Visual Primitives for Abductive Reasoning

Niko Grupen and Ross Knepper
Department of Computer Science, Cornell University, Ithaca, NY
Email: {niko, rak}@cs.cornell.edu

I. Introduction

Humans organize the world into objects that adhere to specific conceptual rules. For example, most adults would agree that it is impossible for two individual objects to occupy the same physical space at the same time, without one being contained in the other. Similarly, it is impossible for one object to exist in some location B, after existing in another location A, without having traversed a path between the two locations [3]. The ability to quickly learn intuitive theories that organize our surroundings into relatable concepts is a trait that enables humans to build intricate mental models of the world. Such a capacity also allows humans to reason about and explain situations they encounter in the world and is thus a key component of the foundational knowledge we call intuitive physics.

It is not surprising then that infants even as young as 4-6 months exhibit the beginnings of physical reasoning [10, 47]. Seminal work in Cognitive Science and Developmental Psychology has shown that infants are able to reason about basic physical phenomena—including object individuation by relative motion [26, 51], object permanence [3], spatiotemporal continuity [48]—despite their limited perceptual capabilities compared to grown adults. Presented with a complex and largely uninterpretable environment, infants seek structure by developing simple physical rules and building upon them.

Despite successes in statistical pattern recognition tasks, modern deep learning systems have not yet matched the reasoning power of a 4-6 month old infant. Finding, testing, and revising possible explanations for observations in a given environment—abductive inference—can inform how an agent builds and refines models of the world. Explanation through abduction enables a robot to build such models in a task-agnostic manner, mirroring the world-modeling capabilities of the developing infant. We posit that task-agnostic models of intuitive physics will enable more robust robotic intelligence by bridging the gap between the experiential learning capabilities of model-free algorithms and the explanatory power of model-based learning.

An effective physical reasoning system must address two problems. The first problem is transforming raw sensory information into a set of concepts, preferably ones that are compositional. The second problem involves reasoning over those concepts in an interpretable way. The former is a problem broached by representation learning [8], where the goal is to extract visual concepts from an observed sensory input [18, 20]. In this paper we focus on the latter, proposing an abductive framework for reasoning over visual primitives. We describe a hierarchical Bayesian model that propagates belief over a set of distinct hypotheses representing possible explanations of an observed scene. Our model generates hypotheses composed of individual visual primitives regarding object motion as well as relational visual primitives that capture physical interactions between two or more objects (i.e. occlusion). Further, we provide a new video dataset, PhysSprites, containing synthetic recreations of the environments traditionally used to study intuitive physics in infants [3, 26, 48]. The PhysSprites dataset, shown in Figure 1, serves as the basis for our preliminary evaluation in Section IV.

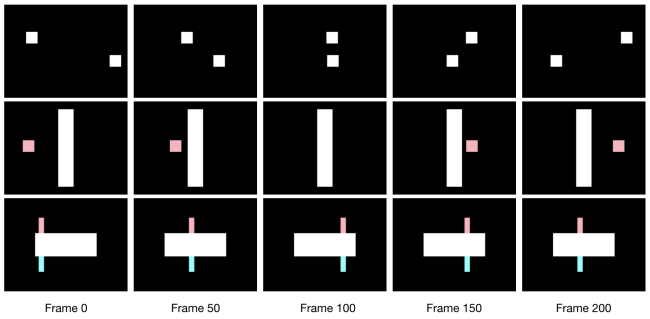

Fig. 1: Sample frames from the PhysSprite dataset. *Top-to-Bottom:* Spatially separated blocks, a single-object occlusion event, a box-and-rod partial occlusion. *Left-to-Right:* frames progress through time.

A summary of our method, including a video complement to our evaluation, can be found at: *https://bit.ly/2Z8HecW*.

II. Related Work

A. Computational Accounts of Intuitive Physics

The method by which infants acquire a model of intuitive physics has been hotly contested dating back to Piaget's initial theory of cognitive development in the infant [41, 42]. From a computational perspective, recent work has identified similarities between the approximate, probabilistic reasoning capabilities of infants and the physics engines found in modern day simulators [5, 7, 49]. Several approaches in the literature have focused on training application-specific models of physics [2, 31, 55]. In some cases, a partial or complete physics simulator can be learned directly from data [6, 11] and used for many different subsequent control tasks [23, 32]. An exciting alternative direction explores the use of probabilistic program induction for one-shot classification tasks [29, 28, 30].

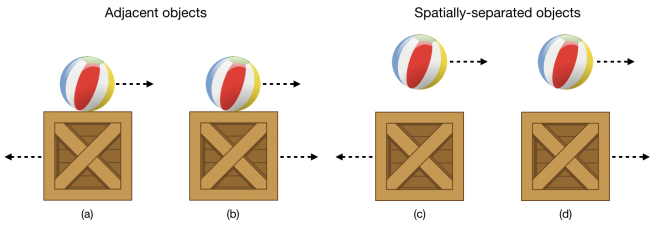

Fig. 2: Adaptation of experiments from Spelke et al. [51]. (a) adjacent objects move relative each other, providing a kinetic cue that they are separate objects; (b) adjacent objects move together, leading infants to perceive a single, unified object; (c), (d) spatially separate objects are perceived as distinct units, regardless of motion.

In contrast, this paper does not explicitly rely on a learned a physics model, but rather employs abductive inference to reason over visual primitives in a manner consistent with a physically-aware infant. Our work is most similar to the work of Teglas et al. [53] in this regard.

B. Abductive Inference

Abductive inference is the logical process of finding an explanation A—preferably the *best* explanation—given an observation B and a general principle $(A \rightarrow B)$ [33, 40]. Abduction differs from both deduction—a conclusion B is guaranteed, given a general principle $(A \rightarrow B)$ and a true observation A—and induction—a general principle $(A \rightarrow B)$, given two individual observations A and B.

For example, when Mary walks barefoot into the yard and feels wet grass under her toes, she can use this observation and knowledge that rain makes the ground wet to infer that it might have rained recently. This explanation is chosen amongst other competing hypotheses, such as sprinklers.

Traditionally, abductive reasoning has been structured as a logical programming problem, in which a logical explanation Δ is recovered for a set of observations Γ, such that:

$$\Sigma \wedge \Delta \models \Gamma \tag{1}$$

where Σ represents domain-specific background knowledge. Logical abduction, which has garnered interest from both the symbolic AI [4, 22, 24, 37] and machine learning communities [13, 35], has been used for a wide variety of perception tasks, such as visual scene understanding [1, 14, 52], image/video interpretation [21, 54], and concept learning [25]. In the specific case of robotic perception, Σ represents background knowledge regarding the effects of the robot's actions on the world and the impact of changes in the world on future observations [50].

Competing accounts have instead posed abductive reasoning as a probabilistic task [12, 15, 16, 36, 43, 44, 45, 46], using Bayesian networks [39] over random variables instead of discrete symbols. We adopt a similar interpretation of abduction, defining a hierarchical Bayesian model that supports reasoning over visual primitives at varying levels of abstraction.

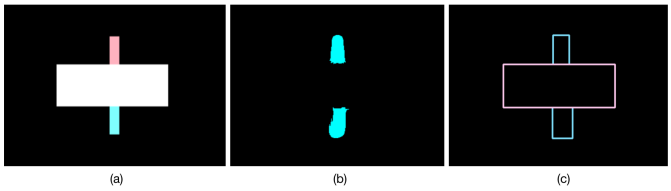

Fig. 3: A visualization of static and motion segmentation: (a) A frame from the box-and-rod environment; (b) A sample motion mask; (c) Segmentation boundaries: motion segmentation in blue and the static segmentation in pink.

III. Approach

Our framework for probabilistic physical inference reasons abductively over a set of visual primitives. Object segmentation extracts a set of visual entities, which are combined to form candidate objects in a scene. Candidate objects are then used to populate a hypothesis space over individual and relational primitives describing each object's motion and interactions with other candidate objects.

A. Object Segmentation

The findings of Spelke et al. [51] suggest that infants individuate objects using both spatial and kinetic cues. In the simplest case, spatial gaps in 2D images constitute strong evidence that the best explanation of a scene contains multiple distinct objects. When objects are adjacent, however, and spatial information is ambiguous, kinetic information takes precedence during individuation, as shown in Figure 2. To this end, we use both motion and static segmentation to identify candidate objects in a scene.

a) Motion Segmentation: Given a set of input image frames $I = \{I_0, I_1, ..., I_n\}$ of dimension $H \times W$, we construct a set of motion masks $M = \{M_0, M_1, ..., M_n\}$ following the method of Pathak et al. [38]. M_i represents a binary segmentation of video frame I_i, where each pixel $p_{x,y} = 1$ if it is undergoing significant motion, or $p_{x,y} = 0$ otherwise.

b) Static Segmentation: We then perform static segmentation using input image frames I and motion masks M, resulting in a set of *entities* $E = \{E_m \cup E_s\}$ from which candidate objects can be constructed.

Due to noise in the segmentation process and the possibility of objects interacting in different ways on a frame-by-frame basis (i.e. contact, occlusions, etc.), we cannot assume that every entity $e \in E$ is an individual object. Therefore, we initialize a set of candidate objects $O = \mathcal{P}(E)$, considering all possible combinations of entities E. To fight exponential complexity in the number of candidate objects, unlikely hypotheses are aggressively pruned. A visualization of the segmentation process is shown in Figure 3. Figure 3 displays a frame from the box-and-rod PhysSprite scene, its motion mask, and the entities resulting from both static and motion segmentation.

B. Reasoning over Visual Primitives

Our goal is to identify a hypothesis $h \in H$ that adequately explains the observations an agent receives from two suc-

cessive image frames $\{I_{t-1}, I_t\} \in I$. We define a set of visual primitives $V = \{V_{ind} \cup V_{rel}\}$ that can be used to describe objects in a scene individually, as well as interactions between objects. In the present 2D case, V_{ind} consists of motion primitives describing an object's movement in one of the four cardinal directions (or lack thereof) and V_{rel} covers occlusion interactions between two or more objects.

Using candidate objects O from the previous segmentation step and visual primitives V, we initialize a hypothesis space H, where each hypothesis $h \in H$ encodes the number of objects in the scene, their positions, and an explanation composed of V. We model this stochastic process as a hierarchical Bayesian network [39], propagating belief in accordance with the following transition and observation models:

a) Transition Model: We model the passage of time as an increase in entropy:

$$P(\hat{H}_t) = \sum_{H_{t-1}} P(H_t|H_{t-1})P(H_{t-1}) = P(H_{t-1}) \circledast f(\bullet) \quad (2)$$

where $\circledast$ denotes convolution with an entropy kernel $f(\bullet)$. We also assume a constant linear velocity model for all objects in the scene, following Newton's equations of motion:

$$\hat{x}_t = x_{t-1} + v_{t-1}\Delta t \quad (3)$$

where x_{t-1} and v_{t-1} denote the object's position and velocity at the previous time-step, respectively.

b) Observation Model: The Bayesian network reasons over visual primitives that either directly involve object motion or those that result in a temporary change of perceived object motion. Thus, optical flow provides a useful representation with which to compare hypotheses with incoming observations. Given input frames $\{I_{t-1}, I_t\}$ and their corresponding flow map $F_t = \textit{flow}(I_{t-1}, I_t)$, we synthesize an image $\hat{I}_h$ for each hypothesis $h \in H$ and compute a hypothetical flow map $\hat{F}_t = \textit{flow}(I_{t-1}, \hat{I}_h)$. The likelihood is then be computed as the inverse of the ℓ_2-loss between the ground truth optical flow and hypothetical flow maps:

$$P(d|h) = \frac{1}{\mathcal{L}^{\ell_2}(F_t, \hat{F}_t)} = \frac{1}{(F_t - \hat{F}_t)^2} \quad (4)$$

where d denotes an incoming observation.

Finally, we compute a posterior distribution over hypotheses via Bayes' Theorem:

$$P(h|d) = \frac{P(d|h)P(h)}{\Sigma_{h' \in H} P(d|h')P(h')} \propto P(d|h)P(h) \quad (5)$$

where each $h \in H$ is a specific hypothesis composed of primitives $v \in V$ pertaining to objects $o \in O$, $P(d|h)$ is the aforementioned observation model describing the likelihood of d, and $P(h)$ is the prior distribution over hypotheses. Unlike a supervised learning algorithm that may learn the posterior $P(h|d)$ directly, we instead shift focus to defining an observation model $P(d|h)$ that effectively describes the likelihood of the input signal, given an $h \in H$. Figure 4 illustrates belief propagation alongside optical flow for single-object and multi-object scenes.

C. Relational Primitives

We extend the model presented in Section III-B to handle relational primitives explaining interactions between two or more objects. Such interactions differ slightly from individual motion primitives, unfolding over the course of a longer sequence of frames. Computationally, however, our interest remains in reverse-engineering the physical reasoning displayed by infants. To this end, we again make use of the relative motion cues provided by optical flow to compute the likelihood of each hypothesis $h \in H$.

At the abstract relational level, the observation model $P(d|h)$ differs in a few important ways. First, it maintains a running motion similarity score over a horizon of T time-steps to account for longer duration relational primitives. Next, in the case of occlusion, we are no longer interested in the similarity between a hypothesized movement $\hat{F}_t$ and the ground truth flow F_t, but rather the difference between an unobstructed motion path $\hat{F} = \{\hat{F}_0, \hat{F}_1, ..., \hat{F}_t\}$ of a candidate object $o \in O$ and the object's true path $F = \{F_0, F_1, ..., F_t\}$. In this case, we no longer need to invert the ℓ_2-loss term, as discontinuities contribute as positive evidence towards an occlusion hypothesis. These steps result in the updated observation model:

$$P(d|h) = \sum_{1}^{T} \mathcal{L}^{\ell_2}(F_t, \hat{F}_t) = \sum_{1}^{T} (F_t - \hat{F}_t)^2 \quad (6)$$

Finally, though relational primitives occur over the course of many time-steps, their explanatory power only takes effect when objects are actually *interacting*. For example, if two objects exist in a particular scene, such as in Figure 1, the probability that one object is occluded does not carry weight if the objects are separated by a wide margin on the image plane. We introduce the following condition on $P(d|h)$ to ensure the posterior over relational primitives is not adjusted until one or more objects intersect:

$$P(d|h) = \begin{cases} \sum_1^T \mathcal{L}^{\ell_2}(F_t, \hat{F}_t), & \text{if } \{o_A, o_B\} \in O \text{ intersect.} \\ 1.0, & \text{otherwise.} \end{cases} \quad (7)$$

In sum, at each time-step, the relational model receives as input a ground truth flow map F_t and the *maximum a posteriori* (MAP) estimate $\hat{h}_{MAP} = \text{argmax}_h P(h|d)$ from the lower-level Bayesian model and computes a posterior following (5) as before with updated observation model (7).

IV. Preliminary Evaluation

Synthetic video datasets have become popular in the representation learning community for providing a simple yet informative testbed for visual learning algorithms [9, 17, 18, 19]. We follow a similar approach with the PhysSprites dataset.

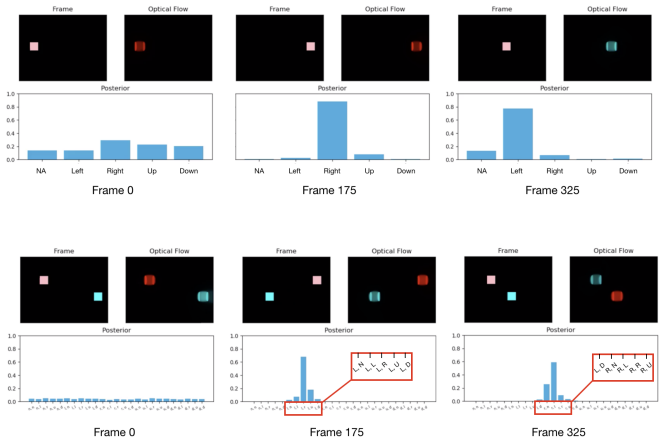

Fig. 4: Sample results for the single-object and multi-object cases.

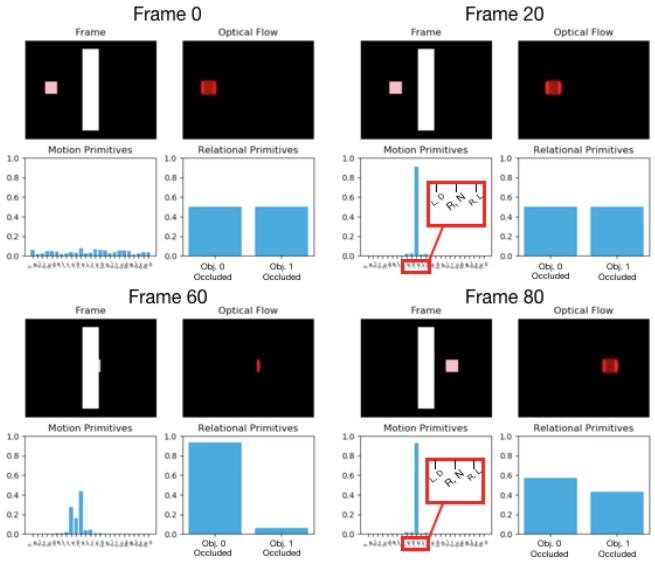

Fig. 5: Belief at various frame intervals throughout a difficult single-object occlusion environment.

PhysSprites, modeled after the work of Matthey et al. [34], contains synthetic recreations of environments used in classical infant studies, testing our model's ability to mirror their results for key traits of intuitive physics. Examples include box-and-rod displays to measure relative motion cues [3, 26] and both single- and multi-block displays testing spatiotemporal continuity [48]. Altogether, the PhysSprites dataset consists of 30 unique video scenes with an average of 642 frames.

a) Individual Primitives: Results show our model effectively reasons over individual primitives regarding object motion. Figure 4 demonstrates the reasoning ability of our model in both single-object motion and multi-object motion cases. Each subsection of Figure 4 displays a sample frame, a flow map corresponding to that frame, and the posterior over motion primitives. Of note are Frames 175 and 325 of the upper part of Figure 4, which show an object move to the right, change course, and head back to the left. Below the images and flow maps, the MAP estimate of the posterior over hypotheses correctly identifies the motion primitive that explains these observations. We find similar results for the multi-object case on the bottom-half of the figure. In the bottom half of Figure 4, the magnified letters are shorthand for visual primitives pertaining to each object. For example, L, R represent the hypothesis that object 0—the blue cube—is moving to the left and object 1—the pink cube—is moving right.

b) Relational Primitives: We additionally evaluate our model's ability to reason over visual primitives that involve interactions between multiple objects. Consider Figure 5, for example, which displays the most difficult occlusion scene in the PhysSprites dataset. In this single-object occlusion scene, the pink block (object 0) disappears completely behind the rectangular occluder (object 1). Notice that the increase in belief that object 0 becomes occluded (relational primitives between frames 20 and 60) corresponds precisely to the decrease in belief regarding the motion of the object. Our model is effectively finding an explanation for the missing object and quickly reassigning belief once it reappears.

V. Discussion and Future Work

We have presented a hierarchical Bayesian framework for reasoning abductively over visual primitives and, through qualitative evaluation, have identified a candidate model for a snapshot of the infant development cycle. We also provided the PhysSprites dataset, which served as the test-bed for preliminary evaluations of our model.

The design of our model offers multiple directions for further research. First, due to its ability to reason over visual primitives, we are optimistic about the prospects of integrating our model with representation learning techniques that map raw visual inputs (i.e. images, videos) to discrete symbols. Further, an intriguing direction for future work will consider alternative definitions of the *best* hypothesis. Possible alternatives include the most *simple* explanation and the most *informative* explanation [27]. Finally, rather than surfacing frame-level hypotheses, we would like our model to abduce a single explanation for a sequence of frames. For example, when asked to explain the movement of the square block in row two of Figure 1, a human observer is not likely to respond: "the block moved right one pixel, the block again moved right one pixel, etc". Instead, the observer would respond: "the block moved right until it disappeared behind the larger block, then appeared on the other side". We will extend this work to aggregate hypotheses into abstract explanations.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1646417. We are grateful for this support.

References

[1] Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. Visual commonsense for

scene understanding using perception, semantic parsing and reasoning. In *2015 AAAI Spring Symposium Series*, 2015.

[2] Pulkit Agrawal, Ashvin V Nair, Pieter Abbeel, Jitendra Malik, and Sergey Levine. Learning to poke by poking: Experiential learning of intuitive physics. In *Advances in Neural Information Processing Systems*, pages 5074–5082, 2016.

[3] Renee L Baillargeon, Elizabeth S. Spelke, and Stanley Wasserman. Object permanence in five-month-old infants. *Cognition*, 20:191–208, 1985.

[4] Chitta Baral. Abductive reasoning through filtering. *Artificial Intelligence*, 120(1):1–28, 2000.

[5] Christopher Bates, Peter Battaglia, Ilker Yildirim, and Joshua B Tenenbaum. Humans predict liquid dynamics using probabilistic simulation. In *CogSci*, 2015.

[6] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. In *Advances in neural information processing systems*, pages 4502–4510, 2016.

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

[8] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[9] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in beta-vae. *arXiv preprint arXiv:1804.03599*, 2018.

[10] Susan Carey and Fei Xu. Infants' knowledge of objects: beyond object files and object tracking. *Cognition*, 80: 179–213, 2001.

[11] Michael B Chang, Tomer Ullman, Antonio Torralba, and Joshua B Tenenbaum. A compositional object-based approach to learning physical dynamics. *arXiv preprint arXiv:1612.00341*, 2016.

[12] Eugene Charniak and Solomon Eyal Shimony. *Probabilistic semantics for cost based abduction*. Brown University, Department of Computer Science, 1990.

[13] James A Crowder and John N Carbone. Abductive artificial intelligence learning models. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, pages 90–96. The Steering Committee of The World Congress in Computer Science, Computer , 2017.

[14] Krishna SR Dubba, Anthony G Cohn, David C Hogg, Mehul Bhatt, and Frank Dylla. Learning relational event models from video. *Journal of Artificial Intelligence Research*, 53:41–90, 2015.

[15] Didier Dubois, Angelo Gilio, and Gabriele Kern-Isberner. Probabilistic abduction without priors. *International Journal of Approximate Reasoning*, 47(3):333–351, 2008.

[16] M Julia Flores, José A Gámez, and Serafín Moral. Abductive inference in bayesian networks: finding a partition of the explanation space. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 63–75. Springer, 2005.

[17] Will Grathwohl and Aaron Wilson. Disentangling space and time in video with hierarchical variational auto-encoders. *arXiv preprint arXiv:1612.04440*, 2016.

[18] Irina Higgins, Loic Matthey, Xavier Glorot, Arka Pal, Benigno Uria, Charles Blundell, Shakir Mohamed, and Alexander Lerchner. Early visual concept learning with unsupervised deep learning. *arXiv preprint arXiv:1606.05579*, 2016.

[19] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, volume 3, 2017.

[20] Irina Higgins, Nicolas Sonnerat, Loic Matthey, Arka Pal, Christopher P Burgess, Matko Bosnjak, Murray Shanahan, Matthew Botvinick, Demis Hassabis, and Alexander Lerchner. Scan: Learning hierarchical compositional visual concepts. *arXiv preprint arXiv:1707.03389*, 2017.

[21] Jerry R Hobbs and Rutu Mulkar-Mehta. Using abduction for video-text coreference. In *Proceedings of BOEMIE 2008 Workshop on Ontology Evolution and Multimedia Information Extraction*, 2008.

[22] Jerry R Hobbs, Mark E Stickel, Douglas E Appelt, and Paul Martin. Interpretation as abduction. *Artificial intelligence*, 63(1-2):69–142, 1993.

[23] Michael Janner, Sergey Levine, William T Freeman, Joshua B Tenenbaum, Chelsea Finn, and Jiajun Wu. Reasoning about physical interactions with object-oriented prediction and planning. *arXiv preprint arXiv:1812.10972*, 2018.

[24] John R Josephson, B Chandrasekaran, Jack W Smith, and Michael C Tanner. A mechanism for forming composite explanatory hypotheses. *IEEE Transactions on Systems, Man, and Cybernetics*, 17(3):445–454, 1987.

[25] Antonis C Kakas and Fabrizio Riguzzi. Abductive concept learning. *New Generation Computing*, 18(3): 243–294, 2000.

[26] Philip J. Kellman and Elizabeth S. Spelke. Perception of partly occluded objects in infancy. *Cognitive Psychology*, 15(4):483 – 524, 1983. ISSN 0010-0285. doi: https://doi.org/10.1016/0010-0285(83)90017-8.

[27] JHP Kwisthout. Two new notions of abduction in bayesian networks. 2010.

[28] Brenden Lake, Ruslan Salakhutdinov, and Joshua Tenenbaum. Concept learning as motor program induction: A large-scale empirical study. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34, 2012.

[29] Brenden Lake, Chia-ying Lee, James Glass, and Josh Tenenbaum. One-shot learning of generative speech

concepts. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36, 2014.

[30] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[31] Adam Lerer, Sam Gross, and Rob Fergus. Learning physical intuition of block towers by example. *arXiv preprint arXiv:1603.01312*, 2016.

[32] Yunzhu Li, Jiajun Wu, Russ Tedrake, Joshua B Tenenbaum, and Antonio Torralba. Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids. *arXiv preprint arXiv:1810.01566*, 2018.

[33] Claire Liang, Julia Proft, Erik Andersen, and Ross A Knepper. Implicit communication of actionable information in human-ai teams. 2019.

[34] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

[35] Raymond J Mooney. Integrating abduction and induction in machine learning. In *Abduction and Induction*, pages 181–191. Springer, 2000.

[36] Rohit J Kate Raymond J Mooney. Probabilistic abduction using markov logic networks.

[37] Charles G Morgan. Hypothesis generation by machine. *Artificial Intelligence*, 2(2):179–187, 1971.

[38] Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2701–2710, 2017.

[39] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988. ISBN 0-934613-73-7.

[40] Charles S. Peirce. The Proper Treatment of Hypotheses: a Preliminary Chapter, toward an Examination of Hume's Argument against Miracles, in its Logic and in its History. MS [R] 692. 1901.

[41] Jean Piaget. *The construction of reality in the child*. Routledge, 2013.

[42] Jean Piaget and Margaret Cook. *The origins of intelligence in children*, volume 8. International Universities Press New York, 1952.

[43] David Poole. Representing bayesian networks within probabilistic horn abduction. In *Uncertainty Proceedings 1991*, pages 271–278. Elsevier, 1991.

[44] David Poole. Learning, bayesian probability, graphical models, and abduction. In *Abduction and Induction*, pages 153–168. Springer, 2000.

[45] Rajat Raina, Andrew Y Ng, and Christopher D Manning. Robust textual inference via learning and abductive reasoning. In *AAAI*, pages 1099–1105, 2005.

[46] Jan-Willem Romeijn. Abducted by bayesians? *Journal of Applied Logic*, 11(4):430–439, 2013.

[47] Elizabeth S. Spelke, Karen Breinlinger, Janet Macomber, and Kristen Jacobson. Origins of knowledge. *Psychological review*, 99:605–32, 11 1992. doi: 10.1037/0033-295X.99.4.605.

[48] Elizabeth S. Spelke, Roberta Kestenbaum, Daniel Simons, and Debra Wein. Spatiotemporal continuity, smoothness of motion and object identity in infancy. *British Journal of Developmental Psychology*, 13:113–142, 06 1995. doi: 10.1111/j.2044-835X.1995.tb00669.x.

[49] Adam N Sanborn, Vikash K Mansinghka, and Thomas L Griffiths. Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological review*, 120(2):411, 2013.

[50] Murray Shanahan. Perception as abduction: Turning sensor data into meaningful representation. *Cognitive science*, 29(1):103–134, 2005.

[51] Elizabeth S Spelke, Claes von Hofsten, and Roberta Kestenbaum. Object perception in infancy: Interaction of spatial and kinetic information for object boundaries. *Developmental Psychology*, 25(2):185, 1989.

[52] Jakob Suchan, Mehul Bhatt, Przemysław Wałega, and Carl Schultz. Visual explanation by high-level abduction: On answer-set programming driven reasoning about moving objects. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[53] Erno Teglas, Edward Vul, Vittorio Girotto, Michel Gonzalez, Joshua B Tenenbaum, and Luca L Bonatti. Pure reasoning in 12-month-old infants as probabilistic inference. *science*, 332(6033):1054–1059, 2011.

[54] Paul Thagard and Cameron Shelley. Abductive reasoning: Logic, visual thinking, and coherence. In *Logic and scientific methods*, pages 413–427. Springer, 1997.

[55] Zhihua Wang, Stefano Rosa, Bo Yang, Sen Wang, Niki Trigoni, and Andrew Markham. 3d-physnet: Learning the intuitive physics of non-rigid object deformations. *arXiv preprint arXiv:1805.00328*, 2018.