2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Graph-based Hierarchical Knowledge Representation for Robot Task Transfer from Virtual to Physical World

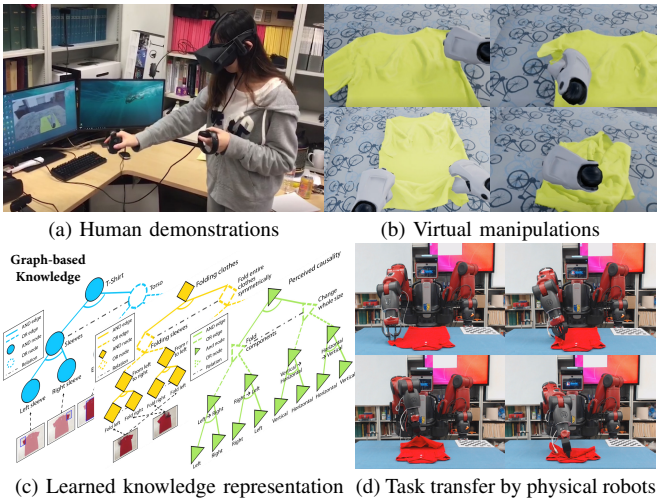
Zhenliang Zhang[1] Yixin Zhu[2] Song-Chun Zhu[2]

***Abstract*— We study the hierarchical knowledge transfer problem using a cloth-folding task, wherein the agent is first given a set of human demonstrations in the virtual world using an Oculus Headset, and later transferred and validated on a physical Baxter robot. We argue that such an intricate robot task transfer across different embodiments is only realizable if an abstract and hierarchical knowledge representation is formed to facilitate the process, in contrast to prior literature of sim2real in a reinforcement learning setting. Specifically, the knowledge in both the virtual and physical worlds are measured by information entropy built on top of a graph-based representation, so that the problem of task transfer becomes the minimization of the relative entropy between the two worlds. An And-Or-Graph (AOG) is introduced to represent the knowledge, induced from the human demonstrations performed across six virtual scenarios inside the Virtual Reality (VR). During the transfer, the success of a physical Baxter robot platform across all six tasks demonstrates the efficacy of the graph-based hierarchical knowledge representation.**

I. Introduction

Robots would be able to rapidly acquire skills for various tasks if a robot could extract and learn the abstract knowledge merely from human demonstrations. Despite virtual training in various virtual environments [1], [2], [3], [4], [5], [6], [7] is readily available in the past few years, there still exists two unsolved challenges. First, since the ground-truth data with its hierarchy is fully accessible in virtual environments, how can we take such an advantage by utilizing the structural data to teach a robot? In this paper, we adopt the And-Or-Graph (AOG) [8] to represent the structure of knowledge in the virtual world and transfer to the physical world. Second, how realistic the virtual world needs to be to afford a positive knowledge transfer and robot execution in the physical world? Here, we study the efficiency of knowledge transfer by comparing different levels of realism in the simulations of virtual environments and the types of interactions.

Our contributions, also summarized in Fig. 1, are three-fold: (i) We model both the virtual world and the physical world from a probabilistic perspective and represent the two worlds' differences as the relative entropy defined on a graph-based representation. (ii) We adopt the AOG grammar model as the knowledge representation and demonstrate its efficacy during the task transfer. (iii) We develop a virtual environment with scene-level discrete spaces for the task of folding clothes. Two crucial factors (*i.e.*, the realism of the simulations, and the levels of interactions) that could affect the knowledge transfer are thoroughly evaluated.



(a) Human demonstrations (b) Virtual manipulations

(c) Learned knowledge representation (d) Task transfer by physical robots

Fig. 1: Overview of the proposed framework for learning abstract knowledge for robot task transfer. (a) Using the Oculus headset and Touch controller, (b) a subject can demonstrate a sequence for the task of clothes folding in a physically realistic VR environment. Our algorithm is able to (c) induce a hierarchical graph-based knowledge representation based on human demonstrations, and (d) transfer it to a physical Baxter robot for execution by minimizing the entropy.

II. Related Work

A. Virtual Training and Sim2Real

Virtual Reality (VR) is ideal for virtual training due to its capability to rapidly construct any types of environments given specific parameters. In the physical world, such a requirement could be difficult (or even infeasible) to construct the apparatus or create various conditions with desired parameters [9], [10], [11]. In this paper, we hope to transfer the robot skill from human demonstrations using a real-time cloth simulator with high realism [12] inside the VR, which nicely combines the advantages of the natural and detailed manipulation provided by VR with the needs for learning robot skills of complex tasks from human demonstrations.

There has been an increasing interest in the field of Sim2Real; it adopts the synthetic and simulated data to assist the learning on a large scale and hopes to transfer the learned models to the physical world. In general, Sim2Real can be categorized into three different strategies: domain adaptation [13], [14], [15], system identification [16], and domain randomization [17]. This paper's focus is perpendicular to the prior literature; we emphasize how to build the hierarchical knowledge representation from human demonstration and verify whether such an abstract knowledge would facilitate the positive task transfer.

[1]Tencent. Email: qkzhang@tencent.com
[2]UCLA Center for Vision, Cognition, Learning, and Autonomy (VCLA) at Statistics Department. Emails: yixin.zhu@ucla.edu, sczhu@stat.ucla.edu

The work reported herein was supported by ONR N00014-19-1-2153, ONR MURI N00014-16-1-2007, and DARPA XAI N66001-17-2-4029.

B. VR for Robotic Tasks

The Artificial Intelligence (AI) community has recently witnessed a trend that shifts towards so-called "embodied AI," wherein the goal is to empower a virtual agent to learn through interacting inside virtual environments. This direction is in stark contrast with the common trend that learns from static image datasets. In this paper, we take a similar stance for virtual agents; however, instead of letting the virtual agents unsupervisedly explore the environment, we emphasize what the agent could learn from human demonstrations and what kind of knowledge representation would help the knowledge transfer to the physical world. To ensure the collection of high-quality data in virtual environment, both state-of-the-art hardware [18], [19] and software [4], [20] systems are adopted.

III. Problem Formulation

A. Probabilistic World Model

Let W^V denote the virtual world, W^P the physical world, O^V a single virtual object, O^P a single physical object. Assuming $\text{num}(O^V) = \text{num}(O^P)$ (*i.e.*, an equal number of objects in both the virtual and physical worlds), the world model is defined as

$$\begin{cases} W^V = \bigcup_{i=1}^{\text{num}(O^V)} O^V \\ W^P = \bigcup_{i=1}^{\text{num}(O^P)} O^P \end{cases}. \tag{1}$$

Let $\{X_1, ..., X_n\}$ denote objects attributes, and assume that any object can be described as a mixture of these attributes

$$\begin{cases} O^V = p^{O^V}(X_1, ..., X_n) \\ O^P = p^{O^P}(X_1, ..., X_n) \end{cases}, \tag{2}$$

where p^{O^V} and p^{O^P} denote the probabilistic distribution of the virtual and the physical object, respectively. Assuming each object's attributes are independent of each other, the virtual and the physical world can be defined as

$$\begin{cases} p^{W^V}(W^V) = \prod_{i=1}^{\text{num}(O^V)} p_i^{O^V}(X_1, ..., X_n) \\ p^{W^P}(W^P) = \prod_{i=1}^{\text{num}(O^P)} p_i^{O^P}(X_1, ..., X_n) \end{cases}. \tag{3}$$

We adopt relative entropy measured by the KL divergence [21] to represent the difference between the virtual and the physical world. The relative entropy between $p^{W^P}(W^P)$ and $p^{W^V}(W^V)$ is denoted as $\mathbb{D}(p^{W^P}(W^P)||p^{W^V}(W^V))$. Since $\mathbb{D}$ is always larger or equal to zero, we have

$$\begin{aligned} &\min\{\mathbb{D}(p^{W^P}(W^P)||p^{W^V}(W^V))\} \\ &= \min\{\sum_{i=1}^{\text{num}(O^V)} \mathbb{D}(p_i^{O^P}(X_1, ..., X_n)||p_i^{O^V}(X_1, ..., X_n))\}, \end{aligned} \tag{4}$$

i.e., the measurement of the similarity between the virtual and the physical world is equivalent to the measurement of the similarity among a set of virtual and physical objects.

We can further decompose the attributes into two sets: a task-related set $\{X_1, ..., X_k\}$ and a task-unrelated set $\{X_{k+1}, ..., X_n\}$, so we have

$$\mathbb{D}(p_i^{O^P}(X_1, ..., X_n)||p_i^{O^V}(X_1, ..., X_n)) = \mathbb{D}(p_i^{O^P}(X_1, ..., X_k)||p_i^{O^V}(X_1, ..., X_k)), \tag{5}$$

where the set of task-unrelated attributes is discarded. By such a simple derivation, our goal is to find a suitable probabilistic distribution of a given virtual object in order to minimize the relative entropy of p^{O^V} and p^{O^P} in the physical world regarding the task-related set of attributes

$$\hat{p}^{O^V} = \arg\min_{p^{O^V}} \mathbb{D}(p^{O^P}(X_1, ..., X_k)||p^{O^V}(X_1, ..., X_k)). \tag{6}$$

B. Graph-based Knowledge for Task Transfer

A good knowledge representation should be able to measure and explain the difference between the virtual and the physical world so that the knowledge transfer between the two worlds is realizable. In this paper, we choose the a specific type of probabilistic graphical model [22], AOG, due to its transparency [23], expressiveness of knowledge representation [24], and ability of contextual adaptation [25].

Let $\mathcal{G}$ denotes the learned graph-based knowledge from the virtual world and K the knowledge error transfer function (KETF) between the virtual and the physical world

$$K = \mathbb{D}(p^{\mathcal{G}^P}(\mathcal{G})||p^{\mathcal{G}^V}(\mathcal{G})) = \mathop{\mathbb{E}}_{p^{\mathcal{G}^P}} log\frac{p^{\mathcal{G}^P}(\mathcal{G})}{p^{\mathcal{G}^V}(\mathcal{G})}. \tag{7}$$

An AOG $\mathcal{G}$ could be induced by a set of "sentences" or instances of a given task, either by object entities and their relations [26], [27], [28], [29], [30], action sequences [31], [32], [33], [34], [35], [23], [36], causal relation [37], [38], [39], [40], or jointly [41]. By setting the "Or" nodes of an AOG, a parse graph is a sampled to provide a deterministic description of the given implementation of a task. Let pg denote a parse graph of the current state of the world W and $p(pg|W)$ its probabilistic distribution

$$p(\mathcal{G}) = \prod_{i=1}^{\text{num}(pg)} (p(pg_i|W)p(W)). \tag{8}$$

Substituting Eq. (8) into Eq. (7), we have

$$K = \mathop{\mathbb{E}}_{p^{\mathcal{G}^P}} \log \frac{\prod_{i=1}^{\text{num}(pg)} (p^{\mathcal{G}^P}(pg_i|W^P)p^{\mathcal{G}^P}(W^P))}{\prod_{i=1}^{\text{num}(pg)} (p^{\mathcal{G}^V}(pg_i|W^V)p^{\mathcal{G}^V}(W^V))}, \tag{9}$$

where $p^{\mathcal{G}^P}(pg_i|W^P)/p^{\mathcal{G}^V}(pg_i|W^V) = 1$ since pg will not be affected by the pixel-level difference between the virtual and the physical world. Hence, the equation can be simplified to

$$K = \mathop{\mathbb{E}}_{p^{\mathcal{G}^P}} \log(\frac{p^{\mathcal{G}^P}(W^P)}{p^{\mathcal{G}^V}(W^V)})^{\text{num}(pg)} = \text{num}(pg) \mathop{\mathbb{E}}_{p^{\mathcal{G}^P}} \log(\frac{p^{\mathcal{G}^P}(W^P)}{p^{\mathcal{G}^V}(W^V)}), \tag{10}$$

where $\text{num}(pg)$ is the total number of parse graphs. Note that the virtual and the physical world share the same pg due to their common logical structure of tasks, but they are grounded to different observations (in terms of pixels). The above equation is in accord with the intuition: When $p^{\mathcal{G}^V}(W^V)$ gets close to $p^{\mathcal{G}^P}(W^P)$, K will be decreased simultaneously; the lower the value of KETF is, the higher the knowledge transfer rate is. Taking Eq. (6) and Eq. (10) together, we can see that increasing the similarity between the attributes of the virtual objects and the attributes of the physical objects will decrease the error of transferred knowledge between the two worlds; this is also well aligned and consistent with our intuition.

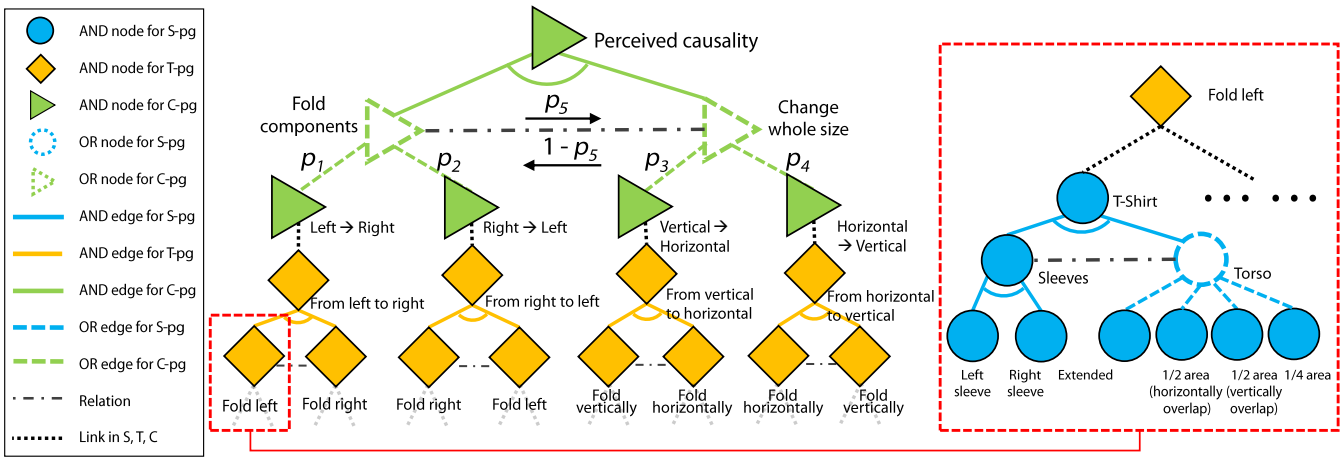


Fig. 2: Illustration of the knowledge representation by an STC-And-Or-Graph (AOG) [8]. A parse or an instance of an AOG is termed as a parse graph (pg). Spatial-pg (S-pg) models the entities and their relations in the scene, Temporal-pg (T-pg) represents the action sequence, and Causal-pg (C-pg) extracts the perceived causality from the human demonstrations. In this example, the probability p_5 determines the order of the sub-tasks, whereas $p_1, \cdots, p_4$ denote the probability of node to be executed.

IV. Bi-Level Hierarchical AOG Learning

Every task encodes two levels of knowledge: A low-level execution layer that carries out the high-level abstract logic layer. The high-level abstract knowledge oversees "what to do," and the low-level execution details govern "how to do." We introduce a bi-level learning scheme defined on AOG.

A. High-Level: AOG Learning

1) AOG as an image grammar: An AOG is defined as $\mathcal{G} = \{\mathcal{V}, \mathcal{R}, \mathcal{P}\}$, where $\mathcal{V} = \mathcal{V}^{\text{And}} \cup \mathcal{V}^{\text{Or}} \cup \mathcal{V}^{\text{Terminal}}$ consists of a disjoint set of And-nodes, Or-nodes, and Terminal-nodes. An And-node is the decomposition of a large entity, indicating all of its child nodes should exist simultaneously. An Or-node is a branching choice; its child nodes can exist only one in a given implementation of a task. A terminal node denotes a specific physical element (for S-AOG), action (for T-AOG), or logic (for C-AOG). $\mathcal{R}$ is the production rule, representing a set of contextual relations among the nodes in $\mathcal{V}$. $\mathcal{P}$ denotes the probabilities for elements in $\mathcal{R}$.

2) Spatial, Temporal, and Causal AOG: A spatial AOG (S-AOG) is a representation of the spatial distribution of every object and the compositional relations of their parts. A temporal AOG (T-AOG) is a representation of actions in a temporal order. A causal AOG (C-AOG) is a representation of the perceived causality, derived from S-AOG and T-AOG.

3) Grammar Learning and Prediction: An AOG can be induced using grammar induction based on observational signals [42], [31]. Given a learned grammar and a partially observed sequence, one can further predict the next STC-unit [32] using certain language parsers, *e.g.*, an Earley parser in NLTK [43]. A learned AOG is shown in Fig. 2.

4) Parameters Learning: After the induction of semantics and syntax, the algorithm also needs to estimate the parameters to assign probabilities of edges between any nodes, *i.e.*, the weight/likelihood of an Or-node to choose a branch. Each observed task execution sequence corresponds to a specific parse graph. Intuitively, the frequency of a node in the entire observed data would indicate the node's probability given its parent node, which can be formulated as an MLE.

Formally, let v denote a child node of an Or-node in an AOG, v_i the i^{th} possible value of this child node, $p(v \to v_i)$ the probability that v appears with the i^{th} value v_i, defined by $p(v \to v_i) = \frac{\#(v \to v_i)}{\sum_{j=1}^{\text{num}(v)} \#(v \to v_j)}$, where $\#(v \to v_i)$ denotes the frequency that v appears with the i^{th} value v_i, and $\text{num}(v)$ is the total number of the possible values of node v.

B. Low-level: Atomic Action Learning

The atomic action refers to an action that cannot be further decomposed as the robot manipulation action. Two crucial ingredients are needed to learn the optimal trajectory so that each atomic action can be easily transferred from the source trajectory performed in the virtual environment to the target trajectory in the physical world: a scale parameter and critical points along the estimated trajectory.

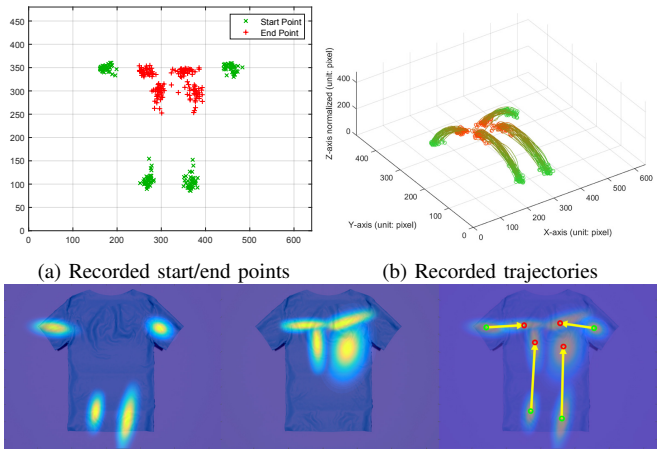
1) Scale Parameter: The entire 3D trajectories of all atomic actions performed in human demonstrations are recorded. Let T^V denote the recorded trajectory in the virtual world for a given task, and T^P the target trajectory to perform in the physical world, we have

$$\begin{cases} T^V = \{\mathbf{q}^{V,s}, \mathbf{q}^{V,e}, \cup_{i=1}^{\text{num}(\mathbf{q}^{V,t})} \mathbf{q}_i^{V,t}\} \\ T^P = \{\mathbf{q}^{P,s}, \mathbf{q}^{P,eend}, \cup_{i=1}^{\text{num}(\mathbf{q}^{P,t})} \mathbf{q}_i^{P,t}\} \end{cases}, \tag{11}$$

where $\mathbf{q}^{V,s}$ and $\mathbf{q}^{V,e}$ are the start/end point of the virtual trajectory, respectively, and $\mathbf{q}_i^{V,t}$ a point on the virtual trajectory excluding the start/end point; $\mathbf{q}^{P,s}$, $\mathbf{q}^{e}$, and $\mathbf{q}_i^{P,t}$ are the corresponding variables in the physical world.

We first estimate the projection of the scale vector to compare the trajectories on the plane (*e.g.*, a tabletop). Let $\mathbf{q}^{V,scale} = \mathbf{q}^{V,e} - \mathbf{q}^{V,s}$ denote the scale vector of the trajectory in the virtual world, and $\mathbf{q}^{P,scale} = \mathbf{q}^{P,e} - \mathbf{q}^{P,s}$ the scale vector in the physical world. Let $\mathbf{n}^V$ be the unit vector of the projection for $\mathbf{q}^{V,scale}$, so that the projection length of $\mathbf{q}^{V,scale}$ is $L^{V,Proj} = \mathbf{q}^{V,scale} \cdot \mathbf{n}^V$. Similarly, the projection length of $\mathbf{q}^{P,scale}$ in the physical world is $L^{P,Proj} = \mathbf{q}^{P,scale} \cdot \mathbf{n}^P$. The scale parameter θ between the trajectories in the two worlds is then given by the ratio

$$\theta = \frac{L^{P,Proj}}{L^{V,Proj}}. \tag{12}$$



(a) Recorded start/end points (b) Recorded trajectories

(c) Estimated start points (d) Estimated end points (e) Estimated trajectories

Fig. 3: Trajectory analysis using Gaussian fitting. Given a human demonstration of cloth-folding sequences in terms of (a) grasp points and (b) trajectories, our algorithm aggregates the raw data and fits the (c) start points and (d) end points with a Gaussian distribution; (e) folding trajectories are further estimated.

2) Trajectory Interpolation: Since recorded trajectories in human demonstrations may contain different numbers of 3D points, the algorithm needs to interpolate the trajectory so that the recorded trajectories share the same structure.

Let $\{\mathbf{q'}_i^{V,start}\}$ denote the set of recorded start points projected onto a 2D plane, $\{\mathbf{q'}_i^{V,end}\}$ the set of end points projected onto a 2D plane (Fig. 3a), and $\{T_i^V\}$ the recorded trajectories (Fig. 3b). We assume the 2D position of the grasp point on the plane follows a Gaussian distribution

$$p(\mathbf{q'}^{V,start}) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp(-\frac{1}{2}(\mathbf{q'}^{V,start} - \mu^{V,start})^T \Sigma^{-1} (\mathbf{q'}^{V,start} - \mu^{V,start})), \quad (13)$$

where $\mathbf{q'}^{V,end}$ shares the same form of Gaussian distribution as $\mathbf{q'}^{V,start}$; see examples in Figs. 3c and 3d.

Let $\text{num}(T_i^V)$ denote the number of points in the trajectory T_i^V, including the start and end points, $n_i = \text{num}(T_i^V)$ the length of a trajectory, and $n^{max} = \max(\text{num}(T_i^V))$ the length of the longest trajectory. We interpolate points in every trajectory so that the total number of points is equal to n^{max}; *i.e.*, for T_i^V, the number of points to be added is $n^{max} - n_i$. Specifically, we procedurally add points to T_i^V by the following steps: (i) Find two neighboring points $\mathbf{q}_j$ and $\mathbf{q}_{j+1}$ that have the largest distance. (ii) Insert $\mathbf{q}^{ins} = 1/2(\mathbf{q}_j + \mathbf{q}_{j+1})$ between $\mathbf{q}_j$ and $\mathbf{q}_{j+1}$. (iii) Loop (i) and (ii) for $n^{max} - n_i$ times and update T_i^V.

3) Trajectory Estimation: We approximate the optimal trajectory by combining interpolated trajectories

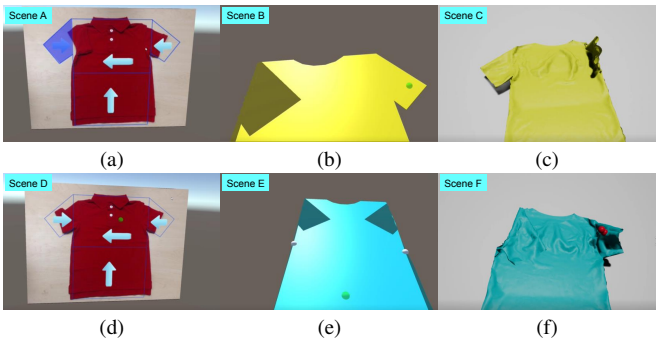
$$p(T_i^V) = \frac{1}{2}(p(\mathbf{q'}_i^{V,start} - \mu^{V,start}) + p(\mathbf{q'}_i^{V,end} - \mu^{V,end})). \quad (14)$$

The approximated trajectory is given by

$$\hat{T} = \sum_{i=1}^{\text{num}(T_i^V)} \omega_i T_i^V, \text{ where } \omega_i = \frac{p(T_i^V)}{\sum_{i=1}^{\text{num}(T_i^V)} p(T_i^V)}. \quad (15)$$

Fig. 3b visualizes an example of the estimated trajectory. Based on Eqs. (12) and (15), the trajectory in the physical world could be approximated by

$$\hat{T}^P = \theta \hat{T}^V. \quad (16)$$



(a) (b) (c)

(d) (e) (f)

Fig. 4: Six virtual environments with different fidelity for learning and evaluating the knowledge of folding clothes. (a) Grasp areas are recorded. (b) Grasp point is visible and recorded. (c) Full physics-based simulation but with only grasp points recorded. (d) Grasp point is recorded but the trajectory is the pre-defined line. (e) Grasp point is recorded but the trajectory is estimated. (f) Full physics-based simulation with the complete trajectory recorded.

V. Experiments[1]

A. Hypothesis Space

We consider two factors that would affect the knowledge transfer between the virtual and the physical world: (i) The realism of the simulated physics of the virtual environment, α, where $0 \leqslant \alpha \leqslant 1$. (ii) The realism of the interaction inside the virtual environment, β, where $0 \leqslant \beta \leqslant 1$. An intuitive expectation would be: the virtual environment with low fidelity (α and β close to 0) would result in a high knowledge transfer error; conversely, the virtual environment in high fidelity leads to a low knowledge transfer error.

B. Hardware and Software

1) System Setup: We build the system based on the Oculus Rift CV1, capable of capturing the human's head and hands' positions and poses. We further use the Oculus Touch controllers as the interface for human subjects to demonstrate cloth folding in the virtual world. During the evaluation, a physical Baxter robot is deployed to evaluate folding clothes' performance in the physical world. Our system is implemented on a desktop with an Intel i7-7700K CPU and an NVIDIA GeForce GTX 1070Ti Graphics Card.

2) Construction of Virtual Environments: We construct six different virtual scenes of clothes-folding tasks with different fidelity in Unity3D and Unreal Engine 4; see Fig. 4. Table I lists the detailed configurations for every scene and the supported features (*i.e.*, grasp point, and trajectory). For each virtual scene, we collect 30 sequences of trials. Every sequence includes the full 3D data of human demonstrations.

TABLE I: Configurations of constructed virtual scenes. "Phy.": physics. "Int.": interaction. "Pre.": the trajectory is pre-defined regardless of how the grasp points are distributed. "Est." trajectory is estimated by the spatial distribution of grasp points.

Scene Description	Fidelity Level		Data Feature		
(3 × Phy. and 2 × Int.)	Physics	Interaction	Procedure	Grasp Point	Trajectory
A. Low-Phy. & Low-Int.	α_1	β_1	√	×	×
B. Median-Phy. & Low-Int.	α_2	β_1	√	√	×
C. High-Phy. & Low-Int.	α_3	β_1	√	√	×
D. Low-Phy. & High-Int.	α_1	β_2	√	√	√ (*Pre.*)
E. Median-Phy. & High-Int.	α_2	β_2	√	√	√ (*Est.*)
F. High-Phy. & High-Int.	α_3	β_2	√	√	√

[1]Preliminary results have been reported as an extended abstract [44].

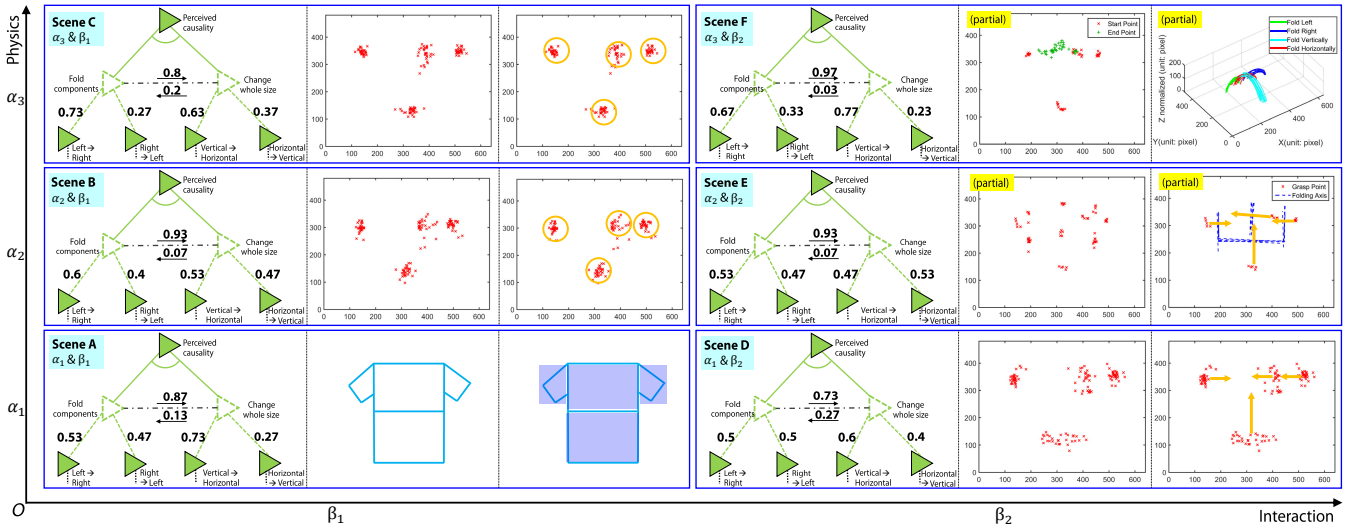


Fig. 5: Visualization of the learned knowledge represented by AOGs in six different scenes with various fidelity of physics and interaction. The horizontal axis is the fidelity of interaction, and the vertical axis is the fidelity of physics. In every block parameterized by α and β, the left column is the induced C-AOG based on the observed data, and the number next to an edge is the branching probability. The middle column is the grasp point collected in virtual scenes. The right column is the final knowledge of atomic actions.

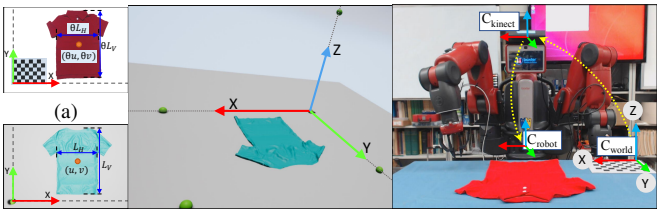


Fig. 6: The transformation of coordinate systems between the physical environment and the virtual environment. (a) Top view of the physical environment. (b) Top view of the virtual environment. θ denotes the scale mapping between the two. (c) The world coordinate system is highlighted by green balls. (d) The plane with a chessboard pattern determines the world coordinate system.

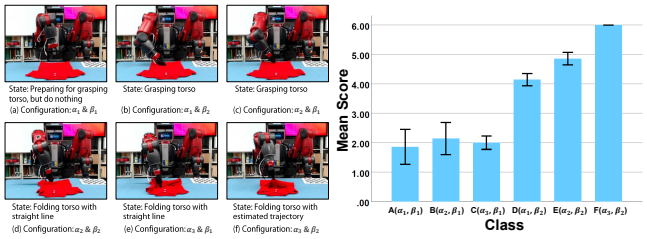


Fig. 7: Results of folding physical clothes using the learned knowledge. (a) Examples of the executions by a physical Baxter robot based on the learned knowledge from virtual scenes. (b) Human subject rating of the robot performance.

C. Evaluation Protocol

1) Subjects: 14 graduate students (7 males and 7 females; ages: 20–28) at UCLA were recruited in a within-subject design. All the subjects do not have similar VR experiences. They are asked to rank the results according to how well the clothes were folded. A better folding result corresponds to a higher score. The final mean score across the six scenes would serve as a reliable indicator of the human utility regarding the states of clothes, showing the extent to which the knowledge has been successfully transferred.

2) Evaluation of Knowledge Transfer: We compute six AOG-based knowledge representation to evaluate how different fidelity affects the learning results. Intuitively, the more complete the information provided, the better knowledge can be learned, and the higher the knowledge transfer rate is. The intuition of knowledge transfer is verified on a physical Baxter robot platform. Fig. 6 illustrates an example of coordinate transfer on "Scene F."

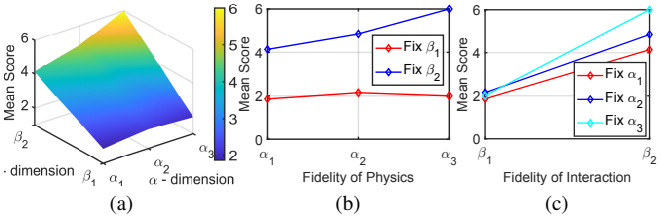
3) Evaluation of Knowledge Generalization: We test the learned knowledge with new examples of clothes to evaluate the model generalization. Specifically, we adopt 3 different categories of clothes (*i.e.*, trousers, shorts, and dress). The evaluation criteria are based on (i) the final state of the clothes and (ii) the sequence of actions.

D. Results

1) Visualization of Extracted Knowledge: Section IV-A shows qualitative results, wherein the left column of every block includes the learned C-AOG and its parameters, the middle column shows the grasp points, the right column visualizes the learned atomic actions. Specifically,

- Scene A (α_1 and β_1): as we can only obtain the folding actions by triggering animations, the learned knowledge is only the manipulation area without any grasp point.
- Scene B (α_2 and β_1): the learned knowledge is the estimated geometrical center of every point cluster.
- Scene C (α_3 and β_1): similar to Scene B, we can use the collected grasp points to estimate the geometrical center.
- Scene D (α_1 and β_2): we obtain both grasp points and folding actions by folding logic encoded in the animation.
- Scene E (α_2 and β_2): both grasp points and the estimated straight lines for folding actions are obtained.
- Scene F (α_3 and β_2): both grasp points and the complete trajectories are recorded.

2) Factors in Knowledge Transfer: We apply the learned knowledge from virtual environments to a physical robot to evaluate knowledge transfer; see an example in Fig. 7a. The robot executions are recorded and presented to human subjects for ratings; the scores are summarized in Fig. 7b.



(a) (b) (c)

Fig. 8: Utility landscape of folding clothes based on human subject ratings. (a) The fitted mesh based on the ranking results of users, which shows the basic landscape of human utility. (b) Fixing the realism of the interaction (β), the mean score varies with the realism of the physics. (c) Fixing the realism of the physics (α), the mean score varies with the realism of the interaction.

A Wilcoxon test is conducted to identify whether different values of β (realism of interaction) significantly affect the ranking score of human subjects, when the value of α (realism of physics) is fixed. The test results indicate that there indeed exists a significant difference between the ranking scores for Group A and Group D ($Z = 3.359, p < .001$) when $\alpha = \alpha_1$, Group B and Group E ($Z = 3.336, p < .05$) when $\alpha = \alpha_2$, and Group C and Group F ($Z = 3.556, p < .05$) when $\alpha = \alpha_3$. On average, Group D is better than Group A, Group E better than Group B, and Group F better than Group C. Such results show the realism of interaction will benefit the knowledge learning rate regardless of the realism of physics.

A Friedman test is performed to identify whether different values of α (realism of physics) significantly affect the ranking score of human subjects, when the value of β (realism of interaction) is fixed. The test results indicates that there is no significant difference among the ranking scores ($\chi^2(2, N = 14) = .571, p = .751$) when $\beta = \beta_1$. But there is a significant difference among the ranking scores ($\chi^2(2, N = 14) = 24.571, p < .001$) when $\beta = \beta_2$. Such a result reveals that Group A, B, and C do not show any statistical difference; *i.e.*, the realism of physics will not affect the knowledge learning when $\beta = \beta_1$.

We further conduct the Wilcoxon test to identify whether different values of α (realism of physics) significantly affect the ranking score of human subjects, when the value of β (realism of interaction) is fixed. The results indicate there are significant differences between the ranking scores for Group D and Group E ($Z = 2.673, p < 0.01$), Group E and Group F ($Z = 3.557, p < 0.001$), and Group D and Group F ($Z = 3.557, p < 0.001$). Such an analysis shows that the realism of physics can have a positive correlation with the knowledge learning rate when $\beta = \beta_2$.

We fit a utility landscape based on collected human data, reflecting the trend of human utility regarding two fidelity factors of the cloth-folding task; see Fig. 8a. We also visualize the effect of each factor when another factor is fixed; see Figs. 8b and 8c. The results are consistent with the derivation discussed in Section III.

3) Knowledge Generalization: We test the knowledge generalization/transfer with a new set of clothes, as shown in the first column of Fig. 9. Specifically, the robot is asked to fold the clothes with an action sequence so that the clothes' size becomes smaller in time. The learned high-level logic

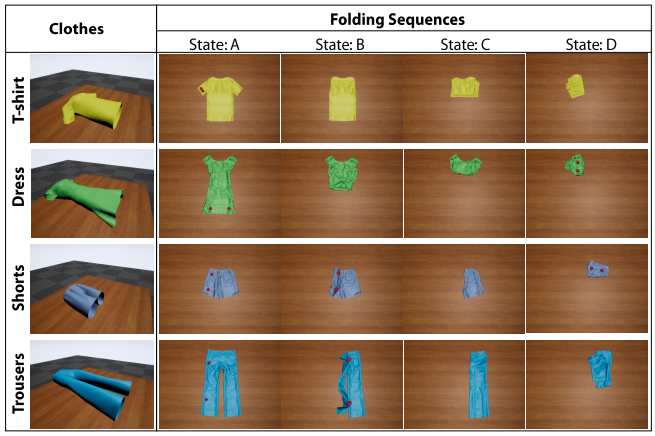
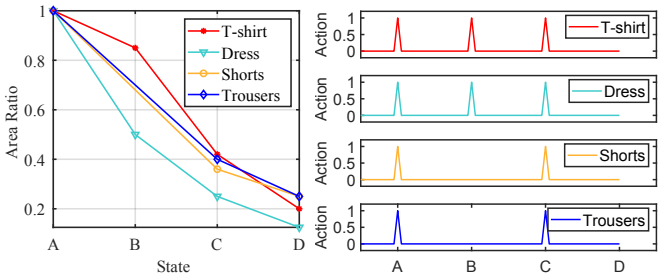


Fig. 9: Knowledge generalization/transfer of folding various unseen clothes. The red balls indicate virtual fingers.



(a) (b)

Fig. 10: Two evaluation criteria for folding clothes. (a) The size of clothes becomes smaller when the folding action sequences unfold. (b) Actions taken in a given sequence for different folding tasks.

guides the robot to determine the task planner to shrink the size of clothes, and the low-level atomic actions shepherd the robot to find appropriate grasping points and action trajectories. We evaluate the knowledge generalization based on two evaluation criteria: the size of clothes and the folding action sequence; see Fig. 10.

VI. Conclusion and Discussion

In this paper, we demonstrate that the AOG-based knowledge representation could help eliminate the gap between the virtual and the physical world. The information entropy provides a new perspective to measure the similarity between the virtual and the physical world. In experiments, we also demonstrate that decreasing the difference between the virtual and the physical world along the crucial dimension (*e.g.*, the realism of physics and interaction) will improve the knowledge transfer rate. We further analyze how the different factors could affect the knowledge transfer rate, indicating that both the realism of the physics and the interaction have a positive correlation with the knowledge transfer rate.

Some limitations are to be addressed. (i) We rely on the off-the-shelf robotic packages to handle the control details, which could be another crucial factor that affects the knowledge transfer rate. (ii) The current state-of-the-art cloth simulation with real-time interaction still has limited effects; one major issue is the self-penetration, which could affect human demonstrations and subject rating. We hope a better real-time simulator would help in the near future.

REFERENCES

[1] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," *arXiv preprint arXiv:1711.03938*, 2017.

[2] E. Kolve, R. Mottaghi, D. Gordon, Y. Zhu, A. Gupta, and A. Farhadi, "Ai2-thor: An interactive 3d environment for visual ai," *arXiv preprint arXiv:1712.05474*, 2017.

[3] M. Savva, A. X. Chang, A. Dosovitskiy, T. Funkhouser, and V. Koltun, "Minos: Multimodal indoor simulator for navigation in complex environments," *arXiv preprint arXiv:1712.03931*, 2017.

[4] X. Xie, H. Liu, Z. Zhang, Y. Qiu, F. Gao, S. Qi, Y. Zhu, and S.-C. Zhu, "Vrgym: A virtual testbed for physical and interactive ai," in *Proceedings of the ACM Turing Celebration Conference*, 2019.

[5] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik *et al.*, "Habitat: A platform for embodied ai research," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[6] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang *et al.*, "Sapien: A simulated part-based interactive environment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[7] F. Xia, W. B. Shen, C. Li, P. Kasimbeg, M. E. Tchapmi, A. Toshev, R. Martín-Martín, and S. Savarese, "Interactive gibson benchmark: A benchmark for interactive navigation in cluttered environments," *Robotics and Automation Letters (RA-L)*, vol. 5, no. 2, pp. 713–720, 2020.

[8] S.-C. Zhu and D. Mumford, "A stochastic grammar of images," *Foundations and Trends® in Computer Graphics and Vision*, vol. 2, no. 4, pp. 259–362, 2007.

[9] C. Li, W. Liang, C. Quigley, Y. Zhao, and L.-F. Yu, "Earthquake safety training through virtual drills," *Proceedings of IEEE Transactions on Visualization & Computer Graph (TVCG)*, vol. 23, no. 4, pp. 1275–1284, 2017.

[10] T. Ye, S. Qi, J. Kubricht, Y. Zhu, H. Lu, and S.-C. Zhu, "The martian: Examining human physical judgments across virtual gravity fields," *Proceedings of IEEE Transactions on Visualization & Computer Graph (TVCG)*, vol. 23, no. 4, pp. 1399–1408, 2017.

[11] D. Wang, J. Kubrlcht, Y. Zhu, W. Lianq, S.-C. Zhu, C. Jiang, and H. Lu, "Spatially perturbed collision sounds attenuate perceived causality in 3d launching events," in *Conference on Virtual Reality and 3D User Interfaces (VR)*, 2018.

[12] M. Macklin, M. Müller, N. Chentanez, and T.-Y. Kim, "Unified particle physics for real-time applications," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 4, p. 153, 2014.

[13] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 53–69, 2015.

[14] G. Csurka, "Domain adaptation for visual applications: A comprehensive survey," *arXiv preprint arXiv:1702.05374*, 2017.

[15] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.

[16] K. J. Åström and P. Eykhoff, "System identification—a survey," *Automatica*, vol. 7, no. 2, pp. 123–162, 1971.

[17] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*, 2017.

[18] H. Liu, Z. Zhang, X. Xie, Y. Zhu, Y. Liu, Y. Wang, and S.-C. Zhu, "High-fidelity grasping in virtual reality using a glove-based system," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2019.

[19] H. Liu, X. Xie, M. Millar, M. Edmonds, F. Gao, Y. Zhu, V. J. Santos, B. Rothrock, and S.-C. Zhu, "A glove-based system for studying hand-object manipulation via joint pose and force sensing," in *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*, 2017.

[20] J. Lin, X. Guo, J. Shao, C. Jiang, Y. Zhu, and S.-C. Zhu, "A virtual reality platform for dynamic human-scene interaction," in *SIGGRAPH ASIA 2016 Virtual Reality meets Physical Reality: Modelling and Simulating Virtual Humans and Environments*, 2016.

[21] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[22] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[23] M. Edmonds, F. Gao, H. Liu, X. Xie, S. Qi, B. Rothrock, Y. Zhu, Y. N. Wu, H. Lu, and S.-C. Zhu, "A tale of two explanations: Enhancing human trust by explaining robot behavior," *Science Robotics*, vol. 4, no. 37, 2019.

[24] Y. Zhu, T. Gao, L. Fan, S. Huang, M. Edmonds, H. Liu, F. Gao, C. Zhang, S. Qi, Y. N. Wu, J. Tenenbaum, and S.-C. Zhu, "Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense," *Engineering*, vol. 6, no. 3, pp. 310–345, 2020.

[25] J. Launchbury, "A darpa perspective on ai," https://youtu.be/-O01G3tSYpU, February 2017.

[26] C. Jiang, S. Qi, Y. Zhu, S. Huang, J. Lin, L.-F. Yu, D. Terzopoulos, and S.-C. Zhu, "Configurable 3d scene synthesis and 2d image rendering with per-pixel ground truth using stochastic grammars," *International Journal of Computer Vision (IJCV)*, vol. 126, no. 9, pp. 920–941, 2018.

[27] S. Qi, Y. Zhu, S. Huang, C. Jiang, and S.-C. Zhu, "Human-centric indoor scene synthesis using stochastic grammar," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[28] S. Huang, S. Qi, Y. Zhu, Y. Xiao, Y. Xu, and S.-C. Zhu, "Holistic 3d scene parsing and reconstruction from a single rgb image," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2018.

[29] Y. Chen, S. Huang, T. Yuan, Y. Zhu, S. Qi, and S.-C. Zhu, "Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2019.

[30] T. Yuan, H. Liu, L. Fan, Z. Zheng, T. Gao, Y. Zhu, and S.-C. Zhu, "Joint inference of states, robot knowledge, and human (false-)beliefs," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2020.

[31] K. Tu, M. Pavlovskaia, and S.-C. Zhu, "Unsupervised structure learning of stochastic and-or grammars," in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

[32] S. Qi, S. Huang, P. Wei, and S.-C. Zhu, "Predicting human activities using stochastic grammar," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2017.

[33] M. Edmonds, F. Gao, X. Xie, H. Liu, S. Qi, Y. Zhu, B. Rothrock, and S.-C. Zhu, "Feeling the force: Integrating force and pose for fluent discovery through imitation learning to open medicine bottles," in *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*, 2017.

[34] S. Qi, B. Jia, and S.-C. Zhu, "Generalized earley parser: Bridging symbolic grammars and sequence data for future prediction," in *Proceedings of International Conference on Machine Learning (ICML)*, 2018.

[35] H. Liu, C. Zhang, Y. Zhu, C. Jiang, and S.-C. Zhu, "Mirroring without overimitation: Learning functionally equivalent manipulation actions," in *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[36] S. Qi, B. Jia, S. Huang, P. Wei, and S.-C. Zhu, "A generalized earley parser for human activity parsing and prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[37] A. Fire and S.-C. Zhu, "Learning perceptual causality from video," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 7, no. 2, pp. 1–22, 2015.

[38] Y. Xu, L. Qin, X. Liu, J. Xie, and S.-C. Zhu, "A causal and-or graph model for visibility fluent reasoning in tracking interacting objects," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[39] M. Edmonds, S. Qi, Y. Zhu, J. Kubricht, S.-C. Zhu, and H. Lu, "Decomposing human causal learning: Bottom-up associative learning and top-down schema reasoning," in *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2019.

[40] M. Edmonds, X. Ma, S. Qi, Y. Zhu, H. Lu, and S.-C. Zhu, "Theory-based causal transfer: Integrating instance-level induction and abstract-level structure learning," in *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[41] Y. Zhu, Y. Zhao, and S.-C. Zhu, "Understanding tools: Task-oriented object modeling, learning and recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[42] Z. Solan, D. Horn, E. Ruppin, and S. Edelman, "Unsupervised learning of natural languages," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 102, no. 33, pp. 11 629–11 634, 2005.

[43] S. Bird and E. Loper, "Nltk: the natural language toolkit," in *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*. Association for Computational Linguistics, 2004, p. 31.

[44] Z. Zhang, J. Guo, D. Weng, Y. Liu, and Y. Wang, "Extracting and transferring hierarchical knowledge to robots using virtual reality," in *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 2020.