Discovering Conceptual Knowledge with Analytic Ontology Templates for Articulated Objects

Jianhua Sun*, Yuxuan Li*, Longfei Xu†, Jiude Wei†, Liang Chai, Cewu Lu‡

Shanghai Jiao Tong University
{gothic, yuxuan_li, xlf2023, wjd_kznwl, chailiang1234, lucewu}@sjtu.edu.cn

Abstract

Human cognition can leverage fundamental conceptual knowledge, like geometry and kinematic ones, to appropriately perceive, comprehend and interact with novel objects. Motivated by this finding, we aim to endow machine intelligence with an analogous capability through performing at the conceptual level, in order to understand and then interact with articulated objects, especially for those in novel categories, which is challenging due to the intricate geometric structures and diverse joint types of articulated objects. To achieve this goal, we propose Analytic Ontology Template (AOT), a parameterized and differentiable program description of generalized conceptual ontologies. A baseline approach called AOTNet driven by AOTs is designed accordingly to equip intelligent agents with these generalized concepts, and then empower the agents to effectively discover the conceptual knowledge on the structure and affordance of articulated objects. The AOT-driven approach yields benefits in three key perspectives: i) enabling concept-level understanding of articulated objects without relying on any real training data, ii) providing analytic structure information, and iii) introducing rich affordance information indicating proper ways of interaction. We conduct exhaustive experiments and the results demonstrate the superiority of our approach in understanding and then interacting with articulated objects.

Extended version — https://arxiv.org/pdf/2409.11702

1 Introduction

Articulated objects (Xiang et al. 2020; Liu et al. 2022a), composed of rigid segments interconnected by joints that enable translation and rotation movements, play an important role in daily life. Learning articulated objects brings essential significance in a wide range of research area, including computer vision, robotics, and embodied AI. Due to intricate geometric structures and diverse joint types in articulated objects (Xiang et al. 2020; Mo et al. 2021), it is challenging to interact with unseen articulated objects for machine intelligence, especially for those in novel categories.

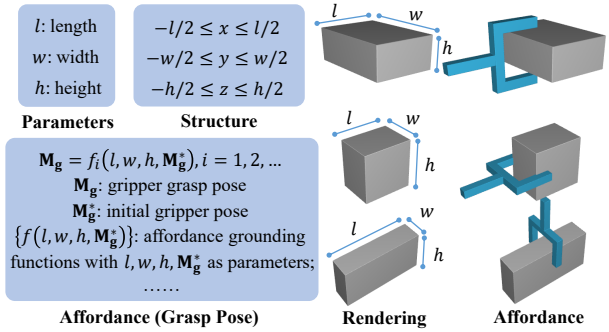

Figure 1: A brief schematic of AOT with *cuboid* as an example. Its structure delineates the cuboid shape, its parameters determine the size and aspect ratios, its affordances can include possible grasp poses and its renderer draws specific cuboid instances in 3D space. Here (x, y, z) refer to the coordinates in 3D space, $f_i(\cdot)$ refers to a function that transforms the initial gripper pose to a grasp pose according to AOT's parameters, the poses $\mathbf{M_g}$ and $\mathbf{M_g^*}$ are in the form of affine transformation matrices.

On the other hand, we humans also come into contact with a large amount and variety of objects in our daily lives. Particularly, even for objects in novel categories that an individual has never seen before, *i.e.* with no prior knowledge at either object or part level, studies from cognitive and brain science (Carey and Xu 2001; Scholl and Leslie 1999; Leslie et al. 1998; Piaget 1955; Ullman 2000; Biederman 1987; Hummel and Biederman 1992) show that human intelligence is still able to find the essence of their geometric structures and further rationally interact with them, relying on fundamental conceptual knowledge. This finding establishes a possible way for machine intelligence to interact with novel articulated objects from the conceptual perspective. However, little previous work has delved into this area in the data-driven era.

Motivated by this idea, we seek to equip intelligent agents with fundamental and generalized concepts, allowing them to discover conceptual knowledge on articulated objects, and finally enabling them to effectively understand and properly interact with such objects. Particularly, these concepts may include geometric ones like *cuboid* or *ring* and kinematic

*These authors contributed equally.
†These authors contributed equally.
‡Corresponding Author.

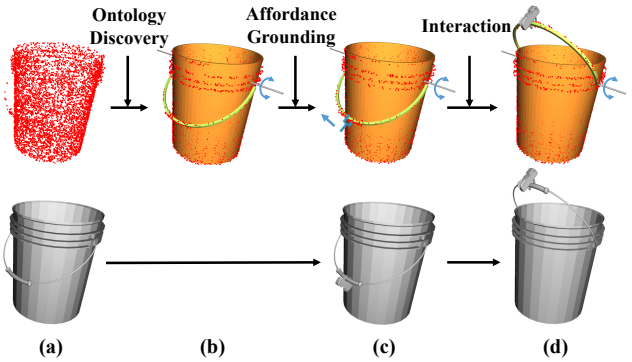

Figure 2: A brief illustration of the AOTNet workflow by articulated concept discovery (a-c). The top row refers to the processing pipeline in AOTNet, and the bottom row refers to the state of the real articulated object in corresponding steps.

ones like *revolute* or *prismatic*. Achieving this goal entails two key questions to answer: i) how to describe a conceptual ontology for machine intelligence and ii) how to leverage the conceptual descriptions to discover conceptual knowledge about the structure and affordance of articulated objects.

In this paper, we present Analytic Ontology Template (AOT) to answer the first question. AOT is a template description of a conceptual ontology in a program-like style, consisting of four main components:

- **Structure**: A set of differentiable mathematical expressions to describe the intrinsic basis of the ontology.
- **Parameter**: The value of variables in the structure description to identify specific instances of this template.
- **Affordance**: Some generalized knowledge of this ontology about where and how to interact with it.
- **Renderer**: A tool for rendering instances of this ontology template to data of certain formats (*e.g.* point clouds, meshes).

Fig. 1 demonstrates a brief schematic of AOT.

The analytical nature of AOT stems from the mathematical expressions that comprise it, exhibiting two significant features. i) Parameterized. By varying the parameters, countless instances of an ontology template can be created. In turn, a specific instance can be resolved into parameters of an ontology template. ii) Differentiable. This enables AOTs to be directly adopted in neural networks.

Regarding the second question, we investigate it in the context of a task where an agent is required to change a novel articulated object from its initial state to a target final state. The entire process, to which we refer as articulated concept discovery for simplicity, involves the discovery of both the geometric and kinematic structure concepts of the articulated part, as well as the identification of affordance concepts pertaining to where and how to properly interact with it. The success rate of an intelligent agent on this task can directly reflect the quality of the structure and affordance concepts discovery.

Coupled with the architecture and features of AOT, we propose a baseline approach called AOTNet powered by a set of AOTs to interact with articulated objects by articulated concept discovery. Fig. 2 gives a brief illustration of the AOTNet workflow. It first discovers the conceptual ontologies on the target, which is carried out by identifying the ontology template type and the corresponding parameters of AOT instances. This step discovers the conceptual essence of the target, *e.g.* the concept *ring* for the handle in Fig. 2. Neural networks involved in this process can be trained with synthetic data obtained by rendering AOT instances with varying parameters. The above process gives analytic estimations of the articulation structure in terms of conceptual ontology. Then, by analytically grounding the affordances of each ontology onto the target, an agent is able to appropriately interact with it guided by the knowledge.

In light of the above points, our AOT-driven approach benefits in understanding and interacting with articulated objects from the following perspectives. i) It generalizes well across novel articulated object categories and does not rely on real training data, due to the strong generalization capability of AOTs at the conceptual level. This holds significance since abundant high quality 3D articulated data (Xiang et al. 2020; Liu et al. 2022a) are expensive and labor-intensive to acquire and annotate, and it is difficult to comprehensively collect data to cover the diverse categories of articulated objects in the real world. ii) It can provide analytic articulation information via conceptual ontologies with mathematically defined structures and specific parameters. iii) Affordance knowledge can be precisely grounded which brings valuable guidance for accurate and controllable interaction. To show the superiority of our AOT-driven approach, we conduct exhaustive experiments on PartNet-Mobility dataset with SAPIEN environment (Xiang et al. 2020).

In summary, the main contributions of this work are:

- We propose Analytic Ontology Template (AOT) to describe conceptual ontologies for machine intelligence. It is capable of analytically describing both geometric and kinematic concepts along with affordances while remaining fully differentiable, allowing it to be incorporated into neural networks.
- We introduce an AOT-driven baseline, AOTNet, to equip intelligent agents with AOTs to discover conceptual knowledge about the structure and affordance on articulated objects. To the best of our knowledge, AOTNet is the first method to operate at the conceptual level, and demonstrates the effectiveness and benefits of interacting with novel articulated objects at this level.
- We evaluate our approach on hundreds of articulated objects across a wide range of novel categories. Through the remarkable results, we first demonstrate the possibility for machine intelligence to mimic this kind of human capability, *i.e.* leveraging generalized concepts to appropriately understand and then interact with a novel object.

2 Analytic Ontology Templates

In this section, we present Analytic Ontology Templates, *a.k.a.* AOT, to describe generalized geometric and kinematic conceptual ontologies for machine intelligence. We start by

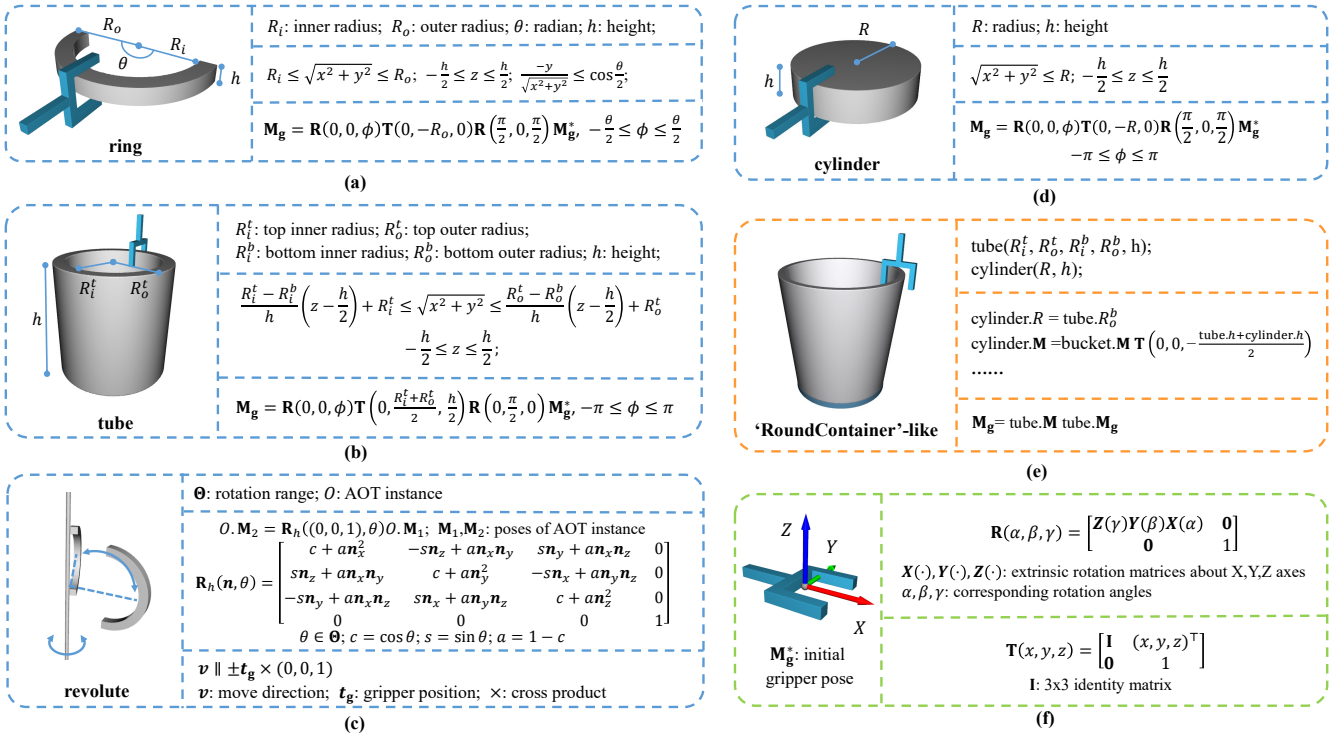

Figure 3: Specific examples of basic geometric (a,b,d), kinematic (c) and composite (e) ontologies, where (a-d) are defined from scratch and (e) is built upon existing ones. Each block in (a-e) is a collection of the AOT name and rendering (left), parameters (top-right), structure (mid-right) and affordances (bottom-right, partial). (f) includes definitions of viewing angle and initial gripper pose $\mathbf{M_g^*}$ (left), rotation matrix $\mathbf{R}$ and translation matrix $\mathbf{T}$ for poses (right). Note that i) the AOT name is just a referent (like a template id) of a concept and bears no semantic meanings; ii) the definitions of basic and composite AOTs are shown in the form of mathematical expressions and pseudo codes for easy understanding; iii) common parameters such as position and rotation in world coordinates are not shown for a clear view, please see Sec. 2.3 for details.

introducing the basic design philosophy in Sec. 2.1, which helps to better understand the motivation behind the specific design. Then, in Sec. 2.2 we describe the detailed architecture of AOT, as well as some particular examples. Finally, several discussions on AOT are presented in Sec. 2.3.

2.1 Design Philosophy

As the name suggests, the design philosophy of AOT covers two major points: i) the essential components of a program template and ii) the analytic nature.

Regarding the first point, it is desirable for a template to provide adequate valid information for a conceptual ontology. Its core components should consist of a detailed delineation of the ontology basis, along with associated parameters that represent diverse variations. Then, the template ought to be incorporated with adequate knowledge of affordances to facilitate interaction, *e.g.* grasp poses. Besides, a function is necessary to draw instances of the template in 3D space, *i.e.* render a template instance to data, for practical use such as network training, inference and visualization.

For the second point, the analytic nature implies that the template should be comprised of mathematical analytical expressions. An analytic description of a concept provides valuable knowledge for understanding its structure and facilitating interaction. Take the concept of *ring*, which frequently appears in the design of handles, as an example, by discovering the *ring* concept on a handle as an analytic description, we can identify its size and pose, as well as the detailed parameters for the orientation of its hinge. Further, the affordance knowledge like grasp poses of it can be grounded according to the parameters mathematically, which helps intelligent agents conduct more precise and controllable interactions. Moreover, by ensuring all mathematical expressions being differentiable, AOTs can be incorporated into neural networks for training and optimization.

2.2 Architecture

As a program template, we introduce class-like scripts to design AOTs, which facilitate easy extension as they enable a new AOT to be defined by inheritance through composition of existing ones in an object-oriented programming fashion. To satisfy the two points mentioned in Sec. 2.1, we design the template with four components. Several specific examples are demonstrated in Fig. 3.

Structure. The structure is the backbone of an ontology, referring to the essential commonality among different entities belonging to the same ontology. It is a deterministic description organized by a series of mathematical expressions.

We exclusively employ differentiable operations for all the mathematical expressions involved, enabling AOTs to participate in training and optimization. This component can be seen as the constructor of a program template.

Parameter. The parameters refer to values of variables in the structure description, causing the variances of different instances belonging to the same ontology. With specific parameters, a certain instance of this template can be created. In turn, a specific target can be resolved into parameters of an ontology template.

Affordance. The affordance indicates the generalized knowledge of this ontology regarding where and how to interact with it. Affordances of geometric ontologies may include grasp poses, contact points for pushing, *etc.* For kinematic ontologies, it usually includes the directions of force that cause movement. It is also described by mathematical expressions with certain parameters. These knowledge can provide valuable guidance to facilitate interaction.

Renderer. The renderer draws instances of this template in 3D space with certain data formats, like point clouds or meshes. These data can be used in many practical ways such as network training, inference and visualization.

2.3 Discussion

Transformation to World Coordinates. In Fig. 3, we omit the common parameters of 6-dof pose (including translation/position $\boldsymbol{t} \in \mathbb{R}^3$ and rotation $\boldsymbol{r} \in \mathbf{SO}(3)$) in the world coordinate system for simplicity. Generally, it is simple to represent the pose of AOT instances in world coordinates. By representing the 6-dof pose with an affine transformation matrix $\mathbf{M} \in \mathbb{R}^{4\times4}$, each formula F within the AOT structure and affordances can be reformulated from $F((x, y, z, 1)) \leq 0$ to $F(\mathbf{M}^{-1}(x, y, z, 1)) \leq 0$, thereby transforming an AOT instance to the world coordinate system from its own coordinate system. For rotationally symmetric geometries, we impose symmetry constraints on $\boldsymbol{r}$ to ensure uniqueness in the representation.

Efforts on Defining AOT. One of the benefits of our concept-level approach driven by AOT is that it only needs to create AOTs rather than collect real training data, and the efforts of creating AOTs are considerably low. In Tab. 1, we compare with building CAD model (Xiang et al. 2020) and real-world scan (Liu et al. 2022a) datasets on both time and money costs, and our approach shows advantages in both aspects in total. This can be attributed to i) AOTs are capable of adapting to different object categories since they are developed to describe generalized concepts, and therefore a small amount of AOTs can cover the conceptual knowledge on many everyday objects, ii) new AOTs can be easily expanded in an inheritance fashion by composing existing ones and iii) a single AOT can be used to effortlessly create an infinite number of diverse synthetic data by rendering instances with different parameters.

3 AOTNet

Taking into account the architecture and the features of AOT, we design AOTNet as a baseline to equip intelligent agents with these generalized concepts and empower them for articulated concept discovery. We first give an overview in Sec. 3.1. In Sec. 3.2, we demonstrate the detailed architecture of AOTNet. Finally, the implementation details are presented in Sec. 3.3.

	CAD	Scan	AOT
T (min/unit)	> 120	20	~ 60
C_L (\$/unit)	> 100	3	~ 10
C_D (\$)	-	~ 1000	-
C_O (\$)	-	✓	-
N (unit/category)	~ 30	~ 30	~ 4

Table 1: Budget comparison between building a CAD model dataset, a real-world scan dataset and analytic ontology templates. T, C_L, C_D, C_O refer to time, labor cost, device cost and object cost respectively. Besides the labor costs, building a real-world scan dataset incurs extra costs of purchasing a scanner and real world objects for scanning. N refers to the number of units on average to cover an object category in terms of articulated concept discovery.

3.1 Overview

Task setting. We study articulated concept discovery in the context of a task that given the initial and final status of a novel articulated object, an agent is asked to shift it from the initial state to the final state through interaction. This requires an intelligent system to possess the capability to comprehend the geometric and kinematic structure of an articulated part and also proper ways to interact with it, which are important topics in both the robotics and the computer vision communities. We develop AOTNet in a challenging setting that the status of the object only includes raw point clouds without any other information.

Pipeline. Our design aims to leverage the features of AOT and demonstrate the benefits of the AOT-driven approach. The workflow is twofold, including ontology discovery and subsequent affordance grounding (see Fig. 2).

As the task mainly takes into account the articulated part that undergoes state changes, we begin the workflow by discovering the geometric and kinematic ontologies on this part. The discovery process finds qualified AOT instances to capture the conceptual essence of the raw point clouds, involving two steps: ontology identification and parameter estimation. Leveraging the renderer, neural networks in these steps can be trained with synthetic data generated by rendering instances with various parameters.

After ontology discovery, the conceptual essence of the target is captured by AOT instances with analytic structures and parameter information. According to the analytic results, we can further accurately ground the affordance knowledge on the object through mathematical calculations. In this manner, an intelligent agent is able to perform appropriate interaction guided by the knowledge.

Discussion. Our AOT-driven approach has strong benefits in understanding and interacting with articulated objects from three perspectives. First, AOTNet works and generalizes well across novel categories without relying on real training data, benefiting from the generalization capability of AOTs

at the conceptual level. In comparison, it is necessary for a conventional object-level approach to first prepare abundant high quality articulated data for training, while acquiring, annotating, and comprehensively covering the various categories of articulated objects poses significant challenges. Second, our approach can provide analytic articulation information via AOT instances with mathematical structure and specific parameters. This also improves the interpretability of our approach. Third, affordance knowledge can be accurately grounded which brings valuable guidance for accurate and controllable interaction.

3.2 Architecture

We first introduce ontology identification and parameter estimation in the ontology discovery part, and then discuss about how grounded affordances facilitate the interaction. As the main purpose of AOTNet is to provide a baseline to demonstrate the feasibility and effectiveness of the AOT-driven paradigm, we do not delve into complex network designs. Experiments will show that promising performance can still be achieved even with a simple design.

Ontology Identification. Ontology identification is first introduced to identify which ontology template the input raw point clouds belong to with a classifier. The architecture of the classifier is a point cloud encoder followed by an MLP classifier. The input for geometric ontology is a single point cloud while that for kinematic ontology is a pair of point clouds at both initial and final states since at least two frames are needed to determine a dynamic process. Therefore, for the kinematic ontology, we first encode each point cloud into features separately, and then concatenate the features together to feed them into the MLP for classification.

Parameter Estimation. This step estimates the parameter values of an AOT to which the input belongs.

Geometric AOTs encompass basic ones and composite ones, where the latter are formed by combining the basic ones. Since the number of parameters of composite ones increases linearly with the number of basic ones composing them, it is difficult to directly regress all the parameters of a complex composite AOT. To this end, we perform parameter estimation of geometric AOTs at the basic ontology level. Particularly, an estimator for a geometric AOT consists of four components: i) a point cloud encoder E to encode the input point cloud $\mathcal{P}$ into deep features, ii) a latent code $\mathcal{R}_w$ for the whole AOT represented by learnable parameters, iii) a set of latent codes $\{\mathcal{R}_i|i = 1, 2, ...\}$ for each basic ontology composing the AOT represented by learnable parameters, and iv) a set of MLPs $\{\mathrm{MLP}_i\,|i = 1, 2, ...\}$ to regress the parameters $\{P_i|i = 1, 2, ...\}$ of each basic ontology by $P_i = \mathrm{MLP}_i([E(\mathcal{P}), \mathcal{R}_w, \mathcal{R}_i])$, where $[\cdot]$ denotes concatenation. The set of parameters $\{P_i|i = 1, 2, ...\}$ is merged together to produce the final estimation. Note that $i = 1$ if the whole AOT is a basic one. Some parameters of basic ontology are discarded to meet the constraints in the whole AOT. To train the estimator, apart from the MSE loss on each parameter, we also introduce a Point2Mesh loss (PyTorch3D 2023b) between the mesh of the AOT instance and the input point cloud. The utilization of the Point2Mesh loss greatly benefits from the differentiable feature of AOT, meaning that the process of rendering the mesh of an AOT instance with certain parameters is differentiable. As a result, this loss can be backpropagated to the estimated parameters, enabling the optimization of the entire estimator.

To estimate the parameters of kinematic AOTs, we use the same network architecture as the classifier and modify its output to the kinematic parameters. Particularly, the output for prismatic is its direction, and that for revolute is the axis direction and the position of the pivot point. We use cosine distance as the axis alignment loss and L2 distance to the ground truth axis as the pivot loss.

Optionally, when full-surrounding point clouds are available through multi-view observations or point cloud completion, we can optimize the estimated parameters of AOT with the aforementioned Point2Mesh loss for better geometry ontology discovery. Furthermore, the mesh rendered by AOT can be deformed to better fit the object point cloud and obtain more refined geometric details according to algorithms like (Hanocka et al. 2020; PyTorch3D 2023a). Nevertheless, our experiments have shown that AOTNet can achieve good results without performing such operations.

Affordance Grounding and Interaction. As mentioned above, we have discovered the descriptions of the conceptual essence of the input data, including both geometric and kinematic concepts as well as the parameters. At this point, corresponding affordances in the AOTs can be grounded mathematically. According to this information, interactions can be conducted on the object to effectively change the state of the articulation as required. In our implementation, we just use a simple interaction strategy that first grasps the articulated part according to the geometric affordances and then applies force along the direction according to the kinematic affordances. The grasp pose is sampled from all possible grasp affordances. We also check for collisions before adopting the pose, and resample if necessary.

3.3 Implementation Details

Training Data Preparation. All the neural networks involved in AOTNet are trained with synthetic point clouds of AOT instances with different parameters, which are generated by the renderer. Particularly, the renderer samples points in 3D space according to the mathematical expressions of the structure to acquire the point cloud of a geometric ontology. For the kinematic ontology, the renderer produces a pair of point clouds, which represent the states of a geometric ontology before and after the movement. Some additional noise and corruptions are introduced to the point clouds for data augmentation.

Networks. The point cloud encoders used in both ontology identification and parameter estimation are transformer encoders (Yu et al. 2022), which extract 128 groups of points with size 32 from the input with 2048 points, and send them into a standard transformer encoder with 12 6-headed attention layers. All MLPs used in our experiments are triple-layered with ReLU activation.

4 Experiments

We evaluate the full AOTNet with a closed-loop validation through the success rate of interaction on the proposed task,

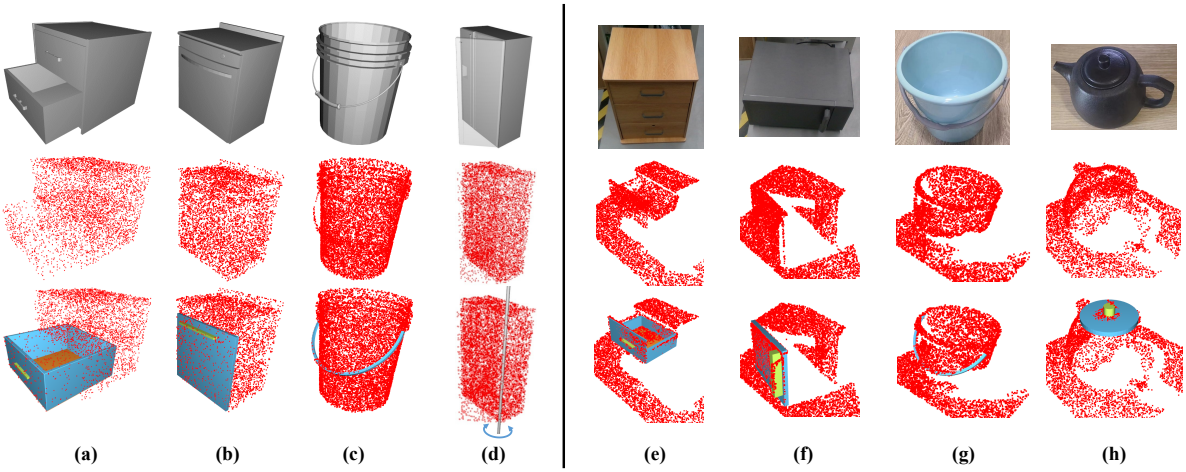

Figure 4: Visualization of ontology discovery results for real-world objects in a simulation environment (a-d) and the physical world (e-h). The first row shows the target objects. The second row gives the input point clouds of the objects. The third row shows the discovery results of the actionable part with AOTNet.

i.e. changing an articulated object from its initial state to a target final state. The success rate can reveal the quality of articulated concept discovery, including ontology discovery and affordance grounding.

Experiment Settings. We conduct our experiments in SAPIEN (Xiang et al. 2020) environment and follow (Mo et al. 2021) for the environment setting. We use a flying Franka Panda gripper as the agent. To generate the input partial point cloud scans, we mount an RGB-D camera with known intrinsic parameters 5-unit-length away pointing to the center of the target object.

To satisfy the compatibility with a single-gripper, we evaluate on a total of 17 specific interaction tasks involving 15 representative object categories chosen from PartNet-Mobility Dataset, including turning faucets, opening table doors, flipping bucket handles, *etc.* In total, 653 objects with 1152 articulation structures are used as test samples. For each specific task, we set the initial state of the target joint as closed (0% of the joint range), and the final state as fully open (100% of the joint range). In practice, we consider a task successful if more than 80% of the total moving range is achieved.

Baselines. To compare with the interaction strategy of AOTNet, we adopt two popular reinforcement learning approaches, Soft Actor-Critic (SAC) (Haarnoja et al. 2018) and Truncated Quantile Critics (TQC) (Kuznetsov et al. 2020), as baselines. For both RL approaches, we provide three different state representations:

- Raw: raw input point cloud of the target.
- Raw + JP: information in Raw, and both ground-truth joint specifications and 6-dof poses of the articulation structure.
- AOT: AOT instances (template type and corresponding parameters) discovered on the target by AOTNet, including both geometric and kinematic ones.

Note that the pose, velocity and gripper contact of the agent are the common states shared among the three settings. The training samples for SAC/TQC are selected from PartNet-Mobility assets except those in test samples. We also provide the *Human* evaluation result as a reference, in which we first invite ten college students to discover the AOTs on the test cases according to their cognition and then use the interaction strategy in AOTNet with these manually discovered AOTs for interaction.

Main Results. Tab. 2 gives the average success rate of our approach and other baseline methods. The success rate of human evaluation is very close to 100%, demonstrating that AOTs can adequately reflect the conceptual knowledge of structure and affordance on articulated objects when they are discovered based on human cognition. For AOTNet, it achieves a remarkable result of 63.7% success rate which greatly outperforms RL-based methods, indicating that the ontology discovery and affordance grounding in AOTNet enjoy good accuracy. Given that AOTNet does not involve real object-level training data and instead only uses algorithm-generated synthetic concept-level data for training, all these test articulated objects are novel for AOTNet, this also demonstrates the superiority of learning at the conceptual level. Fig. 4 shows visualizations of ontology discovery.

Contribution of Affordance Information. The full AOTNet performs interaction according to grounded affordances, while the RL-based approaches explore the interaction strategies according to the input state representations from actual interactions. According to Tab. 2, AOTNet performs better than all the RL settings, especially Raw+JP which introduces joint and pose ground truths. This ablation experiment demonstrates that the affordance knowledge is well grounded, and provides valuable guidance for accurate interaction. In comparison, it is relatively more difficult to explore proper affordances and corresponding interaction strategies with RL from the input representations.

Contribution of Analytic Articulation Information. The

Method	Human	AOTNet	SAC			TQC		
			Raw	Raw + JP	AOT	RAW	Raw + JP	AOT
Avg. Succ. Rate	97.5	63.7	19.6	41.5	34.3	22.6	46.4	39.0

Table 2: Success rate of articulated object interaction.

RL results in Tab. 2 demonstrate that the performance of the three settings shows a similar trend for both SAC and TQC. The Raw setting does not perform well compared with the others considering there is no joint and 6-dof pose information. With ground truths as additional state representations, Raw+JP achieves the best performance among the three. The result of AOT setting is much higher than Raw, indicating one of the benefits of an AOT-driven approach that an analytic description can be given for the raw input, which enables RL agents to better leverage the state representations to learn interaction strategies.

5 Related Work

Conceptual Object Understanding in Human Cognition. Researchers in cognitive science have been studying visual object understanding for several decades (Humphreys, Price, and Riddoch 1999; Ullman 2000; Palmeri and Gauthier 2004; Biederman 1987; Hummel and Biederman 1992). Their investigations have aimed to unravel the intricate processes involved in perceiving, recognizing, and comprehending objects within the human mind, wherein they have discovered the significant role of conceptual knowledge (Biederman 1987; Habel and Eschenbach 2006; Palmeri and Gauthier 2004; Rosch 1975). For example, Biederman (Biederman 1987) has found that the perceptual recognition of objects can be conceptualized to be a process in which an object is segmented into an arrangement of simple geometric components, such as blocks, cylinders, wedges, and cones. Compelling evidence provided in (Palmeri and Gauthier 2004; Dixon et al. 2002; Goldstone and Barsalou 1998) also indicates a strong relation between perception and conceptual knowledge. Studies on infants (Carey and Xu 2001; Scholl and Leslie 1999) further reveals the pivotal significance of conceptual knowledge in object understanding, since infants are much less susceptible to empirical factors. These findings establish a possible way for machine intelligence to understand objects at the conceptual level, but little previous work in the computer vision community has paid attention to this area.

Articulated Object Understanding. Articulation structure is an important topic in both vision and robotics community and has been investigated by many researchers. Many articulated object assets (Xiang et al. 2020; Liu et al. 2022a; Martín-Martín, Eppner, and Brock 2019; Calli et al. 2015; Sun et al. 2024b,a) have been proposed in recent years. Thanks to previous work ShapeNet (Chang et al. 2015) that collects a large amount of CAD object models and PartNet (Mo et al. 2019) that gives hierarchical part semantic segmentation annotations on a subset of ShapeNet, and Shape2Motion (Wang et al. 2019) further label the joint information of CAD models in PartNet for articulation research. PartNet-Mobility (Xiang et al. 2020) takes a step further to collect a large-scale set of articulated objects from 3D Warehouse which include models of real-world manufacturers' products and possess real-world diversity and complexity. On the other hand, some researches (Liu et al. 2022a) aim to collect assets via real-world scanning to build articulated object datasets.

Based on these datasets, articulated objects have been studied from many aspects. From the perception aspect, current works concentrate on multiple areas such as recognition (Zeng et al. 2021; Jain et al. 2021), part segmentation (Yi et al. 2018), pose estimation (Li et al. 2020; Liu et al. 2022b) and tracking (Heppert et al. 2022; Weng et al. 2021). For example, Yi et al. (Yi et al. 2018) develop a network to co-segment the input objects into their articulated parts according to a 3D CAD model similar to the input. Li et al. (Li et al. 2020) propose a normalized coordinate space to estimate 6D pose and joint state for articulated objects. There are also many researches on discovering how to interact with articulated objects (Katz and Brock 2008; Xiang et al. 2020; Mo et al. 2021; Liu et al. 2022a; Wang et al. 2022; Geng et al. 2023; Ling et al. 2024). For example, Mo et al. (Mo et al. 2021; Wang et al. 2022) extract highly localized actionable information for articulated objects with movable parts. Still, most of current approaches work on object or part level and heavily rely on high-quality 3D data for training. In comparison, our AOT-driven paradigm demonstrates promising performance at conceptual level and needs no real training data for its generalization capability.

6 Conclusion

In this paper, we introduce Analytic Ontology Templates as the description for generalized concepts, in order to enable machine intelligence to appropriately perceive, comprehend and interact with articulated objects at the conceptual level, especially for those in novel categories. Our main contributions are as follows. First, we propose Analytic Ontology Template as a parameterized and differentiable template description of a generalized conceptual ontology, and affordances can be numerically described in the template for accurate and controllable interactions. Second, an AOT-driven pipeline called AOTNet is designed accordingly to equip intelligent agents with these concepts, and then empower the agents to effectively discover the conceptual knowledge of structure and affordance on objects. Third, we comprehensively evaluate the effectiveness of AOT and AOTNet on hundreds of articulated objects across a wide range of categories in terms of articulated concept discovery. The experiments suggest the benefits of the AOT-driven paradigm and the possibility for machine intelligence to mimic the human capability of learning at the conceptual level.

References

Biederman, I. 1987. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2): 115.

Calli, B.; Singh, A.; Walsman, A.; Srinivasa, S.; Abbeel, P.; and Dollar, A. M. 2015. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 international conference on advanced robotics (ICAR)*, 510–517. IEEE.

Carey, S.; and Xu, F. 2001. Infants' knowledge of objects: Beyond object files and object tracking. *Cognition*, 80(1-2): 179–213.

Chang, A. X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; Xiao, J.; Yi, L.; and Yu, F. 2015. ShapeNet: An Information-Rich 3D Model Repository. arXiv:1512.03012.

Dixon, M. J.; Desmarais, G.; Gojmerac, C.; Schweizer, T. A.; and Bub, D. N. 2002. The role of premorbid expertise on object identification in a patient with category-specific visual agnosia. *Cognitive Neuropsychology*, 19(5): 401–419.

Geng, H.; Xu, H.; Zhao, C.; Xu, C.; Yi, L.; Huang, S.; and Wang, H. 2023. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7081–7091.

Goldstone, R. L.; and Barsalou, L. W. 1998. Reuniting perception and conception. *Cognition*, 65(2-3): 231–262.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, 1861–1870. PMLR.

Habel, C.; and Eschenbach, C. 2006. Abstract structures in spatial cognition. *Foundations of Computer Science: Potential—Theory—Cognition*, 369–378.

Hanocka, R.; Metzer, G.; Giryes, R.; and Cohen-Or, D. 2020. Point2Mesh: a self-prior for deformable meshes. *ACM Transactions on Graphics*, 39(4).

Heppert, N.; Migimatsu, T.; Yi, B.; Chen, C.; and Bohg, J. 2022. Category-Independent Articulated Object Tracking with Factor Graphs. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3800–3807. IEEE.

Hummel, J. E.; and Biederman, I. 1992. Dynamic binding in a neural network for shape recognition. *Psychological review*, 99(3): 480.

Humphreys, G. W.; Price, C. J.; and Riddoch, M. J. 1999. From objects to names: A cognitive neuroscience approach. *Psychological research*, 62: 118–130.

Jain, A.; Lioutikov, R.; Chuck, C.; and Niekum, S. 2021. Screwnet: Category-independent articulation model estimation from depth images using screw theory. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 13670–13677. IEEE.

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

Kuznetsov, A.; Shvechikov, P.; Grishin, A.; and Vetrov, D. 2020. Controlling overestimation bias with truncated mixture of continuous distributional quantile critics. In *International Conference on Machine Learning*, 5556–5566. PMLR.

Leslie, A. M.; Xu, F.; Tremoulet, P. D.; and Scholl, B. J. 1998. Indexing and the object concept: developing-what'andwhere'systems. *Trends in cognitive sciences*, 2(1): 10–18.

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

Ling, S.; Wang, Y.; Wu, R.; Wu, S.; Zhuang, Y.; Xu, T.; Li, Y.; Liu, C.; and Dong, H. 2024. Articulated object manipulation with coarse-to-fine affordance for mitigating the effect of point cloud noise. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 10895–10901. IEEE.

Liu, L.; Xu, W.; Fu, H.; Qian, S.; Yu, Q.; Han, Y.; and Lu, C. 2022a. AKB-48: a real-world articulated object knowledge base. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14809–14818.

Liu, L.; Xue, H.; Xu, W.; Fu, H.; and Lu, C. 2022b. Toward real-world category-level articulation pose estimation. *IEEE Transactions on Image Processing*, 31: 1072–1083.

Martín-Martín, R.; Eppner, C.; and Brock, O. 2019. The RBO dataset of articulated objects and interactions. *The International Journal of Robotics Research*, 38(9): 1013–1019.

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

Mo, K.; Zhu, S.; Chang, A. X.; Yi, L.; Tripathi, S.; Guibas, L. J.; and Su, H. 2019. 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*, 909–918.

Palmeri, T. J.; and Gauthier, I. 2004. Visual object understanding. *Nature Reviews Neuroscience*, 5(4): 291–303.

Piaget, J. 1955. *The child's construction of reality*. Routledge.

PyTorch3D. 2023a. Deform Source Mesh to Target Mesh. https://github.com/facebookresearch/pytorch3d/blob/main/docs/tutorials/deform_source_mesh_to_target_mesh.ipynb. Accessed: 2024-12-14.

PyTorch3D. 2023b. Distance Between a Pointcloud and a Mesh. https://pytorch3d.readthedocs.io/en/latest/modules/loss.html#pytorch3d.loss.point_mesh_face_distance. Accessed: 2024-12-14.

Rosch, E. 1975. Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3): 192.

Scholl, B.; and Leslie, A. 1999. Explaining the infant's object concept. *What is cognitive science*, 26–73.

Sun, J.; Li, Y.; Wei, J.; Xu, L.; Wang, N.; Zhang, Y.; and Lu, C. 2024a. Arti-PG: A Toolbox for Procedurally Synthesizing Large-Scale and Diverse Articulated Objects with Rich Annotations. arXiv:2412.14974.

Sun, J.; Li, Y.; Xu, L.; Wang, N.; Wei, J.; Zhang, Y.; and Lu, C. 2024b. ConceptFactory: Facilitate 3D Object Knowledge Annotation with Object Conceptualization. arXiv:2411.00448.

Ullman, S. 2000. *High-level vision: Object recognition and visual cognition*. MIT press.

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

Wang, Y.; Wu, R.; Mo, K.; Ke, J.; Fan, Q.; Guibas, L. J.; and Dong, H. 2022. Adaafford: Learning to adapt manipulation affordance for 3d articulated objects via few-shot interactions. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, 90–107. Springer.

Weng, Y.; Wang, H.; Zhou, Q.; Qin, Y.; Duan, Y.; Fan, Q.; Chen, B.; Su, H.; and Guibas, L. J. 2021. Captra: Category-level pose tracking for rigid and articulated objects from point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13209–13218.

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

Yi, L.; Huang, H.; Liu, D.; Kalogerakis, E.; Su, H.; and Guibas, L. 2018. Deep part induction from articulated object pairs. *ACM Transactions on Graphics*, 37(6): 1–15.

Yu, X.; Tang, L.; Rao, Y.; Huang, T.; Zhou, J.; and Lu, J. 2022. Point-BERT: Pre-Training 3D Point Cloud Transformers with Masked Point Modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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