred model consisting of part segmentation, kinematic tree, and 1DoF joint parameters, we use inverse kinematics (IK) to compute the transformation between our canonical point frame and a target pose, using new locations for a sparse set of points on the object. We then re-animate the point cloud through the inferred joint parameters. Fig. 7 shows the results. We see that the re-animation quality significantly improves by inferring the kinematic linking of parts.

Run-time. On RoboArt sequences, ours takes 45 minutes on a RTX 3090 GPU while WatchItMove takes 2 hours. On Sapiens, ours takes 5min (same as MultiBodySync).

Limitation. Our method heavily relies on motion cues. It might fail to distinguish two rigid parts if they undergo the same motion in the sequence (*e.g*. a humanoid robot moves both arms synchronously). We leave this to future work and plan to tackle it by incorporating appearance information.

5. Conclusion

We presented a novel method for building arbitrary rearticulable models from a point cloud sequence. Our approach jointly infers part segmentation, screw-parametric joints, and kinematic tree structures in a category-agnostic manner. We validated our method's efficacy on two challenging datasets, showing superior results over previous leading works. We further showed that the inferred model could be retargeted at any novel pose, demonstrating the potential for reanimation and manipulation.

Acknowledgements: This material is based upon work supported by an NSF CAREER Award (IIS-2143873) and the USDA/NSF AIFARMS National AI Institute (USDA #2020-67021-32799). Prof. Wang's lab is supported, in part, by Amazon research award and Insper innovation grant.

References

[1] Ben Abbatematteo, Stefanie Tellex, and George Konidaris. Learning to generalize kinematic models to novel objects. In *CoRL*, 2019. 2

[2] Hameed Abdul-Rashid, Miles Freeman, Ben Abbatematteo, George Konidaris, and Daniel Ritchie. Learning to infer kinematic hierarchies for novel object instances. In *ICRA*. IEEE, 2022. 2

[3] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv*, 2013. 5

[4] Xiaobai Chen, Aleksey Golovinskiy, and Thomas Funkhouser. A benchmark for 3d mesh segmentation. *TOG*, 28(3), 2009. 7

[5] Zhiqin Chen, Kangxue Yin, Matthew Fisher, Siddhartha Chaudhuri, and Hao Zhang. Bae-net: Branched autoencoder for shape co-segmentation. In *ICCV*, 2019. 3

[6] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms*. MIT press, 2022. 6

[7] David F Crouse. On implementing 2d rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4), 2016. 5

[8] Jian S Dai. Euler–rodrigues formula variations, quaternion conjugation and intrinsic connections. *Mechanism and Machine Theory*, 92, 2015. 4

[9] Andrea F Daniele, Thomas M Howard, and Matthew R Walter. A multiview approach to learning articulated motion models. In *Robotics Research*. Springer, 2020. 2

[10] Karthik Desingh, Shiyang Lu, Anthony Opipari, and Odest Chadwicke Jenkins. Factored pose estimation of articulated objects using efficient nonparametric belief propagation. In *ICRA*. IEEE, 2019. 2

[11] David S Hayden, Jason Pacheco, and John W Fisher. Nonparametric object and parts modeling with lie group dynamics. In *CVPR*, 2020. 2, 6

[12] Eric Heiden, Ziang Liu, Vibhav Vineet, Erwin Coumans, and Gaurav S Sukhatme. Inferring articulated rigid body dynamics from rgbd video. *arXiv*, 2022. 2

[13] Ruizhen Hu, Lubin Fan, and Ligang Liu. Co-segmentation of 3d shapes via subspace clustering. In *Computer graphics forum*, volume 31. Wiley Online Library, 2012. 3

[14] Ruizhen Hu, Wenchao Li, Oliver Van Kaick, Ariel Shamir, Hao Zhang, and Hui Huang. Learning to predict part mobility from a single static snapshot. *TOG*, 36(6), 2017. 2

[15] Jiahui Huang, He Wang, Tolga Birdal, Minhyuk Sung, Federica Arrigoni, Shi-Min Hu, and Leonidas J Guibas. Multibodysync: Multi-body segmentation and motion estimation via 3d scan synchronization. In *CVPR*, 2021. 2, 3, 4, 5, 6, 7, 8

[16] Xiaoxia Huang, Ian Walker, and Stan Birchfield. Occlusion-aware reconstruction and manipulation of 3d articulated objects. In *ICRA*. IEEE, 2012. 2

[17] Ajinkya Jain, Rudolf Lioutikov, Caleb Chuck, and Scott Niekum. Screwnet: Category-independent articulation model estimation from depth images using screw theory. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021. 2, 3

[18] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparametrization with gumble-softmax. In *ICLR*, 2017. 5

[19] Hanxiao Jiang, Yongsen Mao, Manolis Savva, and Angel X Chang. Opd: Single-view 3d openable part detection. *arXiv*, 2022. 2

[20] Zhenyu Jiang, Cheng-Chun Hsu, and Yuke Zhu. Ditto: Building digital twins of articulated objects from interaction. In *CVPR*, 2022. 2, 3

[21] Gunnar Johansson. Visual perception of biological motion and a model for its analysis. *Perception & psychophysics*, 14, 1973. 1

[22] Roy Jonker and Anton Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4), 1987. 5

[23] Yuki Kawana, Yusuke Mukuta, and Tatsuya Harada. Unsupervised pose-aware part decomposition for 3d articulated objects. *arXiv*, 2021. 2

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 2014. 7

[25] Suren Kumar, Vikas Dhiman, Madan Ravi Ganesh, and Jason J Corso. Spatiotemporal articulated models for dynamic slam. *arXiv*, 2016. 2

[26] Hao Li, Guowei Wan, Honghua Li, Andrei Sharf, Kai Xu, and Baoquan Chen. Mobility fitting using 4d ransac. In *Computer Graphics Forum*, volume 35. Wiley Online Library, 2016. 2

[27] Xiaolong Li, He Wang, Li Yi, Leonidas J Guibas, A Lynn Abbott, and Shuran Song. Category-level articulated object pose estimation. In *CVPR*, 2020. 2

[28] Qihao Liu, Weichao Qiu, Weiyao Wang, Gregory D Hager, and Alan L Yuille. Nothing but geometric constraints: A model-free method for articulated object pose estimation. *arXiv*, 2020. 3

[29] Xueyi Liu, Xiaomeng Xu, Anyi Rao, Chuang Gan, and Li Yi. Autogpart: Intermediate supervision search for generalizable 3d part segmentation. In *CVPR*, 2022. 3

[30] Xingyu Liu, Mengyuan Yan, and Jeannette Bohg. Meteornet: Deep learning on dynamic 3d point cloud sequences. In *ICCV*, 2019. 6

[31] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *TOG*, 34(6), 2015. 2

[32] Kevin M Lynch and Frank C Park. *Modern robotics*. Cambridge University Press, 2017. 3

[33] Matthew Matl. Urdfpy. https://github.com/mmatl/urdfpy, 2019. 7

[34] 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 *ICCV*, 2021. 2

[35] Atsuhiro Noguchi, Umar Iqbal, Jonathan Tremblay, Tatsuya Harada, and Orazio Gallo. Watch it move: Unsupervised discovery of 3d joints for re-posing of articulated objects. In *CVPR*, 2022. 2, 3, 4, 6, 7

[36] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019. 2

[37] Mateusz Pawlik and Nikolaus Augsten. Efficient computation of the tree edit distance. *ACM Transactions on Database Systems (TODS)*, 40(1), 2015. 7

[38] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 30, 2017. 6, 7, 8

[39] Shengyi Qian, Linyi Jin, Chris Rockwell, Siyi Chen, and David F Fouhey. Understanding 3d object articulation in internet videos. In *CVPR*, 2022. 3

[40] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *SIGGRAPH Asia*, 2017. 2

[41] Andrei Sharf, Hui Huang, Cheng Liang, Jiapei Zhang, Baoquan Chen, and Minglun Gong. Mobility-trees for indoor scenes manipulation. In *Computer Graphics Forum*, volume 33. Wiley Online Library, 2014. 2

[42] Oana Sidi, Oliver van Kaick, Yanir Kleiman, Hao Zhang, and Daniel Cohen-Or. Unsupervised co-segmentation of a set of shapes via descriptor-space spectral clustering. In *SIGGRAPH Asia*, 2011. 2, 3

[43] Stefano Stramigioli and Herman Bruyninckx. Geometry and screw theory for robotics. *Tutorial during ICRA*, 2001, 2001. 6

[44] Dimitrios Tzionas and Juergen Gall. Reconstructing articulated rigged models from rgb-d videos. In *ECCV*. Springer, 2016. 3

[45] Oliver Van Kaick, Kai Xu, Hao Zhang, Yanzhen Wang, Shuyang Sun, Ariel Shamir, and Daniel Cohen-Or. Co-hierarchical analysis of shape structures. *TOG*, 2013. 3

[46] Weiyue Wang, Ronald Yu, Qiangui Huang, and Ulrich Neumann. Sgpn: Similarity group proposal network for 3d point cloud instance segmentation. In *CVPR*, 2018. 3

[47] Xiaogang Wang, Bin Zhou, Yahao Shi, Xiaowu Chen, Qinping Zhao, and Kai Xu. Shape2motion: Joint analysis of motion parts and attributes from 3d shapes. In *CVPR*, 2019. 3

[48] 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 *CVPR*, 2022. 2

[49] 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 *ICCV*, 2021. 2

[50] Wenxuan Wu, Zhi Yuan Wang, Zhuwen Li, Wei Liu, and Li Fuxin. Pointpwc-net: Cost volume on point clouds for (self-) supervised scene flow estimation. In *ECCV*, 2020. 6, 8

[51] Yuefan Wu, Zeyuan Chen, Shaowei Liu, Zhongzheng Ren, and Shenlong Wang. Casa: Category-agnostic skeletal animal reconstruction. In *NeurIPS 35 (NeurIPS)*, 2022. 2

[52] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *CVPR*, 2020. 7

[53] Zhan Xu, Yang Zhou, Evangelos Kalogerakis, Chris Landreth, and Karan Singh. Rignet: Neural rigging for articulated characters. *arXiv*, 2020. 2, 7

[54] Han Xue, Liu Liu, Wenqiang Xu, Haoyuan Fu, and Cewu Lu. Omad: Object model with articulated deformations for pose estimation and retrieval. *arXiv*, 2021. 2

[55] Zihao Yan, Ruizhen Hu, Xingguang Yan, Luanmin Chen, Oliver Van Kaick, Hao Zhang, and Hui Huang. Rpm-net: recurrent prediction of motion and parts from point cloud. *arXiv*, 2020. 3

[56] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *CVPR*, 2022. 2

[57] Li Yi, Haibin Huang, Difan Liu, Evangelos Kalogerakis, Hao Su, and Leonidas Guibas. Deep part induction from articulated object pairs. *arXiv*, 2018. 2, 3, 6

[58] Qing Yuan, Guiqing Li, Kai Xu, Xudong Chen, and Hui Huang. Space-time co-segmentation of articulated point cloud sequences. In *Computer Graphics Forum*, volume 35. Wiley Online Library, 2016. 3

[59] Vicky Zeng, Tabitha Edith Lee, Jacky Liang, and Oliver Kroemer. Visual identification of articulated object parts. In *IROS*. IEEE, 2021. 3

[60] Kaizhong Zhang and Dennis Shasha. Simple fast algorithms for the editing distance between trees and related problems. *SIAM journal on computing*, 18(6), 1989. 7

[61] Silvia Zuffi, Angjoo Kanazawa, David Jacobs, and Michael J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2