FlowBot++: Learning Generalized Articulated Objects Manipulation via Articulation Projection

Harry Zhang, Ben Eisner, David Held
Robotics Institute, School of Computer Science
Carnegie Mellon University, United States
{haolunz, baeisner, dheld}@andrew.cmu.edu

Abstract: Understanding and manipulating articulated objects, such as doors and drawers, is crucial for robots operating in human environments. We wish to develop a system that can learn to articulate novel objects with no prior interaction, after training on other articulated objects. Previous approaches for articulated object manipulation rely on either modular methods which are brittle or end-to-end methods, which lack generalizability. This paper presents FlowBot++, a deep 3D vision-based robotic system that predicts dense per-point motion and dense articulation parameters of articulated objects to assist in downstream manipulation tasks. FlowBot++ introduces a novel per-point representation of the articulated motion and articulation parameters that are combined to produce a more accurate estimate than either method on their own. Simulated experiments on the PartNet-Mobility dataset validate the performance of our system in articulating a wide range of objects, while real-world experiments on real objects' point clouds and a Sawyer robot demonstrate the generalizability and feasibility of our system in real-world scenarios. Videos are available on our anonymized website here.

Keywords: Articulated Objects, 3D Learning, Manipulation

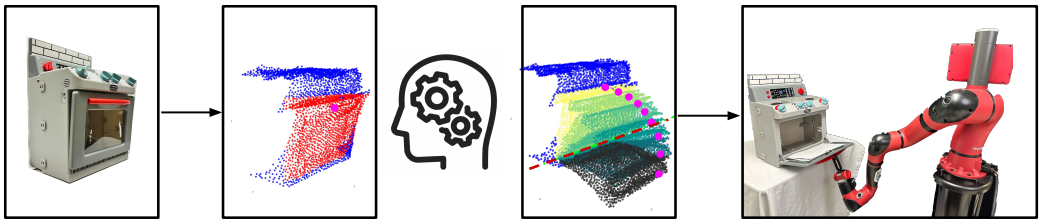


Figure 1: FlowBot++ in Action. The system first observes a point cloud observation of an articulated object and estimates the object's Articulation Flow and Articulation Projection to infer the articulation axis. Then the inferred axis is used to output a smooth trajectory to actuate the object.

1 Introduction

Understanding and manipulating articulated objects such as doors and drawers is a key skill for robots operating in human environments. Unlike previous work [1, 2] that learns to build the kinematic structure of an object through experience, we aim to teach a robot to manipulate novel articulated objects, transferring knowledge from prior experience with other articulated objects. While humans can manipulate novel articulated objects, constructing robotic manipulation agents that can generalize in the same way poses significant challenges, since the complex structure of such objects requires three-dimensional reasoning of their parts and functionality. Due to the large number of categories of such objects and intra-class variations of the objects' structure and kinematics, it is difficult to train perception and manipulation systems that can generalize across such variations.

One approach used in previous work estimates the kinematic tree of an articulated object, using a modular pipeline of segmentation, connectivity estimation, and articulation-parameter estimation [3, 4, 5]; however, such a modular pipeline can be brittle for unseen objects, since a failure in any one component will cause failures in downstream modules. To learn a generalizable perception and manipulation pipeline, robots need to be robust to variations in the articulated objects' geometries and kinematic structures. Another recent approach [6] predicts a dense per-point estimate of how each point would move under articulated motion, without needing to predict the part's articulation parameters. Experiments demonstrate that this approach shows superior generalization to unseen articulated objects. However, the motion estimation needs to be run at each time step, which often yields jerky motion, costly computation, and poor performance in the face of heavy occlusions.

Our approach combines the advantages of both of these approaches: we jointly predict dense per-point motion of how that point would move if the object were to be articulated[1] (similar to Eisner et al. [6]) as well as a dense per-point representation of the corresponding part's articulation parameters. These two predictions are then integrated to produce a more accurate estimate of the object's motion than either method on their own. Each per-point prediction is grounded in a reference frame centered on that point, which avoids the need for the network to make a prediction relative to an unknown, global coordinate system. By estimating per-point predictions, we leverage the advantages of prior work [6] that has shown that per-point predictions enable enhanced generalization to different object geometries and kinematics.

Our system, FlowBot++, leverages these predictions to produce a smooth sequence of actions that articulate the desired part on the object. We train a ***single*** 3D perception module across many object categories; we evaluate the system on zero-shot manipulation of novel objects, without requiring additional videos or interactions with each test object. We show that the trained model generalizes to a wide variety of objects – both in seen categories as well as unseen object categories. The contributions of this paper include:

1. A novel per-point 3D representation of articulated objects that densely models the objects' instantaneous motion under actuation as well as its articulation parameters.
2. A novel approach to integrate the per-point motion prediction and per-point articulation prediction into a single prediction that outperforms either prediction individually.
3. Demonstrations of a robot manipulation system (FlowBot++) that uses the aforementioned per-point predictions to manipulate novel articulated objects in a zero-shot manner. Our experiments evaluate the performance of our system in articulating a wide range of objects in the PartNet-Mobility dataset as well as in the real-world.

2 Related Work

Articulated Object Manipulation: Manipulation of unseen articulated objects and other objects with non-rigid properties remains an open research area due to the objects' complex geometries and kinematics. Previous work proposed manipulating such objects using analytical methods, such as the immobilization of a chain of hinged objects, constraint-aware planning, and relational representations [7, 3, 8]. With the development of larger-scale datasets of articulated objects [9, 10], several works have proposed learning methods based on large-scale simulation, supervised visual learning and visual affordance learning [11, 12, 13]. Several works have focused on visual recognition and estimation of articulation parameters, learning to predict the pose [4, 5, 14, 15, 16] and identify articulation parameters [1, 17] to obtain action trajectories. Statistical motion planning methods could also be applied in the scenarios of complex hinged objects manipulation [18, 19, 20]. More closely related to our work is that of Eisner et al. [6], which learns to predict per-point motion of articulated objects under instantaneous motion. While the per-point motion affordance learning of Eisner et al. [6] outperforms previous approaches, it does not explicitly model the articulation parameters. Thus, the affordance needs to be estimated in each time step, yielding jerky robot trajectories. Concur-

[1] Following the convention of Eisner et al. [6], the system predicts the motion under the "opening" direction, although it can be reversed to perform closing.

rently with our work, Nie et al. [2] proposed a similar idea of estimating articulation parameters from interactions to guide downstream manipulation tasks. However, our method does not require *a priori* interactions with the manipulated objects.

Flow for Policy Learning: Optical flow [21] is used to estimate per-pixel correspondences between two images for object tracking and motion prediction and estimation. Current state-of-the-art methods for optical flow estimation leverage convolutional neural networks [22, 23, 24]. Dong et al. [25], Amiranashvili et al. [26] use optical flow as an input representation to capture object motion for downstream manipulation tasks. Weng et al. [27] use flow to learn a policy for fabric manipulation. Previous work generalized optical flow beyond pixel space to 3D for robotic manipulation [6, 28, 29]. We learn a 3D representation for articulated objects, *Articulation Flow*, which describes per-point correspondence between two point clouds of the same object and *Articulation Projection*, a dense projection representation of the object's articulation parameters.

3 Background

In this paper, we study the task of manipulating an articulated object. We assume that an articulated object could be of two classes: prismatic, which is parameterized by a translational axis, or revolute, which is parameterized by a rotational axis. Eisner et al. [6] derived the optimal instantaneous motion to articulate objects using physics and we review their reasoning here to better motivate our method. Suppose we are able to attach a gripper to some point $p \in \mathcal{P}$ on the surface $\mathcal{P} \subset \mathbb{R}^3$ of a child link with mass m. At this point, the policy can apply a 3D force $\mathbf{F}$, with constant magnitude $||\mathbf{F}||$ to the object at that point. We wish to choose a contact point and force direction that maximize the acceleration of the articulation's child link. Both articulation types are parameterized via axes $\mathbf{v}(t) = \boldsymbol{\omega} t + v$, where $\boldsymbol{\omega}$ is a unit vector that represents the axis direction and v represents the origin of the axis[2] (dashed lines in Fig. 2). Based on Newton's Second Law, Eisner et al. [6] derived that for *prismatic joints*, a force $\mathbf{F}$ maximizes the articulated part's acceleration if the force's direction is parallel to the part's axis $\boldsymbol{\omega}$; the force's point of exertion does not matter as long as the direction is parallel to $\boldsymbol{\omega}$. For *revolute joints*, a force $\mathbf{F}$ maximizes the articulation part's acceleration when $\mathbf{F}$ is tangent to the circle defined by axis $\mathbf{v}$ and radius $\mathbf{r}$, which connects the point of exertion to the nearest point on the axis. Selecting the point that maximizes $\mathbf{r}$ produces maximal linear acceleration. Thus for revolute joints, the optimal choice for the force $\mathbf{F}$ is to pick a point on the articulated part that is farthest from the axis of rotation $\mathbf{v}$ and apply a force tangent to the circle defined by $\mathbf{r}$ and axis $\mathbf{v}$. Similar to Eisner et al. [6], we predict each point's **Articulation Flow**, defined as follows: for each point p on each link on the object, define a vector f_p in the direction of the motion of that point caused by an infinitesimal displacement $\delta\theta$ of the joint in the opening direction, normalized by the largest such displacement on the link:

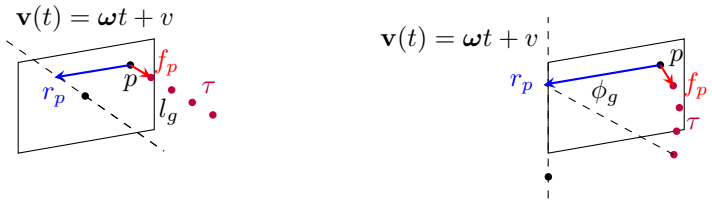
$$\text{Articulation Flow: } f_p = \begin{cases} \boldsymbol{\omega}, & \text{if prismatic} \\ \frac{\boldsymbol{\omega} \times \mathbf{r}}{||\boldsymbol{\omega} \times \mathbf{r}_{\max}||} & \text{if revolute} \end{cases} \tag{1}$$

where $\mathbf{r}$ is the radius for a contact point and $\mathbf{r}_{\max}$ represents the longest radius to the axis on a revolute object. Eisner et al. [6] showed how the predicted Articulation Flow can be used to derive a policy to manipulate previously unseen articulated objects. Below, we will treat the notation for the articulation axis $\mathbf{v}(t)$ and $\mathbf{v}$ interchangably.

4 FlowBot++: From New Representations to Smooth Trajectories

Eisner et al. [6] predicts the per-point Articulation Flow, which allows the method to avoid needing to directly predict the overall kinematic structure. However, the Articulation Flow needs to be re-predicted each timestep, since it captures the ideal instantaneous motion of the articulated part under an infitesimal displacement $\delta\theta$; after the robot takes a small action, the Articulation Flow

[2] There could be infinitely many possible prismatic axes, but the convention of CAD models is defining an axis through the centroid of the prismatic part.



(a) Prismatic (b) Revolute

Figure 2: For each point p on the object, Articulation Flow f_p [6] represents its instantaneous motion under a force in the opening direction; our new representation, Articulation Projection r_p, represents the displacement that projects p to the articulation axis $\mathbf{v}$. We train a network to predict both f_p and r_p and combine their predictions to get a smoother and more robust estimate. The purple points represent an interpolated prismatic trajectory of length l_g and an interpolated revolute trajectory of ϕ_g angle rotation. This corresponds to the trajectory prediction described in Sec. 4.2.

f_p for revolute joints will change direction, since the direction of motion caused by an infitesimal displacement $\delta\theta$ will have changed. Since the motion derived from Articulation Flow by construction only models small movements, this representation lacks the ability to predict a full, multi-step trajectory from the initial observation.

4.1 A New Representation of Articulated Objects

In order to overcome the limitations of previous work [6, 12], if we could estimate the articulation parameters $\mathbf{v}$ (or equivalently, $\boldsymbol{\omega}$ and v), then we would be able to derive the full trajectory of any arbitrary point p on the object would move under articulation of the object under kinematic constraints of $\mathbf{v}$. To achieve this, we formally define a 3D dense visual representation of articulation parameters on top of Articulation Flow; we illustrate each representation graphically in Fig. 2. In this work, we introduce another 3D representation that densely represents the object's articulation parameters, **Articulation Projection** - for each point on the articulated part of the object, we define a vector r_p that represents the displacement from the point itself to its projection onto the part's articulation axis $\mathbf{v}(t)$. For each point $p \in \mathbb{R}^3$ of articulated object $\mathcal{P}$, we define the Articulation Projection r_p mathematically as follows:

$$\text{Articulation Projection: } r_p = \text{proj}_{\mathbf{v}} p - p = \left(\boldsymbol{\omega}\boldsymbol{\omega}^T - \mathbf{I}\right)(p - v) \tag{2}$$

where $\boldsymbol{\omega}$ is a unit vector that represents the axis direction and v represents the origin of the axis and $\mathbf{I}$ represents a 3x3 identity matrix. Mathematically, r_p is calculated as the difference between the vector $\overrightarrow{vp}$ and its projection onto $\mathbf{v}$, which is equivalently *a perpendicular vector from each point to the axis of rotation defined by direction $\boldsymbol{\omega}$ and origin v*, which is represented using the blue vectors in Fig. 2. As we will see, Articulation Projection can be used to predict a longer articulation trajectory while still inheriting the generalization properties of prior work [6].

4.2 Manipulation via Learned Articulation Flow and Articulation Projection

As mentioned above, previous work [6] used the Articulation Flow f_p to compute the instantaneous motion direction, which needs to be re-predicted every time step. In contrast, if we were able to infer the articulation parameters (rotation axis $\boldsymbol{\omega}$ and its origin v), we could then analytically compute a multi-step articulation trajectory using the initial observation of the object without the need to stop and re-compute each step to actuate the articulated part as in previous work [6]. However, Eisner et al. [6] has shown that estimating the Articulation Flow results in more generalizable prediction than estimating the articulation parameters directly. We propose to estimate an articulation trajectory from the Articulation Projection. Because the Articulation Projection is a dense per-point prediction with a reference frame defined at each point (similar to Articulation Flow [6]), it inherits the generalization properties of Articulation Flow that allows it to be used for unseen articulated objects, thus hopefully obtaining the best of both worlds. Further, we will combine both predictions in Section 4.3 to obtain an even more accurate prediction.

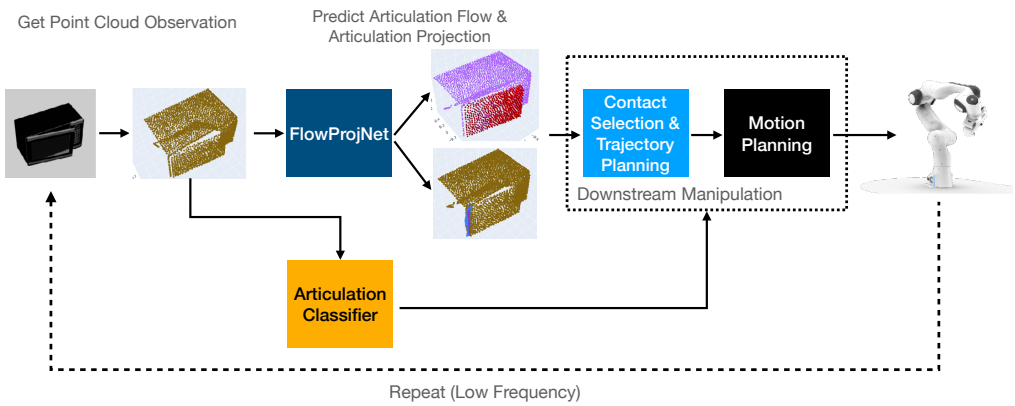


Figure 3: FlowBot++ System Overview. Our system in deployment first takes as input a partial point cloud observation of an articulated object (a microwave shown here) and uses FlowProjNet to jointly estimate the object's Articulation Flow (**top**) and Articulation Projection (**bottom**, points displaced by AP shown). The estimates will then be used in the downstream manipulation pipeline that interpolates and follows the planned trajectory smoothly. Unlike FlowBot3D [6], we do not repeat the estimation every step. Instead, we repeat this loop in a much lower frequency (once every H steps) to improve the smoothness of the planned trajectory.

For *revolute objects*, we infer the rotation axis using both the predicted Articulation Flow (Eq. 1) as well as the Articulation Projection (Eq. 2). We then estimate an H-step trajectory of the contact point under articulated motion in the opening direction. To obtain the articulation parameter $\boldsymbol{\omega}$, we calculate the cross product between the articulation flow f_p and the articulation projection r_p respectively as the articulation axis $\boldsymbol{\omega}$ should be perpendicular to both vectors. To compute the origin of the rotation, we use the point displaced by the articulation projection vector. For each point p, $p + r_p$ projects the point onto the rotation axis $\mathbf{v}$, effectively giving us an estimate of the origin of the rotation:

$$\hat{\boldsymbol{\omega}} = \frac{r_p \times f_p}{||r_p \times f_p||}, \quad \hat{v} = p + r_p \tag{3}$$

These predictions of $\hat{v}$ are visualized as the blue points in the bottom output branch of Fig. 3. Using the inferred axis direction $\hat{\boldsymbol{\omega}}$ and origin $\hat{v}$, we can interpolate a smooth trajectory: given a contact point p, the proposed trajectory should lie on a circle in a plane perpendicular to the inferred axis parameterized by the normalized direction $\hat{\boldsymbol{\omega}}$ and the origin $\hat{v}$. Formally, with Rodrigues' Formula of Lie algebra, given an angle of rotation ϕ about the normalized vector $\hat{\boldsymbol{\omega}}$ we are able to define a rotation matrix as follows:

$$\mathbf{R}(\phi) = \mathbf{I} + \sin\phi[\hat{\boldsymbol{\omega}}]_\times + (1 - \cos\phi)[\hat{\boldsymbol{\omega}}]_\times^2 \tag{4}$$

where $\mathbf{I}$ is identity matrix and $[\hat{\boldsymbol{\omega}}]_\times$ is the skew-symmetric matrix of $\hat{\boldsymbol{\omega}}$, and the rotated point p about this inferred axis by angle ϕ is given by: $p' = \mathbf{R}(\phi)(p - \hat{v}) + \hat{v}$. In order to obtain a trajectory, we interpolate K angles between 0 and a goal angle ϕ_g, and the smooth, K-step trajectory[3] that rotates the contact point p by angle ϕ_g about the estimated axis parameterized by $\hat{\boldsymbol{\omega}}$ and $\hat{v}$ is calculated via:

$$\tau_{\text{revolute}} = \left\{ \mathbf{R}\left(\frac{i}{K}\phi_g\right)(p - \hat{v}) + \hat{v} \right\}_{\forall i \in [0,K]} \tag{5}$$

We learn a separate articulation type classifier f_ψ which takes as input the point cloud observation of the object and classifies if the object is prismatic or revolute to guide the trajectory planning step. For *prismatic objects*, since the motion direction and the axis direction should be the same, we just use the normalized flow prediction f_p to compute the prismatic articulation axis, and we use the estimated axis to compute a smooth, K-step trajectory indexed by i to translate the contact point p

[3] In practice, the goal translation distance and goal angle is set large enough so that the termination condition (part being fully open) of the policy can be triggered.

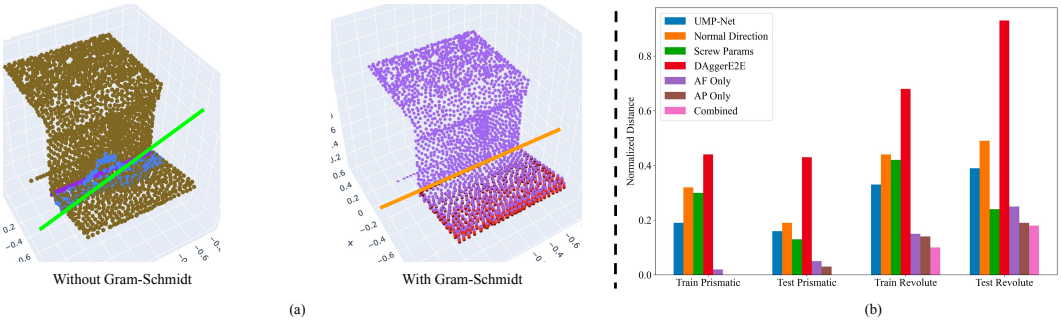


Figure 4: Gram-Schmidt Correction & Performance of Different Methods. In **(a)**, without Gram-Schmidt, the inferred articulation axis $\hat{\boldsymbol{\omega}}$ (green) is not accurate; blue points show the displacement of points by the Articulation Projection r_p. Using Gram-Schmidt, we use the Articulation Flow f_p (red vectors) to correct the axis direction; $\hat{\boldsymbol{\omega}}$ is now perpendicular to the Articulation Flow and aligns better with the ground-truth axis direction. In **(b)**, we show the bar plot of our method compared with several baseline methods on training and testing prismatic/revolute objects using normalized distance (↓). Note the performance gain after correction via Gram-Schmidt (AP Only vs Combined). Some values are not visible because they are < 0.05.

by a goal translation distance l_g:

$$\hat{\boldsymbol{\omega}} = \frac{f_p}{||f_p||}, \quad \tau_{\text{prismatic}} = \left\{ p + \frac{i}{K} l_g \hat{\boldsymbol{\omega}} \right\}_{i \in [0,K]} \tag{6}$$

In practice, to make the prediction more robust, we use a segmentation mask to average the predictions over all points in a given part to get an aggregated version of $\hat{\boldsymbol{\omega}}$ and $\hat{v}$. We also deploy an MPC-style controller that replans after taking the first H steps of trajectory for better performance.

4.3 Jointly Learning Articulation Flow and Articulation Projection

Empirically, the Articulation Projection prediction is sometimes less accurate than the Articulation Flow prediction. On the other hand, the Articulation Flow represents the instantaneous motion and thus needs to predicted at each timestep, which prevents multi-step planning. To obtain the best of both worlds, we correct the Articulation Projection r_p prediction with the flow estimate using Gram-Schmidt on r_p to make it perpendicular to the Articulation Flow f_p; the correction is computed as: $\tilde{r}_p = r_p - \text{proj}_{f_p} r_p$, which is used in place of r_p in Eq. 3. The effect of using Gram-Schmidt is shown in Fig. 4a.

We learn to estimate the Articulation Flow f_p and the Articulation Projection r_p jointly using a deep neural network, which we refer to as FlowProjNet, denoted as f_θ. We assume that the robot has a depth camera and records point cloud observations $O_t \in \mathbb{R}^{3 \times N}$, where N is the total number of observable points from the sensor. The task is for the robot to articulate a specified part through its entire range of motion. For each configuration of the object, there exists a unique ground-truth Articulation Flow $\mathbf{f}_t$ and ground-truth Articulation Projection $\mathbf{r}_t$, given by Equations 1 and 2. Thus, our learning objective is to find a function $f_\theta(O_t)$ that predicts Articulation Flow and Articulation Projection directly from point cloud observations. We define the objective of minimizing the joint L2 error of the predictions:

$$\mathcal{L}_{\text{MSE}} = ||(\mathbf{f}_t \oplus \mathbf{r}_t) - f_\theta(O_t)||_2^2 \tag{7}$$

where $\oplus$ represents concatenation. We train FlowProjNet via supervised learning with this loss. We describe the method to obtain ground-truth labels in Appendix E. We implement FlowProjNet using a segmentation-style PointNet++ [30] as a backbone, and we train a *single* model across all categories, using a dataset of synthetically-generated data tuples based on the ground-truth kinematics provided by the PartNet-Mobility dataset [10]. Details of our dataset construction and model architecture can be found in Appendix E.

5 Experiments

We conduct a wide range of simulated and real-world experiments to evaluate FlowBot++.

	Novel Instances in Train Categories												Test Categories										
	AVG.												AVG.										
UMP-Net [12]	0.18	0.18	0.17	0.32	0.32	0.05	0.06	**0.12**	0.24	0.23	0.18	0.08	0.15	0.23	0.14	0.04	0.00	0.25	0.27	0.09	0.21	**0.13**	0.19
Normal Direction	0.41	0.52	0.67	0.16	0.19	0.51	0.60	0.13	0.14	0.45	0.71	0.31	0.40	0.79	0.43	0.07	0.00	0.64	0.15	0.70	1.00	0.41	0.29
Screw Parameters [1]	0.42	0.44	0.42	0.44	0.16	0.64	0.52	0.21	0.57	0.43	0.58	0.11	0.24	0.28	0.25	0.14	0.12	0.26	0.12	0.11	0.12	0.32	0.26
DAgger E2E [31]	0.52	0.25	0.86	0.51	0.71	0.55	1.00	0.86	0.80	0.74	0.58	0.40	0.64	0.53	0.52	1.00	0.65	0.75	0.54	0.63	0.68	0.87	0.74
FlowBot3D (AF Only) [6]	0.17	0.42	0.22	0.16	0.17	0.03	**0.00**	0.20	0.51	0.07	**0.00**	0.08	0.21	0.17	0.29	**0.00**	0.06	**0.21**	0.10	**0.06**	0.16	0.29	0.73
AP Only (Ours)	0.11	0.07	0.01	0.05	0.08	0.06	0.22	0.17	0.15	0.14	0.13	0.08	0.13	0.12	0.18	0.09	0.00	0.29	0.02	0.14	0.07	0.21	0.27
FlowBot++ (Ours - Combined)	**0.07**	**0.04**	**0.01**	**0.04**	**0.08**	**0.02**	0.19	0.17	**0.14**	**0.07**	0.09	**0.02**	**0.10**	**0.00**	**0.11**	0.09	**0.00**	0.23	**0.02**	0.09	**0.09**	0.20	**0.18**

Table 1: Normalized Distance Metric Results (↓): Normalized distances to the target articulation joint angle after a full rollout across different methods. The lower the better.

5.1 Simulation Results

To evaluate our method in simulation, we implement a suction gripper in PyBullet. We consider the same subset of PartNet-Mobility as in previous work [12, 6]. Each object starts in the "closed" state (one end of its range of motion), and the goal is to actuate the joint to its "open" state (the other end of its range of motion). During our experiments, we use the same Normalized Distance metric defined in prior work [6, 12].

Baseline Comparisons: We compare our proposed method with several baseline methods: UMP-Net [12], Normal Direction, Screw Parameters [1], DAgger End2End [31], FlowBot3D (AF Only) [6] which only uses **A**rticulation **F**low, and our model without the Gram-Schmidt correction ("AP Only"), which only uses the inferred **A**rticulation **P**arameters. Please refer to Appendix B for more details. Each method above consists of a single model trained across all PartNet-Mobility training categories.

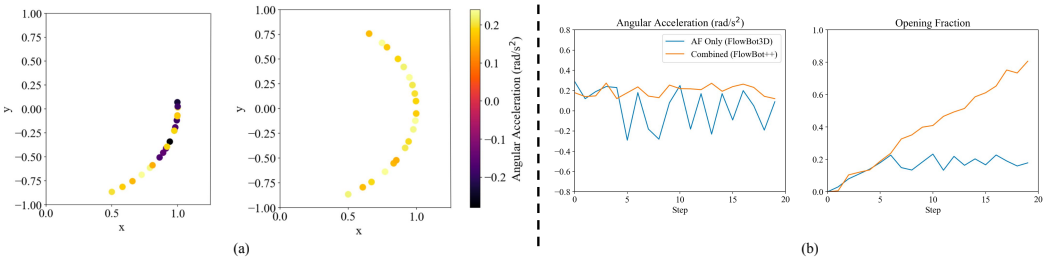


Figure 5: Comparison of Angular Acceleration and Opening Trajectory between FlowBot3D and FlowBot++. In **(a)**, we show two 20-step trajectories' scatter plots of the contact point on a revolute object using FlowBot3D (left) and FlowBot++ (right), colored using the signed intensity of the angular acceleration. In **(b)**, we plot 20 steps vs. the average angular acceleration and opening fraction across 300 trials involving 15 revolute objects. Both plots show that FlowBot++ is able to produce more consistent motions and open the objects further under the same number of steps.

Analysis: Results are shown in Table 1; please refer to Appendix D for comparisons using other metrics. Overall, FlowBot++ outperforms all of the baselines, including FlowBot3D [6] in terms of the normalized distance across many object categories, including unseen test categories. Moreover, without Gram-Schmidt correction (AP Only), the performance degrades but is still slightly better than FlowBot3D [6]. To illustrate why, we show an example in Fig. 5a in which we plot the xy locations of the contact point in the first 20 execution steps as well as the angular acceleration when using each method for articulation. As shown, both FlowBot3D and FlowBot++ perform reasonably well in the first 5 steps due to the absence of occlusions. However, after 5 steps, when the robot starts to heavily occlude the object, FlowBot3D's prediction of each step begins to make the contact point go back and forth, yielding little progress in the opening direction. In contrast, FlowBot++ plans a longer, multi-step trajectory at the start of the motion, interpolated via Eq. 5. This trajectory is much more consistent and smooth in terms of the direction of the motion. This trend is further shown in Fig. 5b, where we show the average angular acceleration and opening fraction across 15 revolute objects that were challenging for FlowBot3D [6]; FlowBot++ is able to achieve a larger

opening fraction and a smoother trajectory due to its lower replanning frequency, which also leads it to suffer less from occlusions since it can make a longer-scale prediction before contact.

Execution Time Comparisons: Here, we provide a comparison of execution time to open objects between FlowBot3D and FlowBot++. In simulation, under the same setup, execution wall-clock time gets reduced from 17.1 seconds per object on average to 1.2 seconds per object on average. The reason for the increased speed is twofold: First, FlowBot++ deploys an MPC-style controller, which only replans after H steps ($H = 7$ in our experiments), compared to FlowBot3D's closed-loop controller which requires replanning every step, since only the instantaneous motion is predicted. Second, with fewer replanning steps, FlowBot++ is less prone to be affected by occlusions because, in each replanning step, it rolls out a longer trajectory, as opposed to FlowBot3D, in which occlusions inevitably affect each step's prediction quality. The reduced prediction quality of FlowBot3D causes the robot to move the object back and forth, extending the execution time.

5.2 Real-World Experiments

We conduct qualitative real-world experiments to assess the sim2real capabilities of FlowBot++. We run FlowBot++ trained in simulation directly on denoised point clouds collected on real-world objects [32] without any retraining. We test our method on 6 real-world prismatic (Drawer) and revolute (Fridge, Oven, Toilet, Microwave, Trashcan) objects. In real-world trials, we provide hand-labeled segmentation masks to the network, which are used to filter and aggregate the results. We use a heuristically-defined grasping policy to select a grasp point and execute the planned trajectories on a Sawyer Black Robot equipped with a parallel-jaw gripper. Following the robot controller design in [28], once a contact point is chosen, we could transform the contact point using Eq. 5 because the gripper is now modeled as rigidly attached to the contact point, forming a full robot end-effector trajectory. Example predictions are shown in Fig. 6, in which all points on the segmented part are collectively transformed using Eq. 5, showing a full trajectory of the articulation. Real-world experiments show promising results that our networks transfer the learned 3D geometry information to real-world data. We show side-by-side comparisons between FlowBot++ and FlowBot3D on our website. FlowBot3D [6] completely failed on real-world oven, while FlowBot++ is able to predict a feasible full-trajectory to open the oven door with only one planning step. We also observe that FlowBot++ is able to produce smoother trajectories without undesired motions and movements of the objects and finish the execution in a shorter time frame, corroborating our findings in simulation. Details of the real-world experiments and system implementation are documented in Appendix F.

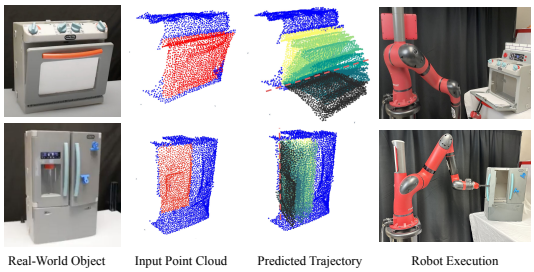


Figure 6: Real-World Experiments. We show two examples of FlowBot++ trained in simulation predicting and executing a full opening trajectory on real-world objects without any fine-tuning or retraining.

6 Conclusions and Limitations

In this work, we propose a novel visual representation for articulated objects, namely Articulation Projection, as well as a policy, FlowBot++, which leverages this representation, combined with Articulation Flow, to successfully manipulate articulated objects, outperforming previous state-of-the-art methods. We demonstrate the effectiveness of our method in both simulated and real environments and observe strong generalization to unseen objects.

Limitations: While our method shows strong performance on a range of object classes, there is still room for improvement. When both Articulation Flow and Articulation Projection predictions are incorrect, the predicted trajectory cannot be further corrected, causing failures. Furthermore, in order to aggregate all the per-point predictions on a part, a segmentation mask is required, adding another potential point of failure. Still, this work represents a step forward in the manipulation of unseen articulated objects, and we hope it provides a foundation for future work in this direction.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1849154. We are grateful to Prof. Daniel Seita for his helpful feedback and discussion on the paper.

References

[1] 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)*, pages 13670–13677, May 2021.

[2] N. Nie, S. Y. Gadre, K. Ehsani, and S. Song. Structure from action: Learning interactions for articulated object 3d structure discovery. *arXiv preprint arXiv:2207.08997*, 2022.

[3] D. Berenson, S. Srinivasa, and J. Kuffner. Task space regions: A framework for pose-constrained manipulation planning. *The International Journal of Robotics Research*, 30(12): 1435–1460, 2011.

[4] L. Yi, H. Huang, D. Liu, E. Kalogerakis, H. Su, and L. Guibas. Deep part induction from articulated object pairs. *arXiv preprint arXiv:1809.07417*, 2018.

[5] 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. *arXiv preprint arXiv:2006.14865*, 2020.

[6] B. Eisner, H. Zhang, and D. Held. Flowbot3d: Learning 3d articulation flow to manipulate articulated objects. *Robotics: Science and Systems (RSS)*, 2022.

[7] J.-S. Cheong, A. F. Van Der Stappen, K. Goldberg, M. H. Overmars, and E. Rimon. Immobilizing Hinged Polygons. *Int. J. Comput. Geom. Appl.*, 17(01):45–69, Feb. 2007.

[8] D. Katz, Y. Pyuro, and O. Brock. Learning to manipulate articulated objects in unstructured environments using a grounded relational representation. In *Robotics: Science and Systems IV*. Robotics: Science and Systems Foundation, June 2008.

[9] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 909–918, 2019.

[10] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang, and Others. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11097–11107, 2020.

[11] K. Mo, L. J. Guibas, M. Mukadam, A. Gupta, and S. Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6813–6823, 2021.

[12] Z. Xu, H. Zhanpeng, and S. Song. Umpnet: Universal manipulation policy network for articulated objects. *IEEE Robotics and Automation Letters*, 2022.

[13] T. Mu, Z. Ling, F. Xiang, D. Yang, X. Li, S. Tao, Z. Huang, Z. Jia, and H. Su. ManiSkill: Learning-from-Demonstrations benchmark for generalizable manipulation skills. *arXiv e-prints*, pages arXiv–2107, 2021.

[14] 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.

[15] R. Hu, W. Li, O. Van Kaick, A. Shamir, H. Zhang, and H. Huang. Learning to predict part mobility from a single static snapshot. *ACM Trans. Graph.*, 36(6):1–13, Nov. 2017.

[16] X. Li, H. Wang, L. Yi, L. J. Guibas, A. Lynn Abbott, and S. Song. Category-Level articulated object pose estimation, 2020.

[17] V. Zeng, T. E. Lee, J. Liang, and O. Kroemer. Visual identification of articulated object parts. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2443–2450. IEEE, 2020.

[18] V. Narayanan and M. Likhachev. Task-oriented planning for manipulating articulated mechanisms under model uncertainty. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3095–3101, May 2015.

[19] F. Burget, A. Hornung, and M. Bennewitz. Whole-body motion planning for manipulation of articulated objects. In *2013 IEEE International Conference on Robotics and Automation*, pages 1656–1662, May 2013.

[20] S. Chitta, B. Cohen, and M. Likhachev. Planning for autonomous door opening with a mobile manipulator. In *2010 IEEE International Conference on Robotics and Automation*, pages 1799–1806, May 2010.

[21] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artif. Intell.*, 17(1):185–203, Aug. 1981.

[22] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox. FlowNet: Learning optical flow with convolutional networks, 2015.

[23] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

[24] Z. Teed and J. Deng. RAFT: Recurrent All-Pairs field transforms for optical flow. In *Computer Vision – ECCV 2020*, pages 402–419. Springer International Publishing, 2020.

[25] S. Dong, D. K. Jha, D. Romeres, S. Kim, D. Nikovski, and A. Rodriguez. Tactile-rl for insertion: Generalization to objects of unknown geometry. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6437–6443. IEEE, 2021.

[26] A. Amiranashvili, A. Dosovitskiy, V. Koltun, and T. Brox. Motion perception in reinforcement learning with dynamic objects. In A. Billard, A. Dragan, J. Peters, and J. Morimoto, editors, *Proceedings of The 2nd Conference on Robot Learning*, volume 87 of *Proceedings of Machine Learning Research*, pages 156–168. PMLR, 2018.

[27] T. Weng, S. M. Bajracharya, Y. Wang, K. Agrawal, and D. Held. Fabricflownet: Bimanual cloth manipulation with a flow-based policy. In *Conference on Robot Learning*, pages 192–202. PMLR, 2022.

[28] C. Pan, B. Okorn, H. Zhang, B. Eisner, and D. Held. Tax-pose: Task-specific cross-pose estimation for robot manipulation. In *Conference on Robot Learning*, pages 1783–1792. PMLR, 2023.

[29] D. Seita, Y. Wang, S. J. Shetty, E. Y. Li, Z. Erickson, and D. Held. Toolflownet: Robotic manipulation with tools via predicting tool flow from point clouds. In *Conference on Robot Learning*, pages 1038–1049. PMLR, 2023.

[30] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[31] S. Ross, G. Gordon, and D. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.

[32] H. Zhang, J. Ichnowski, Y. Avigal, J. Gonzales, I. Stoica, and K. Goldberg. Dex-Net AR: Distributed deep grasp planning using a commodity cellphone and augmented reality app. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 552–558, May 2020.

[33] Q.-Y. Zhou, J. Park, and V. Koltun. Open3d: A modern library for 3d data processing. *arXiv preprint arXiv:1801.09847*, 2018.

[34] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. June 2017.

[35] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.