Learning to See the Physical World

Jiajun Wu (Stanford University; jiajunwu@cs.stanford.edu)
DOI: 10.1145/3465074.3465081

Introduction

I am fascinated by how rich and flexible human intelligence is. From a quick glance at the scenes in Figure 1A, we effortlessly recognize the 3D geometry and texture of the objects within, reason about how they support each other, and when they move, track and predict their trajectories. Stacking blocks, picking up fruits—we also plan and interact with scenes and objects in many ways.

The goal of my dissertation research is to build machines that see, interact with, and reason about the physical world just like humans (Wu, 2020). This problem, physical scene understanding, involves three key topics that bridge research in computer science, AI, robotics, cognitive science, and neuroscience:

- **Perception** (Figure 1B): How can structured, physical object and scene representations arise from raw, multi-modal sensory input (e.g., videos, audios)?

- **Physical interactions** (Figure 1C): How can we build dynamics models that quickly adapt to complex, stochastic real-world scenarios, and how can they contribute to planning and motor control? Modeling physics helps robots to build bridges from a single image and to play games such as Jenga.

- **Reasoning** (Figure 1D): How can physical models integrate structured, often symbolic, priors such as symmetry and use them for commonsense reasoning?

Physical scene understanding is challenging, because it requires a holistic interpretation of scenes and objects, including their 3D geometry, physics, functionality, and modes of interactions, beyond the scope of a single discipline such as computer vision. Structured priors and representations of the physical world are essential: we need proper representations and learning paradigms to build data-efficient, flexible, and generalizable intelligent systems that understand physical scenes.

Our approach to constructing representations of the physical world is to integrate bottom-up recognition models, deep networks, and efficient inference algorithms, with top-down, structured graphical models, simulation engines, and probabilistic programs. In the dissertation, we develop and extend techniques in these areas (e.g., proposing new deep networks and physical simulators); we also explore ways to combine them, building upon studies across vision, learning, graphics, and robotics, with inspiration from cognitive science and neuroscience. Only by exploiting knowledge from all these areas and disciplines, may we build machines that have human-like, physical understanding of complex, real-world scenes.

Perception

Motivated by human perception—rich, generalizable, data-efficient—my research on perception focuses on building structured, object-based models to characterize the appearance...
and physics of objects.

My dissertation research covers various components of the appearance model. On bottom-up recognition, we have developed a general pipeline for 3D shape reconstruction from a single color image (Wu, Wang, et al., 2017) via modeling intrinsic images—depth, surface normals, and reflectance maps (Figure 2A). Our research is inspired by the classic research on multi-stage human visual perception (Marr, 1982), and has been extended to integrating learned priors of 3D shapes (i.e., ‘what shapes look like?’) for more realistic 3D reconstructions, and to tackling cases where the object in the image is not from the training categories.

Complementary to these bottom-up recognition models, we have also explored learning top-down graphics engines directly. We have proposed 3D generative adversarial networks, first applying generative-adversarial learning to 3D shapes for shape synthesis. We have later extended the model as visual object networks, which synthesize object shape and texture simultaneously, enforcing various consistencies with a distributed representation for object shape, 2.5D sketches, viewpoint, and texture (Figure 2B) (Zhu et al., 2018). We have generalized our models to scenes, recovering structured scene representations that not only capture object shape and texture, but enable 3D-aware scene manipulations (Figure 2C) (Yao et al., 2018).

Beyond object appearance, intuition of object physics assists humans in scene understanding. We have developed computational models that learn to infer object physics directly from visual observations. The Galileo model marries a physics engine with deep recognition nets to infer physical object properties (e.g., mass, friction). With an embedded physical simulator, the Galileo model discovers physical properties simply by watching objects move in unlabeled videos; it also predicts how they interact based on the inferred physical properties. The model was tested on a real-world video dataset, Physics 101, of 101 objects interacting in various physical events.

The dissertation also involves integrating geometry and physics perception (Figure 2D), with two primary results as “physical primitive decomposition” (PPD) and “visual de-animation” (VDA) (Wu, Lu, et al., 2017). In PPD, we decompose an object into parts with distinct geometry and physics, by learning to explain both the object’s appearance and its behaviors in physical events; in VDA, our model learns to jointly infer physical world states and to simulate scene dynamics, integrating both a physics engine and a graphics engine. Our recent work has extended these models to complex indoor scenes, exploiting stability for more accurate 3D scene parsing.

Physical Interactions

My dissertation research also includes learning to approximate simulation engines (forward models) themselves. We have explored building physical models in various forms—image-based, object-based, and particle-based; analytical, neural, and hybrid—and have demonstrated their power in challenging, highly underactuated control tasks (Figure 3). Compared with off-the-shelf simulators, a learned dynamics simulator flexibly adapts to novel environments and captures the stochasticity in scene dynamics. Our visual dynamics model demonstrates this in the pixel domain, where it learns to synthesize multiple possible future frames from a single color image by automatically discovering independent movable parts and their motion distributions (Xue et al., 2016) (Figure 3A). We have later extended the
A. Modeling visual dynamics allows us to generate multiple possible future frames from a single image (Xue et al., 2016). B. We have developed a hybrid model that captures object-based dynamics by integrating analytical models and neural nets. It assists the robot in accomplishing a highly underactuated task: pushing the right disk to the target (green) by only interacting with the left disk (Ajay et al., 2019). C. D. Particle-based dynamics models support controlling soft robots (Hu et al., 2019) and manipulating deformable objects and liquids (Li et al., 2019).

We have also explored the idea of learning a hybrid dynamics model, augmenting analytical physics engines with neural dynamics models (Ajay et al., 2019) (Figure 3B). Such a hybrid system achieves the best of both worlds: it performs better, captures uncertainty in data, learns efficiently from limited annotations, and generalizes to novel shapes and materials. These dynamics models can be used in various control tasks: they help to solve highly underactuated control problems (pushing disk A, which in turn pushes disk B to the target position), to control and co-design soft robots (Hu et al., 2019), to manipulate fluids and rigid bodies on a Kuka robot (Li et al., 2019), and to interact and play games such as Jenga that involve complex frictional micro-interactions.

Reasoning
The physical world is rich but structured: natural objects and scenes are compositional (scene are made of objects which, in turn, are made of parts); they often have program-like structure (e.g., symmetry). We have also been exploring ways to bridge structured, often symbolic, priors into powerful deep recognition models.

A test of these neuro-symbolic representations is how well they support solving various reasoning tasks such as analogy making and question answering. Our recent work demonstrated that, when combined with deep visual perception modules, a symbolic reasoning system achieves impressive performance on visual reasoning benchmarks, outperforming end-to-end trained neural models. We have also extended it to jointly learn visual concepts (e.g., colors, shapes) and their correspondence with words from natural supervision (question-answer pairs) via curriculum learning, without human annotations (Mao, Gan, Kohli, Tenenbaum, & Wu, 2019).

Next Steps
With big data, large computing resources, and advanced learning algorithms, the once separated areas across computer science (vision, learning, symbolic reasoning, NLP, rule learning and program induction, planning, and control) has begun to reintegrate. We should now take an more integrative view towards these areas and actively explore their interactions for a more general AI landscape.

One such direction is to achieve more fundamental integration of perception, reasoning,
and planning. While most computational models have treated them as disjoint modules, we observe that having them communicate with each other facilitates model design and leads to better performance. Another direction is to integrate symbolic priors with deep representation learning via program synthesis for concept and structure discovery. Neuro-symbolic methods enjoy both the recognition power from neural nets and the combinatorial generalization from symbolic structure; therefore, they have great potential in scaling up current intelligent systems to large-scale, complex physical scenes in the real life, for which pure bottom-up, data-driven models cannot work well due to the exponentially increasing complexity. Beyond physical objects and scenes, I also want to build computational models that understand an agent's goals, beliefs, intentions, and theory of mind, and use these knowledge for planning and problem solving, drawing inspiration from intuitive psychology.

References


Jiajun Wu is an Assistant Professor of Computer Science at Stanford University, working on computer vision, machine learning, and computational cognitive science. Before joining Stanford, he was a Visiting Faculty Researcher at Google Research. He received his PhD in Electrical Engineering and Computer Science at Massachusetts Institute of Technology. Wu's research has been recognized through the AAAI/ACM SIGAI Doctoral Dissertation Award, the ACM Doctoral Dissertation Award Honorable Mention, the MIT George M. Sprowls PhD Thesis Award in Artificial Intelligence and Decision-Making, the 2020 Samsung AI Researcher of the Year, the IROS Best Paper Award on Cognitive Robotics, and fellowships from Facebook, Nvidia, Samsung, and Adobe.