Learning to See the Physical World

Jiajun Wu

Stanford University

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.

My research goal is to build machines that see, interact with, and reason about the physical world just like humans. This problem, which we call **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, compositional, physical object and scene representations arise from raw, multi-modal sensory input (e.g., video, audio, tactile signals)?
- **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 physical interactions helps robots to build bridges from a single image and to play challenging games such as Jenga.
- **Reasoning** (Figure 1D): How can physical models integrate structured, often symbolic, priors such as categorization, hierarchy, symmetry, and repetition, 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.

My 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 **symbolic programs**. In my research, I develop and extend techniques in these areas (e.g., proposing new deep networks and physical simulators); I further explore innovative ways to combine them, building upon studies across vision, learning, graphics, and robotics. I believe that only by exploiting knowledge from all these areas, may we build machines that have a human-like, physical understanding of complex, real-world scenes.

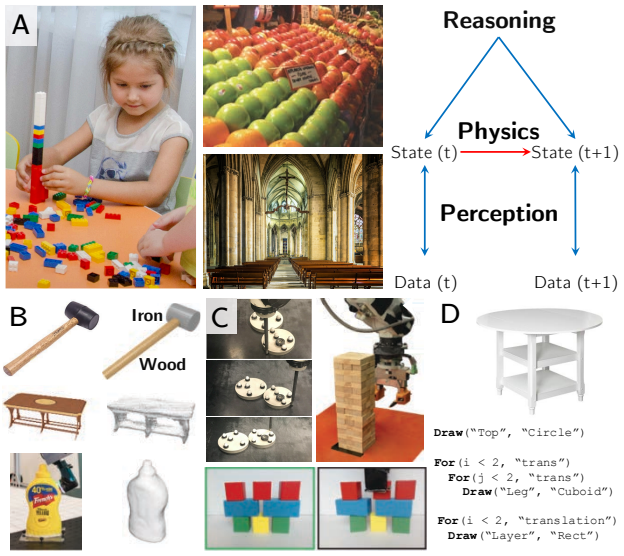

Figure 1: **Physical scene understanding** involves **(I) perception**, building physical object representations from multi-sensory data, **(II) physical interaction**, capturing scene dynamics for planning and control, and **(III) commonsense reasoning**, understanding high-level, structured, often symbolic priors in objects and scenes.

My research is also highly interdisciplinary: I build computational models with inspiration from human cognition, developmental psychology, neuroscience, robotics, and computational linguistics; I also explore how these models can, in turn, assist in solving tasks in these fields.

My new faculty highlight talk will survey my research experience and future plans on these three research topics.