Understanding Tools:
Task-Oriented Object Modeling, Learning and Recognition

Yixin Zhu * yixin.zhu@ucla.edu
Yibiao Zhao * ybzhao@ucla.edu
Song-Chun Zhu sczhu@stat.ucla.edu

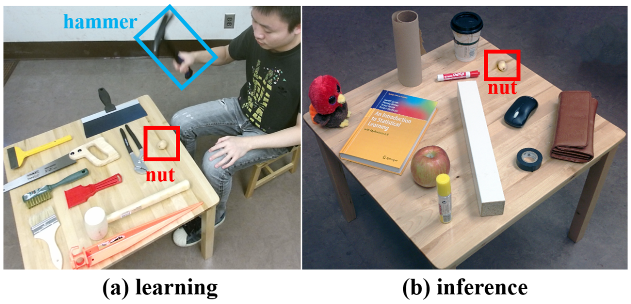
Center for Vision, Cognition, Learning, and Art
University of California, Los Angeles, CA 90095, USA

Abstract

In this paper, we present a new framework – task-oriented modeling, learning and recognition which aims at understanding the underlying functions, physics and causality in using objects as "tools". Given a task, such as, cracking a nut or painting a wall, we represent each object, e.g. a hammer or brush, in a generative spatio-temporal representation consisting of four components: i) an affordance basis to be grasped by hand; ii) a functional basis to act on a target object (the nut), iii) the imagined actions with typical motion trajectories; and iv) the underlying physical concepts, e.g. force, pressure, etc. In a learning phase, our algorithm observes only one RGB-D video, in which a rational human picks up one object (i.e. tool) among a number of candidates to accomplish the task. From this example, our algorithm learns the essential physical concepts in the task (e.g. forces in cracking nuts). In an inference phase, our algorithm is given a new set of objects (daily objects or stones), and picks the best choice available together with the inferred affordance basis, functional basis, imagined human actions (sequence of poses), and the expected physical quantity that it will produce. From this new perspective, any objects can be viewed as a hammer or a shovel, and object recognition is not merely memorizing typical appearance examples for each category but reasoning the physical mechanisms in various tasks to achieve generalization.

1. Introduction

In this paper, we rethink object recognition from the perspective of an agent: how objects are used as "tools" in actions to accomplish a "task". Here a task is defined as changing the physical states of a target object by actions, such as, cracking a nut or painting a wall. A tool is a physical object used in the human action to achieve the task, such as a hammer or brush, and it can be any daily objects and is not restricted to conventional hardware tools. This leads us to a new framework – task-oriented modeling, learning and recognition, which aims at understanding the underlying functions, physics and causality in using objects as tools in various task categories.



(a) learning **(b) inference**

Figure 1. Task-oriented object recognition. (a) In a learning phase, a rational human is observed picking a hammer among other tools to crack a nut. (b) In an inference phase, the algorithm is asked to pick the best object (i.e. the wooden leg) on the table for the same task. This generalization entails physical reasoning.

Fig. 1 illustrates the two phases of this new framework. In a learning phase, our algorithm observes only one RGB-D video as an example, in which a rational human picks up one object, the hammer, among a number of candidates to accomplish the task. From this example, our algorithm reasons about the essential physical concepts in the task (e.g. forces produced at the far end of the hammer), and thus learns the task-oriented model. In an inference phase, our algorithm is given a new set of daily objects (on the desk in (b)), and makes the best choice available (the wooden leg) to accomplish the task.

From this new perspective, any objects can be viewed as a hammer or a shovel, and this generative representation allows computer vision algorithms to generalize object recognition to novel functions and situations by reasoning

[*] Yixin Zhu and Yibiao Zhao contribute equally to this work.

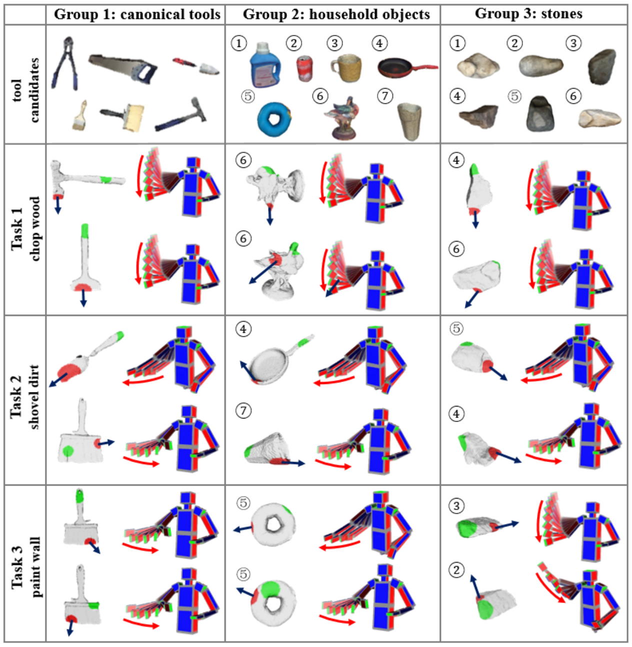


Figure 2. Given three tasks: chop wood, shovel dirt, and paint wall. Our algorithm picks and ranks objects for each task among objects in three groups: 1) conventional tools, 2) household objects, and 3) stones, and output the imagined tool-use: affordance basis (the green spot to grasp with hand), functional basis (the red area applied to the target object), and the imagined action pose sequence.

the physical mechanisms in various tasks, and go beyond memorizing typical examples for each object category as the prevailing appearance-based recognition methods do in the literature.

Fig. 2 shows some typical results in our experiments to illustrate this new task-oriented object recognition framework. Given three tasks: chop wood, shovel dirt, and paint wall, and three groups of objects: conventional tools, household objects, and stones, our algorithm ranks the objects in each group for a task. Fig. 2 shows the top two choices together with imagined actions using such objects for the tasks.

Our task-oriented object representation is a generative model consisting of four components in a hierarchical spatial-temporal parse graph:

i) An *affordance basis* to be grasped by hand;

ii) A *functional basis* to act on the target object;

iii) An *imagined action* with pose sequence and velocity;

iv) The *physical concepts* produced, e.g. force, pressure.

In the learning phase, our algorithm parses the input RGB-D video by simultaneously reconstructing the 3D meshes of tools and tracking human actions. We assume that the human makes rational decisions in demonstration: picks the best object, grasps the right place, takes the right action (poses, trajectory and velocity), and lands on the target object with the right spots. These decisions are nearly optimal against a large number of compositional alternative choices. Using a ranking-SVM approach, our algorithm will discover the best underlying physical concepts in the human demonstration, and thus the essence of the task.

In the inference stage, our algorithm segments the input RGB-D image into objects as a set of candidates, and computes the task-oriented representation – the optimal parse graph for each candidate and each task by evaluating different combinations. This parse graph includes the best object and its tool-use: affordance basis (green spot), functional basis (red spot), actions (pose sequence), and the quantity of the physical concepts produced by the action.

This paper has four major contributions:

1. We propose a novel problem of task-oriented object recognition, which is more general than defining object categories by typical examples, and is of great importance for object manipulation in robotics applications.
2. We propose a task-oriented representation which includes both the visible object and the imagined use (action and physics). The latter are the 'dark matter' [48] in computer vision.
3. Given an input object, our method can imagine the plausible tool-use and thus allows vision algorithms to reason innovative use of daily object – a crucial aspect of human and machine intelligence.
4. Our algorithm can learn the physical concepts from a single RGB-D video and reason about the essence of physics for a task.

2. Task-oriented Object Representation

Tools and tool-uses are traditionally studied in cognitive science [29, 4, 35, 2] with verbal definitions and case studies, and an explicit formal representation is missing in the literature.

In our task-oriented modeling and learning framework, an object used for a task is represented in a joint spatial, temporal, and causal parse graph $pg = (pg_s, pg_t, pg_c)$ including three aspects shown in Fig. 3:

i) A *spatial parse graph* pg_s represents object decomposition and 3D relations with the imagined pose;

ii) A *temporal parse graph* pg_t represents the pose sequence in actions; and

iii) A *causal parse graph* pg_c represents the physical quantities produced by the action on the target object.

In this representation, only the object is visible as input, all other components are imagined.

2.1. Tool in 3D space

An object (or tool) is observed in a RGB-D image in the inference stage, which is then segmented from the background and filled-in to become a 3D solid object denoted

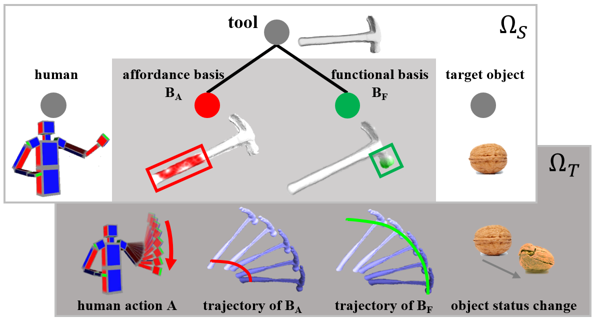


Figure 3. The task-oriented representation of a hammer and its use in a task (crack a nut) in a joint spatial, temporal, and causal space.

by $\mathbf{X}$. The 3D object is then decomposed into two key parts in the spatial parse graph pg_s:

1) Affordance basis $\mathbf{B_A}$, where the imagined human hand grasps the object with certain pose. Through offline training, we have collected a small set of hand poses for grasping. The parse graph pg_s encodes the 3D positions and 3D orientations between the hand poses and the affordance basis during the tool-use, using 3D geometric relations between the hand pose and the affordance basis, as it is done in [45]. The parse graph pg_s will have lower energy or high probability when the hand hold the object comfortably (see the trajectory of affordance basis $\mathbf{B_A}$ in Fig.3).

2) Functional basis $\mathbf{B_F}$, where the object (or tool) is applied to a target object (the nut) to change its physical state (i.e. fluent). The spatial parse graph pg_s also encodes the 3D relations between the functional basis $\mathbf{B_F}$ and the 3D shape of the target object during the action. We consider three types of the functional basis: (a) a single contact spot (e.g. hammer); (b) a sharp contacting line segment or edge (e.g. axe and saw); and (c) flat contacting area (e.g. shovel).

We define a space $\Omega_S = \{pg_s\}$ as the set of all possible spatial parse graph pg_s which is a product space of all the possible objects, their affordance bases, functional bases, hand poses, and 3D relations above.

2.2. Tool-use in time

An tool-use is a specific action sequence that engages the tool in a task, and is represented by a temporal parse graph pg_t. pg_t represents the human action $\mathbf{A}$ as a sequence of 3D poses. In this paper, since we only consider hand-hold objects, we collect some typical action sequences for the arm and hand movements using tools by RGB-D sensors, such as, hammering, shoveling, etc. These actions are then clustered into average pose sequences. For each of the sequence, we record the trajectories of the hand pose (or affordnace basis) and the functional basis.

We define a space $\Omega_T = \{pg_t\}$ as the set of possible pose sequences and their associated trajectories of the affordance basis B_A and functional basis B_F.

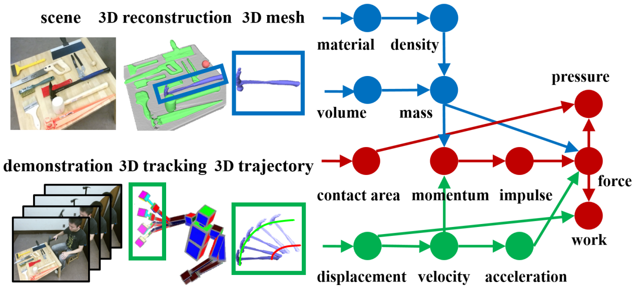


Figure 4. Thirteen physical concepts involved in tool-use and their compositional relations. By parsing human demonstration, the physical concepts of material, volume, concept area, and displacement are estimated from 3D meshes of tool (blue), trajectories of tool-use (green) or jointly (red). The higher-level physical concepts can be further derived recursively.

2.3. Physical concept and causality

We consider of thirteen basic physical concepts involved in tool-use, which can be extracted or derived from the spatial and temporal parse graphs as Fig. 4 illustrates.

Firstly, as the blue dots and lines in Fig. 4 illustrates, we reconstruct the 3D mesh from the input 3D object and thus calculate its volume, and by estimating its material category, we get its density. From volume and density we further calculate the mass of the objects and its parts (when different materials are used).

Secondly, as the green dots and lines Fig. 4 illustrates, we can derive the displacement from the 3D trajectory of affordance basis and functional basis, and then calculate the velocity and acceleration of the two bases.

Thirdly, as red dots and line shows, we can estimate the contact spot, line and area from the functional basis and target object, and further compute the momentum, and impulse. We can then also compute basic physical concepts, such as forces, pressure, and work etc.

Physical concept operators ∇. We define a set of operators, including addition $\nabla_+(\cdot,\cdot)$, subtraction $\nabla_-(\cdot,\cdot)$, multiplication $\nabla_\times(\cdot,\cdot)$, division $\nabla_/(\cdot,\cdot)$, negation $\nabla_{\text{neg}}(\cdot)$, space integration $\nabla_{\int_S}(\cdot)$, time integration $\nabla_{\int_T}(\cdot)$, space derivation $\nabla_{\partial_S}(\cdot)$ and time derivation $\nabla_{\partial_T}(\cdot)$. For example, the concept of the force and acceleration are defined as: force $= \nabla_\times(\text{mass}, \text{acceleration})$, acceleration $= \nabla_{\partial_t}(\text{velocity})$

The causal parse graph pg_c includes the specific physical concepts used in a tool-use which is often an instantiated sub-graph of the concept graph in Fig. 4.

Since the law of physics is universally applicable, the major advantage of using physical concepts is the ability to generalize to novel situations.

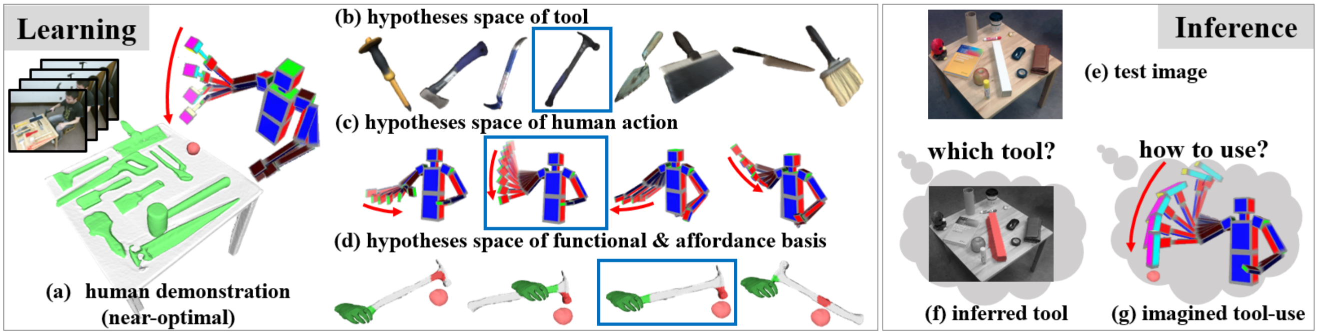


Figure 5. An illustration of learning and inference. (a)-(d) We assume the human choice (shown in blue bounding box) of tool and tool-use (action and affordance / functional bases) is near-optimal, thus most of other combinations of tool and tool-use (action, affordance / functional bases) in the hypotheses spaces should not outperform human demonstration. Based on this assumption, we treat the human demonstration as positive example, and random sample other tools and tool-uses in the hypothesis spaces as negative examples. (e) During the inference, given an image of static scene in a novel situation, (f) the algorithm infers the best tool and imagines the optimal tool-use.

3. Problem definition

3.1. Learning physical concept

Given a task, the goal of the learning algorithm is to find the true physical concept that best explains why a selected tool and tool-use is optimal.

Rational choice assumption states that human choices are rational and near-optimal. As shown in Fig.5 (a-d), we assume that human chooses the optimal tool and tool-use pg^* (in blue box) based on the true physical concept, so that most of other tools and tool-uses in the hypothesis spaces should not outperform the demonstration.

For instance, let us assume the true physical concept to explain the choice of a tool is to maximize "mass", then other tools should not offer more "mass" than the selected one. If there is a heavier tool not picked by human, it implies that "mass" is not the true physical concept.

During learning stage, we consider the selected tool and tool-use as the only positive training example, and we randomly sample n different combinations of tools and tool-uses pg_i, $i = 1 \cdots n$ in the hypothesis spaces as negative training samples.

Ranking function. Based on the rational choice assumption, we pose the tool recognition as a ranking problem [17], so that the human demonstration should be better than other tools and tool-uses with respect to the learned ranking function.

The goal of the learning is to find a ranking function indicating the essential purposes of tool-use in a given task.

$$R(pg) = \boldsymbol{\omega} \cdot \boldsymbol{\phi}(pg), \tag{1}$$

where $\boldsymbol{\omega}$ are the weighting coefficients of the physical concepts. Intuitively, each coefficient reflects the importance of its corresponding physical concept for the task.

Learning ranking function is equivalent to find the weight coefficients so that the maximum number of pairwise constraints is fulfilled.

$$\forall i \in \{1, \cdots, n\} : \boldsymbol{\omega} \cdot \boldsymbol{\phi}(pg^*) > \boldsymbol{\omega} \cdot \boldsymbol{\phi}(pg_i) \tag{2}$$

In this way, these constraints enforce the human demonstration pg^* has the highest ranking score compared with the other negative samples pg_i under the true physical concept.

We approximate the solution by introducing nonnegative slack variables, similar to SVM classification [17]. This leads to the following optimization problem

$$\begin{aligned} \min \quad & \frac{1}{2} \boldsymbol{\omega} \cdot \boldsymbol{\omega} + \lambda \sum_i^n \xi_i^2 & (3)\\ \text{s.t.} \quad & \forall i \in \{1, \cdots, n\} : & \\ & \boldsymbol{\omega} \cdot \boldsymbol{\phi}(pg^*) - \boldsymbol{\omega} \cdot \boldsymbol{\phi}(pg_i) > 1 - \xi_i^2 & (4)\\ & \xi_i \geq 0, & (5) \end{aligned}$$

where ξ_i is a slack variable for each constraint, and λ is the trade-off parameter between maximizing the margin and satisfying the rational choice constraints.

This is a general formulation for the task-oriented modeling and learning problem, where the parse graph pg includes objects $\mathbf{X}$, human action $\mathbf{A}$ and affordance / functional basis B_A / B_F. In this way, this framework subsumes following special cases: i) object recognition based on appearance and geometry $\boldsymbol{\phi}(\mathbf{X})$, ii) action recognition $\boldsymbol{\phi}(\mathbf{A})$, iii) detecting furniture by their affordance $\boldsymbol{\phi}(B_A)$, and iv) physical concept $\boldsymbol{\phi}(pg_c)$. In this paper, we only focus on learning physical concepts.

In our experiment, we only consider the scenario that the learner only observes one demonstration of the teacher choosing one tool from a few candidates. Instead of feeding a large dataset for training, we are more interested in how much the algorithm can learn from such a small sample learning problem. Therefore, we only infer a single physical concept for functional and affordance basis respectively by iterating over the concept space, while this formulation can be naturally generalized to more sophisticated scenarios for future study.

3.2. Recognizing tools by imagining tool-uses

Traditional object recognition methods assume that visual patterns of the objects in both training and testing sets share the same distribution. However, such assumption does not hold in tool recognition problem. The visual appearances of tools at different situations have fundamental differences. For instance, a hammer and a stone can be used to crack a nut, despite the fact the their appearances are quite different.

In order to address this challenge, we propose this algorithm to recognize tools by essential physical concepts and imagine tool-uses during the inference.

Recognize tools by essential physical concepts. Fortunately, as domain general mechanisms, the essential physical concepts in a given task are invariant across different situations. For instance, a hammer and a stone can be categorized as the same tool to crack a nut due to the similar ability to provide enough "force". In the inference, we use the learned ranking function to recognize the best tool.

$$pg^* = \arg\max \boldsymbol{\omega} \cdot \boldsymbol{\phi}(pg), \tag{6}$$

Imagine tool-use beyond observations. Given an observed image of tool without actually seeing the tool-use, our algorithm first imagines different tool-uses (human action and affordance / functional bases), and then combines the imagined tool-uses with observed tools to recognize the best tool by evaluating the ranking function.

The imagined tool-uses are generated by sampling human action and affordance/functional bases from the hypothesis spaces as shown in Fig.5(c-d). We first assign the trajectories of imaged human hand movement to the affordance basis, then compute the trajectory of functional basis by applying the relative 3D transformation between the two bases. Lastly, we calculate the physical concepts recursively as discussed in Section.2.3, and evaluate the ranking function accordingly.

The ability of imagining tool-use is particularly important for an agent to predict how they can use a tool, and physically interact with their environment.

Moreover, such ability of imagining tool-use enables the agent to actively explore different kinds of tool-use instead of to simply mimic the observed tool-use in human demonstration. Although the tool-use in human demonstration is assumed to be optimal, other tool-uses may be better in different situations. For example, the way you use a stone to crack a nut may be quite different from the way you use a hammer.

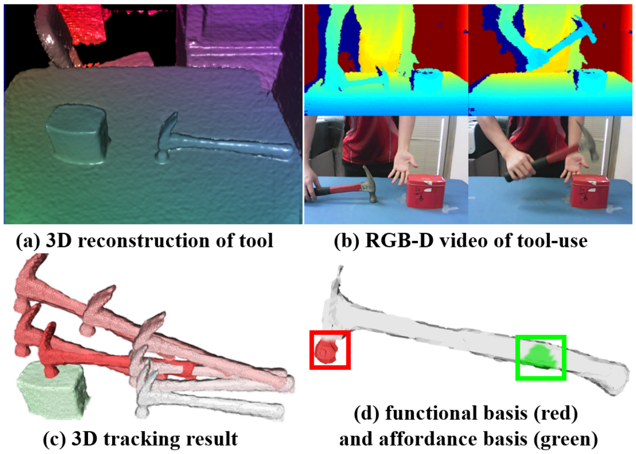

(a) 3D reconstruction of tool
(b) RGB-D video of tool-use
(c) 3D tracking result
(d) functional basis (red)
and affordance basis (green)

Figure 6. Spatial-temporal parsing of human demonstration. (a) Using KinectFusion, we first reconstruct 3D scene, including the tool and the target object. (b) Given a RGB-D video of tool-use by human demonstration, (d) affordance / functiona basis can be detected by (c) 3D tracking.

3.3. Parsing human demonstration

In this section we show how we use the off-the-shelf computer vision algorithms to parse the input RGB-D video of human demonstration.

3D reconstruction. We apply the KinectFusion algorithm [27] to generate a 3D reconstruction of the static scene, including a tool and an object. KinectFusion is GPU optimized such that it can run at interactive rates. Each frame of depth image captured by RGB-D sensors has a lot of missing data. By moving the sensor around, the KinectFusion algorithm fills these holes by combining temporal frames into a smooth 3D point cloud / mesh (Fig.6 (a)). In this work, we only focus on medium sized tool that can be held in one hand, and can be well reconstructed by a consumer-level RGB-D sensor. By fitting the plane of the table, the tool and the target object then can be extracted from background.

3D tracking of tool and target object. Tracking the 3D mesh of tool and target object allows the algorithm to perceive the interactions and detect status changes. In this work, we use an off-the-shelf 3D tracking algorithm based on Point Cloud Library [31]. The algorithm first performs object segmentation using the first depth frame of the RGB-D video, and then invokes particle filtering [26] to track each object segment as well as estimating the 3D orientation frame by frame (Fig.6 (c)).

3D hand tracking. 3D tracking of hand positions and orientations are achieved by 3D skeleton tracking [34]. The skeleton tracking outputs a full body skeleton, including 3D position and orientation of each joint. Without loss of generality, we assume the interacting hand to be the right hand.

Contact detection. Given the tracked 3D hand pose

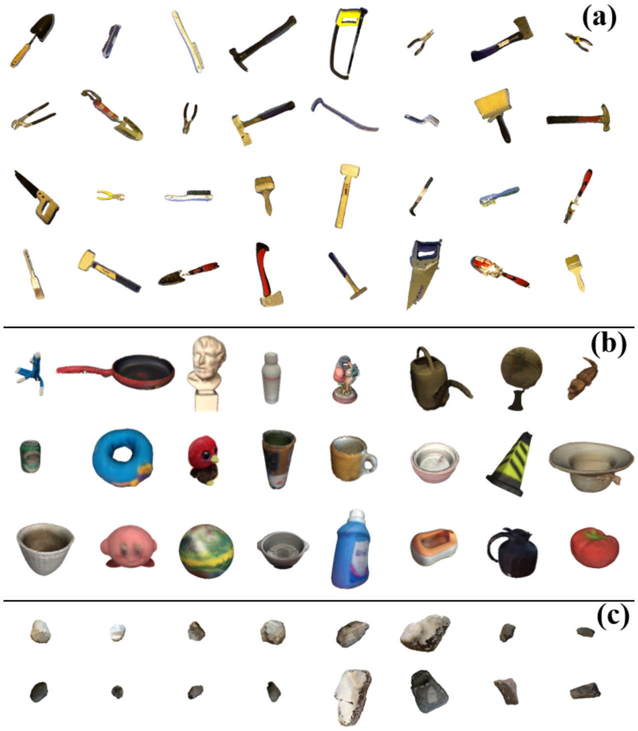


Figure 7. Sample tool instances in dataset. (a) typical tools (b) household objects (c) natural stones.

/ tool / target object, we perform touch detection (Fig.6 (d)) by measuring the euclidean distance among them. The touch detection between the human hand and the tool localizes the 3D location of the affordance basis, while the touch detection between the tool and the target object yields the 3D location of the functional basis.

4. Experiment

In this section, we first introduce our dataset, and evaluate our algorithm in three aspects: (i) learning physical concepts; (ii) recognizing tools; and (iii) imagining tool-uses.

4.1. Dataset

We designed a new Tool & Tool-Use (TTU) dataset for evaluating the recognition of tools and task-oriented objects. The dataset contains a collection of static 3D object instances, together with a set of human demonstrations of tool-use.

The 3D object instances include 452 static 3D meshes, ranging from typical tools, household objects and stones. Some of these object instances are shown in Fig.7. Some typical actions are illustrated in Fig.5. Each action contains a sequence (3-4 seconds) of full body skeletons. Both 3D meshes and human actions are captured by consumer-level RGB-D sensors.

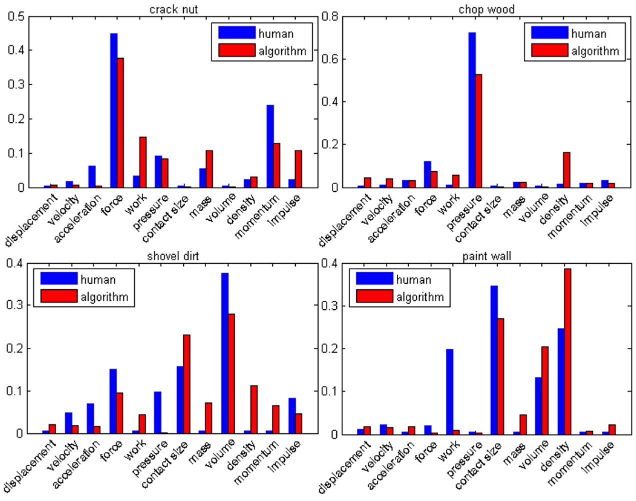


Figure 8. Learning essential physical concepts of tool-use. The red bars represent human judgments about what the essential physical concepts are for each task. The blue bars represent weight coefficients of different physical concepts learned by our algorithm.

4.2. Learning physical concept

We first evaluate our learning algorithm by comparing with human judgments. Forty human subjects annotated the essential physical concepts for four different tasks, the distribution of annotated the essential physical concepts is shown as the blue bars in Fig.8. Interestingly, human subjects have relative consistent common knowledge that force and momentum are useful for cracking nuts, and pressure is important for chopping wood. Our algorithm learned very similar physical concepts as the red bars shown in Fig.8. For the other two tasks i.e. shovel dirt and paint wall, although the human judgments are relatively ambiguous, our algorithm still produces relative similar results of learned physical concepts.

Fig.9 shows an example of learning physical concept for cracking a nut. Given a set of RGB-D images of ten tool candidates in Fig.9 (a) and a human demonstration of tool-use in Fig.9 (b), our algorithm imagines different kinds of tool-use as shown in Fig.9 (c), and ranks them with respect to different physical concepts. By assuming human demonstration is rational and near-optimal, our learning algorithm selects physical concepts by minimizing the number of violations as the red area on the left of Fig.9 (c). For instance, the plot of "force" shows ranked pairs of tool and tool-use with respect to the forces applied on the functional basis. The force produced by human demonstration (the black vertical line) is larger than most of the generated tool-uses, thus it is near-optimal. The instances on the right of Fig.9 (c) are sampled tools and tool-uses. The red ones are the cases outperform human demonstration, while the gray ones are the cases underperform human demonstration.

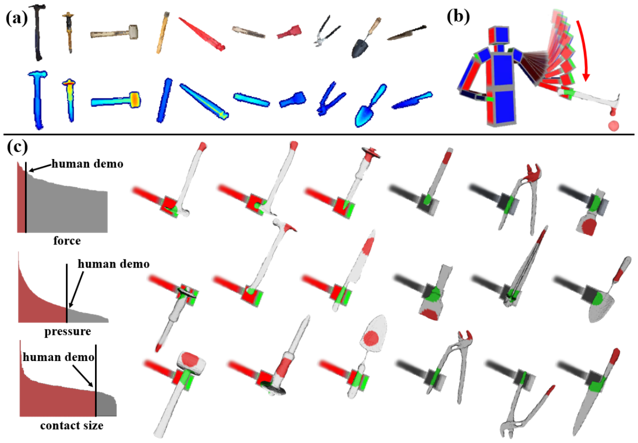


Figure 9. Learning physical concept from single human demonstration for cracking a nut. (a) A set of tool candidates are given by RGB-D images. (b) The human demonstration of tool-use is assumed to be near-optimal. (c) The algorithm sorts all the samples of tool-use with respect to different physical concepts. The black vertical bar represents the human demonstration of tool-use, while the red area and gray area represent samples that outperform and underperform human demonstration receptively. We showed six sampled tool and tool-use, three of which outperform human demonstration, and the others underperform human demonstration. In this cracking nut example, the "forces" is selected as the essential physical concept because there are minimum number of samples that violate the "rational choice assumption" in this case.

4.3. Inferring tools and tool-uses

In the Fig.2, we illustrate qualitative results of inferred tool and tool-use for three tasks, i.e. chop wood, shovel dirt, and paint wall. By evaluating in three scenarios: (a) typical tools, (b) household objects, (c) natural stones, we are interested in the generalization ability of the learned model.

4.3.1 Recognizing tools

We asked four human subjects to rank tool candidates shown in Fig.2. For the task of chopping wood in Fig.10, we plot tool candidates in terms of their average ranking by human subjects (x-axis) and their ranking generated by our algorithm (y-axis).

The three columns show different testing scenarios. We can see that our model learned from canonical cases of tool-use can be easily generalized to recognize tools in novel situation, i.e. household objects and natural stones. The correlation between algorithm ranking and human ranking is consistent across these three scenarios. Sometimes, the algorithm works even better on the stone scenarios.

The three rows represent different levels of tool-use: (a) the "tool-ranking with random use" evaluates the ranking of tools by calculating the expected scores of random tool-use; (b) the "tool-ranking with inferred use" evaluates the ranking of tools by calculating their optimal tool-use inferred by our algorithm; (c) the "tool-ranking with best use" evaluates the ranking of tools by their best uses given by human subjects. The Table.1 summarizes the correlation between human ranking and algorithm ranking on three tasks.

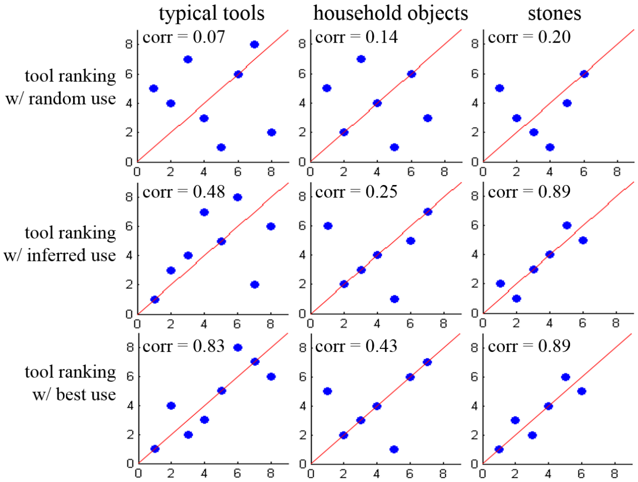


Figure 10. Recognizing tools for chopping wood. The scatters show tool candidates ranked by our algorithm (y-axis) with respect to the average ranking by human subjects (x-axis). The three columns show different testing scenarios, while the three rows represent different levels of tool-use imagined by inference algorithm.

Table 1. Accuracy of tool recognition. This table shows the correlation between the ranking generated by our algorithm and the average ranking annotated by human subjects. The three rows represent different levels of tool-use imagined by our inference algorithm. The qualitative and quantitative ranking results of tool candidates are illustrated in Fig.2 and Fig.10 respectively.

correlation of ranking	chop wood			shovel dirt			paint wall		
algorithm vs. human	tool	object	stone	tool	object	stone	tool	object	stone
tool + random use	0.07	0.14	0.20	0.52	0.32	0.09	0.12	0.11	0.31
tool + inferred use	0.48	0.25	0.89	0.64	0.89	0.14	0.10	0.64	0.20
tool + best use	0.83	0.43	0.89	0.64	0.89	0.14	0.10	0.64	0.20

4.3.2 Imagining tool-uses

We also evaluated the imagined tool-uses in three aspects: human action A, affordance basis B_A and functional basis B_F.

The evaluation of human action is based on the classification of action directions, which are "up", "down", "forward", "backward", "left" and "right". The classification accuracy for this problem over all the experiments is 89.3%. The algorithm can reliably classify the action of cracking a nut as "down". But there are some ambiguities in classifying the action of shoveling dirt, because "left" and "right" are physically similar.

The Fig.11 illustrates three example of imagined affor-

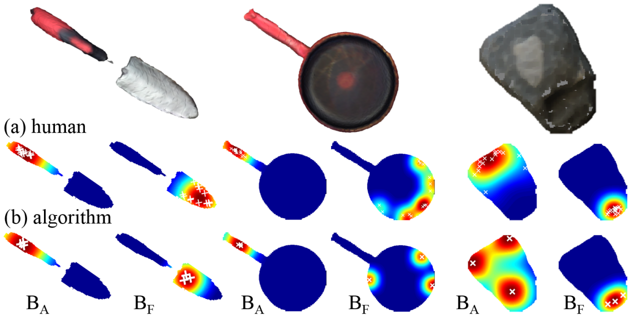


Figure 11. Comparison of human predicted tool-use (a) and algorithm imagined tool-use (b) for shoveling dirt.

Table 2. Error of imagining tool-use for affordance / functional bases (B_A and B_F) . The table shows the 3D distances between their positions imagined by our algorithm and the positions annotated by human subjects. The specific positions for sample tool candidates are shown in Fig.11.

3D distance (cm)	chop wood			shovel dirt			paint wall		
algorithm vs. human	tool	object	stone	tool	object	stone	tool	object	stone
B_A - top 1	1.75	3.02	3.19	1.17	2.03	3.28	0.43	2.48	2.86
B_A - top 3	1.04	2.17	2.81	0.97	0.52	2.21	0.31	2.32	2.67
B_F - top 1	0.48	5.97	3.91	6.98	6.38	0.23	2.35	2.74	2.65
B_F - top 3	0.27	5.92	3.95	2.85	3.29	0.31	1.43	2.64	2.71

dance basis B_A and functional basis B_F. Comparing to human annotations, the algorithm finds very similar positions of affordance basis B_A and functional basis B_F respectively. In Table.2 we show the 3D distances between the positions imagined by our algorithm and the positions annotated by human subjects in centimeter.

5. Discussions

In this paper, we present a new framework for task-oriented object modeling, learning and recognition. An object for a task is represented in a spatial, temporal, and causal parse graph including: i) spatial decomposition of the object and 3D relations with the imagine human pose; ii) temporal pose sequence of human actions; and iii) causal effects (physical quantities on the target object) produced by the object and action. In this inferred representation, only the object is visible, all other components are imagined 'dark' matters. This framework subsumes other traditional problems, such as: (a) object recognition based on appearance and geometry; (b) action recognition based on poses; (c) object manipulation and affordance in robotics. We argue that objects, especially man-made objects, are designed for various tasks in a broad sense [29, 4, 35, 2], and therefore it is natural to study them in a task-oriented framework.

In the following we briefly review related work in the literature of cognitive science, neuroscience, and vision robotics.

5.1. Related work

1) *Cognitiove Science and psychology*. The perception of tools and tool-uses has been extensively studied in cognitive science and psychology. Our work is motivated by the astonishing ability of animal tool-uses [11, 5, 47, 4, 35, 32]. For example, Santos et al.[33] trained two species of monkeys on a task to choose one of two canes to reach food under various conditions that involve physical concepts. Weir et al.[46] reported that New Caledonian crows can bend a piece of straight wire into a hook and successfully used it to lift a bucket containing food from a vertical pipe. These discoveries suggest that animals can reason about the functional properties, physical forces and causal relations of tools using domain general mechanisms. Meanwhile, the history of human tool designing reflects the history of human intelligence development [22, 9, 10, 42]. One argument in cognitive science is that an intuitive physics simulation engine may have been wired in the brain through evolution [3, 39, 41], which is crucial for our capabilities of understanding objects and scenes.

2) *Neuroscience*. Studies in neuroscience [20, 8, 7] found in fMRI experiments that cortical areas in the doral pathway are selectively activated by tools in contrast to faces, indicating a very different pathway and mechanism for object manipulation from that of object recognition. Therefore studying this mechanism will lead us to new directions for computer vision research.

3) *Robotics and AI*. There is also a large body of work studying tool manipulation in robotics and AI. Some related work focus on learning affordance parts or functional object detectors, e.g. [37, 44, 23, 38, 30, 15, 43, 24, 25]. They, however, are still learning high level appearance features, either selected by affordance / functional cues, or through human demonstrations [1], not reason the underlying physical concepts.

4) *Computer vision*. The most related work in computer vision is a recent stream that recognizes functional objects (e.g. chairs) [14, 36, 12, 19, 45, 52, 21, 18] and functional scene (e.g. bedroom) [49, 13, 6, 16] by fitting imagined human poses. The idea of integrating physical-based models has been used for object tracking [40, 28] and scene understanding [50, 51] in computer vision. But our work goes beyond affordance.

5.2. Limitation and future work

In this paper, we only consider handhold physical objects as tools. We do not consider other tools, such as, electrical, digital, virtual or mental tools. Our current object model is also limited by rigid bodies, and can not handle deformable or articulated objects, like scissors, which requires fine-grained hand pose and motion. All these request richer and finer representations which we will study in future work.

Acknowledgment The authors thank Dr. Tao Gao and Prof. Josh Tenenbaum for constructive discussions and motivations. The authors also thank Dr. Brandon Rothrock and Siyuan Qi for his assistance in the experiment. This work is supported by DARPA MSEE FA 8650-11-1-7149, ONR MURI N00014-10-1-0933 and NSF IIS-1423305.

References

[1] B. D. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 57(5):469–483, 2009.

[2] C. Baber. *Cognition and tool use: Forms of engagement in human and animal use of tools*. CRC Press, 2003.

[3] P. W. Battaglia, J. B. Hamrick, and J. B. Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences (PNAS)*, 110(45):18327–18332, 2013.

[4] B. B. Beck. *Animal tool behavior: the use and manufacture of tools by animals*. Garland STPM Pub., 1980.

[5] R. W. Byrne and A. Whiten. Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans (oxford science). 1989.

[6] W. Choi, Y.-W. Chao, C. Pantofaru, and S. Savarese. Understanding indoor scenes using 3d geometric phrases. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 33–40. IEEE, 2013.

[7] S. H. Creem-Regehr and J. N. Lee. Neural representations of graspable objects: are tools special? *Cognitive Brain Research*, 22(3):457–469, 2005.

[8] F. Fang and S. He. Cortical responses to invisible objects in the human dorsal and ventral pathways. *Nature neuroscience*, 8(10):1380–1385, 2005.

[9] S. H. Frey. What puts the how in where? tool use and the divided visual streams hypothesis. *Cortex*, 43(3):368–375, 2007.

[10] K. R. Gibson, K. R. Gibson, and T. Ingold. *Tools, language and cognition in human evolution*. Cambridge University Press, 1994.

[11] J. Goodall. The chimpanzees of gombe: Patterns of behavior. 1986.

[12] H. Grabner, J. Gall, and L. Van Gool. What makes a chair a chair? In *IEEE International Conference on Computer Vision (ICCV)*, pages 1529–1536. IEEE, 2011.

[13] A. Gupta, S. Satkin, A. A. Efros, and M. Hebert. From 3d scene geometry to human workspace. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1961–1968. IEEE, 2011.

[14] S.-B. Ho. Representing and using functional definitions for visual recognition. 1987.

[15] R. Jain and T. Inamura. Learning of tool affordances for autonomous tool manipulation. In *IEEE/SICE International Symposium on System Integration (SII)*, pages 814–819. IEEE, 2011.

[16] Y. Jiang, H. Koppula, and A. Saxena. Hallucinated humans as the hidden context for labeling 3d scenes. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2993–3000. IEEE, 2013.

[17] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 133–142. ACM, 2002.

[18] V. G. Kim, S. Chaudhuri, L. Guibas, and T. Funkhouser. Shape2pose: Human-centric shape analysis. *ACM Transactions on Graphics (TOG)*, 33(4):120, 2014.

[19] H. Kjellström, J. Romero, and D. Kragić. Visual object-action recognition: Inferring object affordances from human demonstration. *Computer Vision and Image Understanding (CVIU)*, 115(1):81–90, 2011.

[20] J. W. Lewis. Cortical networks related to human use of tools. *The Neuroscientist*, 12(3):211–231, 2006.

[21] D. Lin, S. Fidler, and R. Urtasun. Holistic scene understanding for 3d object detection with rgbd cameras. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1417–1424. IEEE, 2013.

[22] W. C. McGrew. *Chimpanzee material culture: implications for human evolution*. Cambridge University Press, 1992.

[23] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor. Learning object affordances: From sensory–motor coordination to imitation. *IEEE Transactions on Robotics (T-RO)*, 24(1):15–26, 2008.

[24] A. Myers, A. Kanazawa, C. Fermuller, and Y. Aloimonos. Affordance of object parts from geometric features. In *Workshop on Vision meets Cognition, CVPR*, 2014.

[25] A. Myers, C. L. Teo, C. Fermüller, and Y. Aloimonos. Affordance detection of tool parts from geometric features. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2015.

[26] S. Nakano, G. Ueno, and T. Higuchi. Merging particle filter for sequential data assimilation. *Nonlinear Processes in Geophysics*, 14(4):395–408, 2007.

[27] R. A. Newcombe, A. J. Davison, S. Izadi, P. Kohli, O. Hilliges, J. Shotton, D. Molyneaux, S. Hodges, D. Kim, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 127–136. IEEE, 2011.

[28] I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *British Machine Vision Conference (BMVC)*, volume 1, page 3, 2011.

[29] F. Osiurak, C. Jarry, and D. Le Gall. Grasping the affordances, understanding the reasoning: toward a dialectical theory of human tool use. *Psychological review*, 117(2):517, 2010.

[30] A. Pieropan, C. H. Ek, and H. Kjellstrom. Functional object descriptors for human activity modeling. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1282–1289. IEEE, 2013.

[31] R. B. Rusu and S. Cousins. 3d is here: Point cloud library (pcl). In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–4. IEEE, 2011.

[32] G. Sabbatini, H. M. Manrique, C. Trapanese, A. D. B. Vizioli, J. Call, and E. Visalberghi. Sequential use of rigid and pliable tools in tufted capuchin monkeys (sapajus spp.). *Animal Behaviour*, 87:213–220, 2014.

[33] L. R. Santos, H. M. Pearson, G. M. Spaepen, F. Tsao, and M. D. Hauser. Probing the limits of tool competence: experiments with two non-tool-using species (cercopithecus aethiops and saguinus oedipus). *Animal cognition*, 9(2):94–109, 2006.

[34] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013.

[35] R. W. Shumaker, K. R. Walkup, and B. B. Beck. *Animal tool behavior: the use and manufacture of tools by animals*. JHU Press, 2011.

[36] L. Stark and K. Bowyer. Achieving generalized object recognition through reasoning about association of function to structure. *IEEE Trans. on Patt. Anal. Mach. Intell (TPAMI)*, 13(10):1097–1104, 1991.

[37] M. Stark, P. Lies, M. Zillich, J. Wyatt, and B. Schiele. Functional object class detection based on learned affordance cues. In *Computer Vision Systems*, pages 435–444. Springer, 2008.

[38] A. Stoytchev. Behavior-grounded representation of tool affordances. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3060–3065. IEEE, 2005.

[39] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011.

[40] A. Q. T.H.Pham, A Kheddar and A. Argyros. Towards force sensing from vision: Observing hand-object interactions to infer manipulation forces. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015.

[41] T. Ullman, A. Stuhlmüller, N. Goodman, and J. Tenenbaum. Learning physics from dynamical scenes. In *Proceedings of the Thirty-Sixth Annual Conference of the Cognitive Science society*, 2014.

[42] K. Vaesen. The cognitive bases of human tool use. *Behavioral and Brain Sciences*, 35(04):203–218, 2012.

[43] K. M. Varadarajan and M. Vincze. Affordance based part recognition for grasping and manipulation. *Workshop on Autonomous Grasping, ICRA*, 2011.

[44] K. M. Varadarajan and M. Vincze. Afrob: The affordance network ontology for robots. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1343–1350. IEEE, 2012.

[45] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu. Modeling 4d human-object interactions for event and object recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3272–3279. IEEE, 2013.

[46] A. A. Weir, J. Chappell, and A. Kacelnik. Shaping of hooks in new caledonian crows. *Science*, 297(5583):981–981, 2002.

[47] A. Whiten, J. Goodall, W. C. McGrew, T. Nishida, V. Reynolds, Y. Sugiyama, C. E. Tutin, R. W. Wrangham, and C. Boesch. Cultures in chimpanzees. *Nature*, 399(6737):682–685, 1999.

[48] D. Xie, S. Todorovic, and S.-C. Zhu. Inferring" dark matter" and" dark energy" from videos. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2224–2231. IEEE, 2013.

[49] Y. Zhao and S.-C. Zhu. Scene parsing by integrating function, geometry and appearance models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3119–3126. IEEE, 2013.

[50] B. Zheng, Y. Zhao, J. C. Yu, K. Ikeuchi, and S.-C. Zhu. Beyond point clouds: Scene understanding by reasoning geometry and physics. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3127–3134. IEEE, 2013.

[51] B. Zheng, Y. Zhao, J. C. Yu, K. Ikeuchi, and S.-C. Zhu. Detecting potential falling objects by inferring human action and natural disturbance. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3417–3424. IEEE, 2014.

[52] Y. Zhu, A. Fathi, and L. Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *European Conference on Computer Vision (ECCV)*, pages 408–424. Springer, 2014.