Distributional Depth-Based Estimation of Object Articulation Models

Anonymous Author(s)
Affiliation
Address
email

Abstract: We propose a method that efficiently learns distributions over articulation models directly from depth images without the need to know articulation model categories a priori. By contrast, existing methods that learn articulation models from raw observations require objects to be textured, and most only predict point estimates of the model parameters. Our core contributions include a novel representation for distributions over rigid body transformations and articulation model parameters based on Screw theory, von Mises-Fisher distributions and Stiefel manifolds. Combining these concepts allows for an efficient, mathematically sound representation that inherently satisfies several constraints that rigid body transformations and articulations must adhere to. In addition, we introduce a novel deep-learning based approach, DUST-net, that efficiently learns such distributions and, hence, performs category-independent articulation model estimation while also providing model uncertainties. We evaluate our approach on two benchmarking datasets and three real-world objects and compare its performance with two current state-of-the-art methods. Our results demonstrate that DUST-net can successfully learn distributions over articulation models and their parameters for novel objects across articulation model categories with better accuracy than state-of-the-art methods.

Keywords: Articulated Objects, Model Learning, Uncertainty Estimation

1 Introduction

Articulated objects, such as drawers, staplers, refrigerators, and dishwashers, are ubiquitous in human environments. These objects consist of multiple rigid bodies connected via mechanical joints such as hinge joints or slider joints. Robots in human environments will need to interact with these objects often while assisting humans in performing day-to-day tasks. To interact safely with such objects, a robot must reason about their articulation properties while manipulating them. An ideal method for learning such properties might estimate these parameters directly from raw observations, such as RGB-D images while requiring limited or no a priori information about the task. The ability to additionally provide a confidence over the estimated properties, would allow such a method to be leveraged in the development of safe motion policies for articulated objects [1].

The majority of existing methods to learn articulation models for objects from visual data either need fiducial markers to track motion between object parts [2–5] or require textured objects [6–10]. Recent deep-learning based methods address this by predicting articulation properties for objects from raw observations, such as depth images [11–14] or PointCloud data [15, 16]. However, the majority of these methods [11, 12, 15, 16] require knowledge of the articulation model category for the object, e.g., whether it has a hinge joint (revolute) or a sliding joint (prismatic), which may not be available in many realistic settings. Alleviating this requirement, Jain et al. [14] introduced ScrewNet, which uses a unified representation based on screw transformations to represent different articulation types and performs category-independent articulation model estimation directly from raw depth images. However, ScrewNet [14] and related methods [11–13, 15, 16] only predict point estimates for an object's articulation model parameters. Nonetheless, reasoning about the uncertainty in the estimated parameters can provide significant advantages for ensuring success in robot manipulation

Submitted to the 5th Conference on Robot Learning (CoRL 2021). Do not distribute.

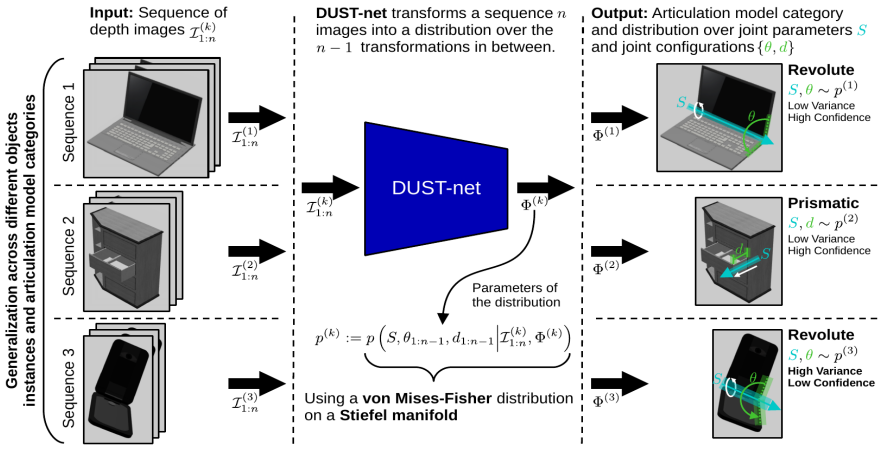

Figure 1: DUST-net uses a sequence of images $\mathcal{I}_{1:n}$ to compute the parameters, Φ, of the conditional distribution over the joint parameters S and configurations $\{\theta, d\}_{1:n-1}$. This distribution allows for inference and reasoning, such as uncertainty and confidence, over both the parameters and the configurations. Using a von Mises-Fisher distribution on a Stiefel manifold allows for an efficient reparameterization that inherently obeys multiple constraints that define rigid body transformations.

tasks, and allows for further advancements such as robust planning [1], active learning using human queries [17], and the learning of behavior policies that provide safety assurances [18]. Motivated by these advantages, we propose a method for learning articulation models, which estimates the uncertainty over model parameters using a novel distribution over the set of screw transformations by representing them as elements of a Stiefel manifold [19]. We introduce DUST-net, **D**eep **U**ncertainty estimation on **S**crew **T**ransforms-**net**work, a novel deep learning-based method that, in addition to providing point estimates of the object's articulation model parameters, leverages raw depth images to provide uncertainty estimates that can be used to guide the robot's behavior without requiring to knowledge of the object's articulation model category a priori.

DUST-net garners numerous benefits over existing methods. First, DUST-net estimates articulation properties for objects with uncertainty estimates, unlike current methods [11–16]. These uncertainty estimates, apart from helping robots to manipulate objects safely [1], can allow a robot to take information-gathering actions when it is not confident and enhance its chances of success in completing the task. Second, in contrast state-of-the-art methods [11, 12, 15, 16], DUST-net can estimate model parameters without the need to to know the articulation model category a priori, by leveraging the unified representation for different articulation model types. Third, the representation helps DUST-net to be more computationally and data-efficient than other state-of-the-art methods [11, 12], as it uses a single network to estimate model parameters for all common articulation models, unlike other methods that require a separate network for each articulation model category [11, 12, 15, 16]. Empirically, DUST-net outperforms other methods even when trained using only half the training data in comparison. Fourth, the distributional learning setting yields more robustness to outliers and noise. Fifth, DUST-net is able to reliably estimate distributions over articulation model parameters for objects in the robot's camera frame. By contrast, ScrewNet [14], the most closely related approach to ours, can only predict point estimates for articulation model parameters in the object's local frame.

We evaluate DUST-net through experiments on two benchmarking datasets: a simulated articulated objects dataset [11] and the PartNet-Mobility dataset [20–22], as well as three real-world objects: a microwave, a drawer, and a toaster oven. We compare DUST-net with two state-of-the-art methods, namely ScrewNet [14] and an MDN-based method proposed by Abbatematteo et al. [11], as well as two baseline methods. The experiments demonstrate that the samples drawn from the distributions learned by DUST-net result in significantly better estimates for articulation model parameters in comparison to the point estimates predicted by other methods. Additionally, the experiments show that DUST-net can successfully and accurately capture the uncertainty over articulation model parameters resulting from noisy inputs.

2 Related Work

Articulation model estimation from visual observations: A widely used approach for estimating articulation models is based on the probabilistic framework proposed by Sturm et al. [2]. It uses the time-series observations of 6D poses of different parts of an articulated object to learn the relationship between them [2, 5, 6, 10]. More recently, Abbatematteo et al. [11] and Li et al. [12] proposed methods to learn articulation properties for objects from raw depth images given articulation model category. In a related body of work on object parts mobility estimation, Wang et al. [15] and Yan et al. [16] proposed approaches to segment different parts of the object in an input point cloud and estimate their mobility relationships, given a known articulation model category. Alleviating the requirement of having a known articulation model category, Jain et al. [14] recently proposed ScrewNet that performs category-independent articulation model estimation from depth images. However, these methods only predict point estimates for the articulation model parameters, while DUST-net predicts a distribution over their values.

Rigid Body Pose Estimation: Our contributions are related to existing work on estimating distributions describing the orientation of rigid bodies. Gilitschenski et al. [23], Arun Srivatsan et al. [24], Srivatsan et al. [25] and Rosen et al. [26] propose strategies that can be used to estimate the rigid body transformation of an object using a combination of Bingham and Guassian distributions, and the von Mises-Fisher distribution, respectively. The mathematical model used by our approach is inspired by these works, but 1) extends them to also represent uncertainty over the configuration of articulated object components about screw axes, and 2) integrates them into a deep learning model that is capable of learning these configurations from raw depth images. In addition, while these approaches use distributions over orientations and rigid body transformations to produce estimates, DUST-net directly outputs a distribution that can be used to facilitate further applications such as uncertainty-aware behavior planning.

Interactive perception (IP): Katz and Brock [3] introduced IP as a method to leverage a robot's interaction with objects to generate a rich perceptual signal for articulation model estimation for planar objects, and extended it to learn 3D articulation models for objects [4]. Martín-Martín et al. [8] used hierarchical recursive Bayesian filters to make estimation more robust and developed online methods for articulation model estimation from RGB images [7–9]. A comprehensive survey on IP methods in robotics was presented by Bohg et al. [27]. While IP presents a powerful tool for estimating articulation properties for objects, a wide majority of existing IP methods require textured objects, unlike DUST-net, which learns these properties using depth images.

Further approaches: Articulation motion models can be viewed as geometric constraints imposed on multiple rigid bodies. Such constraints can be learned from human demonstrations by leveraging different sensing modalities [13, 28–31]. Recently, Daniele et al. [30] proposed a multimodal learning framework that incorporates both vision and natural language information for articulation model estimation. However, these approaches predict point estimates for the articulation model parameters, unlike DUST-net, which predicts a distribution over the articulation model parameters.

3 Preliminaries

Problem Formulation: Given a sequence of n depth images $\mathcal{I}_{1:n}$ of motion between two parts of an articulated object, we estimate the parameters of a probability distribution $p(\phi|\mathcal{I}_{1:n})$ representing uncertainty over the parameters ϕ of the articulation model $\mathcal{M}$ governing the motion between the two parts. Following Jain et al. [14], we define the model parameters ϕ as the parameters of the screw axis of motion, $\mathsf{S} = (\mathbf{l}, \mathbf{m})$, where both $\mathbf{l}$ and $\mathbf{m}$ are elements of $\mathbb{R}^3$. This unified parameterization can be used in articulation models with at most one degree-of-freedom (DoF), namely rigid, revolute, prismatic, and prismatic [14]. Additionally, we estimate the parameters of a distribution $p(q_{1:n-1}|\mathcal{I}_{1:n})$ representing uncertainty over the configurations $q_{1:n-1}$ identifying the rigid body transformations between the two parts in the given sequence of images $\mathcal{I}_{1:n}$ under model $\mathcal{M}$ with parameters ϕ. Configurations $q_i, i \in \{1...n-1\}$ correspond to a set of tuples, $q_i = (\theta_i, d_i)$, defining a rotation around and a displacement along the screw axis S. We assume that the relative motion between the two object parts is determined by a single articulation model.

Background: DUST-net uses a reparameterization of the space of rigid body transformations that allows distributions over an object's articulation model parameters to be defined naturally. Here, we

briefly describe the mathematical foundation leveraged in the proposed distribution over articulation parameters.

Screw Transformations: Chasles' theorem states that *"Any displacement of a body in space can be accomplished by means of a rotation of the body about a unique line in space accompanied by a translation of the body parallel to that line"* [32]. Such a line is called a screw axis, S. We represent this line using Plücker coordinates, given as $(\mathbf{l}, \mathbf{m})$ for a $l = \mathbf{p} + x\mathbf{l}$, with moment vector $\mathbf{m} = \mathbf{p} \times \mathbf{l}$, [32, 33]. The constraints $\|\mathbf{l}\| = 1$ and $\langle \mathbf{l}, \mathbf{m} \rangle = 0$ ensure that the degrees of freedom of the line in space are restricted to four. The rigid body displacement in $SE(3)$ as a screw transform is then defined as $\sigma = (\mathbf{l}, \mathbf{m}, \theta, d)$, where the linear displacement d and the rotation θ are connected through the pitch h of the screw axis, $d = h\theta$.

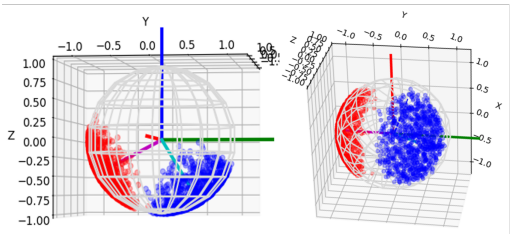

Figure 2: Matrix von Mises-Fisher distribution over $\mathrm{V}_{3,2}$, X, Y, Z axes are shown in red, blue and green colors, respectively. Magenta and cyan colors denote vectors corresponding to the first and second column of the matrix $M \in \mathrm{V}_{3,2}$ representing the mode of the distribution

Stiefel manifold: The *Stiefel manifold* $\mathrm{V}_{k,m}$ is the space whose points are sets of k orthonormal vectors in $\mathbb{R}^m$, called k-frames in $\mathbb{R}^m$ ($k \leq m$) [19]. Points on the Stiefel manifold $\mathrm{V}_{k,m}$ are represented by the set of $m \times k$ matrices X such that $X^T X = I_k$, where I_k is the $k \times k$ identity matrix; thus $\mathrm{V}_{k,m} = \{X_{m,k}; X^T X = I_k\}$. Some special cases of the Stiefel manifold are the unit hypersphere $\mathrm{V}_{1,m}$ in $\mathbb{R}^m$ for $k = 1$, and the orthogonal group $O(m)$ for $m = k$.

Matrix von Mises-Fisher distribution: A random matrix X on $\mathrm{V}_{k,m}$ is said to have the matrix von Mises-Fisher distribution (or matrix Langevin distribution), if its density function is given by $\mathcal{F}(\mathbf{X}|m, \mathbf{F}) = \dfrac{1}{{}_0F_1(\frac{m}{2}, \frac{1}{4}\mathbf{F}^T\mathbf{F})} \exp(\mathrm{Tr}(\mathbf{F}^T\mathbf{X}))$, where $\mathbf{F}$ is any $m \times k$ matrix and ${}_0F_1$ is a hypergeometric function with matrix argument [19]. We can write the general (unique) singular value decomposition (SVD) of $\mathbf{F}$ as $\mathbf{F} = \Gamma\Lambda\Omega^T$, where $\Gamma \in \tilde{\mathrm{V}}_{k,m}$, $\Omega \in O(k)$, $\Lambda = diag(\lambda_1, ..., \lambda_k)$, $\lambda_1 \geq ... \geq \lambda_k \geq 0$, $\tilde{\mathrm{V}}_{k,m}$ denotes the set of matrices $\Gamma \in \mathrm{V}_{k,m}$ with the property that all the elements of the first row of the matrix Γ are positive, and $O(k)$ denoting the orthogonal group in k dimensions. It can be shown that ${}_0F_1(\frac{m}{2}, \frac{1}{4}\mathbf{F}^T\mathbf{F}) = {}_0F_1(\frac{m}{2}, \frac{1}{4}\Lambda^2)$. For more details, we refer to [19].

4 Approach

Given a sequence of depth images $\mathcal{I}_{1:n}$ of motion between two parts of an articulated object, DUST-net estimates parameters of the joint probability distribution $p(\phi, q_{1:n-1}|\mathcal{I}_{1:n})$ representing uncertainty over the articulation model parameters ϕ governing the motion between the two parts and the observed configurations $q_{1:n-1}$. When deciding how to learn this distribution, two goals arise. While some parameters, such as the translation of an object part along a screw axis, are defined on Euclidean space, the set of valid screw axes exhibits constraints that prevent standard distributions defined on $\mathbb{R}^6$ from being applied without complicating the learning process. For example, a standard representation for distributions over screw axes is as the product of a Bingham distribution over the line's orientation and a multivariate normal distribution over its position in space [32]. However, this representation produces non-unique estimation targets. A rotation of θ about the screw axis with orientation $\mathbf{l}$ results in the same transformation as a rotation of $-\theta$ about the screw axis with orientation $-\mathbf{l}$. Similarly, a displacement d along $\mathbf{l}$ results in the same transformation as a displacement $-d$ along $-\mathbf{l}$. This leads to ambiguities in the targets in the estimation problem and can hinder the performance of the trained estimator. By selecting a representation that accounts for these symmetries, these non-unique estimation targets are removed. Second, once a suitable parameterization is chosen, we seek a parametric form for the joint distribution which can be learned by a deep network.

First, we consider the problem of parameterizing the set of screw axes. As noted earlier, we define the model parameter ϕ as the parameters of the screw axis of motion $\mathsf{S} = (\mathbf{l}, \mathbf{m})$. However, this parameterization requires that $\mathbf{l}$ has unit norm, and that $\mathbf{l}$ and $\mathbf{m}$ are orthogonal. To eliminate these constraints, we rewrite the moment vector of a screw axis as $\mathbf{m} = \|\mathbf{m}\| \hat{\mathbf{m}}$, where $\|\mathbf{m}\|$ and $\hat{\mathbf{m}}$

represent its magnitude and a unit vector along it respectively, and the Plücker coordinates for the screw axis as $\mathsf{S} = (\mathbf{l}, \hat{\mathbf{m}}, \|\mathbf{m}\|)$. The Plücker coordinates can then be seen as an unconstrained point in the space $\mathbb{S} := \mathrm{V}_{2,3} \times \mathbb{R}^+$, where $(\mathbf{l}, \hat{\mathbf{m}}) \in \mathrm{V}_{2,3}$ and $\|\mathbf{m}\| \in \mathbb{R}^+$. Consequently, a distribution over screw transformations can be defined as the product of any probability distribution defined on $\mathrm{V}_{2,3}$ and a distribution over $\mathbb{R}^+$. Importantly, because of the one-to-one mapping from elements of $\mathrm{V}_{2,3} \times \mathbb{R}^+$ to screw axes, the non-unique estimation targets described above are eliminated. Based on this parametrization of screw axes, we define the set of valid configuration parameters as follows. We restrict the range of values for the rotation about the screw axis to be $\theta \in [0, 2\pi)$ and restrict the displacement along the axis to be $d \in \mathbb{R}^+$. Note that these constraints do not reduce the representational power of the screw transform $(\mathbf{l}, \mathbf{m}, \theta, d)$ to denote a general rigid body transform, but merely ensure a unique representation.

Having described the parameterization of the set of screw axes and configurations, we now consider the task of defining a joint probability distribution over their values. We propose to represent the distribution over predicted screw axis parameters, $p(\mathsf{S} \mid \mathcal{I}_{1:n})$ whit $\mathsf{S} \in \mathbb{S}$, as a product of a matrix von Mises-Fisher distribution $\mathcal{F}(\cdot|3, \mathbf{F})$ defined on the Stiefel manifold $\mathrm{V}_{2,3}$ and a truncated normal distribution $\mathcal{N}^+(\cdot|\mu, \sigma)$ with truncation interval $[0, +\infty)$ over $\mathbb{R}^+$. Formally,

$$p(\mathsf{S} \mid \mathcal{I}_{1:n}) = p\left(l, \hat{\mathbf{m}}, \|\mathbf{m}\| \;\middle|\; \mathcal{I}_{1:n}, \mathbf{F}, \mu_{\mathbf{m}}, \sigma^2_{\mathbf{m}}\right) = \mathcal{F}\left(\mathbf{l}, \hat{\mathbf{m}} \mid 3, \mathbf{F}\right) \mathcal{N}^+\left(\|\mathbf{m}\| \mid \mu_{\mathbf{m}}, \sigma^2_{\mathbf{m}}\right), \quad (1)$$

where $\mathbf{F}$ is a 3×2 matrix representing the parameters of the matrix von Mises-Fisher distribution over $\mathrm{V}_{2,3}$, and $\mu_{\mathbf{m}}$ and $\sigma_{\mathbf{m}}$ denote the mean and standard deviation of the truncated normal distribution.

Given the sequence of n images, we also wish to estimate the posterior over configurations $q_{1:n-1} = \{\theta_{1:n-1}, d_{1:n-1}\}$ corresponding to the rotations about and displacements along the screw axis S. We define the joint posterior representing the uncertainty over the screw axis S and the configurations $\{\theta_{1:n-1}, d_{1:n-1}\}$ about it as a product of the aforementioned distribution and a set of distributions defined over the configuration parameters,

$$p(\mathsf{S}, \theta_{1:n-1}, d_{1:n-1} \mid \mathcal{I}_{1:n}, \Phi) = p(\mathsf{S}; \mathbf{F}, \mu_{\mathbf{m}}, \sigma^2_{\mathbf{m}})\; \Psi(\theta_{1:n-1}; \psi)\; \Upsilon(d_{1:n-1}; \upsilon) \quad (2)$$

where $\Phi = \{\mathbf{F}, \mu_{\mathbf{m}}, \sigma^2_{\mathbf{m}}, \psi, \upsilon\}$ is the set of parameters for the distribution and Ψ and Υ represent the set of distributions having parameters ψ and υ over the configurations $\theta_{1:n-1}$ and $d_{1:n-1}$, respectively. In this work, we consider Ψ and Υ to be products of truncated normal distributions such that $\Psi = \prod_{i=1}^{n-1} \mathcal{N}^+(\theta_i|\mathrm{M}^i_\theta, \sigma^2_\theta)$ and $\Upsilon = \prod_{i=1}^{n-1} \mathcal{N}^+(d_i|\mathrm{M}^i_d, \sigma^2_d)$ with $\mathrm{M}_\theta = \{\mu^1_\theta, ..., \mu^{n-1}_\theta\}$, $\mathrm{M}_d = \{\mu^1_d, ..., \mu^{n-1}_d\}$, σ_θ, and σ_d denoting the set of means and the standard deviations of the set of truncated normal distributions over the configurations $\theta_{1:n-1}$ and $d_{1:n-1}$, respectively.

Distribution parameter matrix F: The parameter matrix for the matrix von Mises-Fisher distribution over $\mathrm{V}_{3,2}$ is a 3×2 matrix, $\mathbf{F}$. This presents two possible parameterizations of the matrix: first, to estimate each of the 6 elements of the 3×2 matrix $\mathbf{F}$ and second, to estimate the matrices Γ, Λ, and Ω defining the SVD of $\mathbf{F}$, given by $\mathbf{F} = \Gamma\Lambda\Omega^T$. The second parameterization decouples the two objectives of distribution mode alignment with the ground truth labels and uncertainty representation; the mode of the distribution is given by $M = \Gamma\Omega^T$, and the concentration matrix for the distribution is given by $K = \Omega\Lambda\Omega^T$. This decoupling allows the network to independently optimize both objectives, whereas in the first parameterization, changes in the elements of $\mathbf{F}$ causes changes in both components.

By definition, Λ is a 2×2 diagonal matrix with two independent parameters, and $\Omega \in O(2)$ is a rotation matrix in two dimensions with one independent parameter, the rotation angle ω. The matrix $\Gamma \in \tilde{\mathrm{V}}_{3,2}$ can be constructed from a rotation matrix $\mathrm{R} \in O(3)$ by keeping only the first two columns of R. Hence, the matrix Γ can be defined by three independent Euler angles, (α, β, γ) denoting rotation according to the ZYX convention in the rotating frame. Euler angles can suffer from the problem of gimble lock [32], which we resolve by restricting the Euler angles to be in the ranges $\alpha \in [0, 2\pi), \beta \in [0, \pi)$, and $\gamma \in [0, 2\pi)$.

Normalization factor: One of the main challenges of using the matrix von Mises-Fisher distribution is the calculation of its normalization factor ${}_0F_1(\frac{m}{2}, \frac{1}{4}\Lambda^2)$, which is a hypergeometric function of matrix argument [19]. In this work, we approximate this hypergeometric function using a truncated series in terms of zonal polynomials, which are multivariate symmetric homogeneous polynomials and form a basis of the space of symmetric polynomials [19]. Through our preliminary experiments, we found that this truncated series is a good approximation of ${}_0F_1$ as it converges to a finite value, if the singular values of the F, i.e. λ_1 and λ_2 are less than $\lambda_{max} = 50$.

Architecture: DUST-net sequentially connects a ResNet-18 CNN [34] and a 2-layer MLP. ResNet-18 extracts task-relevant features from the input images, which are used by the MLP to predict a set of parameters Φ for the distribution $p(\mathsf{S}, \theta_{1:n-1}, d_{1:n-1} \mid \mathcal{I}_{1:n}, \Phi)$. The network is trained end-to-end, with ReLU activations for the hidden fully-connected layers. The first four output (out of 40) of the last linear layer of MLP, corresponding to the parameters (α, β, γ) and ω representing the matrices Γ and Ω respectively, are fed through a ReLU-6 layer to ensure that the predictions map to their respective ranges. Remaining output is fed through a Softplus layer for non-negative output.

Training: The training data for the model consists of sequences of depth images of objects parts moving relative to each other and the corresponding screw transforms $\mathbf{y} = (l, \hat{\mathbf{m}}, \|\mathbf{m}\|, \theta_{1:n-1}, d_{1:n-1})$. The objects and depth images are rendered in Mujoco [35]. We train DUST-net by maximizing the log-probability of the labels $\mathbf{y}$ under the distribution $p(\mathbf{y} \mid \mathcal{I}_{1:n}, \Phi)$: $\mathcal{L}(\mathbf{y}, \Phi) = -\log p(\mathbf{y} \mid \Phi)$. We assume that the observed configurations in $\mathcal{I}_{1:n}$ share the same variance. We use the precision parameters rather than the standard deviations, $\sigma_{\mathbf{m}}, \sigma_\theta$ and σ_d to represent the distribution during training for better numerical stability. Following the discussion on training MDNs by Makansi et al. [36], we separate the training in three stages. In the first stage, we assume the dispersion of the matrix von Mises-Fisher distribution to be fixed with $\Lambda = \mathrm{diag}(\lambda_0, \lambda_0),\ \lambda_0 = 1$ and learn parameters corresponding to Γ and Ω matrices. In the second stage, we fix the Λ matrix and learn the rest of the parameters in the set Φ. Finally, we train to predict the complete set Φ.

5 Experiments

In this section, we evaluate DUST-net on its ability to learn articulation model parameters and uncertainty estimates. We conducted three sets of experiments evaluating DUST-net's performance under different criteria: (1) how accurate point estimates of the articulation model parameters drawn from DUST-net's estimated distribution are in comparison to the existing methods, (2) how effectively DUST-net captures the uncertainty over parameters arising from noisy input, and (3) how effectively DUST-net transfers from simulation to a real-world setting. We evaluated DUST-net's performance on two simulated benchmarking datasets: the objects dataset provided by Abbatematteo et al. [11], and the PartNet-Mobility dataset [20–22], as well as a set of three real-world objects. From the simulated articulated object dataset [11], we considered the cabinet, microwave, and toaster oven for revolute articulations and the drawer object class for prismatic articulations. From the PartNet-Mobility dataset [20–22], we considered five object classes: the dishwasher, oven, and microwave object classes for the revolute articulation model category, and the storage furniture object class consisting of either a single column of drawers or multiple columns of drawers, for the prismatic articulation model category. Among the three sets of experiments, we conducted the first two sets of experiments on the simulated datasets, while the last set of experiments were conducted on the real-world object dataset. In all the experiments, we assumed that the input depth images are semantically segmented and contain non-zero pixels corresponding only to the two objects between which we wish to estimate the articulation model.

We compared DUST-net's performance in estimating point estimates for articulation model parameters with two state-of-the-art methods, ScrewNet [14] and an MDN-based approach proposed by Abbatematteo et al. [11]. ScrewNet estimates the object's articulation model parameters in a local frame located at the center of the object, whereas DUST-net does so directly in the camera frame. We compare our method with ScrewNet predicting parameters both in the object local frame and the camera

	MAAD / SL	MAAD		Screw Loss	MAAD	SL	MAAD	SL	Uncertainty				
	$\mathbf{l}$	$\hat{\mathbf{m}}$	$\|\mathbf{m}\|$	$D(\mathsf{S}_{GT}, \mathsf{S}_{pred})$	θ_i	$\theta_{\mathbf{l}}$	d_i	$d_{\mathbf{l}}$	$\lambda_{\mathbf{l}}$	$\lambda_{\hat{\mathbf{m}}}$	$\beta_{\|\mathbf{m}\|}$	β_θ	β_d
vm-SoftOrtho	**0.139**	**0.154**	0.068	0.956	**0.012**	**0.117**	0.003	0.006	**56.2**	**55.8**	9.8	47.9	89.5
Direct **F**	0.240	0.261	0.062	0.104	0.010	0.208	0.002	0.006	8.4	7.9	9.8	48.5	75.3
ScrewNet	0.846	0.929	0.486	0.475	0.115	0.217	0.111	0.118	-	-	-	-	-
Abbatematteo et al. [11]	0.194	-	-	0.111	0.223	-	0.045	-	-	-	-	-	-
DUST-net	0.151	0.163	**0.052**	**0.059**	**0.012**	0.122	**0.002**	**0.006**	53.8	54.0	**18.3**	**128.1**	**219.1**
ScrewNet (Local)	0.178	0.443	0.068	0.033	0.057	0.118	0.015	0.015	-	-	-	-	-

Table 1: Mean error values on the MAAD and Screw Loss(SL) metrics for the simulated articulated objects dataset [11]. Point estimates obtained using DUST-net result in better estimates for articulation model parameters in comparison to other methods. In all tables, angular values $\{\mathbf{l}, \hat{\mathbf{m}}, \theta_i, \theta_{\mathbf{l}}\}$ and distances $\{\|\mathbf{m}\|, D, d_i, d_{\mathbf{l}}\}$ are reported in radian and meter, respectively. Numerical values are reported for the uncertainty parameters $\{\lambda_i, \beta_j\}$. Symbol $-$ represents value not reported.

	MAAD / SL	MAAD		Screw Loss	MAAD	SL	MAAD	SL	Uncertainty				
	$\mathbf{l}$	$\hat{\mathbf{m}}$	$\lVert\mathbf{m}\rVert$	$D(\mathsf{S}_{GT}, \mathsf{S}_{pred})$	θ_i	$\theta_{\mathbf{l}}$	d_i	$d_{\mathbf{l}}$	$\lambda_{\mathbf{l}}$	$\lambda_{\hat{\mathbf{m}}}$	$\beta_{\lVert\mathbf{m}\rVert}$	β_θ	β_d
vm-SoftOrtho	0.284	0.243	0.221	1.137	0.030	0.086	0.012	0.027	26.9	31.1	5.7	54.5	60.9
Direct **F**	**0.214**	**0.212**	0.257	0.219	0.030	0.064	0.012	**0.024**	8.1	7.3	4.9	59.5	70.9
ScrewNet	0.846	0.929	0.486	0.475	0.115	0.217	0.111	0.118	-	-	-	-	-
Abbatematteo et al. [11]	0.989	-	-	**0.095**	0.141	-	0.085	-	-	-	-	-	-
DUST-net	0.220	0.219	**0.178**	0.189	**0.029**	**0.063**	**0.012**	0.029	**49.3**	**48.3**	**7.7**	**72.0**	**131.9**
ScrewNet (Local)	0.260	1.23	0.314	0.151	0.060	0.106	0.040	0.009	-	-	-	-	-

Table 2: Mean error values on the MAAD and Screw Loss(SL) metrics for the PartNet-Mobility dataset [20–22]. Point estimates obtained using DUST-net result in better estimates for articulation model parameters in comparison to other methods.

	λ_1	λ_2	$\beta_{\lVert\mathbf{m}\rVert}$	β_θ	β_d	λ_1	λ_2	$\beta_{\lVert\mathbf{m}\rVert}$	β_θ	β_d	λ_1	λ_2	$\beta_{\lVert\mathbf{m}\rVert}$	β_θ	β_d	λ_1	λ_2	$\beta_{\lVert\mathbf{m}\rVert}$	β_θ	β_d
Label Noise	No noise					15	15	50	50	50	12	12	50	50	50	10	10	50	50	50
SynArt	53.8	53.9	18.3	128.0	219.0	8.2	8.2	14.6	53.7	51.9	6.8	6.8	10.5	41.6	49.6	3.8	3.8	10.3	41.9	47.4
PartNet	49.3	48.3	7.7	72.0	132.0	6.4	6.3	9.4	29.5	29.2	4.9	4.7	8.9	34.0	37.9	3.2	3.1	9.4	31.2	32.1

Table 3: Testing variation of DUST-net's confidence over predicted articulation model parameters with input noise. DUST-net's confidence over its predicted parameters decreases monotonically as input noise is increased showing that DUST-net's predicted distribution captures the network's confidence over the predicted articulation parameters effectively.

frame. Additionally, we propose two baseline methods that estimate distributions over articulation model parameters and compare to them DUST-net. The first baseline method (vm-SoftOrtho) can be viewed as an extension of ScrewNet to a distributional setting. It represents the uncertainty over the screw axis orientation vector $\mathbf{l}$ and the direction of moment vector $\hat{\mathbf{m}}$ using two independent von Mises-Fisher distributions and imposes a soft orthogonality constraint over the modes of the two distributions. The distributions over the moment vector magnitude $\lVert\mathbf{m}\rVert$ and configurations $q_{1:n-1}$ are considered to be normal distributions. This method suffers from the same drawback as ScrewNet, i.e., the use of a soft orthogonality constraint during training, and therefore cannot predict a valid set of screw axis parameters directly, unlike DUST-net. The second baseline method (Direct F) uses the same probability distribution as DUST-net to represent the uncertainty over the articulation model parameters, but estimates the individual elements of the $\mathbf{F}$ matrix directly. As a result, it fails to capture the uncertainty over model parameters accurately.

5.1 Accuracy of Point Estimates

The first set of experiments evaluated DUST-net's accuracy in predicting point estimates for articulation model parameters. We use the mode of the estimated distribution as the point estimate for model parameters. We used two metrics to evaluate accuracy: Mean Absolute (Angular) Deviation (MAAD) and Screw Loss (Metric proposed in ScrewNet [14]). MAAD metric indicates how close the individual screw parameters are to targets, whereas the Screw Loss indicates how close the complete predicted screw transforms is to the target transforms. The MAAD metric calculates the angular distance between the orientation of the predicted and ground-truth axis orientation vectors $\mathbf{l}$ and the orientation vectors of the screw axis moment vectors $\hat{\mathbf{m}}$. For the remaining parameters ($\lVert\mathbf{m}\rVert$, $\theta_{1:n-1}, d_{1:n-1}$), it calculates the mean absolute deviation between the predicted and ground-truth values. The screw loss reports the angular distance between the predicted and ground-truth screw axis orientation vectors $\mathbf{l}$ as orientation error and the length of the shortest perpendicular between the predicted and ground-truth screw axes as the distance between them. Configuration errors $\theta_{1:n-1}$ are reported as the difference between the predicted rotation about the predicted screw axis and the true rotation, whereas errors over $d_{1:n-1}$ are calculated as the Euclidean distance between the points displaced by the predicted and true displacements along respective axes.

Results for the synthetic articulated objects dataset and the PartNet-Mobility dataset are shown in Tables 1 and 2, respectively. Results demonstrate that under both metrics, the estimates obtained from DUST-net are considerably more accurate than those obtained from the state-of-the-art methods. DUST-net also correctly estimates very high distribution concentration parameters for the true, noise-free labels. The first baseline, vm-SoftOrtho, performs comparably with DUST-net on both datasets when only MAAD estimates are considered. However, Tables 1 and 2 show that it produces a very high distance ($\approx$ 1m) between the predicted and ground-truth screw axes. This error arises due to the soft-orthogonality constraint used by vm-SoftOrtho, as DUST-net and the second baseline method, both of which handle the constraint implicitly, do not report high errors on that metric. Meanwhile, the second baseline, Direct F, performs comparably with DUST-net on both metrics for both datasets, but fails to capture the uncertainty over parameters with the required accuracy.

		MAAD / SL	MAAD		Screw Loss	MAAD	SL	MAAD	SL	Uncertainty				
		$\mathbf{l}$	$\hat{\mathbf{m}}$	$\lVert\mathbf{m}\rVert$	$D(\mathsf{S}_{GT}, \mathsf{S}_{pred})$	θ_i	$\theta_\mathbf{l}$	d_i	$d_\mathbf{l}$	$\lambda_\mathbf{l}$	$\lambda_{\hat{\mathbf{m}}}$	$\beta_{\lVert\mathbf{m}\rVert}$	β_θ	β_d
Toaster	ScrewNet	2.42	2.48	0.74	0.76	0.45	1.26	0.01	0.00	-	-	-	-	-
Oven	DUST-net	**0.17**	**0.31**	**0.52**	**0.59**	**0.44**	**0.64**	**0.01**	**0.01**	2.5	0.1	11.6	10.8	75.5
Microwave	ScrewNet	0.79	0.81	**0.13**	0.52	1.19	0.54	0.01	0.01	-	-	-	-	-
	DUST-net	**0.41**	**0.42**	0.22	**0.43**	**0.46**	**0.40**	**0.00**	**0.00**	0.7	0.6	19.7	14.3	39.9
Drawer	ScrewNet	0.69	**0.24**	0.49	**0.24**	**0.72**	0.97	0.08	**0.08**	-	-	-	-	-
	DUST-net	**0.42**	0.50	**0.32**	0.74	0.75	**0.56**	**0.07**	**0.08**	0.2	0.1	12.3	31.6	55.2

Table 4: Mean error values on the MAAD and Screw Loss metric for estimation of articulation model parameters for real-world objects when network was trained solely using simulated data

5.2 Uncertainty Estimation

The second set of experiments evaluated how effectively DUST-net's predicted distribution captures the network's confidence over the predicted articulation parameters. We evaluate this by adding artificial noise to the training labels form the two simulated datasets while training DUST-net. As more noise is added, we expect the confidence estimates produces by DUST-net to decrease as well. We add noise to the labels by sampling perturbations from a matrix von Mises-Fisher distribution with varying singular values λ_1 and λ_2 of the distribution parameter matrix $\mathbf{F}$ and the truncated normal distributions with varying precision parameters $\beta_j, j \in \{\lVert\mathbf{m}\rVert, \theta, d\}$. Results are shown in Table 3. In the noiseless case, the singular values of the matrix von Mises-Fisher distribution increases until they reach their maximum allowed value at $\lambda_{max} = 50$, while the precision parameters $\beta_j, j \in \{\lVert\mathbf{m}\rVert, \theta, d\}$ for truncated normal distributions over remaining parameters become arbitrarily large. When label noise is added, our results show that DUST-net's confidence over its predicted parameters decreases monotonically as more noise is added to the labels, supporting our hypothesis.

5.3 Transferring knowledge to Real-world objects

Lastly, we evaluated how effectively DUST-net transfers from simulation to a real-world setting. DUST-net was trained solely on the simulated articulated object dataset [11]. Afterward, we used it to infer the articulation model parameters for three real-world objects. Results (Table 4) report that while DUST-net outperforms ScrewNet in estimating the model parameters for real-world objects, the estimated parameters are not yet accurate enough to be used directly for manipulating these objects. This sub-par performance stems from the significant differences between the training (clean and information-rich simulation data) and test datasets, which consists of noisy depth images acquired with a Kinect sensor and contain high salt-and-pepper noise, spurious features, and incomplete objects. However, a noteworthy insight from the results is that DUST-net also reports very low confidence over the predicted parameters. This clearly delineates why it is beneficial to estimate a distribution over the articulation model parameters instead of plain point estimates. For cases when the input data is noisy and lacks sufficient information, a method predicting point estimates such as ScrewNet yield poor estimates and a robot will have no way to know if the estimates are reliable for manipulating the object or not. In contrast, DUST-net's reported confidence over the predictions allows the robot to develop safe motion policies for articulated objects [1, 18].

6 Conclusion

We introduced DUST-net, which utilizes a novel distribution over screw transforms on a Stiefel manifold to perform category-independent articulation model estimation with uncertainty estimates. We evaluated our approach on two benchmarking datasets and three real-world objects and compared its performance with two current state-of-the-art methods [11, 14]. Results show that DUST-net can estimate articulation models, their parameters, and model uncertainty estimates for novel objects across articulation model categories successfully with better accuracy than the state-of-the-art methods. At present, DUST-net can only predict parameters for 1-DOF articulation models directly. For multi-DoF objects, an additional image segmentation step is required to mask out all non-relevant object parts. This procedure can be repeated iteratively for all object part pairs to estimate relative models between object parts that can be combined later to construct a complete kinematic model for the object [10]. An interesting extension of DUST-net could estimate parameters for multi-DoF objects directly by learning a segmentation network along with it. Another exciting direction of future work is to use DUST-net in an active learning setting where, if the robot is not confident enough about the estimates of the articulation model parameters, it can actively take information-gathering actions to reduce uncertainty.

References

[1] A. Jain and S. Niekum. Efficient hierarchical robot motion planning under uncertainty and hybrid dynamics. In *Conference on Robot Learning*, pages 757–766, 2018.

[2] J. Sturm, C. Stachniss, and W. Burgard. A probabilistic framework for learning kinematic models of articulated objects. *Journal of Artificial Intelligence Research*, 41:477–526, 2011.

[3] D. Katz and O. Brock. Manipulating articulated objects with interactive perception. In *2008 IEEE International Conference on Robotics and Automation*, pages 272–277. IEEE, 2008.

[4] D. Katz, M. Kazemi, J. A. Bagnell, and A. Stentz. Interactive segmentation, tracking, and kinematic modeling of unknown 3d articulated objects. In *2013 IEEE International Conference on Robotics and Automation*, pages 5003–5010. IEEE, 2013.

[5] S. Niekum, S. Osentoski, C. G. Atkeson, and A. G. Barto. Online bayesian changepoint detection for articulated motion models. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1468–1475. IEEE, 2015.

[6] S. Pillai, M. R. Walter, and S. Teller. Learning articulated motions from visual demonstration. *arXiv preprint arXiv:1502.01659*, 2015.

[7] R. Martín-Martín and O. Brock. Online interactive perception of articulated objects with multi-level recursive estimation based on task-specific priors. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2494–2501. IEEE, 2014.

[8] R. Martín-Martín, S. Höfer, and O. Brock. An integrated approach to visual perception of articulated objects. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5091–5097. IEEE, 2016.

[9] R. Martín-Martín and O. Brock. Coupled recursive estimation for online interactive perception of articulated objects. *The International Journal of Robotics Research*, page 0278364919848850, 2019.

[10] A. Jain and S. Niekum. Learning hybrid object kinematics for efficient hierarchical planning under uncertainty. *IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[11] B. Abbatematteo, S. Tellex, and G. Konidaris. Learning to generalize kinematic models to novel objects. In *Proceedings of the Third Conference on Robot Learning*, 2019.

[12] X. Li, H. Wang, L. Yi, L. J. Guibas, A. L. Abbott, and S. Song. Category-level articulated object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3706–3715, 2020.

[13] Q. Liu, W. Qiu, W. Wang, G. D. Hager, and A. L. Yuille. Nothing but geometric constraints: A model-free method for articulated object pose estimation. *arXiv preprint arXiv:2012.00088*, 2020.

[14] A. Jain, R. Lioutikov, C. Chuck, and S. 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.

[15] X. Wang, B. Zhou, Y. Shi, X. Chen, Q. Zhao, and K. Xu. Shape2motion: Joint analysis of motion parts and attributes from 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8876–8884, 2019.

[16] Z. Yan, R. Hu, X. Yan, L. Chen, O. Van Kaick, H. Zhang, and H. Huang. Rpm-net: Recurrent prediction of motion and parts from point cloud. *ACM Trans. Graph.*, 38(6), Nov. 2019. ISSN 0730-0301. doi:10.1145/3355089.3356573. URL https://doi.org/10.1145/3355089.3356573.

[17] Y. Cui and S. Niekum. Active reward learning from critiques. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6907–6914. IEEE, 2018.

[18] A. Taylor, A. Singletary, Y. Yue, and A. Ames. Learning for safety-critical control with control barrier functions. In A. M. Bayen, A. Jadbabaie, G. Pappas, P. A. Parrilo, B. Recht, C. Tomlin, and M. Zeilinger, editors, *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pages 708–717, The Cloud, 10–11 Jun 2020. PMLR. URL http://proceedings.mlr.press/v120/taylor20a.html.

[19] Y. Chikuse. *Statistics on special manifolds*, volume 174. Springer Science & Business Media, 2003. doi:https://doi.org/10.1007/978-0-387-21540-2.

[20] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang, L. Yi, A. X. Chang, L. J. Guibas, and H. Su. SAPIEN: A simulated part-based interactive environment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[21] K. Mo, S. Zhu, A. X. Chang, L. Yi, S. Tripathi, L. J. Guibas, and H. 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)*, June 2019.

[22] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[23] I. Gilitschenski, G. Kurz, S. J. Julier, and U. D. Hanebeck. Unscented orientation estimation based on the bingham distribution. *IEEE Transactions on Automatic Control*, 61(1):172–177, 2015.

[24] R. Arun Srivatsan, M. Xu, N. Zevallos, and H. Choset. Probabilistic pose estimation using a bingham distribution-based linear filter. *The International Journal of Robotics Research*, 37 (13-14):1610–1631, 2018.

[25] R. A. Srivatsan, G. T. Rosen, D. F. N. Mohamed, and H. Choset. Estimating se (3) elements using a dual quaternion based linear kalman filter. In *Robotics: Science and systems*, 2016.

[26] D. M. Rosen, L. Carlone, A. S. Bandeira, and J. J. Leonard. Se-sync: A certifiably correct algorithm for synchronization over the special euclidean group. *The International Journal of Robotics Research*, 38(2-3):95–125, 2019.

[27] J. Bohg, K. Hausman, B. Sankaran, O. Brock, D. Kragic, S. Schaal, and G. S. Sukhatme. Interactive perception: Leveraging action in perception and perception in action. *IEEE Transactions on Robotics*, 33(6):1273–1291, 2017.

[28] C. Pérez-D'Arpino and J. A. Shah. C-learn: Learning geometric constraints from demonstrations for multi-step manipulation in shared autonomy. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 4058–4065. IEEE, 2017.

[29] Y. Liu, F. Zha, L. Sun, J. Li, M. Li, and X. Wang. Learning articulated constraints from a one-shot demonstration for robot manipulation planning. *IEEE Access*, 7:172584–172596, 2019.

[30] A. F. Daniele, T. M. Howard, and M. R. Walter. A multiview approach to learning articulated motion models. In *Robotics Research*, pages 371–386. Springer, 2020.

[31] G. Subramani, M. Zinn, and M. Gleicher. Inferring geometric constraints in human demonstrations. In *Conference on Robot Learning*, pages 223–236. PMLR, 2018.

[32] B. Siciliano and O. Khatib. *Springer handbook of robotics*. Springer, 2016.

[33] Y.-b. Jia. Plücker Coordinates for Lines in the Space [Lecture Notes], August 2019.

[34] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[35] E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.

[36] O. Makansi, E. Ilg, O. Cicek, and T. Brox. Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7144–7153, 2019.